https://doi.org/10.2352/EI.2023.35.14.COIMG-172

A Globally Optimal Fast Iterative Linear Maximum Likelihood Classifier

Prasanna Reddy Pulakurthi [1], ***Sohail A. Dianat*** [1], ***Majid Rabbani*** [1], ***Suya You*** [2], ***Raghuveer M. Rao*** [2]
[1] ***Rochester Institute of Technology, Department of Electrical and Microelectronic Engineering, USA***
[2] ***DEVCOM Army Research Laboratory, USA***

Abstract

A novel iterative linear classification algorithm is developed from a maximum likelihood (ML) linear classifier. The main contribution of this paper is the discovery that a well-known maximum likelihood linear classifier with regularization is the solution to a contraction mapping for an acceptable range of values of the regularization parameter. Hence, a novel iterative scheme is proposed that converges to a fixed point, the globally optimum solution. To the best of our knowledge, this formulation has not been discovered before. Furthermore, the proposed iterative solution converges to a fixed point at a rate faster than the traditional gradient descent technique. The performance of the proposed iterative solution is compared to conventional gradient descent methods on linear and non-linearly separable data in terms of both convergence speed and overall classification performance.

Index terms - Maximum Likelihood Classifier, Contraction Mapping

1. Introduction

In data analytics and machine learning, it is often desirable to classify data into their corresponding classes using a particular cost function. Examples of popular classifiers include support vector machines (SVMs) [1], conditional maximum entropy [2], logistical regression [3], maximum likelihood (ML) classifier [4], naive Bayes classifier [5], and artificial neural networks [6]. These classifiers are helpful in a wide range of applications, including spam filtering, medical diagnosis, fraud detection, sentiment analysis, image classification, etc., to name a few.

In maximum likelihood linear classification, the goal is to find the values of the model parameters (such as the weights and biases in a linear classifier) that maximize the likelihood of the model given the training data. This is done by minimizing the negative log likelihood of the model, which is equivalent to maximizing the likelihood itself.

In the case of a linear classifier, the model assigns a label (e.g., "positive" or "negative") to an input sample based on the value of a linear combination of the input features and the model parameters. For example, in the case of binary classification, the model might assign a label of $+1$ (class 1) to a sample if the linear combination of the input features and the model parameters is greater than some threshold and a label of -1 (class 2), otherwise.

To train a maximum likelihood linear classifier, we need to specify a loss function that measures how well the model is able to predict the labels of the training data given the model parameters. One common loss function for this purpose is the logarithmic loss. This loss function is minimized during training in order to find the values of the model parameters that maximize the likelihood of the model given the training data.

This optimization problem can be solved using a variety of optimization algorithms, such as gradient descent [7] or a quasi-newton method [8]. However, these optimization algorithms involve finding the gradient, which is a computationally expensive and slowly converging solution which has the risk of getting trapped in a local extremum instead of converging to the cost function's globally optimum solution.

This paper proves that a maximum likelihood linear classifier with regularization [4] can be formulated as the solution to a contraction mapping operator. Hence an iterative algorithm can be constructed to converge to its fixed point where convergence to the globally optimum solution is guaranteed, and the convergence speed is faster than the traditional gradient-based techniques. The main contributions of this research can be summarized as follows:

It is mathematically proven that the ML classifier in [4] is the solution to a contraction mapping, and a novel iterative approach is presented to find its fixed point. The performance and convergence speed improvement of the proposed solution is shown via simulations on synthetic data and digit images.

2. Proposed Classification Method

Consider a two-class set of labeled data $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in R^d$ is the i^{th} d-dimensional input vector and the scalar $y_i \in \{1,-1\}$ denotes its corresponding label. This data may or may not be linearly separable. Consider a linear transformation on the input given by,

$$z_i = \boldsymbol{\theta}^T \boldsymbol{x}_i + b. \tag{1}$$

Here $\boldsymbol{\theta}$ is the vector of model parameters representing a normal vector to the separating hyperplane, and b is the bias. The data is assigned to class 1 if $z_i \geq 0$, and to class 2, otherwise. The linear classifier aims to determine the model parameter vector $\boldsymbol{\theta}$ and bias b to maximize the probability of accurate predictions. In [4], this is accomplished by minimizing the negative log likelihood loss function. This loss function is formulated as the negative log of sigmoid operation on (y z_i) and forms the log likelihood loss function as,

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},b) &= \sum_{i=1}^{N} -\ln\left(\frac{1}{1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b)}\right) \\ &= \sum_{i=1}^{N} \ln\left(1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))\right). \end{aligned} \tag{2}$$

The ML solution is found in [4] by minimizing a loss function that weighs the log likelihood function against a regularization term:

$$J(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},b) = L(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},b) + \lambda||\boldsymbol{\theta}||^2 = \sum_{i=1}^{N} \ln\left(1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))\right) + \lambda||\boldsymbol{\theta}||^2. \quad (3)$$

Where λ is the regularization parameter, $||\boldsymbol{\theta}||^2$ is the squared L_2 norm of the model parameters, and the summation is over all the training data samples. The regularization parameter is added to the objective function to help prevent overfitting. By adding the regularization term, the model is less likely to fit to noise in the training data and thus is more likely to generalize well to unseen data. The goal is to find the values of the model parameters $\boldsymbol{\theta}$ that minimize the loss function in Equation 3 given the training data. The optimal solution for the model parameters $\boldsymbol{\theta}$ is then given by,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},b),$$

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^{N} \ln\left(1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))\right) + \lambda||\boldsymbol{\theta}||^2\right].$$

The gradients of J with respect to $\boldsymbol{\theta}$ and b are derived as,

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = 2\lambda\boldsymbol{\theta} - \sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))}, \quad (4)$$

$$\frac{\partial J}{\partial b} = -\sum_{i=1}^{N} \frac{y_i\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))}. \quad (5)$$

Setting the gradient of J with respect to $\boldsymbol{\theta}$ given in Equation 4 equal to zero ($\frac{\partial J}{\partial \boldsymbol{\theta}} = 0$) yields the following nonlinear vector equation for optimal $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \frac{1}{2\lambda}\sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i+b))} = f(\boldsymbol{\theta}). \quad (6)$$

The parameter vector $\boldsymbol{\theta}$ of the hyperplane is found in [4] by formulating a gradient search technique to minimize the objective function in Equation 3 that is typically slow and can potentially get trapped in a local minimum.

A novel contribution of this paper is to prove that the expression in Equation 6 becomes a contraction mapping for certain choices of the regularization parameter λ, and hence has a unique solution that can be found using an iterative algorithm with an arbitrary initial condition $\boldsymbol{\theta}(0)$.

From proof in Section 3, $f(\boldsymbol{\theta})$ is a contraction mapping if $\lambda > \sum_{i=1}^{N}||\boldsymbol{x}_i||^2/8$. The iterative algorithm to solve for $\boldsymbol{\theta}^*$ is given by,

$$\boldsymbol{\theta}(n+1) = \frac{1}{2\lambda}\sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(-y_i(\boldsymbol{\theta}^T(n)\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}^T(n)\boldsymbol{x}_i+b))}, \quad (7)$$

$$b = \frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{\theta}^T(n)\boldsymbol{x}_i). \quad (8)$$

The parameter b in Equation 8 is updated every K iterations, where K is typically a small integer in the range of 3 to 5. This is based on the fact that the solution to the fixed point of the contraction mapping converges after a small number of iterations. The pseudo-code for the proposed iterative algorithm is given in Algorithm 1.

Algorithm 1 Proposed Iterative Solution to the Linear Maximum Likelihood Classifier.

Input: B batch size, N_E number of iterations.
Initialize $n = 0$, $K = 3$, parameter vector $\boldsymbol{\theta}(0)$ by a small random value, $\lambda = \sum_{i=1}^{N}||\boldsymbol{x}_i||^2/8 + \varepsilon$.
while $n < N_E$ **do**
  Sample a mini-batch data $\{\boldsymbol{x}_i\}_{i=1}^{B}$ and its corresponding labels $\{y_i\}_{i=1}^{B}$ and set $k = 0$.
  while $k < K-1$ **do**
    $\boldsymbol{\theta}(n) \leftarrow f(\boldsymbol{\theta}(n))$
    $k = k + 1$
  end
  $\boldsymbol{\theta}(n+1) \leftarrow f(\boldsymbol{\theta}(n))$
  $b = \frac{1}{B}\sum_{i=1}^{B}(y_i - \boldsymbol{\theta}^T(n)\boldsymbol{x}_i)$
  $n = n + 1$
end

3. Contraction Mapping Proof

$f(\boldsymbol{\theta})$ is contraction mapping if it satisfies the property $||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)|| \leq \rho||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||$ and $0 < \rho < 1$, which means that the distance between any two points under the function is strictly less than the distance between the two points themselves. The value ρ is called the contraction factor. It must be strictly less than 1 in order for the function to be a contraction mapping.

If $f(\boldsymbol{\theta})$ is a contraction mapping, then the iterative solution $\boldsymbol{\theta}(n+1) = f(\boldsymbol{\theta}(n))$ always converges to the fixed point solution. Using Equation 6 we can write $||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)||$ as,

$$||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)|| = \frac{1}{2\lambda}\left\|\sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(-y_i(\boldsymbol{\theta}_2^T\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}_2^T\boldsymbol{x}_i+b))} - \sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(-y_i(\boldsymbol{\theta}_1^T\boldsymbol{x}_i+b))}{1+\exp(-y_i(\boldsymbol{\theta}_1^T\boldsymbol{x}_i+b))}\right\|.$$

Denoting $h_i = -y_i(\boldsymbol{\theta}_2^T\boldsymbol{x}_i + b)$ and $w_i = -y_i(\boldsymbol{\theta}_1^T\boldsymbol{x}_i + b)$, and noting that $|y_i| = 1$, the above can be written as,

$$\begin{aligned} &||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)|| \\ &= \frac{1}{2\lambda}\left\|\sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(h_i)}{1+\exp(h_i)} - \sum_{i=1}^{N} \frac{y_i\boldsymbol{x}_i\exp(w_i)}{1+\exp(w_i)}\right\| \\ &\leq \frac{1}{2\lambda}\sum_{i=1}^{N}\left\|\frac{y_i\boldsymbol{x}_i\exp(h_i)}{1+\exp(h_i)} - \frac{y_i\boldsymbol{x}_i\exp(w_i)}{1+\exp(w_i)}\right\| \end{aligned}$$

$$
\begin{aligned}
||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)|| &\leq \frac{1}{2\lambda}\sum_{i=1}^{N}\left\|\frac{y_i\boldsymbol{x}_i(\exp(h_i)-\exp(w_i))}{(1+\exp(h_i))(1+\exp(w_i))}\right\| \\
&= \frac{1}{2\lambda}\sum_{i=1}^{N}|y_i|\,||\boldsymbol{x}_i||\left|\frac{\exp(h_i)-\exp(w_i)}{(1+\exp(h_i))(1+\exp(w_i))}\right| \\
&= \frac{1}{2\lambda}\sum_{i=1}^{N}||\boldsymbol{x}_i||\,|h_i-w_i|\left|\frac{e^{h_i}-e^{w_i}}{(h_i-w_i)(1+e^{h_i})(1+e^{w_i})}\right| \\
&= \frac{1}{2\lambda}\sum_{i=1}^{N}||\boldsymbol{x}_i||^2\,||\boldsymbol{\theta}_2-\boldsymbol{\theta}_1||\left|\frac{e^{h_i}-e^{w_i}}{(h_i-w_i)(1+e^{h_i})(1+e^{w_i})}\right|.
\end{aligned} \tag{9}
$$

Because,

$$
\begin{aligned}
|h_i - w_i| &= |-y_i(\boldsymbol{\theta}_2^T\boldsymbol{x}_i+b) - (-y_i(\boldsymbol{\theta}_1^T\boldsymbol{x}_i+b))| \\
&= |y_i(-\boldsymbol{\theta}_2^T\boldsymbol{x}_i - b + \boldsymbol{\theta}_1^T\boldsymbol{x}_i + b)| \\
&= |y_i|\,||\boldsymbol{x}_i||\,||\boldsymbol{\theta}_2-\boldsymbol{\theta}_1||.
\end{aligned} \tag{10}
$$

If $g(u,v) = \left|\frac{e^u-e^v}{(u-v)(1+e^u)(1+e^v)}\right|$, then it can be shown that $0 \leq g(u,v) \leq 1/4$ as follows:

Since the function g is symmetric across u and v, that is, $g(u,v) = g(v,u)$, the maximum occurs at $u = v = u^*$.

$$
g(u,u) = \lim_{v\to u}\left|\frac{e^u-e^v}{u-v}\right| \times \left|\frac{1}{(1-e^u)(1-e^v)}\right|.
$$

Apply l'Hôpital's rule,

$$
\begin{aligned}
g(u,u) &= \left|\frac{\frac{\partial(e^u-e^v)}{\partial v}}{\frac{\partial(u-v)}{\partial v}}\right| \times \left|\frac{1}{(1-e^u)^2}\right| \\
&= \left|\frac{0-e^v}{0-1}\right| \times \left|\frac{1}{(1+e^u)^2}\right| \\
g(u) &= \left|\frac{e^u}{(1+e^u)^2}\right|.
\end{aligned}
$$

Now we can find the maximum of $g(u)$ by solving for u using $\frac{dg(u)}{du} = 0$. The maximum is found at $u = 0$ and $g(0) = 1/4$. Therefore, $0 \leq g(u,v) \leq 1/4$. Substituting $g(u,v) = 1/4$ in Equation 9 we get,

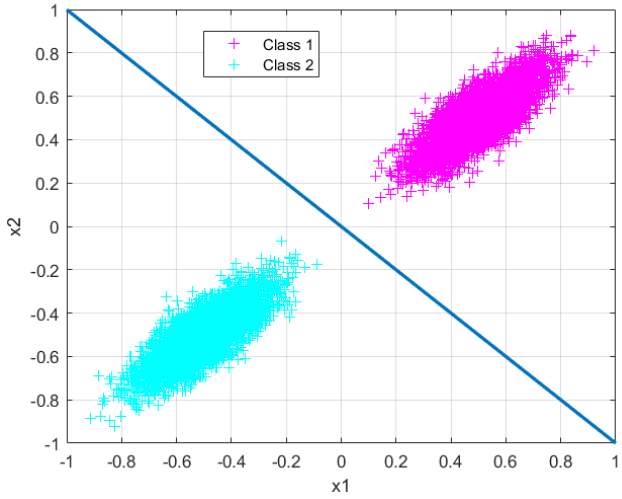
$$
||f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)|| \leq \frac{1}{8\lambda}\sum_{i=1}^{N}||\boldsymbol{x}_i||^2||\boldsymbol{\theta}_2-\boldsymbol{\theta}_1||. \tag{11}
$$

Here the contraction factor $\rho = \frac{1}{8\lambda}\sum_{i=1}^{N}||\boldsymbol{x}_i||^2$ and $f(\boldsymbol{\theta})$ is a contraction mapping if $\rho < 1$.

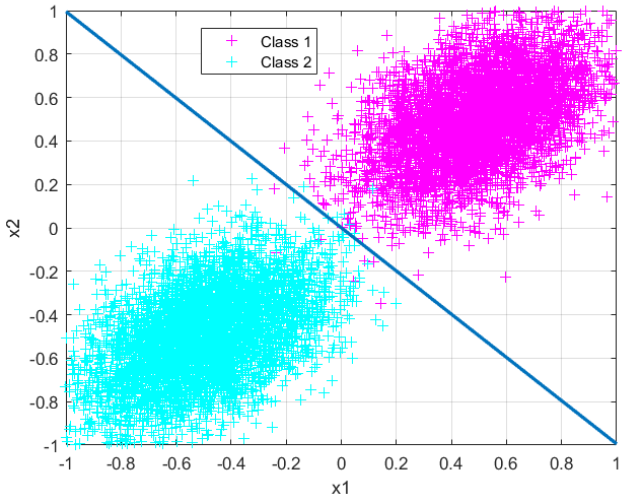
Therefore, $\frac{1}{8\lambda}\sum_{i=1}^{N}||\boldsymbol{x}_i||^2 < 1$, or, alternatively, $\lambda > \sum_{i=1}^{N}||\boldsymbol{x}_i||^2/8$.

4. Results

Classification results: Figure 1a shows the results of using the ML linear classifier formulated in [4] but obtained from our proposed contraction mapping fixed point solution. The synthetic two-class data has been generated by taking 1,000 data points for each class from a 2-D joint Gaussian distribution with a correlation coefficient of 0.5. Figure 1b shows the classification results on non-linearly separable data, which illustrates the robustness of the proposed iterative solution.



(a) Classification results for linearly separable data.



(b) Classification results on non-linearly separable data.

Figure 1. *Classification results.*

Similar to other binary classifiers, the proposed classifier can be extended to multiclass classification using approaches such as the one-vs-all method.

Contraction mapping solution vs. gradient descent: The proposed iterative solution's merit is that it converges significantly faster than the steepest descent methods. The Adam optimizer [9] is the most popular method used in the machine learning literature and is a relevant comparison choice. Both these methods are used to classify images of digits zero and one from the MINIST [10] data set. MNIST dataset consists of ten classes of handwritten digits from 0 to 9 with a training set of 60,000 images, a test set of 10,000, and an image size of 28×28. Now the dimension of the data is $28 \times 28 = 784$, which is significantly larger compared to the previous two-dimensional Gaussian example. Figure 2 shows sample images of ones and zeros.

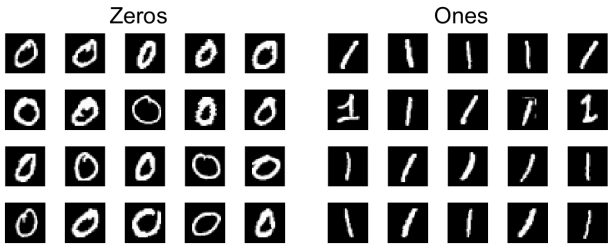


Figure 2. *Sample images of zeros and ones from MNIST dataset.*

The convergence speed of the iterative solution versus the gradient descent method on the binary image classification of zeros and ones is shown in Figure 3. The figure shows that the proposed iterative algorithm converges significantly faster than the Adam optimizer for a wide range of learning rates η.

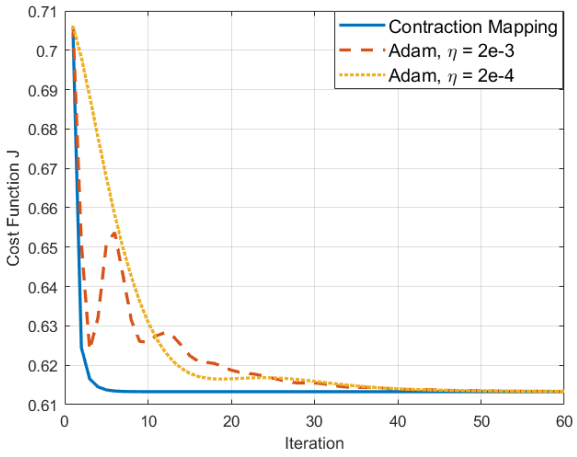


Figure 3. *Convergence speed of the contraction mapping solution versus gradient descent.*

5. Conclusion

This research presents a new method for training a maximum likelihood (ML) linear classifier with regularization. It is first shown that the ML classifier can be cast in the framework of a contraction mapping, and a novel iterative technique is proposed to find its fixed point. This approach is shown to have faster convergence to the globally optimal solution compared to traditional gradient-based techniques. The main contribution of this research is the discovery that the ML linear classifier is the solution to a contraction mapping, which has not been previously reported.

Acknowledgments

This research was funded in part by the DEVCOM Army Research Laboratory (Grant P12312-1).

References

[1] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik, "Support vector clustering," J. Mach. Learn. Res., vol. 2, pp. 125–137, Mar 2002.

[2] Kamal Nigam, "Using maximum entropy for text classification," in IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, pp. 61–67.

[3] Raymond E Wright, "Logistic regression.," 1995.

[4] Jing Jiang, Domain adaptation in natural language processing, University of Illinois at Urbana-Champaign, 2008.

[5] Sona Taheri and Musa Mammadov, "Learning the naive bayes classifier with optimization models," International Journal of Applied Mathematics and Computer Science, vol. 23, no. 4, 2013.

[6] John J Hopfield, "Artificial neural networks," IEEE Circuits and Devices Magazine, vol. 4, no. 5, pp. 3–10, 1988.

[7] Sebastian Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016.

[8] Wiki How English, "Quasi-newton method,".

[9] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

[10] Yann LeCun and Corinna Cortes, "MNIST handwritten digit database," 2010.

Author Biography

Prasanna Reddy Pulakurthi is a Ph.D. student at the Rochester Institute of Technology. His current area of research includes Image Processing, Computer Vision, Machine Learning, and Deep Learning.

Dr. Sohail A. Dianat received a B.S. degree in Electrical Engineering from the Arya-Mehr University of Technology in Tehran, Iran, and his M.S. and D.Sc. degrees in Electrical Engineering from George Washington University. In September 1981, he joined the Rochester Institute of Technology, where he is currently a professor of Electrical Engineering and Imaging Science. Dr. Dianat has taught an assortment of undergraduate and graduate courses in the areas of digital signal/image processing and digital communication. His current research interests include digital signal/image processing and wireless communication.

Dr. Majid Rabbani received a B.S. degree from the Arya-Mehr University of Technology in Tehran, Iran, and his M.S. and D.Sc. degrees from the University of Wisconsin Madison. He has more than 40 years of experience in the area of digital image and video processing and analysis. He retired from Kodak after a 33-year career in research in 2016 with the rank of Kodak Fellow and currently holds the title of Professor of Practice at RIT. Rabbani is a Fellow of IEEE (1997), a Fellow of SPIE (1993), a past chair of the SPIE's Fellows Committee, a Kodak Fellow, and a Distinguished Inventor with 44 issued patents.

Dr. Raghuveer Rao has a B.E. degree in Electronics and Communication Engineering from Mysore University, an M.E. degree in Electrical Communication Engineering from the Indian Institute of Science, and a Ph.D. in Engineering from the University of Connecticut. He joined the Rochester Institute of Technology in 1987, where he was a Professor of Electrical Engineering and a member of the Imaging Science faculty until 2008. Since November 2008, he has been with the Army Research Laboratory in Adelphi, MD, where he is currently the Chief of the Image Processing Branch. Dr. Rao has held visiting appointments with the Indian Institute of Science, Princeton University, the Air Force Research

Laboratory, and the Naval Surface Warfare Center. He is a recipient of the IEEE Signal Processing Society Paper Award and an elected fellow of IEEE and SPIE.

Dr. Suya You received the Ph.D. degree from Huazhong University of Science and Technology, China, in 1994. He is a research assistant professor in the Computer Science Department at the University of Southern California (USC). His expertise is in the fundamental and applied aspects of digital media processing and applications. He also holds research positions at the Integrated Media Systems Center (IMSC), a U.S. National Science Foundation (NSF) Engineering Research Center (ERC) at USC, and the Center for Interactive Smart Oilfield Technologies (CISOFT), a USC-Chevron Center of Excellence for Research and Academic Training on Interactive Smart Oilfield Technologies. His current research focuses on mobile augmented reality, large-scale scene modeling and visualization, game technique for simulation and training, and multisensor data fusion for remote operations. He is the author or co-author of more than 100 papers and a coholder of several patents and technology disclosures in these areas. He is a member of the IEEE.