Recognition of 3D Objects in Arbitrary Pose Using a Fuzzy Associative Database Algorithm

Aaron Mavrinac, Xiang Chen, and Ahmad Shawky

*Abstract***— Once the human vision system has seen a 3D object from a few different viewpoints, depending on the nature of the object, it can generally recognize that object from new arbitrary viewpoints. This useful interpolative skill relies on the highly complex pattern matching systems in the human brain, but the general idea can be applied to a computer vision recognition system using comparatively simple machine learning techniques. An approach to the recognition of 3D objects in arbitrary pose relative to the vision equipment with only a limited training set of views is presented. This approach involves computing a disparity map using stereo cameras, extracting a set of features from the disparity map, and classifying it via a fuzzy associative map to a trained object.**

I. Introduction

HUMANS are generally able to recognize 2D shapes, regardless of changes in orientation, scale, or skew, after having seen the shape in one such configuration. This shape recognition has a very wide range of applications, and accordingly, much work has gone into automating it with computers. The basic theory is that shapes can be extracted from otherwise cluttered and cumbersome images, from which some set of quantifiers efficiently describing the shapes can be obtained and compared to known values through some algorithm for classification. The nature of these quantifiers and the classification algorithm are a subject of much research; most use quantifiers *invariant* to the aforementioned transformations (rotation, scale, skew, etc.) such as Fourier descriptors, moment invariants, and Hough transformations, and most use machine learning methods such as fuzzy logic and neural networks for classification.

Humans are also generally able to recognize 3D objects, regardless of their orientation, after having seen a sufficient number of different views (depending, of course, on the nature of the object itself). To generalize from the 2D case, it is possible to automate this process in a similar manner by obtaining quantifiers describing the 3D surface rather than the 2D shape. Such quantifiers can be extracted from *range images*, or in the case of stereo vision, *disparity maps*. However, a single such image gives information only from a certain perspective; this is commonly referred to as 2.5D. To approach full 3D information, range images must be taken from different perspectives around the object. For classification to continue to work as generalized from the 2D case, the sets of quantifiers from each perspective must be combined to fully describe the object, and the classification algorithm must be designed to operate on this type of information.

In this paper, we expand on previous work in object recognition using invariant values on 2D images [5], justifying the selection of proper invariant descriptors for 3D shapes based on disparity maps and modifying the classification scheme to reflect the new object description. The result is a system capable of recognizing a trained object based on a disparity map taken by a stereo camera rig from any view, where training requires only a few different such views.

II. Prior Work

A. 3D Recognition

There are several cases where 2D moment invariants have been used for recognition of 3D objects. In both [10] and [8], moment invariants are computed on a series of intesity images of the object taken from a variety of positions around it; it is demonstrated that with a sufficient number of images and proper handling of the multi-image input in an artificial neural network scheme, 2D moments are applicable to 3D recognition. However, these methods do not examine 3D information about the object directly, and require a large number of explicitly-ordered views to operate. In addition to the cost of capturing these views, objects are not identified from an arbitrary unknown pose.

Methods have also been proposed which operate on invariants of 3D range data. In [9], the concept of computing characteristic vectors of multiple images is extended to range images, allowing for object recognition in arbitrary pose unaffected by illumination. In [6], local feature histograms, invariant to translations and rotations as well as being robust to partial occlusions, are computed directly on range images; recognition is then performed using histogram matching or probabilistic recognition.

There are a number of alternate possibilities which employ other descriptors entirely. One example is [7], in which chromaticity distributions from a variety of images of the object are used to identify the object; this method of recognition, while pose-invariant, is adversely affected by variations in illumination, though the work attempts to alleviate these problems.

B. Neuro-Fuzzy Recognition

Neuro-fuzzy classifiers are used to solve a wide range of recognition problems [19]. In particular, a number of fuzzy

The authors are with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario, Canada (email: {mavrin1,xchen,shawky}@uwindsor.ca).

This research was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

LVQ schemes have been proposed for prototype-based classification and recognition. Methods such as those described in [13], [14], and [15] employ a fuzzy neighbourhood function on training data with specific classes, whereas others, such as [16], attach fuzzy labels to the training data themselves.

Fuzzy associative memory models [12], [17] have been employed to store rules for classifications based on fuzzy LVQ, notably in [5], upon which the system we describe here is based.

III. Preliminary Theory

A. Disparity Map

In order to quantify the 3D shape of an object in a manner useful for recognition, some representation of the shape must be generated by the sensor. A stereo vision system provides data which can be analyzed in a variety of ways to obtain 3D information, but the crucial point, in this case, is for the representation to lend itself to some analog of the 2D work in [5]. Fortunately, a representation exists to which a similar recognition scheme may be applied, and it is in fact relatively easy to obtain.

For the purpose of this description and throughout this work, the following convention is used for the world and image coordinate systems: lowercase x and y represent image coordinates with origin at the upper left corner of the image and positive axes right and down respectively, and uppercase X, Y, and Z represent world coordinates (which, unless otherwise specified, are mutually orthogonal with Z perpendicular to the rectified image planes and have their origin at the optical center of the left camera). Figure 1 illustrates their relationship.

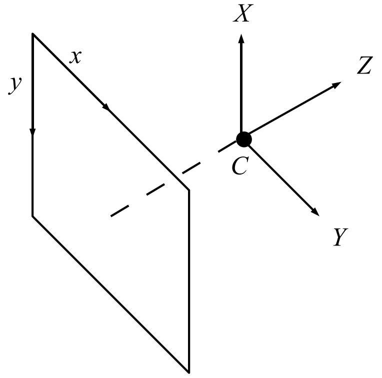


Fig. 1. Coordinate System Convention

We assume a stereo vision system capable of generating rectified stereo images, wherein the epipolar lines are parallel and horizontally aligned as if captured by parallel cameras. In the general case, this requires internal and external (stereo) calibration of the cameras. For a thorough geometrical treatment see [1], [20], and for some practical methods see [2], [3], [4].

Given a pixel of coordinates (x_1, y_1) in one image of an epipolar-rectified stereo pair, and a corresponding pixel (x_2, y_2) in the other (where $y_1 = y_2$), their disparity d is defined as $x_2 - x_1$ [20]. This can be used to triangulate the depth to the original 3D point in the environment (from the optical center of one camera) in the world coordinate system according to the following relation:

$$Z = \frac{bf\lambda}{d} \tag{1}$$

where b is the baseline (distance between the two optical centers), f is the focal length, and λ is a parameter relating the pixel width to real-world measurements.

A *disparity map* is a 2D matrix D containing the disparity of each pixel in one image with respect to the corresponding pixel, if any, in the other. Thus, if pixel (i, j) in the first image corresponds to pixel (k, j) in the second, $D_{ij} = k - i$. The disparity map essentially results in a *range image* when its values are normalized and/or quantized to a range of grayscale values which can be displayed and manipulated as such. This provides an important visualization tool and allows existing image invariant computation algorithms to function unmodified on the data.

With a calibrated stereo vision system, the parameters b, f, and λ are known and Equation 1 may be used to calculate the actual depth (Z coordinate) of the real points represented by pixels in the disparity map. However, for the purposes of 3D object recognition this is not necessary. Instead, the *invariant descriptors* (see Section III-C) are computed from the disparity map, or more specifically, from its associated range image.

B. Correspondence

In order to construct a disparity map for the first image in a stereo pair, it is necessary to establish correspondences in the second image for each pixel in the first. Correlation-based methods such as the sum of square difference (SSD) and normalized cross-correlation (NCC) criteria may be used for this purpose.

Correlation-based correspondence consists of maximizing, for each left-image pixel $\mathbf{p}_l$, a similarity criterion c on the displacement $\mathbf{d} = [d_1, d_2]^T$, selecting $\overline{\mathbf{d}} + \mathbf{p}_l$ as the corresponding right-image pixel.

$$c(\mathbf{d}) = \sum_{k=-W}^{W} \sum_{l=-W}^{W} \psi(I_l(i+k, j+l), I_r(i+k-d_1, j+l-d_2)) \tag{2}$$

In this case, since the images I_l and I_r are rectified and correspondences are therefore found on the same horizontal line, d_1 can be constrained to zero [21]. We use here the SSD criterion for ψ, that is, for two pixel values u and v, $\psi(u, v) = -(u - v)^2$.

C. Invariant Descriptors

We examined a variety of invariant descriptors calculated from 2D images, evaluating their usefulness in describing different range views of an object qualitatively and quantitatively. Three in particular were selected to work collectively to describe a set of range views.

1) Compactness: The first useful descriptor is the *compactness*, a Fourier descriptor which describes a distribution of intensity values in an enclosed region. When applied to a disparity map, it describes the disparity (range) distribution invariant to translation and rotation. The compactness of a greyscale image can be calculated as follows, adapted from [22]:

$$C = \frac{\left(\sum_{y=1}^{h}\sum_{x=1}^{w} f_{\text{boundary}}(x,y)\right)^2}{\sum_{y=1}^{h}\sum_{x=1}^{w} f(x,y)} \qquad (3)$$

where $f(x,y)$ is the value of the image at pixel (x,y) and $f_{\text{boundary}}(x,y)$ defines pixels on the perimeter of a region (object).

2) First Hu Moment: The second descriptor is the first of Hu's seven invariant moments [11], which are invariant to translation, rotation, and scale. Only the lowest-order moment is applied to the disparity maps as it is robust against the inherent noise from imperfect correspondences and occlusions. It is calculated as follows:

$$I_1 = \frac{M_{20} - \overline{x}M_{10} + M_{02} - \overline{y}M_{01}}{M_{00}^2} \qquad (4)$$

where:

$$M_{ij} = \sum_x \sum_y x^i y^i I(x,y) \qquad (5)$$

3) Histogram: The final descriptor is the histogram, a Fourier descriptor which describes the overall distribution of intensities in an image. When applied to a disparity map, it describes rather the range distribution. The histogram is not a scalar value like the previous two descriptors, but may be compared for two different images as follows [6]:

$$\chi^2(I_1, I_2) = \sum_{i=0}^{M} \frac{(h_{1i} - h_{2i})^2}{h_{1i} + h_{2i}} \qquad (6)$$

where I_1 and I_2 are the images, h_{1i} and h_{2i} are the ith elements of the first and second histogram, respectively, and M is the final element in the histogram, which may be 255 in this case as the upper limit of the normalized range for a disparity map.

IV. Fuzzy Associative Database Algorithm

Fuzzy set theory lends itself particularly well to the problem of recognition based on a set of imprecise descriptors with much variation and overlap. However, it is generally impractical to develop a rule set for classification directly, since it is not immediately obvious what each descriptor represents about the object and how they combine. In such cases, one may train and optimize the parameters of the fuzzy system using a neural network, in a configuration known as a neuro-fuzzy system [19].

We describe here a *fuzzy associative database* similar to that found in [5], adapted for invariant values of disparity maps and for multiple training images expected to differ as a result of the viewpoint change. The basic approach is to store a table of fuzzy sets associated with the corresponding membership functions, where each class (type of object to be recognized) has one fuzzy set for each invariant value, which are constructed from fuzzified invariant values extracted from the disparity maps of the object from several different viewpoints (the training set). Recognition can then be accomplished by comparing input invariant values to the fuzzy sets in each class and determining which matches best.

A. Original FAD Algorithm

The original fuzzy associative database algorithm, described fully in [5], is used for invariant recognition of multiple planar objects in 2D. It consists of a fuzzy database (FD) and a fuzzy search engine (FSE) which are trained using invariant values extracted from the binary images of 2D objects.

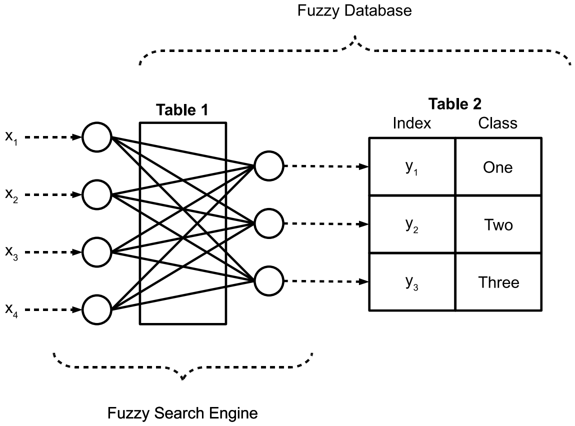


Fig. 2. FAD Network with 4 Invariant Values and 3 Classes

The key difference between this and the 3D recognition problem is the use of multiple images (different views) for training and recognition. In the 2D case, the invariant values are sufficient to characterize all possible planar views of the object, and therefore result in relatively compact membership functions of the corresponding fuzzy sets. In the 3D case, multiple views are necessary to capture the full structure of the object, and although the values are invariant to certain planar transformations of the object from a given view, across different views the resulting membership functions may be quite different. This may result in large areas of overlap among the input membership functions and these must therefore be scaled relative to themselves and one another to better describe the object characteristics.

It is possible and beneficial to emphasize the more unique and descriptive portions of the fuzzy sets before they are used for training or recognition. The fuzzy adaptive database is modified for the multi-image case as described in the following sections.

B. Supervised Training

During the supervised training stage, the invariant descriptors are computed from a disparity map of an object of known class. These are first fuzzified into a fuzzy set with a Gaussian membership function:

$$F(x, m, \sigma) = e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{7}$$

where x is the universe of discourse, m (the mean) is the input crisp value and σ is the standard deviation of the Gaussian, which is determined by trial and error.

Data from multiple views is thus entered, and the fuzzy sets are joined via a union operator. This results in a joint fuzzy set in each invariant value describing the object in an unbiased fashion from multiple viewpoints. In other words, the fuzzy set describes the entire range of acceptable invariant data associated with the object class. The value σ is chosen so that this statement is as true as possible without any more overlap with other classes than is necessary.

The net result so far, assuming a good training set and a good value of σ, is that the fuzzy system comprised of the fuzzy sets for each invariant, for a given class, should return a strong response to input invariants generated by a disparity map of any viewpoint of an object of the correct class. However, it is also highly likely at this point that there is much overlap among the different classes for certain invariants, and there is no practical way to directly account for such ambiguities.

In order to correct for this, once the fuzzy sets (and the corresponding membership functions) have been constructed for all training examples, they are adaptively scaled, essentially competing for the ranges of each invariant which best describe their classes. To accomplish this, the crisp invariants from the training set are first clustered according to the following algorithm [18]:

1) Taking values of the network inputs as the initial values to form the weight vector;
2) Determine the winner unit based on the minimum distance;
3) Updating the weight vectors of the winner as follows;

$$w_i(N+1) = w_i(N) + \alpha(\rho - w_i(N)) \tag{8}$$

where N is the number of training epochs (iterations), ρ is the network inputs (crisp invariant values in our case), and α is the learning rate (for example $\alpha = e^{-0.13q-0.69}$ where q is the number of trainees in a specific class).

After the cluster centers are found, each fuzzy input is scaled by a measure of the distance from the crisp input data to the associated cluster center as shown below:

$$A_{ij} = A_{ij} e^{-\left(\frac{|w_i - \rho_{ij}|}{|w_i + \rho_{ij}|}\right)} \tag{9}$$

where w_i is the location of the cluster center in the ith class, A_{ij} is the jth fuzzy input data of the ith class, and ρ_{ij} is the jth crisp input data in the ith class. As the distance between the cluster center w_i and input ρ_{ij} increases, A_{ij} approaches zero, thus reducing the contribution of data that is far from the cluster center of the class.

Figures 3 and 4 show an example of scaling on a simple fuzzy membership function.

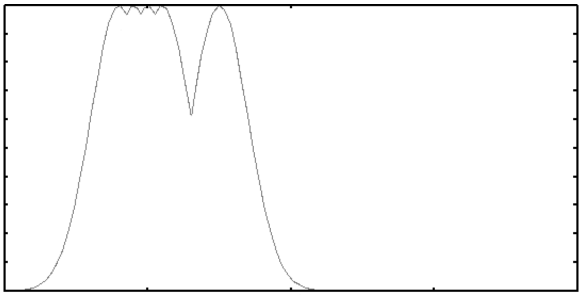


Fig. 3. Fuzzy Membership Function Before Scaling

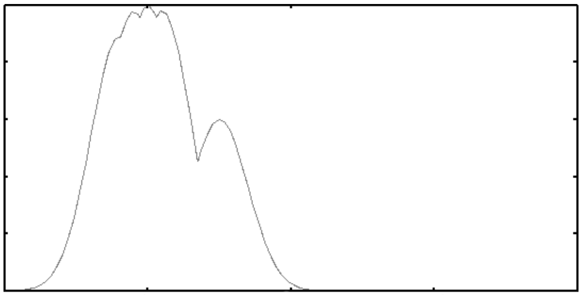


Fig. 4. Fuzzy Membership Function After Scaling

C. Recognition

Once the fuzzy associative database has been constructed, recognition is a relatively simple process. The system takes crisp invariant values computed from a disparity map of the object to be recognized (in any allowable orientation).

The crisp invariants are compared exhaustively to the FAD fuzzy set for each class, returning the total of the responses from each fuzzy set. The inference method found to best quantify the similarity for individual invariant values is a simple crisp value response, according to a standard inference equation:

$$\mu_a = \vee[\mu_j(x) \wedge I(x)] \tag{10}$$

where $\mu_j(x) \wedge I(x)$ represents the fuzzy intersection between the trained fuzzy set for invariant j and the fuzzified invariant from input image I, and the leading $\vee$ (union) indicates the fuzzy union over all invariant values. The class with the highest overall degree of membership μ_a is returned as the probable object class.

D. System Overview

The operation of the system is summarized in two flowchart diagrams. The first (Figure 5) shows the basic process of capturing images, creating the disparity map, and computing the invariant descriptors, mostly covered in section III. The second (Figure 6) shows the actual recognition network, including training, as described in subsections IV-B and IV-C.

Note in Figure 6 that the invariant fuzzy set scaling and clustering process takes place after all views have been captured by the vision system (with the fuzzified invariant membership functions stored unscaled), so that the resulting database incorporates descriptive characteristics of the 3D object from all of the views.

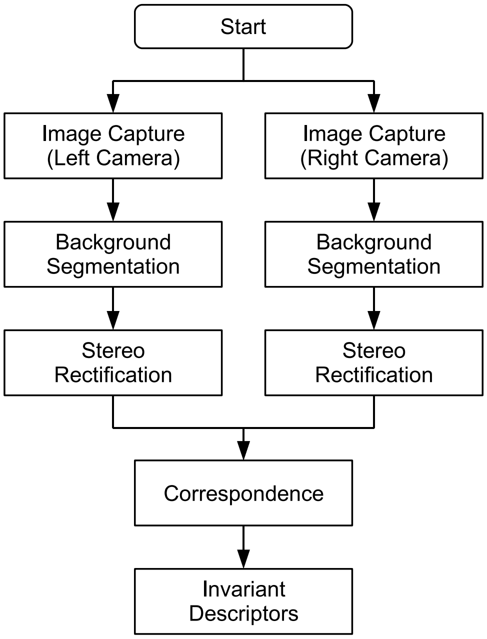


Fig. 5. Capture Process

V. Experimental Results

A. Apparatus

Testing was conducted using a vision platform consisting of two high-resolution CCD cameras, mounted on a robotic arm and calibrated for stereo triangulation. No particular constraints were applied to camera or object positioning other than generally placing the objects reasonably within the field of view of the system. The platform is shown in Figure 7.

B. Computing Invariant Values

In a practical system, conditions may not be ideal for generating proper invariant descriptors without some prior processing of the disparity maps. Since we want to recognize objects from different viewpoints, it must also be assumed that the objects might be found in different places in the field of view of the system, and with a background scene present this has a serious effect on the resultant disparity maps and invariant descriptors.

Fortunately, given a static background, it is a relatively simple task to compare each pixel to a stored image of the background itself and segment out everything but the object. Many methods exist in the computer vision and image processing literature, some more complex than others; we have employed a simple thresholding technique, with experimentally-tuned parameters t, F, and B, outlined below:

1) For each pixel $p_{i,j}$ and stored background pixel $s_{i,j}$, if $|p_{i,j} - s_{i,j}| > t$, mark as foreground.
2) Mark as background all foreground pixels in regions with contiguous area less than F.
3) Mark as foreground all background pixels in regions with contiguous area less than B.

The descriptors we use for recognition are invariant to translation, among other things, so once background subtraction has been performed it is of no concern where in the image the object lies, so long as it is fully within the image.

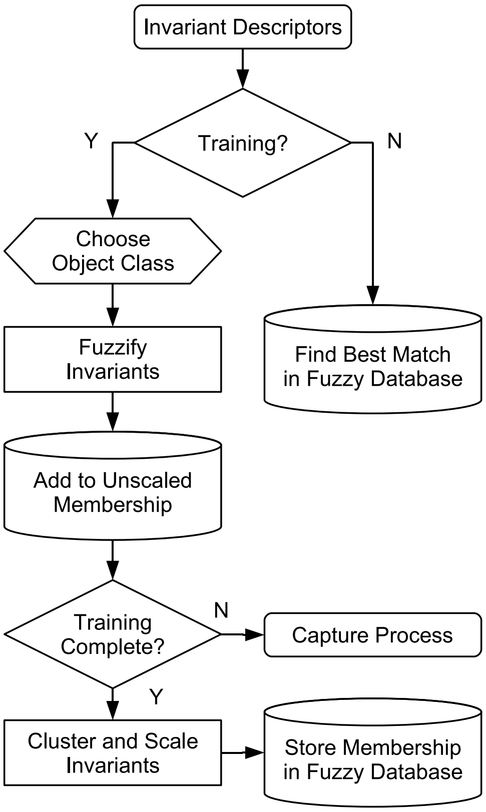

Invariant Descriptors
Training?
Y
N
Choose Object Class
Fuzzify Invariants
Add to Unscaled Membership
Find Best Match in Fuzzy Database
Training Complete?
N
Capture Process
Y
Cluster and Scale Invariants
Store Membership in Fuzzy Database

Fig. 6. Recognition Network

C. Results

The system was tested using the training set of Table I on a set of 200 disparity maps taken from different viewpoints of 3 different objects.

The recognition rates of the experiment using Gaussian fuzzification, three training views, and the simple crisp-value inference method are shown in Table II. Test A used no data scaling whereas Test B employed the LVQ self-scaling method. A very high recognition rate was achieved in all three classes, despite noise in the generated disparity maps and ambiguity in the shapes of the objects.

VI. Conclusions

After examining a variety of possible invariant descriptors for recognition of 3D objects based on disparity maps, we have found a particular combination of three to yield the best recognition results: compactness, the first Hu moment, and the histogram difference, as detailed in subsection III-C.

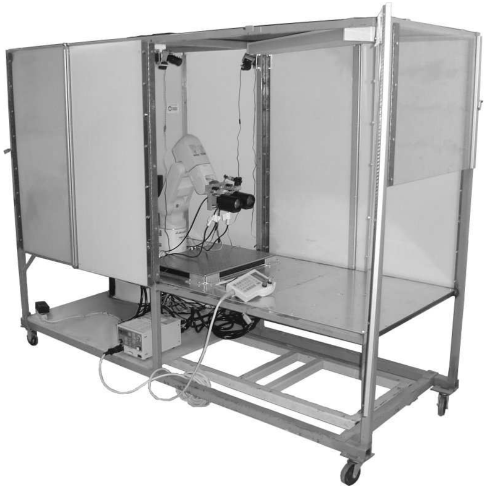
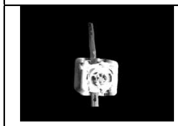
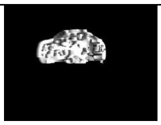
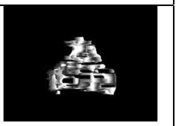
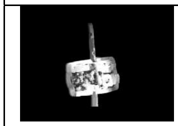
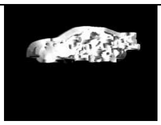
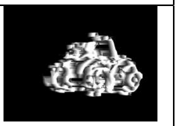
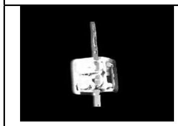
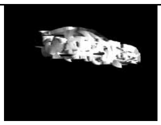
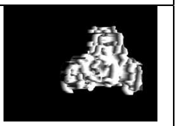


Fig. 7. Vision Platform

TABLE I
EXPERIMENT TRAINING SET

Class 1	Class 2	Class 3
		
		
		

The recognition method used a neural network to optimize fuzzy membership functions for the invariant descriptors against one another, which successfully mitigated misclassification introduced by ambiguities in the individual functions. After training the recognition system with just three views of an object, as described in section V, a very high recognition rate was achieved on disparity maps generated from arbitrary views.

The recognition could be made more robust by introducing additional invariant descriptors to the same general concept. One way to achieve this would be to improve the correlation correspondence algorithm to yield a smoother and more accurate range image; this could potentially allow the use of higher-order moment invariants. Another possibility would be to apply some form of normalization to the stereo images or the disparity maps so that additional descriptors not invariant to certain properties could be used. Finally, it may be possible to optimize recognition further by weighting the contribution of the individual invariant descriptor membership functions to the classification.

TABLE II
RECOGNITION RESULTS

Test	Class 1	Class 2	Class 3
A	94.00%	93.81%	86.67%
B	98.00%	98.97%	100.00%

REFERENCES

[1] J. J. Koenderink and A. J. van Doorn, "Geometry of Binocular Vision and a Model for Stereopsis," *Biological Cybernetics*, vol. 21, pp. 29–35, 1976.

[2] R. Y. Tsai, "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 364–374, 1986.

[3] R. Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[4] Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[5] S. Shahir, X. Chen, and M. Ahmadi, "Fuzzy Associative Database for Multiple Planar Object Recognition," *Proc. Intl. Symp. on Circuits and Systems*, vol. 5, pp. 805–808, 2003.

[6] G. Hetzel, B. Leibe, P. Levi, and B. Schiele, "3D Object Recognition for Range Images using Local Feature Histograms," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 394–399, 2001.

[7] S. Lin and S. W. Lee, "Using Chromaticity Distributions and Eigenspace Analysis for Pose-, Illumination-, and Specularity-Invariant Recognition of 3D Objects," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 426–431, 1997.

[8] N. Rui, J. Guangrong, Z. Wencang, and F. Chen, "3D Object Recognition from 2D Invariant View Sequence Under Translation, Rotation and Scale by Means of ANN Ensemble," *Proc. IEEE Intl. Wkshp. on VLSI Design and Video Technology*, pp. 292–295, 2005.

[9] R. J. Campbell and P. J. Flynn, "Eigenshapes for 3D Object Recognition in Range Data," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 505–510, 1999.

[10] M. Y. Mashor, M. M. Osman, M. R. Arshad, "3D Object Recognition Using 2D Moments and HMLP Network," *Proc. Intl. Conf. on Computer Graphics, Imaging and Visualization*, pp. 126–130, 2004.

[11] M. K. Hu, "Visual Pattern Recognition By Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[12] S.-G. Kong and B. Kosko, "Adaptive Fuzzy Systems for Backing Up a Truck-and-Trailer," *IEEE Trans. on Neural Networks*, vol. 3, no. 2, pp. 211–223, 1992.

[13] N. B. Karayiannis and P. I. Pai, "Fuzzy Algorithms for Learning Vector Quantization," *IEEE Trans. on Neural Networks*, vol. 7, pp. 1196–1211, 1996.

[14] B. Kusumoputro, H. Budiarto, and W. Jatmiko, "Fuzzy-Neuro LVQ and its Comparison with Fuzzy Algorithm LVQ in Artificial Odor Discrimination System," *ISA Trans.*, vol. 41, no. 4, pp. 395–407, 2002.

[15] K. L. Wu and M. S. Yang, "A Fuzzy-Soft Learning Vector Quantization," *Neurocomputing*, vol. 55, no. 3, pp. 681–697, 2003.

[16] C. Thiel, B. Sonntag, and F. Schwenker, "Experiments with Supervised Fuzzy LVQ," *Proc. 3rd IAPR Wkshp. on Artificial Neural Networks in Pattern Recognition*, pp. 125–132, 2008.

[17] B. Kosko, Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence, Prentice Hall, 1992.

[18] T. Kohonen, *Self-Organizing Maps*, Springer, 1995.

[19] D. Nauck, F. Klawonn, and R. Kruse, *Foundations of Neuro-Fuzzy Systems*, Wiley, 1997.

[20] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, The MIT Press, 1993.

[21] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, 1998.

[22] R. C. Gonzalez, *Digital Image Processing*, Prentice-Hall, 2002.