

# A Fuzzy Associative Approach for Recognition of 3D Objects in Arbitrary Pose

Aaron Mavrincac, Ahmad Shawky, and Xiang Chen

**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 given 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 [10], 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. PRELIMINARY THEORY

### A. Disparity Map

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, which is beyond the scope of this work; for a thorough geometrical treatment see [3], [29], and for some practical methods see [4], [5], [6].

Throughout this paper, the following convention is used for the world and image coordinate systems: lowercase  $x$  and  $y$  represent image coordinates starting at the upper left corner, 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.

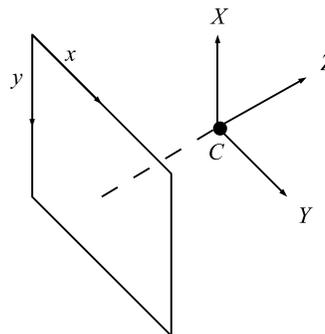


Fig. 1. Coordinate System Convention

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  $x_2 - x_1$  [29]. This can be used to triangulate the depth to the original 3D point in the environment (from

the optical centers of the cameras) in the world coordinate system according to the following relation:

$$Z = \frac{bf\lambda}{d} \quad (1)$$

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

A *disparity map* is a representation of the disparities of each pixel in one image with respect to the corresponding pixel, if any, in the other. It can be thought of as a two-dimensional matrix  $D$  of dimensions equal to the pixel dimensions of the digital image. If a pixel  $(i, j)$  in the image corresponds to a pixel  $(i, k)$  in the other image, the entry  $D_{i,j}$  is equal to the difference  $k - j$ .

For the purposes of this method, it is not necessary to know  $b$ ,  $f$ , or  $\lambda$  directly, since they are scalar parameters which are fixed in a given stereo camera rig (they are, however, found at least indirectly as part of calibration). For visual representation, it is normalized such that each entry falls within the range  $[0, 255]$  and displayed as a grayscale 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  $\bar{\mathbf{d}} + \mathbf{p}_l$  as the corresponding right-image pixel.

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

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 [31]. We use here the SSD criterion for  $\psi$ , 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*, 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 grayscale image can be calculated as follows, adapted from [32]:

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

where  $f(x, y)$  is the value of the image at pixel  $(x, y)$  and  $f_{boundary}(x, y)$  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 [19], 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:

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

where:

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

3) *Histogram*: The final descriptor is the histogram, 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 [13]:

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

where  $I_1$  and  $I_2$  are the images,  $h_{1i}$  and  $h_{2i}$  are the  $i$ th 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.

### D. Neuro-Fuzzy System

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 [28].

In this case, the neural network is used to competitively scale membership functions representing the same invariant descriptor across different classes, reducing the effect of ambiguities which might lead to misclassification. This technique is similar to *adaptive scaling* in that it responds indirectly to the recognition performance of the system during supervised training.

### III. MAIN RESULTS

We describe here a *fuzzy associative database* similar to that found in [10], adapted for multiple training images expected to differ as a result of the viewpoint change. The basic approach is to store a table of membership functions, where each class (type of object to be recognized) has one membership function 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 membership functions in each class and determining which matches best.

#### A. 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 Gaussian membership function:

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

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

Data from multiple views is thus entered, and the membership functions are joined via a union operator. This results in a complete membership function in each invariant value describing the object in an unbiased fashion from multiple viewpoints. In other words, the membership function describes the entire range of acceptable invariant data associated with the object class. The value  $\sigma$  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  $\sigma$ , is that the fuzzy system comprised of the membership functions 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 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 [27]:

- 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;

$$w_i(N+1) = w_i(N) + \alpha(\rho - w_i(N)) \quad (8)$$

where  $N$  is the number of training epochs (iterations),  $\rho$  is the network inputs (crisp invariant values in our case), and  $\alpha$  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:

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

where  $w_i$  is the location of the cluster center in the  $i$ th class,  $A_{ij}$  is the  $j$ th fuzzy input data of the  $i$ th class, and  $\rho_{ij}$  is the  $j$ th crisp input data in the  $i$ th class. As the distance between the cluster center  $w_i$  and input  $\rho_{ij}$  increases,  $A_{ij}$  approaches zero, thus reducing the contribution of data that is far from the cluster center of the class.

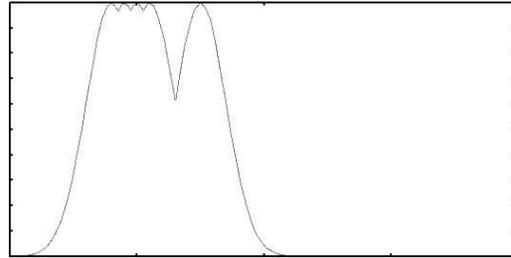


Fig. 2. Fuzzy Membership Function Before Scaling

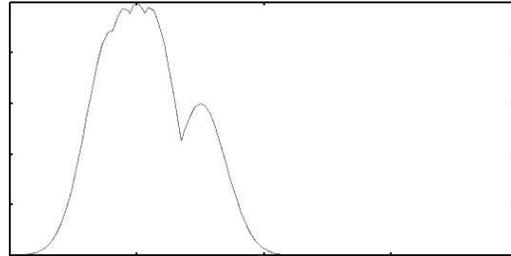


Fig. 3. Fuzzy Membership Function After Scaling

Figures 2 and 3 show an example of scaling on a simple fuzzy membership function.

#### B. 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 fuzzy associative database membership functions for each class, returning the total of the responses from each membership function. The inference method found to best quantify the similarity for individual invariant values is a simple crisp value response, according to the following inference equation:

$$\mu_a = \vee[\mu_j(x) \wedge I(x)] = b \quad (10)$$

The class with the highest resultant total is returned as the probable object class.

### C. Other Considerations

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.

### D. System Overview

The operation of the system is summarized in two flowchart diagrams. The first (Figure 4) shows the basic process of capturing images, creating the disparity map, and computing the invariant descriptors, mostly covered in section II. The second (Figure 5) shows the actual recognition network, including training, as described in subsections III-A and III-B.

## IV. EXPERIMENTAL RESULTS

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 6.

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, self-scaling, three training views, and the simple crisp-value inference method are shown in Table II. A very high recognition rate was achieved in all three classes,

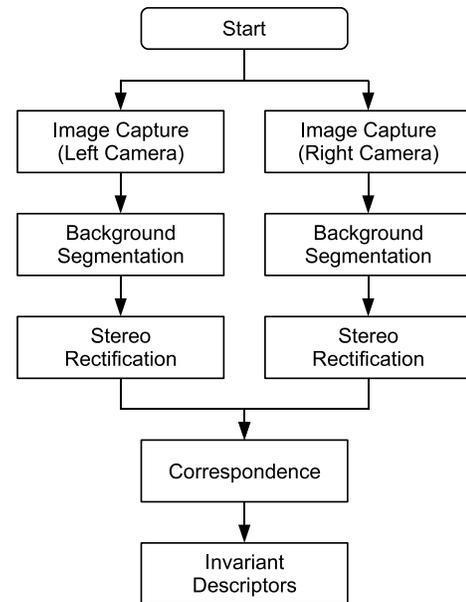


Fig. 4. Capture Process

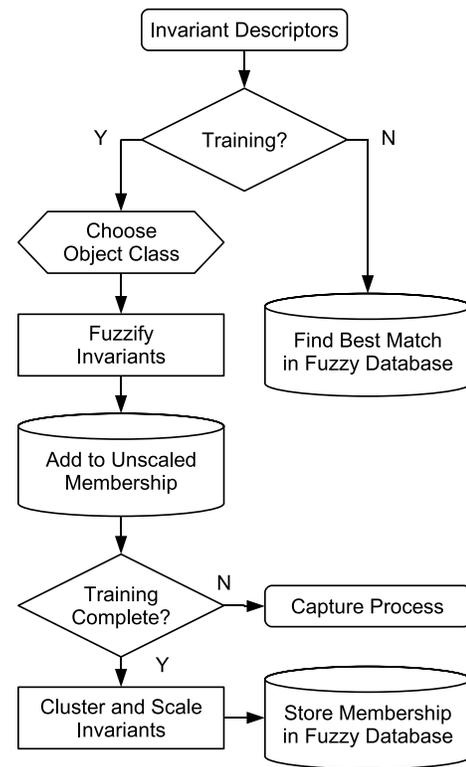
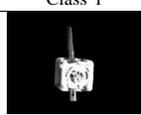
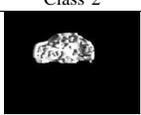
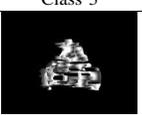


Fig. 5. Recognition Network



Fig. 6. Vision Platform

TABLE I  
EXPERIMENT TRAINING SET

Class 1	Class 2	Class 3
		
		
		

despite noise in the generated disparity maps and ambiguity in the shapes of the objects.

TABLE II  
RECOGNITION RESULTS

Class 1	Class 2	Class 3
98.00%	98.97%	100.00%

## V. 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 II-C.

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 IV, 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.

## ACKNOWLEDGMENT

This work was supported in part by NSERC-Discovery Research Grant.

## REFERENCES

- [1] C. Strelca and L. Van Gool, "PDE-Based Multi-View Depth Estimation," *Proc. 1st Intl. Symp. on 3D Data Processing Visualization and Transmission*, pp. 416–425, 2002.
- [2] N. Grammalidis, L. Bleris, and M. G. Strintzis, "Using the Expectation-Maximization Algorithm for Depth Estimation and Segmentation of Multi-View Images," *Proc. 1st Intl. Symp. on 3D Data Processing Visualization and Transmission*, pp. 686–689, 2002.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] W. Teoh and X. Zhang, "An Inexpensive Stereoscopic Vision System for Robots," *Proc. IEEE Intl. Conf. on Robotics and Automation*, vol. 1, pp. 186–189, 1984.
- [8] R. P. Wildes, "Direct Recovery of Three-Dimensional Scene Geometry from Binocular Stereo Disparity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 761–774, 1991.
- [9] P. An, C. Lu, and Z. Zhang, "Object Segmentation Using Stereo Images," *Proc. Intl. Conf. on Communications, Circuits and Systems*, vol. 1, no. 27–29, pp. 534–538, 2004.
- [10] 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.
- [11] A. Calway, "Recursive Estimation of 3D Motion and Surface Structure from Local Affine Flow Parameters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 562–574, 2005.
- [12] X. Sun and E. Dubois, "A Method for the Synthesis of Intermediate Views in Image-Based Rendering Using Image Rectification," *Proc. IEEE Canadian Conf. on Electrical and Computer Engineering*, pp. 991–994, 2002.
- [13] 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.
- [14] 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.

- [15] 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.
- [16] 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.
- [17] J. Sun, F. Zhao, and J. Zhao, "A New Method for Recognition of 3D Objects," *Proc. 1st Intl. Symp. on Systems and Control in Aerospace and Astronautics*, pp. 475–477, 2006.
- [18] 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.
- [19] M. K. Hu, "Visual Pattern Recognition By Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.
- [20] G. I. Salama and A. L. Abbott, "Moment Invariants and Quantization Effects," *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 157–163, 1998.
- [21] W. Xioahung and R. Rongchun, "A New Method for Image Normalization," *Proc. Intl. Symp. on Intelligent Multimedia, Video and Speech Processing*, pp. 365–369, 2001.
- [22] M. Bleyer and M. Gelautz, "A Layered Stereo Algorithm Using Image Segmentation and Global Visibility Constraints," *Proc. IEEE Intl. Conf. on Image Processing*, pp. 2997–3000, 2004.
- [23] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Intl. Jnl. of Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002.
- [24] R. Berkan and S. Trubatch, *Fuzzy Systems Design Principles: Building Fuzzy IF-THEN Rule Bases*, IEEE Press, 1997.
- [25] P. Antognetti and V. Milutinović, *Neural Networks: Concepts, Applications and Implementations, Volume 1*, Prentice Hall, 1991.
- [26] B. Müller, J. Reinhardt, and M. T. Strickland, *Neural Networks: An Introduction*, Springer, 1995.
- [27] T. Kohonen, *Self-Organizing Maps*, Springer, 1995.
- [28] D. Nauck, F. Klawonn, and R. Kruse, *Foundations of Neuro-Fuzzy Systems*, Wiley, 1997.
- [29] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, The MIT Press, 1993.
- [30] Y. Ma, S. Soatto, J. Koščeká, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*, Springer-Verlag, 2004.
- [31] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, 1998.
- [32] R. C. Gonzalez, *Digital Image Processing*, Prentice-Hall, 2002.