Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition

Lei Xu, Adam Krzyżak, *Member, IEEE,* and Ching Y. Suen, *Fellow, IEEE*

Abstract— **Method of combining the classification powers of several classifiers is regarded as a general problem in various application areas of pattern recognition, and a systematic investigation has been made. Possible solutions to the problem can be divided into three categories according to the levels of information available from the various classifiers. Four approaches are proposed based on different methodologies for solving this problem. One is suitable for combining individual classifiers such as Bayesian, *k-NN* and various distance classifiers. The other three could be used for combining any kind of individual classifiers. On applying these methods to combine several classifiers for recognizing totally unconstrained handwritten numerals, the experimental results show that the performance of individual classifiers could be improved significantly. For example, on the U.S. zipcode database, the result of 98.9% recognition with 0.90% substitution and 0.2% rejection can be obtained, as well as a high reliability with 95% recognition, 0% substitution and 5% rejection. These results compared favorably to other research groups in Europe, Asia, and North America.**

I. Introduction

RECENTLY, in the area of character recognition, the concept of combining multiple classifiers is proposed as a new direction for the development of highly reliable character recognition systems [1], and some preliminary results have indicated that the combination of several complementary classifiers will improve the performance of individual classifiers [1]–[4].

We believe that the combination of multiple classifiers is a general problem that is interesting not only to the character recognition area but also to various application areas of pattern recognition. The main reasons come from two aspects. First, in almost any one of the current pattern recognition application areas such as character recognition, speech recognition, remote sensing, geophysical prospecting and medical applications as well as many others [1]–[18], [29], there are a number of classification algorithms available. These algorithms are based on different theories and methodologies. Broadly speaking, we have now two large groups of methods, namely, feature-vector-based methods and syntactic-and-structural methods. Furthermore, each group includes many algorithms that are based on a variety of methodologies, e.g., for the first group alone, there exist Bayes classifier, *k-NN* classifier, various distance classifiers and neural network based classifiers ... etc. Usually, for a specific application problem, each of these classifiers could attain a different degree of success, but maybe none of them is totally perfect, or even not as good as expected for practical applications. So there is a need to study the methodology of integrating the results of a number of different classification algorithms so that a better result could be obtained. The second aspect is that for a specific recognition problem, usually numerous types of features could be used to represent and recognize patterns. To make a strong impression on this fact, some examples are presented as follows.

1) In character recognition, the usuable features may come from density of point measurements, moments, characteristic loci, mathematical transforms (Fourier, Walsh, Hadamard ...), they may also come from skeletons or contours (such as loop, endpoint, junction, arc, concavities and convexities, stroke ...) [1], [2].
2) In applications related to texture analysis such as remote sensing and scene analysis, the usable features may come from co-occurrence matrix, Fourier descriptors, power spectrum, moments, contrasts, as well as various structural primitives [7], [29].
3) In waveform analysis and recognition such as seismic signal, EEG and ECG, speech recognition and speaker identification, underwater acoustics as well as recognition of curve-like images, the usable features may come from power spectrum, AR modeling, function approximation, zero crossing, hidden Markov modeling, and many types of structural line segments [8]–[15].

No doubt, many other examples in various pattern recognition application areas can still be found.

These features are represented in very diversified forms, e.g., they may be continuous variables, binary values, discrete labels, structural primitives ... , it is very difficult to lump them together into one single classifier to make decision. As a result, many classifiers are needed to handle the different types of features. More specifically, there are three different cases where different ways of processing are required. In the first case, the features belong to types that are drastically different (e.g., continuous variables and structural primitives), and classifiers based on different theories and methodologies (e.g., feature based methods and syntactic methods) are needed to treat such features. This case was discussed earlier in the first aspect. In the second case, the features may be different not in the form of representation but also in the physical meanings, e.g., for a set of features that are represented in the form of continuous variables, suppose that some of them are

Manuscript received October 21, 1990; revised August 13, 1991. This work was supported in part by The Natural Sciences and Engineering Research Council of Canada, and in part by the National Networks of Centres of Excellence program of Canada.

The authors are with the Centre for Pattern Recognition and Machine Intelligence, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, PQ H3G 1M8, Canada.

IEEE Log Number 9104551.

0018-9472/92$03.00 © 1992 IEEE

features representing the pattern's volume, some representing the pattern's temperature, etc., then to lump these features into a vector for one classifier, we need first to normalize the scales of these features. This job is usually also quite difficult. However, it may become much easier if we integrate the results of several classifiers each of which uses features representing the same type of physical property and thus being of the same scale. Finally, in the third case, even if normalization is not required in lumping a lot of variables into one very high-dimensional vector, it may still be a good idea to divide the high-dimensional vector into several vectors with lower dimensions as input to several classifiers, since it is well known that high-dimension vectors will not only increase computational complexity but will also produce implementation problems and accuracy problems.

Of course, the combination of multiple classifiers is by no means the only solution for the problem involving a variety of features. The hybrid systems of statistical and syntactical or structural methods have been developed [5], [16] to cover the first case mentioned previously. Various ways of feature selection and dimension reduction [19] have been proposed to solve the problems associated with the third case. Multistage system [20], multilevel hierarchical classifier [21], especially tree classifiers [22], [23] have been investigated extensively. However, even though they may succeed up to a certain degree and certain aspects, they too encounter many of their own difficulties. Thus, a new direction of combining multiple classifiers is certainly worthwhile to be explored further.

It appears that the study on the combination problem is now only at its preliminary stage. It has started in the character recognition area. Due to the complexity of handwritten character recognition, recently it has been realized that classifiers based on different methodologies or different features are usually complementary to each other [16], [17]. Thus efforts have been made to develop various complementary classifiers or those called expert modules [1], [2], [5] for handwritten character recognition. Then it naturally raises the question of obtaining a consensus on the results of each individual classifier or expert. Presently, three kinds of effort have been made toward this direction. One makes use of the majority voting principle [1], [2], i.e., each individual classifier represents one score that is either as a whole assigned to one class label or divided into several labels. The label, which receives more than half of the total scores, is taken as the final result. The second uses a kind of a candidate subset combining and re-ranking approach [3], [4], namely, each individual classifier produces a subset of ranked candidate labels, and the labels in the union of all subsets are re-ranked based on their old ranks in each subset. The third applies Dempster–Shafer (D–S) Theory on the special case of combining several individual distance classifiers [6]: the distance calculated by each individual classifier is transformed by some way into a confidence value between [0, 1], which is used as the basic probability assignment of the only one focal element, the simplest combining rule based on D–S theory is used in the special case to combine the contribution of each individual to give the final result. The results of these efforts on handwritten character recognition or on line script recognition are quite interesting and inspiring [1]–[6].

In this paper, we propose to conduct a more systematical investigation into the problem of multiclassifier combination. Generally speaking, we could consider that it consists of two parts. The first part, being closely dependent on the specific applications, includes the problems of "How many classifiers are chosen for a specific application problem? What kind of classifiers should be used? And for each classifier what types of features should be chosen?", as well as other problems that relate to the construction of those individual and complementary classifiers. Papers [1], [2] have already described a lot of work on the recognition of totally unconstrained handwritten numerals. This paper does not intend to study the problems of this part. The second part, which is general and common to various applications, includes the problems related to the question—how to combine the results of different existing classifiers so that a better result can be obtained. This paper will concentrate on problems related to this second part.

In Section II, we first summarize the problems of combining multiclassifiers into three categories according to the levels of information produced by various classifiers. Then in the following section, several approaches have been proposed to tackle these problems. These approaches include new versions as well as a general form of voting principle, an averaged Bayes classifier and its version, a combination approach in Bayesian formalism, and a combination approach in Dempster–Shafer formalism. The latter two especially are new approaches adapted from the literature of evidence gathering and uncertainty reasoning. The approaches proposed in this paper are applied to the problems of recognizing totally unconstrained handwritten numerals, the four experts presented in [1], [2] are used as the individual classifiers. The obtained combination results are significantly better than any individual classifier, e.g., on the same database as [1], [2] and by combining the four individual classifiers given there, the result could give 98.9% recognition, 0.9% substitution and 0.2% rejection. If it is required to suppress the substitution rate, the results could give 95% recognition, 0% substitution and 5% rejection; while the individual classifier with the best performance among the four can only provide the result of 93.9% recognition, 1.6% substitution and 4.5% rejection.

II. The Problem of Combining Multiple Classifiers

A. Three Levels in Classifier's Output Information

Given a pattern space P consisting of M mutually exclusive sets $P = C_1 \cup \cdots \cup C_M$ with each of C_i, $\forall i \in \Lambda = \{1, 2, \cdots M\}$ representing a set of specified patterns called a class (e.g., $M = 10$ for the problem of numerals recognition). For a sample x from P, the task of a classifier (denoted e) is to assign x one index $j \in \Lambda \cup \{M+1\}$ as a label to represent that x is regarded as being from class C_j if $j \neq M+1$, with $j = M+1$ denoting that e has no idea about which class x comes from, or in other words, x is rejected by e. Regardless what internal structure a classifier has and on what theory and methodology it bases, we may simple regard a classifier as a

function box that receives an input sample x and outputs a label j, or in short denoted by $e(x) = j$.

Although j is the output information we only want at the final stage of classification, practically many of the existing classification algorithms usually supply or are able to supply some other related information. For example, a Bayes classifier may also supply M values of post-probabilities $P(i/x), i = 1, \cdots, M$ for each possible label. In fact, the final label j is the result of maximum selection from the M values and this selection certainly discards some information that is considered useless for the final output when there is only a single classifier. However such discarded information may be useful for multiclassifier combination. Depending on whether some output information other than one label j is used and the other kind of information is used, we will have different types of multiclassifier combination problems.

Generally speaking, the output information that various classification algorithms supply or are able to supply can be divided into three levels.

1) The abstract level: a classifier e only outputs a unique label j, or for some extension, e outputs a subset $J \subset \Lambda$.
2) The rank level: e ranks all the labels in Λ or (a subset $J \subset \Lambda$) in a queue with the label at the top being the first choice.
3) The measurement level: e attributes each label in Λ a measurement value to address the degree that x has the label.

Among the three levels, the measurement level contains the highest amount of information and the abstract level contains the lowest. From the measurements attributed to each label, we could rank all the labels in Λ according to a rank rule (e.g., ascending or descending). By choosing the label at the top rank, or directly by choosing the label with the maximal or minimal value at the measurement level, we can assign a unique label to x. In other words, from the measurement level to the abstract level there is an information reduction process or abstraction process.

Many classification algorithms are able to supply output information from the measurement level, e.g., Bayes classifier supplies the post-probabilities $P(i/x), i \in \Lambda$ and various distance classifiers supply the distance between x and each prototype sample of each class as the measurements. In other words, processing at the measurement level is an intermediate stage of many classifiers. However, some of classifiers may be able to supply the output information only from the abstract level, e.g., the pure syntactic classifier.

B. Three Types of Problems for Multiple Classifier Combination

According to which of the aforementioned three output information levels a combination is based upon, various problems of combining multiple classifiers could be summarized into the following three types:

Type 1: The combination is made based on the output information of the abstract level. Given K individual classifiers e_k, $k = 1, \cdots, K$ each of which assigns input x to a label j_k, i.e., produces an event $e_k(x) = j_k$, the problem is to use these events to build an integrated classifier E, which gives x one definitive label j, i.e., $E(x) = j$, $j \in \Lambda \cup \{M+1\}$.

Type 2: The combination is made based on the output information of the rank level. For an input x, each e_k produces a subset $L_k \subseteq \Lambda$ with all the labels in L_k ranked in a queue, the problem is to use these events $e(x) = L_k$, $k = 1, \cdots, K$ to build an E with $E(x) = j$, $j \in \Lambda \cup \{M+1\}$.

Type 3: The combination is made based on the output information of the measurement level. For an input x, each e_k produces a real vector $M_e(k) = [m_k(1), \cdots, m_k(M)]^t$ (where $m_k(i)$ denotes a kind of degree that e_k considers that x has label i), the problem is to use these events $e(x) = M_e(k)$, $k = 1, \cdots, K$ to build an E with $E(x) = j$, $j \in \Lambda \cup \{M+1\}$.

The three types of problems described previously cover the different scopes of applications. On the problem of Type 1, the individual classifiers could be very different from each other in their theories or methodologies (e.g., e_k may base on a statistical method, while e_l on a syntactic method). In fact, any kind of classifier will at least supply the output information at the abstract level, so it could be said that the problem of Type 1 covers all kinds of pattern recognition areas. Thus, this type of problem should be most interesting. In contrast, the problem of Type 3 requires that all the individual classifiers should be able to supply the output information at the measurement level. Furthermore, if there are any measurement vectors of different kinds (say, $M_e(k)$ is a vector of postprobabilities, while $M_e(l)$ a vector of some kind of distances), the measurements should be able to be transformed into the same kind of measurement, since a reasonable combination operation on these measurements could be made only when they have the same measure scale. The problem of Type 2 has a generality between Type 1 and Type 3, it requires that all the individual classifiers be able to supply the output information at the rank level. Thus, its individual classifier could not be a pure syntactic classifier that only outputs one label, but could be any classifier that is able to supply output information at the measurement level since a ranked list L_k could be easily obtained from the correspondent measurement vector $M_e(k)$.

The combination problem studied in [1], [2] belongs to Type 1, and the problem studied in [3], [4] belongs to Type 2. In this paper, we will study the problems of Type 3 in Section III. Then, in the following four sections, we concentrate on the problem of Type 1 because we think that it is the most useful one due to its generality.

III. Averaged Bayes Classifier and Its Versions

A. Averaged Bayes Classifier

We use this section to discuss the combination problem of Type 3. First, we look at a special case that all individual classifiers are Bayes classifiers.

For a Bayes classifier e, its classification of an input x is actually based on a set of real value measurements–postprobabilities:

$$P(x \in C_i/x), i = 1, \cdots, M \tag{1}$$

where $x \in C_i$ denotes that x comes from class C_i. In the convention of statistical pattern recognition literature, these probabilities are simply denoted by $P(C_i/x)$, $\forall i \in \Lambda$. They represent the probabilities that x comes from each of the M classes under the condition x.

In principle, these probabilities are not related to each classifier e_k. But in practice, that each e_k classifies x is not really based on those true values of (1), which are not available. Instead, for each x, e_k estimates by itself a set approximations of those true values. These approximations depend on what features e_k are used and how e_k is trained. To clarify such a dependence, we denote them as follows:

$$P_k(x \in C_i/x), i = 1, \cdots, M, k = 1, \cdots, K. \tag{2}$$

For any e_k, a definitive decision is made as

$$e_k(x) = j \text{with } P_k(x \in C_j/x) = \max_{i \in \Lambda} P_k(x \in C_i/x). \tag{3}$$

Now, we don't care about the results of (3). Instead, we use the approximations of (2) for combining the classification results on the same x by all K classifiers. One simple approach here we propose is to use the following average value as a new estimation of combined classifier E:

$$P_E(x \in C_i/x) = \frac{1}{K} \sum_{k=1}^{K} P_k(x \in C_i/x), i = 1, \cdots, M. \tag{4}$$

The final decision made by this E is given by

$$E(x) = j, \text{ with } P_E(x \in C_j/x) = \max_{i \in \Lambda} P_E(x \in C_i/x) \tag{5}$$

that is, a Bayes decision is based on these newly estimated post-probabilities. So, we call such a combined E as an averaged Bayes classifier. If we expect that the classified results are more reliable, we could use the following equation to replace (5) to take into account the trade-off between the substitution rate and the rejection rate in (6) (shown at the bottom of the page) with $0 \leq \alpha \leq 1$ being a threshold.

Another alternative is to use the median value of $P_k(x \in C_i/x)$, $i = 1, \cdots, M$, denoted by $P_m(x \in C_i/x)$, to replace the correspondent average value. Since $\sum_{i=1}^{M} P_m(x \in C_i/x) \neq 1$, we use the following normalized values as the new estimations:

$$P_E(x \in C_i/x) = \frac{P_m(x \in C_i/x)}{\sum_{i=1}^{M} P_m(x \in C_i/x)}. \tag{7}$$

B. Extensions to Other Classifiers

The previous approach could be extended to cover several cases when some e_k's belong to another kind of classifiers.

First we consider the case that e_k is a $k - NN$ classifier. In this case, the classification process consists of two steps. The first one is to find the k_{nn} nearest prototype samples to the present input x with

$$k_{nn} = \sum_{i=1}^{M} k_i, \quad k_i \geq 0$$

where k_i represents the number of prototype samples from class C_i. The second step is to classify x into class C_j according to $k_j = \max_i k_i$, that is,

$$E(x) = j, \quad \text{when } k_j = \max_{i \in \Lambda} k_i. \tag{8}$$

Since the measurements k_i, $i = 1, \cdots, M$ have a different scale from the measurements in the form of post-probabilities, it is not reasonable to use (5) directly. The following formula introduces one way to transform k_i's into the approximations as

$$P_k(x \in C_i/x) = \frac{k_i}{k_{nn}}, i = 1, \cdots, M \tag{9}$$

and then these approximations could be put into (5) for the subsequent combination computing.

Second, we consider the case e_k is some kind of distance classifier, i.e., for each x, e_k classifies x according to some distance measures (e.g., Euclidean, Mahalanobis, and other pseudodistances etc.) $d_k(i)$ between x and the centers (or prototypes) of each class C_i, $i = 1, \cdots, M$. If one could design some functions

$$p_k(i) = f_i(d_k(i), i = 1, \cdots, M) \tag{10}$$

for example:

$$p_k(i) = \frac{1/d_k(i)}{\sum_{i=1}^{M} 1/d_k(i)} \tag{11}$$

to derive a set of $p_k(i)$'s which obey the three basic axioms of probability theory, one could use these $p_k(i)$ as apparent post-probabilities and put them into (4) for combination.

Generally, any classifiers in which some kind of apparent post-probabilities are computable could be combined by means of (5).

IV. Combining Multiple Classifiers By Voting Principle

A. The Earlier Works

From now on, this paper will concentrate on the combination problem of Type 1 since this type is the most general and useful one.

As indicated in Section II, the problem is to produce a new event $E(x) = j$ from the given events $e_k(x) = j_k$, $k = 1, \cdots, K$, where the following equation may not necessarily hold:

$$e_1(x) = e_2(x) = \cdots = e_K(x). \tag{12}$$

$$E(x) = \begin{cases} j, & \text{if } P_E(x \in C_j/x) = \max_{i \in \Lambda} P_E(x \in C_i/x) \leq \alpha \\ M+1, & \text{otherwise} \end{cases} \tag{6}$$

That is, conflicts may exist among the decisions of K classifiers. A simple and common rule used for resolving this kind of conflicts in human social life is voting by majority. This rule has been adapted for multiclassifier combination by [1], [2], [5] in the recognition of unconstrained numerals.

In [5], eleven individual classifiers are proposed based on template matching, structural and statistical methods respectively. If six out of 11 vote for the same label, then the label is taken as the final result. In [2], two classifiers are used based on structural features extracted from skeleton and contour respectively. Three rules specific for the combination of the two classifiers are proposed. These rules maybe regarded as the special examples of the majority voting rule. In [1], two classifiers are added to [2]. Among the four, each of the last three outputs only one label, i.e., $e_2(x) = j_2$, $e_3(x) = j_3$, $e_4(x) = j_4$; while the first one outputs a subset of labels, i.e., $e_1(x) = J$ with $\#|J| \leq 3$. The e_k, $k = 2, 3, 4$ represents one vote that is assigned to its output label j_k; while $e_1(x)$ divides one vote into $\#|J|$ fractions with each label in J receiving $1/(\#|J|)$ vote. The decision is made such that the label that receives more than half of the votes (i.e., two) is taken as the final output.

B. Variants of Voting Principle and a General Expression

For convenience, we represent the event $e_k(x) = i$ in the form of a binary characteristic function:

$$T_k(x \in C_i) = \begin{cases} 1, & \text{when } e_k(x) = i \text{ and } i \in \Lambda \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The most conservative voting rule is the following

$$E(x) = \begin{cases} j, & \text{if } \exists j \in \Lambda, \cap_{k=1}^{K} T_k(x \in C_j) > 0 \\ M+1, & \text{otherwise} \end{cases} \tag{14}$$

that is, the combined classifier E decides that x comes from C_j iff all the K classifiers decided that x comes from C_j simultaneously, otherwise it rejects x. In (15), "$\cap$" denotes the operator of logical AND or binary multiplication, and in the following (15), "$\cup$" denotes the operator of logical OR or binary summation.

A slight modification of (14) could lead to a version that is less conservative. The version is shown in (15) (at the bottom of the page), which results in an E that decides $x \in C_i$ as long as some classifiers support $x \in C_i$ and no other classifier supports a different $x \in C_j$, $j \neq i$. Or in other words, (15) means that the classifiers that reject x have no impact on the combined E unless all the classifiers reject x.

The majority voting rule used in [1] could be expressed by the following formula:

$$E(x) = \begin{cases} j, & \text{if } T_E(x \in C_j) = \max_{i \in \Lambda} T_E(x \in C_i) > \frac{K}{2} \\ M+1, & \text{otherwise.} \end{cases} \tag{16}$$

where

$$T_E(x \in C_i) = \sum_{k=1}^{K} T_k(x \in C_i), i = 1, \cdots, M. \tag{17}$$

By slightly modifying this formula, a more general version is established as follows

$$E(x) = \begin{cases} j, & \text{if } T_E(x \in C_j) = \max_{i \in \Lambda} T_E(x \in C_i) \geq \alpha * K \\ M+1, & \text{otherwise.} \end{cases} \tag{18}$$

where $0 < \alpha \leq 1$. Note that (16) is the special case of (18) with $\alpha = 0.5 + \epsilon$, and $\epsilon > 0$ is arbitrarily small. Equation (12) is equivalent to the special case of (18) with $\alpha = 1.0$.

In (18), the thresholding operation only considers that the maximal votes of the final selected label must be large enough. There may exist cases that there are more than two labels that receive the maximal vote or the vote of the maximal are not considerably larger than the vote of the second maximal. In these cases, even the maximal vote of the final selected label may be quite large, the decision still may not be reliable since there exists an opponent that may also receive a large vote. To tackle this problem, a new majority voting rule is proposed in (19) (shown at the bottom of the page) and (20):

$$\begin{aligned} \max_1 &= \max_{i \in \Lambda} T_E(x \in C_i) \\ \max_2 &= \max_{i \in \Lambda - \{j\}} T_E(x \in C_i) \end{aligned} \tag{20}$$

where $0 < \alpha \leq 1$. Since K, the number of classifiers, is constant, the votes of $\max_2$ could be regarded as the implicit objections to the label j. Thus, rule (19) in fact requires that the pure supports received by the finally selected label must be large enough. It is not difficult to see that rule (15) is equivalent to the special case of (18) with $\max_2 = 0$.

All the aforementioned variants could be included in a general expression as

$$E(x) = \begin{cases} j, & \text{if } T_E(x \in C_j) = \max_1 \geq \alpha * K + d_t(x) \\ M+1, & \text{otherwise.} \end{cases} \tag{21}$$

$$E(x) = \begin{cases} j, & \text{if } \exists j \in \Lambda, \cap_{k=1}^{K} \{T_k(x \in C_j) \cup (1 - \cup_{q=1}^{M} T_k(x \in C_q)\} > 0 \\ M+1, & \text{otherwise} \end{cases} \tag{15}$$

$$E(x) = \begin{cases} j, & \text{if } T_E(x \in C_j) = \max_1 \text{ and } \max_1 - \max_2 \geq \alpha * K \\ M+1, & \text{otherwise.} \end{cases} \tag{19}$$

where $d_t(x)$ is a function of $T_E(x \in C_i)$, $i = 1, \cdots, M$. Rule (17) is a special case of (21) with $d_t(x) = 0$ and rule (20) is a special case of (21) with $d_t(x) = \max_2$. $d_t(x)$ could also be another function, e.g., the median value of $T_E(x \in C_i)$, $i = 1, \cdots, M$. Observe that the threshold of (21) consists of two parts: a constant part that is independent of x, and a dynamic one that varies with input x.

Finally, we should also point out that at the beginning of the subsection, although we only let $T_k(x \in C_i)$ be the event of type $e_k(x) = i$, i.e., each classifier outputs a single label. For the general case $e_k(x) = J$ with J being a subset of labels (e.g., expert no. 1 in [1]), all the previous equations also apply by defining a nonbinary characteristic function for the event $e_k(x) = J$ as follows:

$$T_k(x \in C_i) = \begin{cases} \frac{1}{\#|J|}, & \text{when } i \in J \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

V. The Combination of Multiple Classifiers in Bayesian Formalism

A. Confusion Matrix, Prior Knowledge, and Beliefs

In the previous section, those voting methods that combine the results of individual classifiers are only based on the label outputted by each classifier (i.e., the event $e_k(x) = j$). Each of $e_k(x) = j_k$'s is equally treated as one vote without considering the error of each e_k itself. This and the next section will take these errors into consideration.

The errors of each classifier e_k are usually described by its confusion matrix that is given by

$$PT_k = \begin{pmatrix} n_{11}^{(k)} & n_{12}^{(k)} & \cdots & n_{1M}^{(k)} & n_{1(M+1)}^{(k)} \\ n_{21}^{(k)} & n_{22}^{(k)} & \cdots & n_{2M}^{(k)} & n_{2(M+1)}^{(k)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{M1}^{(k)} & n_{M2}^{(k)} & \cdots & n_{MM}^{(k)} & n_{M(M+1)}^{(k)} \end{pmatrix} \tag{23}$$

for $k = 1, 2, \cdots, K$; where each row i corresponds to class C_i and each column j corresponds to the event $e_k(x) = j$. Thus, an element $n_{ij}^{(k)}$ denotes that $n_{ij}^{(k)}$ samples of class C_i have been assigned a label j by e_k.

The confusion matrix PT_k of a trained classifier e_k could be obtained by using e_k to classify a test sample set that reflects the distribution of pattern space. It follows from (23) that the total number of samples in the test set is

$$N^{(k)} = \sum_{i=1}^{M} \sum_{j=1}^{M+1} n_{ij}^{(k)} \tag{24}$$

in which the number of samples in each class C_i is

$$n_{i\cdot}^{(k)} = \sum_{j=1}^{M+1} n_{ij}^{(k)}, i = 1, \cdots, M \tag{25}$$

and the number of samples that are assigned j by e_k is

$$n_{\cdot j}^{(k)} = \sum_{i=1}^{M} n_{ij}^{(k)}, j = 1, \cdots, M+1. \tag{26}$$

For an event $e_k(x) = j$ of an error-bearing classifier e_k, its truth (i.e., x does come from class C_j) has uncertainty. With the knowledge of its confusion matrix PT_k, such an uncertainty could be described by the conditional probabilities that propositions $x \in C_i$, $i = 1, \cdots, M$ are true under the occurrence of the event $e_k(x) = j$, that is

$$P(x \in C_i / e_k(x) = j) = \frac{n_{ij}^{(k)}}{n_{\cdot j}^{(k)}} = \frac{n_{ij}^{(k)}}{\sum_{i=1}^{M} n_{ij}^{(k)}}, i = 1, \cdots, M. \tag{27}$$

From another viewpoint, the confusion matrix PT_k could be regarded as the prior knowledge of an expert. Upon receipt of the evidence–the occurrence of event $e_k(x) = j$, the expert expresses his beliefs with uncertainty on each of M mutually exclusive propositions $x \in C_i$, $\forall i \in \Lambda$ by a real numeral bel(.) called belief value. The higher the bel(.) he gives to a proposition, the more likely it is true. With the knowledge of PT_k, he expresses his bel(.)'s on each proposition $x \in C_i$ in the form of a conditional probability as given by (27), viz.:

$$\text{bel}(x \in C_i / e_k(x), EN) = P(x \in C_i / e_k(x) = j_k), i = 1, \cdots, M. \tag{28}$$

That is, bel(.) is defined as the probability under the condition of $e_k(x) = j_k$ and the environment EN. Where EN denotes the common classification environment that consists of any events that are independent of any of events $e_k(x) = j_k$, $k = 1, \cdots, K$, e.g., the environment at least contains the occurrence of a specific input pattern x.

Such a belief expression given in (28) is exactly that used by Pearl [24] as well as others who adopt Bayesian formalism for evidence gathering and uncertainty reasoning in AI literature. In his recent book [24], Pearl described that "in this formalism, propositions are given numeral parameters signifying the degree of belief accorded to them under some body of knowledge, and the parameters are combined and manipulated according to the rules of probability theory." The advantages of adopting Bayesian formalism and various methods for manipulating uncertainty reasoning along the formalism have been studied extensively in [24]. In our case, there are M propositions $x \in C_i$, $\forall i \in \Lambda$, the numeral parameters are the conditional probabilities given by (27) and the body of knowledge consists of event $e_k(x)$, matrix PT_k as well as the environment EN.

B. Belief Integration Based on Bayesian Formula

With K classifiers $e_1, \cdots, e_K$, we will have K matrices $PT_1, \cdots, PT_K$. When these classifiers are used on the same input x, K events $e_k(x) = j_k$, $k = 1, \cdots, K$ will happen. As discussed previously, each $e_k(x) = j_k$ and its corresponding PT_k could supply a set of bel$(x \in C_i / e_k(x), EN)$, $i = 1, \cdots M$, each of which supports one of the M propositions. A natural question is how to integrate these individual supports

to give the combined values in (29) (shown at the bottom of the page).

From (28) and (29), we have (30) (shown at the bottom of the page).

If classifiers $e_1, \cdots, e_K$ perform independent of each other (e.g., we can consider that classifiers are independent when they use independent feature sets, or they are trained by independent training sets), then the events $e_1(x) = j_1, \cdots, e_K(x) = j_K$ will be independent of each other either under the condition of $x \in C_i$ as well as EN or the condition of solely EN. Thus we have

$$\frac{P(e_1(x) = j_1, \cdots, e_K(x) = j_K / x \in C_i, EN)}{P(e_1(x) = j_1, \cdots, e_K(x) = j_K / EN)} = \frac{\prod_{k=1}^{K} P(e_k(x) = j_k / x \in C_i, EN)}{\prod_{k=1}^{K} P(e_k(x) = j_k / EN)} = \frac{\prod_{k=1}^{K} P(x \in C_i / e_k(x) = j_k)}{\prod_{k=1}^{K} P(x \in C_i / EN)}.$$

Since

$$\frac{P(e_k(x) = j_k / x \in C_i, EN)}{P(e_k(x) = j_k / EN)} = \frac{P(x \in C_i / e_k(x) = j_k)}{P(x \in C_i / EN)}$$

by putting the previous into (30), we have

$$\text{bel}(i) = P(x \in C_i / EN) \frac{\prod_{k=1}^{K} P(x \in C_i / e_k(x) = j_k)}{\prod_{k=1}^{K} P(x \in C_i / EN)} \tag{31}$$

where $P(x \in C_i / e_k(x) = j_k)$ could be estimated by (27) with j being replaced by j_k, $P(x \in C_i / EN)$ represents the probability that $x \in C_i$ is true under occurrence of x and the common environment EN. It should be noticed that the occurrence of x is a necessary condition for events $e_k(x) = j_k$, $k = 1, \cdots, K$ and thus should be placed behind all the condition bars "/" in (27)–(29) and (30) and (39).

A better estimation of $P(x \in C_i / EN)$ should be the postprobabilities $P(x \in C_i / x)$. This means that (31) provides an alternative way to solve the combination problem of Type 3, which we have studied in Section III. The alternative way is not intended to be further studied in this paper. For the purpose of this paper, i.e., the combination problem of Type 1, the postprobabilities $P(x \in C_i / x)$'s are not available. For practical implementation, we use the following (32) as an approximation of (31):

$$\text{bel}(i) = \eta \prod_{k=1}^{K} P(x \in C_i / e_k(x) = j_k) \tag{32}$$

with η as a constant that ensures that $\sum_{i=1}^{M} \text{bel}(i) = 1$ (since $x \in C_i$, $i = 1, 2, \cdots, M$ are mutually exclusive and exhaustive). That is, we have

$$\frac{1}{\eta} = \sum_{i=1}^{M} \prod_{k=1}^{K} P(x \in C_i / e_k(x) = j_k). \tag{33}$$

Finally, depending on these bel(i) values, we can classify x into a class according to the decision rule given here:

$$E(x) = \begin{cases} j, & \text{if bel}(j) = \max_{i \in \Lambda} \text{bel}(i); \\ M+1, & \text{otherwise.} \end{cases} \tag{34}$$

In making the trade-off between the substitution rate and the rejection rate, (34) could be modified into (35)

$$E(x) = \begin{cases} j, & \text{if bel}(j) = \max_{i \in \Lambda} \text{bel}(i) \geq \alpha; \\ M+1, & \text{otherwise.} \end{cases} \tag{35}$$

where $0 < \alpha \leq 1$ is a threshold.

VI. The Combination of Multiple Classifiers in Dempster–Shafer Formalism

In this section, we study the combination problem of Type 1 to consider the errors of individual classifiers by adapting Dempster–Shafer's evidence theory. The combination is made in the situation that only the recognition, substitution and rejection rates of each individual classifier are used as the prior knowledge. These rates, which usually represent the performance indexes of a classifier, are easily obtained by testing the classifiers with a test sample set.

A. Dempster–Shafer Theory

For convenience, we first briefly introduce the key points of Dempster–Shafer theory.

Given a number of exhaustive and mutually exclusive propositions A_i, $i = 1, \cdots, M$, which form a universal set $\Theta = \{A_1, \cdots, A_M\}$. A subset $\{A_{i_1}, \cdots, A_{i_q}\} \subset \Theta$ represents a proposition denoting the disjunction $A_{i_1} \cup \cdots \cup A_{i_q}$. Each element $A_i \subset \Theta$ corresponds to a one-element subset $\{A_i\}$, called a singleton. All the possible subsets of Θ form a superset 2^{Θ}, i.e., each subset $A \subset \Theta$ is an element of 2^{Θ}, i.e., $A \in 2^{\Theta}$.

The D–S theory uses a numeric value in the range [0, 1] inclusive to indicate belief in a proposition (subset) $A \subset \Theta$ based on the occurrence of an evidence e. This value, conventionally denoted by bel(A), indicates the degree to which the evidence e supports the proposition A. The value of bel(A) is calculated from another function called a basic probability assignment (BPA), which represents the individual impact of each evidence on the subsets of Θ. A BPA (denoted m) is a generalization of a probability mass distribution. It

$$\text{bel}(i) = \text{bel}[x \in C_i / e_1(x), \cdots, e_K(x), EN] = P[x \in C_i / e_1(x) = j_1, \cdots, e_K(x) = j_K, EN], i = 1, \cdots, M \tag{29}$$

$$\text{bel}(i) = P(x \in C_i / e_1(x) = j_1, \cdots, e_K(x) = j_K, EN) = \frac{P(e_1(x) = j_1, \cdots, e_K(x) = j_K / x \in C_i, EN) P(x \in C_i / EN)}{P(e_1(x) = j_1, \cdots, e_K(x) = j_K / EN)} \tag{30}$$

assigns values in [0, 1] to every element of 2^Θ (i.e., each subset of Θ, instead of each element of Θ as in probability theory) such that the numeric values sum up to 1. Usually, $m(A) \geq 0$ is used to denote the value assigned to subset A.

There are three distinct features about BPA.

1) $m(A)$ is the portion of the total belief committed exactly to A, which cannot be further subdivided among the subsets of A and does not include the portions of the total belief committed to subsets of A.
2) The singletons $\{A_i\}$, $\forall i \in \Lambda$ are only parts of the elements in 2^Θ, so it is possibly $\sum_{i=1}^{M} m(A_i) < 1$, and since A_i and $\neg A_i = \Theta - A_i$ are only two elements of 2^Θ, possibly $m(A_i) + m(\neg A_i) < 1$. This avoids the basic axioms of Bayesian formalism, or in other words, BPA supplies an incomplete probabilistic model.
3) A subset $A \in 2^\Theta$ with $m(A) > 0$ is called a focal element. When A is the only focal element in 2^Θ, we have $m(\Theta) = 1 - m(A)$, i.e., $m(\Theta)$ absorbs the unassigned portions of the total belief after commitment of belief to various proper subsets of Θ.

Since a subset A represents the disjunction of all the elements in A, the truth of $B \subset A$ implies the truth of A, i.e., all the portions committed exactly to every subset of A will also support A. Hence, bel(A) is given by

$$\text{bel}(A) = \sum_{B \subseteq A} m(B) \tag{36}$$

with the special cases.

1) When $A = A_i$ is a singleton, then bel(A) = bel(A_i) = $m(A_i)$.
2) When $A = \Theta$, bel$(\Theta) = 1$.

When two or more evidences exist, there will be two or more sets of BPA's and bel(.)'s given to the subsets of the same Θ. Under the condition that each of these evidences is independent of each other, Dempster's combination rule could be used to combine them into a new BPA and bel(.), which represents the combined impact of the two evidences.

Let bel_1, bel_2 and m_1, m_2 denote two belief functions and the corresponding BPA's respectively. The Dempster rule defines a new BPA $m = m_1 \oplus m_2$, which represents the combined effect of m_1 and m_2, i.e., for $A \neq \varnothing$ (where $\varnothing$ denotes the empty set):

$$\begin{aligned} m(A) = m_1 \oplus m_2(A) &= k \sum_{X \cap Y = A, A \neq \varnothing} m_1(X) m_2(Y) \\ k^{-1} = 1 - \sum_{X \cap Y = \varnothing} m_1(X) m_2(Y) &= \sum_{X \cap Y \neq \varnothing} m_1(X) m_2(Y) \end{aligned} \tag{37}$$

where $X \subseteq \Theta$, $Y \subseteq \Theta$ are any subsets. The m is a BPA if $k^{-1} \neq 0$; if $k^{-1} = 0$, then $m_1 \oplus m_2$ does not exist and m_1, m_2 are said to be totally contradictory, i.e., the two evidences are in conflict. If the two evidences are not in conflict, the combined belief function, denoted by $\text{bel}_1 \oplus \text{bel}_2$ may be computed from $m_1 \oplus m_2$ directly by (36).

Because Dempster rule is associative and commutative, it could be used to combine multiple evidences by sequentially (and with an arbitrary order) using (37) to obtain

$$m = m_1 \oplus m_2 \oplus \cdots \oplus m_K. \tag{38}$$

Obviously, the combination m exists when any of two among the K evidences are not in conflict.

Moreover, the values of m could also be calculated by the following formula:

$$\begin{aligned} m(A) &= k \sum_{X_1 \cap \cdots \cap X_K = A} \prod_{i=1}^{K} m_i(X_i) \\ k^{-1} &= 1 - \sum_{X_1 \cap \cdots \cap X_K = \varnothing} \prod_{i=1}^{K} m_i(X_i) \\ &= \sum_{X_1 \cap \cdots \cap X_K \neq \varnothing} \prod_{i=1}^{K} m_i(X_i). \end{aligned} \tag{39}$$

After obtaining m, we can calculate its correspondent bel by (36).

B. Modeling Multiclassifier Combination by D–S Theory

In our problem, the M exhaustive and mutually exclusive propositions are given by $A_1 = x \in C_i$, $\forall i \in \Lambda$, which respectively denote that input sample x comes from C_i, $\forall i \in \Lambda$, and the universal proposition is $\Theta = \{A_1, \cdots, A_M\}$. When applied to the same input x, K classifiers $e_1, \cdots, e_K$ will produce K evidences $e_k(x) = j_k$, $k = 1, 2, \cdots, K$ with each $e_k(x) = j_k$ denoting that x is assigned a label $j_k \in \Lambda \cup \{M+1\}$ by classifier e_k.

Given that $\epsilon_r^{(k)}$, $\epsilon_s^{(k)}$ are respectively the recognition rate and the substitution rate of e_k (usually $\epsilon_r^{(k)} + \epsilon_s^{(k)} < 1$ due to the rejection action). For each $e_k(x) = j_k$, when $j_k \in \Lambda$, one could have uncertain beliefs that the proposition $A_{j_k} = x \in C_{j_k}$ is true with a degree $\epsilon_r^{(k)}$ and is not true with a degree $\epsilon_s^{(k)}$; when $j_k = M + 1$ (i.e., x is rejected by e_k), one has no ideas about anyone of the M propositions $A_i = x \in C_i$, $\forall i \in \Lambda$, which could be regarded as the full support of the universal proposition Θ.

We can define a BPA function m_k on Θ for evidence $e(x) = j_k$ in the following way:

1) When $j_k = M + 1$, m_k has only a focal element Θ with $m_k(\Theta) = 1$. Since e_k says nothing about anyone of the M propositions, this is a degenerated case.
2) When $j_k \in \Lambda$, m_k has only two focal elements A_{j_k} and $\neg A_{j_k} = \Theta - \{A_{j_k}\}$ with $m_k(A_{j_k}) = \epsilon_r^{(k)}$ and $m_k(\neg A_{j_k}) = \epsilon_s^{(k)}$ since e_k only gives A_{j_k} and $\neg A_{j_k}$ the support of degrees $\epsilon_r^{(k)}$, $\epsilon_s^{(k)}$ respectively. Moreover, e_k says nothing about any other propositions, so we have $m_k(\Theta) = 1 - \epsilon_r^{(k)} - \epsilon_s^{(k)}$.

As a result, with the existence of all the evidences $e_k(x)$, $k = 1, \cdots, K$, we will have K *bap*'s m_k, $k = 1, \cdots, K$. Our problem is to use the Dempster rule to obtain a combined BPA $m = m_1 \oplus m_2 \oplus \cdots \oplus m_{K'}$ and to use this new *BPA* to calculate bel(A_i) and bel$(\neg A_i)$ for $\forall i \in \Lambda$ based on all the K

evidences. Hereafter, we could form the combined classifier by using decision rules derived from these beliefs.

Before discussing how to compute effectively the combined m in the next subsection, we make two preparations.

First, we discard those evidences $e_k(x) = j_k$ with $j_k = M + 1$ since it follows from (37) and (39) that a BPA as $m_k(\Theta) = 1$ has no influence on the result of the combination. After discarding such evidences, assume that there are $K' \leq K$ evidences $e_k(x)$, $k = 1, \cdots, K'$ with each $j_k \in \Lambda$. On the K' evidences, without losing generality, we could further rule out the following three special cases.

1) $K' = 0$. It means that all of K classifiers reject the input sample x. Obviously, the final decision is simple to reject x too.
2) There is one classifier e_k, which has the recognition rate $\epsilon_r^{(k)} = 1$. This means that e_k itself can classify any input sample with absolute correctness. Thus, all the other classifications are no longer necessary.
3) There is one classifier e_k, which has the substitution rate $\epsilon_s^{(k)} = 1$. This means that e_k always makes wrong decisions, i.e., the classifier is generally not much useful.[1] We will not use such classifiers here.

As a result, we can concentrate on the general case that there are K' evidences $e_k(x) = j_k$, $k = 1, \cdots, K'$ with $0 < \epsilon_r^{(k)} < 1$, $0 \leq \epsilon_s^{(k)} < 1$.

Second, we show that in the general case previously the combination m does exist, i.e., $m_1, \cdots, m_{K'}$ are not in conflict. To show this, it is sufficient to show $k^{-1} \neq 0$. It follows from (39) that we only need to show that there is a combination $X_1 \cap X_2 \cap \cdots \cap X_{K'} \neq \varnothing$ with $m_1(X_1)m_2(X_2)\cdots m_{K'}(X_{K'}) \neq 0$. For $k = 1, 2, \cdots, K'$, since $0 < m_k(A_{j_k}) = \epsilon_r^{(k)} < 1$, we know that for at least one of $m_k(\neg A_{j_k}) \neq 0$, $m_k(\Theta) \neq 0$ should be true. Let $X_1 = A_{j_1}$, and for $k = 2, \cdots, K'$, let $X_k = A_{j_k}$ if $A_{j_k} = A_{j_1}$, otherwise, let X_k be one of $\neg A_{j_k}$, Θ such that $m_k(X_k) \neq 0$. As a result, we have $X_1 \cap X_2 \cap \cdots \cap X_{K'} = A_{j_1} \neq 0$ with $m_1(X_1)m_2(X_2)\cdots m_{K'}(X_{K'}) \neq 0$. Therefore, we proved $k^{-1} \neq 0$.

C. An Effective Combination Scheme

For the computation of $m = m_1 \oplus m_2 \oplus \cdots \oplus m_{K'}$, the direct use of both (37) and (39) yields the computation cost that increases exponentially with M. Usually, it may be computationally prohibitive, especially when we have a large number of classifiers. However, for our problem, because of a special feature that the BPA of each evidence only has two focal elements; one is singleton and the other is negation of this singleton, we can arrive at a computing method with computation cost $O(M)$. The method consists of two main steps described in the following.

[1] Except the special case of two class problem, in this case the complement of the decision of such a classifier is always right, then we reach the same situation as (2) by exchanging the roles of $\epsilon_s^{(k)}$, $\epsilon_r^{(k)}$. However, for a problem involving more than two classes, the complement decision of such a classifier does not supply much useful information.

In the first step, we collect the evidences into groups with those impacting the same proposition in each group, and then combine the evidences in each group respectively.

For all the evidences $e_k(x) = j_k$, $k = 1, \cdots, K'$, suppose that among $j_1, \cdots, j_{K'}$ there are $K_1 \leq \min(M, K')$ different values $j'_1, \cdots j'_{K_1}$, thus all the K' evidences are collected into K_1 groups $E_1, E_2, \cdots, E_{K_1}$ with each $e(x) = j_k$ being put into group E_k if $e_k(x) = j_k = j'_k$. For each group E_k, since all its evidences $e_{k_1}(x) = j'_k, \cdots, e_{k_p}(x) = j'_k$ impact on the same propositions $A_{j'_k}$ and $\neg A_{j'_k}$, we can recursively use (37) to make a combined BPA m_{E_k} from BPA's $m_{k_1}, \cdots, m_{k_p}$, which are provided by $e_{k_1}, \cdots, e_{k_p}$, i.e.,

$$m_{E_k} = m_{k_1} \oplus \cdots \oplus m_{k_p} = [\cdots [[m_{k_1} \oplus m_{k_2}] \oplus m_{k_3}] \oplus \cdots] \oplus m_{k_p} \tag{40}$$

or

$$m_2 = m_{k_1} \oplus m_{k_2}, m_3 = m_2 \oplus m_{k_3}, \cdots, m_{E_k} = m_p = m_{p-1} \oplus m_p.$$

For example:

$$\begin{aligned}
k_2 &= \frac{1}{1 - m_{k_1}(A_{j'_k})m_{k_2}(\neg A_{j'_k}) - m_{k_1}(\neg A_{j'_k})m_{k_2}(A_{j'_k})} \\
&= \frac{1}{1 - \epsilon_r^{(k_1)}\epsilon_s^{(k_2)} - \epsilon_s^{(k_1)}\epsilon_r^{(k_2)}} \\
m_2(\Theta) &= k_2 m_{k_1}(\Theta)m_{k_2}(\Theta) \\
&= k_2(1 - \epsilon_r^{(k_1)} - \epsilon_s^{(k_1)})(1 - \epsilon_r^{(k_2)} - \epsilon_s^{(k_2)}) \\
m_2(\neg A_{j'_k}) &= k_2 m_{k_1}(\neg A_{j'_k})[m_{k_2}(\Theta) + m_{k_2}(\neg A_{j'_k})] \\
&\quad + k_2 m_{k_2}(\neg A_{j'_k})m_{k_1}(\Theta) \\
&= k_2 \epsilon_s^{(k_1)}(1 - \epsilon_r^{(k_2)}) \\
&\quad + k_2 \epsilon_s^{(k_2)}(1 - \epsilon_r^{(k_1)} - \epsilon_s^{(k_1)}) \\
m_2(A_{j'_k}) &= k_2 m_{k_1}(A_{j'_k})[m_{k_2}(\Theta) + m_{k_2}(A_{j'_k})] \\
&\quad + k_2 m_{k_2}(A_{j'_k})m_{k_1}(\Theta) \\
&= k_2 \epsilon_r^{(k_1)}(1 - \epsilon_s^{(k_2)}) + k_2 \epsilon_r^{(k_2)}(1 - \epsilon_s^{(k_1)} - \epsilon_r^{(k_1)}) \\
m_2(A) &= 0, \text{ for all other } A \subset \Theta.
\end{aligned} \tag{41}$$

It is important to notice that the new BPA m_2 has also only the same two focal elements as before. Consequently, it follows from (40) that $m_3, \cdots, m_{p-1}, m_{E_k}$ could be calculated recursively in the same way.

For, $r = 3, \cdots, p-1, p$, the general recursive formula is given by

$$\begin{aligned}
k_r &= \frac{1}{1 - m_{r-1}(A_{j'_k})m_{k_r}(\neg A_{j'_k}) - m_{r-1}(\neg A_{j'_k})m_{k_r}(A_{j'_k})} \\
m_r(\Theta) &= k_r m_{r-1}(\Theta)m_{k_r}(\Theta) \\
m_r(\neg A_{j'_k}) &= k_r m_{r-1}(\neg A_{j'_k})[m_{k_r}(\Theta) + m_{k_r}(\neg A_{j'_k})] \\
&\quad + k_r m_{k_r}(\neg A_{j'_k})m_{r-1}(\Theta) \\
m_r(A_{j'_k}) &= k_r m_{r-1}(A_{j'_k})[m_{k_r}(\Theta) + m_{k_r}(A_{j'_k})] \\
&\quad + k_r m_{k_r}(A_{j'_k})m_{r-1}(\Theta) \\
m_r(A) &= 0, \text{ for all other } A \subset \Theta.
\end{aligned} \tag{42}$$

As a result, we obtain the new combined BPA $m_{E_k} = m_p$, which again has only the same two focal elements as before.

Therefore, after the combination, E_k is equivalent to a new classifier that produces an event $E_k(x) = j'_k$ with a new recognition rate $\epsilon_r^{(k)} = m_{E_k}(A_{j'_k})$ and the substitution rate $\epsilon_s^{(k)} = m_{E_k}(\neg A_{j'_k})$.

In summary, the first step has converted the K' classifiers $e_1, \cdots, e_{K'}$ and their correspondent evidences $e_k(x) = j_k$, $k = 1, \cdots, K'$ into the $K_1 \leq \min(K', M)$ combined classifiers $E_1, \cdots, E_{K_1}$ with their correspondent evidences $E_k(x) = j'_k$, $k = 1, \cdots, K_1$ and $j'_1 \neq j'_2 \neq \cdots \neq j'_{K_1}$.

The second step is to further combine the BPA's m_{E_k}, $k = 1, \cdots, K_1$ into a final combined BPA

$$m = m_{E_1} \oplus m_{E_2} \oplus \cdots \oplus m_{E_{K_1}} \tag{43}$$

and then to calculate the correspondent $\text{bel}(A_i)$, $\text{bel}(\neg A_i)$ for $\forall i \in \Lambda$.

In this case, since any two m_{E_i}, m_{E_j}, $j \neq i$ will affect different focal elements, the combined $m_{E_i} \oplus m_{E_j}$ will affect five focal elements instead of two focal elements as each m_{E_i}, m_{E_j} did. Furthermore, the number of focal elements of $m_{E_1} \oplus m_{E_2}$, $m_{E_1} \oplus m_{E_2} \oplus m_{E_3}, \cdots, m_{E_1} \oplus m_{E_2} \oplus \cdots \oplus m_{E_{K_1}}$ will increase rapidly as more BPA's are combined. In such situation, the direct use of (37) is computationally quite expensive.

Fortunately, we have

$$\{j'_1, j'_2, \cdots, j'_{K_1}\} = \Lambda = \{1, 2, \cdots, M\}$$

when $K_1 = M$, i.e., $\{j'_1, j'_2, \cdots, j'_{K_1}\}$ is just a permutation of $\{1, 2, \cdots, M\}$. Thus, these E_k, $k = 1, \cdots, K_1$ are just the same as one of the intermediate products in paper [26], called simple evidence functions. Moreover, even when $K_1 < M$, we can append $M - K_1$ functionless evidences E_k, $k = K_1 + 1, \cdots, M$ with $m_{E_k}(\Theta) = 1$, $m_{E_k}(A_{j'_k}) = 0$, $m_{E_k}(\neg A_{j'_k}) = 0$ for $k = K_1 + 1, \cdots, M$ and $j'_1 \neq j'_2 \neq \cdots \neq j'_{K_1} \neq j'_{K_1+1} \neq \cdots \neq j'_M$. Again, $\{j'_1, j'_2, \cdots, j'_M\} = \Lambda$ is a permutation of $\{1, 2, \cdots, M\}$, and then E_k, $k = 1, \cdots, M$ become the so called simple evidence functions too. On the other hand, it is easy to see that for any m_i, we have $m_i = m_i \oplus m_{E_k}$, $k \geq K_1 + 1$. Thus

$$\begin{aligned} &m_{E_1} \oplus m_{E_2} \oplus \cdots \oplus m_{E_{K_1}} \oplus m_{E_{K_1+1}} \oplus \cdots \oplus m_{E_M} \\ &\quad = m_{E_1} \oplus m_{E_2} \oplus \cdots \oplus m_{E_{K_1}} = m. \end{aligned}$$

That is, the appended E_k, $k = K_1 + 1, \cdots, M$ have no influences on the final combined m. So, with these facts, we can directly borrow those formulae developed in [26] for combining the simple evidence functions to serve our purpose.

From the formulae (6)–(8) provided in paper [26], by noticing that in our case $m_{E_k}(\Theta) = 1$, $m_{E_k}(A_{j'_k}) = 0$, $m_{E_k}(\neg A_{j'_k}) = 0$ for $k = K_1 + 1, \cdots, M$, we obtain after some derivation the following formulae for computing the final combined beliefs for proposition $A_{j'_k}, \neg A_{j'_k}$, $k = 1, 2, \cdots, M$:

$$\begin{aligned} A &= \sum_{k=1}^{K_1} \frac{m_{E_k}(A_{j'_k})}{1 - m_{E_k}(A_{j'_k})} \\ B &= \prod_{k=1}^{K_1} [1 - m_{E_k}(A_{j'_k})] \\ C &= \prod_{k=1}^{K_1} m_{E_k}(\neg A_{j'_k}) \end{aligned} \tag{44}$$

$$k^{-1} = \begin{cases} (1 + A)B - C, & \text{if } K_1 = M \\ (1 + A)B, & \text{if } K_1 < M \end{cases} \tag{45}$$

and (46) and (47) (shown at the bottom of the page).

In the formulae, we have factors $(1/(1 - m_{E_k}(A_{j'_k})))$, $k = 1, \cdots, K_1$, which are not meaningless under the condition that $m_{E_k}(A_{j'_k}) < 1$, $k = 1, \cdots, K_1$. For each k, recall (40) that m_{E_k} is calculated by $m_{E_k} = m_{k_1} \oplus \cdots \oplus m_{k_p}$ and that in our problem $m_{k_i}(A_{j'_k}) = \epsilon_r^{(k_i)} < 1$, $m_{k_i}(\neg A_{j'_k}) = \epsilon_s^{(k_i)} < 1$, $i = 1, \cdots, p$, it follows from (39) that $m_{E_k}(A_{j'_k}) < 1$. So, we see that all the previous equations are always meaningful.

Now, let's examine the computational complexity of the whole procedure. First, the complexity for A, B, C is respectively at most of order $O(M)$, thus, the computation spent on all the equations of (44)–(47) is also of order $O(M)$. Second, in (40) and (41), all the computations are constant with respect to M. So the total computation of the whole procedure is of order $O(M)$.

D. Decision Rules

Recall that $\{j'_1, j'_2 \cdots, j'_M\}$ is just a permutation of $\{1, 2, \cdots, M\}$, we see that belief values $\text{bel}(A_i)$, $\text{bel}(\neg A_i)$, $i = 1, 2, \cdots, M$ are all given by (44)–(47). With these

$$\text{bel}(A_{j'_k}) = \begin{cases} k\left\{\frac{m_{E_k}(A_{j'_k})}{1 - m_{E_k}(A_{j'_k})} B + \frac{m_{E_k}(\Theta)}{m_{E_k}(\neg A_{j'_k})} C\right\}, & \text{if } K_1 = M \text{ or } K_1 = M - 1 \& k = M \\ k \frac{m_{E_k}(A_{j'_k})}{1 - m_{E_k}(A_{j'_k})} B, & \text{in all the other cases} \end{cases} \tag{46}$$

$$\text{bel}(\neg A_{j'_k}) = \begin{cases} k\left\{\left(A - \frac{m_{E_k}(A_{j'_k}) - m_{E_k}(\neg A_{j'_k})}{1 - m_{E_k}(A_{j'_k})}\right) B - C\right\}, & \text{if } K_1 = M \\ k\left(A - \frac{m_{E_k}(A_{j'_k}) - m_{E_k}(\neg A_{j'_k})}{1 - m_{E_k}(A_{j'_k})}\right) B, & \text{if } K_1 < M \& k \leq K_1 \\ k\left(A - \frac{m_{E_k}(A_{j'_k})}{1 - m_{E_k}(A_{j'_k})}\right) B, & \text{in all the other cases} \end{cases} \tag{47}$$

values, we can finally define the combined classifier E by the following rules:

$$E(x) = \begin{cases} j, & \text{if } \mathrm{bel}(A_j) = \max_{i\in\Lambda} \mathrm{bel}(A_i) \\ M+1, & \text{otherwise.} \end{cases} \quad (48)$$

In making the trade-off between the substitution rate and the rejection rate, (48) could be modified into (49):

$$E(x) = \begin{cases} j, & \text{if } \mathrm{bel}(A_j) = \max_{i\in\Lambda} \mathrm{bel}(A_i) \geq \alpha \\ M+1, & \text{otherwise} \end{cases} \quad (49)$$

where $0 < \alpha \leq 1$ is a threshold.

The previous rules didn't take into consideration the beliefs $\mathrm{bel}(\neg A_i)$'s, which also contain useful information for the final decision. The following rules are proposed in order to include this information.

1) The first rule

$$E(x) = \begin{cases} j, & \text{if } d_j = \max_{i\in\Lambda} d_i > \alpha \\ M+1, & \text{otherwise} \end{cases} \quad (50)$$

where $0 < \alpha < 1$ and $d_i = \mathrm{bel}(A_i) - \mathrm{bel}(\neg A_i)$, $i = 1, \cdots, M$ reflects the pure total support by the proposition A_i.

2) The rule tries to pursue the highest recognition rate under the constraint of a bounded substitution rate. Equation (51) is shown at the bottom of the page.

3) In (52), which is shown at the bottom of the page, where $0 < \alpha_1, \alpha_2 \leq 1$ are predefined thresholds.

VII. Applications of the Combination Approaches to the Recognition of Totally Unconstrained Handwritten Numerals

A. Individual Classifiers and Database

The four classifiers proposed in [1] (where they are called four experts) are used here to show the significant benefits obtained by using the combination approaches proposed in the earlier sections of this paper. As in [1], the four classifiers are named expert#1, expert#2, expert#3, and expert#4, and are denoted by e_1, e_2, e_3, and e_4. The first three are based on the features extracted from the skeletons, while e_4 is based on the features derived from contours. See [1], for more details.

The data used here come from the U.S. Zipcode database of the Concordia OCR research team. This database contains 17 140 run-length coded binarized digits. The samples were originally collected from the dead letter envelopes by the U.S. Postal Services at different locations. After some preprocessings (see [1] for details), 4000 samples (400×10 digits, i.e., each of the 10 numerals has 400 samples) were used for training the four experts, and then, a new set of 2000 samples (200×10 digits) was used for testing them. The following results are obtained from the testing set.

TABLE I
The Results of Four Experts

	Recogn.	Substi.	Reject.	Reliab.
e_1	86.05%	2.25%	11.70%	97.45%
e_2	93.10%	2.95%	3.95%	96.98%
e_3	92.95%	2.15%	4.90%	97.74%
e_4	93.90%	1.60%	4.50%	98.32%

In Table I, Recogn., Substi., Reject., and Reliab. are abbreviations of recognition, substitution, rejection and reliability rates respectively. The reliability rate is defined by

$$\text{Reliability} = \frac{\text{Recognition}}{100\% - \text{Rejection}}. \quad (53)$$

In addition, if e_1 assigns a subset of labels to an input, the input is regarded as being rejected in Table I. Moreover, in the following, this nonunique recognition of e_1 is also always regarded as the rejection except for some cases specifically indicated.

In the following three subsections, we will show the results of experiments by the combination approaches proposed in Sections IV–VI, then in the last subsection, we will make some comparisons with the three approaches. Among e_1, e_2, e_3, e_4, only e_3 could supply the output information at the measurement level, hence we could only consider the combination problem of Type 1. Thus, we have not conducted any experiments by using the averaged Bayes classifier proposed in Section III.

B. Experiments by the Approach Based on Dempster–Shafer Formalism

Extensive experiments have been carried out to test the performance of the approach proposed in Section VI. These experiments could be divided into two groups. In the first group, we use the recognition, substitution and rejection rates provided in Table I as prior knowledge, and use the approach for combining the results of the four experts on all the 2000 samples of the test set given in Table I. But in the second group, we divide the 2000 samples into two sets. The first 1000 samples are used to test e_i, $i = 1, 2, 3, 4$ in order to obtain the estimations of the recognition, substitution and rejection rates, i.e., these 1000 samples are used for estimating these rates. Then, the four classifiers are tested by the remaining 1000 samples and combined by the approach given in Section VI using the rates learned from the first 1000 samples. In each group, a number of experiments have been conducted using

$$E(x) = \begin{cases} j, & \text{if } \mathrm{bel}(A_j) = \max_{i\in\Lambda}\{\mathrm{bel}(A_i)/\forall i, \mathrm{bel}(\neg A_i) \leq \alpha\} \\ M+1, & \text{otherwise.} \end{cases} \quad (51)$$

$$E(x) = \begin{cases} j, & \text{if } \mathrm{bel}(A_j) = \max_{i\in\Lambda}\{\mathrm{bel}(A_i)/\forall i, \mathrm{bel}(A_i) \geq \alpha_1, \mathrm{bel}(\neg A_i) \leq \alpha_2\} \\ M+1, & \text{otherwise.} \end{cases} \quad (52)$$

TABLE II
RESULTS OF TRADE-OFF AS α INCREASES

α	Recogn.	Substi.	Reject.	Reliab.
0.00	98.95%	0.85%	0.20%	99.15%
0.90	97.60%	0.30%	2.10%	99.69%
0.95	95.85%	0.05%	4.10%	99.95%
0.99	93.95%	0.00%	6.05%	100.0%
e_4:	93.90%	1.60%	4.50%	98.32%

TABLE III
RESULTS OF TRADE-OFF AS α INCREASES ACCORDING TO RULE (52)

α	Recogn.	Substi.	Reject.	Reliab.
0.00	98.80%	0.80%	0.40%	99.20%
0.90	96.45%	0.25%	3.30%	99.74%
0.91	96.15%	0.10%	3.75%	99.90%
0.92	96.15%	0.10%	3.75%	99.90%
0.93	95.85%	0.05%	4.10%	99.95%
0.95	95.45%	0.05%	4.50%	99.95%
0.99	91.80%	0.00%	8.20%	100.00%
e_4:	93.90%	1.60%	4.50%	98.32%

the different decision rules given by (48)–(52) with different values of thresholds to provide different outcomes to allow trade-offs between the substitution and rejection rates.

Experiments of the First Group

1) Results from decision rule (48): The results of applying (48) to process the combined beliefs given by (44)–(47) are

$$\text{Recogn.}: 98.95\% \quad \text{Substi.}: 0.85\%$$
$$\text{Reject.}: 0.20\%, \quad \text{Relab.}: 99.15\%.$$

2) Results from decision rule (49): By using (49) to replace (48), we can determine the trade-offs between high recognition rate and high reliability rate. The following table shows some profiles about the trade-offs as α increases.

   In the Table II, the last row is the performance of e_4 alone. Refer to Table I, we see that e_4 gives the best performance among the four individual classifiers. Here and later, we list this performance again to show the improvements obtained by the combined classifier E. As shown in the table, the combined E is absolutely superior to e_4 (and thus to all other individual classifiers) in all aspects. The most interesting thing is that E could keep a high reliability and a high recognition simultaneously.
3) Results from decision rule (52). This rule produced the outcomes similar to those by (51), the difference is that it uses the threshold α to control the pure support $\text{bel}(A_i) - \text{bel}(\neg A_i)$ for adjusting the trade-offs between the recognition rate and the reliability rate. Table III is the counterpart of Table II.

Experiments with the Second Group

As indicated earlier in this section, the experiments of this group use the first 1000 samples as training set to obtain the initial recognition, substitution and rejection rates,

TABLE IV
RESULTS OF INDIVIDUAL CLASSIFIERS ON THE FIRST 1000 SAMPLES

α	Recogn.	Substi.	Reject.	Reliab.
e_1	87.0%	1.5%	11.5%	98.31%
e_2	94.4%	2.4%	3.2%	97.52%
e_3	95.0%	1.2%	3.8%	99.79%
e_4	94.8%	0.9%	4.3%	99.06%

TABLE V
RESULTS OF INDIVIDUAL CLASSIFIERS ON THE LAST 1000 SAMPLES

α	Recogn.	Substi.	Reject.	Reliab.
e_1	85.1%	3.0%	11.9%	96.59%
e_2	91.8%	3.5%	4.7%	96.33%
e_3	90.9%	3.1%	6.0%	96.70%
e_4	93.0%	2.3%	4.7%	97.59%

TABLE VI
OUTCOME OF E BY RULE (48) ON THE LAST 1000 SAMPLES

$i \mid o$	0	1	2	3	4	5	6	7	8	9	rej.
0:	98	0	0	1	0	0	0	0	1	0	0
1:	0	100	0	0	0	0	0	0	0	0	0
2:	0	0	99	0	0	0	0	0	1	0	0
3:	0	0	1	98	0	0	0	0	0	0	0
4:	0	0	0	0	97	0	0	0	0	1	1
5:	0	0	0	0	0	100	0	0	0	0	0
6:	1	0	0	0	0	0	99	0	0	0	0
7:	0	0	1	0	1	0	0	98	0	0	0
8:	0	0	0	0	0	0	0	0	100	0	0
9:	0	0	0	0	0	0	0	1	1	97	1

Recogn.: 98.6% Substi.: 1.2% Reject.: 0.2%, Relab.: 98.8%

and then use the last 1000 samples as the testing set to check the performance. Thus for convenience of comparing the combined classifier with each individual classifier, the following Tables IV and V are given to show the performances of individual classifiers on the first 1000 samples and the last 1000 samples respectively.

By comparing Tables IV and V, one could see that the performances of e_i, $i = 1, 2, 3, 4$ on the first 1000 samples are better than those on the last 1000 samples, i.e., we selected the difficult half of the original data set for testing our combination approach. In the following, we use the rates given in Table IV as the prior knowledge for the combination test on the last 1000 samples.

1) Results from decision rule (48): The results of applying (48) to process the combined beliefs given by (44)–(48) are shown in the following confusion matrix.
2) Results from decision rule (49): Table VII lists the results by using (49) on the last 1000 samples for different values of α. Table VIII also clearly shows the significant improvements of E over e_4.

   As an example, we also present the confusion matrix in Table VIII of the combined E with $\alpha = 0.96$ as follows to give some classification details.
3) Results from decision rule (49): Table IX corresponds to Table IV.

C. Experiments by the Approach Based on Bayesian Formalism

Several experiments were conducted to verify the approach

TABLE VII
RESULTS OF TRADE-OFFS AS α INCREASES, USING THE LAST 1000 SAMPLES

α	Recogn.	Substi.	Reject.	Reliab.
0.00	98.6%	1.2%	0.2%	98.80%
0.90	97.0%	0.4%	2.6%	99.59%
0.92	97.0%	0.4%	2.6%	99.59%
0.94	96.9%	0.4%	2.7%	99.59%
0.96	95.0%	0.0%	5.0%	100.0%
e_4:	93.0%	2.3%	4.7%	97.59%

TABLE VIII
OUTCOME OF E BY RULE (49) WITH $\alpha = 0.96$ USING THE LAST 1000 SAMPLES

$i \mid o$	0	1	2	3	4	5	6	7	8	9	rej.
0:	94	0	0	0	0	0	0	0	0	0	6
1:	0	99	0	0	0	0	0	0	0	0	1
2:	0	0	97	0	0	0	0	0	0	0	3
3:	0	0	0	92	0	0	0	0	0	0	8
4:	0	0	0	0	92	0	0	0	0	0	8
5:	0	0	0	0	0	97	0	0	0	0	3
6:	0	0	0	0	0	0	97	0	0	0	3
7:	0	0	0	0	0	0	0	93	0	0	7
8:	0	0	0	0	0	0	0	0	98	0	2
9:	0	0	0	0	0	0	0	0	0	91	9

Recogn.: 95.0% 11 Substi.: 0.0% Reject.: 5.0%, Relab.: 100.0%

proposed in Section V, They consist of three groups. In the first group, we used the confusion matrices of e_i, $i = 1, 2, 3, 4$ on all the 2000 testing samples as the prior knowledge, and then combined the results of the four classifiers on the same 2000 samples.

1) Experiments with the first group: The following table shows the results of the combinations obtained from (34) and (35) for a specified range of α values. Again, from the list, we see that the combined E is significantly better than any of the individual classifiers.
2) Experiments with the second group: In the second group, we use the confusion matrices of e_i, $i = 1, 2, 3, 4$ on the first 1000 samples as the prior knowledge, and then combine the results of the four classifiers on the last 1000 samples; i.e., the first 1000 samples constituted the learning set and the last 1000 samples the testing set.
Table XI is the counterpart of Table X.
In this table, the recognition rate of the combined E is lower than that of e_4, and it seems that the combined result is even worse than the individual classifier e_4. However, the recognition rate is not the only index for evaluating the performance of classification. In Table XI, the substitution rate is significantly reduced and the reliability is significantly increased. Thus, from this point of view, we see that the performance is still improved considerably.
Furthermore, we can also observe that as α increases, the combined recognition rate and substitution rate decrease while the reliability increases. Thus, there is a trade-off when we choose the value of α. In practice, we may have two ways to do this. One is to previously set an upper bound for the substitution rate, we adjust α such that the resulting substitution rate is below the bound and the recognition rate is as high as possible. The other way is to adjust α such that the substitution rate is reduced as much as possible until it can't significantly reduce any more.
The similar phenomena can also be observed from the results of Tables XII–XVI. The similar trade-offs also apply. In fact, this kind of trade-off applies also to the results given in the previous Section VII-B and Section VII-B.
3) Experiments with the third group: In the third group, each time we leave out 10 samples (one for each digit) as testing set and use the 1990 samples as the training set to learn from the confusion matrices of e_i, $i = 1, 2, 3, 4$, the same process is repeated 200 times until each of the 2000 samples has been taken as testing sample once.

Table XII is the counterpart of Table X. It shows the advantage of the combined E over individual classifiers.

TABLE IX
RESULTS OF TRADE-OFFS AS α INCREASES WITH RULE (49) ON THE LAST 1000 SAMPLES

α	Recogn.	Substi.	Reject.	Reliab.
0.00	98.6%	1.1%	0.3%	98.80%
0.90	96.1%	0.3%	3.6%	99.69%
0.91	96.1%	0.3%	3.6%	99.69%
0.92	95.7%	0.3%	4.0%	99.69%
0.93	95.5%	0.3%	4.2%	99.69%
0.94	95.0%	0.0%	5.0%	100.0%
e_4:	93.0%	2.3%	4.7%	97.59%

TABLE X
RESULTS OF TRADE-OFF AS α INCREASES

α	Recogn.	Substi.	Reject.	Reliab.
0.00000	99.20%	0.80%	0.00%	99.20%
0.90000	98.85%	0.50%	0.65%	99.50%
0.99000	98.35%	0.20%	1.45%	99.80%
0.99900	97.75%	0.15%	2.10%	99.85%
0.99990	96.35%	0.05	3.6%	99.95%
0.99999	94.05%	0.00%	5.95%	100.00%
e_4:	93.90%	1.60%	4.50%	98.32%

TABLE XI
RESULTS OF TRADE-OFFS AS α INCREASES, ON THE LAST 1000 SAMPLES

α	Recogn.	Substi.	Reject.	Reliab.
0.00000	92.3%	0.9%	6.8%	99.03%
0.90000	92.2%	0.7%	7.1%	99.25%
0.99000	91.8%	0.6%	7.6%	99.35%
0.99900	91.7%	0.4%	7.9%	99.57%
0.99990	91.5%	0.3%	8.2%	99.67%
0.99999	91.0%	0.3%	8.7%	99.67%
e_4:	93.0%	2.3%	4.7%	97.59%

D. Experiments by the Approach Based on Voting Principle

The experiments in this subsection consist of two groups. One includes the results obtained from a testing set of all the 2000 samples. The other includes the results on a testing set of the last 1000 samples. In fact, the combination approach based on voting principle does not need the learning procedure for

TABLE XII
RESULTS OF TRADE-OFFS AS α INCREASES

α	Recogn.	Substi.	Reject.	Reliab.
0.00000	96.55%	1.45%	2.00%	98.52%
0.90000	96.25%	1.10%	2.65%	98.87%
0.99000	95.80%	0.90%	3.30%	99.07%
0.99900	0.99900	0.70%	3.95%	99.27%
0.99990	94.15%	0.60%	5.25%	99.37%
0.99999	92.30%	0.60%	7.10%	99.35%
e_4:	93.90%	1.60%	4.50%	98.32%

TABLE XIII
THE RESULTS OF VOTING RULE (18) FOR DIFFERENT THRESHOLDS

	Recog.	Subst.	Reject.	Reliab.
$0.0 \leq \alpha \leq 0.25$	98.90%	0.90%	0.20%	99.10%
$0.25 < \alpha \leq 0.50$	97.95%	0.35%	1.70%	99.64%
$0.50 < \alpha \leq 0.75$	92.90%	0.00%	7.10%	100.00%
e_4:	93.90%	1.60%	4.50%	98.32%

TABLE XIV
RESULTS OF VOTING RULE (19) FOR DIFFERENT THRESHOLDS

	Recog.	Subst.	Reject.	Reliab.
$\alpha = 0.0$	98.90%	0.90%	0.20%	99.10%
$0.0 \leq \alpha \leq 0.25$	98.60%	0.50%	0.90%	99.50%
$0.25 < \alpha \leq 0.50$	95.45%	0.05%	4.50%	99.95%
$0.50 < \alpha \leq 0.75$	88.60%	0.00%	11.4%	100.0%
e_4:	93.90%	1.60%	4.50%	98.32%

obtaining prior knowledge. The second group is provided here only to facilitate a comparison with previous investigations described in Section VII-D.

In both groups, the two general voting rules represented by (18) and (19) are used for a number of threshold values.

Experiments with the First Group (on the 2000 Samples)

1) Results obtained from voting rule (18): Table XIII gives the results produced by (18) with different values of α. It follows from this table that the combined E by voting rule could also improve individual classifiers significantly.
   Note that $0.0 \leq \alpha < 0.25$ means that any value between [0.0, 0.25] produces the same result because the votes are digits among $\{0, 1, 2, 3, 4\}$.
2) Results obtained from voting rule (20).

Experiments with the Second Group (on the Last 1000 Samples)

1) Results obtained from voting rule (19): In the following, Tables XV and XVI correspond to Tables XIII and XIV respectively, the difference is that all the results here are obtained from the last 1000 samples.
2) Results obtained from voting rule (20): As an example, the confusion matrix of the combined E with $0.25 < \alpha \leq 0.50$ is given in Table XVI.

Remarks: In all the experiments described Section VII-D, the nonunique assignment $e_1(x) = J$ is treated as a rejection if J has more than one element. It is worthwhile to point out

TABLE XV
RESULTS OF VOTING RULE (18), ON THE LAST 1000 SAMPLES

	Recog.	Subst.	Reject.	Reliab.
$0.0 \leq \alpha \leq 0.25$	98.6%	1.2%	0.2%	98.80%
$0.25 < \alpha \leq 0.50$	97.6%	0.5%	1.9%	99.49%
$0.50 < \alpha \leq 0.75$	91.7%	0.0%	8.3%	100.00%
e_4:	93.0%	2.3%	4.7%	97.59%

TABLE XVI
RESULTS OF VOTING RULE (20), USING THE LAST 1000 SAMPLES

	Recog.	Subst.	Reject.	Reliab.
$\alpha = 0.0$	98.6%	1.2%	0.2%	98.80%
$0.0 \leq \alpha \leq 0.25$	98.3%	0.7%	1.0%	99.29%
$0.25 < \alpha \leq 0.50$	94.3%	0.0%	5.7%	100.00%
$0.50 < \alpha \leq 0.75$	86.1%	0.0%	13.9%	100.00%
e_4:	93.0%	2.3%	4.7%	97.59%

TABLE XVII
OUTCOME OF E BY RULE (20) WITH $0.25 < \alpha \leq 0.50$, ON THE LAST 1000 SAMPLES

$i \mid o$	0	1	2	3	4	5	6	7	8	9	rej.
0:	94	0	0	0	0	0	0	0	0	0	6
1:	0	99	0	0	0	0	0	0	0	0	1
2:	0	0	95	0	0	0	0	0	0	0	5
3:	0	0	0	91	0	0	0	0	0	0	9
4:	0	0	0	0	92	0	0	0	0	0	8
5:	0	0	0	0	0	95	0	0	0	0	5
6:	0	0	0	0	0	0	97	0	0	0	3
7:	0	0	0	0	0	0	0	93	0	0	7
8:	0	0	0	0	0	0	0	0	98	0	2
9:	0	0	0	0	0	0	0	0	0	89	11

Recogn.: 9.43% 11 Substi.: 0.0% Reject.: 5.7%, Relab.: 100.0%

that if (22) is used (i.e., as treated in [1]), then the results could improve somewhat due to the voting contribution from $e_1(x) = J$.

For example, on the 2000 samples, we could have

	Recogn.	Substi.	Reject.	Reliab.
$\alpha = 0.51$	93.05%	0.0%	6.95%	100.0%

which is better than its counterpart in Table XIII, i.e.,

	Recogn.	Substi.	Reject.	Reliab.
$0.50 < \alpha \leq 0.75$	92.9%	0.0%	7.1%	100.0%.

In other words, $e_1(x) = J$ does give some useful information for combination. In Section VIII, we will discuss how to generalize our combination approaches so that the case of $e_1(x) = J$ could be included.

E. Comparison of the Three Approaches: A Case Study

The three previous subsections have shown that each of the three combination approaches improves the performance of the individual classifiers significantly. In this subsection, we further compare the three approaches with each other and investigate the characteristics of each approach.

Through reorganizing some results of the previous subsection, we may get four ordered lists Tables XVIII–XXIV (will be given in the sequel) for comparing the performance of the three approaches.

TABLE XVIII
THE MAXIMUM RECOGNITION RATES: FROM THE 2000 SAMPLE SET

approach:	B	DS	V_d	V	DS_d	B_l
$\text{recogn}_{\max}$	99.2%	>98.95%	>98.9%	=98.9%	>98.8%	>96.55%

TABLE XIX
THE MAXIMUM RECOGNITION RATES: FROM THE LAST 1000 SAMPLES

approach	DS	DS_d	V_d	V	P
$\text{recogn}_{\max}$	98.6% =	98.6% =	98.6% =	98.6% =	92.3%

TABLE XX
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES = 100%: FROM THE 2000 SAMPLE SET

approach	B	DS	V	DS_d	V_d
$\text{recogn}_{\max}$	94.05% >	93.95% >	92.9% >	91.9% >	88.6

TABLE XXI
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES = 100%: FROM THE LAST 1000 SAMPLES

approach	DS	DS_d	V_d	V
$\text{recogn}_{\max}$	95.0% =	95.0% >	94.3% >	91.7%

TABLE XXII
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES $\geq$ 99.9%: FROM THE 2000 SAMPLE SET

approach	B	DS_d	DS	V_d	V
$\text{recogn}_{\max}$	96.35% >	96.15% >	95.85% >	95.45% >	92.9%

TABLE XXIII
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES $\geq$ 99.9%: FROM THE LAST 1000 SAMPLES

approach	DS	DS_d	V_d	V
$\text{recogn}_{\max}$	95.0% =	95.0% >	94.3% >	91.7%

TABLE XXIV
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES $\geq$ 99.0%: FROM THE 2000 SAMPLE SET

approach	B	DS	V_d	V	DS_d	B_l
$\text{recogn}_{\max}$	99.2% >	98.95% >	98.9% =	98.9% >	98.8% >	95.8%

In all the tables, $\text{recogn}_{\max}$ represents the maximum recognition rate achieved by an approach, i.e., the recognition rate of the approach at threshold value $\alpha = 0$. In the last three tables, the best recognition rate of an approach is the highest one among a given subset of all the recognition rates obtained from the experiments of the previous three subsections, with the correspondent reliability rate of each one in the subset being not lower than the given value (e.g., 99% in Table XXIV and XXV). For example, in Table II, under the constraint that the reliability is not lower than 0.999, we have the best recognition rate 95.85%. Thus in Table XXI, under item DS the rate is 95.85%.

The notation B represents the approach based on Bayesian formalism in Section V with decision rule given by (34) and (35). DS denotes the approach based on Dempster–Shafer formalism in Section VI with decision rule (48) and (49). DS_d is a version of DS with the decision rule replaced by (50). V denotes the approach based on voting principle in Section IV with decision rule (21). V_d is a version of V with decision rule replaced by (19).

In part (a) of all the tables, the training method of B, DS, DS_d is the same as that in the first group of Sections VII-B and VII-C, i.e., the 2000 samples are not only used to learn from the confusion matrices or the rates of individual classifiers, but also to test the combination approaches. Whereas in part (b) of the tables, the training method is the same as that in the second group of in Sections VII-B and VII-C, i.e., the first 1000 samples are used for learning and the last 1000 samples for testing. In addition, in Tables XVIII and XXIV, we use B_l to represent a procedure of B with the training given in the third group of Section VII, i.e., in each case, 1990 samples are used for learning, and the other 10 samples (one per digit) for testing; the process is repeated 200 times until each of the 2000 samples has been tested once.

From Tables XXI–XXV, we can summarize the following observations on the characteristics of each approach.

1) If the confusion matrices of individual classifiers are well learned, then the performance of the approach based on Bayesian formalism is the best, e.g., in part (a) of all the four tables, B is ranked at the top. However, the approach is unstable. The rough learning will degenerate the performance of the approach rapidly, e.g., in part (b) of all the four tables, B is either ranked the lowest or simply not ranked because it could not reach the required reliability.
2) The approach based on Dempster–Shafer formalism is quite robust. Inaccurate learning does not influence the performance substantially. For example, from Tables XVIII and XIX, the recognition rates of DS, DS_d only drop a little, while from Tables XX to XXI, their recognition rates even go up. In addition, we could also see that the performances of DS, DS_d are not much different, which means that the decision rules given by (48)–(50) work equally well.
3) Both the approach based on D–S formalism and the approach based on voting principle behave well. On the average, the approach based D–S formalism is better than the approach based on voting principle, especially when high reliability is required (see Tables XX–XXIII).
4) For the approach based on the voting principle, the version V_d (i.e., using decision rule (19)) performs better than the version V (i.e., using decision rule (18)), especially when high reliability is desired (see Tables XX–XXIII). In addition, it also follows from Table XIII that (18) is more flexible than its special case (16). For example, (16), which was originally used in [1], could only give the recognition rate of 92.90% with 0% substitution. However, (18) could also give other two choices as shown in Table XIII.

Before closing this section, we must emphasize that the previous observations are just obtained from a case study on the 2000 sample data set presently available. They should be tested with more data sets. Strictly speaking, the experimentally obtained Recogn., Substi., Reject. and Reliab. rates

TABLE XXV
THE BEST RECOGNITION RATES UNDER THE CONSTRAINT THAT THEIR CORRESPONDENT RELIABILITY RATES $\geq$ 99.0%: FROM THE LAST 1000 SAMPLES

approach	V_d	V	DS	DS_d	B
$\text{recogn}_{\max}$:	98.3% >	97.60% >	95.0% =	95.0% >	92.3%

should all be regarded as random variables, and thus the comparisons of their values should be analyzed statistically; e.g., in Table XVIII, 99.2% > 98.95% should be tested under a given significance level. This more rigorous analysis is certainly a direction that deserves further pursuit in future. Here, we unfortunately do not have enough information to do so. First, in our experiments, for each algorithm and under a fixed α, we have only one sample for random variables Recogn., Substi., Reject. and Reliab., e.g., as given in Table XVIII, for $\alpha = 0$, we only have one sample value 99.2% and 98.95 for random variables Recogn.(B) and Recogn.(DS) respectively. To check statistically either always Recogn.(B) > Recogn.(DS) or on the average E[Recogn.(B)] > E[Recogn.(DS)], we at least need enough number samples of Recogn.(B), Recogn.(DS) for forming some test statistics to conduct statistical analysis. Second, even when there are some ways to increase the number of such samples,[2] we also need to known the population distributions of Recogn.(B), Recogn.(DS). We need some further studies as well as evidences before we can appropriately make assumptions on the distributions of Recog., Substi., Reject. and Reliab. for every algorithm. Third, even provided that we are given the assumption that all the distributions are Gaussian, and provided that we modeled the problem as, say, a simple Hypothesis testing: H_0 : E[Recogn.(B)] = E[Recogn.(DS)] and H_1 : E[Recogn.(B)] $\neq$ E[Recogn.(DS)], we still need to known whether it is appropriate to make the test under the condition with known Var[Recogn.(B)] = Var[Recogn.(DS)] or unknown Var[Recogn.(B)] $\neq$ Var[Recogn.(DS)]. So, we see that the aforementioned interesting issues deserve investigation with more data sets and experiments in the future.

VIII. CONCLUSION

The combination of several independent classifiers is a general problem that occurs in various application areas of pattern recognition. According to the levels of output information by various classifiers, the problems of combining multiclassifiers can be divided into three types. Type 3 covers the individual classifiers that can output measurement values based on which the decision is made. The averaged Bayes classifier and its version are proposed in Section III to solve the problem of this type. This approach is suitable for combining individual classifiers such as Bayesian classifier, k-NN classifier and various distance classifiers. Type 2 covers the individual classifiers that can output an ordered list of possible decisions (i.e., a number of labels), the method proposed in [4] aims at tackling the problems of this type. Type 1 covers the individual classifiers that output one label (one decision), i.e., any classifiers including those discussed in Type 2 and Type 3. Three approaches have been proposed in this paper to solve the problems of Type 1. One is based on the voting principle that is commonly used in social life. The other two are developed in accordance with Bayesian formalism and Dempster–Shafer formalism—two well known formalisms used in evidence gathering and uncertainty reasoning.

A simple approach based on the voting principle is the majority voting approach that was originally proposed in [1], [2], [5]. In our paper, two new versions of the voting principle are also presented. They are proved better than the simple majority voting approach used in the experiments presented in Section VII. Moreover, a general formula is given; it summarizes all those versions that are based on the voting principle. Although simple and useful, these approaches can not consider the classification error of each event $e_k(x) = j_k$. To fill this gap, other two approaches are developed. They regard each event $e_k(x) = j_k$ as an evidence with uncertain supports to the possible decisions (i.e., labels). One approach uses the confusion matrix of each individual classifier as the prior knowledge to manage this uncertainty. It gathers the evidences $e_k(x) = j_k$, $k = 1, \cdots, K$ derived from Bayes formula. The other approach uses the recognition rate and substitution rate of each individual classifier as prior knowledge to manage the uncertainty. Dempster's combination formula is used for gathering the evidences.

The experimental results on the recognition of totally unconstrained handwritten numerals have shown that the performances of individual classifiers could be improved significantly by the combination approaches proposed in this paper except the one in Section III, which could not be tested because it was not suitable for our application problem. The experiments have also shown the features of each approach and the details are given in Section VII-E. Based on our case study in Section VII-E, roughly speaking, the first recommendation would be the approach based on D–S formalism since it can obtain high recognition and reliability rates simultaneously and robustly. However, this recommendation does not mean that we should abandon the other approaches, which actually work also pretty well. Here, we should emphasize that the observations here are obtained from our experiments on recognition of unconstrained handwritten numerals. They may be true or may not be totally true when generalized to the problems or even the same problem when other data bases are used. However, the combining approaches studied in this paper are general and not constrained to any specific application problem, we believe that they will improve the performances of individual classifiers in general.

We argue that the research addressed to the problem of combining multiclassifiers may provide new insight to the literature of statistical pattern recognition. Previously, the main efforts focus on the design of one good classifier and the reduction of a high-dimensional feature vector so that a desired classification rate can be attained. Now, we can also change our focus. Instead of designing one high performance classifier (the job is usually extremely difficult), we can build a number of classifiers that use the low dimension feature vectors of different and complementary types. Each classifier

[2]For example, try to find other data sets or just divide the present 2000 character samples into many subsets (say, 50 in one subset). In the latter case, from each subset, one can obtain one sample for Recogn.(B), Recogn.(DS), and thus in total we can have 40 samples for Recogn.(B), Recogn.(DS).

itself may not have a superb performance. However the appropriate combination of these individual classifiers may produce a performance of high quality. There are also many new problems to be studied, we list some of them here:

1) All the approaches described in this paper are based on the assumption that individual classifiers are independent of each other. How to generalize these approaches or develop a new approach to combine dependent classifiers?
2) How many classifiers are appropriate for a special problem with a given number of feature variables? And how to distribute these variables to each classifier?
3) In this paper, the recognition rate and the substitution rate of each individual classifier were fixed after their training phase. It is well-known that these rates can be changed by applying different rejection thresholds to the individual classifiers. How to adjust these thresholds in the combination phase such that the best combination can be achieved?
4) Is it possible to develop a method to analyze the recognition rate of the combined classifier theoretically instead of experimentally?

Even concentrated only on the approaches proposed in this paper, there are also open problems to be tackled. For example, on testing the approach proposed in Section III and using (31) directly to solve the combination problem of Type 3. In addition, at the end of Section VII-D, we remarked that the use of the nonunique assignment $e_1(x) = J = \{j_1, j_2, \cdots, j_p\} \subset \Lambda$ by (22) can further improve the performance. The approaches proposed in Sections V and VI can be extended to cover this case. Here, we briefly introduce the key points of such extensions:

1) For the approach given in Section V, the extension needs two modifications. First, use (22) to let the event $e_k(x) = J$ distribute its score to learn from the confusion matrix of classifier e_k. Second, when e_k classifies an unknown sample x with $e_k(x) = \{j_1, j_2, \cdots, j_p\}$, then the following formula is used to replace (27):

$$P(x \in C_i / e_k(x) = \{j_1, j_2, \cdots, j_p\}) = \frac{\sum_{r=1}^{p} n_{ij_r}^{(k)}}{\sum_{r=1}^{p} \sum_{i=1}^{M} n_{ij_r}^{(k)}}, \quad i = 1, \cdots, M.$$

2) For the approach given in Section VI, the extension can be made by directly using (40) or (39) for solving the combined BPA's and bel's. However, the computational complexity will increase exponentially with M. If there is only one classifier that has an output of nonunique assignments, as in the case in [1] and Section VII of this paper, we can borrow directly the implementation scheme of Dempster rule for hierarchical evidence recently proposed by Schafer *et al.* [28]. Furthermore, the scheme can also be used directly for our purpose to some special cases where several classifiers have the output of nonunique assignments. In these cases, for any two of such classifiers, say $e_k(x) = J$, $e_l(x) = I$, we have the truth on either $I \cap J = \varnothing$ or on one of $I \subseteq J$, $J \subseteq I$. It is not difficult to see that these cases correspond to some hierarchical structures.

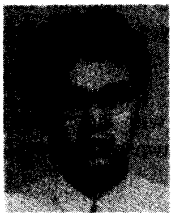
ACKNOWLEDGMENT

The authors wish to thank Christine Nadal for providing the classification results of the four individual classifiers e_1, e_2, e_3, e_4, which form the basis of the experiments performed in Section VII. The authors would also like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] C. Y. Suen, C. Nadal, T. A. Mai, R. Legault, and L. Lam, "Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts," *Frontiers in Handwriting Recognition,* C. Y. Suen, Ed., in *Proc. Int. Workshop on Frontiers in Handwriting Recognition,* Montreal, Canada, Apr. 2–3, 1990, pp. 131–143.
[2] C. Nadal, R. Legault, and C. Y. Suen, "Complementary algorithms for the recognition of totally unconstrained handwritten numerals," in *Proc. 10th Int. Conf. Pattern Recog.,* vol. A, June 1990, pp. 434–449.
[3] J. J. Hull, A Commike, and T. K. Ho, "Multiple algorithms for handwritten character recognition," in *Frontiers in Handwriting Recognition,* C. Y. Suen, Ed., *Proc. Int. Workshop Frontiers in Handwriting Recognition,* Montreal, PQ, Canada, Apr. 2–3, 1990, pp. 117–129.
[4] T. K. Ho, J. J. Hull, and S. N. Srihari, "Combination of structural classifiers," in *Proc. 1990 IAPR Workshop Syntactic and Structural Pattern Recog.*, June 1990, pp. 123–137.
[5] J. J. Hull, S. N. Srihari, E. Cohen, C. L. Kuan, P. Cullen, and P. Palumbo, "A blackboard-based approach to handwritten zip code recognition," in *Proc. US Postal Service Adv. Tech. Conf.,* 1988, pp. 1018–1032.
[6] E. Mandler and J. Schuermann, "Combining the classification results of independent classifiers based on the Dempster/Shafer theory of evidences," in *Pattern Recognition and Artificial Intelligence,* Gelsema and Kanal, Eds. Amsterdam: Elsevier Science, North-Holland, 1988, pp. 381–393.
[7] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE,* vol. 67, pp. 786–804, 1979.
[8] S. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE,* vol. 73, no. 11, pp. 1625–1650, 1985.
[9] F. Telinek, "Continuous speech recognition by statistical methods," *Proc. IEEE,* vol. 64, no. 4, pp. 532–556, 1976.
[10] L. Rabiner and B. Juang, "Introduction to hidden Markov models," *IEEE ASSP Mag.,* vol. 3, no. 1, pp. 4–16, Jan. 1986.
[11] J. D. Gibson, "Adaptive prediction for speech encoding," *IEEE ASSP Mag.,* vol. 1, p. 12, 1984.
[12] D. O'shaughnessy, "Speaker recognition," *IEEE ASSP Mag.,* vol. 3, no. 4, pp. 4–17, 1986.
[13] E. Skordalakis, "Syntactic ECG processing: A review," *Pattern Recog.,* vol. 9, no. 4, pp. 305–313, 1985.
[14] T. Pavlidis, "Waveform segmentation through functional approximation," *IEEE Trans. Comput.,* vol. C-20, pp. 59–67, 1971.
[15] C. H. Chen, Ed., *Artificial Intelligence and Signal Processing in Underwater Acoustic and Geo-Physics Problems,* Special Issue, *Pattern Recognition,* no. 6, 1985.
[16] B. Duerr, W. Haettich, H. Tropf, and G. Winkler, "A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals," *Pattern Recog.,* vol. 12, no. 3, pp. 189–199, 1980.
[17] P. Ahmed and C. Y. Suen, "Computer recognition of totally unconstrained handwritten ZIP codes," *Int. J. Pattern Recog. Artificial Intell.,* vol. 1, no. 1, pp. 1–15, 1987.
[18] C. H. Chen, Ed., "Special section on digital signal and waveform analysis," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 4, no. 2, pp. 97–141, 1982.
[19] P. A. Devijver and J. Kittler, *Pattern Recognition: A statistical Approach.* London: Prentice Hall, 1982.
[20] L. Lam and C. Y. Suen, "Structural classification and relaxation matching of totally unconstrained handwritten zip-code numbers," *Pattern Recog.,* vol. 21, no. 1, pp. 19–31, 1988.
[21] C. L. Kuan and S. N. Srihari, "A stroke-based approach to handwritten numeral recognition," in *Proc. US Postal Service Adv. Techn. Conf.,* 1988, pp. 1033–1041.
[22] P. Arentiero, R. Chin and P. Beaudet, "An automated approach to the design of tree classifiers," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 4, no. 1, pp. 51–57, 1982.

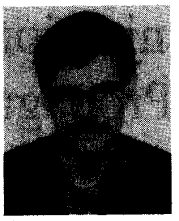
[23] Q. R. Wang and C. Y. Suen, "Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no. 4, pp. 406–417, 1984.
[24] J. Pearl, *Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann Publishers, 1988.
[25] J. Gordon and E. H. Shortliffe, "A method for managing evidential reasoning in hierarchical hypothesis space," *Artificial Intell.*, vol. 23, pp. 323–357, 1985.
[26] J. A. Barnett, "Computational Methods for a mathematical theory of evidence," in *Proc. 7th Int. Joint Conf. Artificial Intell.*, Vancouver, BC, 1985, pp. 868–875.
[27] G. Shafer, *A Mathematical Theory of Evidence.* Princeton, NJ: Princeton University Press, 1976.
[28] G. Shafer and R. Logan, "Implementing Dempster's rule for hierarchical evidence," *Artificial Intell.*, vol. 33, pp. 271–298, 1987.
[29] D. H. Ballard and C. M. Brown, *Computer Vision.* Englewood Cliffs, NJ: Prentice-Hall, 1982.



Lei Xu received the Masters and Ph.D. degrees in pattern recognition and signal processing from, in 1984 and 1987 respectively, Tsinghua University.

From June 1987 to June 1988, he was a postdoctoral fellow at Department of Mathematics, Peking University, where he has been an Associate Professor since September 1988. He was also visited the Department of Information Technology and the Lappeenranta University of Technology, Finland. At the Department of Computer Science, Concordia University, Montreal, PQ, Canada, he served as a Senior Researcher and Visiting Scholar, each for one year respectively. He is an author of more than 60 papers on signal processing, pattern recognition, artificial intelligence, computer vision, and neural networks. His present research interests are mainly focused on neural networks and computer vision. He has also experiences of serving as reviewers for *Neural Networks*, IEEE TRANSACTIONS ON NEURAL NETWORKS, *Artificial Intelligence*, and several Chinese high-level scientific journals.

Dr. Xu was one of 40 winners of the First Fok Ying Tung Education Foundation Prize for young teachers of the universities of the Peoples Republic China, 1988. He was also the second of the 10 winners of the Second Beijing Young Scientists Prize awarded by Beijing Association for Science and Technology in 1988.



Adam Krzyżak (M'86) was born in Szczecin, Poland, on May 1, 1953. He received the M.Sc. and Ph.D. degrees in computer engineering from the Technical University of Wrocław, Poland, in 1977 and 1980, respectively.

In 1980 he became an Assistant Professor in the Institute of Engineering Cybernetics, Technical University of Wrocław, Poland. From November 1982 to July 1983, he was a Postdoctorate Fellow in the School of Computer Science. McGill University, Montreal, PQ, Canada. Since August 1983, he has been with the Department of Computer Science, Concordia University, Montreal, where he is currently an Associate Professor. He has publications in the areas of pattern recognition, image processing, identification and estimation of control systems as well as in various applications of probability theory and statistics. He edited the book, *Computer Vision and Pattern Recognition* (Singapore: World Scientific, 1989).

Dr. Krzyżak is a recipient of the International Scientific Exchange Award.

Ching Y. Suen (M'66–SM'78–F'86) received the M.Sc. (Eng.) degree from the University of Hong Kong and the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada.

In 1972, he joined the Department of Computer Science of Concordia University, Montreal, PQ, Canada, where he became a Professor in 1979 and served as Chairman from 1980 to 1984. Presently he is the Director of CENPARMI, the new Centre for Pattern Recognition and Machine Intelligence of Concordia. During the past 15 years, he was also appointed to visiting positions in several institutions in different countries. He is the author/editor of 10 books on subjects ranging from *Computer Vision and Shape Recognition, Frontiers in Handwriting Recognition,* to *Computational Analysis of Mandarin and Chinese.* His latest book is entitled *Operational Expert System Applications in Canada,* (New York, Pergamon Press). He is the author of 250 papers and his current interests include pattern recognition and machine intelligence, expert systems, optical character recognition and document processing, and computational linguistics.

Dr. Suen is an active member of several professional societies and is a Fellow of the IEEE. He is an Associate Editor of several journals related to his areas of interest. He is the Past President of the Canadian Image Processing and Pattern Recognition Society, Governor of the International Association for Pattern Recognition, and President of the Chinese Language Computer Society. He is also the Editor-in-Chief of *Computer Processing of Chinese & Oriental Languages,* an international journal of the Chinese Language Computer Society. He is the winner of the 1992 ITAC/NSERC award for his outstanding contributions to pattern recognition, expert systsms, and computational linguistics.