Gaussian Mixture Density Modeling, Decomposition, and Applications

Xinhua Zhuang, *Senior Member, IEEE*, Yan Huang, K. Palaniappan, *Member, IEEE*, and Yunxin Zhao, *Senior Member, IEEE*

***Abstract*— Gaussian mixture density modeling and decomposition is a classic yet challenging research topic. We present a new approach to the modeling and decomposition of Gaussian mixtures by using robust statistical methods. The mixture distribution is viewed as a (severely) contaminated Gaussian density. Using this model and the model-fitting (MF) estimator, we propose a recursive algorithm called the *Gaussian mixture density decomposition* (GMDD) algorithm for successively identifying each Gaussian component in the mixture. The proposed decomposition scheme has several distinct advantages that are desirable but lacking in most existing techniques. In the GMDD algorithm the number of components does not need to be specified *a priori*, the proportion of noisy data in the mixture can be large, the parameter estimation of each component is virtually initial independent, and the variability in the shape and size of the component densities in the mixture is taken into account.**

Gaussian mixture density modeling and decomposition has been widely applied in a variety of disciplines that require signal or waveform characterization for classification and recognition, including remote sensing, target identification, spectroscopy, electrocardiography, speech recognition, or scene segmentation. We apply the proposed GMDD algorithm to the identification and extraction of clusters, and the estimation of unknown probability densities. Probability density estimation by identifying a decomposition using the GMDD algorithm, that is, a superposition of normal distributions, is successfully applied to the difficult biomedical problem of automated cell classification. Computer experiments using both real data and simulated data demonstrate the validity and power of the GMDD algorithm for various models and different noise assumptions.

I. Introduction

Finite mixture densities have served as important models for complex processes throughout the history of statistics. The underlying motivation for studying finite mixture densities is the versatility of the model for representing physical phenomena. Many practical applications involve the extraction of *regularities* from a limited set of data samples. The *regularities* to be discovered in the data are typically not so regular given noisy sparse observations. Consequently, the distribution of the data cannot be accurately modeled using any simple structured probabilistic distribution. For example, a primary requirement of medical diagnosis is to identify the range of values for an appropriate feature set that distinguishes abnormal from normal cells. Abnormal cells may differ from normal cells in such characteristics as shape, color, texture, size, motility, chemical composition, metabolism, and other measurable factors. However, as is common with most practical applications, the statistical tendency of abnormal cells cannot be easily characterized by any simple structured density. Hence, a mixture model consisting of a number of component densities can be used in order to construct a satisfactory model for distinguishing among different classes of cells.

The most important class of finite mixture densities are *Gaussian mixtures*. The reasons for the importance and widespread use of Gaussian mixture densities are not incidental, but include the fact that a univariate Gaussian distribution has a simple and concise representation requiring only two parameters, the mean μ and the variance σ^2. The Gaussian density is symmetric, unimodal, isotropic, and assumes the least prior knowledge (as measured in terms of the uncertainty or entropy of the distribution) in estimating an unknown probability density with a given mean and variance [9]. These characteristics of the Gaussian distribution along with its well-studied status give Gaussian mixture density models the power and effectiveness that other mixture densities can hardly surpass. Despite many successes in the past few decades, however, difficulties still remain in the modeling and decomposition of general Gaussian mixture densities [2], [1].

The maximum-likelihood method for estimating the parameters of mixture densities has been one of the most widely used approaches to the problem [12]. However, the method of maximum-likelihood for directly estimating the parameters of a Gaussian mixture density given only data samples has many practical implementation difficulties for the following reasons.

1) The number of components in the mixture is required to be known *a priori.*
2) For mixtures with a large number of components the total number of parameters to be estimated (the unknown mean vectors and unknown covariance matrixes) can be very large in proportion to the available data samples.
3) Mixture components may be merged due to identical mean vectors or even with distinct mean vectors.
4) There may be singularities in the log-likelihood function, since the likelihood need not be bounded above.

Manuscript received January 1, 1994; revised January 11, 1996. This work was supported by the National Science Foundation under Grant IRI 9112649 and Grant IRI 95-02074, and by NASA Applied Information Systems Research program award NRA-93-OSSA-09 and NAG-2573. The associate editor coordinating the review of this paper and approving it for publication was Robert M. Haralick.

X. Zhuang and Y. Huang are with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211 USA (e-mail: zhuang@ece.missouri.edu).

K. Palaniappan is with the Universities Space Research Association, Laboratory for Atmospheres, NASA/Goddard Space Flight Center, Greenbelt, MD 20771 USA.

Y. Zhao is with the Beckman Institute, University of Illinois, Urbana, IL 61801 USA.

Publisher Item Identifier S 1057-7149(96)06033-2.

5) The log-likelihood function can have nonunique solutions in which several choices for the parameter values each have the largest local maximum.
6) There may be several local maxima of the log-likelihood function that is not the global maximum or the largest local maxima.
7) The final estimates maximizing the likelihood function are sensitive to the initial values used by numerical algorithms to solve the maximum-likelihood equations.

Even when the number of components is known *a priori* or restricted, then without appropriate constraints the maximum-likelihood estimates of the mean vectors and covariance matrixes in a Gaussian mixture can be completely erroneous [3]. A multiresolution approach that does not require an initial estimate of the number of Gaussian components based on a multiple-scale representation of the original signal for Gaussian mixture decomposition was developed by Palaniappan [11].

The new approach to mixture density modeling involves the use of robust statistics [15]. Robustness techniques are used to deal with model deviation issues. Many model assumptions that are frequently used in exploratory data analysis, such as normality, linearity and independence, are at best approximations to reality. In particular, gross errors can occur quite often and unexpectedly in real data samples. It is not unusual for even a single outlier to have a disasterous effect on the estimated statistics because of the undue influence of any outlier. Given the incidence and impact of outliers, robust statistics aim to:

- describe the structure that best fits the bulk of the data;
- identify data point deviations (outliers) or substructure deviations that can be further analyzed if required;
- identify and provide warning about unduly influential data points ("leverage points"), in the case of an unbalanced distribution of data that would severely impact the statistical estimates, as in regression analysis.

The distribution of data, presence of substructures, and the influence of each data point are of great concern in robust statistical analysis.

Robust statistics can be naturally applied to mixture density modeling to improve the accuracy and reliability with which individual densities can be identified. A mixture density is observed as a composition of simple structured densities or data structures. Now with respect to a particular density or structure, all of the other densities or structures can be readily classified as part of the outlier category in the sense that these other observations obey different statistics. Thus, a Gaussian mixture density can be viewed as a *contaminated Gaussian density* with respect to each Gaussian component in the mixture. When all of the observations for a single Gaussian density are grouped together, then the remaining observations (Gaussian components and noise) can be considered to form an unknown outlier density. Each Gaussian component can be estimated separately one at a time in an iterative fashion by using the contaminated Gaussian density model. The iterative estimation of Gaussian densities enables the number of components in the resulting mixture to be successively reduced.

The outlier density obtained in the process of robust estimation using the contaminated Gaussian model, however, might comprise a large proportion of the observations, resulting in a severely contaminated Gaussian density. The partial modeling approach, known as the model-fitting (MF) estimator was proposed to estimate the parameters of a severely contaminated Gaussian density [17], [18], [20]. The MF estimator is statistically a highly robust estimator and is made less initial dependent. The MF estimator will be used to detect each Gaussian component assuming a Gaussian mixture model for the observations. The recursive mixture decomposition algorithm proceeds by using the MF estimator to identify a valid Gaussian component, removing the inlier samples (observations associated with the identified Gaussian component) in order to speed up convergence, and then continuing with the decomposition using the smaller data set with fewer observations. This recursive procedure for the Gaussian mixture density modeling and decomposition problem using the MF estimator is called the Gaussian mixture density decomposition (GMDD) algorithm.

In the following section, the MF estimator is rederived based solely upon the minimum error Bayesian classification method. Zhuang *et al.* [18] showed that the solution to a general regression estimation problem corresponds to a local maximum of a family of partially modeled log-likelihood functions. This insight is used to develop a new enhanced MF estimation algorithm by integrating a one-dimensional (1-D) sequential search for a desired partial model with the Monte Carlo random search for an appropriate initial. The combination of the two search techniques makes the detection of a Gaussian component virtually initial independent.

The paper is organized as follows. In Section II, we present the enhanced version of the MF estimator and the GMDD algorithm. In Section III, we apply the GMDD algorithm to the analysis of cluster data. In Section IV, the same algorithm is used to estimate an unknown probability density in the practical context of cervical smear cell classification. Conclusions are presented in the final section. The appendix briefly describes the Kolmogorov–Smirnov (K–S) normality test, which is used to determine the validity of extracted Gaussian components and clusters.

II. Minimum-Error Bayesian Classification, Enhanced MF-Estimator, and GMDD Algorithm

A. Minimum Error Bayesian Classification and Partial Modeling

As discussed in Section I, the basic problem involved in Gaussian mixture density modeling and decomposition is to discover each valid Gaussian component G that is characterized by $N(m, C)$ in the given data set X. Let X consist of N samples $x^1, \cdots, x^N$, which belong to an n-dimensional Euclidean space R_n. It is assumed that each sample $x^k \in X$ is generated by an unknown Gaussian distribution $N(m, C)$ with probability $(1-\epsilon)$ plus an unknown outlier distribution $h(\cdot)$ with probability ϵ. These data samples are then identically distributed with the common probability density f, namely

$$f(x^k) = \frac{1-\epsilon}{(\sqrt{2\pi})^n\sqrt{|C|}}\exp\left(-\frac{1}{2}d^2(x^k)\right) + \epsilon h(x^k) \quad (1)$$

where $d^2(x^k)$ represents the squared Mahalanobis distance of x^k from the unknown mean vector m, namely

$$d^2(x^k) = (x^k - m)C^{-1}(x^k - m). \quad (2)$$

The density $f(\cdot)$ becomes a pure Gaussian density when $\epsilon = 0$. Accordingly, $f(\cdot)$ is called a contaminated Gaussian density when $\epsilon > 0$ [6].

The contaminated Gaussian density model $f(\cdot)$ directly incorporates uncertainty and can be justified as an appropriate model in many practical applications. *Supernature* chooses x^k from $N(m, C)$ with probability $(1-\epsilon)$ or from $h(\cdot)$ with probability ϵ. But the experimenter is only able to observe x^k. Ideally, a sample x^k is classified as an *inlier* if it is realized from $N(m, C)$ or as an outlier otherwise. Let G be the subset of all observations from the Gaussian component that contains all inliers and B be its complement. The subset $B = X - G$, then contains all of the outliers in the given set of observations. This can be stated as

$$\begin{aligned} G &= \{x^i\colon x^i \text{ generated by } N(m, C)\} \\ B &= \{x^j\colon x^j \text{ not generated by } N(m, C)\}. \end{aligned} \quad (3)$$

To be fruitful, we need reasonable sound constraints, as emphasized in the previous section. By assuming that *supernature* is impartial we reduce uncertainty and then the ideal classification (3) would imply that the likelihood of any inlier being generated by $N(m, C)$ is greater than the likelihood of any outlier being generated by $N(m, C)$. Therefore, for any $x^i \in G$ and $x^j \in B$, it follows that

$$g_i > g_j \quad (4)$$

where g_k stands for $1/((\sqrt{2\pi})^n\sqrt{|C|})\exp(-\frac{1}{2}\ d^2(x^k))$. Defining

$$\begin{aligned} g &= \min\{g_i\colon x^i \in G\} \\ b &= \max\{g_j\colon x^j \in B\} \end{aligned} \quad (5)$$

then by (4), the minimum likelihood of any inlier g must be greater than the maximum likelihood of any outlier b resulting in the following constraint:

$$g > b. \quad (6)$$

The discovery of a valid Gaussian component in X can be viewed as the result of a feasible classification procedure that divides the only observed data set into inliers and outliers. The ideal classification (3) generated by *supernature* cannot be exactly retrieved by the experimenter even if all the parameters $m, C, \epsilon, h(x^1), \cdots, h(x^N)$ are known. We are thus looking for such a practical classification that attains the minimum misclassification error.

Using (1), the probability of a sample x^k being an inlier is given by

$$\lambda_k = \frac{(1-\epsilon)g_k}{f_k} \quad (7)$$

where f_k stands for $f(x^k)$. According to Tou and Gonzalez [16], the misclassification error probability will be minimal if the experimenter uses the following Bayesian classification:

$$\begin{aligned} G &= \{x^i\colon \lambda_i > 0.5\} = \left\{x^i\colon g_i > \frac{\epsilon h(x^i)}{1-\epsilon}\right\} \\ B &= \{x^j\colon \lambda_j \le 0.5\} = \left\{x^j\colon g_j \le \frac{\epsilon h(x^j)}{1-\epsilon}\right\} \end{aligned} \quad (8)$$

to approximate the ideal classification (3). The minimum-error Bayesian classification (8) states that an inlier is more likely to have been generated by $N(m, C)$ than by the unknown outlier distribution $h(\cdot)$ and that an outlier is less likely to have been generated by $N(m, C)$ than by $h(\cdot)$. Due to the constraint (6), the Bayesian classification (8) enforces those N unknown outlier density values $h(x^k)$'s to fall in the half-open interval $[(1-\epsilon)^b/\epsilon, (1-\epsilon)^g/\epsilon)$. Interestingly enough, any combination of N values from that interval would realize the same Bayesian classification, as can be easily verified. One possible combination that assumes the least configurational information about the unknown outlier distribution would have all N unknowns to be identical, that is,

$$h(x^1) = h(x^2) = \cdots = h(x^N) = \delta. \quad (9)$$

It is important to emphasize that there exists an infinite number (a continuum) of partial models (that is, δ's), each of which lies in the interval and provides the same satisfactory Bayesian classification (8) with regard to the N observed data samples. Using the partial modeling assumption (9) for the outlier density function, (1) reads as

$$f_k = (1-\epsilon)g_k + \epsilon\delta. \quad (10)$$

If the samples $x^k, k = 1, \cdots, N$, are further assumed to be independent of each other, then the log-likelihood function of observing $x^1, \cdots, x^N$ conditioned on m, C, ϵ, and δ can be expressed as

$$Q = \log \mathrm{P}(x^1, \cdots, x^N | m, C, \epsilon, \delta) = \sum_k \log f_k. \quad (11)$$

Using (10), Q can be written as

$$Q = N\log(1-\epsilon) + \sum_k \log\{g_k + \frac{\epsilon\delta}{1-\epsilon}\}. \quad (12)$$

Defining the model-fitting function

$$q(m, C; t) = \sum_k \log\{g_k + t\} \quad (13)$$

it is easy to verify that the maximization of Q at δ with respect to m and C is equivalent to maximizing $q(m, C; t)$ at t with respect to m and C, provided that $t = \epsilon\delta/(1-\epsilon)$. Without confusion, we shall refer to each "t" (≥ 0) as a partial model from now on.

B. Enhanced MF Estimator and the GMDD Algorithm

In order to fully justify the proposed Bayesian classification (8) and the partial modeling of the unknown outlier density function (9), the following issues need to be discussed.

- Is it possible to ensure that the maximization of $q(m, C; t)$ with respect to m and C, as t varies, will yield a valid Gaussian component G, that is largely immune to the proportion of outliers?
- Is it possible to ensure that the detection of a valid Gaussian component can be made to be virtually independent of the initialization of the mean vector m?

The first issue was positively answered in Zhuang *et al.* [18] by showing that the MF estimator is data-confusion resistant and is not adversely affected when the proportion of bad data increases. Now we discuss the second issue of initial independence. Due to the large range of choices available for a satisfactory partial model "t," we propose to integrate a 1-D sequential search for partial models t using a Monte Carlo random search method [14] for different initializations of m. The random search of initials, when performed at each acceptable partial model, leads to the detection of a valid Gaussian component. The two-step search process also improves the speed of convergence and the efficiency of searching for the true Gaussian component. In the following, we present the enhanced MF estimator within the context of detecting a valid Gaussian component. The extension to the general regression domain is straightforward and will not be presented.

A sequence of partial models is set up, $t_0 = 0 < t_1 < \cdots < t_L$, where t_L denotes an upper bound for all potentially desirable partial models. The determination of the upper bound depends upon the spatial arrangements of data samples and can be estimated. At each selected partial model "t_s," $s = 0, 1, \cdots, L$, we iteratively maximize $q(m, C; t_s)$ with respect to m and C by using the gradient ascent rule beginning with a randomly chosen initial mean $m^{(0)}$. Having solved $\max_{m,C}\ q(m, C; t_s)$ for $m(t_s)$ and $C(t_s)$, we calculate the inlier set: $G(t_s) = \{x^i\colon g_i(m(t_s), C(t_s)) > t_s\}$ followed by the Kolmogorov–Smirnov (K–S) normality test on $G(t_s)$. If the test succeeds, then a valid Gaussian component, $G(t_s)$, has been determined. Otherwise, we proceed to the next partial model if the upper bound t_L has not been reached. This provides a brief description of the enhanced MF estimator. The details for solving each $\max_{m,C}\ q(m, C; t_s)$ is given in Section II-C.

In the recursive GMDD algorithm, after detecting a valid Gaussian component, we subtract it from the current data set and apply the MF estimator to the new size-reduced data set to search for another valid Gaussian component. Individual Gaussian components continue to be estimated recursively until the size of a new data set gets too small.

C. Solving $\max_{m,C}\ q(m, C; t_s)$

The gradient ascent rule is used to solve each maximization of $\max_{m,C}\ q(m, C; t_s)$ for a specified t_s. The gradients $\nabla_m q(m, C; t_s)$ and $\nabla_C q(m, C; t_s)$ can be derived as

$$\begin{aligned}
\nabla_m q(m, C; t_s) &= \nabla_m \sum_k \log(g_k + t_s)\\
&= \sum_k \frac{1}{g_k + t_s}\nabla_m g_k\\
&= \sum_k \frac{g_k}{g_k + t_s}\nabla_m(\log g_k)\\
&= \sum_k \lambda_k \nabla_m(\log g_k)\\
&= \sum_k \lambda_k \left\{ \nabla_m \log \frac{1}{\sqrt{2\pi}^n \sqrt{|C|}} \right.\\
&\quad \left. - \nabla_m \frac{(x_k - m)'C^{-1}(x_k - m)}{2} \right\}\\
&= -\frac{1}{2}\sum_k \lambda_k \nabla_m\{(x_k - m)'C^{-1}(x_k - m)\}\\
&= \sum_k \lambda_k C^{-1}(x_k - m)\\
&= C^{-1}\sum_k \lambda_k(x_k - m) \qquad (14)
\end{aligned}$$

$$\begin{aligned}
\nabla_C q(m, C; t_s) &= \sum_k \lambda_k \nabla_C(\log g_k)\\
&= \sum_k \lambda_k \left\{ \nabla_C \log \frac{1}{\sqrt{2\pi}^n \sqrt{|C|}} \right.\\
&\quad \left. - \nabla_C \frac{(x_k - m)'C^{-1}(x_k - m)}{2} \right\}\\
&= -\frac{1}{2}\sum_k \lambda_k\{\nabla_C \log|C|\\
&\quad + \nabla_C(x_k - m)'C^{-1}(x_k - m)\}\\
&= -\frac{1}{2}\sum_k \lambda_k\{C^{-1} - C^{-2}(x_k - m)(x_k - m)'\}\\
&= -\frac{1}{2}C^{-1}\left\{\sum_k \lambda_k - C^{-1}\right.\\
&\quad \left. \cdot \sum_k \lambda_k(x_k - m)(x_k - m)'\right\} \qquad (15)
\end{aligned}$$

where

$$\lambda_k = \frac{g_k}{g_k + t_s}, g_k = \frac{1}{(\sqrt{2\pi})^n\sqrt{|C|}}\exp(-\frac{1}{2}d^2(x^k)).$$

Motivated by the gradient expressions (14) and (15), the following fast iterative scheme is used to solve each $\max_{m,C} q(m, C; t_s)$ for $m(t_s)$ and $C(t_s)$:

Step 1. Randomly choose $m^{(0)}$, then form $C^{(0)}$ by $C^{(0)} = 1/N\ \Sigma_k(x^k - m^{(0)})(x^k - m^{(0)})'$.

Step 2. Given the jth estimates $m^{(j)}$ and $C^{(j)}$ at the jth iterative step, $g_k^{(j)}$ and $\lambda_k^{(j)}, k = 1, \cdots, N$ are first calculated, and the $(j+1)$th estimates $m^{(j+1)}$ and $C^{(j+1)}$ are then calculated as the solution to the following weighted least-squares problem:

$$\max_{m,C} \sum_k \lambda_k^{(j)} \log g_k \rightarrow m^{(j+1)}, C^{(j+1)}. \tag{16}$$

From (14) and (15), $m^{(j+1)}$ and $C^{(j+1)}$ can be derived as

$$\begin{aligned} m^{(j+1)} &= \frac{1}{\lambda^{(j)}} \sum_k \lambda_k^{(j)} x^k \\ C^{(j+1)} &= \frac{1}{\lambda^{(j)}} \sum_k \lambda_k^{(j)} (x^k - m^{(j+1)})(x^k - m^{(j+1)})' \end{aligned} \tag{17}$$

where $\lambda_k^{(j)} = g_k^{(j)}/(g_k^{(j)} + t_s)$ and $\lambda^{(j)} = \sum_k \lambda_k^{(j)}$.

III. Cluster Analysis

An important application of the GMDD algorithm is to the problem of identifying clusters or classes in multidimensional datasets. Cluster analysis plays a central role in pattern recognition, particularly when only unlabeled training data samples are available. A cluster can be loosely defined as a set of samples whose density is larger than the density of the surrounding volume. Clustering methods attempt to partition a set of observations by grouping them into a number of statistical classes. The objective of clustering unlabeled data is to obtain an organization and description of the data that is both meaningful and compact. The general class of methods for extracting information from unlabeled samples are known as *unsupervised learning* algorithms. Once data samples are labeled using a clustering algorithm, the resulting classes can be used to design a pattern classifier for classifying unknown samples. Consequently, the effectiveness of cluster analysis is crucial to the performance of the classifier in identifying the class to which noisy observations belong. There are several major difficulties encountered in cluster analysis that are also common to mixture data analysis. They include the following:

- The characteristics and locations of clusters are usually not known *a priori*, so a parameterization of the clusters is needed.
- The number of clusters in the sample space is rarely known *a priori*.
- In real applications, well-defined clusters are atypical. Contamination of the data due to non-Gaussian noise and mixing with outliers makes the statistical identification and estimation of cluster parameters a difficult problem not only to solve but to formulate in a proper framework without using simplifying assumptions. However, Duda and Hart [3] convincingly argue that modeling observed data samples as being realizations from a Gaussian mixture distribution is an effective approximation for a variety of complex probability distributions.

A number of clustering algorithms have been developed and modified over the past several decades by researchers in the statistical and pattern recognition fields as well as applied areas like medical diagnosis, remote sensing, environmental science, psychology, manufacturing, marketing and finance [7]. One of the most widely used clustering algorithms is the K-means algorithm along with the variation known as ISODATA [7], [16]. Recently, the minimum volume ellipsoid (MVE) robust estimator was introduced in statistics [13]. The MVE estimator was extended by Jolion *et al.* [8] to a clustering method known as the general minimum volume ellipsoid (GMVE) clustering algorithm, and was successfully applied to several computer vision problems. The GMVE algorithm uses a random sampling approach in order to avoid combinatorial intractability. The reliability of the initial guess for the GMVE algorithm is also extremely important. Due to these two limitations of the GMVE algorithm, results of experiments for cluster analysis are compared only with the performance of the K-means algorithm.

Cluster analysis in a general sense can be considered to be a special case of mixture analysis. As argued by Duda and Hart [3], a great variety of data clusters can be approximated by assuming that data samples are realized from a Gaussian mixture density. The cluster identification problem does have certain additional requirements. For example, large clusters are preferred over many small clusters, and clusters with a large degree of overlap should be merged into a single cluster. Such perceptual constraints on the geometry of clusters can usually be translated into an appropriate statistical significance level in a normality test.

The application of the GMDD algorithm to cluster analysis is straightforward. The output of the GMDD algorithm will be a number of clusters and possibly an *unassigned set* containing samples that do not belong to any of the detected clusters. Depending upon the application, the unassigned set of observations can be discarded, or each sample from this set assigned to a cluster based on a criterion such as the $k-$nearest neighbor rule.

A. Cluster Analysis Using Simulated Data

Two-dimensional (2-D) mixture datasets are used for computer simulation experiments in order to illustrate the results clearly. Three experiments using three different mixture models are performed as discussed below.

Gaussian Mixture Densities: The first case compares the performance of the GMDD and K-means algorithms in reliably segmenting data generated from *noiseless* Gaussian mixture distributions with known parameters. Figs. 1 and 2 show the clustering results using noiseless data generated from mixtures with three and six distinct Gaussian components, respectively. The K-means algorithm requires an *a priori* estimate of the number of clusters to be identified. Even with the additional information on the correct number of mixture components and noiseless samples, the K-means algorithm is able to identify the clusters well only when the intercluster distance is large, as shown in Fig. 1. The three clusters making up classes B, C, and F are particularly poorly resolved by the K-means algorithm, as shown in Fig. 2. The GMDD algorithm is successful even when the clusters are not well separated as shown in Fig. 2. The initializations used by the K-means algorithm are true means in each case.

Contaminated Gaussian Mixture Densities: The second case compares the performance of the GMDD and K-means

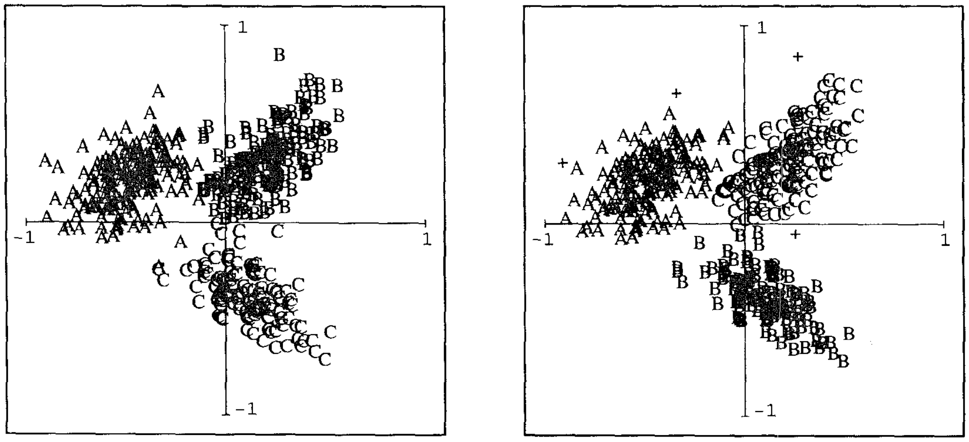


Fig. 1. Clustering by K-means (left) and GMDD algorithm (right): three clean Gaussian components.

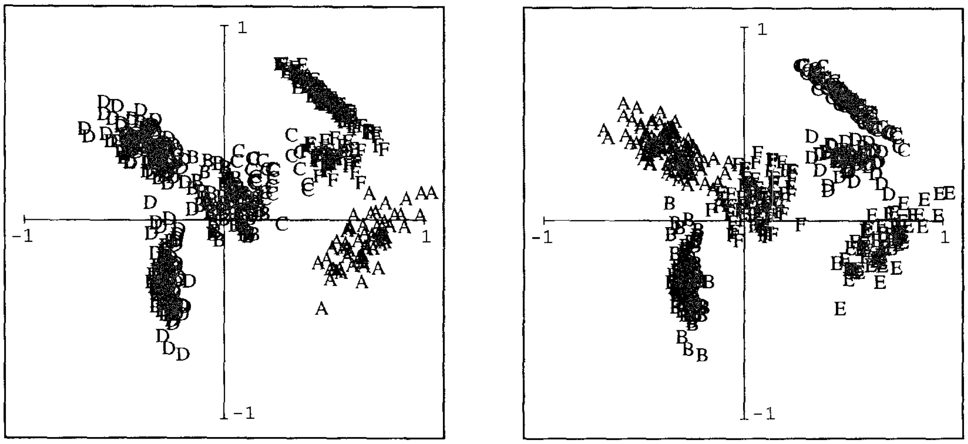


Fig. 2. Clustering by Kmeans (left) and GMDD algorithm (right): six clean Gaussian components.

algorithms in robustly grouping data generated from *noisy* Gaussian mixture distributions with known parameters. The contaminating noise is used to simulate the effects of outlier behavior. A very noisy background consisting of 200 data points uniformly distributed within the region defined by $[-1,1] \times [-1,1]$ is added to a pure Gaussian mixture with five components. The performance of the GMDD and K-means algorithms on the contaminated Gaussian mixture data can be compared in Fig. 3. The performance of the K-means algorithm has severely degenerated in that it tries to classify all of the corrupt data while the GMDD algorithm is much more robust as it tries to ignore outliers. The data values that are unassigned to any cluster are marked by the $+$ symbol in the results for the GMDD clustering. The unassigned set matches very closely the set of outliers contaminating the mixture model. The correct number of mixture components and true means as initializations are provided to the K-means algorithm in this case, too.

Special Case: A special case is designed to examine the sensitivity of the GMDD algorithm to the Gaussian distribution assumption. A pair of heavily overlapping Gaussian distributed clusters are constructed so that the two distributions form a single larger cluster; note that the combined cluster does not quite follow a Gaussian distribution. The GMDD algorithm is able to detect and estimate the combined cluster as a single cluster (see the cluster marked as "A" in Fig. 4) as the significance level used in the normality test is appropriately lowered.

IV. Unknown Probability Density Estimation

In this section, we apply the GMDD algorithm to modeling and estimating an unknown probability density in the practical context of automated cervical smear cell classification.

Estimating an unknown density function from data samples is fundamentally important in pattern recognition [3]. In practice, many pattern recognition systems are built by acquiring knowledge from training samples. Such knowledge usually represents the overall tendency or statistics of training samples, and many learning procedures thus involve estimating the probability density based on data samples. It is not exaggerated to say that the demand for unknown probability density estimation is tremendous. To date, however, there are only a few parametric or nonparametric techniques available. Moreover, most existing techniques are either computationally costly or too simple to fit the densities encountered in practice.

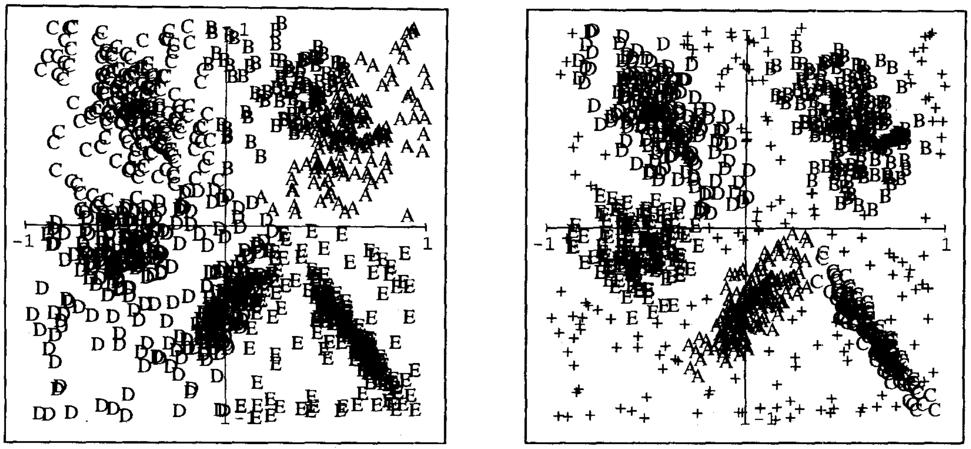


Fig. 3. Clustering by K-means (left) and GMDD algorithm (right): Five clean Gaussian components in a uniform noise background.

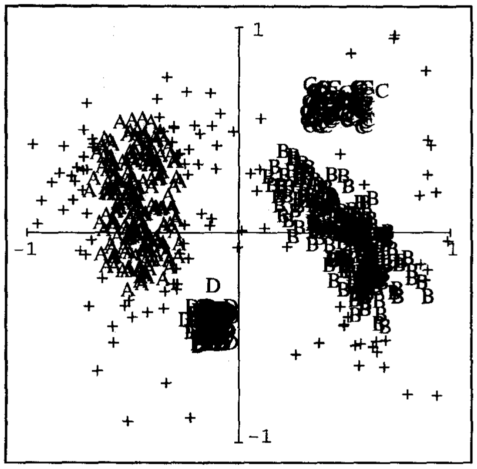


Fig. 4. The GMDD algorithm detects a combined cluster (marked by "A"), which actually does not quite follow Gaussian distribution.

As emphasized before, multimodal Gaussian mixture densities provide a realistic probability structure in modeling or fitting an unknown density function [15], [4]. The purpose of unknown probability density estimation is somehow different from that of cluster analysis. A compact and approximate representation of unknown probability density is of great concern. The proper organizing of components, such as separability between components, sizableness of each individual component, is not so important, provided the composition of those components is a good approximation of the actual probability density.

The automated cervical cancer slide prescreening has been the objective of extensive research over 30 years [10]. Basically, the goal is to design a machine that inspects the patient pap smear slides to detect abnormal or cancerous cells.

It is apparent that the most crucial component in the system is the cell classifier, whose objective is to distinguish a submitted object of interest among artifact, normal cell, and abnormal cell, denoted as C_0, C_1, C_2, respectively.

A widely accepted classification method is the Bayesian paradigm. It was shown that the Bayesian method is a very flexible and powerful framework for this particular system, being able to integrate much other diagnostic information [19].

Let $\mathbf{F}$ denote an M-dimensional feature vector extracted from an object of interest in the slides. Our goal is to design a Bayesian cell classifier so as to minimize the misclassification error probability, namely

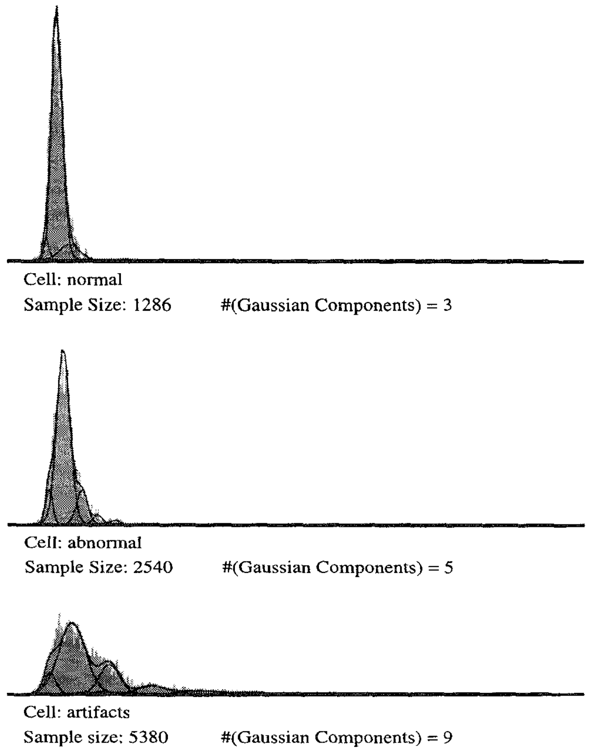
$$\arg \max_{0\leq j\leq 2} P(C_j|\underline{\mathrm{F}}) = \arg \max_{0\leq j\leq 2} P(C_j)P(\mathbf{F}|C_j) \qquad (18)$$

where $P(C_j)$ represents the occurring frequency of cell type C_j, which shall be evaluated in advance using clinical statistics. Hence, we need only to model each $P(\mathbf{F}|C_j)$, for $j = 0, 1, 2$, as a Gaussian mixture density and apply the GMDD algorithm to estimating them by using the training samples collected and labeled by cytotechnologists from the slides of a representative group of patients.

It is profitable to make some observations of this specific problem. The past experience indicates that the training sample set is noise contaminated. Thus, the significance level of the K–S test is reasonably relieved so that rough Gaussian densities can also pass the normality test.

For clarity, we show only the case of 1-D features. The actual Bayesian classifier implemented in the system uses four-dimensional (4-D) or five-dimensional (5-D) feature vectors. The computer experiments are arranged as follows. For each cell type C_j, $j = 0, 1, 2$, $P(\mathbf{F}|C_j)$ with $\mathbf{F}$ being a specific feature is estimated.

Figs. 5 and 6 show two sets of experimental results for two different cell features, i.e., the "compactness" and the "integrated optical density." The definition for those features and the method to measure them are omitted here. In the figures, each extracted Gaussian component is shown in light curves, and the overall estimate of each $P(\mathbf{F}|C_j)$ by composing its Gaussian components is drawn in solid curve. The training sample size and the detected number of Gaussian components in each case are also illustrated in the figures. Comparison to the raw data histogram reveals that the GMDD algorithm does estimate each mixture density very well even though the statistics of those training sample sets vary considerably. It is also seen that many noises in data samples are rejected in the reconstruction.

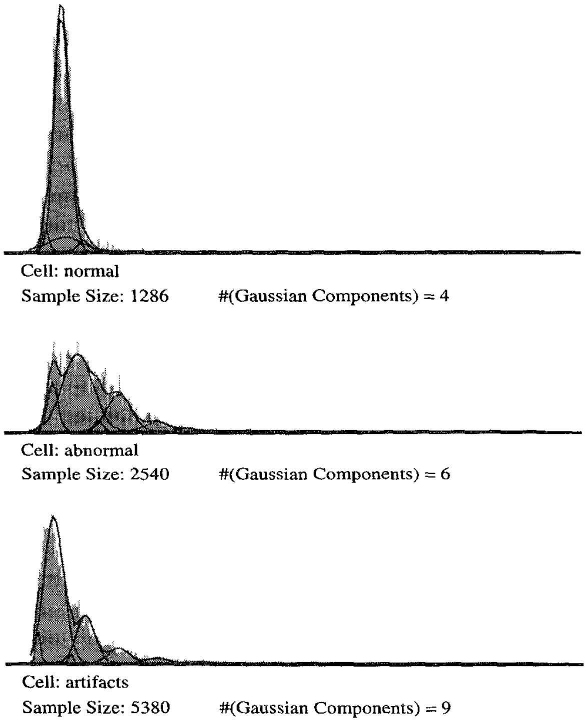


Cell: normal
Sample Size: 1286 #(Gaussian Components) = 3

Cell: abnormal
Sample Size: 2540 #(Gaussian Components) = 5

Cell: artifacts
Sample size: 5380 #(Gaussian Components) = 9

Fig. 5. Unknown probability density estimation using feature "compactness" for three cell types.



Cell: normal
Sample Size: 1286 #(Gaussian Components) = 4

Cell: abnormal
Sample Size: 2540 #(Gaussian Components) = 6

Cell: artifacts
Sample Size: 5380 #(Gaussian Components) = 9

Fig. 6. Unknown probability density estimation using feature "integrated optical density" for three cell types.

V. Conclusion

A robust statistical approach was applied to the problem of Gaussian mixture density modeling and decomposition. The robust estimation approach known as the recursive GMDD algorithm was developed and experimentally verified. The validity and effectiveness of the GMDD algorithm for applications such as robust clustering and unknown probability density estimation were demonstrated. Up to now, the GMDD algorithm has been tested on low-dimensional datasets but not high-dimensional data sets (i.e., 10 dimensions or more). In order to deal with high-dimensional datasets, it will likely be necessary to reduce the range of random initializations in order to maintain the efficiency of the GMDD algorithm.

Appendix A
Kolmogorov–Smirnov Normality Test

Suppose β is a random variable with observed samples $\beta_k, k = 1, \cdots, K$. Let $P(\overline{\beta})$ represent the known cumulative distribution function of β. Namely

$$P(\overline{\beta}) = \text{Prob}(\beta \leq \overline{\beta}).$$

An unbiased estimator $S_K(\overline{\beta})$ of $P(\overline{\beta})$ can be constructed based on the observed samples β_k, $k = 1, \cdots, K$, as follows:

$$S_K(\overline{\beta}) = \frac{1}{K}\#\{\beta_k\colon \beta_k \leq \overline{\beta}\}$$

Different data sets would provide different estimates of the cumulative distribution function. A number of statistics are available to measure the overall difference between a cumulative distribution function estimate and the known theoretical cumulative distribution function. The Kolmogorov–Smirnov (K–S) test statistic D is a particularly simple measure. The K–S statistic is defined as the maximum value of the absolute difference between the estimated and theoretical cumulative distributions. The K–S statistic for comparing a cumulative distribution function estimate $S_K(\overline{\beta})$ to the known cumulative distribution function $P(\overline{\beta})$, is

$$D = \max_{-\infty < \overline{\beta} < \infty} |S_K(\overline{\beta}) - P(\overline{\beta})|$$

What makes the K-S statistic attractive is that the distribution of D can be usefully approximated under the null hypothesis when the data sets are indeed drawn from the same distribution. The distribution of D can be used to assess the significance of any observed nonzero value of D.

The function that is needed for the calculation of the significance of any observed D is the limit of $\text{Prob}((\sqrt{\text{K}}\text{D} > \lambda)$ as $K \rightarrow \infty$ and can be written as

$$\begin{aligned} Q_{KS}(\lambda) &= \lim_{K \rightarrow \infty} \text{Prob}\{\sqrt{K}D > \lambda\} \\ &= 2\sum_{1}^{\infty}(-1)^{r-1}e^{-2r^2\lambda^2} \end{aligned}$$

which is a monotonic function with the limiting values

$$Q_{KS}(\infty) = 0, \quad Q_{KS}(0) = 1.$$

In terms of this function, the significance level of an observed value of D is approximately given by $Q_{KS}(\sqrt{K}D)$. The approximation becomes asymptotically accurate as K becomes large. A good approximation to $Q_{KS}(\sqrt{K}D)$ is given by the first term in the following series expansion:

$$Q_{KS}(\sqrt{K}D) = 2\sum_{r=1}^{\infty}(-1)^{r-1}e^{-2r^2KD^2}$$

That is

$$Q_{KS}(\sqrt{K}D) \approx 2e^{-2KD^2}.$$

due to the rapid convergence of the series. Therefore, for a significance level of α, the K–S statistic $D \geq \sqrt{(-1/2K)\ln\ \alpha/2}$ can be used to reject the null hypothesis.

Appendix B
Gaussian Distributed Cluster

Let G consist of K pattern vectors $x^1, \cdots, x^K$, each of which belongs to R_n. The question to be answered is: Are these K pattern vectors generated by a known Gaussian distribution $N(m, C)$? If the pattern vectors are, then the squared Mahalanobis distances $\beta_k = d^2(x^k), k = 1, \cdots, K$, represent K observed samples from a chi-square distribution with n degrees of freedom. The theoretical cumulative distribution function $P(\overline{\beta})$ for the chi-square distribution with n degrees of freedom is calculated using the relation

$$\overline{\beta} = \chi^2_{n,P(\overline{\beta})}.$$

The unbiased estimator $S_K(\overline{\beta})$ of $P(\overline{\beta})$ can be calculated by using the procedure in Appendix A. The K–S statistic can then be approximated by

$$D_K = \max_j |P(\overline{\beta}_j) - S_K(\overline{\beta}_j)|.$$

As a result, the significance level of the null hypothesis that the data set G is generated by $N(m, C)$ is approximated by $Q_{KS}(\sqrt{K}D_K)$.

The above testing procedure assumes that the mean vector m and the covariance matrix C are known. Quite often, the more practical question is whether the K pattern vectors were generated by a Gaussian distribution whose mean vector m and covariance matrix C each are unknown *a priori*? The unknown squared Mahalanobis distances $\overline{\beta}_k = d^2(x^k), k = 1, \cdots, K$, which represent K samples from a chi-square distribution with n degrees of freedom, need to be appropriately estimated first. The unbiased estimators of m and C are given by

$$m = \frac{1}{K}\sum_k x^k$$

$$C = \frac{1}{K-1}\sum_k (x^k - m)(x^k - m)'.$$

Using the above unbiased estimates for the mean vector and covariance matrix, each squared Mahalanobis distance $\overline{\beta}_k = d^2(x^k)$ can be straightforwardly estimated. Proceeding as before, the K–S statistic, D_K and the significance level $Q_{KS}(\sqrt{K}D)$ are each calculated appropriately. If the null hypothesis is supported using the above K–S test, then we would accept that the set of observations G describes a valid Gaussian distributed cluster.

References

[1] G. R. Dattatreya and L. N. Kanal, "Estimation of mixing probabilities in multiclass finite mixtures," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 1, pp. 149–158, 1990.

[2] H. Derin, "Estimating components of univariate Gaussian mixtures using Prony's method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, no. 1, pp. 142–148, 1987.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[4] K. Fukunaga, *Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[5] F. R. Hampel, E. M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley, 1986.

[6] P. J. Huber, *Robust Statistics.* New York: Wiley, 1981.

[7] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[8] J. M. Jolion, P. Meer, and S. Bataouche, "Robust clustering with applications in computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 8, pp. 791–802, 1991.

[9] J. N. Kapur, *Maximum-Entropy Models in Science and Engineering.* New York: Wiley, 1989.

[10] J. S. Lee *et al.*, "A processing strategy for automated pap smear screening," *Analyt. Quan. Cytol. Histol.*, vol. 14, no. 5, Oct., 1992.

[11] K. Palaniappan, "Multiresolution, adaptive methods, and classification for image analysis of dna autoradiographs," Ph.D. dissertation, Dept. Electr. Comput. Eng., Univ. Illinois, Urbana-Champaign, IL, 1991.

[12] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, 1984.

[13] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection.* New York: Wiley, 1987.

[14] R. Y. Rubinstein, *Monte Carlo Optimization, Simulation and Sensitivity of Queuing Networks.* New York: Wiley, 1986.

[15] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley, 1985.

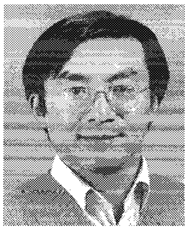
[16] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles.* Reading, MA: Addison-Wesley, 1974.

[17] X. Zhuang and R. M. Haralick, "A highly robust estimator for computer vision," in *Proc. 10th Int. Conf. Pattern Recog.*, Atlantic City, NJ, vol. 1, June 1990, pp. 545–550.

[18] X. Zhuang, T. Wang, and P. Zhang, "A highly robust estimator through partially likelihood function modeling and its application in computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 1, pp. 19–35, 1992.

[19] X. Zhuang, J. S. Lee, Y. Huang, and A. Nelson, "Staining independent Bayes classifier for automated cell pattern recognition," in *Proc. IS&T/SPIE Symp. Electron. Imaging: Sci. Tech.*, San Jose, CA, 1993.

[20] X. Zhuang and Y. Huang, "Robust 3D-3D pose estimation," in *Proc. 4th Int. Conf. Comput. Vision,* Berlin, Germany, 1993, pp. 567–571; also in *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 8, pp. 818–824, 1994.



Xinhua Zhuang (SM'92) is currently Professor of Electrical and Computer Engineering at the University of Missouri, Columbia. He has been a consultant to Siemens, Panasonic, NeoPath Inc., and NASA. He has been affiliated with a number of schools and research institutes including Hannover University, Germany, Zhejiang University, China, the University of Washington, USA, the University of Illinois, USA, the University of Michigan, USA, the Virginia Polytechnic Institute and State University of Virginia, USA, and the Research Institute of Computers. He has authored more than 150 articles in the areas of signal processing, speech recognition, image processing, machine vision, pattern recognition, and neural networks, and has been a contributor to six books.

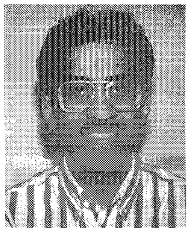
He has received a number of awards, including a NATO Advisory Group of Aerospace Research and Development (AGARD) fellowship, National Science Foundation grants of China, K. C. Wong Education Foundation grants of Hong Kong, National Science Foundation grants, and NASA HPCC grants.

Professor Zhuang serves as Associate Editor of IEEE Transactions on Image Processing.

Yan Huang received the B.E. degree in computer science and engineering from Zhejiang University, China, in 1990, and the M.S. and Ph.D. degrees (both in electrical engineering) from the University of Missouri, Columbia, in 1992 and 1995, respectively.

He worked for NeoPath, Inc., Amiable Technologies, Inc., and NASA Goddard Space Flight Center during the summers of 1992, 1993, and 1994, respectively. He is currently with AT&T Bell Laboratories. He has over 20 technical publications including 10 journal publications. His research interests include video and image compression, computer vision, image precessing, pattern recognition, and neural networks.

Dr. Huang received a Superior Graduate Achievement award from the University of Missouri in 1995.



K. Palaniappan (S'84, M'84, S'85, M'85, S'86, M'88, S'89, M'90) received the B.A.Sc and M.A.Sc degrees in systems design engineering from the University of Waterloo, UIUC, Waterloo, Canada, and the Ph.D degree in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 1991. As an undergraduate student he had work terms with various companies including Ontario Hydro, Canadian Ministry of Environment, Bell Canada, and Bell Northern Research. In 1981, he was selected as an IAESTE exchange student with Preussen Elektra in Landesbergen, Germany.

He has been a teaching and research assistant in the Department of Systems Design Engineering at University of Waterloo, in the Coordinated Science Laboratory at UIUC, in the Department of Electrical and Computer Engineering at UIUC, and in the Department of Microbiology at UIUC. In the summer of 1986, he was with Bell North Research in Montreal, Canada, working on document image analysis. Since 1991, he has been working at NASA Goddard Space Flight Center (GSFC) (affiliated with USRA as a senior research scientist).

Dr. Palaniappan received a NASA Outstanding Achievement Award (Mesoscale Dynamics and Precipitation Branch) in 1993. He received the University Space Research Association's Creativity and Innovation Science Award in 1993, National Science and Engineering Research Council of Canada Scholarship from 1982–1988, University of Waterloo Engineering Scholarship from 1982–1984, and an Ontario Scholar Award in 1978.

Yunxin Zhao (SM'94) received the B.S. degree in 1982 from Beijing Institute of Posts and Telecommunications, Beijing, China, and the M.S.E.E. and Ph.D. degrees in 1985 and 1988, respectively, from the University of Washington, Seattle.

She was with Speech Technology Laboratory, Panasonic Technologies, Inc., from October 1988 to August 1994, mainly working on speaker-independent continuous speech recognition. She is currently Assistant Professor with the Department of Electrical and Computer Engineering, Beckman Institute, and the Coordinated Science Laboratory at University of Illinois, Urbana-Champaign. Her research interests lie in the general area of human–computer interaction, with main focus on automatic speech recognition and related multidisciplinary research topics. She has performed research on computer network performance analysis, time-frequency signal analysis, speech and image processing, and recognition.