Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance

Louisa Lam and Ching Y. Suen, *Fellow IEEE*

***Abstract*—Recently, it has been demonstrated that combining the decisions of several classifiers can lead to better recognition results. The combination can be implemented using a variety of strategies, among which majority vote is by far the simplest, and yet it has been found to be just as effective as more complicated schemes in improving the recognition results. However, all the results reported thus far on combinations of classifiers have been experimental in nature. The intention of this research is to examine the mode of operation of the majority vote method in order to gain a deeper understanding of how and why it works, so that a more solid basis can be provided for its future applications to different data and/or domains. In the course of our research, we have analyzed this method from its foundations and obtained many new and original results regarding its behavior. Particular attention has been directed toward the changes in the correct and error rates when classifiers are added, and conditions are derived under which their addition/elimination would be valid for the specific objectives of the application. At the same time, our theoretical findings are compared against experimental results, and these results do reflect the trends predicted by the theoretical considerations.**

***Index Terms*— Character recognition, classifier combination, decision combination, majority vote problem.**

I. Introduction

IN THE domain of OCR, there has been a recent movement toward combining the decisions of several classifiers in order to arrive at improved recognition results. The combination can be implemented using a variety of strategies. In [11] and [14], the combined decision is obtained by a majority vote of the individual classifiers, while variations of this scheme are implemented in [6], [11], and [15]. When the individual classifiers produce ranked lists of decisions, these rankings can be used to obtain combined decisions [9]. Further developments in deriving a combined decision include statistical approaches [4], [10], formulations based on Bayesian and Dempster–Shafer theories of evidence [4], [15], and neural networks [12]. In all these cases, it was found that using a combination of classifiers has resulted in a remarkable improvement in the recognition results, and this is true regardless of whether the classifiers are independent or make use of orthogonal features.

Among all the combination methods, majority vote is by far the simplest for implementation. It does not assume prior knowledge of the behavior of the individual classifiers (also called *experts*), and it does not require training on large quantities of representative recognition results from the experts. Yet, in a very recent study [12], when five combination strategies (majority vote, Bayesian, logistic regression, fuzzy integral, and neural network) are employed on seven classifiers, the results show that the majority vote is just as effective as the other more complicated schemes in improving the recognition rate for the data set used.

The last point deserves some attention, because all the results reported thus far on the majority vote have been experimental in nature—given a particular set of classifiers on a certain database, certain results have been obtained. Consequently, there is no assurance that similar improvements can be obtained when a different database is used, or when a different set of classifiers are combined. Therefore this method requires a closer scrutiny, so that its behavior can be better understood and used to advantage not only in character recognition, but also in other pattern recognition areas where a multiplicity of algorithms exist, each producing a set of well-defined outcomes. Examples of these areas could be speech recognition and other problems in computer vision where it may not be realistic to expect large volumes of data for training classifier combinations. For these applications, majority vote is the most appropriate option.

This work is concerned with understanding *how* and *why* the combination of expert opinions by majority vote can produce improved recognition results, and the assumptions under which this can be expected to happen. To achieve this purpose, we will examine the topic starting from its logical foundations, that is, from the classical voting problem. For this particular problem, the binomial distribution has been used to determine the probabilities of consensus, but only for an odd number of voters. We will extend these results to even numbers of voters, and compare their results with those for odd numbers so that an ordering of probabilities can be obtained for combining different numbers of experts.

Manuscript received April 1, 1994; revised August 29, 1996. This work was supported by the Natural Sciences and Engineering Research Council of Canada, the National Networks of Centres of Excellence program of Canada, and the FCAR program of the Ministry of Education of the Province of Québec.

L. Lam is with the Department of Science and Mathematics, Hong Kong Institute of Education, Northcote Campus, Hong Kong (e-mail: llam@nc.ied.edu.hk).

C. Y. Suen is with the Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, P.Q., Canada H3G 1M8 (e-mail: suen@cenparmi.concordia.ca).

Publisher Item Identifier S 1083-4427(97)06202-4.

From this model, we will relax the assumptions of equal probabilities among the experts and examine the effects to the consensus created by the addition of new experts. This problem is studied without assuming independence of the experts whenever possible. With this assumption, conditions are derived as to *when* these additions would improve the results of the combination. Interestingly, these conditions are related to the widely studied notion of the odds ratio. Throughout the paper, we will compare our findings against experimental results.

II. The Voting Problem

In the rest of this paper, we assume that n classifiers or experts are deployed, and that for each input sample, each expert produces a unique decision regarding the identity of the sample. This identity could be one of the allowable classes, or a rejection when no such identity is considered possible. In the event that the decision can contain multiple choices, the top choice would be selected. In combining the decisions of the n experts, the sample is assigned the class for which there is a consensus, or when at least k of the experts are agreed on the identity, where

$$k = \begin{cases} \dfrac{n}{2} + 1 & \text{if } n \text{ is even} \\ \dfrac{n+1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

Otherwise, the sample is rejected. Since there can be more than two classes, the combined decision is correct when a majority of the experts are correct, but wrong when a majority of the decisions are wrong *and* they agree. A rejection is considered neither correct nor wrong, so it is equivalent to a neutral position or an abstention. For the problem we are examining, there is only one correct answer but many wrong ones for each individual. Consequently, there will be cases where the group has no consensus, leading to a rejection. This possibility of a rejection by the group would exist whether each individual has the rejection option or not, and so we include the reject option for each expert for the sake of generality.

While each classifier has the possibilities of being correct, wrong, or neutral, the combined (correct) recognition rate is really the probability of the consensus being correct, assuming each vote to have only two values—correct or not. In other words, errors and rejections can be grouped together as the other possibility when the correct rate is considered. However, in this case, the overall error rate of the combination cannot be calculated directly from the error rate of each classifier. Due to the nature of consensus, the combined decision is wrong only when a majority of the votes are wrong *and* they make the same mistake. Of course, this is a strength of this combination method—due to the large number of possible mistakes, the majority would not often make the same one. As a result of this need for consensus, we can only calculate the probability of the consensus committing a particular error from the individual probabilities of committing the same error. To assess this particular (mistaken) probability of consensus, we can also consider each vote to have only two values—it makes this particular mistake or not.

To avoid confusion in dealing with our problem, we should distinguish between the number of choices available to the voter (expert), and the number of values his/her vote would have. In the example of m possible classes, the expert would have $m+1$ choices for each classification. However, when we consider the recognition rate of the consensus, each vote would have only two values—correct or not. The probability of this vote being correct would coincide with the recognition rate r of the expert, while the other option has probability $1-r$. Analogously, when we consider the probability of the consensus making a particular mistake (misclassifying a sample of "2" as "3," for example), the two values of each vote are whether the expert makes this particular mistake or not, and these values have probabilities s and $1-s$, given that the expert makes this misclassification with probability s.

Consequently, both cases reduce to the problem of determining the probability of consensus when each vote has only two alternatives, and so they can be considered as the same problem with different parameters. With this situation of two alternatives, the subject had been much studied as the classical voting problem under the following assumptions:

A1: The number of voters is odd.

A2: Each voter has the same probability p of voting one way (for example, correctly).

A3: The individual decisions are independent.

Of these assumptions, A1 will be eliminated from the outset, with interesting consequences. Then new results are obtained from the cases when A2 is not assumed, and subsequently we will examine the consequences when A3 is relaxed.

The different premises between the classical problem and the current one are summarized in Table I, where m is the number of possible recognition classes for the current problem.

We will use $P(C_C)$ and $P(C_W)$ to denote the probabilities of the consensus being correct and wrong respectively. One consequence of the last difference shown in Table I lies in the behavior of $P(C_W)$ as $P(C_C)$ changes, or vice versa. In the original problem, $P(C_W)+P(C_C)=1$, so they would change in opposite directions, and maximizing one quantity would minimize the other. This is no longer true when the consensus has the reject option, since $P(C_W)+P(C_C)\leq 1$. For this case, decreasing $P(C_W)$ does not necessarily imply that $P(C_C)$ would increase correspondingly (since the reject probability can increase). The nature and magnitude of the changes are found to be related to whether n is even or odd. These are some of the aspects that will be presented in Section III.

III. General Results on the Probability of Consensus

If we assume that n independent experts have the same probability p of being correct, then the probability of the consensus being correct, denoted by $P_C(n)$, can be computed using the binomial distribution as

$$P_C(n) = \sum_{m=k}^{n} \binom{n}{m} p^m (1-p)^{n-m}$$

where the value of k is as defined in Section II. Condorcet [3] is usually credited with first recognizing this fact, and his

TABLE I
DIFFERENCES BETWEEN CLASSICAL AND CURRENT PROBLEMS

Difference	Classical problem	Problem under study
Number of voters	Odd	Odd or even
Number of choices for each voter	2 (correct or wrong)	1 correct m-1 wrong 1 rejection
Existence of consensus	Always	Not guaranteed: lack of consensus leads to rejection

work became the basis of much modern research in voting and decision making (for example, [1] and [7]). The following theorem, attributed to him, has provided validity to the belief that the judgment of a group is superior to those of individuals provided the individuals have reasonable competence.

Theorem 0: Suppose n is odd and $n \geq 3$. Then the following are true:

1) If $p > 0.5$, then $P_C(n)$ is monotonically increasing in n and $P_C(n) \to 1$ as $n \to \infty$.
2) If $p < 0.5$, then $P_C(n)$ is monotonically decreasing in n and $P_C(n) \to 0$ as $n \to \infty$.
3) If $p = 0.5$, then $P_C(n) = 0.5$ for all n.

The following recursive formula is given in [8], but the derivation is unpublished.

Corollary to Theorem 0: If n is odd and $n \geq 3$, then

$$P_C(n+2) = P_C(n) + p^2 \binom{n}{\frac{n+1}{2}} p^{(n-1)/2}(1-p)^{(n+1)/2} - (1-p)^2 \binom{n}{\frac{n+1}{2}} p^{(n+1)/2}(1-p)^{(n-1)/2}.$$

In the rest of this section, we will generalize the above theorem to even as well as odd values of n. The following theorem and corollaries apply when $n \geq 1$.

Theorem 1:

$$P_C(2n+1) = P_C(2n) + p^{n+1}(1-p)^n \binom{2n}{n} \quad \text{and}$$

$$P_C(2n) = P_C(2n-1) - p^n(1-p)^n \binom{2n-1}{n}$$

Proof:

$$\begin{aligned} P_C(2n+1) &= \sum_{m=n+1}^{2n+1} p^m(1-p)^{2n+1-m}\binom{2n+1}{m} \\ &= \sum_{m=n+1}^{2n+1} p^m(1-p)^{2n+1-m}\left[\binom{2n}{m} + \binom{2n}{m-1}\right] \\ &= (1-p)\sum_{m=n+1}^{2n+1} p^m(1-p)^{2n-m}\binom{2n}{m} \\ &\quad + p\sum_{m=n+1}^{2n+1} p^{m-1}(1-p)^{2n-(m-1)}\binom{2n}{m-1} \\ &= (1-p)\sum_{m=n+1}^{2n} p^m(1-p)^{2n-m}\binom{2n}{m} \\ &\quad + p\sum_{k=n}^{2n} p^k(1-p)^{2n-k}\binom{2n}{k} \\ &\quad \text{since } \binom{2n}{2n+1} = 0 \\ &= (1-p+p)\sum_{k=n+1}^{2n} p^k(1-p)^{2n-k}\binom{2n}{k} \\ &\quad + p^{n+1}(1-p)^n\binom{2n}{n} \\ &= P_C(2n) + p^{n+1}(1-p)^n\binom{2n}{n}, \end{aligned}$$

and

$$\begin{aligned} P_C(2n) &= \sum_{m=n+1}^{2n} p^m(1-p)^{2n-m}\binom{2n}{m} \\ &= \sum_{m=n+1}^{2n} p^m(1-p)^{2n-m}\left[\binom{2n-1}{m} + \binom{2n-1}{m-1}\right] \\ &= (1-p)\sum_{m=n+1}^{2n} p^m(1-p)^{2n-m-1}\binom{2n-1}{m} \\ &\quad + p\sum_{m=n+1}^{2n} p^{m-1}(1-p)^{2n-m}\binom{2n-1}{m-1} \\ &= (1-p)\sum_{m=n+1}^{2n-1} p^m(1-p)^{2n-1-m}\binom{2n-1}{m} \\ &\quad + p\sum_{k=n}^{2n-1} p^k(1-p)^{2n-1-k}\binom{2n-1}{k} \\ &\quad \text{since} \binom{2n-1}{2n} = 0 \end{aligned}$$

TABLE II
VALUES OF $P_C(n)$ FOR DIFFERENT VALUES OF p AND n

p	Values of n								
	2	3	4	5	6	7	8	9	10
0.10	0.0100	0.0280	0.0037	0.0086	0.0013	0.0027	0.0004	0.0009	0.0001
0.15	0.0225	0.0608	0.0120	0.0266	0.0059	0.0121	0.0029	0.0056	0.0014
0.20	0.0400	0.1040	0.0272	0.0579	0.0170	0.0333	0.0104	0.0196	0.0064
0.25	0.0625	0.1562	0.0508	0.1035	0.0376	0.0706	0.0273	0.0489	0.0197
0.30	0.0900	0.2160	0.0837	0.1631	0.0705	0.1260	0.0580	0.0988	0.0473
0.35	0.1225	0.2818	0.1265	0.2352	0.1174	0.1998	0.1061	0.1717	0.0949
0.40	0.1600	0.3520	0.1792	0.3174	0.1792	0.2898	0.1737	0.2666	0.1662
0.45	0.2025	0.4253	0.2415	0.4069	0.2553	0.3917	0.2604	0.3786	0.2616
0.50	0.2500	0.5000	0.3125	0.5000	0.3438	0.5000	0.3633	0.5000	0.3770
0.55	0.3025	0.5748	0.3910	0.5931	0.4415	0.6083	0.4770	0.6214	0.5044
0.60	0.3600	0.6480	0.4752	0.6826	0.5443	0.7102	0.5941	0.7334	0.6331
0.65	0.4225	0.7183	0.5630	0.7648	0.6471	0.8002	0.7064	0.8283	0.7515
0.70	0.4900	0.7840	0.6517	0.8369	0.7443	0.8740	0.8059	0.9012	0.8497
0.75	0.5625	0.8438	0.7383	0.8965	0.8306	0.9294	0.8862	0.9511	0.9219
0.80	0.6400	0.8960	0.8192	0.9421	0.9011	0.9667	0.9437	0.9804	0.9672
0.85	0.7225	0.9393	0.8905	0.9734	0.9527	0.9879	0.9786	0.9944	0.9901
0.90	0.8100	0.9720	0.9477	0.9914	0.9842	0.9973	0.9950	0.9991	0.9984
0.95	0.9025	0.9928	0.9860	0.9988	0.9978	0.9998	0.9996	1.0000	0.9999

$$= (1-p+p)\sum_{k=n}^{2n-1} p^k(1-p)^{2n-1-k}\binom{2n-1}{k} - (1-p)p^n(1-p)^{n-1}\binom{2n-1}{n}$$
$$= P_C(2n-1) - p^n(1-p)^n\binom{2n-1}{n}.$$

The next three corollaries are direct consequences of Theorem 1.

Corollary 1:

$$P_C(2n+1) - P_C(2n-1) = p^n(1-p)^n\binom{2n-1}{n}(2p-1).$$

Proof: From Theorem 1,

$$P_C(2n+1) - P_C(2n-1)$$
$$= p^{n+1}(1-p)^n\binom{2n}{n} - p^n(1-p)^n\binom{2n-1}{n}$$
$$= p^n(1-p)^n\left\{p\left[\binom{2n-1}{n} + \binom{2n-1}{n-1}\right] - \binom{2n-1}{n}\right\}$$
$$= p^n(1-p)^n\left[(p-1)\binom{2n-1}{n} + p\binom{2n-1}{n-1}\right]$$
$$= p^n(1-p)^n\binom{2n-1}{n}(p-1+p)$$

since $\binom{2n-1}{n-1} = \binom{2n-1}{n}$

$$= p^n(1-p)^n\binom{2n-1}{n}(2p-1).$$

We note that this conclusion coincides with that of Corollary to Theorem 0 when the latter result has been simplified.

Corollary 2:

$$P_C(2n+2) - P_C(2n) = p^{n+1}(1-p)^n\binom{2n}{n}\left[\frac{2np+p-n}{n+1}\right].$$

Corollary 3:

$$P_C(2n+2) - P_C(2n-1) = p^n(1-p)^n\binom{2n}{n}\left[\frac{(4n+2)p^2 - 2np - (n+1)}{2(n+1)}\right].$$

From the preceding results, we can deduce the following remarks when $0 < p < 1$.

1) As immediate consequences of Theorem 1,

$$P_C(2n) < P_C(2n+1) \quad \text{and}$$
$$P_C(2n) < P_C(2n-1) \quad \text{for all } n \text{ and } p.$$

2) When even numbers $2n$ of experts are combined, $P_C(2n)$ is monotonically increasing if $p > n/(2n+1)$, which is true for all n if $p \geq 1/2$. Conversely, $P_C(2n)$ is monotonically decreasing with n if $p < n/(2n+1)$, which is true for all n if $p < 1/3$. When $1/3 \leq p < 1/2$, however, the behavior of $P_C(2n)$ would depend on the relative magnitudes of p and $n/(2n+1)$.

Suppose $p = 0.4$. Then $P_C(2) < P_C(4)$ since $p > n/(2n+1) = 1/3$, while $P_C(4) = P_C(6)$ since $p = n/(2n+1) = 2/5$, and $P_C(6) > P_C(8)$ since $p < 3/7$. This represents a departure from the odd cases, where $P_C(n)$ is monotonically decreasing for all $p < 1/2$. These differences can be seen in Table II,

TABLE III
(a) PERFORMANCES OF INDIVIDUAL EXPERTS ON CENPARMI DATABASE AND (b) PERFORMANCES OF SOME COMBINATIONS

Method	Rec.	Subs.	Rej.
E1	86.05	2.25	11.70
E2	92.85	2.45	4.70
E3	92.95	2.15	4.90
E4	93.90	1.60	4.50
E5	96.95	3.05	0.00
E6	98.30	1.70	0.00

(a)

Combination	Rec.	Subs.	Rej.
Two Experts			
E1+4	81.75	0.00	18.25
E1+5	84.15	0.05	15.80
E1+6	85.40	0.10	14.50
E2+4	88.40	0.00	11.60
E2+6	91.95	0.05	8.00
E3+4	88.80	0.00	11.20
E3+6	91.90	0.15	7.95
E4+5	91.60	0.10	8.30
E4+6	92.95	0.10	6.95
E5+6	96.55	0.85	2.60
Three Experts			
E2+3+4	96.40	0.10	3.50
E2+3+6	97.45	0.30	2.25
E2+4+5	97.60	0.15	2.25
E2+5+6	98.30	0.85	0.85
E3+4+6	97.65	0.25	2.10
E4+5+6	98.50	0.85	0.65
Four Experts			
E1+2+3+4	92.50	0.00	7.50
E2+3+4+5	95.45	0.00	4.55
E2+3+4+6	95.70	0.00	4.30
E2+4+5+6	97.00	0.15	2.85
Five Experts			
E1+2+3+4+5	97.55	0.05	2.40
E1+2+3+4+6	97.70	0.10	2.20
E2+3+4+5+6	98.25	0.25	1.50
Six Experts			
E1+2+3+4+5+6	97.05	0.05	2.90

(b)

where values of $P_C(n)$ are shown for different values of n and p.

3) When both even and odd numbers of experts are considered together, we know from Remark 1) that $P_C(2n+2) < P_C(2n+3)$ and $P_C(2n+2) < P_C(2n+1)$ for all p. The relation between $P_C(2n+2)$ and $P_C(2n-1)$ is given by Corollary 3, from which it follows that

$$P_C(2n+2) > P_C(2n-1) \quad \text{iff}$$
$$p > f_1(n) = \frac{n+\sqrt{5n^2+6n+2}}{4n+2}$$

and $P_C(2n+2) < P_C(2n-1)$ iff the reverse inequality holds for p. Since $f_1(n)$ is increasing and approaches $p_u = (1+5^{1/2})/4$ as $n \to \infty, P_C(2n+2) > P_C(2n-1)$ for all n if $p \geq p_u$ ($\doteq 0.8090$).

For example, when $p = 0.75, P_C(8) < P_C(5)$ since $p < [3+(65)^{1/2}]/14$, while the opposite holds for $p = 0.8$.

4) As a consequence of Remarks 1) and 2), we can conclude that when $p \geq p_u$,

$$P_C(2n) < P_C(2n-1) < P_C(2n+2) < P_C(2n+1) < P_C(2n+4) < P_C(2n+3)$$

for all n. For example,

$$P_C(2) < P_C(1) < P_C(4) < P_C(3) < P_C(6) < P_C(5) < P_C(8) < P_C(7) < \cdots$$

and these are shown by the results in Table II where $p \geq 0.85$.

5) Remark 4) defines the ordering of $P_C(n)$ for sufficiently large values of p. We now consider small values of p to consider the probabilities of consensus errors. If $p < 1/3$, the following inequalities are true for all n.

a) $P_C(2n+1) < P_C(2n-1)$ by Theorem 0;
b) $P_C(2n) < P_C(2n-2)$ by Corollary 2; and
c) $P_C(2n) < P_C(2n+1)$ by Theorem 1.

To obtain a complete ordering of these probabilities, it remains to compare $P_C(2n+1)$ with $P_C(2n-2)$. From Theorem 1,

$$P_C(2n+1) = P_C(2n-2) + p^n(1-p)^{n-1}\binom{2n-1}{n} \cdot \left[\frac{(2-4n)p^2+(6n-3)p+(1-n)}{2n-1}\right]$$

so $P_C(2n+1) < P_C(2n-2)$ when $(2-4n)p^2+(6n-3)p+(1-n) < 0$, which is true when

$$p < f_2(n) = \frac{3}{4} - \frac{1}{4}\sqrt{5+\frac{4}{2n-1}}.$$

Since $f_2(n)$ is monotonically increasing with a minimum value $p_l \doteq 0.1208$ when $n = 2, P_C(2n+1) < P_C(2n-2)$ if p is below this value.

For these small values of p, the consensus probabilities are ordered as

$$P_C(2n+2) < P_C(2n) < P_C(2n+1) < P_C(2n-2) < P_C(2n-1)$$

for all n. For example, we would have

$$P_C(8) < P_C(9) < P_C(6) < P_C(7) < P_C(4) < P_C(5) < P_C(2) < P_C(3).$$

From Table II, it can be seen that this is true for $p = 0.1$, but not for $p = 0.15$ which exceeds the threshold value p_l.

6) According to Theorem 1, $P_C(2n-1) - P_C(2n) = p^n(1-p)^n\binom{2n-1}{n}$. By considering the convergence of the series $\Sigma_{n=1}^{\infty}\binom{2n-1}{n}[p(1-p)]^n$, we can conclude that as $n \to \infty, p^n(1-p)^n\binom{2n-1}{n} \to 0$ for $0 < p < 1, p \neq 0.5$. For $p = 0.5$, the sequence $\{(\frac{1}{2})^{2n}\binom{2n-1}{n}\}$ can be shown to converge to 0 by comparison with $\{1/\sqrt{2n+1}\}$. Therefore $P_C(2n)$ approaches the same limit as $P_C(2n-1)$ for all values of p in (0, 1).

IV. Application to Pattern Recognition

The discussion in Section III presents a mathematical model for comparing the recognition rates obtained from the majority vote of n independent experts when each expert has recognition rate p. We assume the experts to have reasonable performance, so that p is greater than the threshold p_u given above, in which case the ordering of the consensus probabilities is stated in Remark 4). In particular, it follows that a combination of an even number n of experts would yield a recognition rate that is lower than those obtained from both $n+1$ and $n-1$ experts.

In pattern recognition applications, it is also an important consideration that the results should have low error or substitution rates. For these error rates, we can consider the consensus probabilities for small p. If the probabilities of each expert making a particular mistake are approximately equal to p, then we can certainly assume that $p<p_l$, in which case a combination of an even number n of experts would be less likely to commit this error than the consensus of $n+1$ or $n-1$ experts. With this assumption of approximate uniformity, the same conclusion regarding the number of experts can be applied to the overall substitution rates, which are after all the summation of the probabilities of particular errors.

The assumption of equal probabilities has made possible the computation of exact differences in the likelihoods of consensus, whether it is the correct or wrong decision. Admittedly, the assumption of equal probability, while convenient in theory, is impossible to achieve in practice—different experts cannot be expected to operate with equal probabilities in real-life situations. For this reason, we will examine the consequences of relaxing this condition in the next section. Actually, the ordering of the probabilities derived in Remarks 4) and 5) has been demonstrated in experiments where the performances of experts do differ. The results of these experiments are described below.

In the first experiment, six classifiers are applied to the recognition of handwritten numerals from the CENPARMI database. While the first four recognizers [14] had been developed independently of one another, the last two experts [5] are very similar in their behavior because they are adapted to the same feature vector. The performances of the individual methods are shown in Table III(a) and part of the combined results are given in Table III(b). The results shown in these tables differ to a certain extent from those presented in [5] for two reasons. The substitution rate of E2 has been decreased, and integer votes are used here. In the previous work, a fraction of a vote is assigned to each candidate class of E1 when this expert cannot differentiate between two or three classes, whereas in the present context this would be considered to be a rejection by this expert.

From these tables, it is clear that the performances of the experts differ significantly, but the combined recognition and substitution rates mostly follow the patterns stated in Remarks 5) and 6). For example, E2 + 6 produces lower recognition and substitution rates than E2 + 3 + 6 or E2 + 5 + 6, while E2 + 3 + 6 has higher rates than E2 + 3 + 4 + 6, which has lower rates than E1 + 2 + 3 + 4 + 6, and so on.

TABLE IV
Performances of CEDAR Classifiers on BS Database

Classifier	Recognition
Binpoly	93.99
Chaincode	96.38
Gabor	95.17
Gradient	96.20
GSC	97.05
Histogram	93.88
Morphology	95.76

In the second experiment, the data consists of the recognition results obtained by seven classification algorithms developed at CEDAR in Buffalo, NY. The test set BS contains 2711 handwritten numerals extracted from United States Postal Service mailpieces, and these are contained on CEDAR CD-ROM together with the recognition results. Fuller descriptions of the classifiers are given in [12]. When only the top choice is considered, the individual algorithms produce the results shown in Table IV with no rejections, i.e., each input sample is assigned its nearest class.

These classifiers have 120 possible combinations, whose recognition results on the BS database are shown in Fig. 1, where the scatter plot shows the recognition versus the substitution rates. The two disjoint clusters of points (one resulting from combinations of even numbers of experts and the other from odd numbers) tend to illustrate the comments made above. Combining the decisions of odd numbers of experts produces higher correct as well as higher error rates, so the corresponding points in Fig. 1 are positioned to the upper right, while the even combinations result in points located to the left and below (representing lower substitution and correct rates). The tendencies of these combinations will be further developed in the next section, in which we will consider the combinations of experts with unequal probabilities of being correct (which is the case for this example).

V. Relaxation of Equal Probability Assumption

In Section III of this paper, we have assumed equal probabilities of correct classification, which has made possible the determination of the exact differences between the consensus probabilities. Since this assumption cannot be expected to be true in practice, in this section we derive comparisons between the consensus probabilities when the assumption of uniform p is relaxed. In particular, we will examine the differences created by the addition of votes to a group of voters (even as well as odd in number). Given the unequal probabilities, the exact differences would depend on the individual probabilities, but the *sign* of these differences can be easily determined when one vote is added. When two votes are added, the change in the consensus probabilities will be expressed precisely in terms of the individual probabilities. Then, by using a theorem in Graph Theory, the sign of this difference is found to depend on the

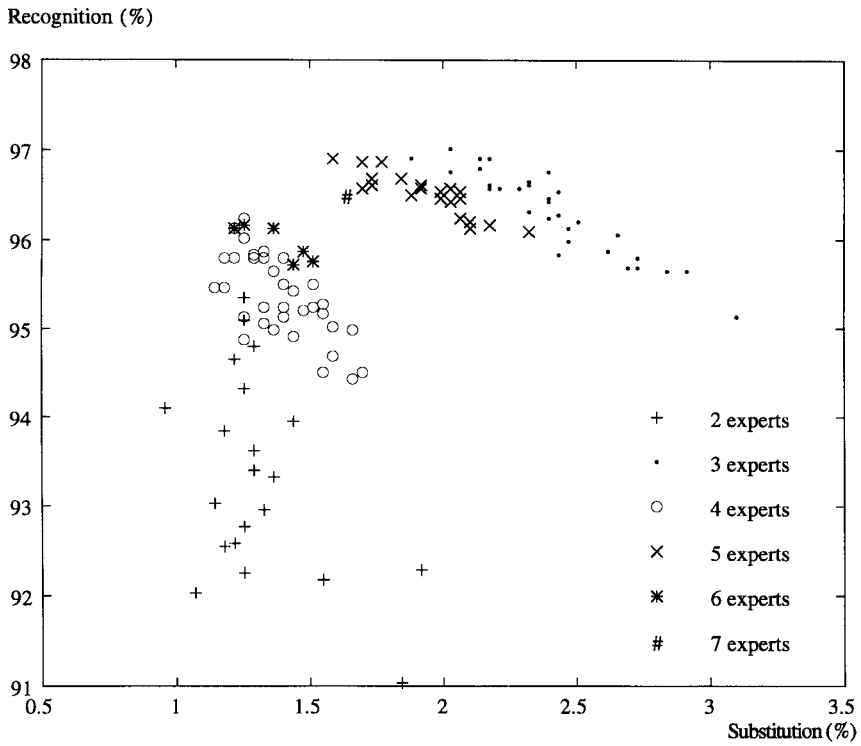

Fig. 1. Combined results of CEDAR classifiers.

TABLE V
EFFECT OF ADDING ONE VOTE TO $2n$ VOTES

Case	Original *2n* votes	New vote	Original Decision	New decision
1	***n*** correct	Correct	Reject	Correct
2	***n*** wrong and agree	Wrong and agree	Reject	Wrong

TABLE VI
EFFECT OF ADDING ONE VOTE TO $2n + 1$ VOTES

Case	Original *2n+1* votes	New vote	Original decision	New decision
1	***n+1*** correct (class c1)	Not c1	Correct	Reject
2	***n+1*** wrong & agree (class c2)	Not c2	Wrong	Reject

familiar notion of the odds ratio. These results will be derived and discussed below.

A. Addition of One Vote

We will first consider the effect (on the probability of the group decision being correct or wrong) of adding one vote to $2n$ and $2n+1$ votes respectively. In each case, the addition of the new vote would make a difference only when the original group decision had been split in a "marginal" way, so that the new vote could tip the balance.

When the original group has $2n$ voters, this would be the case if n of the votes had been in agreement—either they are correct, or they make the same mistake. Depending on the decision of the new vote, the possible changes are summarized in Table V. When a change does occur, it is in the direction of reducing the rejection rate, changing it into a correct decision part of the time and an error in the other cases. In other words, adding one vote to $2n$ (which also changes the number of voters from even to odd) has the effect of reducing the degree of "indecision," changing it into correct or wrong decisions.

On the other hand, the addition of one vote to $2n + 1$ votes would change the group decision only if $n + 1$ of the original votes are in agreement, and the new vote disagrees, thus changing the original majority to a lack of consensus. These possibilities are summarized in Table VI. The end result is that the rejection rate would increase, while both the correct and error rates would decrease. We can consider this to be a result of changing an odd number of voters into an even number, when more "tied" votes may occur.

TABLE VII
EFFECT OF ADDING TWO VOTES TO $2n$ VOTES

Case	Original $2n$ votes	New votes	Original decision	New decision
1	n correct (class c1)	(c1, c1)	Reject	Correct
2	$n+1$ correct (class c1)	(not c1, not c1)	Correct	Reject

The trends shown in Tables V and VI are true regardless of the values of the probabilities of the individual experts. The individual probabilities, however, do determine the magnitudes of the changes. If each expert is correct much more often than (he/she is) wrong, then the probability of Case 1 is expected to be greater than that of Case 2 whether the number of voters is even or odd. Since no assumption on the independence of the experts has been made, these results would always be valid.

B. Addition of 2 Votes to $2n$ Votes

If we consider the addition of two votes to an even (or odd) number of votes as the repeated addition of one vote, then it is not clear what the net effect of the two additions would be, since the second step appears to reverse the trend of the first. For this reason, we have to examine the results when the two votes are added together to an existing group.

Suppose the original voters have probabilities $p_i, 1 \leq i \leq 2n$, of being correct, and for the new votes these probabilities are q_1 and q_2. The addition of the new votes would affect the correct rate only in the cases shown in Table VII.

In Case 1, the two new correct votes would change the original tied vote into a majority, while in Case 2 the new votes would deprive the original decision of a majority. Since the first case causes the correct rate to increase while the second causes it to decrease, the net change to this rate depends on the relative probabilities of the two cases. We will calculate the probability of each case when the expert opinions are assumed to be independent.

Let A denote the set of $\binom{2n}{n}$ vectors of the form $(p'_1, p'_2, \cdots, p'_{2n})$, where for each i,

$$p'_i = \begin{cases} p_i & \text{for } n \text{ terms} \\ 1-p_i & \text{for the other } n \text{ terms} \end{cases}$$

and let B be the set of $\binom{2n}{n+1}$ vectors of the form $(p''_1, p''_2, \cdots, p''_{2n})$, where for each i,

$$p''_i = \begin{cases} p_i & \text{for } n+1 \text{ terms} \\ 1-p_i & \text{for } n-1 \text{ terms.} \end{cases}$$

Then for every vector $\boldsymbol{a}$ in A, there exist exactly n vectors in B that differ from $\boldsymbol{a}$ at only one component. These n vectors of B are obtained by replacing each of the $(1-p_i)$ terms in $\boldsymbol{a}$ by p_i. Similarly, for each $\boldsymbol{b}$ in B, there are $n+1$ vectors in A that differ from $\boldsymbol{b}$ at only one component, each of which is obtained by changing a p_i in $\boldsymbol{b}$ into $(1-p_i)$. Since $\binom{2n}{n} > \binom{2n}{n+1}, |A| > |B|$.

In order to determine the difference between the probabilities of Case 1 and Case 2 in Table VII, we prove and use the following result:

Theorem 2: There exists a one-to-one function f from B into A such that for every $\boldsymbol{b}$ in $B, \boldsymbol{b}$ and $f(\boldsymbol{b})$ differ at only one component.

Proof: Define G to be a graph whose vertices are the set of all vectors in $A \cup B$, and every vertex $\boldsymbol{c}$ in G is adjacent to all vertices that differ from $\boldsymbol{c}$ at only one component. Then G is a bipartite graph in which every vertex of A is adjacent to n vertices in B, and every vertex of B is adjacent to $n+1$ vertices in A. The function f corresponds to a complete matching of B to A, and its existence reduces to the well-known "marriage" problem.

By Hall's Theorem, such a matching exists iff every subset of k vertices in B must be collectively adjacent to at least k distinct vertices in A. We now show that this condition is satisfied in the present context. Suppose the subset of k vertices in B is $B' = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_k\}$. If we list all the vertices in A that B' is collectively adjacent to, we obtain $k(n+1)$ vertices by the comment in the last paragraph. We denote this set of (not necessarily distinct) $k(n+1)$ vertices by A', and we need to show that there are at least k distinct vertices in A'.

If A' has less than k distinct vertices, then at least one vertex $\boldsymbol{a}'$ must appear in A' more than $n+1$ times, implying that $\boldsymbol{a}'$ must be adjacent to more than $n+1$ vertices in B' (since each occurrence of $\boldsymbol{a}'$ corresponds to an adjacent vertex in B'). This is a contradiction, and so the hypothesis of Hall's Theorem must be satisfied, or the function f exists.

We now use this theorem to compare the probabilities of Cases 1 and 2 in Table VII, which we denote by $P\{reject \rightarrow correct\}$ and $P\{correct \rightarrow reject\}$ respectively. The net increase in the correct rate would be

$$P\{reject \rightarrow correct\} - P\{correct \rightarrow reject\} = q_1 q_2 \sum_A \prod_{i=1}^{2n} p'_i - (1-q_1)(1-q_2) \sum_B \prod_{i=1}^{2n} p''_i \quad (5.1)$$

Given the existence of f by Theorem 2, each term in the second sum has a corresponding term in the first sum such that the 2 terms differ at only one component. Hence

$$P\{reject \rightarrow correct\} - P\{correct \rightarrow reject\} > 0 \quad \text{if} \quad q_1 q_2 (1-p_i) - (1-q_1)(1-q_2) p_i \geq 0, \quad \text{or} \quad \frac{q_1 q_2}{(1-q_1)(1-q_2)} \geq \frac{p_i}{1-p_i} \quad \text{for all } i. \quad (5.2)$$

In other words, the addition of two votes to $2n$ votes would increase the correct rate if the product of the odds ratio of the two new votes is not less than the odds ratio of any original vote. Since the odds ratio of any expert should be greater than one when the correct rate is considered, this condition is easy to satisfy. In the event that all the probabilities are equal,

TABLE VIII
EFFECT OF ADDING TWO VOTES TO $2n + 1$ VOTES

Case	Original $2n+1$ votes	New votes	Original decision	New decision
1	n correct (class c1)	(c1, c1)	Reject or Wrong	Correct
2	$n+1$ correct (class c1)	(not c1, not c1)	Correct	Reject or Wrong

this condition coincides with the one derived in Remark 2) of Section III. Furthermore (5.2) is a sufficient, but not necessary, condition. The correct rate is more likely to increase with the addition of two votes, for the following reasons.

i) The first sum in (5.1) contains more terms than the second, and

ii) Each term in the first sum is the product of $n + 2$ probabilities of being correct (and n probabilities of being wrong), while each term in the second sum is the product of $n+1$ and $n+1$ such probabilities, respectively. Since the probability of being correct is usually greater than that of being wrong, the first sum is expected to be greater than the second. In Example 5.2.1 below, it can be seen that the correct rate increases when condition (5.2) is satisfied, even though the two additional votes do not have better performance on their own.

Example 5.2.1: Suppose $n = 2, p_1 = p_2 = 0.8, p_3 = 0.85$ and $p_4 = 0.9$, while $q_1 = 0.7$ and $q_2 = 0.8$. Then (5.2) is satisfied for $1 \leq i \leq 4$, and in this case $P_C(4) = 0.8752$ while $P_C(6) = 0.9140$.

We note that analogous arguments can be applied to consider the change in the probability of making a mistake when we add two votes to $2n$. Suppose the original votes have probabilities $s_1, s_2, \cdots, s_{2n}$ of making this mistake while the new votes have probabilities t_1 and t_2. If we replace the notion of "being correct" in the above discussion with that of "making this mistake," the same process of reasoning would yield the result that the probability of making this mistake increases, or $P\{reject \rightarrow wrong\} - P\{wrong \rightarrow reject\} > 0$, if

$$\frac{t_1 t_2}{(1-t_1)(1-t_2)} \geq \frac{s_i}{1-s_i} \quad \text{for all } i. \tag{5.3}$$

If the probability of making a mistake is very small for each expert, then inequality (5.3) would rarely be true. However, (5.3) is a sufficient, but not necessary, condition. Actually, the sign of $P\{reject \rightarrow wrong\} - P\{wrong \rightarrow reject\}$ cannot be determined *a priori*, as it would depend on the values of n, s_i's and t_i's. This is true due to the occurrence of two opposing factors. In the expression for this difference [analogous to condition (5.1)], the first sum contains more terms than the second. At the same time, it is clear that the individual terms in the first sum are smaller than those in the second when the s_i's and t_i's are very small, since each term in the first sum is the product of $n+2$ of these small values while each term in the second is the product of only $n+1$ of them. As a result, the sign of the difference has to depend on the probabilities involved.

Therefore we can conclude that when two votes are added to $2n$ votes, it is much more probable for the correct rate to increase, while the change in the error rate would depend on the individual error rates.

C. Addition of Two Votes to $2n + 1$ Votes

There are two main differences between this case and that of Section V-B. The first difference is in the changes of decisions that can be caused by the addition of two votes in this instance, and the second is in the conditions equivalent to (5.2) and (5.3) that would apply in this case.

First, the addition of two votes to $2n+1$ can cause a change in the correct rate under the conditions of Table VIII.

As in Section V-B, we let $p_i, 1 \leq i \leq 2n+1$, and q_1, q_2 denote the probabilities of being correct. Let C be the set of $\binom{2n+1}{n}$ vectors of the form $(p'_1, p'_2, \cdots, p'_{2n+1})$ such that

$$p'_i = \begin{cases} p_i & \text{for } n \text{ terms} \\ 1 - p_i & \text{for } n+1 \text{ terms.} \end{cases}$$

Analogously, we let D represent the set of all $\binom{2n+1}{n+1}$ vectors of the form $(p''_1, p''_2, \cdots, p''_{2n+1})$ in which

$$p''_i = \begin{cases} p_i & \text{for } n+1 \text{ terms} \\ 1 - p_i & \text{for } n \text{ terms.} \end{cases}$$

Then C and D have the same cardinality, and the change in the correct rate is

$$\begin{aligned} &P\{reject\ or\ wrong \rightarrow correct\} \\ &\quad - P\{correct \rightarrow reject\ or\ wrong\} \\ &\quad = q_1 q_2 \sum_C \prod_{i=1}^{2n+1} p'_i - (1-q_1)(1-q_2) \sum_D \prod_{i=1}^{2n+1} p''_i \end{aligned} \tag{5.4}$$

which is positive, or the correct rate increases, when the odds ratios satisfy the inequality

$$\frac{q_1 q_2}{(1-q_1)(1-q_2)} > \frac{p_i}{1-p_i} \quad \text{for all } i. \tag{5.5}$$

In the case of equal probabilities, this conclusion coincides with conclusion 1) of Theorem 0. Furthermore, since C and D contain the same number of vectors, we can also conclude that the correct rate would decrease if inequality (5.5) were reversed.

Example 5.3.1: Suppose $n = 2, p_1 = p_2 = 0.8, p_3 = p_4 = 0.85$ and $p_5 = 0.9$, while $q_1 = 0.7$ and $q_2 = 0.8$. Then condition (5.5) is satisfied for $1 \leq i \leq 5$, and in this case $P_C(5) = 0.9692$ while $P_C(7) = 0.9759$.

Example 5.3.2: Suppose $n = 1, p_1 = p_2 = 0.7$, and $p_3 = 0.75$, while $q_1 = q_2 = 0.6$. Then

$$\frac{q_1 q_2}{(1-q_1)(1-q_2)} = 2.25 < \frac{p_i}{1-p_i} \quad \text{for } i = 1, 2, 3.$$

In fact, for this example, the probabilities of being correct are 0.8050 and 0.7971, respectively, before and after the addition of the two new votes.

To consider the probability of making a mistake, we can let s_i $(1 \le i \le 2n+1)$, and t_1, t_2 be the probabilities of making the mistake as before. By similar reasoning, we can conclude that the net change in the probability of making this mistake, denoted by

$$P\{reject\ or\ correct \to wrong\} - P\{wrong \to reject\ or\ correct\}$$

is positive if

$$\frac{t_1 t_2}{(1-t_1)(1-t_2)} > \frac{s_i}{1-s_i} \quad \text{for all } i \tag{5.6}$$

and the change is negative if the reverse inequality holds. Due to the small values of the odds ratios of making a mistake, condition (5.6) would rarely be true.

We now note the second difference between adding two votes to $2n$ and to $2n+1$ votes. This lies in the values of the changes in the correct rates. Suppose the condition for the change being positive is satisfied, i.e., condition (5.5) is true. Then a comparison of the expressions in (5.1) and (5.4) would indicate the former to have a higher value when the p_i's and q_i's have similar values in both expressions. This is due to the fact that the expression in (5.1) contains a number of extra terms with positive signs. In other words, adding two votes to $2n$ votes would be more effective in increasing the correct rate than the addition of two votes to $2n+1$, given similar probabilities of being correct. This can be seen in the results of Examples 1 and 2, where $P_C(7) - P_C(5) = 0.0037$ while $P_C(6) - P_C(4) = 0.0388$. In the case of equal probabilities, this can also be observed in Table II, where, for $p = 0.8$ for example, $P_C(7) - P_C(5) = 0.0246$ while $P_C(6) - P_C(4) = 0.0819$.

A difference also exists between the changes in the probabilities of making a mistake when two votes are added to $2n$ and to $2n+1$ votes. We have already made the observation that the direction of the change in the first case would depend on the values of n, s_i's, and t_i's. When two votes are added to an odd number of votes, however, it is much more likely for the error rate to decrease. In this instance, $P\{reject\ or\ correct \to wrong\} - P\{wrong \to reject\ or\ correct\}$ is expressed as the difference of two sums having an equal number of terms, and the terms in the first sum should be smaller than those in the second when the s_i's and t_i's are very small (as explained at the end of Section V-B). It follows that $P\{reject\ or\ correct \to wrong\} < P\{wrong \to reject\ or\ correct\}$ when two votes are added to an odd number of votes, while this statement cannot be made if the original number of votes is even.

This section can be summarized briefly as follows.

1) Adding one vote to an even number of votes increases both the correct and error rates while reducing the rejection rate. Exactly the opposite results are obtained when one vote is added to an odd number.

2) Adding two votes to an even or odd number of votes would increase the correct rate if the odds ratios satisfy conditions (5.2) and (5.5), respectively. Furthermore, adding two votes to an even number would be more effective in increasing the correct rate than adding the votes to an odd number, given similar probabilities. However, if reducing the error rate is the objective, then more definite gains can be obtained in the second case.

Interestingly, these conclusions can be observed in Fig. 1, where results of combining the CEDAR classifiers are shown. When one vote is added to an even number, the result moves toward the upper right (higher correct as well as error rates), while the movement is in the opposite direction when one more vote is added. This "zigzag" effect agrees with statement 1) above. At the same time, it is clear that the increase from two to four, then to six experts results in mainly an upward trend (increase in recognition rate). This is in marked contrast to the leftward movement (decrease in substitution rate) produced by increasing the number of experts from three to five and seven. These results are reflections of comment 2) above, and they are particularly noteworthy given that the independence of the expert opinions cannot be taken for granted in the experiment.

VI. Variations on Majority Vote

In the previous sections, we have derived many conclusions about the expected behavior of the consensus. For example, it is clear that the performance of the combined decision is an increasing function of the number of experts, provided each expert can perform at an appropriately high level. The number of experts that can be used would naturally depend on practical limitations, and adding new experts cannot always be readily accomplished. For this reason, in this section we consider means to combine the existing experts in more optimal ways, and derive conditions as to which of the strategies would be preferable for a given objective.

Suppose an odd number of experts are available and a higher reliability is desired for the combination. This can be easily accomplished in one of two ways: to eliminate one of the experts from voting, or to double the vote of one of the experts (i.e., assign a double weight to this vote). Either action would change the number of votes from an odd to an even number, so that the majority would produce more reliable results. Of course, doubling one vote is equivalent to the addition of a dependent vote, but it has been shown in Section V that adding one vote to an odd number would decrease both the error and correct rates regardless of independence. Analogously, when an even number of votes are given, the same options can be used to obtain an odd number of votes, when the recognition rate would be higher.

An illustration of these results is given in Fig. 2, which shows the substitution rates produced by all 56 combinations of three and five CEDAR classifiers using majority vote, together with the results obtained by doubling the best classifier of each combination and eliminating the weakest before voting. For clarity, the combinations are represented on the x axis in ascending order of their error rates by majority vote.

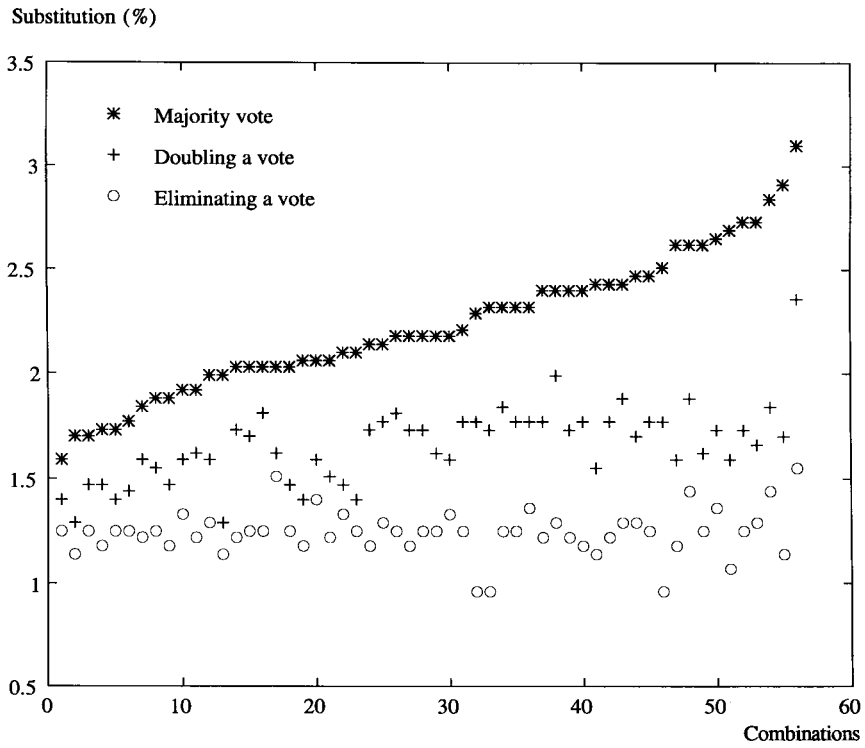

Fig. 2. Substitution rates produced by odd combinations of CEDAR classifiers.

TABLE IX
EFFECT OF ELIMINATING v_1 FROM $2n$ VOTES

Case	Original *2n* votes	v_1	Original decision	New decision
1	*n* correct (class c1)	not c1	Reject	Correct
2	*n* wrong and agree (class c2)	not c2	Reject	Wrong

TABLE X
EFFECT OF DOUBLING v_{2n} AMONG $2n$ VOTES

Case	Original *2n* votes	v_{2n}	Original decision	New decision
1	*n* correct (class c1)	c1	Reject	Correct
2	*n* wrong and agree (class c2)	c2	Reject	Wrong

Intuitively, one would double the "best" and eliminate the "worst" algorithm, where these attributes are measured according to the correct and error rates. For algorithms with no rejections, the choice is obvious; otherwise the choice would depend on the priority placed on higher recognition or lower substitution rates. Apart from this consideration, it remains to be resolved as to which alternative is better—to eliminate a vote or to double one. In the rest of this section, we will derive conditions to provide answers to this question.

A. The Even Case

As before, we suppose that the $2n$ experts are independent and they have correct probabilities $p_i, 1 \leq i \leq 2n$. For ease of notation and without loss of generality, we suppose that votes v_1 and v_{2n} are respectively the votes to be eliminated and doubled. The elimination of v_1 would lead to increases in both the correct and error rates in the cases shown in Table IX.

If we let A = the set of all vectors of the form $(1 - p_1, p_2'', p_3'', \cdots, p_{2n}'')$, where for $2 \leq i \leq 2n$,

$$p_i'' = \begin{cases} p_i & \text{for } n \text{ terms} \\ 1 - p_i & \text{for } n-1 \text{ terms,} \end{cases}$$

then the change in the correct rate resulting from the elimination of v_1 can be represented by

$$P_e\{reject \rightarrow correct\} = \sum_A (1 - p_1) \prod_{i=2}^{2n} p_i''.$$

On the other hand, doubling the vote of v_{2n} would increase both the correct and error rates in the cases indicated in Table X.

If we let B = the set of all vectors of the form $(p'_1, p'_2, \cdots, p'_{2n-1}, p_{2n})$, where for $1 \le i \le 2n-1$,

$$p'_i = \begin{cases} p_i & \text{for } n-1 \text{ terms} \\ 1-p_i & \text{for } n \text{ terms,} \end{cases}$$

then the change in the correct rate when v_{2n} is doubled can be given as

$$P_d\{reject \to correct\} = \sum_B p_{2n} \prod_{i=1}^{2n-1} p'_i,$$

and therefore

$$P_e\{reject \to correct\} - P_d\{reject \to correct\} = \sum_A (1-p_1) \prod_{i=2}^{2n} p''_i - \sum_B p_{2n} \prod_{i=1}^{2n-1} p'_i.$$

In order to compare these two sums, we will define a 1-1 function f of A onto B. The set A can be partitioned as $A = X \cup Y$, where X consists of all the elements in A with $p''_{2n} = p_{2n}$, and Y contains the rest.

If $\boldsymbol{a} \in X$, then $\boldsymbol{a}$ is also an element of B, and we define $f(\boldsymbol{a}) = \boldsymbol{a}$. If $\boldsymbol{a} \in Y$, then $\boldsymbol{a} = (1-p_1, p''_2, \cdots, p''_{2n-1}, 1-p_{2n})$, where for $\boldsymbol{b} = (p''_2, p''_3, \cdots, p''_{2n-1}), p''_i = p_i$ for n terms. Let A_2 be the set of all such vectors $\boldsymbol{b}$, and let B_2 be the set of all vectors $(p'_2, p'_3, \cdots, p'_{2n-1})$ such that

$$p'_i = \begin{cases} p_i & \text{for } n-2 \text{ terms} \\ 1-p_i & \text{for } n \text{ terms.} \end{cases}$$

By rephrasing (in terms of transversal theory) the reasoning used in the proof of Theorem 2, there exists a 1-1 function f_2 of A_2 onto B_2 such that for every $\boldsymbol{b} \in A_2, \boldsymbol{b}$ and $f_2(\boldsymbol{b})$ differ at exactly two entries. In other words, $f_2(\boldsymbol{b})$ is obtained by changing two of the p_i's in $\boldsymbol{b}$ into $(1-p_i)$'s.

Since every $\boldsymbol{a} \in Y$ would have the form $(1-p_1, \boldsymbol{b}, 1-p_{2n})$ with $\boldsymbol{b} \in A_2$, we can define $f(\boldsymbol{a}) = (p_1, f_2(\boldsymbol{b}), p_{2n})$. We note that $f|_Y$ is 1-1 because f_2 has this property, and f is 1-1 on $A = X \cup Y$ since $f(X) \cap f(Y) = \phi$.

It therefore follows that $P_e\{reject \to correct\} - P_d\{reject \to correct\} > 0$ if

$$(1-p_1)p_i p_j (1-p_{2n}) > p_1(1-p_i)(1-p_j)p_{2n} \quad \text{for all } i, j \ne 1, 2n,$$

which is true if

$$\frac{p_i}{1-p_i}\frac{p_j}{1-p_j} > \frac{p_1}{1-p_1}\frac{p_{2n}}{1-p_{2n}}. \tag{6.1}$$

If the reverse inequalities hold, then $P_e\{reject \to correct\} < P_d\{reject \to correct\}$.

Since changing rejects into correct classifications would increase the recognition rate, we can conclude that when conditions (6.1) are satisfied, eliminating v_1 would produce a higher recognition rate than doubling v_{2n}, while the opposite conclusion would be true if the inequalities were reversed.

Naturally, if we denote the odds ratio $p_i/1-p_i$ by r_i, then it is logical to consider these alternatives only when r_1 is small and r_{2n} is large. In addition, when v_{2n} is doubled, then larger values of r_{2n} should imply more improvement. On the other hand, the elimination of v_1 should lead to better results when r_1 is smaller. Therefore conditions (6.1) imply that the significant entity is the product $r_1 r_{2n}$. The ratios r_1 and r_{2n} can vary in opposite directions without affecting the sign of the difference $P_e\{reject \to correct\} - P_d\{reject \to correct\}$, provided conditions (6.1) or their opposites are satisfied. It also means that when $r_1 r_{2n}$ is small enough compared to the other odds ratios, a higher recognition rate can be obtained from eliminating v_1 than doubling v_{2n}, while the reverse is true when $r_1 r_{2n}$ is relatively large.

If s_i $(1 \le i \le 2n)$ is the probability of expert i making a particular mistake, and we consider eliminating v_1 versus doubling v_{2n}, then the same reasoning would lead to

$$P_e\{reject \to wrong\} - P_d\{reject \to wrong\} > 0 \quad \text{if } \frac{s_i}{1-s_i}\frac{s_j}{1-s_j} > \frac{s_1}{1-s_1}\frac{s_{2n}}{1-s_{2n}}$$

and the statement would also be true if all the inequalities were reversed.

Example 6.1.1: Suppose $n = 3, p_1 = 0.72, p_2 = 0.75, p_3 = p_4 = 0.8, p_5 = 0.85$ and $p_6 = 0.9$. Then $r_1 r_6 > r_i r_j$ for $i, j \ne 1, 6$, and we expect $P_e\{reject \to correct\} < P_d\{reject \to correct\}$, which is true since the former equals 0.0502 while the latter has value 0.0549.

For the special case of a two-class recognition problem in which there are no rejections, $s_i = 1 - p_i$, and so $(p_i/1-p_i)(p_j/1-p_j) > (p_1/1-p_1)(p_{2n}/1-p_{2n}) \Leftrightarrow (s_i/1-s_i)(s_j/1-s_j) < (s_1/1-s_1)(s_{2n}/1-s_{2n})$. Therefore if the inequalities are satisfied for the p_i's, it would imply that eliminating v_1 is better than doubling v_{2n}. The opposite conclusion follows when the inequalities are reversed.

In order to test the applicability of the theoretical results to a practical situation where the independence of experts cannot be assumed, we consider the combinations of four experts from Table III(a). The choice of four experts ensures that condition (6.1) or its reverse inequality will always be satisfied. Since experts E5 and E6 are highly correlated, combinations containing both of these experts are not considered. For each of the remaining nine combinations, v_1 refers to the first expert in the combination and v_4 the last, and the value of $d = r_1 r_4 - r_2 r_3$ is shown in Table XI together with the recognition rates when the vote of the expert with the highest (lowest) recognition rate is doubled (eliminated).

It is encouraging that the experimental results generally coincide with the theoretical conclusion: when $d > 0$, doubling v_4 produces higher recognition rate than eliminating v_1, and vice versa. The exceptions are in combinations four and six, in which d has very small magnitudes. It would be more illuminating if recognition results on much larger databases can be used.

B. The Odd Case

Given $2n+1$ experts, it is possible to obtain more reliable results from the combination by eliminating vote v_1 or doubling v_{2n+1}. These actions will create changes in the marginal cases shown in Tables XII and XIII, respectively. In order to compare $P_e\{correct \to reject\}$ with $P_d\{correct \to reject\}$,

TABLE XI
RESULTS OF DOUBLING v_4 VERSUS ELIMINATING v_1

Combination		$d = r_1r_4 - r_2r_3$	Recognition rate	
			Doubling v_4	Eliminating v_1
1	E1+2+3+4	−76.26	96.35	96.40
2	E1+2+3+5	24.86	97.20	97.10
3	E1+2+3+6	185.47	97.70	97.45
4	E1+2+4+5	−3.82	97.60	97.60
5	E1+2+4+6	156.78	98.00	97.70
6	E1+3+4+5	−6.88	97.70	97.40
7	E1+3+4+6	153.73	98.10	97.65
8	E2+3+4+5	209.93	97.85	97.40
9	E2+3+4+6	547.94	98.20	97.65

TABLE XII
EFFECT OF ELIMINATING v_1 FROM $2n + 1$ VOTES

Case	Original $2n+1$ votes	v_1	Original decision	New decision
1	$n+1$ correct (class c1)	c1	Correct	Reject
2	$n+1$ wrong and agree (class c2)	c2	Wrong	Reject

TABLE XIII
EFFECT OF DOUBLING v_{2n+1} AMONG $2n + 1$ VOTES

Case	Original $2n+1$ votes	v_{2n+1}	Original decision	New decision
1	$n+1$ correct (class c1)	not c1	Correct	Reject
2	$n+1$ wrong and agree (class c2)	not c2	Wrong	Reject

we determine the probabilities of occurrence of Case 1 in these tables.

If C = the set of all vectors of the form $(p_1, p_2'', p_3'', \cdots, p_{2n+1}'')$, where for $2 \le i \le 2n+1$,

$$p_i'' = \begin{cases} p_i & \text{for } n \text{ terms} \\ 1 - p_i & \text{for } n \text{ terms} \end{cases}$$

then $P_e\{correct \to reject\} = \Sigma_C\, p_1 \Pi_{i=2}^{2n+1} p_i''$.

Suppose D = the set of all vectors of the form $(p_1', p_2', \cdots, p_{2n}', 1 - p_{2n+1})$ such that for $1 \le i \le 2n$,

$$p_i' = \begin{cases} p_i & \text{for } n+1 \text{ terms} \\ 1 - p_i & \text{for } n-1 \text{ terms.} \end{cases}$$

Then $P_d\{correct \to reject\} = \Sigma_D\,(1 - p_{2n+1}) \Pi_{i=1}^{2n} p_i'$.

In this case $|C| = \binom{2n}{n} > \binom{2n}{n+1} = |D|$. By a reasoning similar to that used in the even case, there exists a 1-1 function f of D into C which allows us to conclude that $P_e\{correct \to reject\} - P_d\{correct \to reject\} > 0$ if

$$\frac{p_1}{1-p_1}\frac{p_{2n+1}}{1-p_{2n+1}} > \frac{p_i}{1-p_i}\frac{p_j}{1-p_j} \quad \text{for all } i, j \ne 1, 2n+1. \quad (6.2)$$

When this is the case, a higher correct rate should result from doubling v_{2n+1} than from eliminating v_1.

Since C contains more terms than D, no conclusion can be drawn when the reverse inequalities hold. However, this difference in the number of terms also implies that it is more likely in general to have a higher recognition rate when v_{2n+1} is doubled than when v_1 is eliminated. In addition, when conditions (6.2) are satisfied, the difference in the probabilities between the two alternatives is expected to be greater in the odd case than the even one, given similar p_i's. This can be seen by comparing the results of Examples 6.1.1 with those of Example 6.2.1 given below.

Example 6.2.1: Suppose $n = 2, p_1 = 0.72, p_2 = 0.75, p_3 = 0.8, p_4 = 0.85$ and $p_5 = 0.9$. Then conditions (6.2) are satisfied, and $P_e\{correct \to reject\} = 0.0895$ while $P_d\{correct \to reject\} = 0.0422$.

Example 6.2.2: For an actual situation, we can consider the five experts E1–E4 and E6 in Table III(a) of Section IV. The odds ratios of these experts satisfy the condition $r_1r_6 > r_ir_j$ for $i, j = 2, 3, 4$, so we expect doubling v_6 to produce a higher recognition rate than eliminating v_1, which is the case experimentally since those results are 97.25% and 95.7% respectively.

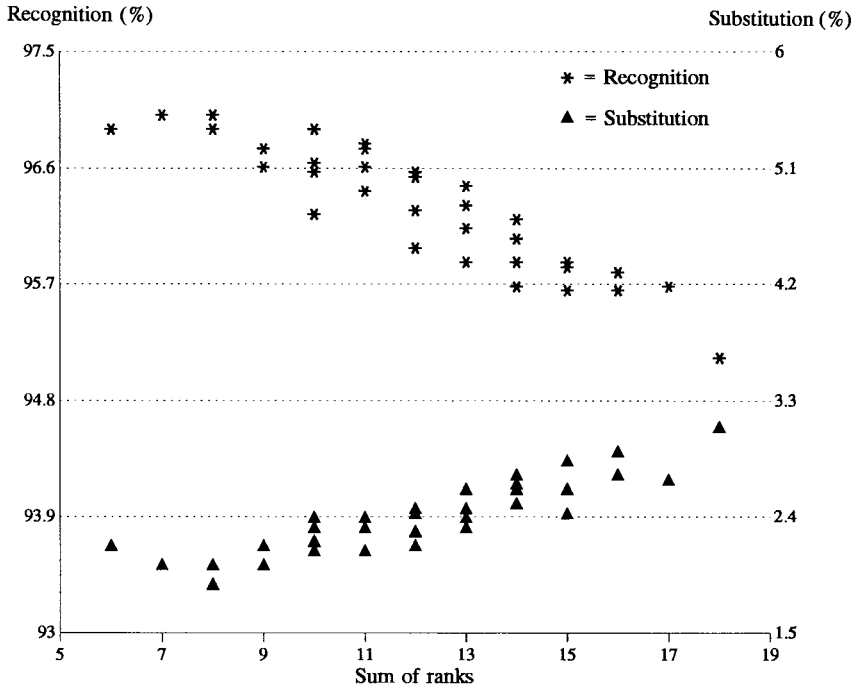

Fig. 3. Performances of combinations of three classifiers.

Analogously, by applying the same reasoning to case 2 of Tables XII and XIII, we conclude that $P_e\{wrong \rightarrow reject\} - P_d\{wrong \rightarrow reject\} > 0$ if

$$\frac{s_1}{1-s_1}\frac{s_{2n+1}}{1-s_{2n+1}} > \frac{s_i}{1-s_i}\frac{s_j}{1-s_j} \quad \text{for all } i,j \neq 1, 2n+1.$$

Therefore, when these conditions are satisfied, eliminating v_1 would result in a lower error rate than doubling v_{2n+1}. Again, due to the different number of terms involved, it is more likely that $P_e\{wrong \rightarrow reject\} > P_d\{wrong \rightarrow reject\}$, which means the elimination of v_1 should produce a lower error rate in general. This conclusion can also be observed in the odd combinations of the classifiers shown in Fig. 2.

The results of this section can be summarized as follows.

1) From an odd (even) number n of experts, an even (odd) number $n \pm 1$ of votes can be easily obtained by doubling vote v_j or eliminating vote v_i. When the majority vote is taken, the recognition and error rates of the new combinations would be both lower (higher) than those of the original, as has been stated in Section V. The advantage in eliminating v_i versus doubling v_j depends on the pairwise products of the odds ratios. If

$$r_i r_j > r_k r_l \quad \text{for all } k,l \neq i,j \tag{6.3}$$

then doubling v_j produces a higher recognition rate than eliminating v_i, for both even and odd values of n.

2) When condition (6.3) is satisfied, the gain in recognition rate is more significant for an odd number than for an even number n of experts, given that the experts have similar levels of performance. This is due to the difference in the number of terms involved in the calculation of the probabilities.

3) When n is odd, doubling v_j should result in a higher recognition rate even when the inequalities are not completely satisfied, because of the different number of terms. For the same reason, however, eliminating v_j should generally produce a lower error rate.

4) When the inequalities in (6.3) are reversed, eliminating v_i would produce a higher recognition rate than doubling v_j when n is even. In the event that n is odd, the outcome would depend on the value of n as well as the individual probabilities of the experts.

VII. Concluding Remarks

The majority voting method has been used to combine the results of classifiers for character recognition, and it has been successful from an experimental point of view. The intention of this study is to gain a deeper understanding of how this method works, and to examine its mode of operation, so that we can have confidence in its performance when applied to different data and/or experts. By this detailed analysis, we have largely achieved our objective of providing a more reliable basis for using this method. This is especially true when the decisions of the individual experts can be assumed to be independent. However, we note that even in the absence of this assumption, the experimental results do reflect the trends predicted by the theoretical considerations.

In the course of our research, we have derived many conclusions about the expected behavior of the consensus. Nevertheless, a number of decisions remain with the user. For example, the choice of an odd or even number of experts would depend on the requirements of the specific application. The former produces a higher recognition rate, and the consensus of $2n-1$ experts would outperform that of $2n$ experts in this

respect. However, it is often the case in pattern recognition applications that errors are much more costly than rejections; for example, in the precision index set by the Institute for Post and Telecommunication Policy (IPTP) of Japan, the cost of an error is ten times that of a rejection [16]. If this is an important factor, then the performance of an even number of experts would be more reliable, and their number should be incremented also by an even number at each subsequent stage of refinement. If only an odd number of experts are available, then the relative merits of doubling the best and eliminating the weakest can be considered.

It should be pointed out that combining the decisions of experts is not exactly a substitute for designing better classifiers. As has been remarked in [5], it is a truism that combinations of better algorithms tend to produce better results. This is graphically depicted here in Fig. 3, in which the performances of all combinations of three CEDAR classifiers are shown. The classifiers are ranked from one to seven according to their performance, with rank 1 for the best classifier, and so on. The values on the x axis represent the sum of the ranks for each combination, the recognition rate for each combination is shown on the y_1 axis (on the left), and the substitution rates are indicated on the y_2 axis on the right. From the behavior of the recognition and error rates as functions of the sum of ranks, it is obvious that the development of superior classifiers should remain an important objective.

Two points that may be related to the present work would be the analysis of the consensus when the experts' decisions are dependent, and the theoretical analysis of other combination methods. When the independence assumption is not applicable, a general theoretical analysis of the behavior of majority vote would be far too complex to be feasible (due to the very large number of variables whose interrelations are unknown), and it remains beyond the scope of this article. If the experts provide point estimates with a multivariate normal joint distribution of errors, then it has been shown [2] that k dependent experts are worth the same as n independent experts, where $n \leq k$**.** Under these assumptions, the equivalent number n of independent experts is a concave (down) function of k, and the upper limit for n (which depends on the common correlation ρ) is quite low. For example, even if $\rho = 0.25, n$ cannot exceed four for any k.

Other combination methods that are more specific and empirical in nature would be less likely to yield general theoretical analysis, which remains a difficult problem. Each combination method may need to be examined from its own perspective. A method which can be explored is the combination by neural networks, because this method is derived from a conceptual and mathematical framework. Recently, a one-layer perceptron (without hidden layers) has been designed in [13] to utilize concepts like weight sharing and weight decay, and it has the capabilities of eliminating redundant classifiers and dynamically selecting classifiers. Further advances may be possible in this direction.

ACKNOWLEDGMENT

The authors would like to thank Prof. J. Opatrny of Concordia University for discussions on the proof of Theorem 2, and the researchers at CEDAR, Buffalo, NY, for making their recognition results publicly accessible.

REFERENCES

[1] D. Black, *The Theory of Committees and Elections*. London, U.K.: University, 1958.
[2] R. T. Clemen and R. L. Winkler, "Limits for the precision and value of information from dependent sources," *Oper. Res.*, vol. 33, pp. 427–442, 1985.
[3] N. C. de Condorcet, *Essai sur l'Application de l'Analyze à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Paris, France: Imprimérie Royale, 1785.
[4] J. Franke and E. Mandler, "A comparison of two approaches for combining the votes of cooperating classifiers," in *Proc. 11th Int. Conf. Pattern Recognition*, The Hague, The Netherlands, 1992, vol. 2, pp. 611–614.
[5] J. Franke, L. Lam, R. Legault, C. Nadal, and C. Y. Suen, "Experiments with the CENPARMI data base combining different classification approaches," in *Proc. 3rd Int. Workshop Frontiers Handwriting Recognition*, Buffalo, NY, May 1993, pp. 305–311.
[6] P. D. Gader, D. Hepp, B. Forester, and T. Peurach, "Pipelined systems for recognition of handwritten digits in USPS zip codes," in *Proc. U.S. Postal Service Advanced Technology Conf.*, Nov. 1990, pp. 539–548.
[7] B. Grofman and G. Owen, "Condorcet models, avenues for future research," in *Information Pooling Group Decision Making: Proc. 2nd Univ. California, Irvine, Conf. Political Economy*, B. Grofman and G. Owen, Eds. Westport, CT: JAI, 1986, pp. 93–102.
[8] B. Grofman, G. Owen, and S. Field, "Thirteen theorems in search of the truth," *Theory Decision*, vol. 15, pp. 261–278, 1983.
[9] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 1, pp. 66–75, 1994.
[10] Y. S. Huang and C. Y. Suen, "An optimal method of combining multiple classifiers for unconstrained handwritten numeral recognition," in *Proc. 3rd Int. Workshop Frontiers Handwriting Recognition*, Buffalo, NY, May 1993, pp. 11–20.
[11] F. Kimura, Z. Chen, and M. Shridhar, "An integrated character recognition algorithm for locating and recognizing zip codes," in *Proc. U.S. Postal Service Advanced Technology Conf.*, Nov. 1990, pp. 605–619.
[12] D.-S. Lee and S. N. Srihari, "Handprinted digit recognition: a comparison of algorithms," in *Proc. 3rd Int. Workshop Frontiers Handwriting Recognition*, Buffalo, NY, May 1993, pp. 153–162.
[13] D.-S. Lee, "A theory of classifier combination: The neural network approach," Ph.D. dissertation, State Univ. New York, Buffalo, 1995.
[14] C. Y. Suen, C. Nadal, T. A. Mai, R. Legault, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proc. IEEE*, vol. 80, pp. 1162–1180, July 1992.
[15] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418–435, May/June 1992.
[16] M. Yoshimura, T. Shimizu, and I. Yoshimura, "A zip code recognition system using the localized arc pattern method," in *Proc. 2nd Int. Conf. Document Analysis and Recognition*, Tsukuba, Japan, Oct. 1993, pp. 183–186.

Louisa Lam received the B.A.. degree from Wellesley College, Wellesley, MA, and the Ph.D. degree in mathematics from the University of Toronto, Toronto, Ont., Canada.

She has been conducting research at the Centre for Pattern Recognition and Machine Intelligence at Concordia University, Montreal, P.Q., Canada, in the areas of handwriting recognition, thinning, and combinations of classifiers. She is currently an Adjunct Associate Professor with the Department of Computer Science, Concordia University, and Senior Lecturer of Mathematics and Computer Science at the Hong Kong Institute of Education, Hong Kong.

Dr. Lam is a member of Phi Beta Kappa.

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

In 1972, he joined the Department of Computer Science, Concordia University, Montreal, P.Q., Canada, where he became Professor in 1979 and served as Chairman from 1980 to 1984. Presently, he is the Director of CENPARMI, the Centre for Pattern Recognition and Machine Intelligence, Concordia University, and Associate Dean—Research, Faculty of Engineering and Computer Science. He has also been appointed to several visiting positions in institutions in other countries. He is the author/editor of 11 books on subjects ranging from computer vision and shape recognition, handwriting recognition and expert systems, to computational analysis of Mandarin and Chinese. He is also the author of more than 250 papers. He is an Associate Editor of several journals related to pattern recognition and artificial intelligence. He founded *Computer Processing of Chinese and Oriental Language*, an international journal of the Chinese Language Computer Society, in 1983, and served as Editor-in-Chief for ten years. His current interests include pattern recognition and machine intelligence, character recognition and expert systems, document processing and computational linguistics.

Dr. Suen is a member of several professional societies and a fellow of the IAPR and the Royal Society of Canada. He is the recipient of several awards, including the 1992 ITAC/NSERC award for outstanding contributions to pattern recognition, expert systems, and computational linguistics.