Pattern Analysis & Applications (1999)2:205–214

Reliability Parameters to Improve Combination Strategies in Multi-Expert Systems

L. P. Cordella[1], P. Foggia[1], C. Sansone[1], F. Tortorella[2] and M. Vento[1]

[1]*Dipartimento di Informatica e Sistemistica, Università degli Studi di Napoli 'Federico II', Napoli, Italy;*
[2]*Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e Matematica Industriale, Università degli Studi di Cassino, Cassino, Italy*

Abstract: Recognition systems based on a combination of different experts have been widely investigated in the recent past. General criteria for improving the performance of such systems are based on estimating the reliability associated with the decision of each expert, so as to suitably weight its response in the combination phase. According to the methods proposed to-date, when the expert assigns a sample to a class, the reliability of such a decision is estimated on the basis of the recognition rate obtained by the expert on the chosen class during the training phase. As a consequence, the same reliability value is associated with every decision attributing a sample to a same class, even though it seems reasonable to take into account its dependence on the quality of the specific sample. We propose a method for estimating the reliability of each single recognition act of an expert on the basis of information directly derived from its output. In this way, the reliability value of a decision is more properly estimated, thus allowing a more precise weighting during the combination phase. The definition of the reliability parameters for widely used classification paradigms is discussed, together with the combining rules employing them for weighting the expert opinions. The results obtained by combining four experts in order to recognise handwritten numerals from a standard character database are presented. Comparison with classical combining rules is also reported, and the advantages of the proposed approach outlined.

Keywords: Classification reliability; Combining rules; Multi-expert systems; Neural nets; OCR; Statistical classifiers

1. INTRODUCTION

In many areas of pattern recognition, especially in the case of applications characterised by large data variability and significant amounts of noise, the performance of a given classification system is sometimes unsatisfactory for the needs of real applications. In these cases, efforts are generally devoted to the selection of description methods and classification algorithms which can keep performance high, even in the presence of highly distorted samples.

To-date, a large variety of algorithms, providing a sample description in terms of a vector of numerical features [1] or in terms of structural primitives [2,3], have been proposed, together with classification paradigms suitable both for statistical approaches, such as the k-NN [4], the Bayesian [1], and the neural [5] methods, and for structural and syntactic approaches, such as graph-matching algorithms [3] and grammar parsing techniques [2].

However, practical experience suggests that in many applications, single recognition systems (experts), albeit very refined, fail to achieve an acceptable performance level. It happens that a recognition system, on the basis of some assumptions about the world of patterns to recognise, adopts description and classification models which, as a whole, are not equally adequate for all the patterns.

The idea of combining various experts with the aim of compensating for the weakness of each single expert while preserving its own strength, has recently been investigated widely [6,7]. The rationale lies in the assumption that, by suitably combining the results of a set of experts according to a rule (combining rule), the performance obtained can be better than that of any single expert. The successful implementation of a Multi-Expert System (MES) implies the use of the most complementary experts possible, and

Received: 3 August 1997
Received in revised form: 24 November 1998
Accepted: 11 December 1998

the definition of a combining rule for determining the most likely class a sample should be attributed to, given the class to which it is attributed by each single expert.

Preliminary experimental results encouraged the approach, and various research groups concentrated on different aspects of the problem [7–9]. Investigations aimed at determining the complementary nature of experts to be used and their optimal number, as well as the optimal combination topology, are reported in the literature [10–13]. Most of the research has been devoted to finding different combining rules able to solve the conflicts, i.e. to determine the most likely class on the basis of the responses of the experts in the case of discordance [14–17]. A quite different approach, specifically developed in the field of neural computation, is to combine different experts each of which is defined over a local region of the input space. Jacobs et al [18] introduced such a strategy with their 'mixture of experts' architecture. It involves a set of function approximators ('expert networks') that are combined by a classifier ('gating network'). These networks are trained simultaneously so as to split the input space into regions where particular experts can specialise. Jordan and Jacobs [19,20] extended this approach to a recursively-defined architecture (the so-called 'hierarchical mixture of experts'), in which a tree of gating networks combines the expert networks into successively larger groupings that are defined over nested regions of the input space. The learning process of the experts is thus carried out in a coupled manner, and the combining scheme is a competitive one, as only one expert at a time can be selected for each input sample.

In every case, the mixture of experts approach does not make it possible to exploit the different information coming from different descriptions of the same sample, as all the experts work on the same input space, i.e. on the same description.

In the following, we focus our attention on a MES made up of experts which generally employ different descriptions of the input sample; in this case, the main problem is how to solve the conflicts arising among the different responses of the experts on the basis of a suitably defined combining rule. One of the simplest combining rules, the 'Majority Voting' rule [11,21], assigns the input sample to the class for which a relative or absolute majority of experts agrees; otherwise, the sample is rejected. More sophisticated criteria for resolving conflicts among the experts require the introduction of a measure of the reliability associated with the response of each expert. For instance, the 'Weighted Voting' methods [8,22] are mainly based on this idea: the votes of all the experts are collected, and the input sample is assigned to the class for which the sum of the votes, each weighted by the estimated reliability of the corresponding expert, is the highest. In a similar way, the reliability could be effectively used for solving the ties occurring in the 'Majority Voting' rule. If two or more classes receive the same number of votes, the tie may be solved by considering the reliability of the experts.

A well-known approach for defining a reliability parameter for a given expert is based on the evaluation of its recognition performance on the training set [14]. For example, if an expert assigns the input sample to the *i*th class, the decision is attributed a reliability proportional to the recognition rate of the training set samples attributed to the *i*th class. A drawback of such a definition is that every decision attributing a sample to a same class is assigned the same reliability regardless of the quality of the sample, and thus of the reliability of the specific decision. The average performance of an expert on the training set, for each given class, is undoubtedly significant for a sample whose representation is similar to that of the majority of samples of that class, but this value does not necessarily reflect the actual reliability of each single classification act.

To overcome this problem, we propose a method to evaluate the reliability of each classification act performed by a given expert, and to use this value to weight its vote in a MES. The definition of the parameter measuring the classification reliability is made on the basis of information directly derived from the output of the expert. Namely, a correspondence can be established between the state of the expert's output and the situations in the feature space which can give rise to unreliable classifications. The operating definition of the parameter which allows us to detect such situations and to quantify classification reliability will depend upon the classifier architecture considered.

To evaluate the effectiveness of the approach, several multi-expert systems have been considered. Each system was obtained by combining various experts according to different combining rules. The experts are handwritten character recognisers made of different descriptor-classifier pairs. In fact, the performance of a recognition system depends not only upon its classification section, but also upon the quality of the descriptions of the samples to be recognised, which are computed by the description section the classifier is provided with.

Tests have been carried out using the digits of the NIST Database 19 [23]. Results obtained with classical combining rules are also reported, and the advantages of our approach outlined.

In Section 2 some widely used criteria for combining experts are briefly reviewed, while in Section 3 our approach is presented: the proposed reliability parameters are defined and their use for improving the performance of some well-known combining rules is discussed. Finally, in Section 4 experimental results obtained by using different multi-experts for handwritten character recognition are presented.

2. COMBINING CRITERIA

As mentioned in the introduction, many ways to combine expert decisions have been proposed in the recent past. Some of them are based on heuristic approaches, such as voting or ranking strategies, while others are based on probability theory, e.g. the Bayesian method [7].

From a theoretical point of view, given a set of experts, the performance of the combining scheme should improve with the amount of information provided by each single expert. In the literature, the various classification algorithms are divided into three types, depending on the output information they are able to supply [14]. *Type 1* classifiers output a unique label, i.e. the label of the presumed class; they are also known as classifiers that work at an *abstract* level. *Type 2* classifiers, which work at a *rank* level, rank all the classes in a queue where the class at the top is the first choice, while *Type 3* classifiers, which work at a *measurement* level, attribute each class a measurement value to represent the degree that the input sample belongs to that class.

Almost all the combining rules are defined with reference to Type 1 classifiers. On the other hand, combining schemes that exploit information from the classifiers at the measurement level allow us to define combining rules that are more sophisticated and potentially more effective.

The most common combining rules are the ones based on a voting strategy, where each expert gives its own opinion (i.e. a vote) about the class of the input pattern. The Relative Majority Voting Rule (MV rule) decides that the input pattern belongs to the ith class C_i if and only if the relative majority of the classifiers votes for the class C_i. If two or more classes obtain the same number of votes, the input sample is rejected. More formally, let us denote by $Y_k(x)$ the output of the kth classifier when the sample x is submitted to it (Y_k will be used when the dependence on the input sample is evident from the context). If the vote of the kth classifier for the class C_i is denoted by

$$V_k^i = \begin{cases} 1 \text{ if } Y_k = i \\ 0 \text{ if } Y_k \neq i \end{cases}$$

the total number of votes received by C_i will be given by

$$V^i = \sum_k V_k^i$$

If $T = \{i | V^i = \max_j V^j\}$ is the set made up of the classes with the maximum number of votes, the output of an MES using the MV rule is

$$Y = \begin{cases} i = \arg\max_j V^j & \text{if card}(T) = 1 \\ 0 \text{ (i.e. reject)} & \text{otherwise} \end{cases} \quad (1)$$

A commonly used version of the MV rule virtually eliminates the reject by using a suitably defined reliability parameter (often referred to as the *confidence degree*) to weight the expert votes when two or more classes receive the maximum number of votes. If we denote as $D_k(x)$ (D_k for short) the value of such a reliability parameter for the kth classifier, the sum of the weighted votes for class C_i is given by

$$W^i = \sum_k D_k V_k^i \quad (2)$$

Thus, the MES output becomes

$$Y = \begin{cases} i = \arg\max_j V^j & \text{if card}(T) = 1 \\ i = \arg\max_{j \in T} W^j & \text{otherwise} \end{cases} \quad (3)$$

An alternative is to use the parameter D_k not in the presence of ties only, but in all cases for weighting the decision of each expert involved in the combination. The output provided by an MES using this rule, known as Weighted Voting Rule (WV rule), is

$$Y = \arg\max_j W^j \quad (4)$$

To evaluate D_k, i.e. the reliability of the vote given by the kth expert, the most common choice is to use the confusion matrix E^k [14]. The generic element $e_{i,j}^k$ ($1 \le i,j \le n$, where n is the number of the classes) of E^k represents the percentage of samples of the class C_i assigned to the class C_j. A reasonable definition of D_k based on E^k is

$$D_k = \frac{e_{i,i}^k}{\sum_j e_{j,i}^k} \quad (5)$$

given $Y_k = i$. In fact, if N_i is the total number of training set samples belonging to C_i, then $N_i e_{i,i}^k$ is the number of samples of C_i which have been correctly classified by the kth classifier, and $\Sigma_j N_j e_{j,i}^k$ is the total number of samples of any class assigned to C_i. It follows that $N_i e_{i,i}^k / \Sigma_j N_j e_{j,i}^k$ is an estimate of the probability that a sample has been correctly classified if it has been assigned to C_i. Thus, under the assumption that $N_i = N_j \; \forall i,j$, Eq. (5) expresses the *a posteriori* probability that the kth classifier gives the correct answer.

It is worth noting that for an MES made of two experts, Eqs (3) and (4), providing the output of the MES in the case where the MV and the WV rule are respectively adopted, can be reduced to the same form, thus implying that the results obtained in the two cases are the same. In the general case, when the number M of experts is greater than two, it can be simply shown that a necessary condition for the two rules to give different results is

$$\begin{cases} \min_k D_k < \dfrac{M-2}{M+2} & \text{for M even} \\ \min_k D_k < \dfrac{M-1}{M+1} & \text{for M odd} \end{cases} \quad (6)$$

A further alternative is given by the Bayesian Combining Rule (BC rule), which uses all the information present in E^k for estimating the *a posteriori* probability that the input sample belongs to the ith class. In particular, if we indicate with ($Y_k = j_k$) the event that the kth expert assigns the input sample x to one of the classes (let us denote its index by j_k), the class C_i selected by the BC rule is the one which maximises the post-probability:

$$Y = \operatorname{argmax}_i P(x \in C_i | Y_1 = j_1, Y_2, = j_2 \ldots, Y_M = j_M) \quad (7)$$

If the experts can be assumed to be independent of each other and the *a priori* probability is the same for all the classes, it can be shown that Eq. (7) can be written as

$$Y = \operatorname{argmax}_i \prod_{k=1}^{M} P(x \in C_i | Y_k = j_k) \quad (8)$$

According to Eq. (5), an estimation of the post-probability for the kth expert is given by

$$P(x \in C_i|Y_k = j_k) = e^k_{i,j_k}/\sum_{h=1}^{n} e^k_{h,j_k} \tag{9}$$

thus, the class that maximises the post-probability (7) satisfies the following relation:

$$Y = \underset{i}{\operatorname{argmax}} \prod_{k=1}^{M} e^k_{i,j_k} \tag{10}$$

which can be adopted for implementing the BC rule.

3. RELIABILITY PARAMETERS

The combining approaches described in the previous section estimate the reliability of the classification by looking at the global performance of the expert on the training set. As pointed out in the introduction, more effective solutions could be obtained by introducing a reliability parameter which can estimate the accuracy of each single classification act of an expert.

To better illustrate this point, let us refer to the feature space representation of a given set of samples. Classification reliability evaluation requires that the situations in the feature space which can give rise to unreliable classifications are characterised and correlated to the state of the classifier output. The low reliability of a classification can be traced back to one of the following situations (see Fig. 1): (a) the sample considered is significantly different from those present in the training set, i.e. its representative point is located in a region of the feature space which is far from those associated with the various classes; (b) the point which represents the sample considered in the feature space lies where the regions pertaining to two or more classes overlap, i.e. where training set samples belonging to more than one class are present.

To distinguish between classifications which are unreliable because a sample is of type *a* or *b*, let us define two reliability parameters, ψ_a and ψ_b, whose values vary in the interval [0,1]. It is assumed that parameter values near to 1 characterise very reliable classifications, while low values correspond to unreliable classifications.

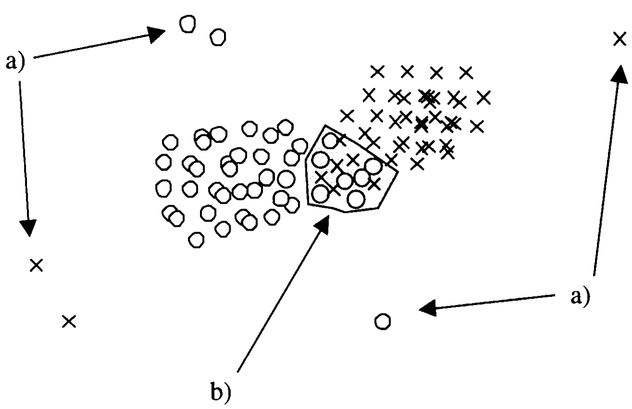

Fig. 1. Two cases of unreliable classification. (a) Samples significantly different from those of the training set; (b) samples lying in a confusion region of the feature space.

The two parameters are associated with each expert, and each parameter is a function of the expert output vector (indeed, of the output of its classification section). A parameter ψ providing an inclusive measure of the reliability of a classification can be computed by combining the values of ψ_a and ψ_b. The form chosen for ψ is:

$$\psi = \min\{\psi_a, \psi_b\} \tag{11}$$

This is certainly a conservative choice because it implies that, for a classification to be considered unreliable, only one reliability parameter needs to assume a low value, regardless of the value assumed by the other one. Of course, other choices could be made (e.g. see the survey on combining operators presented by Bloch [24]), depending on the requirements of the application at hand. However, this matter lies beyond the scope of the present paper, and the implications of alternative choices will not be considered here. In Fig. 2 a general scheme describing our approach is presented.

To exploit the reliability information defined above, we have used the Majority Voting and the Weighted Voting rules, choosing $D_k = \psi_k$ (classification reliability of the *k*th classifier) in the definition of W_i (see Eq. (2)). In this way, we have two new combination strategies: the Reliability Based Majority Voting (*RBMV*) and the Reliability Based Weighted Voting (*RBWV*).

The operating definition of ψ_a and ψ_b depends upon the particular classifier architecture, and requires the classifier to work at the measurement level; thus, it is applicable to Type 3 classifiers. Note that effectively exploiting the information held by the output vector of such classifiers is not a trivial task. In the following, the reliability parameters ψ_a and ψ_b for the Multi-Layer Perceptron (MLP) [25] and the Nearest Neighbour (NN) classifier [26] will be defined. These classifiers have been chosen for our testing purposes because

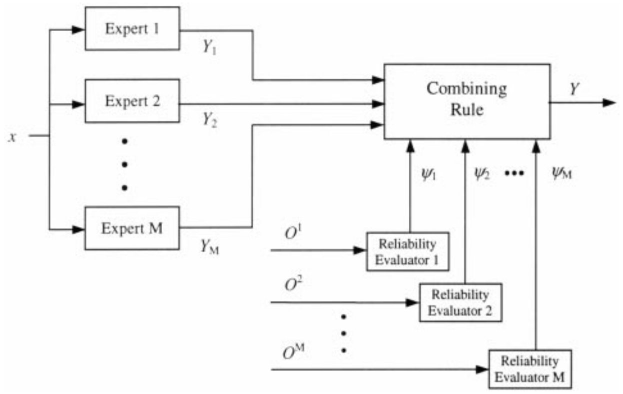

Fig. 2. A general scheme describing our approach. To obtain the decision Y of the MES, the decisions $Y_1, Y_2, \ldots, Y_M$ of the component experts are combined according to a rule which takes into account the reliability parameters $\psi_1, \psi_2, \ldots, \psi_M$ evaluated on the basis of the expert output vectors $O^1, O^2, \ldots, O^M$. In this way, a different value of the reliability can be associated to each classification act of an expert.

they are among the most commonly used. The definition of the reliability in case of a wider set of neural classifiers can be found in Cordella et al [27].

The output layer of the MLP classifier is made of the same number of neurons as classes. From a theoretical point of view, it is expected that if, during training, a sample belonging to the *k*th class is presented to the network input, the *k*th output neuron will assume a value equal to 1, while all the other outputs will assume a value equal to 0 (ideal output vector). In practice, the status of the output vector is generally different from the ideal one (i.e. the values of its elements may be numbers in the interval [0,1]), and the input sample is attributed to a class according to a given rule. The simplest rule is Winner-Takes-All, according to which the input sample is attributed to the class whose output neuron has the highest value.

This network uses the value of the connection weights to obtain the hyperplanes defining the decision regions in the feature space [28]. During the training phase, the network dynamically modifies the decision region boundaries in such a way as to provide, for each sample, an output vector as close as possible to the ideal one. Consequently, test set samples very different from those which contributed to determine the hyperplanes separating the decision regions (see case *a* in Fig. 1) may fall outside every region: in this case, all the output neurons will provide values near to 0. Therefore, an effective definition of the reliability parameter ψ_a can be the following:

$$\psi_a = O_{\text{win}} \tag{12}$$

where O_{win} is the output of the winner neuron. In this way, the nearer to 0 the value of O_{win}, the less reliable the classification is considered.

Samples of type *b*, lying where two or more decision regions overlap, typically generate output vectors with two or more elements having similar values. In these cases, the risk of a classification error is significantly high. Thus, for a given O_{win}, the higher the difference between O_{win} and $O_{2\text{win}}$ (the output of the neuron having the highest value after the winner), the safer the decision of attributing the sample to the winning class. Consequently, a suitable parameter for evaluating the reliability of these situations can be

$$\psi_b = O_{\text{win}} - O_{2\text{win}} \tag{13}$$

Let us note that, since the values of the elements of the output vector are real numbers in the interval [0,1], the reliability parameters ψ_a and ψ_b also assume values in the same interval, as required by their definition.

In conclusion, the classification reliability of the MLP classifier can be measured by

$$\begin{aligned}\psi &= \min\{\psi_a,\psi_b\} = \min\{O_{\text{win}}, O_{\text{win}} - O_{2\text{win}}\} \\ &= O_{\text{win}} - O_{2\text{win}} = \psi_b\end{aligned} \tag{14}$$

The Nearest Neighbour classifier assigns the input sample *x* to the class including the training set sample with the smallest distance from *x*. Note that the training set in the case of an NN classifier is usually referred to as the *reference set*, since there is no explicit training phase of the NN classifier.

Let us indicate as O_i the minimum distance between *x* and the set of the reference samples belonging to class C_i. Thus, the smallest distance of *x* from a reference samples, say O_{win}, is:

$$O_{\text{win}} = \min_i O_i = \min_i \left(\min_{r_{ij}\in C_i} d(r_{ij},x)\right) \tag{15}$$

where r_{ij} is the *j*th reference sample of the *i*th class.

Obviously, test set samples that are significantly different from those present in the reference set will have a significant distance from all the samples of the latter set. Therefore, the reliability parameter ψ_a can be defined as

$$\psi_a = 1 - \frac{O_{\text{win}}}{O_{\max}} \tag{16}$$

where the value of $O_{\max}$ is the highest among the values of O_{win} obtained for samples taken from a set (*training-test set*) disjoint from both the reference set and the test set.

As it is to be expected that the value of O_{win} is higher for samples of type *a*, these will be classified with a low reliability (low values of ψ_a). According to the above definition, the value of ψ_a lies certainly between 0 and 1 only for the samples belonging to the set on which the value of $O_{\max}$ has been computed, since the relation $O_{\text{win}} \le O_{\max}$ may not hold true for other samples. Therefore, to make the definition applicable, the previous expression must become

$$\psi_a = \max\left\{1 - \frac{O_{\text{win}}}{O_{\max}}, 0\right\} \tag{17}$$

On the other hand, samples of type *b* have comparable distances from at least two reference samples belonging to different classes. Consequently, the reliability parameter ψ_b must be a function of both O_{win} and $O_{2\text{win}}$, where $O_{2\text{win}}$ is the distance between *x* and the reference sample with the second smallest distance from *x* among all the reference set samples belonging to a class which is different from that of O_{win} (i.e. $O_{2\text{win}} = \min_{j\neq k} O_j$, where $k = \arg\min_i O_i$):

$$\psi_b = 1 - \frac{O_{\text{win}}}{O_{2\text{win}}} \tag{18}$$

According to this definition, ψ_b assumes values ranging from 0 to 1, and the larger the difference $O_{2\text{win}} - O_{\text{win}}$, the higher the values of ψ_b.

The classification reliability for the NN classifier is therefore given by

$$\begin{aligned}\psi &= \min\{\psi_a,\psi_b\} = \\ &\min\left\{\max\left\{1 - \frac{O_{\text{win}}}{O_{\max}}, 0\right\}, 1 - \frac{O_{\text{win}}}{O_{2\text{win}}}\right\}\end{aligned} \tag{19}$$

4. TESTING THE APPROACH

To validate our approach, the performance of several multi-expert systems using the combination strategies illustrated in Section 2 and based on different definitions of reliability, including ours, was tested. The experts making up each MES are handwritten digit recognition systems. They have been selected not so much because they achieved best performances in the field (systems performing better have been reported in the literature), but rather because they represent a variety of approaches to recognition, and thus can *a priori* be relied on to have some degree of complementarity. Indeed, we are interested in demonstrating that a suitable combination of experts makes it possible to achieve a better performance than that of the best expert in the pool.

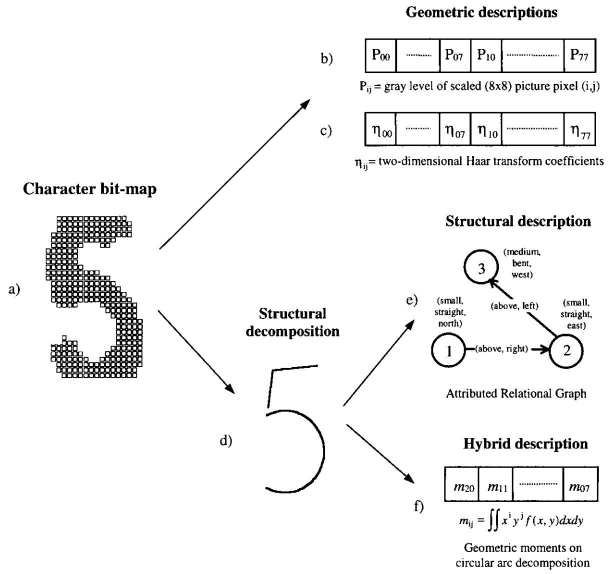

Fig. 3. The four descriptions employed for the experiments. (a) Original character bit-map; (b) general scheme of the description obtained from a scaled grey-level version of a picture; (c) description scheme based on the Haar transform of the bit-map; (d) structural decomposition of (a) in terms of circular arcs; (e) structural description of (d) based on Attributed Relational Graphs; (f) hybrid description scheme based on the evaluation of geometric moments computed on the structural decomposition.

All the tests were performed using the NIST Database 19 [23]. It contains eight sets of images extracted from 3699 Handwriting Sample Forms and digitised at 300 dpi. Each form is composed of 34 fields containing digits, upper-case and lower-case letters, both constrained and unconstrained. Form writers were people on the staff of the U.S. Census Bureau and high school students. Each character, segmented and labelled, is represented by a 128×128 bit-map. NIST strongly suggests using the hsf_4 set as standard OCR reporting set, and allows all other sets to be used as training sets. A cross validation study [23] has demonstrated that hsf_4 is significantly more difficult, from the recognition point of view, than the other sets.

4.1. The Experts Adopted

Each of the experts adopted is made of a descriptor, which characterises the pattern to be recognised as a function of a set of features, and of a classifier which assigns the pattern to one of the possible classes, on the basis of the knowledge acquired during the training phase. To ensure that the experts were as uncorrelated as possible, different descriptor-classifier pairs were chosen.

As regards description, there are two main approaches in the field of handwritten character recognition: in the first (here on called *geometric*) the character is described by means of a set of measurements (*features*) directly extracted from the bit map. The second approach (here on called *structural*) assumes that the character to be recognised can be decomposed into simpler components, and then described in terms of appropriate attributes of the components and of their topological relations. Hybrid techniques that combine the two approaches have also been employed.

We have taken into account four description schemes: two of them are representative of the geometric approach, while the others refer to the structural and the hybrid approach, respectively.

The first geometric description method employed assumes an 8×8 matrix of real numbers in the range [0,1] as character representation. It is obtained by superimposing an 8×8 grid on the character bit-map (see Fig. 3(a)) and computing the average value of the pixels falling in each area. The matrix obtained is finally coded as a 64-element vector (see Fig. 3(b)).

The other geometric description method calculates the two-dimensional Haar transform [29] of the character bit-map. The Haar transform, which is a type of wavelet transform, of a suitably scaled bit-map B of dimensions $N \times N$, where N must be a power of 2, is given by

$$\{\eta_{ij}\} = H \cdot B \cdot H^T \tag{20}$$

where H is the Haar matrix:

$$H = \begin{pmatrix} h_0(0) & h_0(1/N) & \cdots & h_0((N-1)/N) \\ h_1(0) & h_1(1/N) & \cdots & h_1((N-1)/N) \\ \vdots & \vdots & & \vdots \\ h_{N-1}(0) & h_{N-1}(1/N) & \cdots & h_{N-1}((N-1)/N) \end{pmatrix} \tag{21}$$

The $h_k(x)$ are the Haar functions, which are defined for $k = 0,\ldots,N-1$ and for $x \in [0,1]$ by the following equations:

$$h_0(x) = \frac{1}{\sqrt{N}}$$

$$h_k(x) = \frac{1}{\sqrt{N}} \cdot \begin{cases} 2^{p/2} & \text{if} \quad \dfrac{q-1}{2^p} \leq x < \dfrac{q-1/2}{2^p} \\ -2^{p/2} & \text{if} \quad \dfrac{q-1/2}{2^p} \leq x < \dfrac{q}{2^p} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where p and q are the unique integers such that $k = 2^p + q - 1$ and $1 \leq q \leq 2^p$.

In our case, the scaled bit-map B has $N = 64$. The feature vector used consists of the 64 coefficients $(\eta_{00},\ldots,\eta_{07},\eta_{10},\ldots,\eta_{77})$ belonging to the 8×8 submatrix obtained starting from the upper left corner of the transformed image.

The main phases of the process leading to the adopted structural description are briefly summarised below. Characters are first thinned and then, to preserve the original information on character shape, further processed for correcting the shape distortions introduced by thinning. After this correction a character is represented by a set of polygonal lines which are then approximated with pieces of circular arcs (see Fig. 3(d)). The structural description can be represented by an Attributed Relational Graph (ARG) whose nodes represent the component arcs and whose edges represent the relations between the arcs (Fig. 3(e)). Node attributes specify span, relative size and orientation of the corresponding arc, while edge attributes specify topologic relations between arc pair projections on the coordinate axes. More details can be found elsewhere [30].

The fourth description method refers to a hybrid approach and consists of geometric features extracted from a structural decomposition of the character. In particular, after approximating a character with a set of circular arcs, geometric moments are computed on them by means of suitably established recurrent formulae [31]. The moments up to the 7th order are considered (from the experiments it has been noted that the inclusion of higher order moments did not give a significant improvement of the recognition rate); however, since moments of order 0 and 1 have been used to make the remaining moments scale and translation invariant, the final description is made up of a 33 element vector (see Fig. 3(f)).

The aim of obtaining experts as independent as possible has also motivated the choice of the classifiers, which have been selected on the basis of their different characteristics. Three classifiers were implemented by means of MLP networks, while the fourth classifier chosen is of the Nearest Neighbour type.

The above classifiers have been combined with the four previously outlined descriptions, giving rise to the four experts illustrated below and summarised in Table 1. The acronyms used to denote the experts specify the classifier and the associated description type. Only handwritten digits were considered for the test, thus the number of classes is ten for every classifier.

- *The MLP-GS Expert.* The MLP-GS expert combines the MLP classifier with the geometric description based on the scaled bit-map. Therefore, the input layer of the classifier is made of 64 neurons, each one associated to a pixel of the scaled image. The chosen network architecture has a single hidden layer of 30 neurons and an output layer of 10 neurons corresponding to the ten digits. The learning algorithm is the standard Backpropagation one, with a constant learning rate equal to 0.5. The sigmoidal activation function was chosen for all the neurons.
- *The MLP-GH Expert.* The MLP-GH expert uses the geometric description based on the Haar transform. The input layer of the MLP classifier is composed of 64 neurons, each one associated to a coefficient of the Haar transform of the character image. Unlike the previous one, this network has a single hidden layer of 50 neurons, while the output layer is again made up of 10 neurons. The learning algorithm, the learning rate and the activation function are the same as those of the MLP-GS expert.
- *The MLP-H Expert.* In this expert, the classifier works with the hybrid description. Thus the input layer of the classifier is made of 33 neurons. All the remaining network parameters are the same used in the MLP-GS expert.
- *The NN-S Expert.* This uses the structural description associated to the NN classifier. To attribute a sample to a class, a metric is defined in the ARG space, according to which the difference between two characters is measured by the minimum value of the distances between any two possible mappings of the corresponding ARGs. The latter distances are measured as a function of the attribute values of components and relations. Cases where two ARGs are made up of a different number of nodes and/or arcs are also considered. More details about the metric can be found elsewhere [32]. A subset suitably extracted from the training set used in the other cases has been assumed as the reference set: in this way, we have obtained an endurable computational load for the classifier, without affecting its performance.

Table 1. The description and classification models adopted by the experts considered

Expert	Description	Classifier
MLP-GS	Scaled Bit map (8 × 8)	Multi-Layer Perceptron
MLP-GH	Haar transform coefficients	Multi-Layer Perceptron
MLP-H	Geometrical moments on a circular arc-based description	Multi-Layer Perceptron
NN-S	Attributed Relational Graphs	Nearest Neighbor

4.2. Experimental Results

In our tests we used both the hsf_3 and hsf_4 sets from the NIST Database 19. The hsf_3 set was split into two sets: a training set, composed of 34,644 samples, used for training the MLP-based experts; and a so-called training-test set made of 29,252 samples. As already mentioned, a subset of the training set (8000 samples) was assumed as the reference set for the NN-S expert. The training-test set was used both to compute the confusion matrices and to establish the number of cycles for stopping the learning phase of the MLP-based experts, in order to avoid overtraining [33]. The hsf_4 set, made of 58,646 samples, was adopted as the test set. Figure 4 illustrates a few characters of this set.

Eleven different multi-experts have been considered: one of them combines four experts, four of them combine three experts, and each of the remaining six combines two experts. The multi-experts have been designed so as to test all the possible combinations of experts, each obtained by pairing

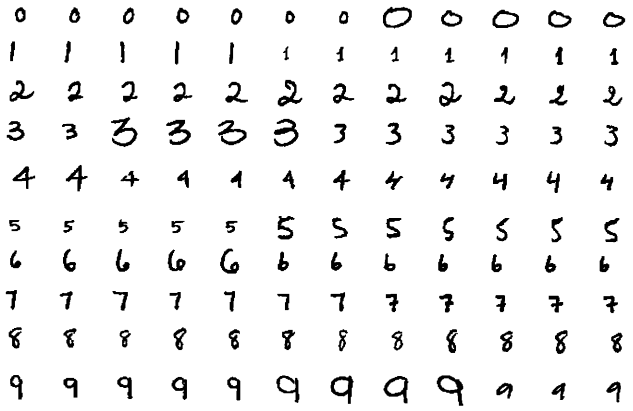

Fig. 4. Some characters of the test set considered (the hsf_4 set of the NIST Database 19).

a classifier with a descriptor. The experimental results obtained with the considered combinations of experts are summarised in Tables 2 and 3.

In particular Table 2 shows, for each MES considered, the recognition rate achieved by the best single classifier and the average recognition rate of the MES component experts. The last two columns, respectively, show the percentage of test set samples that are correctly classified by at least one expert and the percentage of the test set samples correctly classified by all the experts considered in the MES. These parameters represent the performance upper and lower bounds achievable by the MES considered using the MV or the WV rules. The MV rule has been implemented in the version using the definition of D_k given by Eq. (5) for tie breaking. It is worth noting that, by increasing the number of experts, the upper bound increases while the lower bound decreases: in fact, when many experts are considered, the probability that a sample is recognised by at least one of them is high, while the probability that all the experts agree on the same class becomes lower.

Table 2. Some figures characterising the MESs adopted and their component experts

MES ID	No. of experts in the MES	Best single classifier	Average recognition rate	Upper bound	Lower bound
A	2	90.74	89.59	93.65	85.53
B	2	90.74	88.19	94.29	82.08
C	2	90.74	87.43	96.05	77.67
D	2	88.44	87.04	93.39	80.69
E	2	88.44	86.28	95.64	75.78
F	2	85.64	84.88	94.65	73.97
G	3	90.74	88.27	95.60	79.09
H	3	90.74	87.76	97.26	73.94
I	3	90.74	86.83	97.11	71.62
J	3	88.44	86.06	97.08	70.46
K	4	90.74	87.23	97.82	69.36

Table 3. MES recognition rates vs. combining rules

MES ID	Majority voting			Weighted voting			Bayes
	MV	RBMV	Δ_N	WV	RBWV	Δ_N	BC
A	90.39	91.12	22.39	90.39	91.12	22.39	90.12
B	90.35	90.90	13.96	90.35	90.90	13.96	90.58
C	90.74	91.58	15.82	90.74	91.58	15.82	90.98
D	87.79	89.94	38.39	87.79	89.94	38.39	87.75
E	88.11	90.74	34.93	88.11	90.74	34.93	89.41
F	86.53	88.24	21.06	86.53	88.24	21.06	87.73
G	91.19	91.42	5.22	91.19	91.47	6.35	91.04
H	92.16	92.61	8.82	92.16	92.27	2.16	92.41
I	92.00	92.47	9.20	92.00	92.20	3.91	92.24
J	90.79	91.73	14.94	90.79	91.25	7.31	91.59
K	92.28	92.63	6.32	92.28	92.36	1.44	92.30

In Table 3 the recognition rates obtained with the reliability-based combining rules are compared with those achieved by the corresponding MV and WV rules. It can be seen that columns MV and WV contain identical values. This is in accordance with the considerations made in Section 2; in particular, the conditions expressed by Eq. (6) are never verified for our systems. Similarly, with MESs made up of two experts (rows A to F) the RBMV and RBWV columns contain equal values. This happens because also in this case the two RBMV and RBWV rules reduce to the same form.

For each MES and for each combining rule, the recognition percentages together with the value of the parameter Δ_N are also given. Δ_N is defined as the ratio $\Delta R_2/\Delta R_1 \cdot 100$, where ΔR_1 is the difference between the recognition rate obtained by the MV (WV) combining rule and the recognition upper bound (fifth column of Table 2), and ΔR_2 is the difference between the recognition rates obtained using the RBMV (RBWV) and the MV (WV) combining rules. The parameter Δ_N gives a measure of the improvement obtained by introducing the proposed reliability parameter. This measure is normalised with respect to the maximum improvement which can be achieved with the pair MES-combining rule considered. The improvements obtained range from 5.22% to 38.39% for the MV rule and from 2.16% to 38.39% for the WV rule.

The last column of Table 3 shows the results obtained with the BC rule which does not make use of any reliability parameter. Let us recall that this rule uses all the information available in the confusion matrices of the component experts, and thus, in principle, it could achieve recognition rates better than the MV and WV rules. However, the use of the reliability parameter in the RBMV and in the RBWV rules allows us to attain better results than those achieved with the BC rule for most of the multi-expert systems considered. This improvement is more significant when the MES is made of two experts due to the fact that ties are more likely to occur for such systems.

It should be remembered that the recognition systems considered have been selected not because they have an outstanding performance on the database used, but to perform the test on systems adopting different description and classification paradigms. However, the recognition rates obtained are not especially low, considering the quality of the characters in the database.

To better characterise the improvement obtained with the proposed rules, let us consider the comparison between the MV and RBMV rules. With these rules the use of the reliability parameter is limited to the case in which a tie break is needed. In Fig. 5 the percentages of tie breaks correctly solved when using the MV and the RBMV rules are shown for all the multi-expert systems considered. It can be noted that for systems made of two experts, the use of the RBMV rule allows us to recognise more than 70% of the cases in which a tie break occurs, about 9% more than with the MV rule. In the case of systems made of more than two experts the percentage of samples recognized by the RBMV decreases to about 48%, but the average difference between the recognition rates obtained by the RBMV rule and the MV rule increases to more than 13%, with a maximum value of 20%.

5. CONCLUSIONS

A novel parameter measuring the reliability of each single classification act of a recognition system has been introduced, and its use for weighting the votes of the experts making up a multi-expert system has been demonstrated.

The approach has been tested using four handwritten character recognition systems, combined in different ways to form 11 multi-expert systems, on the digits of a large standard database (NIST 19).

Although it is limited to a few percent, the absolute recognition rate improvement achieved by the multi-expert systems when using the two combining rules with the reliability parameter introduced in Section 3 should be considered significant. A more appropriate evaluation of the results obtained should be made by considering that, in a real multi-expert system, an upper bound for the achievable recognition rate is determined by the choice of the component experts. In particular, in our case, the upper bounds of the multi-expert systems used are listed in Table 2. The parameter Δ_N, introduced according to the above considerations, shows that it has been possible to recover up to about 40% of the maximum theoretically achievable improvement given by the difference between the upper bound of the recognition rate and the recognition rate obtained with the MV (WV) rule.

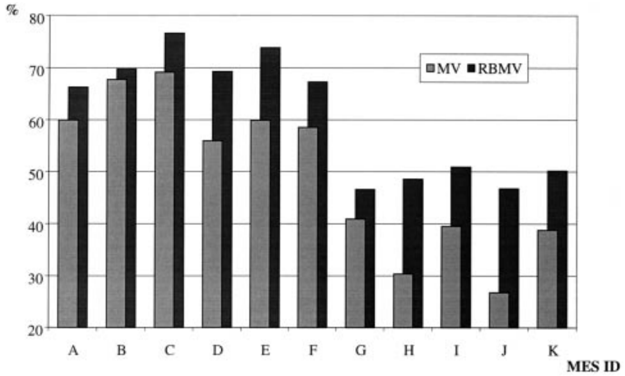

Fig. 5. Percentages of tie breaks which are correctly solved when using the MV and RBMV rules, for each of the multi-expert systems considered.

References

1. Fukunaga K. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press, 1990
2. Fu KS. Syntactic Methods in Pattern Recognition. Academic Press, 1974
3. Pavlidis T. Structural Pattern Recognition. Springer-Verlag, 1977
4. Dasrathy B. Nearest Neighbor Pattern Classification. IEEE Press, 1990
5. Anderson JA, Rosenfeld E (eds). Neurocomputing: Foundations of Research. MIT Press, 1988
6. Suen CY, Nadal C, Legault R, Mai TA, Lam L. Computer Recognition of Unconstrained Handwritten Numeral. Proc IEEE 1992; 80(7): 1162–1180
7. Ackermann B, Bunke H. Combination of Classifiers on the Decision Level for Face Recognition. Technical Report IAM-96–002, Institut für Informatik und angewandte Mathematik, Universität Bern, 1996
8. Ho TK, Hull JJ, Srihari SN. Decision Combination in Multiple Classifier Systems. IEEE Trans Pattern Analysis Machine Intell 1994; 16(1): 66–75
9. Kittler J. Improving Recognition Rates by Classifier Combination: A Theoretical Framework. In: Downtown AC, Impedovo S (eds), Progress in Handwriting Recognition. World Scientific, 1997, pp 231–248
10. Hall DL. Mathematical Techniques in Multi-Sensor Data Fusion. Artech House, 1992
11. Lam L, Suen CY. Increasing Experts for Majority Vote in OCR: Theoretical Considerations and Strategies. Proc 4th International Workshop on Frontiers in Handwriting Recognition 1994, pp 245–254
12. Powalka RK, Sherkat N, Whitrow RJ. Multiple Recognizer Combination Topologies. In: Simner ML, Leedham CG, Thomassen AJWM (eds), Handwriting and Drawing Research: Basic and Applied Issues. IOS Press, 1996, pp 329–342
13. Rahman A, Fairhurst M. Introducing New Multiple Expert Decision Combination Topologies: A Case Study Using Recognition of Handwritten Characters. In Proc 4th International Conference on Document Analysis and Recognition. 1997, pp 886–891
14. Xu L, Krzyzak A, Suen CY. Method of Combining Multiple Classifiers and Their Application to Handwritten Numeral Recognition. IEEE Trans Systems, Man Cybern 1992; 22(3): 418–435
15. Lee D-S, Srihari SN. A Theory of Classifier Combination: The Neural Network Approach. Proc 3rd ICDAR 1995, pp 42–45.
16. Huang YS, Suen CY. A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. IEEE Trans Pattern Analysis Machine Intell 1995; 17(1): 90–94
17. Cho S-B, Kim JH. Combining Multiple Neural Networks by Fuzzy Integral for Robust Classification. IEEE Trans Systems, Man Cybern 1995; 25(2): 380–384

18. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation 1991; 3(1): 79–87
19. Jordan MI, Jacobs RA. Hierarchies of adaptive experts. In: Moody JE, Hanson S, Lippmann RP (eds). Advances in Neural Information Processing System 4. Morgan Kauffmann, 1992; pp 985–992
20. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural Computation 1994; 6(2): 181–214
21. Lam L, Suen CY. A Theoretical Analysis of the Application of Majority Voting to Pattern Recognition. Proc 12th International Conference on Pattern Recognition, vol 2. IEEE Press, 1994, pp 418–420
22. Lam L, Suen CY. Optimal combination of Pattern Classifiers. Pattern Recognition Lett 1995; 16: 945–954
23. Grother PJ, NIST Special Database 19. Technical Report, National Institute of Standards and Technology, 1995
24. Bloch I. Information Combination Operators for Data Fusion: A Comprative Review with Classification. IEEE Trans Systems, Man Cybern – Part A 1996; 26(1): 52–76
25. Rumelhart DE, Mc Clelland JL, Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Vol.1: Foundations. MIT Press, 1986
26. Cover TM, Hart PE. Nearest Neighbor Pattern Classification. IEEE Trans Infor Theory 1967; 13: 21–27
27. Cordella LP, Sansone C, Tortorella F, Vento M, De Stefano C. Neural Networks Classification Reliability. In: Leondes CT (ed) Academic Press theme volumes on Neural Network Systems, Techniques and Applications. Academic Press, 1998, pp 161–199
28. Lippmann RP. An Introduction to Computing with Neural Nets. IEEE ASSP Magazine 1987; 4(2): 4–22
29. Pratt WK. Digital Image Processing. Wiley-Interscience, 1991
30. Cordella LP, De Stefano C, Vento M. A Neural Network Classifier for OCR using Structural Descriptions. Machine Vision Applications 1995; 8: 336–342
31. Foggia P, Sansone C, Tortorella F, Vento M. Combining Statistical and Structural Approaches for Handwritten Character Description. Image Vision Comput 1999; 17(9): 701–711
32. De Stefano C, Foggia P, Tortorella F, Vento M. A Distance Measure for Structural Descriptions using Circle Arcs as Primitives. Proc 13th International Conference on Pattern Recognition, vol II. IEEE Press, 1996, pp 290–294.
33. Hecth-Nielsen R. Neurocomputing. Addison-Wesley, 1990

Luigi P. Cordella was born in Milan, Italy, in 1938. After receiving the Laurea degree in Physics from the University of Rome, he became an Assistant Professor. In 1970 he joined the Institute of Cybernetics of the Italian National Research Council (CNR) in Arco Felice, Naples, where he became head of the Image Analysis Department. In 1983 he joined the Faculty of Engineering of the University of Naples 'Federico II', where he is a Full Professor of Computer Science. From 1989 to 1992 he was Chairman of the Department of Computer Science and Systems of the above university. He has been active in the fields of Electronics, Mathematical Models of Biological Systems, Image Analysis, Pattern Recognition, Computer Applications in Biomedicine, Parallel Computing. His present research interests include Syntactic and Structural Pattern Recognition, Shape Analysis, Document Recognition and Neural Networks. He has published more than 100 papers and is co-editor of five books. Professor Cordella has been chairman or a member of the program committee of several international conferences. He is a Fellow of the IAPR and a member of the IEEE Computer Society.

Pasquale Foggia was born in Naples, Italy, in 1971. He received a Laurea degree with honours in Computer Engineering from the University of Naples 'Federico II' in 1995. He is currently a PhD student at the Dipartimento di Informatica e Sistemistica of the University of Naples 'Federico II'. His research interests are in the fields of classification algorithms, optical character recognition, graph matching and inductive learning. He is a member of the International Association for Pattern Recognition (IAPR).

Carlo Sansone was born in Naples, Italy, in 1969. He received a Laurea degree with honours in Electronic Engineering in 1993 and a PhD degree in Electronic and Computer Engineering in 1997, both from the University of Naples 'Federico II'. He has been Assistant Professor of Computer Sciences and Databases at the University of Naples 'Federico II' and Assistant Professor of Computer Science at the University of Cassino. His research interests are in the fields of classification algorithms, optical character recognition and neural networks theory and applications. Carlo Sansone is a member of the International Association for Pattern Recognition (IAPR).

Francesco Tortorella was born in Salerno, Italy, in 1963. He received a Laurea degree with honours in Electronic Engineering in 1991 and a PhD degree in Electronic and Computer Engineering in 1995, both from the University of Naples 'Federico II', Italy. From 1995 to 1996 he was Assistant Professor of Computer Architectures at the University of Naples 'Federico II'; from 1997 to 1998 he was Assistant Professor of Computer Science at the University of Cassino. In September 1998 he joined the Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e Matematica Industriale, University of Cassino, as a researcher. His current research interests include classification algorithms, optical character recognition, map and document processing and neural networks. Dr. Tortorella is a member of the International Association for Pattern Recognition (IAPR).

Mario Vento was born in Naples, Italy, in 1960. In 1984 he received a Laurea degree with honours in Electronic Engineering, and in 1988 a PhD in Electronic and Computer Engineering, both from University of Naples 'Federico II', Italy. Since 1989, he has been a researcher associated with the Dipartimento di Informatica e Sistemistica at the above University. Currently, he is Associate Professor of Artificial Intelligence and Computer Science at the Faculty of Engineering of the University of Naples. His present research interests are in the field of image analysis and recognition, image description and classification techniques, soft computing, machine learning and artificial intelligence. Mario Vento is a member of the International Association for Pattern Recognition (IAPR).

Correspondence and offprint requests to: L. P. Cordella, Dipartimento di Informatica e Sistemistica, Università degli Studi di Napoli 'Federico II', Via Claudio 21, I-80125 Napoli, Italy. Email: cordel@unina.it