A Framework for Classifier Fusion: Is It Still Needed?

Josef Kittler

Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, Surrey GU2 7XH, UK
J.Kittler@eim.surrey.ac.uk

Abstract. We consider the problem and issues of classifier fusion and discuss how they should be reflected in the fusion system architecture. We adopt the Bayesian viewpoint and show how this leads to classifier output moderation to compensate for sampling problems. We then discuss how the moderated outputs should be combined to reflect the prior distribution of models underlying the classifier designs. We then elaborate how the final stage of fusion should combine the complementary measurement information that might be available to different experts. This process is embodied in an overall architecture which shows why the fusion of raw expert outputs is a nonlinear function of the expert outputs and how this function can be realised as a sequence of relatively simple processes.

1 Introduction

In the two decades since the publication of the Devijver-Kittler text [1], Statistical Pattern Recognition has made significant advances. The following brief review of the progress made, serving as an introduction to the main part of this paper, is biased and idiosyncratic, presented merely to motivate the main discussion. For a detailed account of the achievements of the last twenty years the reader is referred to a recent review by Jain, Duin and Mao [30].

In statistical pattern classification the most notable progress has been made in the area of modelling probability density functions using a mixture of simple components, predominantly gaussians. The general approach is discussed in detail in [2]. Some interesting developments on the theme include the joint modelling of the background and class specific contributions to the mixture model [10] which provides useful information from the point of view of classifier design. One of the most important issues in modelling pdfs by a mixture of components is architecture selection. The problem is that the usual goodness of fit criteria (Kullback-Leibler measure, maximum likelihood) monotonically improve as the number of components increases. The best architecture therefore has to be selected using alternative measures. One possibility is to check the recognition performance achieved with the resulting model by cross validation using another set of observations. However, this does not guarantee that the model is necessarily good, not mentioning that the cross validation is costly in terms of the amount of data that has

[0] Pierre Devijver Award Lecture 2000

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 45–56, 2000.

to be available. A more promising idea that has received a lot of attention is based on the minimum complexity principle (Ockham's rasor). Accordingly, the simplest model that explains the data is the best. This has led to the use of penalised goodness of fit measures. Its simplest form proposed by Akaike [3] which imposed a prior distribution over all the possible models was refined by a number of researchers making the penalty term dependent on the size of the training set [4], dimensionality and the degree of data correlation [5]. However, while these methods protect from overfitting, they do not guarantee that the model will not be underfitted. In this respect a more promising approach to architecture selection is based on the idea of model validation recently proposed by [6].

Over the period the state of the art of the methodology for classifier design has been pushed significantly by other research communities. Notable advances have been made in the area of decision trees [7], [32] and in neural networks [31]. The most recent development in machine learning, Support Vector Machines, is particularly exciting and stimulating [8].

The last decade has also witnessed considerable advances in feature selection. The popular method of optimising feature selection criteria, the *plus - l, take away - r* algorithm has been enhanced by making the numbers of forward and backward search steps, l and r , dynamic (data dependent). This computationally efficient algorithm [9] which is known as *Floating search*, has been found [12] to be the most effective suboptimal search method. Its recent further enhancements are the adaptive floating search method [14] and the oscillating search method [15].

The classical Branch and Bound search algorithm has been accelerated [16] by profiling the effect on the criterion function value of feature knock-outs. The observed profile can be used to make look ahead predictions which in turn are useful in guiding the search process into the most promising part of the search tree [16]. It has been also shown that feature selection can be performed as part of the process of data modelling using gaussian mixtures [10,13]. The effectiveness of evolutionary optimisation approaches in feature selection has been demonstrated in [18,17,19]. The use of fused classifier error as a criterion for feature selection has been suggested in [20,21].

Another area where significant advances have been made is classification in context. Conventional pattern classification involves a single object. However, objects usually do not exist in isolation. Other objects (neighbouring or otherwise) may convey contextual information that can be exploited in decision making. The motivation systematically to incorporate contextual information in the decision process led to the development of techniques which have close affinity to structural pattern recognition methods. Depending whether the classification problem is formulated as message centered (joint labelling of all the objects) or object centered (labelling of a single object using context) the classification problem leads to either graph matching, or probabilistic relaxation. The Bayesian framework for the latter has been developed in [22] and extended to handle relational information in [23,24].

Last but not least, one of the most exciting directions of the last ten years has been classifier fusion. It has been recognised for some time that the classical approach to designing a pattern recognition system which focuses on finding the best classifier has a serious drawback. Any complementary discriminatory information that other classifiers

may encapsulate is not tapped. Multiple expert fusion aims to make use of many different designs to improve the classification performance. Over the last few years a myriad of methods for fusing the outputs of multiple classifiers have been proposed [34,35,36,37, 38,39,40,41,42,43,44]. The methods range from simple Bayesian estimations methods, through trainable multistage strategies where the outputs of component classifiers are considered as features and the fusion is performed by another classifier designed using independent data, to data dependent methods where each classifier has a domain of superior competence and their opinion is called on only when the observation falls into this domain.

Notwithstanding all these advances it is pertinent to question their significance in view of the developments in Support Vector Machines [8]. This novel approach to training classifiers by minimising the structural risk enables the designer to position the class separating boundaries carefully so as to reduce the chances of misclassifying new patterns. The optimal positioning of the boundaries can be achieved for a given training set in the space of any dimensionality without feature selection. Surprisingly, it would appear that one can generate an arbitrary number of additional dimensions (features) without risking overfitting. This facilitates the construction of more effective, nonlinear boundaries between classes without compromising the ability to generalise. One could then make the inference that the design methodology should, at least in principle, lead to classifiers that capture all the discriminatory information in a single design. Seeking and fusing multiple opinions should thus be unnecessary.

The above argument, in one bold sweep, would make the classical pattern recognition system model and all the achievements of the last two decades out of date. When put to test in the context of an application, concerned with personal identity verification [27,28], the results were interesting, but not one sided. The problem was to verify the claimed identity of probe face images and for this purpose we used SVMs, simple Euclidean distance and normalised correlation classifiers. The experiments were performed in eigen face and fisher face spaces under various photometric normalisations. Interestingly, SVMs were able extract the relevant discriminatory information even from the eigenface representation, regardless of the sophistication of the photometric preprocessing or lack of it. However, once the discriminatory information was extracted by the traditional means of discriminant analysis, and the data suitably normalised, very simple classifiers outperformed the powerful SVMs. In fact SVMs did not benefit from these preprocessing steps at all.

There are three important conclusions that can be drawn from the above findings:

- SVMs are very effective in extracting discriminatory information from any representation and can successfully cope with complex intraclass variations
- SVMs designs are not guaranteed to be superior to other carefully designed classifiers and therefore one can argue that fusion is still relevant
- The use of knowledge about the problem domain and the data for the architecture selection is crucial to achieving successful designs

With the rationale for continued interest in classifier fusion re-established it is pertinent to ask whether classifier fusion systems should be designed using powerful machine learning methods, or whether a careful consideration to the issue of fusion architecture selection should be given in view of the third conclusion. In this paper we advocate the

latter. This point has already been argued in the context of neural net classifier design in [33]. In this paper we consider the problem and issues of fusion and discuss how they should be reflected in the fusion system architecture. We adopt the Bayesian viewpoint and show how this leads to classifier output moderation to compensate for sampling problems. We then discuss how the moderated outputs should be combined to reflect the prior distribution of the models underlying the classifier designs. The final stage of fusion combines the complementary measurement information that may be available to different experts. This process is embodied in an overall architecture which shows why the fusion of raw expert outputs is a nonlinear function and how this function can be realised as a sequence of relatively simple processes.

The paper is organised as follows. In the next section we introduce the theoretical model underpinning classifier fusion. The effect of averaging over classifier models is discussed in Section 3. Classifier output moderation is the subject of discussion in Section 4. The resulting architecture is described in Section 5. Section 6 draws the paper to conclusion.

2 Theoretical Framework

Consider a pattern recognition problem where pattern Z is to be assigned to one of the m possible classes $\{\omega_1, ..., \omega_m\}$. Let us assume that we make R vector observations $\mathbf{x}_i \;\; i = 1, .., R$ on the given pattern and that the i-th measurement vector is the input to the i-th expert modality. We shall assume that these observations are provided by different sensors, or perhaps by the same sensor over a period of time. The different sensors can be either physical or logical. When the measurements are acquired by different physical sensors it is reasonable to assume that they will be conditionally statistically independent. However, for logical sensors we may not be able to make such a strong assumption. Indeed, there may often be the case that some of the components of the measurement vectors will be highly correlated or even identical copies. This could happen if the measurement vectors $\mathbf{x}_i \;\; i = 1, .., R$ are formed from a larger pool of features by a selection with replacement which makes the features available for other classifier input vectors. This construction could result in some of the features to be shared. Logical sensors, of course, could also generate features that are weakly correlated. However, we shall not consider all the possible scenarios. Instead, we shall make the assumption that the components of one vector are either statistically conditionally independent from those of another, or they are exact replicas.

In principle, vectors $\mathbf{x}_i$ could share different numbers of features. However, once again we shall not consider such complications as this would make our notation unnecessarily complex. The analysis to be presented could easily be extended to any specific practical situation, if desired. Thus for the sake of simplicity, and without any loss of generality, we shall 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 the R modalities whereas ξ_i is specific to the i-th modality. We shall assume that given a class identity, the modality specific part of the pattern representation ξ_i is conditionally independent from $\xi_j \;\; j \neq i$.

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 by $P(\omega_k)$. We shall consider the models to be mutually exclusive which means that only one model can be associated with each pattern.

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

$$assign\ \theta \rightarrow \omega_j\ if \quad P(\theta = \omega_j|\mathbf{x}_1, .., \mathbf{x}_R) = \max_k P(\theta = \omega_k|\mathbf{x}_1, .., \mathbf{x}_R) \tag{1}$$

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

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

where $p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)$ and $p(\mathbf{x}_1, ..., \mathbf{x}_R)$ are the conditional and unconditional measurement joint probability densities. The latter can be expressed in terms of the conditional measurement distributions as $p(\mathbf{x}_1, ..., \mathbf{x}_R) = \sum_{j=1}^{m} p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_j)P(\omega_j)$ and therefore, in the following, we can concentrate only on the numerator terms of (2).

We commence by expressing $p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)$ as

$$p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k) = p(\xi_1, ..., \xi_R|\mathbf{y}, \theta = \omega_k)p(\mathbf{y}|\theta = \omega_k) \tag{3}$$

Recalling our assumption that the modality specific representations $\xi_i \ \ i = 1, .., R$ are conditionally statistically independent, we can write

$$p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k) = [\Pi_{i=1}^{R} p(\xi_i|\mathbf{y}, \theta = \omega_k)]p(\mathbf{y}|\theta = \omega_k) \tag{4}$$

which can further be expressed as

$$p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k) = [\Pi_{i=1}^{R} \frac{P(\theta = \omega_k|\mathbf{y}, \xi_i)p(\mathbf{y}, \xi_i)}{P(\omega_k|\mathbf{y})p(\mathbf{y})}]\frac{P(\omega_k|\mathbf{y})p(\mathbf{y})}{P(\omega_k)} \tag{5}$$

and finally

$$p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k) = [\Pi_{i=1}^{R} \frac{P(\theta = \omega_k|\mathbf{x}_i)p(\mathbf{x}_i)}{P(\omega_k|\mathbf{y})p(\mathbf{y})}]\frac{P(\omega_k|\mathbf{y})p(\mathbf{y})}{P(\omega_k)} \tag{6}$$

Let us pause to look at the meaning of the terms defining $p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)$. First of all $P(\theta = \omega_k|\mathbf{x}_i)$ is the $k - th$ class aposteriori probability computed by each of the R classifiers whereas $P(\omega_k|\mathbf{y})$ is $k - th$ class probability based on the shared features. $p(\mathbf{x}_i)$ and $p(\mathbf{y})$ are the mixture measurement densities of the representations used for decision making by each of the experts. Since the measurement densities are independent of the class labels they can be cancelled out by the normalising term in the expression for the aposteriori probability in (2) and we obtain the decision rule

$$assign \quad \theta \rightarrow \omega_j \quad if \quad [\Pi_{i=1}^{R} \frac{P(\theta = \omega_j|\mathbf{x}_i)}{P(\theta = \omega_j|\mathbf{y})}]P(\theta = \omega_j|\mathbf{y}) =$$

$$= max_{k=1}^{m}[\Pi_{i=1}^{R}\frac{P(\theta=\omega_k|\mathbf{x}_i)}{P(\theta=\omega_k|\mathbf{y})}]P(\theta=\omega_k|\mathbf{y}) \tag{7}$$

which combines the individual classifier outputs in terms of a product. Each factor in the product for class ω_k is normalised by the aposteriori probability of the class given the shared representation.

Now let us consider the ratio $\frac{P(\theta=\omega_k|\mathbf{x}_i)}{P(\theta=\omega_k|\mathbf{y})}$ and suppose it is close to one. We can then write $P(\theta=\omega_k|\mathbf{x}_i) = P(\omega_k|\mathbf{y})(1+\Delta_{ki})$. Substituting into (7) and linearising the product by expanding it and neglecting all terms of second order and higher, the decision rule becomes

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

$$= max_{k=1}^{m}[(1-R)P(\theta=\omega_k|\mathbf{y}) + \sum_{i=1}^{R} P(\theta=\omega_k|\mathbf{x}_i)] \tag{8}$$

3 Averaging over Models

The probability level fusion embodied by rules (7) and (8) involves the true aposteriori class probabilities. In practice these functions will be estimated by our experts and will be subject to errors. The design of the experts will normally be based on training data. In general we may have different training sets for each of the sensors. This will specially be true in the case when the different sensors are physical. A typical example here is the problem of person identification using biometrics involving voice characteristic, frontal face images, face profile, lip dynamics, etc. In all these cases the training data sets will be completely different.

Let us consider one such sensor, say sensor i for which the available training set is X_i. An estimate $\tilde{P}(\omega|\mathbf{x}_i)$ of $P(\omega|\mathbf{x}_i)$ that an expert can deliver will be influenced by the training set X_i. We shall represent this explicitly by writing for the estimate $\tilde{P}(\omega|\mathbf{x}_i) = P(\omega|\mathbf{x}_i, X_i)$ The design of each expert will involve the choice of a model, M, and for each model we shall estimate its parameters represented by vector γ_i. Hence an actual estimate will be conditioned on these two factors as $P(\omega|\mathbf{x}_i, X_i, M, \gamma_i)$. It follows that if we wish to obtain as good estimate as possible, we should

- consider as many models as possible, and
- minimise the influence of a particular choice of model parameters.

Mathematically this can be achieved by averaging over all the possible models and their parameters. This can be written as

$$\tilde{P}(\omega|\mathbf{x}_i) = \int\int P(\omega|\mathbf{x}_i, X_i, M, \gamma_i)p(M)p(\gamma_i)dMd\gamma_i \tag{9}$$

where $p(M)$ and $p(\gamma_i)$ are the distributions of models and model parameters respectively.

In (9) the integration over the parameter space is referred to as moderation and it will be discussed in detail in Section 4. The integral over the model space can be estimated as

$$\tilde{P}(\omega|\mathbf{x}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\int P(\omega|\mathbf{x}_i, X_i, M_j, \gamma_i)p(\gamma_i)d\gamma_i \tag{10}$$

Denoting $P(\omega|\mathbf{x}_i, X_i, M_j, \gamma_i) = \hat{P}_j(\omega|\mathbf{x}_i)$ and its integral over γ_i as

$$P_j(\omega|\mathbf{x}_i) = \int \hat{P}_j(\omega|\mathbf{x}_i)p(\gamma_i)d\gamma_i \tag{11}$$

we finally obtain an estimate of the aposteriori class probability based on sensor i as

$$\tilde{P}(\omega|\mathbf{x}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i} P_j(\omega|\mathbf{x}_i) \tag{12}$$

4 Expert Output Moderation

In Section 2 we argued for a moderation of raw expert outputs. The moderation is warranted for pragmatic reasons, namely to minimise the veto effect of overconfident erroneous classifiers. However, as has been pointed out in (11) there is also a good theoretical basis for moderating expert estimates of class aposteriori probabilities. The argument goes as follows. The estimation of the aposteriori class probability $P_j(\omega|\mathbf{x_i})$ by the j-th expert using the output $\mathbf{x_i}$ of sensor i is dependent on the training set X_i of data collected from sensor i. For a particular model which underlies the design of expert j the training data is used to estimate the model parameters γ_{ij}. Note however, that the estimated parameters cannot be considered as the true parameters of the assumed model. By taking into account the distribution of the model parameter vector it should be possible to obtain a more conservative estimate of the decision output which will reduce the risk of overfitting to a particular training set.

Mathematically, expert output $P_j(\omega|\mathbf{x_i})$ is derived by integrating parameter dependent estimates $\hat{P}_j(\omega|\mathbf{x_i}) = P_j(\omega|\mathbf{x_i}, \gamma_{ij})$ over the model parameter space as

$$P_j(\omega|\mathbf{x_i}) = \int P_j(\omega|\mathbf{x_i}, \gamma_{ij})p(\gamma_{ij})d\gamma_{ij} \tag{13}$$

Under the assumption that in the observation space the classes are distributed normally, the moderation converts the Gaussian density into Student's t distribution [25]. Bedworth [26] used this result to derive moderated posterior class probabilities for multilevel fusion and showed, on the standard UCI repository of classification problems [29], that for small training sample sets the results obtained by moderation are superior to non moderated expert output fusion.

It is perhaps true to say that a Student's t distribution converges to the corresponding Gaussian quite rapidly and for training sets of reasonable size there should not be any appreciable difference between moderated and raw expert outputs. However, for some types of classifiers, moderation is pertinent even for sample sets of respectable size. An

important case is the k-Nearest Neighbour ($k-NN$) classifier. Even if the training set is relatively large, say hundreds of samples or more, the need for moderation is determined by the value of k, which may be as low as $k = 1$. Considering just the simplest case, a two class problem, it is perfectly possible to draw all k-Nearest Neighbours from the same class which means that one of the classes will have the expert output set to zero. In the subsequent (product) fusion this will then dominate the fused output and may impose a veto on the class even if other experts are supportive of that particular hypothesis.

We shall now consider this situation in more detail. Suppose that we draw k-Nearest Neighbours and find that κ of these belong to class ω. Then the unbiased estimate $\hat{P}_j(\omega|\mathbf{x_i})$ of the aposteriori probability $P(\omega|\mathbf{x_i})$ is given by

$$\hat{P}_j(\omega|\mathbf{x_i}) = \frac{\kappa}{k} \tag{14}$$

It should be noted that the actual observation κ out of k could arise for any value of $P(\omega|\mathbf{x_i})$ with the probability

$$q(\kappa) = \binom{k}{\kappa} P^{\kappa}(\omega|\mathbf{x_i})[1 - P(\omega|\mathbf{x_i})]^{k-\kappa} \tag{15}$$

Assuming that a priori the probability $P(\omega|\mathbf{x_i})$ taking any value between zero and one is equally likely, we can find an aposteriori estimate of the aposteriori class probability $P(\omega|\mathbf{x_i})$ as

$$P_j(\omega|\mathbf{x_i}) = \frac{\int_0^1 P(\omega|\mathbf{x_i}) P^{\kappa}(\omega|\mathbf{x_i})[1 - P(\omega|\mathbf{x_i})]^{k-\kappa} dP(\omega|\mathbf{x_i})}{\int_0^1 P^{\kappa}(\omega|\mathbf{x_i})[1 - P(\omega|\mathbf{x_i})]^{k-\kappa} dP(\omega|\mathbf{x_i})} \tag{16}$$

where the denominator is a normalising factor ensuring that the total probability mass equals to one. By expanding the term $[1 - P(\omega|\mathbf{x_i})]^{k-\kappa}$ and integrating, it can be easily verified that the right hand side of (16) becomes

$$P_j(\omega|\mathbf{x_i}) = \frac{\kappa + 1}{k + 2} \tag{17}$$

which is the beta distribution. Thus the moderated equivalent of $\frac{\kappa}{k}$ is $\frac{\kappa+1}{k+2}$. Clearly our estimates of aposteriori class probabilities will never reach zero which could cause a veto effect. For instance, for the Nearest Neighbour classifier with $k = 1$ the smallest expert output will be $\frac{1}{3}$. As k increases the smallest estimate will approach zero as $\frac{1}{k+2}$ and will assume zero only when $k = \infty$.

In multiple expert fusion involving k-NN classifiers, moderation can play a very important role. This has been demonstrated in [45] where the performance of a fused system involving a product fusion rule improved dramatically.

5 Fusion System Architecture

The discussions in Sections 2, 3 and 4 lead to an architecture shown in Figure 1. The set of measurements from each physical or logical sensor is the input to a battery of classifiers deploying different models for the computation of raw expert outputs. The

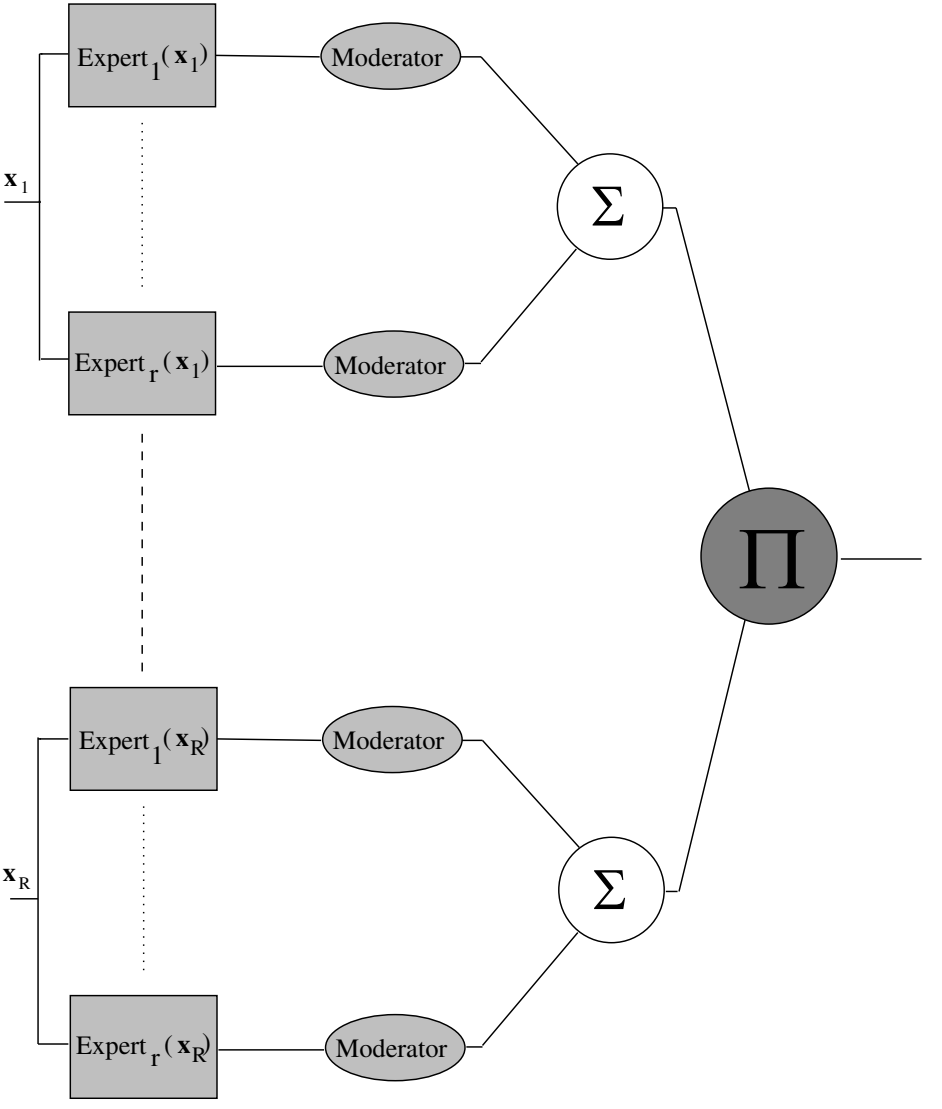

Fig. 1. Fusion system architecture

meaning of different models is understood in a broad sense, including not only different distributional models, but also variations on each of the models considered. These variations can be realised by bagging, by using different initialisations of the respective learning algorithms, choosing different classifier design parameters and architectures. Each classifier is assumed to generate the aposteriori probabilities for each of the m classes. These raw outputs are first moderated before they are combined to produce a better estimate of the class posteriors. Note that the importance of moderation will depend on

the severity of the sampling problem and the degree of averaging. In principle, the more extensive the averaging, the less important the moderation. The combined moderated expert outputs for each of the sensors are then fused to reach a final decision. The fusion can be accomplished by the product rule or sum rule as discussed in Section 2.

Note that the architecture in Figure 1 is very general. If there is only one sensor, then the final result will be obtained just by moderation and averaging. The averaging over different models can take into account the estimation errors by associating weights with each of the moderated outputs.

6 Conclusions

We argued that, in spite of the recent advances in machine learning based on the concept of Support Vectors, the conventional approaches to classifier design, including feature selection, contextual classification and classifier fusion retain their relevance. We considered the problem and issues of classifier fusion in more detail and discussed how they should be reflected in the fusion system architecture. We adopted the Bayesian viewpoint and showed how this led to classifier output moderation to compensate for sampling problems. We then discussed how the moderated outputs should be combined to reflect the prior distribution of models underlying the classifier designs. We then elaborated how the final stage of fusion should combine the complementary measurement information that might be available to different experts. This process is embodied in an overall architecture which shows why the fusion of raw expert outputs is a nonlinear function of expert outputs and how this function can be realised as a sequence of relatively simple processes.

Acknowledgements

The support of EPSRC Grant GR/M61320 and EU Framework V Project Banca is gratefully acknowledged.

References

1. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice-Hall, Englewood Cliffs, N.J. (1982).
2. Titterington, D., Smith, A., Makov, U.: Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, Chichester (1985).
3. Akaike, H.: A New Look at Statistical Model Identification. *IEEE Trans. Automatic Control* **19** (1994) 716-723.
4. Schwarz, G.: Estimating the Dimension of a Model. *The Annals of Statistics* **6** (1978) 461-464.
5. Sardo, L., Kittler, J.: Minimum Complexity Estimator for RBF Networks Architecture Selection. *Proc. International Conference on Neural Networks*, Washington (1996) 137-142.
6. Sardo, L., Kittler, J.: Model Complexity Validation for PDF Estimation Using Gaussian Mixtures. *Proc. 14th International Conference on Pattern Recognition*, Brisbane (1998) 195-197.
7. Quinlen, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, Calif. (1993).

8. Vapnik, V.N.: The Nature of Statistical Learning Theory. John Wiley, New York (1998).
9. Pudil, P., Novovičová, J., Kittler, J.: Floating Search Methods in Feature Selection. *Pattern Recognition Letters* **15** (1994) 1119–1125.
10. Pudil, P., Novovičová, J., Choakjarerwanit, N., Kittler, J.: Feature selection based on the approximation of class densities by finite mixtures of special type. *Pattern Recognition* **28** (1995) 1389-1397.
11. Pudil, P., Novovičová, J.: Novel Methods for Subset Selection with Respect to Problem Knowledge. *IEEE Transactions on Intelligent Systems - Special Issue on Feature Transformation and Subset Selection* (1998) 66-74.
12. Jain, A.K., Zongker, D.: Feature Selection: Evaluation, Application and Small Sample Performance. *IEEE Transactions on PAMI* **19** (1997) 153-158.
13. Novovičová, J., Pudil, P., Kittler, J.: Divergence based feature selection for multimodal class densities. *IEEE Transactions on PAMI*, **18** (1996) 218-223.
14. Somol, P., Pudil, P., Novovičová J., Paclik, P.: Adaptive floating search methods in feature selection. *Pattern Recognition Letters* **20** (1999) 1157-1163.
15. Somol, P., Pudil, P.: Oscillating Search Algorithms For Feature Selection. *Proc. 15th IAPR International Conference on Pattern Recognition*, Barcelona (2000).
16. Somol, P., Pudil, P., Ferri, F.J., Kittler, J.: Fast Branch and Bound Algorithm For Feature Selection. *Proc 4th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida (2000).
17. Mayer, H.A., Somol, P., Pudil, P., Grim, J., Huber R., Schwaiger, R.: A Comparison of Deterministic and Non-Deterministic Feature Selection Algorithms for k-NN, Gaussian, and Neural Classifiers. *Proc. 4th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida (2000).
18. Ferri, F.J., Kadirkamanathan, V., Kittler, J.: Feature Subset Search Using Genetic Algorithms. *Proc IEE Workshop on Natural Algorithms in Signal Processing* (1993) 23/1-23/7.
19. Mayer, H.A., Somol, P., Huber, R., Pudil, P.: Improving Statistical Measures of Feature Subsets by Conventional and Evolutionary Approaches. *Proc. 3rd IAPR International Workshop on Statistical Techniques in Pattern Recognition*, Alicante, (2000).
20. Alkoot F.M., Kittler, J.: Multiple Expert System Design by Combined Feature Selection and Probability Level Fusion. *Proc. Conf. Fusion 2000* , Paris (2000).
21. Alkoot F.M., Kittler, J.: Feature Selection for an Ensemble of Classifiers. *Proc. 4th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida (2000).
22. Kittler, J., Hancock, E.R.: Combining Evidence in Probabilistic Relaxation. *International Journal of Pattern Recognition and Artificial Intelligence*, **3** (1989) 29-51.
23. Christmas, W.J., Kittler, J., Petrou, M.: Structural Matching in Computer Vision Using Probabilistic Relaxation. *IEEE Trans Pattern Analysis and Machine Intelligence*, **17** (1995) 749-764.
24. Kittler, J.: Probabilistic Relaxation and the Hough Transform. *Pattern Recognition* **33** (2000) 705-714.
25. Shanmugan, K.S., Breipohl, A.M.: Random Signals: Detection, Estimation and Data Analysis. Wiley, New York (1988).
26. Bedworth, M.: High level data fusion. PhD Thesis, Aston University, United Kingdom (1999).
27. Jonsson, K., Kittler, J., Li Y.P., Matas, J.: Support Vector Machine for Face Authentication. In *Proceeding of BMVC'99* (1999) 543-553.
28. Jonsson, K., Kittler, J., Matas, J.: Learning Support Vectors for Face Authentication: Sensitivity to Mis-Registrations. *Proceeding of ACCV'00*, Taipei (2000) 806-811.
29. Murphy, P.: Repository of machine learning databases and domain theories. ftp://ftp.ics.uci.edu/pub/machine-learning-databases (1999).
30. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **PAMI-22** (2000) 4-37.

31. Bishop, C.J., Neural networks for pattern recognition. Clarendon Press, Oxford (1995).
32. Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J.: Classification and Regression Trees, Wadsworth, California (1984).
33. Christmas, W.J., Kittler, J., Petrou, M.: Analytical Approaches to Neural Network Design. *in Multiple Paradigms, Comparative Studies and Hybrid Systems*, eds E S Gelsema, and L N Kanal, North Holland (1994) 325-335.
34. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence* **20** (1998) 226-239.
35. Kittler, J.: Combining Classifiers: A Theoretical Framework. *Pattern Analysis and Applications* **1** (1998) 18-27.
36. Fairhurst, M.C., Abdel Wahab, H.M.S: An interactive two-level architecture for a memory network pattern classifier. *Pattern Recognition Letters* **11** (1990) 537-540.
37. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **16** (1994) 66-75.
38. Wolpert, D.H.: Stacked generalization. *Neural Networks* **5** (1992) 241-260.
39. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. SMC* **22** (1992) 418-435.
40. Kittler, J., Matas, J., Jonsson, K., Ramos Sánchez, M.U.: Combining evidence in personal identity verification systems. *Pattern Recognition Letters* **18** (1997) 845-852.
41. Tumer, K., Ghosh, J.: Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. *Pattern Recognition*, **29** (1996) 341-348.
42. Woods, K.S., Bowyer, K., Kergelmeyer, W.P.: Combination of multiple classifiers using local accuracy estimates. *Proc. of CVPR96* (1996), 391-396.
43. Kittler, J., Hojjatoleslami, A., Windeatt, T.: Strategies for combining classifiers employing shared and distinct pattern representations. *Pattern Recognition Letters* **18** (1997) 1373-1377.
44. Huang, T.S., Suen, C.Y.: Combination of multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans Pattern Analysis and Machine Intelligence* **17** (1995) 90-94.
45. Alkoot, F.M., Kittler, J.: Improving the performance of the product fusion strategy. *Proc. 15th IAPR International Conference on Pattern Recognition*, Barcelona (2000).