Statistical Pose Estimation of Land Targets in SAR

Jose C. Principe, Dongxin Xu, Andrew Learn*, Qun Zhao

Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611
E-mail: principe@cnel.ufl.edu

*Sensors Directorate of the Air Force
Research Laboratory
Wright-Patterson Air Force Base
Dayton, OH 45433

ABSTRACT

This paper explores statistical pose estimation in SAR ATR using a recently proposed training method based on information theory. The theory of training with information theoretic learning is briefly summarized. Different pose estimator topologies and training criteria are employed. Experimental results in the MSTAR I and II show that our proposed method is capable of producing 1-DOF and 2-DOF pose estimations and we show the dependence of the training parameters on performance.

1. INTRODUCTION

The information of the relative position of a target with respect to the sensor, termed the aspect angle of the observation or the pose, is important for automatic target recognition (ATR). The aspect angle naturally divides the complexity of the input space, because targets display many different signatures as a function of the aspect angle. Therefore, a reliable methodology to estimate pose will provide some extra information that will simplify the design of classifiers and improve their performance. In SAR ATR pose estimation is not a simple task due to the enormous variability of the scattering phenomenology.

One possible way to estimate the pose is to perform first point scatter detection followed by a match to a large database of stored targets [1]. This method may be accurate but it is very time consuming. Alternatively some features of the target can be identified such as the leading bright edge of the target and its shadow and pose computed from them [2]. These methods have the problem of using heuristics to determine the features used for pose estimation.

We have devised a way to automatically select the best features for the proposed task by minimizing an information theoretic measure of performance [3]. We define features as projections to a subspace (Figure 1), but unlike PCA (principal component analysis) our projections are allowed to be nonlinear. In essence we are implementing information filtering, because we are removing selectively parts of the input image that are irrelevant for our goal, just like we do in linear filtering operations. But unlike linear filtering, this method can be nonlinear and it is preserving information (in a theoretical sense), unlike optimal filtering that only achieves optimization of error variance (second order

statistics)

Figure 1. Feature extraction as an optimal mapping that preserves information.

This paper briefly describes our pose estimation methodology and shows results obtained in the MSTAR I and II data sets. We address one degree of freedom (1-DOF) and 2-DOF pose estimation. For the 1DOF case we show the accuracy of the method in terms of different target mixtures in training, different training criteria and also different network topologies. Finally, we show that a more realistic 2-DOF pose estimator can be trained with very good results by simply increasing the output space of the network.

2. Renyi's Entropy and Potential Energy

Let $a_i \in R^m, i = 1, \dots, N$, be a set of samples from a random variable $Y \in R^m$ in m-dimensional space. One interesting question is what will be the entropy associated with this set of data points. One answer lies in the estimation of the data pdf by the Parzen Window method using a Gaussian kernel:

$$f_Y(y) = \frac{1}{N} \sum_{i=1}^N G(y - a_i, \sigma^2 I) \quad (1)$$

where $G(\cdot, \cdot)$ is the Gaussian function, σ^2 is the variance, and $I \in R^{m \times m}$ is the identity matrix. When Shannon's entropy is used along with this pdf estimation, the computations become very complex. Fortunately, Renyi's entropy with order 2 may lead to a simpler form. Generally, Renyi's entropy with order α (differential entropy for continuous random variable) is:

$$H_{R\alpha}(Y) = \frac{1}{1-\alpha} \log \int f_Y(y)^\alpha dy, \alpha > 0, \alpha \neq 1 \quad (2)$$

If $S(Y)$ is the Shannon's differential entropy for Y , then $\lim_{\alpha \rightarrow 1} H_{R\alpha}(Y) = S(Y)$, and $H_{R\beta}(Y) \geq S(Y) \geq H_{R\gamma}(Y)$, for

$0 < \beta < 1$ and $1 < \gamma$. So, Shannon's entropy can be viewed as one member of Renyi's entropy family. When $\alpha = 2$, Renyi's entropy $H_{R2}(Y)$ is also called quadratic entropy. Combining eq.1 and eq.2, the entropy measure for a set of discrete data points $H_{R2}(\{a_i\})$ becomes:

$$\begin{cases} H_{R2}(\{a_i\}) = -\log V(\{a_i\}) \\ V(\{a_i\}) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G(a_i - a_j, 2\sigma^2 I) \end{cases} \quad (3)$$

Making the analogy between data samples and “information particles”, $V(\{a_i\})$ can be regarded as the overall potential energy since $G(a_i - a_j, 2\sigma^2 I)$ can be taken as the potential energy of “particle” a_i in the potential field of “particle” a_j , or vice versa. *So, maximizing entropy is in this case equivalent to minimizing the information potential. The information potential $V(\{a_i\})$ can be thought of as a cost function that needs to be minimized to achieve maximum entropy.* The derivative of the potential is a force, in this case an information force that acts on the samples to repel or attract them (depending if the goal is to minimize or maximize the information potential). Taking the derivative of (3) we obtain the information force as $F = -G(a_i - a_j, 2\sigma^2 I)(a_i - a_j)/(2\sigma^2)$.

Suppose that the problem is one of finding a mapping: $R^n \rightarrow R^m: Y = g(X, w)$ where w is a set of unknown parameters such that the entropy in the output space $H_{R2}(\{a_i\})$ is maximized. If we consider the points a_i as the outputs of the mapping $a_i = g(b_i, w)$ when the given input data are $b_i \in R^n$, $i = 1, \dots, N$ the problem is to find w such that the information potential in the output space is minimized. In this case, following the idea of “error back-propagation” [4], the forces will be back-propagated to each parameter in w according to the chain rule:

$$\frac{\partial}{\partial w} V(\{a_i(w)\}) = \sum_{i=1}^N \frac{\partial V}{\partial a_i} \frac{\partial a_i}{\partial w} \quad (4)$$

i.e. we obtain a general, nonparametric, and sample-by-sample methodology to adapt arbitrary nonlinear (smooth) mappings based on entropy maximization [5].

Mutual Information is more useful than entropy for engineering problems because it basically computes the distance between two pdfs. We introduced in [6] an approximation to the Kullback-Leibler divergence based on the Euclidean difference (ED-QMI) of vectors inequality

$$I_{ED}(Y_1, Y_2) = \left(\iint f_{Y_1 Y_2}(z_1, z_2)^2 dz_1 dz_2 \right) + \left(\iint f_{Y_1}(z_1)^2 f_{Y_2}(z_2)^2 dz_1 dz_2 \right) - 2 \left(\iint f_{Y_1 Y_2}(z_1, z_2) f_{Y_1}(z_1) f_{Y_2}(z_2) dz_1 dz_2 \right) \quad (5)$$

Obviously, $I_{ED}(Y_1, Y_2) \geq 0$ and equality holds if and only if Y_1 and Y_2 are statistically independent, so it is also a divergence. Basically (5) measures the Euclidean distance between the joint pdf and the factorized marginals. With the previous definitions it is not difficult to obtain

$$I_{ED}((Y_1, Y_2)|y) = V_{ED}(y) \\ V_{ED}(y) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N V_{ij}^1 V_{ij}^2 - \frac{2}{N} \sum_{i=1}^N V_i^1 V_i^2 + V^1 V^2 \quad (6)$$

where V_{ij} is the information potential of each information particle in the joint space, V_i is the information potential of each particles in the marginal potential field and V^1 is the marginal information potential. It is also not difficult to obtain the formula for the calculation of the information force produced by $V_{ED}(y)$ as

$$c_{ij}^k = V_{ij}^k - V_i^k - V_j^k + V^k, \quad k = 1, 2 \\ F_i^l = \frac{\partial V_{ED}}{\partial y_i^l} = \frac{-1}{N^2 \sigma^2} \sum_{j=1}^N c_{ij}^k V_{ij}^l d_{ij}^l \\ i = 1, \dots, N, \quad l \neq k, \quad l = 1, 2 \quad (7)$$

where c_{ij}^k are cross matrices which serve as force modifiers.

3. A Statistical Formulation for Pose Estimation

Suppose that we have a set of data in pairs $\{(x_i, \theta_i), i = 1, \dots, N\}$, where $x_i \in R^m$ is a vector reformulated from a two-dimensional image and θ_i is a vector of ground truth information relative to the image contents. Here we restrict the pose estimation to a one degree of freedom (1-DOF) problem, where x_i is a SAR image of a land vehicle obtained at a given depression angle, and θ_i is the azimuth (aspect angle) of the vehicle.

Generally, the pose estimation given a particular image $x_i \in R^m$ can be formulated as a *MAP* (maximum a posteriori probability) problem, i.e.

$$\hat{\theta} = \arg \max P(\theta_i | x) \quad (8)$$

where $P(\theta_i | x)$ is the *a posteriori* probability of the aspect angle θ given x . From Equation (8), we can see that the

best estimation of the aspect angle given x is the one which maximizes the *a posteriori* probability. It is known that a multilayer perceptron trained with the back-propagation algorithm may be viewed as a practical method for performing a nonlinear input-output mapping of a general nature. To be specific in the problem of pose estimation, let x denote the m -dimensional input image vector, and let θ denote the 2-dimensional output vector, then the input-output relationship of the network defines a mapping from a m -dimensional Euclidean input space to a 2-dimensional Euclidean output space, which is infinitely continuously differentiable.

The *priori* knowledge about the aspect angle is usually unknown, therefore, according to Jayne's MaxEnt principle [7], the uniform distribution is a reasonable assumption about the probability density function (pdf) of aspect angle θ . By Bayes theorem, Equation (8) can be written as

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax} \frac{P(\theta_i) f_{X|\theta}(x|\theta_i)}{f_X(x)} \\ &= \operatorname{argmax} f_{X|\theta}(x|\theta_i)\end{aligned}\quad (9)$$

where $P(\theta_i)$ is the *a priori* probability of the aspect angle, $f_{X|\theta}(x|\theta_i)$ is the conditional probability density function of x given a particular aspect angle θ_i , and $f_X(x)$ is the marginal pdf of x . Equation (9) implies that the problem of pose estimation becomes one of estimating the conditional pdf of images x for all the possible azimuth θ_i , $i=1, \dots, N$.

Since x is a very high dimensional vector, any assumption about its pdf form is probably not appropriate for pose estimation. So, dimensionality reduction (feature extraction) is necessary in this case, and a non-parametric method should be used for the pdf estimation. Let $\varphi; R^m \rightarrow R^p$ be a mapping (feature extractor), $p \ll m$, then $y = \varphi(x, W)$ is a feature set for x , where W is the parameter set of the feature extractor. The pose estimation problem (8) can be rewritten as:

$$\begin{aligned}\hat{w} &= \operatorname{argmax} P(w_i|y) \\ y &= \varphi(x, W)\end{aligned}\quad (10)$$

4. Network Structure and Learning Criterion

The key problem now is how to construct the mapping $\varphi; R^m \rightarrow R^p$. Neural networks have been successfully used as tools for functional mapping. In this paper, both the perceptron and a linear network are employed. For training a MLP, the mean square error (MSE) is the most popular criterion, which produces the famous error back-propagation algorithm [4].

From the information theoretic point of view, a mapping, which can be regarded as an information transmission chan-

nel, should convey as much information as possible about the input. Here our concern are the information of aspect angles, so y should best represent this information. A quantitative measure suitable for this purpose is mutual information between y and θ_i [6], then the criterion for parameter selection of W can be formulated as

$$W_{opt} = \underset{i}{\operatorname{argmax}} I(y, \theta_i) \quad (11)$$

where $I(y, \theta_i)$ is the mutual information between y and θ_i . Equation (11) means that the optimal set should be the one which maximizes the mutual information between the feature and the aspect angle. In this paper, a perceptron or a linear network (64,000 inputs and 2 outputs) are employed to implement the mapping $y = \varphi(x, W)$. Now the problem can be described as finding the parameters W of the network so that the mutual information between its output and the aspect angle is maximized (Figure 2). For more detailed information, please see [6].

5. Experimental Results

The proposed algorithm and estimation method are validated in the MSTAR/IU Database. In this paper, we are interested in finding out the effect on pose estimation performance of the following three factors: training set, training criterion, and the training network topology. To give a quantitative measure of the results, two variables are employed, the mean and standard deviation of the estimation error, are defined as follow

Figure 2. Pose estimator using the maximization of mutual information.

a. Error mean: mean of pose estimation error

$$\begin{aligned} Error(i) &= |Aspect\ true(i) - Aspect\ estimate(i)| \\ mean &= \frac{1}{n} \sum_{i=1}^n Error(i) \end{aligned}$$

b. Error std: standard deviation of estimation error

$$std = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Error(i) - Errormean)^2}$$

5.1 Training set

In the MSTAR II Database, there are eight classes of targets (T72, BTR-60, 2S1, Brdm_2, D7, T62, Zil131, Zsu_23_4), and for some, the SAR images are collected at four depression angles (15, 17, 30, and 45 degree). How to determine the training set is a problem to be considered first. We study the effect on performance of choosing all the target types or just some types, using training data of single depression angle or of multiple depression angles. For 1-

DOF pose estimation, one task is to find out how the algorithm is able to generalize across different depression angles. Some of our results are given in Table 1 and Table 2. In Table 1, a linear network is trained with data from the T72 class (different serial numbers) at a single depression angle, i.e., 17, 30 or 45 degrees, respectively. The double star in the table means testing with the training vehicle (which obtains the lower error). It is observed that the best depression angle for training is 30 degree, which is mid way between the extremes. This is a compromise between the features captured from 15 and 45 degrees elevation, and produces a maximum pose error of 6.39 degrees.

Table 1: Linear network trained by single depression angle (17, 30 and 45 degree, respectively)

Target Type at depression (deg.)	mean/std Training: T72 at depression:17 deg.	mean/std Training: T72 at depression:30 deg.	mean/std Training: T72 at depression:45 deg.
T72_A04 (15)	3.82 / 5.36	4.05 / 4.12	7.27 / 4.89
T72_A05 (15)	6.29 / 3.78	4.74 / 3.15	7.30 / 5.12
T72_A04 (17)	1.24 / 1.39 **	3.77 / 2.86	6.59 / 5.91
T72_A05 (17)	4.14 / 2.89	3.98 / 2.82	6.07 / 4.33
T72_A64_1(30)	6.29 / 4.76	1.06 / 1.21 **	3.88 / 2.93
T72_A64_2(30)	6.78 / 5.61	3.32 / 2.36	4.71 / 3.60
T72_A64_1(45)	35.19 / 24.45	6.39 / 5.17	1.34 / 1.86 **
T72_A64_2(45)	21.96 / 16.68	5.15 / 3.77	4.41 / 2.92

In Table 2, the network is trained by target type at multiple depression angles (one and two types of targets in the training set). We conclude that training with a mixed of angles (second column) improves performance (largest error for the T72 is now 4.47 degrees). Including a different vehicle in the training set does not help performance for the T72s, but allows the system to generalize better for the other target types (e.g. the error for the zsu_23_4c decreased from 8.52 to 7.59 degrees). If we remember that the pose is known with an accuracy of only 3.5 degrees for all these

vehicles we can conclude that the pose estimator is generalizing rather well. We also tested the robustness of the pose estimator with respect to occlusion (Figure 4) and showed its superior performance due to global features [8].

Table 2: Linear network trained by single depression angle (T72 at 30 degree), mixed angles(T72 at 15 and 45 degree), and mixed types and angles T72 at 15, 45 degree and Brd_2 at 30 degree)

Target Type (Depression Angle:deg)	mean/std Train: T72 at 30 deg.	mean/std Train: T72 at 15 & 45 deg.	mean/std Train: T72 at:15 & 45, and Brd2 at 30 deg.
T72_A04 (15)	4.05 / 4.12	4.04 / 3.03	3.68 / 3.02
T72_A05 (15)	4.74 / 3.15	4.02 / 3.09	4.13 / 2.94
T72_A04 (17)	3.77 / 2.86	3.87 / 3.00	3.56 / 2.74
T72_A05 (17)	3.98 / 2.82	3.69 / 2.83	3.75 / 2.68
T72_A64_1(30)	1.06 / 1.21 **	3.16 / 2.36	2.75 / 2.04
T72_A64_2(30)	3.32 / 2.36	4.17 / 3.10	3.91 / 2.71
T72_A64_1(45)	6.39 / 5.17	3.39 / 2.93	2.84 / 2.45
T72_A64_2(45)	5.15 / 3.77	4.47 / 3.12	4.84 / 3.38
d7 (17)	10.10 / 9.73	8.12 / 6.28	6.97 / 6.30
brdm_2 (17)	5.24 / 4.03	5.63 / 4.72	4.69 / 3.77
zsu_23_4 (17)	6.17 / 6.19	6.91 / 5.72	6.41 / 6.99
brdm_2c(30)	8.70 / 5.74	7.25 / 5.48	5.04 / 3.79
zsu_23_4c(30)	7.67 / 5.38	8.52 / 6.10	7.59 / 5.02

The circle in the left panel of the figure shows the training set performance for the bmp2, while the triangles show the

Figure 3. Results of occlusion. Left panel: pose estimate for train (diamonds) and test (triangles). Middle panel: change of pose estimate with occlusion increment. Right panel: occlusion at increment 36 (90%).

performance of the pose estimator when an image cut from the background is superimposed on the vehicle. The pose estimate degrades smoothly until step 40 (central panel) which corresponds to 90% object occlusion (shown at the

right panel). We verified that the occlusion direction also affects the break down point of the pose estimator.

5.2 Training criterion

We propose to estimate the joint probability density function (PDF) by training the pose estimator to maximize the mutual information (MI) between its output and the known pose. However, the training is time consuming since the MI algorithm complexity is proportional to the square of the number of training samples ($O(N^2)$). Therefore we want to find simpler ways to train the pose estimator. We found out in the experiments that the output of the network trained with MI always stabilizes to a circle, which suggests utilizing this output as the desired response for pose estimation (left panel of Figure 3). Once the desired response is available, training the network with the well established MSE criterion is also possible. The advantage of MSE method is that the computational complexity becomes $O(N)$. We experimentally compared the network performance trained with the MI and MSE criterion in Table 3. It is observed that the two criteria perform at the same level, so we recommend training with the MSE method due to its simpler computational complexity. But we have to remember that the desired response (the circle) was “discovered” by the results obtained with the MI criterion.

Table 3: Linear network trained by MSE (compare with Table 2)

Target Type (Depression Angle:deg)	mean/std Training: T72 at depression:30 deg.	mean/std Training: T72 at depression: 15 & 45 deg.
T72_A04 (15)	5.57/4.57	4.03/3.43
T72_A05 (15)	5.73/4.56	4.13/3.20
T72_A04 (17)	4.99/4.18	4.34/3.41
T72_A05 (17)	4.60/3.65	4.60/3.65
T72_A64_1(30)	2.58/2.13	4.25/3.40
T72_A64_1(45)	12.85/8.48	2.14/1.96
T72_A64_2(45)	6.96/6.06	4.51/3.74
d7 (17)	12.12/11.38	9.81/7.55
brdm_2 (17)	6.73/5.62	7.35/5.52
brdm_2c(30)	8.85/6.42	11.75/8.18
zsu_23_4c(30)	7.99/6.29	9.51/7.32

5.3 Training network topology

For our pose estimation task, we compared the effect of including a nonlinearity in the network topology, that is we trained a perceptron (Table 4). From this table, we found that the performance of the linear network is as good as that

of the perceptron. The reason might be that in such a high-dimensional input space (80x80), a linear network is able to preserve all the pose information. Therefore we suggest the use of a linear network due to its faster and more reliable training.

Table 4: Perceptron trained by back-propagation (compare with Table 2)

Target Type	mean/std Training: T72 at depression: 30 deg.	mean/std Training: T72 at depression: 15 & 45 deg.
T72_A04 (15)	3.73 / 5.11	4.32 / 3.83
T72_A05 (15)	4.39 / 2.84	4.15 / 3.35
T72_A04 (17)	3.91 / 3.20	4.35 / 3.44
T72_A05 (17)	3.89 / 2.89	4.20 / 3.34
T72_A64_1(30)	1.08 / 1.46	4.15 / 3.11
T72_A64_1(45)	8.32 / 9.04	1.54 / 2.23
T72_A64_2(45)	5.40 / 4.31	5.07 / 3.68
d7 (17)	14.77 / 12.62	8.91/7.10
brdm_2 (17)	7.37 / 5.36	6.43 / 4.86
brdm_2c(30)	12.52 / 8.20	10.42 / 7.55
zsu_23_4c(30)	9.45 / 7.84	9.13 / 6.60

5.4 2-DOF pose estimation

Finally we would like to check the performance of the pose estimator in a more difficult 2-degree of freedom pose estimation (azimuth and elevation) [9]. The additional degree of freedom was incorporated by adding a third output node to the network. This produces a three-dimensional output space and generates a cylindrical representation surface for azimuth and elevation, as shown in Figure 4. Azimuth is represented by a unit circle in two dimensions as before, and elevation is represented by height in the third dimension.

Three targets in the MSTAR data set were used for training and testing: 2S1, Brdm-2, and Zsu-23/4. (These were the only targets that included 15, 30, and 45 degree elevation coverage.) The training set incorporated 2S1 images from 0-180 degrees, separated by approximately 3.5 degrees. Images of all three vehicles made up the test set. Test results for all three vehicles are shown in Table 5. The output space for the 2S1 test is shown in Figure 5. Figure 5 clearly shows the ability of the network to distinguish between elevations. Three distinct circles appear, one for each elevation that was tested. Table 5 indicates that the network also generalizes well. While performance is better for the 2S1

(the vehicle used for training) as expected, pose estimation errors are low for the other vehicle types as well.

Figure 4. Represen-

Figure 5. Test results with the statistical pose estimator for the target 2S1. Dots are training set, while triangles rep-

6. Conclusions

This paper reports our results of pose estimation for MSTAR/IU database, based on our proposed method of pose estimation using information theoretic features [2], [6]. The need for such an effective algorithm stems from our goal of

creating accurate and robust classifiers, because knowing the pose of a target will streamline the size and training of the classifier module.

Table 5: Azimuth and elevation test results

Target type	Azimuth mean/std	Elevation mean/std
2S1	2.33/1.91**	1.58/1.23**
Brdm_2	5.80/5.30	3.33/2.40
zsu_23_4	5.02	4.52/2.48

The preliminary results of our method show a maximum mean error of 8 degrees of azimuth, when the training images are taken at 3.5 degree increments. This result is obtained when the training vehicle is NOT in the training set, and the elevation of the test images are different from the training set. This is comparable to other methods, but here the advantage is that the result is obtained with very little computation (the advantage of adaptive systems is that all the hard work is done in the system training that is off-line). By comparison, we think that a linear network trained with MSE criterion is suitable to perform the task of pose estimation. The 2-DOF examples show that the system is able to deal with two degrees of freedom quite easily, and even improved the specificity. Further work on this approach is warranted by these results.

Acknowledgments: This work was partially supported by a DARPA -IU grant F33615-97-1-1-19.

REFERENCE

- [1] Wissinger J., Ristroph R., Diemunsch J., Severson W., Fruedenthal E., "MSTAR's extensible search engine and model based inferencing toolkit", Proc. SPIE, vol 3721, 554-571, 1999.
- [2] Voicu L., Patton R., Myler H., "Multicriterion vehicle pose estimation for SAR ATR", Proc. SPIE, vol 3721, 497-506, 1999.
- [3] Fisher J., Principe J., "Entropy manipulation of arbitrary nonlinear mappings", Proc. IEEE Workshop Neural Nets for Sig Proc. VII, 14-23, 1997.
- [4] Rumelhart D., Hinton G., Wiliams R., "Learning internal representations by error backpropagation", Nature, vol 323, 533-536, 1986.
- [5] Principe J., Xu D., Fisher J., "Information theoretic Learning", in Unsupervised adaptive filtering, Haykin (Ed.), Wiley, 2000 (in press).
- [6] Xu D., "Energy, entropy and information potential in neurocomputing", Ph.D. dissertation, U. of Florida, 1998.
- [7] Jaynes E., "Information theory and statistical mechanics", Phys. Rev., vol 106, 620-630, 1957.
- [8] Principe J., Information Theoretic Feature Extraction for SAR/ATR, Progress report, AirForce grant F 33615-97-1-1-19, 1999.
- [9] Learn, Andrew W. "Target Pose Estimation from Radar Data Using Adaptive Networks". MS thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 1999