AUTOMATIC TARGET RECOGNITION IN LASER RADAR IMAGERY

Magnús Snorrason*, Harald Ruda and Alper Caglayan**
Charles River Analytics
55 Wheeler St., Cambridge, MA 02138
mss@cra.com / (617)491-3474x524

ABSTRACT

This paper presents an Automatic Target Recognition (ATR) system for laser radar (LADAR) imagery, designed to classify objects at multiple levels of discrimination (target detection, classification, and recognition) from single LADAR images. Segmentation is performed in both the range and non-range LADAR channels and results combined to increase object detection rate or decrease false positive detection rate. Through use of the range data, object subimages are projected and rotated to canonical orientations, providing invariance to translation, scale and rotations in 3-D. Global features are extracted for rapid target detection and local receptive field features are computed for target recognition. 100% detection and recognition rates are shown for a small set of real LADAR data.

1. INTRODUCTION

The fundamental question addressed in this research is how to adapt the traditional computer vision approach of segmentation-feature-extraction-classification to imagery produced with a LADAR sensor.

The problem at hand is to automatically classify objects in a battlefield that are imaged from above. The objects must be classified at multiple levels of discrimination: such as, "man-made vs. natural object"; if man-made, then "building vs. mobile target"; if mobile, then "tank vs. other vehicle"; if tank, then "T72, M60, ZSU, or BMP".

LADAR is a promising sensor technology for this application since it provides high resolution 3-D information about the scene, as well as traditional 2-D images. Depending on the sensor design, a LADAR sensor can collect three channels of data: range, active (infrared reflectance), and passive (infrared emittance) data. Range data is a measure of physical distance, unlike the other two channels which measure energy; therefore it is not affected by illumination from other energy sources, internal heat of the target, or weather.

The novelty of our approach is in the combined use of range and non-range data to generate segmented image objects which are invariant to rotation and translation in 3-D. Features computed from these objects are then used as inputs to a hierarchical classifier based on the Fuzzy-ARTMAP neural network [1]. Figure 1 shows a block diagram of the segmentation, projection, feature extraction, and hierarchical classification components that have been implemented in the *Khoros* image processing/graphical software development environment. Not shown in this figure are various knowledge bases for exogenous information and decision fusion. To summarize the key points of our work:

- Segmentation of objects from their background using information from both the range and non-range (active or passive) channels, allowing both a logical "OR" of the information sources to decrease the chance of missing a target in segmentation, and a logical "AND" to decrease the number of false positive detections.
- Production of an orthogonal set of 2-D "virtual views" (invariant to rotation and translation in 3-D) from a single LADAR view using the known imaging geometry to construct an inertial coordinate system and computing images as seen from a virtual observer which can be moved within that coordinate system
- Hierarchical classification of object signatures to first determine whether the object is a potential target or not (target detection), then determine the general class of each target object, such as building, bridge, or land vehicle (target classification), and finally to recognize the target object within each class.

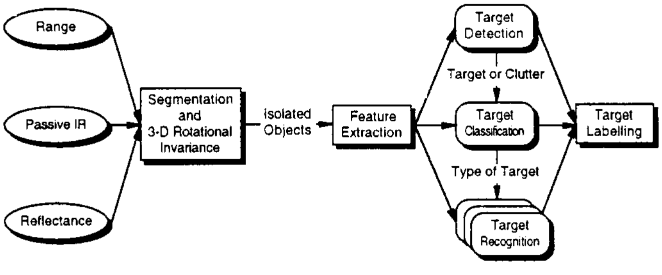

Figure 1: System level block diagram showing main data streams from the LADAR sensor to a list of labelled objects.

2. VIRTUAL VIEWS

The requirements for constructing a spherical coordinate system containing both the scene and the sensor (inertial system) from a single image are to know the range ρ to each pixel in the field-of-view, the depression viewing angle θ for each row, and the azimuth viewing angle ψ for each column. The first requirement is met by definition if the image is produced with LADAR or some other range sensor. The other two requirements are generally fulfilled by recording the instantaneous orientation of the sensor with respect to the horizon (for θ) and with respect to the heading of the sensor or of the carrier on which the sensor is mounted (for ψ).

Once a spherical (ρ,θ,ψ) coordinate system has been produced, it is simple to transform to the Cartesian coordinates of (x,y,z). It is particularly convenient to work in Cartesian coordinates for producing a top view (looking down parallel to the z axis), since orthographic projections can be computed for that view without using trigonometric functions.

* Member IEEE. ** Senior Member IEEE.

0-7803-2431-5/95 $4.00 © 1995 IEEE

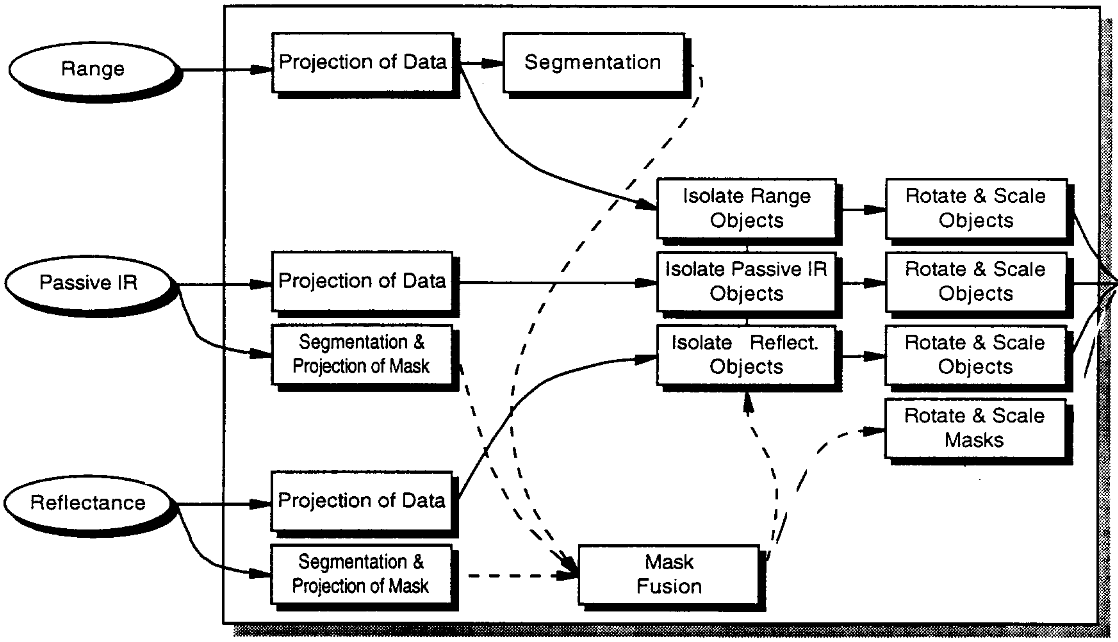

Figure 2: Details of the Segmentation and 3-D Rotational Invariance block from figure 1. Solid lines indicate continuous valued data, dashed lines indicate binary data. This architecture holds for any projection; section 3 discusses the case of projection to top view.

Since a LADAR sensor is an active sensor, it can be considered both illuminant and observer. Therefore, the coordinates of the illuminant, the observer, and every illuminated point in the scene are known. This is sufficient information to produce a "virtual" view by moving the observer to a new location in the coordinate system and recomputing the view as seen from that point.

Unlike the observer, the illuminant cannot be moved after the data has been collected, so any occluded or unlit areas in the original view will remain invisible in all virtual views, producing "data shadows". Finally, the virtual view will typically contain new occlusions which obscure data points that were visible in the original view. This is the data loss associated with the given projection.

3. SEGMENTATION AND PROJECTION TO TOP VIEW

Figure 2 shows how the range, passive emittance and active reflectance streams are processed separately. Depending upon the design of the sensor, one or more of these channels will be available, and it must be possible to tailor the algorithm to make full use of the sensor. The segmentation of the image is performed in each stream, as that may be the only stream available. However, in those cases where more than one channel is available, the channels can be combined in order to improve performance beyond that of a single channel.

In terms of segmentation, the range channel is treated differently from the reflectance and emittance channels. The range channel provides *location* data that is fundamentally different from the data provided by the other two channels, which provide *energy* images. The location data becomes the basis of segmentation by height and the energy images are segmented with more traditional techniques. In short, the range channel extracts and projects z coordinate values, then performs segmentation; the other channels perform segmentation and then the projection before combining with the range segmentation.

The range channel uses segmentation by height. By using gray scale to code height, it is possible to locate the z coordinate of the ground plane by peak detection in an appropriately preprocessed histogram. The ground plane is then removed by thresholding at a fixed height above it. In this interpretation, anything that projects significantly out of the ground is an object. This segmentation method produces very accurate object outlines, but it can produce a large number of false positive objects since it does not discriminate between man-made and natural objects.

The other two channels use a multi-scale dispersion segmentation method that was developed after empirical analysis of LADAR infrared reflectance images. The regions of those images that contain man-made objects such as military vehicles have reflectance values that are either significantly higher or lower than average; or they have highly varying reflectance characteristics. Computing dispersion [2] over a small window size (3x3 pixels) emphasizes the latter, while the dispersion measure for a large window (11x11 pixels) emphasizes the former. One image is produced for each window size, and the images are added and thresholded. Preliminary analysis indicates that this method works for infrared emittance images as well.

There are at least two alternatives for combining segmentations in the "Mask Fusion" block. If the aim is to find the maximum number of possible objects, then the separate segmentations are "OR"ed together and then cleaned up with morphological operations. It is also possible to eliminate some false-positive detections by correlating objects in more than one channel; then the channels are cleaned up before being "AND"ed together. If all three channels are available, combinations of these methods can be used.

Rotational invariance is achieved by rotating each object around its center of gravity in the image plane to one of four canonical positions. Similar approaches have been used in previous object recognition systems [3, 4] based on the implicit assumption that object orientation has no discriminant quality for separating object classes. In the original oblique view, this is not true in general. For example, upright human beings seen from any other angle than straight above tend to be taller than they are wide. Consequently, in any view other than top view, vertical orientation of the object's major axis could be an indicator for detecting standing or moving humans. In the top view, how-

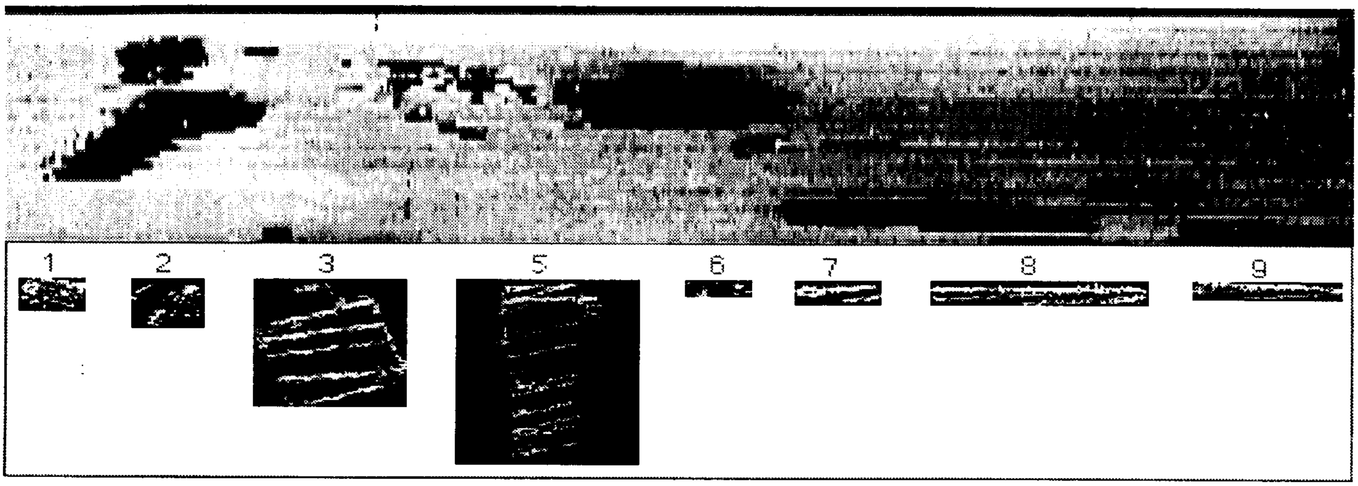

Figure 3, top: LADAR image showing five buildings and a few trees. Gray scale codes histogram equalized *z* coordinates with "darker" representing "higher". Bottom: Eight segmented objects shown in virtual top view after rotation to canonical orientation. Objects 1, 2, and 6 are trees, object 3 is center building, 5 is building at far left, 7, 8, and 9 are buildings at lower right.

ever, the assumption is valid since the orientation of the object's major axis represents the direction along the ground surface in which the object is facing, and in general that is arbitrary.

Object recognition can be done based on just the top view, using objects such as those shown in figure 3. However, seeing the object "in elevation" (i.e., from side, front, or back), provides information which is complementary to the information gained from the top view. It also counteracts any problems of data shadows associated with the top view. With the addition of one other perpendicular canonical view, rotational invariance can be extended from the 2-D domain to 3-D by training the classifier only on the canonical top and side views of each object and ignoring all other 3-D orientations.

Figure 2 applies to a projection to any virtual view using the location data. Our present implementation uses the top view exclusively. The location data is used in many places to perform this top-projection, but the system as shown will still work if there is no location data available. In that case, the projection becomes an identity operation and all images remain in the original oblique view (see section 5 for an example).

4. FEATURE EXTRACTION

The next step after segmentation is to transform subimages, each containing one segmented image object, into a form which can be used as input by the neural classifier. This must be a 2-D to 1-D transform since classifiers work with vectors but not with matrixes. The transform should also compress the information, since even a small subimage produces a vector with thousands of elements if every pixel is preserved. In addition to these basic requirements, the transform should maximize the difference between objects of different classes while minimizing the difference between objects of the same class.

Since the range data provides a direct measure of physical size, a number of useful features can be computed from the binary object mask produced by the "Mask Fusion" block in figure 2. The most obvious are area, length, width, eccentricity, and higher order moments. Using gray scale to code height, features such as mean height, maximum height, and standard deviation of height are computed. The passive infrared and active reflectance channels provide two more sets of features that can be interpreted as thermal and color features respectively. More esoteric statistics, such as fractal dimension and entropy are also computed from the grayscale codes.

All of those features are "global" in the sense that each one is a function of the whole object. Local receptive field features are computed with a spatial kernel that is significantly smaller than the object. We have found that local features, although computationally more expensive, are necessary for good performance at the discrimination level of target recognition. Gabor kernels produce highly compressed representations of objects, capturing all at once: local average of gray scale values, local spatial frequency, and the local orientation of contrast gradients (lines and edges). As shown in the next section, target recognition based on Gabor features performs very well when used with reflectance data and no projection.

However, Gabor features are not as effective for data that has been projected to a virtual view, due to data shadows separating scan lines. Classification should be invariant to the overall orientation of scan lines, just as it is invariant to 3-D orientation of objects. However, any kernel that is sensitive to local orientation is also sensitive to the orientation of scan lines in subimages such as shown in figure 3. Our solution is to use coarse coding [5] with unoriented kernels (circularly symmetric Gaussians), normalized for the number of valid data points. This approach retains spatial relationships of locally averaged gray scale values, without requiring the data shadows to be filled in or interpolated.

5. CLASSIFICATION

The hierarchical classifier consists of a set of Fuzzy-ARTMAP neural networks [1], each trained with a specific subset from the total set of extracted features. It is hierarchical because that allows the recognition problem to be broken down into manageable components: first separate target objects from clutter objects with one classifier; then divide the target objects into classes such as ground vehicles, aircraft, buildings, and bridges, with another classifier; then subdivide each class into recognizable targets such as tanks and other vehicles, with one classifier per target class. Each classifier adds its prediction for a given object to the object's label and the confidence for that prediction (based on Fuzzy-ARTMAP's "category choice"). Interpreting the confidence as probability, an example of one object's final label might be: P(target) = 0.93, P(ground vehicle | target) = 0.78, P(tank | ground vehicle) = 0.88.

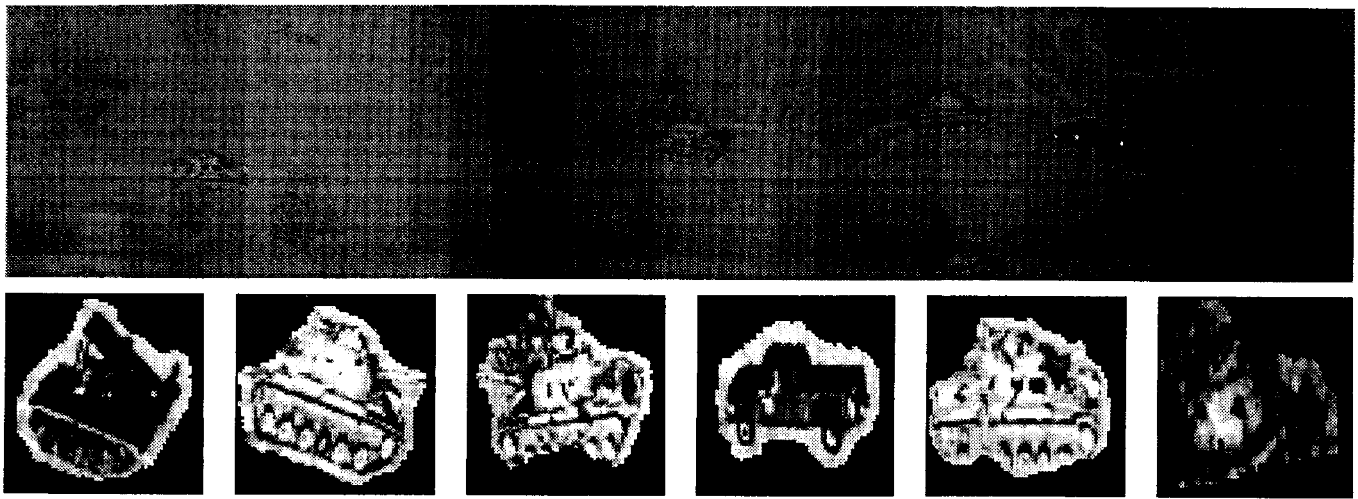

Figure 4, top: LADAR image (gray scale coding reflectance) showing 4 tanks and 2 other vehicles. Bottom: Each target segmented, rotated and scaled, but not projected.

The classification results reported here come from a data set of 9 active reflectance images containing a total of 44 objects, 20 ground vehicles and 24 clutter objects. Figure 4 shows one of the images and the isolated target objects from that image. Note that this data set is only using one of the three channels shown in figure 2; in particular the range data is not used, so no projections are performed. The rotation to canonical orientation is therefore relative to the original viewing angle and therefore does not give true 3-D rotational invariance. This data set is not adversely affected because the vehicles are all lined up close to a side view.

Half the objects were randomly chosen for training and the other half for testing. There were 20 targets: 13 tanks, 3 trucks, 2 Armored Personnel Carriers (APC), 1 missile launcher, and 1 artillery unit. Non-targets included trees, patches of ground, image artifacts, etc. This separation into targets and non-targets is precisely the task of *target detection.*

Starting with a set of 164 features and iteratively retraining and testing a Fuzzy-ARTMAP neural network classifier with different subsets of features, we were able to reach 100% correct target detection on the test set. As few as five features suffice to yield this performance, or as many as 38 can be used. However, adding more features to the set of 38 degrades performance. The reasons are the small size of the training set and that each feature is given an equal *a priori* probability or importance in influencing the outcome of the classification, so adding uncorrelated features which do not substantially help the classification dilutes the effect of the more useful features.

The five features that constitute the minimal set yielding 100% correct classifications are:

1. Eccentricity of object shape
2. Standard moment m_{01} (x coordinate of shape centroid)
3. Maximum gray level
4. Kurtosis (4th order statistic) of gray levels
5. A measure of fractal dimension of gray levels

We also implemented a *target recognition* level classifier for separating the set of objects determined to be targets by the target detector described above into tanks and non-tanks. Non-tanks here include APCs, trucks, missile launcher, and artillery unit. This classifier also used a Fuzzy-ARTMAP neural network and thus learned the training set perfectly.

On the test set, in this case 10 objects (six tanks and four others), the system achieved 100% performance using a set of local Gabor features. The use of oriented local features is appropriate here, since the images are in the original unprojected view, circumventing the problem with separated scan lines discussed in section 4. We used a sampling of 9 locations (3x3 grid) in each subimage, and 12 orientations per location for a total of 108 features. Each feature has a sine and cosine pair that was converted to a phase and magnitude pair; the phase component was then discarded.

Previous work with the same data set using a hybrid neural network / expert system [6] has shown that both target detection and recognition are non-trivial in this data, due to complexity of clutter, varying contrast polarity between objects and background, and partial occlusions of many objects.

REFERENCES

[1] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., Rosen, D.B (1992). Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Trans. on Neural Networks,* 3(5), 698-713.

[2] Jain, A.K. (1989). *Fundamentals of Digital Image Processing,* Englewood Cliffs, NJ: Prentice-Hall.

[3] Zhou, Y.T. and Hecht-Nielsen, R. (1993). Target Recognition Using Multiple Sensors. *Neural Networks for Signal Processing III, Proc. of the 1993 IEEE-SP Workshop.* 411-420, Piscataway, NJ: IEEE Service Center.

[4] Carpenter, G.A., Grossberg, S., and Lesher, G.W. (1992). A What-and-Where Neural Network for Invariant Image Preprocessing. *IJCNN-92, Baltimore,* III, 303-308.

[5] Waxman, A.M, Seibert, M., Bernardon, A.M., and Fay, D.A. (1993). Neural Systems for Automatic Target Learning and Recognition. *The Lincoln Laboratory Journal,* 6(1), 77-116.

[6] Snorrason, M., Caglayan, A.K., and Buller, B.T. (1993). Using Self-Organized and Supervised Learning Neural Networks in Parallel for Automatic Target Recognition. *Neural Networks for Signal Processing III, Proc. of the 1993 IEEE-SP Workshop.* 537-546. Piscataway, NJ: IEEE Service Center.