Multiple Classifier System Approach to Model Pruning in Object Recognition

Josef Kittler and Ali R. Ahmadyfard

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford GU2 7XH, United Kingdom

Abstract. We propose a multiple classifier system approach to object recognition in computer vision. The aim of the approach is to use multiple experts successively to prune the list of candidate hypotheses that have to be considered for object interpretation. The experts are organised in a serial architecture, with the later stages of the system dealing with a monotonically decreasing number of models. We develop a theoretical model which underpins this approach to object recognition and show how it relates to various heuristic design strategies advocated in the literature. The merits of the advocated approach are then demonstrated experimentally using the SOIL database. We show how the overall performance of a two stage object recognition system, designed using the proposed methodology, improves. The improvement is achieved in spite of using a weak recogniser for the first (pruning) stage. The effects of different pruning strategies are demonstrated.

1 Introduction

There are several papers [4,14,15,16,17,8,11,9] concerned with multiple classifier system architectures suggesting that complex architectures, in which the decision process is decomposed into several stages involving coarse to fine classification, result in improved recognition performance. In particular, by grouping classes and performing initially coarse classification, followed by a fine classification refinement which disambiguates the classes of the winning coarse group, one can achieve significant gains in performance. [9] applies this approach to the problem of handwritten character recognition and suggests that class grouping should maximise an entropy measure. Similar strategies have been advocated in [4,14,15,16,17]. The popular decision tree methods can be seen to exploit the same phenomenon.

The aim of this paper is to demonstrate that these heuristic processes do have a theoretical foundation. We propose a framework for analysing the benefit of hierarchical class grouping. Using this framework we develop a theoretical basis for multiple expert fusion in serial coarse to fine object recognition system architectures. The analysis will suggest and explain a number of strategies that can be adopted to build such architectures.

We apply the proposed design methodology to the problem of 3D object recognition using 2D views. This problem has been receiving a lot of attention over the last two decades, resulting in a spectrum of techniques which exploit, for instance, colour [20, 5,12], shape [2,19] and object appearance [18,10,13]. Although none of the existing

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 342–353, 2004.

methods provide a panacea on their own, we argue that a combination of several object recognition techniques can be very effective.

More specifically we demonstrate that the proposed approach accomplishes a sequential pruning of the list of object model hypotheses, with the later stages of the system having to deal with a monotonically decreasing number of models. The merits of the advocated approach are then demonstrated experimentally using the SOIL database. We show, how the overall performance of a two stage object recognition system based on the expounded principles improves. The improvement is achieved in spite of using a weak recogniser for the first (pruning) stage. The effects of different pruning strategies are demonstrated.

The paper is organised as follows. In Section 2 the problem of object recognition using hierarchical class grouping is formulated. We derive an expression for the additional decision error, over and above the Bayes error, as a function of estimation error. In Section 3 we discuss various model pruning strategies that naturally stem from this analysis. In Section 4 one of these strategies is applied to the problem of 3D object recognition using a two stage decision making system. Section 5 draws the paper to conclusion.

2 Mathematical Notation and Problem Formulation

Consider an object recognition problem where object Z is to be assigned to one of m possible models $\{\omega_i,\ i = 1, ...m\}$. Let us assume that the given scene object is represented by a measurement vector, $\mathbf{x}$. In the measurement space each object category ω_k is modelled by the probability density function $p(\mathbf{x}|\omega_k)$ and let the a priori probability of object occurrence be 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 instance.

Now according to the Bayesian decision theory, given measurements $\mathbf{x}$, the instance, Z, should be assigned to model 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.

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$P(\theta = \omega_j|\mathbf{x}) = \max_k P(\theta = \omega_k|\mathbf{x}) \tag{1}$$

In practice, for each interpretation, a decision making system will provide only an estimate $\hat{P}(\omega_i|\mathbf{x})$ of the true aposteriori class probability $P(\omega_i|\mathbf{x})$ given measurement $\mathbf{x}$, rather than the true probability itself. Let us denote the error on the estimate of the ith model class aposteriori probability at point $\mathbf{x}$ as $e(\omega_i|\mathbf{x})$ and let the probability distribution of errors be $p_i[e(\omega_i|\mathbf{x})]$. Clearly, due to estimation errors, the object recognition based on the estimated aposteriori probabilities will not necessarily be Bayes optimal. In the appendix we derive the probability, $e_S(\mathbf{x})$ of the decision relating to object $\mathbf{x}$ being suboptimal, and refer to it as the switching error probability. We shown in (14) that this probability primarily depends on the margin $\Delta P_{si}(\mathbf{x}) = P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})$ between the aposteriori probabilities of the Bayes optimal hypothesis ω_s and the next most probable model ω_i, as well as on the width (variance) of the distribution of estimation error.

Now how do these labelling errors translate to recognition error probabilities? We know that for the Bayes minimum error decision rule the error probability at point $\mathbf{x}$ will be $e_B(\mathbf{x})$. If our pseudo Bayesian decision rule, i.e. the rule that assigns patterns according to the maximum estimated aposteriori class probability, deviates from the Bayesian rule with probability $e_S(\mathbf{x})$, the local error of the decision rule will be given by

$$\alpha(\mathbf{x}) = e_B(\mathbf{x})[1 - e_S(\mathbf{x})] + e_S(\mathbf{x})[1 - e_B(\mathbf{x})] \tag{2}$$

The error, $\alpha(\mathbf{x})$, will be close to Bayesian only if $e_S(\mathbf{x})$ is negligible. Thus we want the label switching error to be as small as possible.

Conventionally, the multiple classifier fusion paradigm attempts to ameliorate the switching error probability by reducing the variance of estimation errors. This is achieved by combining multiple estimates obtained by a number of diverse object recognition experts. In this paper we adopt a completely different approach that strives to increase the margin between the posteriors of the competing model hypotheses in order to reduce the error probability $e_S(\mathbf{x})$ by alternative means. The basic idea is to group models into superclasses in such a way that the margin between the posteriors of the resulting model sets widens. The number of groups is a free parameter. For our purposes we divide the classes into two groups and perform a coarse classification of the input pattern to one of these two groups. Then, in the next stage, we refine the classification and continue dividing the the most probable super class in the two subsets by considering the remaining alternatives.

In general, there will be m hypotheses that can be grouped hierarchically into two groups at each stage of the hierarchy. Let us denote the two groups created at stage k by Ω^k and $\bar{\Omega}^k$. The set Ω^k will be divided in the next stage into two subsets, and so on. Thus the class sets Ω^k will satisfy

$$\Omega^k \epsilon \Omega^j \quad j < k \tag{3}$$

Further, let us denote the probability of classifying measurement vector $\mathbf{x}$ from superclass Ω^k suboptimally by $w^k(\mathbf{x})$. Referring to (14), in this two (super)class case, the switching error probability at stage k is given simply by

$$w^k(\mathbf{x}) = \int_{\Delta P_{\Omega^k}(\mathbf{x})}^{\infty} p[\eta_{\Omega^k}(\mathbf{x})] d\eta_{\Omega^k}(\mathbf{x}) \tag{4}$$

where $\Delta P_{\Omega^k}(\mathbf{x})$ is the margin between the posteriors of thw two super classes at stage k and $\eta_{\Omega^k}(\mathbf{x})$ is the associated estimation error. Assuming that the Bayes optimal hypothesis is contained in set Ω^k, it will end up in superclass Ω^{k+1} with probability $1 - w^k(\mathbf{x})$ Similarly, at the $(k+1)^{st}$ stage the probability of making a suboptimal decision is $w^{k+1}(\mathbf{x})$, while the Bayes optimal decision will be made with probability $1 - w^{k+1}(\mathbf{x})$. The complete n-stage hypothesis refinement process is illustrated in Figure 1. By reference to Figure 1 the total switching error probability of the hierarchical decision making process can be written as

$$\begin{aligned} e_S(\mathbf{x}) = w^1(\mathbf{x}) + \sum_{i=2}^{n-1} [\Pi_{j=1}^{i-1}(1 - w^j(\mathbf{x}))] w^i(\mathbf{x}) \\ + [\Pi_{j=1}^{n-1}(1 - w^j(\mathbf{x}))] w^n(\mathbf{x}) \end{aligned} \tag{5}$$

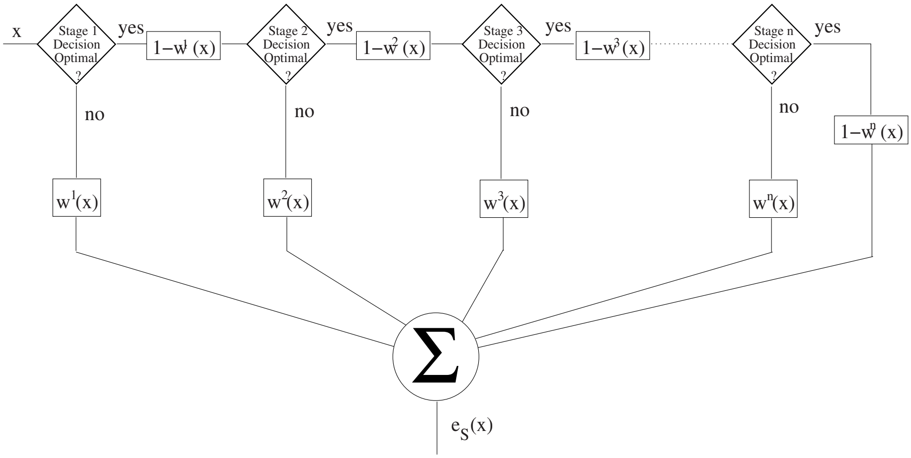

Fig. 1. The total probability of label switching, $e_S(\mathbf{x})$ in a coarse to fine multistage object recognition system

Note that the final stage will involve only the closest competitors, classes $\Omega^n = \{\omega_s, \omega_i\}$. The probability $w^n(\mathbf{x})$ of label switching will be given by

$$w^n(\mathbf{x}) = \int_{\Delta Q_{si}(\mathbf{x})}^{\infty} p[\eta_\omega(\mathbf{x})] d\eta_\omega(\mathbf{x}) \tag{6}$$

where $\eta_\omega(\mathbf{x}) = 2e(\omega_i|\mathbf{x})$ is the combined estimation error for the two posteriors, since in the two class $e(\omega_s|\mathbf{x}) = -e(\omega_i|\mathbf{x})$.

In equation (6) we denote the aposteriori probabilities, $Q(\omega_r|\mathbf{x}),\ r = s, i$ for model classes ω_s and ω_i by different symbols to indicate that these functions differ from $P(\omega_r|\mathbf{x}),\ r = s, i$ by a scaling factor $P(\omega_s|\mathbf{x}) + P(\omega_i|\mathbf{x})$ since they have to sum up to one. Note that if functions $Q(\omega_r|\mathbf{x})$ are estimated via probability densities, the estimation errors will be scaled up versions of the original errors $e(\omega_r|\mathbf{x})$. However, if these functions are estimated directly from the training data, the errors will be different and can be assumed to have the same distribution as the original unscaled errors $e(\omega_r|\mathbf{x})$. If this is the case, then one can see why this two stage approach may produce better results. The probability mass under the tail of the error estimation distribution will rapidly decay as the margin (the tail cut off) increases. If the error distributions are the same but the margins increase by scaling, the probability of label switching will go down.

3 Discussion

Let us consider the implication of expression (5). Assuming that the estimation errors have identical distribution at all the stages of the sequential decision making process, the label switching error $w^i(\mathbf{x})$ at stage i will be determined entirely by the margin (difference) between the aposteriori probabilities of classes $P(\Omega^i|(\mathbf{x}))$ and $P(\bar{\Omega}^i|(\mathbf{x}))$.

By grouping model classes at the top of the hierarchy we can increase this margin and therefore control the additional error. In this way we can ensure that the additional errors $w^i(\mathbf{x})$ in all but the last stage of the decision making process are negligible. In the limiting case, when $w^i(\mathbf{x}) \to 0, \quad i = 1, ..., n-1$ the switching error $e_S(\mathbf{x})$ will be equal to $w^n(\mathbf{x})$. At that point the set Ω^n is likely to contain just a single class. Thus the last stage decision will involve two classes only. Note that whereas the margin between the aposteriori probabilities of the two model classes, say ω_s and ω_i, at the top of the hierarchy, was $P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})$, in the last stage, it will become

$$\delta = \frac{P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})}{P(\omega_s|\mathbf{x}) + P(\omega_i|\mathbf{x})} \tag{7}$$

Thus the margin will be significantly magnified and consequently the additional error $e_S(\mathbf{x})$ significantly lower than what it would have been in a single stage system.

The expression (5) immediately suggests a number of grouping strategies. For instance, in order to maintain the margin as large as possible in all stages of the decision making process it would clearly be most effective to group all but one class in one super class and the weakest class in the complement super class. This strategy has been suggested, based on heuristic arguments, in [21]. The disadvantage of this strategy is that it would involve $m-1$ decision steps.

Computationally more effective is to arrive at a decision after $log_2 m$ steps. This would lead to grouping which maintains a balance of the two class sets Ω^i and $\bar{\Omega}^i$. Another suggestion [9] is to split the classes so as to minimise an entropy criterion. However, all these strategies exploit the same underlying principle embodied by our model.

4 Experimental Results

In this section we illustrate the merits of model grouping within the context of 3D object recognition. The core of our object recognition system is a region-based matching scheme proposed in [1]. In this method an object image is represented by its constituent regions segmented from the image. The regions are represented in the form of an Attributed Relational Graph (ARG). In this representation each region is described individually and by its relation with its neighbouring regions expressed in terms of binary measurements. We use the representative colour of each region as its unary measurement and we characterize geometric relations between region pairs using binary measurements.
The matching process is performed using probabilistic relaxation labelling [3]. In this approach, for each region from a test image, we compute the probability that the region corresponds to a particular node of the ARG representing the combined set of object models. We model an object using an image taken from the frontal view. The label probabilities for a region in the test image are initialized by measuring the similarity between the unary measurements corresponding to the two regions being matched. These probabilities are then updated by taking into account the consistency of labelling at the neighbouring regions.

We tested the idea of grouping from two different aspects: label and model grouping. By label grouping we mean that for each region in the test image we classify the union of labels associated with the regions in the database into two sets: candidate labels and rejected labels. The process of region matching then proceeds using only the set of candidate labels. We propose two label pruning schemes: pruning at the initialization stage and at the end of each iteration of the relaxation labelling. At initialization, for each region in the test image we compile a list of candidate labels. This list is based on the degree of similarity between unary measurements. At the end of each iteration of the relaxation labelling process we note the label probabilities associated with each region in the test image and drop those labels whose probabilities are below a predefined threshold.

Model grouping realizes the same idea at a higher level. Let us consider our model-based recognition system as a set of serial classifiers where each classifier is in fact an object recognition expert. Each expert takes the list of model candidates from the previous expert and delivers a pruned list of model candidates to the next classifier. The pruning of models is performed by matching the test image features against features extracted from the model candidates. The objective for the last expert is to select the winning candidate.

In a simple case of this scenario we consider just two recognition experts in a tandem. The first expert performs a course grouping of the object hypotheses based on an entropy criterion. This initial classification is performed using colour cues. We opt for the *MNS* method of Matas et al [12] for this purpose. In this method the colour structure of an image is captured in terms of a set of colour descriptors computed on multimodal neighbourhoods detected in the image[12]. We use the similarity between the descriptors from the scene image and each of the m object models to find the aposteriori probabilities of the object in the scene image belonging to the various model classes in the database.

Having provided the set of a posteriori probabilities $\mathcal{P} = \{p(\omega_i|\overline{x}), \forall i \in \{1 \cdots m\}\}$, we rank them in the descending order. Our objective is to compile a list of hypothesised objects based on their likelihood of being in the scene ($\mathcal{P}$). For this purpose we use the entropy of the system as a criterion. Let us consider the list,Ω, of model hypotheses arranged according to the descending order of their probabilities. If Ω is split into two groups Ω^1 and $\bar{\Omega}^1$ comprising the K most likely objects in the scene and the remaining objects in the database respectively, the entropy of the system is evaluated as follows [7]:

$$E = \alpha E(\Omega^1) + (1 - \alpha) E(\bar{\Omega}^1) \tag{8}$$

where $E(\Omega^1)$ and $E(\bar{\Omega}^1)$ are the entropies associated with groups Ω^1 and $\bar{\Omega}^1$ respectively and α is the probability that the present object in the scene exists in the group Ω^1. By searching the range of possible configurations,($r = 1 \cdots m$), the grouping with the minimum entropy is selected and the group of the hypothesised objects,Ω^1, is passed to the next expert. The second expert is the ARG matcher[1] described earlier. The whole recognition system is referred to as the MNS-ARG method.

We designed two experiments to demonstrate the effect of both label pruning and model pruning on the performance of the ARG method. We compared three recognition systems from the recognition rate point of view: ARG with/without label pruning and MNS-ARG (with label pruning). The experiments were conducted on the SOIL-47 (Surrey Object Image Library) database which contains 47 objects each of which has been imaged from 21 viewing angles spanning a range of up to ±90 degrees. Figure 2 shows the frontal view of the objects in the database. The database is available online[6]. In

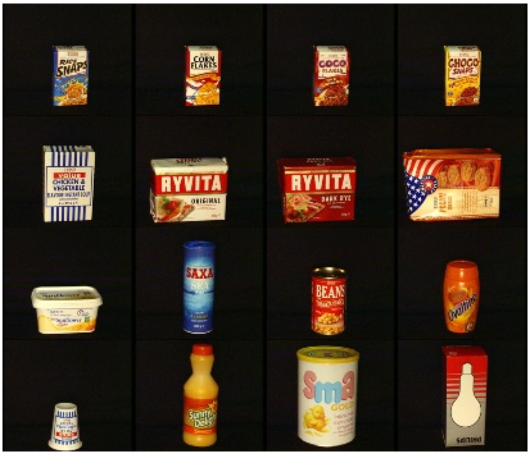

Fig. 2. The frontal view of some objects in the SOIL47 database

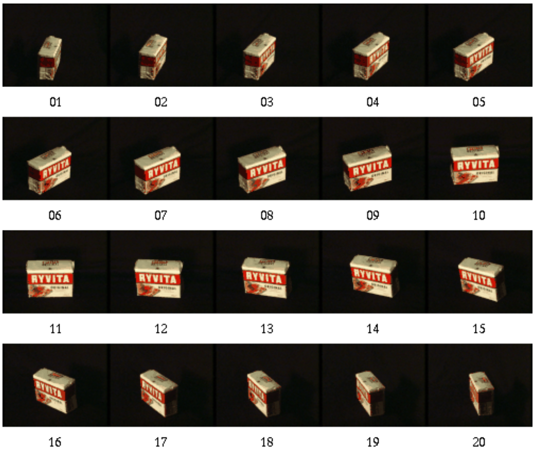

Fig. 3. An object in the database imaged from 20 viewing angles

this experiment we model each object using its frontal image while the other 20 views of the objects are used as test images (Fig. 3). The size of images used in this experiment is 288 × 360 pixels.

In the first experiment we applied the ARG matching for two different cases: with label pruning and without label pruning. The recognition performance for these two cases is shown in Fig. 4. As can be seen, the performance of the ARG matching is considerably enhanced by label pruning. It is worth noting that as this experiment showed the label pruning also speeds up the process of the relaxation labelling significantly.

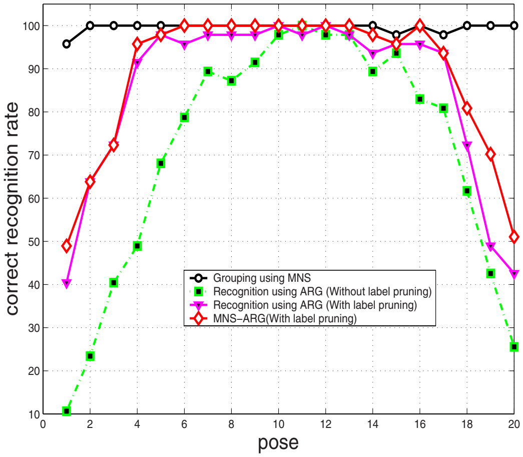

Fig. 4. The percentage of correct recognition for the ARG and the MNS-ARG methods

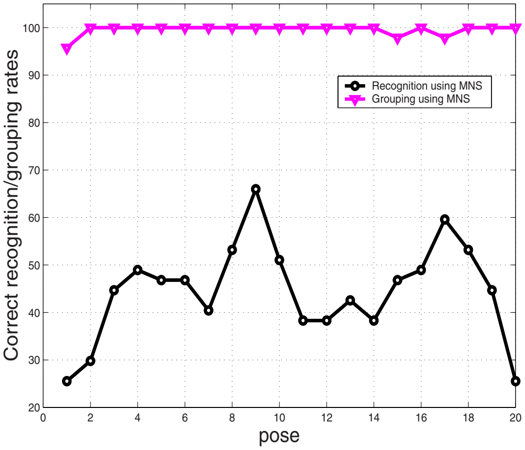

Fig. 5. The likelihood of the correct model being in the list of hypothesised objects generated by the MNS method

In the second experiment, for each test image we applied the MNS method to determine the hypothesised objects matched to it. The results of the experiment are shown in Fig. 5. In this figure we plot the percentage of cases in which the list of hypothesised objects includes the correct model. This rate has been shown as a function of the object pose. For comparison we plot the percentage of cases in which the correct object has the highest probability among the other candidates. It is referred to as the recognition rate. The results illustrate that the recognition rate for the MNS method is not very high. This is not surprising as many grossery items contain similar surface colours.

In contrast, as seen from Fig. 5 in the majority of cases the hypothesised list includes the correct object. It is worth noting that the average size of the list of hypothesised objects is 16 which is near to one third of the database size(47 objects).

The ARG method was then applied to identify the object model based on the list of hypothesised objects generated by the MNS method. This recognition procedure was applied to all test images in the database. In Fig. 4 we have plotted the recognition rate for the MNS-ARG method as a function of object pose. For comparison we have shown the recognition rate when ARG method is applied as a stand alone expert. As a base line we added the rate of correct classification of the MNS method. The results show that the object grouping using the MNS method improves the recognition rate particularly for extreme object views. For such views the hypotheses at a node of the test graph do not receive a good support from its neighbours (problem of distortion in image regions). Moreover a large number of labels involved in the matching increases the entropy of labelling. When the number of candidate labels for a test node declines by virtue of model pruning the entropy of labelling diminishes. Consequently it is more likely for a test node to take its proper label (instead of the null label).

Similar to label pruning, the grouping using the MNS method not only gains the recognition rate but also it reduces the computational complexity of the entire recognition system. This experiment showed that the MNS-ARG method can be performed almost three times faster than the stand alone ARG method.

5 Conclusion

We proposed a multiple classifier system approach to object recognition in computer vision. Multiple experts are used successively to prune the list of candidate hypotheses that have to be considered for object interpretation. The experts are organised in a serial architecture, with the later stages of the system dealing with a monotonically decreasing number of models. We developed a theoretical model which underpins this approach to object recognition and show how it relates to various heuristic design strategies advocated in the literature. The merits of the advocated approach were then demonstrated on a two stage object recognition system. Experiments on the SOIL database showed worthwhile performance improvements, especially for object views far from the frontal, which was used for modelling. The improvements were achieved in spite of using a weak recogniser for the first (pruning) stage. The beneficial effects of different pruning strategies were demonstrated.

Acknowledgements. This work was supported by EU Project Vampire.

References

1. A. Ahmadyfard and J. Kittler. Enhancement of ARG object recognition method. *Proceeding of 11 the European Signal Processing Conference*,volume 3, pages 551–554, September 2002.
2. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 509–522, 2002.

3. W.J. Christmas, J. Kittler, and M. Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 749–764, 1995.
4. M C Fairhurst and A F R Rahman. Generalised approach to the recognition of structurally similar handwritten characters using multiple expert classifiers. *IEE Proceeding on Vision, Image and Signal Processing*, 144(1):15–22, 2 1997.
5. G. Finlayson, B. Funt, and J. Barnard. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17((5):522–529, 1995.
6. http://www.ee.surrey.ac.uk/Research/VSSP/demos/ colour/soil47/.
7. K. Ianakiev and V. Govindaraju. Architecture for classifier combination using entropy measures. In *IAPR International Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 340–350, June 2000.
8. G Kim and V Govindaraju. A lexicon driven approach to handwritten word recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):366–379, 1997.
9. I Krassimir and V Govindaraju. An architecture for classifier combination using entropy measure. In J Kittler and F Roli, editors, *Proceedings of Multiple Classifier Systems 2000*, pages 340–350, 2000.
10. D.G. Lowe. Three-dimensional object recognition form single two-dimensional image. *Artificial Intelligence*, pages 355–395, 1987.
11. S Madhvanath, E Kleinberg, and V Govindaraju. Holistic verification for handwritten phrases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1344–1356, 1999.
12. J. Matas, D. Koubaroulis, and J. Kittler. Colour image retrieval and object recognition using the multimodal neighbourhood signature. In *Proceedings of ECCV*, pages 48–64, 2000.
13. H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, pages 5–24, 1995.
14. A F R Rahman and M C Fairhurst. Exploiting second order information to design a novel multiple expert decision combination platform for pattern classification. *Electronic Letters*, 33:476–477, 1997.
15. A F R Rahman and M C Fairhurst. A new hybrid approach in combining multiple experts to recognise handwritten numerals. *Pattern Recognition Letters*, 18:781–790, 1997.
16. A F R Rahman and M C Fairhurst. An evaluation of multi-expert configurations for for the recognition of handwritten numerals. *Pattern Recognition*, 31:1255–1273, 1998.
17. A F R Rahman and M C Fairhurst. Enhancing multiple expert decision combination strategies through exploitation of a priori information sources. In *IEE Proceeding on Vision Image and Signal Processing*, volume 146, pages 40–49, 1999.
18. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
19. Z Shao and J Kittler. Shape representation and recognition using invariant unary and binary relations. *Image and Vision Computing*, 17:429–444, 1999.
20. M.J. Swain and D.H Ballard. Colour indexing. *Intl. Journal of Computer Vision*, 7(1):11–32, 1991.
21. M. Turner and J. Austin. A neural relaxation technique for chemical graph matching. In *Proceeding of Fifth International Conference on Artificial Neural Networks*, 1997.

Appendix: Probability of Suboptimal Decision Making

In order to investigate the effect of estimation errors on decision making, let us examine the class aposteriori probabilities at a single point $\mathbf{x}$. Suppose the aposteriori 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

$e_B(\mathbf{x}) = 1 - P(\omega_s|\mathbf{x})$. However, our classifier only estimates these aposteriori class probabilities. The associated estimation errors may result in suboptimal decisions, and consequently in an additional recognition error. To quantify this additional error we have to establish what the probability is for the recognition system to make a suboptimal decision. This situation will occur when the aposteriori class probability estimates for one of the other model classes becomes maximum. Let us derive the probability $e_{S_i}(\mathbf{x})$ of the event occurring for class $\omega_i,\ i \neq s$, i.e. when

$$\hat{P}(\omega_i|\mathbf{x}) - \hat{P}(\omega_j|\mathbf{x}) > 0\,\forall j \neq i \tag{9}$$

Note the left hand side of (9) can be expressed as

$$P(\omega_i|\mathbf{x}) - P(\omega_j|\mathbf{x}) + e(\omega_i|\mathbf{x}) - e(\omega_j|\mathbf{x}) > 0 \tag{10}$$

Equation (10) defines a constraint for the two estimation errors $e(\omega_k|\mathbf{x}),\ \ k = i, j$ as

$$e(\omega_i|\mathbf{x}) - e(\omega_j|\mathbf{x}) > P(\omega_j|\mathbf{x}) - P(\omega_i|\mathbf{x}) \tag{11}$$

The event in (9) will occur when the estimate of the aposteriori probability of class ω_i exceeds the estimate for class ω_s, while the other estimates of the aposteriori class probabilities $\omega_j,\ \forall j \neq i, s$ remain dominated by $\hat{P}(\omega_i|\mathbf{x})$. The first part of the condition will happen with the probability given by the integral of the distribution of the error difference in (11) under the tail defined by the margin $\Delta P_{si}(\mathbf{x}) = P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})$. Let us denote this error difference by $\eta_\omega(\mathbf{x})$. Then the distribution of error difference $p[\eta_\omega(\mathbf{x})]$ will be given by the convolution of the error distribution functions $p_i[e(\omega_i|\mathbf{x})]$ and $p_s[e(\omega_s|\mathbf{x})]$, i.e.

$$p[\eta_\omega(\mathbf{x})] = \int_{-\infty}^{\infty} p_i[\eta_\omega(\mathbf{x}) + e(\omega_s|\mathbf{x})] p_s[e(\omega_s|\mathbf{x})] de(\omega_s|\mathbf{x}) \tag{12}$$

Note that errors $e(\omega_r|\mathbf{x}),\ \forall r$ are subject to various constraints (i.e. $\sum_r e(\omega_r|\mathbf{x}) = 0,\ \ -P(\omega_r|\mathbf{x}) \leq e(\omega_r|\mathbf{x}) \leq 1 - P(\omega_r|\mathbf{x})$). We will make the assumption that the constraints are reflected in the error probability distributions themselves and therefore we do not need to take them into account elsewhere (i.e. integral limits, etc). However, the constraints also have implications on the validity of the assumptions about the error distributions in different parts of the measurement space. For instance in regions where all the classes are overlapping, the Gaussian assumption may hold but as we move to the parts of the space where the aposteriori model class probabilities are saturated, such an assumption would not be satisfied. At the same time, one would not be expecting any errors to arise in such regions and the breakdown of the assumption would not be critical. Returning to the event in (9), the probability of the first condition being true is given by $\int_{\Delta P_{si}(\mathbf{x})}^{\infty} p[\eta_\omega(\mathbf{x})] d\eta_\omega(\mathbf{x})$

Referring to equation (11), for each j the second condition will hold for $j \neq s, i$ with probability $\int_{-\infty}^{\Delta Pij(\mathbf{x}) + e(\omega_i|\mathbf{x})} p_j[e(\omega_j|\mathbf{x})] de(\omega_j|\mathbf{x})$, with the exception of the last term, say $e(\omega_k|\mathbf{x})$ which is constrained by

$$e(\omega_k|\mathbf{x}) = -\sum_{\substack{j = 1 \\ j \neq k}}^{m} e(\omega_j|\mathbf{x}) \tag{13}$$

Thus, finally, the probability of assigning point $\mathbf{x}$ to model class ω_i instead of the Bayes optimal class ω_s will be given by

$$\begin{gathered} e_{S_i}(\mathbf{x}) = \int_{\Delta P_{si}(\mathbf{x})}^{\infty} p[\eta_\omega(\mathbf{x})] d\eta_\omega(\mathbf{x}) \\ \bullet \int_{-\infty}^{\Delta P_{ij}(\mathbf{x})+e(\omega_i|\mathbf{x})} p_j[e(\omega_j|\mathbf{x})] de(\omega_j|\mathbf{x})........ \\ ... \int_{-\Delta P_{ik}(\mathbf{x})-e(\omega_i|\mathbf{x})-\sum_{\substack{t=1 \\ t \neq k \\ t \neq l}}^{m} e_t(\omega_t|\mathbf{x})}^{\Delta P_{il}(\mathbf{x})+e(\omega_i|\mathbf{x})} p_l[e(\omega_l|\mathbf{x})] de(\omega_l|\mathbf{x}) \end{gathered} \tag{14}$$

and the total probability of label switching will be given by

$$e_S(\mathbf{x}) = \sum_{\substack{i=1 \\ i \neq s}}^{m} e_{S_i}(\mathbf{x}) \tag{15}$$