

# A Novel Confidence-Based Framework for Multiple Expert Decision Fusion

A. F. R. Rahman and M. C. Fairhurst  
Electronic Engineering Laboratories,  
University of Kent,  
Canterbury, Kent, CT2 7NT, UK  
[A.F.R.Rahman|M.C.Fairhurst]@ukc.ac.uk

## Abstract

A novel confidence-based parallel multiple expert decision combination framework is introduced. The traditional approaches to parallel multiple source decision fusion either take no account of the confidence of each decision by each participating expert in the combined framework or only exploit the confidences associated with each decision. But it is entirely possible to incorporate more additional *a priori* information in the form of various confidence indices that can be estimated from the performances of various participating experts. The confidence-based parallel multiple expert decision combination framework proposed here addresses this shortcoming. Very encouraging results have been obtained by implementing this proposed framework in combining decisions of multiple experts applied to the problem of handwritten and machine printed character recognition.

## 1 Introduction

In image classification, there are numerous ways to combine the decisions supplied in parallel by multiple experts. The various rules proposed for this depend on the amount of information that is available from the individual experts. Different types of information can be efficiently used in forming different approaches to different types of classifier combination as suggested by Xu *et al.*[23]. If only labels are available, then a majority voting scheme can be implemented[11, 9]. Efforts have also been made to calculate some type of label ranking in order to select the most likely decision[10, 3]. Some experts supply continuous values relating to the *a posteriori* probabilities. In these cases, several generalised combination schemes can be adopted which try to select the likeliest class by averaging or by performing linear combinations of these values[23]. All these strategies implicitly assume a parallel combination of multiple experts, where these experts may or may not be mutually independent.

## 2 Traditional Approaches to Parallel Decision Fusion: The Shortcomings

Traditional approaches to parallel decision combination have always suffered from one main shortcoming, since in most cases, these approaches depend upon the voting of the participating experts in some form or other. Some of the most commonly employed approaches are the Majority Voting Scheme[15], the Max Rule[12], the Median Rule[14] and the Min Rule[13]. In the Majority Voting Scheme, the votes cast by each expert are pooled and the class hypothesis voted by the majority is selected. In this case, there is no reflection of the confidences of each decision in the final combined decision, which often leads to the wrong classification (see, for example, [16]). The other rules mentioned above do take note of the confidences of each decision, but base the final decision on the relative values of the confidences attached to individual decisions. So ultimately, the decision combination is based on evaluating the confidences in recognising individual patterns, but this still does not include information about the strengths of individual experts in classifying various classes or any other special considerations.

It is entirely possible to evaluate other confidence indices extracted from the participating experts beside the confidence index associated with each decision and incorporate them in the decision combination process[19, 20]. Traditional approaches to decision combination fail to include these diverse sources of discriminating information while trying to arrive at a consensus. The parallel multiple expert framework proposed here addresses this deficiency and attempts to incorporate useful *a priori* information in the decision making process in designing a more robust decision combination scheme.

## 3 A Theoretical Framework for Confidence-based Decision Fusion

If there are  $n$  independent experts having the same probability  $p$  of being correct, then the probability of a consensus decision being correct, denoted by  $P_{Con}(n)$  can be computed using the binomial distribution as:

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

where  $k$  can be defined as:

$$k = \begin{cases} \frac{n}{2} + 1 & \text{If } n \text{ is even,} \\ \frac{n+1}{2} & \text{if } n \text{ is odd.} \end{cases} \quad (2)$$

This derivation is attributed to Condorcet[4], who first emphasised the fact that decisions of a group of experts are superior to that of the individual experts, provided the individuals have reasonable competence. Detailed investigations have been carried out to analyse this majority voting scenario[18, 8], but in general the problem is simplified by assuming that the number of experts is odd, each expert has the same probability of selecting the correct candidate and that individual decisions are independent. Assuming  $P_{Concorrect}$  and  $P_{Conincorrect}$  are the probability that the consensus is correct and wrong

respectively, it is straightforward to state that  $P_{Concorrect} + P_{Conincorrect} = 1$ . Unfortunately, this framework does not take into account of the scenario when a reject option is allowed, which is a very important consideration when this scheme is applied to cases where the number of experts is small, since now  $P_{Concorrect} + P_{Conincorrect} \leq 1$ .

This analysis demonstrates the scenario of combining multiple experts when the probability of success of the various experts participating in a decision combination scheme are identical. In practice, however, the scenario is far from ideal. The participating experts have their own performance indices, creating a scenario where theoretical modelling becomes very difficult. The traditional approach to decision combination attempts to address this problem by assigning due emphasis to the probability of success (confidence of a classification) of each expert for each pattern (sample). Although this approach rectifies the problem to some extent, it falls far short of a complete solution.

The proposed parallel classifier combination scheme advocates an approach based on incorporating more confidence indices (probability measures), providing more information about the diverse performance criteria of the participating experts than is made by traditional approaches. It is obvious that each expert cooperating in the framework needs to be able to generate an absolute class index from the sample pattern passed on to it. There are three additional sources of information that any decision is associated with, as follows:

- The sample Confidence Index ( $\alpha$ ): This is associated with the pattern currently being evaluated, and demonstrates the confidence of an expert in assigning a particular class label to a particular sample.
- The Class Confidence Index ( $\beta$ ): This is associated with a class for a particular expert, and denotes the *a priori* confidence of an expert in correctly classifying samples coming from a particular class, and
- The overall confidence Index ( $\gamma$ ): This is associated with a particular expert, and denotes the overall *a priori* confidence of an expert in classifying all the samples from all the classes for a particular task domain.

Obviously the extent to which these additional sources of information are utilised depends on the experts employed at that particular point of the decision hierarchy. An information packet can be defined as  $IP(\alpha_{ijk}, \beta_{ij}, \gamma_i)$ , where  $\alpha_{ijk}$  is the sample confidence index assigned by the  $i^{th}$  expert to the  $k^{th}$  sample coming from the  $j^{th}$  class,  $\beta_{ij}$  is the confidence of the  $i^{th}$  expert in correctly classifying the  $j^{th}$  class and  $\gamma_i$  is the overall confidence index of the  $i^{th}$  expert. This information packet  $IP(\alpha_{ijk}, \beta_{ij}, \gamma_i)$  and the decision process  $d(X, IP_i)$  employed by the  $i^{th}$  expert, where  $X$  is the current pattern, characterise the performance at that particular point. Assuming:

$\omega(X)$	the original class associated with the current pattern $X$ ,
$P_{ei}$	the probability that $\omega(X)$ does not contain the true class, i.e. $P_{ei} = P[\omega(X) \neq X]$ ,
$P_{ci}$	the probability that $\omega(X)$ contains the true class, i.e. $P_{ci} = P[\omega(X) = X]$ ,

then the overall error probability  $P_{eT}$  of the combined parallel network depends on the following criteria:

- When the class index associated with highest confidence is wrong, provided the next nearest confidence index is lower than the highest index by more than the threshold value ( $\psi_c$ ). The error probability  $P_e(a)$  in this case can be expressed as:

$$P_e(a) = P[(d(X, P_i) \neq X) / (\alpha_{max} - \alpha_{max-1} \leq \psi_c)] \quad (3)$$

- When more than one sample confidence index is within the range of the threshold value, and the corresponding class index having the highest value is wrong, provided the next nearest sample confidence index is lower than the highest class confidence index by more than the threshold value ( $\theta_c$ ). The error probability  $P_e(b)$  in this case can be expressed as:

$$P_e(b) = P[(d(X, P_i) \neq X) / ((\alpha_{max} - \alpha_{max-1} \leq \psi_c)(\beta_{max} - \beta_{max-1} \leq \theta_c))] \quad (4)$$

- When more than one sample confidence index is within the range of the threshold value, more than one class confidence index is within the range of the threshold value, and the corresponding overall confidence index having the highest value is wrong, provided the next nearest overall confidence index is lower than the highest overall confidence index by more than the threshold value ( $\theta_c$ ). The error probability  $P_e(c)$  in this case can be expressed as:

$$P_e(c) = P[(d(X, P_i) \neq X) / ((\alpha_{max} - \alpha_{max-1} \leq \psi_c)(\beta_{max} - \beta_{max-1} \leq \theta_c)(\gamma_{max} - \gamma_{max-1} \leq \phi_c))] \quad (5)$$

The overall error probability of the proposed parallel combination, therefore, can be expressed as:

$$P_{eT} = P_e(a) + P_e(b) + P_e(c) \quad (6)$$

Accordingly the overall correct classification probability  $P_{cT}$  of the  $n$ -expert parallel network is:

$$P_{cT} = 1 - P_{eT} \quad (7)$$

The overall behaviour of any arbitrary expert participating in the proposed parallel hierarchy has been discussed so far, but nothing has been mentioned about the rejection capacity of the system. When individual experts are allowed to reject a particular number of patterns, this has an effect on the overall performance of the system. In general, as a rejection capacity is incorporated, the error rate falls at the cost of rejections, but as a result, the absolute recognition rate also falls. Although the purity of the recognition enhances, on the whole fewer patterns are correctly classified. For the proposed parallel combination approach, rejection by individual experts is an integral part of the system design. It is interesting to analyse the behaviour of the configuration if all the experts have rejection capability. Assuming:

$P_{ri}$  the probability that the  $i^{th}$  expert rejects the current pattern  $X$ ,

the overall reject probability  $P_{rT}$  of the combined configuration can be expressed as:

$$P_{rT} = P_{r1} + P_{r2} + \dots + P_{rn} \quad (8)$$

But on the other hand, the overall error probability of the configuration is now affected by the rejection capability of the configuration. The overall error probability of the configuration can now be expressed as:

$$P_{eT} = (P_{e1} - P_{r1}) + (P_{e2} - P_{r2}) + \dots + (P_{en} - P_{rn}) \quad (9)$$

So this proposed framework takes the consensus decision based on a hierarchy of confidence comparisons. Consensus about the decision is primarily based on comparing sample confidences, failing which the secondary comparison is based on comparing class confidences and finally, failing everything else, the final decision is based on overall confidence comparison. A sample is rejected if the framework fails to find evidence to support any of the class hypothesis in terms of the three confidence criteria.

## 4 Implementation and Results

In the proposed combination approach, the individual top choices of the experts are of utmost importance, since the decision combination algorithm puts due emphasis on all the top choices selected by the individual experts. Although extensive utilisation of *a priori* information is incorporated in the hierarchical decision making process, the confidence of individual experts in identifying top choices is very important (see, for example, [6]).

It is also very important to clarify how the various confidence indices are evaluated. Once trained, their performances are evaluated on a defined evaluating data set. As a completely independent data set is required to evaluate the performance of individual experts before an attempt is made to combine them, the test sets of the various databases are split into two equal parts. One part is used as an evaluating dataset and the other part is used as the test set. The confidence values associated with the behaviour of the experts, the class confidence values and the overall confidence values are calculated at this stage. The parallel combination of decisions can then take these classifiers and combine their decisions in a way that ensures that there is an enhanced recognition rate and a low error rate at the cost of rejecting a certain percentage of the input patterns. The sample confidence values associated with the individual samples are estimated when the actual recognition takes place during the combined recognition phase.

### 4.1 Selection of Individual Experts

To compare the performances of different multiple expert configurations, it is important to have a group of primary experts which have comparable inter-expert performance indices, but which, at the same time, use different types of features and classification criteria. The following primary experts were chosen for this purpose.

- *Binary Weighted Scheme(BWS)*: This employs a technique based on  $n$ -tuple sampling or memory network processing[7]. The image array is divided into a certain number of samples, each consisting of a fixed number of pixels. Each of these samples is connected to a memory element, which in turn computes a single valued Boolean function.
- *Frequency Weighted Scheme(FWS)*: This is similar to the *BWS*, but in this case the memory elements calculate the relative frequencies of the sampled features, thereby indicating the probability distribution of the group of points or  $n$ -tuples[5].
- *Multi-layer Perceptron Network(MLP)*: This is the standard multilayer perceptron neural net structure, employing the standard error backpropagation algorithm[22].
- *Moment-based Pattern Classifiers(MPC)*: These statistical algorithms make use of the  $n^{\text{th}}$  order mathematical moments derived from the binarised patterns. Different discriminating functions may be used to identify possible cluster formation[21].

## 4.2 Selecting Databases

Three databases have been chosen in all the experiments and simulations discussed here. Two of these databases contain handwritten characters and the third consists of machine printed characters. The first handwritten database is one compiled at the University of Essex, UK (Database A)[17]. The second database (Database B) is compiled by the U.S. National Institute of Standards and Technology, and is popularly known as the NIST Database[1]. The third is compiled at the University of Kent at Canterbury, UK (Database C)[2]. This is a compilation of printed characters and was extracted from machine printed post-codes supplied by the British Post Office. All these databases contain samples of alpha-numeric characters (the numerals 0 to 9 and upper case letters A to Z, with no distinction made between the characters '1/I' and '0/O').

## 4.3 Performance of Individual Experts

Table 1 presents the performance of the various individual experts on the different databases. Results in the cases of application of these experts on the digit classes and the digit plus upper case letter classes have been presented. The direct implication of incorporating a rejection capability is that the confidence associated with the individual decisions are more robust. But the price to pay for this added confidence in the top choice decisions delivered by the individual experts is the reduction in the absolute recognition rate. This is consistent with the scenario presented by the current parallel combination approach, since the configuration visualised in this case incorporates very powerful rejection recovery approaches. This ability of the parallel configuration in utilising the rejections by various individual experts cooperating in a unified framework indicates strong support for the decision to incorporate the rejection capability in the individual experts.

## 4.4 Performance of Decision Combination

Since there are four experts in the individual expert pool, it is actually possible to combine all four experts in parallel. Table 2 illustrates the classification performance when

Expert	Database	Classes	Accepted	Recognised	Error	Rejected
FWS	C	Digits	98.43	97.34	1.09	1.57
		Digits + Upper Case	98.41	96.52	1.89	1.59
	A	Digits	96.75	90.79	5.96	3.25
		Digits + Upper Case	97.30	78.86	18.44	2.70
	B	Digits	97.35	78.76	18.59	2.65
		Digits + Upper Case	94.44	71.88	22.56	5.56
MPC	C	Digits	98.23	95.78	2.45	1.77
		Digits + Upper Case	97.79	93.61	4.18	2.21
	A	Digits	97.20	91.63	5.57	2.80
		Digits + Upper Case	98.88	79.96	18.92	1.12
	B	Digits	97.62	85.78	11.84	2.38
		Digits + Upper Case	95.43	72.58	22.85	4.57
BWS	C	Digits	98.93	97.87	1.06	1.07
		Digits + Upper Case	97.81	94.66	3.15	2.19
	A	Digits	95.75	86.47	9.28	4.25
		Digits + Upper Case	95.00	71.22	23.78	5.00
	B	Digits	95.50	72.31	23.19	4.50
		Digits + Upper Case	96.80	65.89	30.91	3.20
MLP	C	Digits	98.54	97.56	0.98	1.46
		Digits + Upper Case	98.58	95.71	2.87	1.42
	A	Digits	96.89	90.72	6.17	3.11
		Digits + Upper Case	97.20	80.39	16.81	2.80
	B	Digits	95.13	82.31	12.82	4.87
		Digits + Upper Case	97.20	71.27	25.93	2.80

Table 1: Performance of individual classifiers

Average Recognition Performance						
Database	Digits					
	Recognition		Error		Rejection	
	Traditional	Proposed	Traditional	Proposed	Traditional	Proposed
C	98.71	98.89	1.29	0.98	0.00	0.13
A	93.42	94.21	6.58	5.02	0.00	0.77
B	88.84	90.57	11.16	6.98	0.00	2.45

Table 2: Performance of combination: Digit classes

combining four experts in a unified parallel framework. Results on the digit classes for various databases have been reported. It is clearly seen from this Table that the combination of four experts has resulted in lower overall rejection rate, lower overall error rate and consequently higher absolute recognition rate. This is due to the fact that more information regarding the class membership and recognition confidence along with much more *a priori* knowledge about the behaviour of the cooperating experts are incorporated in the decision making process. Since there are alternatives in the decision making process, the overall rejection therefore also falls. This has resulted in a shift of the error rate towards the rejection rate and, as a direct result, overall recognition rate has increased. Table 3 shows the performance achieved when the same combinations of four experts are applied to the recognition of the complete alpha-numeric datasets from the three different databases. Similar observations to those derived from Table 2 can also be derived from Table 3. Application of four experts in the parallel combination framework has resulted in lower error rates, lower rejection rates and consequently higher absolute recognition

Average Recognition Performance						
Database	Digits + Upper Case Classes					
	Recognition		Error		Rejection	
	Traditional	Proposed	Traditional	Proposed	Traditional	Proposed
C	97.75	97.75	2.25	1.61	0.00	0.64
A	82.34	84.34	17.66	13.03	0.00	2.63
B	78.29	78.38	21.71	17.92	0.00	3.89

Table 3: Performance of combination: Digit plus upper case classes

rates. The important point about these observations is that these trends are not random, rather they are consistently repeated over the various datasets.

For comparative purposes, the results of decision combination on a traditional top-choice majority voting scheme have also been presented in Tables 2 and 3. It is clearly seen from these results that the proposed decision combination approach can outperform the majority voting scheme in all the cases except one, in which case the results are identical. This demonstrates the strength of the proposed framework in combining decisions from multiple sources.

## 5 Conclusion

A novel confidence-based parallel multiple expert decision fusion framework has been proposed. It has been demonstrated that such a framework takes into account additional information about the relative strengths of the participating experts while arriving at the combined decision, thereby making the overall decision more robust. The framework has been tested by implementing various parallel multiple expert decision combination configurations for recognising handwritten and machine printed characters and considerable performance enhancement over the traditional majority voting scheme has been achieved.

## Acknowledgement

This work was supported by the UK Engineering and Physical Sciences Research Council.

## References

- [1] NIST Special Databases 1-3, 6-8, 19, 20, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA.
- [2] Image Processing & Computer Vision, Electronic Engineering, University of Kent, Canterbury, CT2 7NT, UK.
- [3] S. C. Bagui and N. R. pal. A multistage generalization of the rank nearest neighbor classification rule. *Pattern Recognition Letters*, 16(6):601–614, 1995.
- [4] N. C. de Condorcet. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Imprimerie Royale, Paris, France, 1785.



- [5] M. C. Fairhurst and M. A. G. Mattaso Maia. Performance comparison in hierarchical architectures for memory network pattern classifiers. *Pattern Recognition Letters*, 4(2):121–124, 1986.
- [6] M. C. Fairhurst and A. F. R. Rahman. A generalised approach to the recognition of structurally similar handwritten characters. *IEE Proc. on Vision, Image and Signal Processing*, 144(1):15–22, 1997.
- [7] M. C. Fairhurst and T. J. Stonham. A classification system for alphanumeric characters based on learning network techniques. *Digital Processes*, 2:321–339, 1976.
- [8] R. Farquharson. *Theory of Voting*. Yale University, USA, 1969.
- [9] J. Frank and E. Mandler. A comparison of two approaches for combining the votes of cooperating classifiers. In *Proc. 11th ICPR*, pages 611–614, The Hague, Netherlands, 1992.
- [10] T. K. Ho, J. J. Hull, and S. N. Srihari. On multiple classifier systems for pattern recognition. In *Proc. 11th ICPR*, pages 84–87, The Hague, Netherlands, 1992.
- [11] F. Kimura and M. Shridhar. Handwritten numeral recognition based on multiple algorithms. *Pattern Recognition*, 24(10):969–983, 1991.
- [12] J. Kittler and M. Hatef. Improving recognition rates by classifier combination. In *Proc. 5th Int. Workshop on Frontiers of Handwriting Recognition*, pages 81–102, 1996.
- [13] J. Kittler, M. Hatef, and R. P. W. Duin. Combining classifiers. In *Proc. 13th ICPR, Vienna, Austria*, pages 897–901, 1996.
- [14] J. Kittler, A. Hojjatoleslami, and T. Windeatt. Weighting factors in multiple expert fusion. In *Proc. British Machine Vision Conference*, pages 41–50, 1997.
- [15] L. Lam and C. Y. Suen. A theoretical-analysis of the application of majority voting to pattern-recognition. In *Proc. 12th IAPR Int. Conf. on Pattern Recognition, Conf. B: Pattern Recognition and Neural Networks*, volume 2, pages 418–420, Jerusalem, Israel, 1994.
- [16] L. Lam and C. Y. Suen. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):553–568, 1997.
- [17] S. Lucas and A. Amiri. Statistical syntactic methods for high-performance OCR. *IEE Proc. on Vision, Image and Signal Processing*, 143(1):23–30, 1996.
- [18] P. K. Pattanaik. *Voting and Collective Choice*. Cambridge University Press, London, UK, 1971.
- [19] A. F. R. Rahman and M. C. Fairhurst. An evaluation of multi-expert configurations for recognition of handwritten numerals. *Pattern Recognition*, 31(9):1255–1273, 1998.
- [20] A. F. R. Rahman and M. C. Fairhurst. Machine-printed character recognition revisited: Re-application of recent advances in handwritten character recognition research. *Special Issue on Document Image Processing and Multimedia Environments, Image Vision & Computing*, 1998. Accepted for publication, in press.
- [21] T. H. Reiss. *Recognizing planer objects using invariant image features*. Springer-Verlag, Berlin Heidelberg, Germany, 1993.
- [22] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, in *Parallel Distributed Processing*, volume 1, pages 318–362. MIT Press, Cambridge, MA, 1986. D. E. Rumelhart and J. L. McClelland(Eds.).
- [23] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. on Systems, Man and Cybernetics*, 23(3):418–435, 1992.