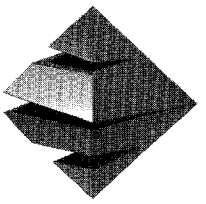
Pattern Analysis & Applic. (1998)1:18–27



Combining Classifiers: A Theoretical Framework

J. Kittler

Centre for Vision, Speech and Signal Processing, School of Electronic Engineering, Information Technology and Mathematics, University of Surrey, Guildford, UK

Abstract: The problem of classifier combination is considered in the context of the two main fusion scenarios: fusion of opinions based on identical and on distinct representations. We develop a theoretical framework for classifier combination for these two scenarios. For multiple experts using distinct representations we argue that many existing schemes such as the product rule, sum rule, min rule, max rule, majority voting, and weighted combination, can be considered as special cases of compound classification. We then consider the effect of classifier combination in the case of multiple experts using a shared representation where the aim of fusion is to obtain a better estimate of the appropriate *a posteriori* class probabilities. We also show that the two theoretical frameworks can be used for devising fusion strategies when the individual experts use features some of which are shared and the remaining ones distinct. We show that in both cases (distinct and shared representations), the expert fusion involves the computation of a linear or nonlinear function of the *a posteriori* class probabilities estimated by the individual experts. Classifier combination can therefore be viewed as a multistage classification process whereby the *a posteriori* class probabilities generated by the individual classifiers are considered as features for a second stage classification scheme. Most importantly, when the linear or nonlinear combination functions are obtained by training, the distinctions between the two scenarios fade away, and one can view classifier fusion in a unified way.

Keywords: Compound decision theory; Multiple expert fusion; Pattern classification

1. INTRODUCTION

The problem of classifier combination has always been of interest to the pattern recognition community. Initially, the goal of classifier combination was to improve the efficiency of decision making by adopting multistage combination rules, whereby objects are classified by a simple classifier using a small set of inexpensive features in combination with a reject option. For the more difficult objects more complex procedures, possibly based on additional, more costly features, are employed [1–4]. In other studies, successive classification stages gradually reduce the set of possible classes [5–8]. Multistage classifiers may also be used to stabilise the training of classifiers based on a small sample size, e.g. by the use of bootstrapping [9].

More recently, it has been observed that the accuracy of pattern classification can also be improved by multiple expert fusion. In other words, the idea is not to rely on a single decision making scheme. Instead, several designs (experts) are used for decision making. By combining the opinions of the individual experts, a consensus decision is derived. Various classifier combination schemes have been devised, and it has been experimentally demonstrated that some of them consistently outperform a single best classifier.

An interesting issue in the research concerning classifier ensembles is the way they are combined. If only labels are available a majority vote [7,10] or a label ranking [11,12] may be used. If continuous outputs like *a posteriori* probabilities are supplied, an average or some other linear combination has been suggested [13,14]. It depends upon the nature of the input

Received: 8 October 1997
Received in revised form: 6 January 1998
Accepted: 10 January 1998

classifiers and the feature space as to whether this can be theoretically justified. A review of these possibilities is presented in Hansen and Salamon [15]. If the classifier outputs are interpreted as fuzzy membership values, belief values or evidence, fuzzy rules [16,17], belief functions and Dempster–Shafer techniques [10,14,18,19] are used. Finally, it is possible to train the output classifier separately using the outputs of the input classifiers as new features [20,21]. Woods et al [22], on the other hand, take the view that different classifiers are competent to make decisions in different regions, and their approach involves partitioning the observation space into such regions. For a recent review of the literature see Kittler [23].

From the point of view of their analysis, there are basically two classifier combination scenarios. In the first scenario, all the classifiers use the same representation of the input pattern. In this case, each classifier, for a given input pattern, can be considered to produce an estimate of the same *a posteriori* class probability.

In the second scenario, each classifier uses its only representation of the input pattern. In other words, the measurements extracted from the pattern are unique to each classifier. An important application of combining classifiers in this scenario is the possibility to integrate physically different types of measurements/features. In this case, it is no longer possible to consider the computed *a posteriori* probabilities to be estimates of the same functional value, as the classification systems operate in different measurement spaces.

In this paper, we develop a theoretical framework for classifier combination approaches for these two scenarios. For multiple experts using distinct representations, we argue that many existing schemes can be considered as special cases of compound classification, where all the representations are used jointly to make a decision. We note that under different assumptions and using different approximations, we can derive the commonly used classifier combination schemes such as the product rule, sum rule, min rule, max rule, majority voting and weighted combination schemes. We address the issue of the sensitivity of various combination rules to estimation errors, and point out that the techniques based on the benevolent sum-rule fusion are more resilient to errors than those derived from the severe product rule.

We then consider the effect of classifier combination in the case of multiple experts using a shared representation. We show that here the aim of fusion is to obtain a better estimate of the appropriate *a posteriori* class probabilities. This is achieved by the means of reducing the estimation error variance. We also show that the two theoretical frameworks for the case of distinct and shared representation, respectively, can be used for devising fusion strategies when the individual experts use features some of which are shared and the remaining ones distinct.

We show that in both cases (distinct and shared representations), the expert fusion involves the computation of a linear or nonlinear function of the *a posteriori* class probabilities estimated by the individual experts. Classifier combination can therefore be viewed as a multistage classification process, whereby the *a posteriori* class probabilities generated by the individual classifiers are considered as features for a second stage classification scheme. Most importantly, when the linear or nonlinear combination functions are obtained by training, the distinctions between the two scenarios fade away, and one can view classifier fusion in a unified way. This probably explains the success of many heuristic combination strategies that have been suggested in the literature without any concerns about the underlying theory.

The paper is organised as follows. In Section 2 we discuss combination strategies for experts using independent (distinct) representations. In Section 3 we consider the effect of classifier combination for the case of shared (identical) representation. The findings of the two sections are discussed in Section 4. Finally, Section 5 offers a brief summary.

2. DISTINCT REPRESENTATIONS

It has been observed that classifier combination is particularly effective if the individual classifiers employ different features [12,14,24]. Consider a pattern recognition problem where pattern Z is to be assigned to one of the m possible classes $\{\omega_1,\ldots,\omega_m\}$. Let us assume that we have R classifiers, each representing the given pattern by a distinct measurement vector. Denote the measurement vector used by the i-th classifier by $\mathbf{x}_i$. In the measurement space each class ω_k is modelled by the probability density function $p(\mathbf{x}_i/\omega_k)$, and its *a priori* probability of occurrence is denoted $P(\omega_k)$. We shall consider the models to be mutally exclusive, which means that only one model can be associated with each pattern.

Now according to the Bayesian theory, given measurements $\mathbf{x}_i, = 1,\ldots,R$, the pattern, Z, should be assigned to class ω_j, i.e. its label θ should assume value $\theta = \omega_j$, provided the *a posteriori* probability of that interpretation is maximum, i.e.

$$\text{assign} \quad \theta \rightarrow \omega_j \quad \text{if}$$

$$P(\theta = \omega_j|\mathbf{x}_{1,\ldots,}\mathbf{x}_R) = \max_k P(\theta = \omega_k|\mathbf{x}_1,\ldots,\mathbf{x}_R) \quad (1)$$

Let us rewrite the *a posteriori* probability $P(\theta = \omega_k|\mathbf{x}_1,\ldots,\mathbf{x}_R)$ using the Bayes theorem. We have

$$P(\theta = \omega_k|\mathbf{x}_1,\ldots,\mathbf{x}_R) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k)P(\omega_k)}{p(\mathbf{x}_1,\ldots,\mathbf{x}_R)} \quad (2)$$

where $p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k)$ and $p(\mathbf{x}_1,\ldots,\mathbf{x}_R)$ is the unconditional measurement joint probability density. Since the latter is class independent, in the following, we can concentrate only on the numerator terms of Eq. (2).

Let us assume that measurements $\mathbf{x}_j$, $\forall\ j$ are conditionally statistically independent. This assumption may seem to be rather strong, but as the classifiers use distinct representations, it will often be satisfied, especially if the representations are derived from completely different sensing modalities [25]. Under this assumption

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k) = \Pi_{i=1}^{R} p(\mathbf{x}_i|\theta = \omega_k) \quad (3)$$

where $p(\mathbf{x}_i|\theta = \omega_k)$ is the measurement process model of the i-th representation. Substituting from Eq. (3) into Eq. (2) and eventually into Eq. (1), we obtain the decision rule

$$\textit{assign} \quad \theta \rightarrow \omega_j \quad \textit{if} \quad (4)$$

$$P(\omega_j)\prod_{i=1}^{R} p(\mathbf{x}_i|\theta = \omega_j) = \max_{k=1}^{m} P(\omega_k) \prod_{i=1}^{R} p(\mathbf{x}_i|\theta = \omega_k)$$

or in terms of the *a posteriori* probabilities yielded by the respective classifiers

$$\textit{assign} \quad \theta \rightarrow \omega_j \quad \textit{if}$$

$$P^{-(R-1)}(\omega_j) \prod_{i=1}^{R} P(\theta = \omega_j|\mathbf{x}_i)p(\mathbf{x}_i)$$

$$= \max_{k=1}^{m} P^{-(R-1)}(\omega_k) \prod_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i)p(\mathbf{x}_i) \quad (5)$$

The decision rule (5) quantifies the likelihood of a hypothesis by combining the *a posteriori* probabilities generated by the individual classifiers by means of a product rule. It is effectively a severe rule of fusing the classifier outputs, as it is sufficient for a single recognition engine to inhibit a particular interpretation by outputting a close to zero probability for it. We shall adopt the approach used in Kittler et al [26] to show that, under certain assumptions, this severe rule can be developed into a benevolent information fusion rule which has the form of a sum. Benevolent fusion rules are less affected by one particular expert than severe rules. Thus, even if the soft decision outputs of a few experts for a particular hypothesis are close to zero, the hypothesis may be accepted, provided it receives a sufficient support from all the other experts.

To develop such a benevolent rule, let us express the product of the *a posterior* probabilities and mixture densities on the right-hand side of Eq. (5) $P(\theta = \omega_k|\mathbf{x}_i)p\mathbf{x}_i)$ as

$$P(\theta = \omega_k|\mathbf{x}_i)p(\mathbf{x}_i) = P(\theta = \omega_k)p_i(1 + \delta_{ki}) \quad (6)$$

where p_i is a nominal reference value of the mixture density $p(\underline{\mathbf{x}}_i)$. A suitable choice of p_i is, for instance, $p_i = \max_{\underline{x}} p(\underline{\mathbf{x}}_i)$. Substituting Eq. (6) for the *a posteriori* probabilities in Eq. (5), we find

$$P^{-(R-1)}(\omega_k) \prod_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i)p(\mathbf{x}_i) = P(\omega_k)$$

$$\prod_{i=1}^{R} p_i \prod_{i=1}^{R} (1 + \delta_{ki}) \quad (7)$$

If we expand the product and neglect any terms of second and higher order, we can approximate the right-hand side of Eq. (7) as

$$P(\omega_k) \prod_{i=1}^{R} p_i \prod_{i=1}^{R} (1 + \delta_{ki}) \doteq P(\omega_k)$$

$$\prod_{i=1}^{R} p_i + P(\omega_k) \prod_{i=1}^{R} p_i \sum_{i=1}^{R} \delta_{ki} \quad (8)$$

Substituting Eqs (8) and (6) into Eq. (5) and eliminating $\Pi_{i=1}^{R} p_i$, we obtain a sum decision rule

$$\textit{assign}\ \theta \rightarrow \omega_j\ \textit{if}\ (1 - R)P(\omega_j) + \sum_{i=1}^{R} \frac{P(\omega_j|\mathbf{x}_i)p(\mathbf{x}_i)}{p_i}$$

$$= \max_{k=1}^{m}[(1 - R)P(\omega_k) + \sum_{i=1}^{R} \frac{P(\omega_k|\mathbf{x}_i)p(\mathbf{x}_i)}{p_i}] \quad (9)$$

This approximation will be valid provided that δ_{ki} satisfies $|\delta_{ki}| \ll 1$. It can easily be established that this condition will be satisfied if $P(\omega_k|\mathbf{x}_i)p(\mathbf{x}_i)/p_iP(\omega_i) - 1$ is small in absolute value sense. Note that this condition will hold when the amount of information about class identity of the object gained by observing $\mathbf{x}_i$ is small and the observation is representative for the distinction of $\mathbf{x}_i$, which means that $p(\mathbf{x}_i)$ will be close to the reference value p_i. However, whatever approximation error is introduced when the conditions do not hold, we shall see later that the adoption of the approximation has some other benefits which will justify even the introduction of relatively gross errors at this step.

Before proceeding any further, it may be pertinent to ask why we did not cancel out the unconditional probability density functions $p(\mathbf{x}_i)$ from the decision rule. The main reason is that this term conveys very useful information about the confidence of the classifier in the observation made. It is clear that a pattern representation for which the value of the probability density is very small for all the classes will be an outlier, and should not be classified by the respective classifier. By retaining this information, in the case of the product rule (5), we have the option of suppressing

the effect of outliers on the decision making process by setting the *a posteriori* probabilities for all the classes to a constant, i.e.

$$if \frac{p(\mathbf{x}_i)}{p_i} \leq threshold\ then\ P(\omega_k|\mathbf{x}_i) = const.\ \forall k \tag{10}$$

In contrast, the sum information fusion rule will automatically control the influence of such outliers on the final decision. In other words, the classifier combination rule in Eq. (9) is a weighted average rule, where the weights reflect the confidence in the soft decision values computed by the individual classifiers. Thus, our decision rule (9) can be expressed as

$$\begin{aligned} &assign \quad \theta \rightarrow \omega_j \quad if \\ &(1-R)P(\omega_j) + \sum_{i=1}^{R} \mathrm{w}(\mathbf{x}_i)P(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m}[(1-R) \\ &P(\omega_k) + \sum_{i=1}^{R} \mathrm{w}(\mathbf{x}_i)P(\omega_k|\mathbf{x}_i)] \end{aligned} \tag{11}$$

The main practical difficulty with the weighted average classifier combiner as specified in Eq. (11) is that not all classifiers will have the inner capability to output such information. For instance, it would not be provided by a multilayer perceptron and many other classification methods. We shall therefore limit our objectives somewhat, and identify the weights w_i which will reflect the relative confidence in the classifiers in expectation. This can be done easily by selecting weight values by means of minimising the empirical classification error count produced by the decision rule

$$\begin{aligned} &assign \quad \theta \rightarrow \omega_j \quad if \\ &(1-R)P(\omega_j) + \sum_{i=1}^{R} \mathrm{w}_i P(\omega_j|\mathbf{x}_i) \\ &= \max_{k=1}^{m}[(1-R)P(\omega_k) + \sum_{i=1}^{R} \mathrm{w}_i P(\omega_k|\mathbf{x}_i)] \end{aligned} \tag{12}$$

in which the data dependence of the weights has been suppressed. In other words, we find w_i, $i = 1, R$ such that

$$e = \frac{1}{N}\sum_{k=1}^{N} \eta(Z_k) \tag{13}$$

where Z_k, $k = 1, N$ is the k-th training sample and $\eta(Z_k)$ takes values

$$\eta(Z_k) = \begin{cases} 0 & \beta_k = \theta_k \\ 1 & otherwise \end{cases} \tag{14}$$

is minimised. In Eq. (14), β_k is the true class label of pattern Z_k and θ_k is the class label assigned to it by the decision rule (12). The optimisation can easily be achieved by an exhaustive search through the weight space.

For equal *a priori* class probabilities, the decision rule (12) simplifies to

$$\begin{aligned} &assign\ \theta \rightarrow \omega_j\ if \\ &\sum_{i=1}^{R} \mathrm{w}_i P(\omega_j|\mathrm{x}_i) = \max_{k=1}^{R} \sum_{i=1}^{R} \mathrm{w}_i P(\omega_k|\mathrm{x}_i) \end{aligned} \tag{15}$$

2.1. Error Sensitivity

In practice, the individual experts will not output the true *a posteriori* probabilities $P(\omega_k|\mathbf{x}_i)$, $i = 1,R$ but instead their estimates $\hat{P}(\omega_k|\mathbf{x}_i)$, where

$$\hat{P}(\omega_k|\mathbf{x}_i) = P(\omega_k|\mathbf{x}_i) + \epsilon(\mathbf{x}_i) \tag{16}$$

and $\epsilon(\mathbf{x}_i)$ is the estimation error. Replacing the *a posteriori* class probabilities in decision rule (12) with their hatted counterparts, and substituting from Eq. (16), we have

$$\begin{aligned} &assign \quad \theta \rightarrow \omega_j \quad if \\ &(1-R)P(\omega_j) + \sum_{i=1}^{R} \mathrm{w}_i[P(\omega_j|\mathbf{x}_i) + e_{ji}] = \max_{k=1}^{m} \\ &\left\{(1-R)P(\omega_k) + \sum_{i=1}^{R} \mathrm{w}_i[P(\omega_k|\mathbf{x}_i) + e_{ki}]\right\} \end{aligned} \tag{17}$$

which can be rewritten as

$$\begin{aligned} &assign\ \theta \rightarrow \omega_j\ if \\ &(1-R)P(\omega_j) + \left[\sum_{i=1}^{R} \mathrm{w}_i P(\omega_j|\mathbf{x}_i)\right] \\ &\left[1 + \frac{\sum_{i=1}^{R} \mathrm{w}_i e_{ji}}{\sum_{i=1}^{R} \mathrm{w}_i P(\omega_j|\mathbf{x}_i)}\right] = \max_{k=1}^{m} \\ &\left\{(1-R)P(\omega_k) + \left[\sum_{i=1}^{R} \mathrm{w}_i P(\omega_k|\mathbf{x}_i)\right] \cdot \right. \\ &\left. \left[1 + \frac{\sum_{i=1}^{R} \mathrm{w}_i e_{ki}}{\sum_{i=1}^{R} \mathrm{w}_i P(\omega_k|\mathbf{x}_i)}\right]\right\} \end{aligned} \tag{18}$$

A comparison of Eqs (12) and (18) shows that each term in the *error free* classifier combination rule (12) is affected by error factor

$$\left[1 + \frac{\sum_{i=1}^{R} \mathrm{w}_i e_{ki}}{\sum_{i=1}^{R} \mathrm{w}_i P(\omega_k|\mathbf{x}_i)}\right] \tag{19}$$

Thus, in the weighted average rule the compounded effect of errors, which is computed as a sum, is scaled by the sum of the weighted *a posteriori* probabilities. A judicial choice of weights (by training) and the implied error averaging process will result in the damp-

ening of the errors. Thus, the weighted sum decision rule can be expected to be resilient to estimation errors, and also to approximation errors that we may have inadvertently introduced in developing it. This contrasts with the inordinate sensitivity to errors exhibited by the product rule [26]. Although the product rule can be expected to perform better when no estimation errors are present, for large errors the superior performance of the sum rule has been confirmed experimentally [27,28]. It follows, therefore, that the weighted average classifier combination rule is not only a very simple and intuitive technique of improving the reliability of decision making based on different classifier opinions, but it is also remarkably robust.

It can readily be shown that the decision rules (5) and (9) simplify to the following commonly used combination strategies:

assign $\theta \rightarrow \omega_j$ *if*

Product Rule

$$P^{-(R-1)}(\omega_j)\Pi_{i=1}^{R}P(\theta = \omega_j|\mathbf{x}_i) =$$

$$\max_{k=1}^{m} P^{-(R-1)}(\omega_k)\Pi_{i=1}^{R}P(\theta = \omega_k|\mathbf{x}_i) \tag{20}$$

This rule follows directly from Eq. (5).

Sum Rule

$$(1-R)P(\omega_j) + \sum_{i=1}^{R} P(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m}$$

$$\left[(1-R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_k|\mathbf{x}_i)\right] \tag{21}$$

This rule follows from Eq. (9) under the assumption of equal weighting of the outputs of the respective experts, i.e. $w(\mathbf{x}_i) = 1\ \forall i$ and $\forall \mathbf{x}_i$.

Max Rule

$$\max_{i=1}^{R} P(\theta = \omega_j|\mathbf{x}_i) = \max_{k=1}^{m} \max_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \tag{22}$$

This rule approximates the sum rule in Eq. (21) under the assumption that all the classes are *a priori* equiprobable, and the sum will be dominated by the expert decision output which lends the maximum support for a particular hypothesis.

Min Rule

$$\min_{i=1}^{R} P(\theta = \omega_j|\mathbf{x}_i) = \max_{k=1}^{m} \min_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \tag{23}$$

This rule approximates the product rule (20) under the assumption that all the classes are *a priori* equiprobable and the product will be dominated by the expert decision output which lends the minimum support for a particular hypothesis.

Majority Vote Rule

$$\sum_{i=1}^{R} \Delta_{ji} = \max_{k=1}^{m} \sum_{i=1}^{R} \Delta_{ki} \tag{24}$$

This rule is obtained from the sum rule in Eq. (21) under the assumption that all the classes are *a priori* equiprobable and the individual expert outputs $P(\theta=\omega_k|\mathbf{x}_i)$ are hardened into outputs Δ_{ki} as $\Delta_{ki}=1$ *if* $P(\theta=\omega_k|\mathbf{x}_i) = \max_{l=1}^{m} P(\theta=\omega_l|\mathbf{x}_i)$ and zero otherwise.

As the combination strategies *max rule* and *vote* are related to the *sum rule* [26], they are less sensitive to estimation errors, and are therefore likely to perform better than the *min-rule* which can be derived from the *product rule*.

3. IDENTICAL REPRESENTATIONS

In many situations we wish to combine the results of multiple classifiers which use an identical representation for the input pattern $\mathbf{x}$. A typical example of this situation is a battery of k-NN classifiers which employ different numbers of nearest neighbours to reach a decision. Alternatively, neural network classifiers trained with different initialisations or different training sets [21,29,30] also fall into this category. The combination of ensembles of neural networks has been studied elsewhere [13,15–18,20].

By means of classifier combination, one is able to obtain a better estimate of the *a posteriori* class probabilities, and in consequence, a reduced classification error. A typical estimator is the averaging estimator

$$\dot{P}(\omega_i|\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} \hat{P}_j(\omega_i|\mathbf{x}) \tag{25}$$

where $\hat{P}_j(\omega_i|\mathbf{x})$ is the *a posteriori* class probability estimate given pattern $\mathbf{x}$, delivered by the jth estimator and $\dot{P}(\omega_i|\mathbf{x})$ is the combined estimate based on N observations.

Assuming that the errors $e_j(\omega_i|\mathbf{x})$ between the true class *a posteriori* probabilities $P(\omega_i|\mathbf{x})$ and their estimates are unbiased, i.e.

$$E\{e_j(\omega_i|\mathbf{x})\} = E\{\hat{P}_j(\omega_i|\mathbf{x}) - P(\omega_i|\mathbf{x})\} = 0\ \forall i, j, \mathbf{x} \tag{26}$$

the combined estimate $\dot{P}(\omega_i|\mathbf{x})$ will be an unbiased estimate of $P(\omega_i|\mathbf{x})$. Suppose the standard deviations $\sigma_j(\omega_i|\mathbf{x})$ $\forall i,j$ of errors $e_j(\omega_i|\mathbf{x})$ are equal, i.e.

$$\sigma_j(\omega_i|\mathbf{x}) = \sigma(\mathbf{x}) \ \forall i,j \tag{27}$$

Then, provided the errors $e_j(\omega_i|\mathbf{x})$ are independent, the variance of the error distribution for the combined estimate $\dot{\sigma}^2(\mathbf{x})$ will be

$$\dot{\sigma}^2(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{N} \tag{28}$$

Now, if the standard deviations $\sigma_j(\omega_i|\mathbf{x})$ of the errors are not identical, then the combined estimate should take that into account by weighting more the contributions of the estimates associated with a lower variance, i.e.

$$\dot{P}(\omega_i|\mathbf{x}) = \frac{1}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2(\omega_i|\mathbf{x})}} \sum_{j=1}^{N} \frac{1}{\sigma_j^2(\omega_i|\mathbf{x})} \hat{P}_j(\omega_i|\mathbf{x}) \tag{29}$$

Provided the errors are unbiased and independent, the combined estimate in Eq. (29) will also be unbiased, and its variance $\dot{\sigma}_{ji}^2(\omega_i|\mathbf{x})$ will be

$$\dot{\sigma}^2(\mathbf{x}) = \frac{1}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2(\omega_i|\mathbf{x})}} \tag{30}$$

From Eq. (30), it can be seen that the variance of the error distribution of the combined estimator will be dominated by the low variance terms.

The weighted estimator (29) represents a general case which may be written as

$$\dot{P}(\omega_i|\mathbf{x}) = \sum_{j=1}^{N} w_{ij}(\mathbf{x}) \hat{P}_j(\omega_i|\mathbf{x}) \tag{31}$$

with the weights $w_{ij}(\mathbf{x})$ satisfying

$$\sum_{j=1}^{N} w_{ij}(\mathbf{x}) = 1 \tag{32}$$

It will assume a specific form in particular circumstances. For instance, if the properties of the individual estimators are class independent, the weights will satisfy

$$w_{ij}(\mathbf{x}) = w_j(\mathbf{x}) \tag{33}$$

If, in addition, the variances of the error distributions of the individual estimators $\sigma_j^2(\omega_i|\mathbf{x})$ are independent of the position in the pattern space the weights will satisfy

$$w_{ij}(\mathbf{x}) = w_j \tag{34}$$

It also subsumes the case when the variances are all identical with

$$w_{ij}(\mathbf{x}) = \frac{1}{N} \tag{35}$$

Recall that when the respective variances of the individual estimators are known, the weights can be determined using the formula

$$w_{ij}(\mathbf{x}) = \frac{\frac{1}{\sigma_j^2(\omega_i|\mathbf{x})}}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2(\omega_i|\mathbf{x})}} \tag{36}$$

If this information is not available, it may be possible to estimate the appropriate weights so that the classification error obtained with the estimator in Eq. (31) is minimised. To adopt this approach, it will be necessary to have another independent set of training data.

Note that the estimator (31) is defined as a linear combination of the individual estimates. This immediately suggests that it may be possible to obtain even a better combined estimate of the class *a posteriori* probabilities by means of a nonlinear combination function as

$$\dot{P}(\omega_i|\mathbf{x}) = F(\hat{P}_1(\omega_i|\mathbf{x}),\ldots,\hat{P}_N(\omega_i|\mathbf{x})) \tag{37}$$

In fact, estimators which aim to enhance their resilience to outliers by adopting a rank order statistic such as the median,

$$\dot{P}(\omega_i|\mathbf{x}) = med_{j=1}^{N} \hat{P}_j(\omega_i|\mathbf{x}) \tag{38}$$

fall into this category. Such nonlinear estimators do not require any additional training. However, if sufficient additional training data is available, a suitable nonlinear function may be found by means of general function approximation (i.e. neural network methodology), or by other design alternatives. The effective local variance of the resulting estimator could be estimated from the input variances by function linearisation techniques.

To investigate the effect of classifier combination, let us examine the distribution of the *a posteriori* probabilities at a single point $\mathbf{x}$. Suppose the *a posteriori* probability of class ω_s is maximum, i.e. $P(\omega_s|\mathbf{x}) = \max_{i=1}^{m} P(\omega_i|\mathbf{x})$, giving the local Bayes error $e_B = 1 - \max_{i=1}^{m} P(\omega_i|\mathbf{x})$. However, our classifiers only estimate these *a posteriori* class probabilities, and the associated estimation errors may result in suboptimal decisions and consequently in an additional classification error. To quantify this additional error, we have to establish what the probability is for the recognition system to make a labelling error. This situation will occur when any of the *a posterior* class probability estimates for a

class other than ω_s become maximum over all the classes. Let us derive the probability of the event occurring for class ω_i, i.e. when

$$\dot{P}(\omega_i|\mathbf{x}) - \dot{P}(\omega_j|\mathbf{x}) > 0 \ \forall j \neq i \tag{39}$$

Note that the left-hand side of Eq. (39) can be expressed as

$$P(\omega_i|\mathbf{x}) - P(\omega_j|\mathbf{x}) + \epsilon(\omega_i|\mathbf{x}) - \epsilon(\omega_j|\mathbf{x}) > 0 \tag{40}$$

where $\epsilon(\omega_i|\mathbf{x})$ is the error of the combined estimate. Equation (40) defines a constraint for the two estimation errors $\epsilon(\omega_k|\mathbf{x})$ $k = i,j$ as

$$\epsilon(\omega_i|\mathbf{x}) - \epsilon(\omega_j|\mathbf{x}) > P(\omega_j|\mathbf{x}) - P(\omega_i|\mathbf{x}) \tag{41}$$

Now, on the left-hand side of Eq. (41) we have two identically distributed random variables. Let us assume that the distributions are Gaussian. This, in practice, will approximate the true distribution of estimation errors very coarsely as both ends of the [0,1] interval from which the *a posteriori* class probabilities can assume values will clip the errors. Nevertheless, the analysis under even such a simplistic assumption will give an indication of the benefits of classifier combination.

Since the error distributions are Gaussian, the distribution of the difference of the two random variables will also be Gaussian, with a twice as large variance. The probability of constraint (41) being satisfied is given by the area under the Gaussian tail with a cut-off point at $P(\omega_j|\mathbf{x}) - P(\omega_i|\mathbf{x})$. More specifically, this probability, which we shall denote $Q_{ij}(\Delta P_{ji}(\mathbf{x}))$, is given by

$$Q_{ij}(\Delta P_{ji}(\mathbf{x})) = \begin{cases} \frac{1}{2} - \frac{1}{2} erf\left(\frac{\Delta P_{ji}(\mathbf{x})}{2\dot{\sigma}}\right) & \Delta P_{ji}(\mathbf{x}) \geq 0 \\ \frac{1}{2} + \frac{1}{2} erf\left(\frac{|\Delta P_{ji}(\mathbf{x})|}{2\dot{\sigma}}\right) & \Delta P_{ji}(\mathbf{x}) < 0 \end{cases} \tag{42}$$

where $\Delta P_{ji}(\mathbf{x}) = P(\omega_j|\mathbf{x}) - P(\omega_i|\mathbf{x})$ and $erf\left(\frac{\Delta P_{ji}(\mathbf{x})}{2\dot{\sigma}}\right)$ is the error function, defined as

$$erf\left(\frac{\Delta P_{ji}(\mathbf{x})}{2\dot{\sigma}}\right) = \frac{1}{\sqrt{\pi}\dot{\sigma}} \int_0^{\Delta P_{ji}(\mathbf{x})} \exp^{-\frac{1}{2}\frac{\gamma^2}{2\dot{\sigma}^2}} d\gamma \tag{43}$$

Now, the event in Eq. (39) will occur with probability

$$Q_i(\mathbf{x}) = \prod_{\substack{j=1 \\ j \neq i}}^{m} Q_{ij}(\Delta P_{ji}(\mathbf{x})) \tag{44}$$

Hence, the pattern $\mathbf{x}$ will be misclassified with probability

$$Q(\mathbf{x}) = \sum_{\substack{i=1 \\ i \neq s}}^{m} Q_i(\mathbf{x}) \tag{45}$$

In fact, the additional error probability $Q(\mathbf{x})$ will be dominated by the second most probable class, which will be involved in defining the decision boundary. This can be observed by considering all the classes with very low *a posteriori* probabilities. For those, the probability $Q_j(\mathbf{x})$ will be brought to zero by the term $Q_{js}(\Delta P_{sj}(\mathbf{x}))$, which will be extremely small because of the large difference in $\Delta P_{sj}(\mathbf{x})$. Only the class ω_k whose *a posteriori* probability is comparable to $P(\omega_k|\mathbf{x})$ will contribute a non-negligible probability value, because of its small $\Delta P_{sk}(\mathbf{x})$ and negative $\Delta P_{jk}(\mathbf{x})$ with respect to all the other classes ω_j, $\forall j \neq k,s$, which will produce a multiplicative factors $Q_{kj}(\Delta P_{jk}(\mathbf{x}))$ close to unity. Hence, $Q(\mathbf{x})$ will effectively be determined by $Q_{ks}(\Delta P_{sk}(\mathbf{x}))$.

The average additional (over and above the Bayes error) misclassification error will then be

$$\bar{e} = \int Q(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{46}$$

Recalling Eq. (42), each probability $Q_{ij}(\Delta P_{ji}(\mathbf{x}))$ in Eq. (44) depends heavily upon the variance of the error of the *a posteriori* class probability estimate. With the number of multiple experts increasing, the estimate variance goes down by a factor of N. However, the probability of the additional error goes down much more dramatically. In comparison with a single expert $N = 1$, the probability of the pointwise error, assuming that only $P(\omega_s|\mathbf{x})$ and $P(\omega_k|\mathbf{x})$ are comparable, will be reduced by a factor

$$\frac{1 - erf\left(\frac{\Delta P_{sk}(\mathbf{x})}{2\dot{\sigma}}\right)}{1 - erf\left(\frac{\Delta P_{sk}(\mathbf{x})}{2\frac{\dot{\sigma}}{N}}\right)} \tag{47}$$

Note that these improvements are achieved only near the decision boundaries, as far from the boundaries the probability of a pattern $\mathbf{x}$ being misclassified is negligible. Thus these impressive improvements will be diluted by the averaging process in Eq. (46), where over extensive regions the local probability of additional error will effectively be zero, because of the large difference between the maximum class *a posteriori* probability and all the others.

For discriminant function classifiers the benefit of combining multiple experts using an identical representation has been investigated by Tumer and Ghosh

[31,32]. They showed that the classifications error will be reduced as a result of the effective discriminant function of the combiner being closer to the Bayesian decision boundary. An earlier study of the effect of combining multiple experts which base their decisions on their estimates of the class *a posteriori* probabilities can be found elsewhere [33,34].

A linear combiner of classifier outputs has been applied to the problem of combining evidence in an automatic personal identity verification system [25]. The system fuses multiple instances of biometric data to improve performance. In this application, a single classifier computes *a posteriori* class probabilities for several instances of input data over a short period of time, which are then combined. For this reason, an equal weight combination was appropriate. A combination strategy involving unequal weights has been used [35] to fuse the *a posteriori* class probabilities of several classifiers employed in the detection of microcalcifications in mammographic images. The weights were estimated by training. The combination of classifiers which produce statistically dependent outputs is discussed in Bishop [33]. The approach also leads to a linear combination, where the weights reflect the correlations between individual expert outputs.

4. DISCUSSION

In practical situations, one is also likely to face a problem where a part of the representation used by the respective experts is shared and a part is distinct.

Let us assume that the components of each pattern vector $\mathbf{x}_i$ can be divided into two groups, forming vectors $\mathbf{y}$ and ξ_i, i.e. $\mathbf{x}_i = [\mathbf{y}^T, \xi_i^T]^T$, where the vector of measurements $\mathbf{y}$ is shared by all of the R classifiers, whereas ξ_i is specific to the i-th classifier. We shall assume that given a class identity, the classifier specific part of the pattern representation ξ_i is conditionally independent from ξ_j $j \neq i$.

Let us now return to the joint probability density $p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k)$ in Eq. (3), and express it as

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k) = p(\xi_1,\ldots,\xi_R|\mathbf{y},\theta = \omega_k)p(\mathbf{y}|\theta = \omega_k) \quad (48)$$

Recalling our assumption that the classifier specific representations ξ_i $i = 1,\ldots,R$ are conditionally statistically independent, we can write

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k) = [\Pi_{i=1}^R p(\xi_i|\mathbf{y},\theta = \omega_k)]\, p(\mathbf{y}|\theta = \omega_k) \quad (49)$$

which, assuming that the shared measurements are conditionally independent from the classifier specific ones can be expressed as

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k) = \left[\Pi_{i=1}^R \frac{P(\theta = \omega_k|\mathbf{y},\xi_i)p(\mathbf{y},\xi_i)}{P(\omega_k|\mathbf{y})p(\mathbf{y})}\right] \frac{P(\omega_k|\mathbf{y})p(\mathbf{y})}{P(\omega_k)} \quad (50)$$

and finally,

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_R|\theta = \omega_k) = \left[\Pi_{i=1}^R \frac{P(\theta = \omega_k|\mathbf{x}_i)p(\mathbf{x}_i)}{P(\omega_k|\mathbf{y})p(\mathbf{y})}\right] \frac{P(\omega_k|\mathbf{y})p(\mathbf{y})}{P(\omega_k)} \quad (51)$$

In Eq. (51), $P(\omega_k|\mathbf{y})$ is the k-th class probability based on the shared freatures, and $p(\mathbf{y})$ is the corresponding mixture measurement density. We thus obtain the decision rule

$$\text{assign } \theta \rightarrow \omega_j \text{ if } \left[\Pi_{i=1}^R \frac{P(\theta = \omega_j|\mathbf{x}_i)}{P(\theta = \omega_j|\mathbf{y})}p(\mathbf{x}_i)\right] P(\theta = \omega_j|\mathbf{y})p(\mathbf{y}) = max_{k=1}^m \left[\Pi_{i=1}^R \frac{P(\theta = \omega_k|\mathbf{x}_i)}{P(\theta = \omega_k|\mathbf{y})}p(\mathbf{x}_i)\right] P(\theta = \omega_k|\mathbf{y})p(\mathbf{y}) \quad (52)$$

in which $p(\mathbf{y})$ in the denominator was cancelled out on the grounds that the numerator term $p(\mathbf{y})$ serves as an outlier indicator adequately. The rule combines the individual classifier outputs in terms of a product. Each factor in the product for class ω_k is normalised by the *a posteriori* probability of the class given the shared representation.

A linearisation of the product in Eq. (52) using the methodology introduced in Section 2 yields the corresponding weighted sum rule [35]

$$\text{assign } \theta \rightarrow \omega_j \text{ if } w_y P(\theta = \omega_j|\mathbf{y}) + \sum_{i=1}^R w_i P(\theta = \omega_j|\mathbf{x}_i) = max_{k=1}^m \left[w_y P(\theta = \omega_k|\mathbf{y}) + \sum_{i=1}^R w_i P(\theta = \omega_k|\mathbf{x}_i)\right] \quad (53)$$

Note that the classifier combination rules (52) and (53) are expressed in terms of the *a posteriori* class probabilities returned by the individual classifiers using mixed representations and the *a posteriori* class probability based on the shared representation. Each classifier provides an independent estimate of the latter. It is therefore sensible to average these values to obtain a more reliable estimate, as discussed in Section 3. This problem has been considered by Kittler et al [36], and the combination strategies developed have

been applied to the problem of automatic detection of microcalcifications in digital mammograms.

The combination strategies discussed in Sections 2 and 3 can be viewed as a multistage process, whereby the input data is used to compute the relevant *a posteriori* class probabilities which, in turn, are used as features in the next processing stage. The problem is then to find class separating surfaces in this new feature space. The *sum rule* and the *averaging* estimator and their weighted versions then implement linear separating boundaries in this space. The other combination strategies implement nonlinear boundaries. The idea can then be extended further, and the problem of combination posed as one of training the second stage using these probabilities so as to minimise the recognition error. This is the approach adopted by various multistage combination strategies as exemplified by the behaviour knowledge space method of Huang and Suen [37] and the techniques in [20,21]. In the behaviour knowledge space method, the space of the classifier outputs is tessellated into small bins, and the computed *a posteriori* class probabilities are used as indices to address these bins. The training data is mapped into these cells via the *a posteriori* class probabilities and their true class labels stored. A pattern of unknown class membership is then classified by indexing into one of the bins, and identifying the class which receives the majority vote.

When linear or nonlinear combination functions are acquired by means of training, there is very little distinction between the two basic scenarios. Moreover, such solutions are able to handle the fusion of measurements which are not conditionally statistically independent. Consequently, it is possible to view classifier combination in a unified way. This probably explains the successes achieved with heuristic combination schemes derived without any serious concerns about their theoretical legitimacy.

5. CONCLUSIONS

The problem of combining classifiers was considered. Recent developments in the methodology of multiple expert fusion were reviewed. The review was organised according to the two main fusion scenarios: fusion of opinions based on identical, and on distinct representations. A theoretical framework for classifier combination approaches for these two scenarios was then developed. For multiple experts using distinct representations, we argued that many existing schemes could be considered as special cases of compound classification, where all the representations are used jointly to make a decision. Under different assumptions and using different approximations, we derived the commonly used classifier combination schemes such as the product rule, sum rule, min rule, max rule, median rule and majority voting, and weighted combination schemes. We addressed the issue of the sensitivity of various combination rules to estimation errors, and pointed out that the techniques based on the benevolent sum-rule fusion are more resilient to errors than those derived from the severe product rule.

We then considered the effect of classifier combination in the case of multiple experts using a shared representation. We showed that here the aim of fusion was to obtain a better estimation of the appropriate *a posteriori* class probabilities. This can be achieved by the means of estimation-error variance reduction. We also showed that the two theoretical frameworks for the case of distinct and shared representations, respectively, could also be used for devising fusion strategies when the individual experts use features some of which are shared, and the remaining ones distinct.

We showed that in both cases (distinct and shared representations), the expert fusion involves the computation of a linear or nonlinear function of the *a posteriori* class probabilities estimated by the individual experts. Classifier combination can therefore be viewed as a multistage classification process, whereby the *a posteriori* class probabilities generated by the individual classifiers are considered as features for a second stage classification scheme. Most importantly, when the linear or nonlinear combination functions are obtained by training, the distinctions between the two scenarios fade away, and one can view classifier fusion in a unified way. This probably explains the success of many heuristic combination strategies that have been suggested in the literature without any concerns about the underlying theory.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council Grant GR/J89255.

References

1. Pudil P, Novovicova J, Blaha S, Kittler J. Multistage pattern recognition with reject option. Proceedings from the 11th IAPR International Conference on Pattern Recognition, Volume II, Conference B: Pattern Recognition Methodology and Systems 1992; 92–95
2. El-Shishini H, Abdel-Mottaleb MS, El-Raey M, Shoukry A. A multistage algorithm for fast classification of patterns. Pattern Recognition Letters 1989; 10(4): 211–215
3. Zhou JY, Pavlidis T. Discrimination of characters by a multistage recognition process. Pattern Recognition 1994; 27(11): 1539–1549

4. Kurzynski MW. On the identity of optimal strategies for multistage classifiers. Pattern Recognition Letters 1989; 10(1): 36–46
5. Fairhurst MC, Abdel Wahab HMS. An interactive two-level architecture for a memory network pattern classifier. Pattern Recognition Letters 1990; 11(8): 537–540
6. Denisov DA, Dudkin AK. Model-based chromosome recognition via hypotheses construction/verification. Pattern Recognition Letters 1994; 15(2): 299–307
7. Kimura F, Shridhar M. Handwritten numerical recognition based on multiple algorithms. Pattern Recognition 1991; 24(10): 969–983
8. Tung CH, Lee HJ, Tsai JY. Multi-stage pre-candidate selection in handwritten Chinese character recognition systems. Pattern Recognition 1994; **27**(8): 1093–1102
9. Skurichina M, Duin RPW. Stabilizing classifiers for very small sample sizes. Proceedings 11th IAPR International Conference Pattern Recognition, Vienna, 1996
10. Franke J, Mandler E. A comparison of two approaches for combining the votes of cooperating classifiers. Proceedings 11th IAPR International Conference on Pattern Recognition, Volume II, Conference B: Pattern Recognition Methodology and Systems, 1992; 611–614
11. Bagui SC, Pal NR. A multistage generalization of the rank nearest neighbor classification rule. Pattern Recognition Letters 1995; 16(6): 601–614
12. Ho TK, Hull JJ, Srihari SN. Decision combination in multiple classifier systems. IEEE Transactions PAMI 1994; 16(1): 66–75
13. Hashem S and Schmeiser B. Improving model accuracy using optimal linear combinations of trained neural networks. IEEE Transactions Neural Networks 1995; 6(3): 792–794
14. Xu L, Krzyzak A, Suen CY. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions SMC 1992; 22(3): 418–435
15. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans PAMI 1990; 12(10): 993–1001
16. Cho SB, Kim JH. Combining multiple neural networks by fuzzy integral for robust classification. IEEE Transactions Systems, Man Cybernetics 1995; 25(2): 380–384
17. Cho SB, Kim JH. Multiple network fusion using fuzzy logic. IEEE Transactions Neural Networks 1995; 6(2): 497–501
18. Rogova G. Combining the results of several neural network classifiers. Neural Networks 1994; 7(5): 777–781
19. Tresp V, Taniguchi M. Combining estimators using non-constant weighting functions. In Advances in Neural Information Processing Systems 7, Tesauro G, Touretzky DS, Leen TK. (eds). MIT Press, 1995
20. Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. In Advances in Neural Information Processing Systems 7, Gesauro G, Touretzky DS, Leen TK. (eds). MIT Press, 1995
21. Wolpert DH. Stacked generalization. Neural Networks 1992; 5(2): 241–260
22. Woods KS, Bowyer K, Kergelmeyer WP. Combination of multiple classifiers using local accuracy estimates. Proceedings CVPR96 1996, 391–396
23. Kittler J. Improving recognition rates by classifier combination: A review. Proceedings IAPR 1st Int Workshop on Statistical Techniques in Pattern Recognition, Prague, 1997; 205–210
24. Ali KM, Pazzani MJ. On the link between error correlation and error reduction in decision tree ensembles. Technical Report 95-38, ICS-UCI, 1995
25. Kittler J, Matas J, Jonsson K, Ramos Sánchez MV. Combining evidence in personal identity verification systems. Pattern Recognition Letters 1997; 18: 845–852
26. Kittler J, Hatef M, Duin RPW. Combining classifiers. Proc 13th Int Conf Pattern Recognition, Volume II, Track B, Vienna, 1996; 897–901
27. Tax DMJ, Duin RPW, van Breukelen M. Comparison between product and mean classifier combination rules. Proceedings IAPR 1st Int Workshop on Statistical Techniques in Pattern Recognition, Prague, 1997; 165–170
28. Tax DMJ, Duin RPW, van Breukelen M, Kittler J. Combining multiple classifiers by averaging or multiplying. Machine Learning (submitted)
29. Ho TK. Random decision forests. Third International Conference on Document Analysis and Recognition, Montreal, Canada, August 14–16 1995; 278–282
30. Cao J, Ahmadi M, Shridhar M. Recognition of handwritten numerals with multiple feature and multistage classifier. Pattern Recognition 1995; 28(2): 153–160
31. Tumer K, Ghosh J. Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognition 1996; 29: 341–348
32. Tumer K, Ghosh J. Classifier combining: Analytical results and implications. Proceedings of the National Conference on Artificial Intelligence, Portland, OR, 1996
33. Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press, 1995
34. Kittler J. Improving recognition rates by classifier combination: A theoretical framework. In Progress in Handwriting Recognition, Downton AC, Impedovo S. (eds). World Scientific, 1997; 231–247
35. Kittler J, Hojjatoleslami A, Windeatt T. Weighting factors in multiple expert fusion. Proceedings of the British Machine Vision Conf Colchester, UK, 1997; 41–50
36. Kittler J, Hojjatoleslami A, Windeatt T. Strategies for combining classifiers employing shared and distinct pattern representations. Pattern Recognition Letters 1997 (to appear)
37. Huang TS, Suen CY. Combination of multiple experts for the recognition of unconstrained handwritten numerals. IEEE Transactions PAMI 1995; 17: 90–94

Josef Kittler graduated from the University of Cambridge in Electrical Engineering in 1971 where he also obtained his PhD in Pattern Recognition in 1974 and the ScD degree in 1991. He joined the Department of Electronic and Electrical Engineering of Surrey University in 1986 where he is a Professor, in charge of the Centre for Vision, Speech and Signal Processing.

He has worked on various theoretical aspects of pattern recognition and on many applications including automatic inspection, ECG diagnosis, remote sensing, robotics, speech recognition, and document processing. His current research interests include pattern recognition, image processing and computer vision.

He has co-authored a book with the title *Pattern Recognition: A statistical approach*, published by Prentice-Hall. He has published more than 300 papers. His is a member of the editorial boards of *Pattern Recognition Journal, Image and Vision Computing, Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence*, and *Machine Vision and Applications*.

Correspondence and offprint requests to: J. Kittler, Centre for Vision, Speech and Signal Processing, School of Electronic Engineering, Information Technology and Mathematics, University of Surrey, Guildford GU2 5XH, UK. Email: J.Kittler@ee.surrey.ac.uk