Available online at www.sciencedirect.com
ScienceDirect
Applied Soft Computing 8 (2008) 437–445
Applied Soft Computing
www.elsevier.com/locate/asoc

Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition

K. Sirlantzis *, S. Hoque, M.C. Fairhurst

Department of Electronics, University of Kent, Canterbury, United Kingdom

Received 6 May 2003; received in revised form 30 June 2005; accepted 9 August 2005
Available online 18 March 2007

Abstract

In this paper we present two methods to create multiple classifier systems based on an initial transformation of the original features to the binary domain and subsequent decompositions (quantisation). Both methods are generally applicable although in this work they are applied to grey-scale pixel values of facial images which form the original feature domain. We further investigate the issue of diversity within the generated ensembles of classifiers which emerges as an important concept in classifier fusion and propose a formal definition based on statistically independent classifiers using the κ statistic to quantitatively assess it. Results show that our methods outperform a number of alternative algorithms applied on the same dataset, while our analysis indicates that diversity among the classifiers in a combination scheme is not sufficient to guarantee performance improvements. Rather, some type of trade off seems to be necessary between participant classifiers' accuracy and ensemble diversity in order to achieve maximum recognition gains.

Keywords: Diversity; Bit-plane decomposition; Multiple classifier systems; Face recognition

1. Introduction

The benefits of recognition systems which combine the output of multiple classification devices to improve performance have been established in recent years through a large number of mainly empirical studies [1–3]. More in-depth analysis of the internal workings of such systems aiming to identify conditions that influence their successful implementation have led to an increasingly interesting discussion about the requirement for *diversity* among the component classifiers. In particular, the definition and evaluation of diversity is currently attracting the attention of numerous researchers in the field (see for exam-pie [4]).

Contemporary applications such as those aiming to exploit biometric data are of increasing interest. Face recognition occupies a prominent position as one of the more intuitive, albeit potentially complex, tasks related to such applications. Difficulties can range from changes in the environmental conditions, which typically cannot be controlled, to the variety of facial expressions and aging effects that can influence the recognition ability of a specific system [5]. To address these issues effectively, intelligently designed classification engines are needed, among which multi-classifier schemes seem to emerge as a preferred choice [6,7].

As a result of these considerations, in this paper we present two methods to create multiple classifier systems for face recognition based on an initial transformation of the original feature vectors to the binary domain and a subsequent quantisation. Both methods are generally applicable although in this work they are applied to grey-scale pixel values of facial images which form the original feature domain. In both cases the final recognition is the result of the output combination of classifiers trained on subspaces of the intermediate feature spaces spawned by the binary strings. The way these subspaces are formed is what, in essence, distinguishes the two methods, which have been previously applied by the authors to handwriting recognition problems with significant success (see for example [8]).

However, the main focus of the work presented in the current paper is to provide a definition for the diversity among the classifiers in a multi-classifier scheme, propose a measure to evaluate the levels of this diversity and finally explore the relations of the estimators with performance characteristics of

* Corresponding author.
E-mail addresses: k.sirlantzis@kent.ac.uk (K. Sirlantzis), s.hoque@kent.ac.uk (S. Hoque), m.c.fairhurst@kent.ac.uk (M.C. Fairhurst).

1568-4946/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2005.08.002

the individual classifiers and their combination. Of course, the scope of such an investigation is too wide to be covered comprehensively here and hence we restrict our aim to accumulating evidence for identifying possible hypotheses about the role diversity plays in the final *generalisation* ability (recognition of unknown samples) of the combination schemes.

In this sense, it should be noted that the two methodologies we present for the creation of the classifiers to be combined form appropriate vehicles for the type of investigations we are interested in, due to the following reasons. First, they construct, as will become clear in the following discussion, two profoundly different sequences of individual classifiers, the first being trained in a series of feature subsets of decreasing information content and the second in a series of training sets of similar information content. Second, it is expected that a number of statistical characteristics of the sets of classifiers produced, such as the homogeneity of the individual recognition rates and of the diversity of subsets in each case, will be significantly different, due to differences in the training set creation process. Therefore, they will provide the opportunity of studying the quantities of interest in significantly diverse conditions.

To this end, we report a series of cross-validation experiments to estimate the chosen diversity measure as well as the accuracy and variability of the classifiers and the corresponding combinations. In the remainder of the paper we first present formal descriptions of the grey-level image decomposition schemes we propose, followed by a definition of an interesting diversity measure along with a discussion of related research. Then, we describe the multiple classifier systems implemented and the data used in our experiments, and finally, we discuss the results obtained leading to our concluding remarks.

2. Binary quantisation of images

Any grey-scaled image can be split into a series of binary layers. This concept of splitting a multilevel (monochrome or colour) image into a series of binary images is called *bit-plane decomposition*. The idea was originally introduced by Schwarz and Barker [9] as an approach to data compression.

For this decomposition (quantisation), the intensity levels of the image are initially represented using binary notation. In its simplest representation, if

$$U = (u_{xy,k}), \quad x = 1, \ldots, N,\ y = 1, \ldots, M$$

denote a grey-level image with resolution $N \times M$ pixels, the grey-level $u_{xy,k}$, corresponding to the pixel with coordinates x and y of a *k-bit* image, can be expressed in the form of a base-2 polynomial:

$$u_{xy,k} = b_{xy,k}2^{k-1} + b_{xy,k-1}2^{k-2} + \cdots + b_{2,k}2^1 + b_{1,k}2^0. \qquad (1)$$

The $b_{xy,i}$ ($i = 1, \ldots, k$) can be extracted using

$$b_{xy,i} = (\theta_{xy,i-1} - 2\theta_{xy,i}),$$

where

$$\theta_{xy,j} \triangleq \mathrm{Int}\left[\frac{u_{xy,k}}{2^j}\right], \quad \mathrm{Int}[x] \triangleq \text{integerpart of } x.$$

For 'σ' distinct intensity levels, each pixel of the image is represented by a k ($\leq \lceil \log_2 \sigma \rceil$) bit binary code. It is also possible to use other forms of binary notation (for example Grey coding, Excess-3, etc.) to express the intensity values before quantisation, and classifier performance is somewhat dependent on this choice (see [10] for details). Bits can be extracted from this binary representation of the pixel intensity to form a number of decomposed layers. Two distinct ways of achieving this are: the Ordered (or Sequential), and the Random methods.

2.1. Ordered quantisation

This quantisation approach is also referred to as *bit-plane decomposition*. In this approach, the image is decomposed into k layers where layer 'i' consists of the ith order bits of the grey-level values. Formally, the formation of the decomposed layers can be expressed by:

$$L_i = (l_{xy,i}), l_{xy,i} \in \{0, 1\}, \quad x = 1, \ldots N,\ y = 1, \ldots, M,\ i = 1, \ldots, k$$

where

$$l_{xy,i} = \begin{cases} 1, & \text{if } b_{xy,i} = 1 \\ 0, & \text{otherwise} \end{cases}$$

and L_i denotes the ith layer decomposed images. Thus, for example, layer '1' is formed by collecting all the *Least Significant Bits* (LSB) and layer '8' the *Most Significant Bits* (MSB), of an 8-bit binary coded grey-scale image. Fig. 1

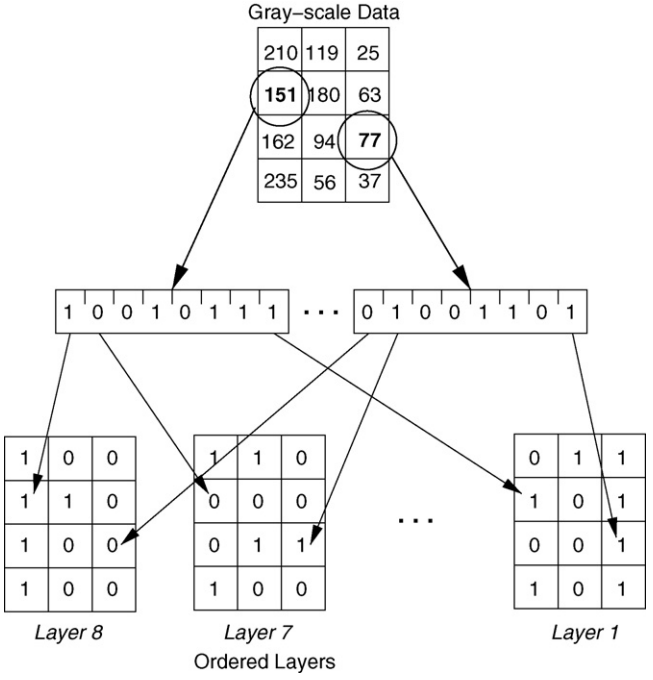

Fig. 1. Ordered feature quantisation schemes.

illustrates this scheme, while Fig. 3(b) shows the outcome of this quantisation process when applied to a grey-scaled face image (Fig. 3(a)).

2.2. Random quantisation

This approach is motivated by the Random Subspace Method (RSM) for creating ensembles of classification trees presented in [2]. In the RSM, randomly chosen features from the original feature space in a pattern recognition task are used to form subspaces. These are subsequently used to train classifiers with an aim of generating diversity among them. Similarly, in the random quantisation approach, bits are chosen arbitrarily for the generation of random layers. To create a randomly decomposed layer a matrix of random numbers W is first generated where each element denotes which bit is to be chosen from the binary representations of the corresponding pixels' intensity levels. The size of this template is equal to that of the image resolution. If

$$W_i = (w_{xy,i}), w_{xy,i} \sim \mathcal{U}[1,k], \quad x = 1,\ldots,N,\ y = 1,\ldots,M$$

denotes one such random template for an image of resolution $N \times M$ and k-bit pixel intensities, where $\sim\mathcal{U}[1,k]$ denotes Uniformly distributed random integers from 1 to k, then the produced randomly quantised layer images can be generally defined as:

$$R_i = (r_{xy,i}), r_{xy,i} \in \{0,1\}, \quad x = 1,\ldots N,\ y = 1,\ldots,M,\ i = 1,\ldots$$

where R_i denotes the image generated by ith random template and thus constitutes the ith random layer, and

$$r_{xy,i} = \begin{cases} 1, & \text{if } b_{xy,d} = 1,\ d = w_{xy,i} \\ 0, & \text{otherwise} \end{cases}$$

where $b_{xy,d}$ are the corresponding coefficients in Eq. (1).

Use of a template ensures that the same bits are always chosen to form a given random layer. Since many such templates (i.e., random matrices) can be generated, this quantisation approach can provide a large number of binary layers which can be subsequently used in a multi-classifier environment. Fig. 2 demonstrates this scheme, and Fig. 3(c) shows the outcome of this quantisation process when applied to a grey-scaled face image.

3. Independence and diversity

In this section, we discuss how to define diversity in a pool of classifiers used in the combination scheme. Strongly related to the notion of "diversity" is the concept of statistical independence, from which we shall begin. Assume that m classifiers $f_1, \ldots, f_m$ are available to classify points (feature vectors) from a given k-class classification problem. Then, for any given feature vector $\mathbf{x}$, each classifier f_i assigns $\mathbf{x}$ a class label $\omega_j \in \Omega = \{\omega_1, \ldots, \omega_k\}$.

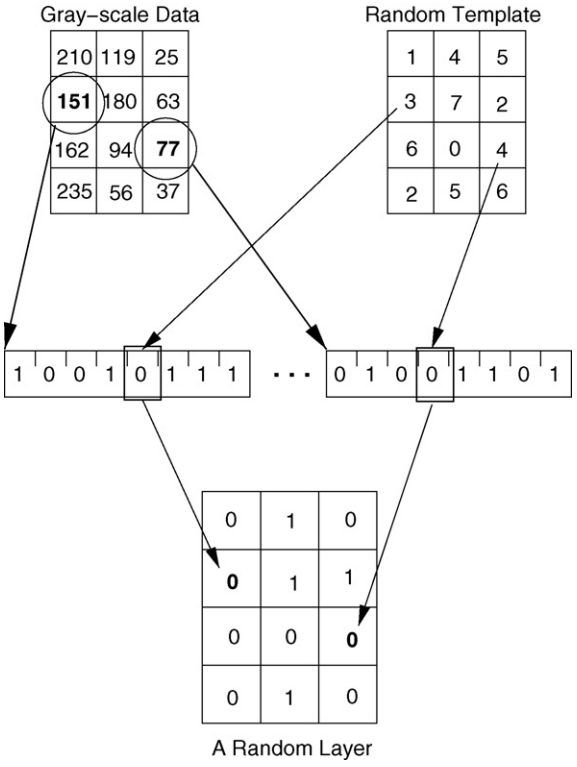

Fig. 2. Random feature quantisation schemes.

The independence of classifiers can be defined by the usual statistical notion of independent experiments. Intuitively, independence means that the output of f_i is unaffected by outputs of $f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_m$. Let $O_i \in \Omega$ denote the output of f_i and $P(O_i)$ the probability that the output from f_i is O_i. After applying $f_i, \ldots, f_m$, the classifiers are said to be *independent* if the probability of observing the compound outcome $(O_1, \ldots, O_m)$ equals the product of probabilities $P(O_i)$, i.e., $P(O_1, \ldots, O_m) = P(O_1), \ldots, P(O_m)$ for all $O_i \in \Omega$, $1 \le i \le m$. It is easy to see that the independence of $f_1, \ldots, f_m$ is equivalent to the independence of random variables $f_1(\mathbf{X}), \ldots, f_m(\mathbf{X})$, where $\mathbf{X}$ is a random feature vector.

3.1. Measurement of agreement among classifiers

In practice, in the majority of cases, the sequence of classifiers are to some degree dependent on each other, and it is then desirable to estimate the strength of the association among their outputs. A metric serving this purpose is the measurement of agreement among the classifiers based on their output, and this estimator forms a natural measure for the classifier diversity within a combination setting. Suppose that m classifiers $f_1, \ldots, f_m$ $(m \ge 2)$ are used to classify $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for a given k-class classification problem $(k \ge 2)$. For $l = 1, \ldots, n$ and $j = 1, \ldots, k$, denote by y_{lj} the number of classifiers which assign $\mathbf{x}_l$ to class j. That is,

$$y_{lj} = \sum_{i=1}^{m} I(f_i(\mathbf{x}_l) = j)$$

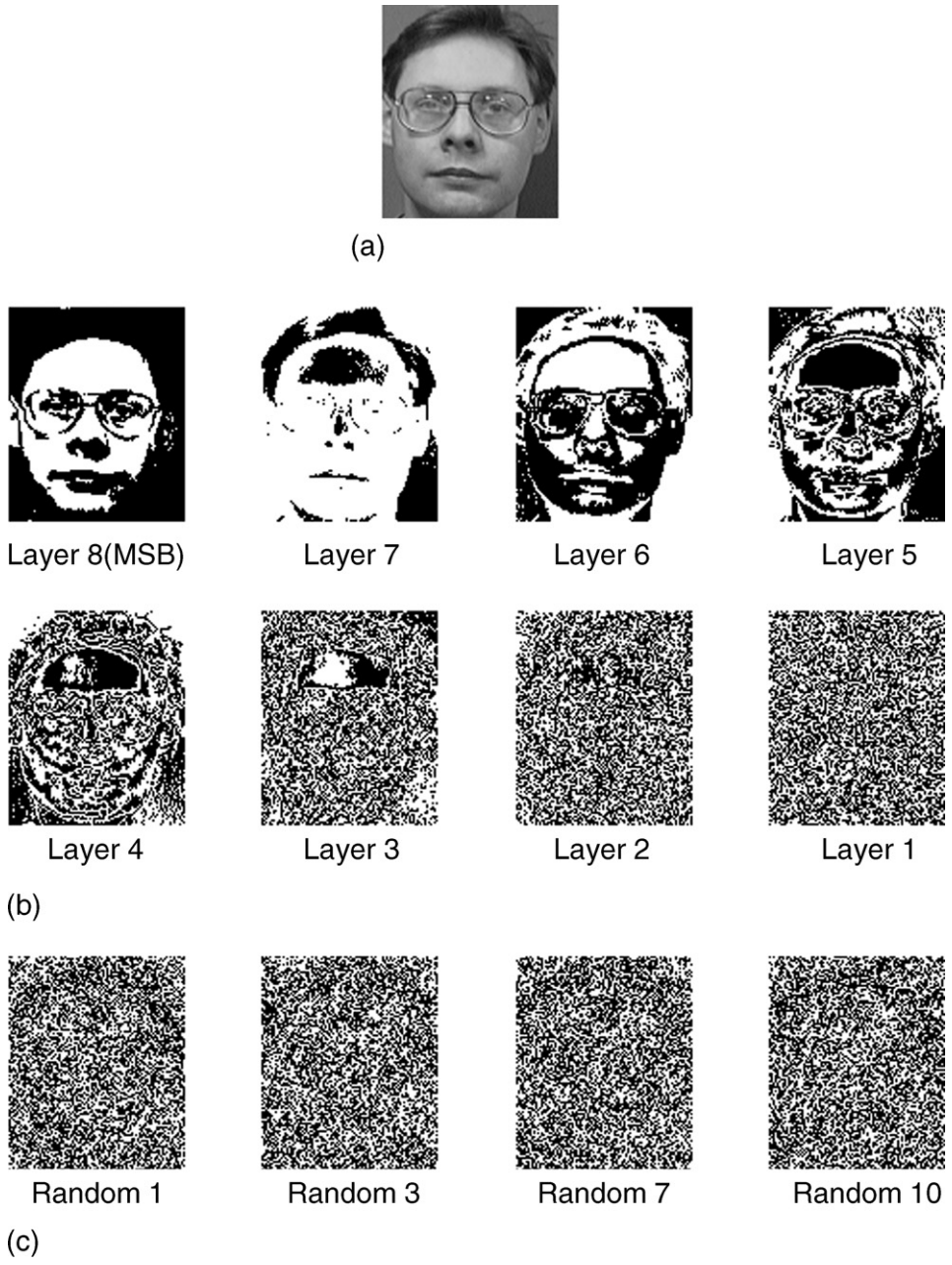

Fig. 3. Effect of quantisation of the grey-scale image shown in (a). (a) Original image, (b) ordered layer quantisation outcome and (c) random layer quantisation outcome.

where $I(f = a)$ equals 1 if $f = a$ and 0 otherwise. Note that $\sum_{j=1}^{k} y_{lj} = m$ for each l. The following displays y_{lj} for m classifications of each of n points (samples) into one of k classes.

Points	Class 1	⋯	Class j	⋯	Class k
$\mathbf{x}_1$	y_{11}	⋯	y_{1j}	⋯	y_{1k}
⋮	⋮		⋮		⋮
$\mathbf{x}_l$	y_{l1}	⋯	y_{lj}	⋯	y_{lk}
⋮	⋮		⋮		⋮
$\mathbf{x}_n$	y_{n1}	⋯	y_{nj}	⋯	y_{nk}

Then, a quantity that naturally measures the degree of agreement among the classifiers' outputs is the following "kappa" statistic [11]:

$$\kappa = 1 - \frac{nm^2 - \sum_{l=1}^{n}\sum_{j=1}^{k} y_{lj}^2}{nm(m-1)\sum_{j=1}^{k} p_j q_j},$$

where $p_j = \sum_{l=1}^{n} y_{lj}/nm$ represents the overall proportion of outputs of classifiers in support of category j, $q_j = 1 - p_j$, $j = 1, \ldots, k$. Clearly κ expresses a special type of relationship among classifiers. In fact, it quantifies the level to which the classifiers agree in their decisions *beyond any agreement that could occur due to chance*. Guidelines for the evaluation of n are given in

[12] as follows: $\kappa \leq 0.75$ indicates excellent agreement beyond chance, $0.4 \leq \kappa \leq 0.75$ indicates good agreement beyond chance, and $\kappa < 0.4$ indicates poor agreement beyond chance, and hence, a very 'diverse' group of classifiers.

The majority of works found in the literature to date define diversity and use estimators based on an "oracle" type of classifier output (e.g., [13,4,14,15]), that is "0/1" to denote wrong and correct classification decisions. However, we take the view that to define the type of association of f_i's based on "0/1" outputs of classifiers, representing erroneous and correct decisions, respectively, for a random sample of feature vectors may lead to ill-defined estimators of diversity since the differences of the $k-1$ possible erroneous decisions is not accounted for. Therefore, the full range of the crisp output (i.e., class labels) of the classifiers in the pool of our trained individual classifiers is employed here in the definition of independence and the resulting diversity estimation.

4. Related work

Ensembles of classifiers have been applied successfully in different domains such as handwriting recognition [16], medical diagnosis [17] and biometrics [18], among others. Despite the existence of a significant body of experimental evidence about a wide range of different ways to combine classifiers, there is still no rigorous theoretical approach to established what are the vital ingredients for constructing a successful ensemble. Intuitively, it is accepted that the incorporation of team members which are accurate and do not make similar errors (i.e., they are 'diverse' in their predictions) will lead to performance improvements. It is, therefore, not surprising that in order to systematically construct accurate ensembles, more research is now focussing on establishing the properties which an ensemble team should posses to achieve the desired gains. A significant part of these studies aim to establish different measures of diversity for classifier ensembles and their relationship with the ensemble accuracy [4] and other characteristics of the individual classifiers and the selection of suitable feature sets [19].

In particular, the relationship between the gain obtained from using an ensemble and the diversity among the classifiers has been the subject of a number of research papers [4,15]. These investigations are centered, mainly, on the correlation between the diversity measure and either the ensemble accuracy or the gain over the average base classifier or the best performing classifier of the ensemble team [4,14]. As a result a number of recommendations were reported in favour of the use of particular diversity measures. Despite these recommendations, there is still no agreement on the best diversity measure to use, or whether indeed, diversity measures can be used effectively in building ensembles which maximise performance. Where diversity measures have been used for ensemble feature selection, it has been reported that the process was sensitive to the choice of the diversity measure used, and choice of the best diversity measure to use was dependent on other factors such as the data being processed [20]. In this paper we take a different approach, and present two feature extraction methods which aim to infuse diversity to the group of the base classifiers which are going to be used in our multiple classifier schemes.

5. The multiple classifier schemes

We start this section about the multi-classifier systems implemented with a short presentation of the individual component classifier used and we proceed with a more formal description of classifier ensembles and the decision fusion rules we used.

The Moving Window Classifier (MWC) is used as the basic individual classifier in the implementation of our multiple classifier systems. The choice of MWC is due to its fast and accurate recognition performance as well as its easy hardware implementation. Details of the MWC scheme, which is an enhanced n-tuple based classification system, can be found in [21–23]. One MWC is trained using each of the decomposed layers and finally a fusion mechanism (here, the *Sum* and the *Majority Vote* rules [3]) is used to combine the individual classifier outputs in the ensemble and generate the final class decision.

In the present work the individually trained classifiers used are arranged in a parallel structure to form a multi-classifier recognition system. The choice of this simple architecture to form the ensemble was preferred so that our experimental results could be more easily studied and interpreted. Fig. 4 gives a schematic representation of the systems illustrating the information flow in the parallel combination architecture through the components of the schemes from the layered

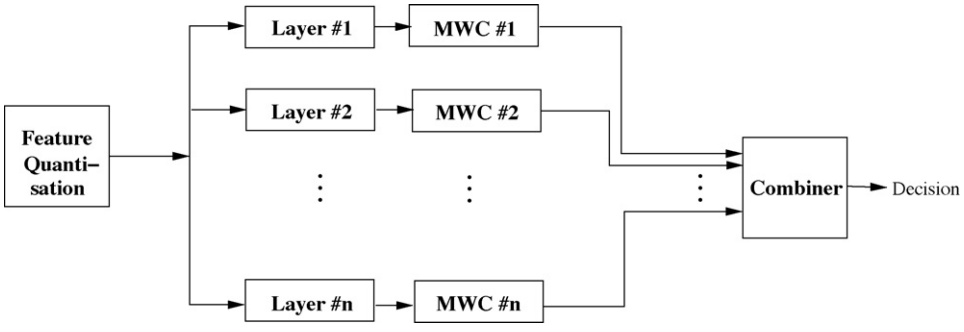

Fig. 4. The multiple classifier schemes with the associated quantised layers.

Fig. 5. Ten different images of 'subject 4' in the ORL data set.

sampling of the binary feature quantisations to the component classifiers and finally the output combination stage.

Although a variety of fusion rules have been devised by researchers [3], the choice of the most appropriate scheme is usually dependent on the degree of diversity of the feature spaces on which the participant classifiers are trained, and the nature of their outputs.

To better understand this, let us consider that a pattern q is to be assigned to one of the k possible classes $\{\omega_1, \ldots, \omega_k\}$ and there are m independent classifiers to each of which q is represented by a feature vector $\mathbf{x}_i$, $i = 1, \ldots, m$. Let each class ω_j be modelled by the probability density function $P(\mathbf{x}_i|\omega_j)$. Following a Bayesian perspective, each classifier is considered to provide an estimate of the true class posterior probability $P(\omega_j|\mathbf{x}_i)$ given $\mathbf{x}_i$. The pattern q should be assigned, consequently, to the class having the highest *posterior* probability. Assuming equal a priori probabilities for all the classes, the corresponding decision rule is:

$$\text{assign } \theta \rightarrow \omega_j \text{ if } P(\omega_j|\mathbf{x}_1, \ldots, \mathbf{x}_m) = \max_{l=1}^{k}(P(\omega_l|\mathbf{x}_1, \ldots, \mathbf{x}_m),$$

where θ is the class label of the pattern under consideration q.

The idea underlying multiple classifier fusion is to obtain a better estimator of the posterior probability by combining the resulting estimates of the individual members of the ensemble. Then, the corresponding 'Sum' combination rule can be formally expressed as follows:

$$\text{assign } \theta \rightarrow \omega_j \text{ if } \sum_{i=1}^{m} P(\omega_j|\mathbf{x}_i) = \max_{l=1}^{k}\left[\sum_{i=1}^{m} P(\omega_l|\mathbf{x}_i)\right].$$

The 'Majority Vote' rule uses 'hardened' decisions Δ_{ij} which can be obtained from the posterior probability estimates provided by the participant classifiers as follows:

$$\Delta_{ij} = \begin{cases} 1, & \text{if } P(\omega_j|\mathbf{x}_i) = \max_{l=1}^{k} P(\omega_l|\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases}$$

Then, after combination

$$\text{assign } \theta \rightarrow \omega_j \text{ if } \sum_{i=1}^{m} \Delta_{ij} = \max_{l=1}^{k}\left[\sum_{i=1}^{m} \Delta_{il}\right].$$

6. Experiments and discussion

For our experimental investigations of the two presented multi-classifier schemes and the estimation of the diversity among the constituent classifiers, we chose a task from the face recognition domain using grey-scale images. The reasons underlying our choice will become apparent from the following description of the characteristics of the particular database used. In addition, face recognition is a complex task domain which still poses challenges to classification algorithms, and therefore it forms an ideal example where the possible benefits of our proposed multi-classifier schemes can be observed and evaluated. Also, important conclusions can be drawn about the role of classifier diversity in the performance of the combination system under the adverse conditions of such a realistic task domain.

The ORL face image database[1] has been used in the experiments. This database consists of 400 images, 10 each of 40 different subjects (4 female and 36 male), captured over the span of a 2-year period from subjects aged between 18 and 81. All the subjects are in an upright, frontal position without restrictions on facial expression. Limited lateral movement and limited tilt is present. Subjects were photographed under different lighting conditions, but always against a dark homogeneous background. Some subjects were captured both with and without glasses. The images have been manually cropped and re-scaled to a resolution of 112×92, 8-bit grey-levels. Fig. 5 illustrates a typical set of the 10 images per subject from the ORL data set. A set of five images of each subject was

[1] The database can be downloaded from: http://www.uk.research.att.com/facedatabase.html.

Table 1
Mean error rates (%) of individual classifiers over the 5 cross-validation runs

Quantisation method	Layer used for training									
	1	2	3	4	5	6	7	8	9	10
Ordered (O)	96.5	95.3	62.7	30.7	11.9	6.1	5.4	9.3	–	–
Random (R)	12.5	14.2	12.9	11.9	12.1	12.9	13.5	13.2	12.0	11.6

Table 2
Performance after fusion (error rates) of ordered layers only

Layers combined	Error rates (%)					Absolute gain
	Before fusion			After fusion: fusion rule used		
	Mean of pool	Standard deviation	Best of pool	SUM rule	Majority Vote	
LI–L8	39.76	37.54	5.40	3.00	5.70	2.40
LI–L4	71.35	27.92	30.70	33.10	64.50	−2.40
L3–L6	27.90	22.98	6.10	5.50	12.90	0.60
L4–L8	12.68	9.82	5.40	2.90	4.50	2.50
L5–L8	8.18	3.45	5.40	3.10	4.70	2.30
L3, L4	46.80	17.44	30.7	31.70	–	−1.70
L4, L5	21.30	10.43	11.9	11.80	–	0.10
L6, L7	5.75	2.28	5.40	3.60	–	1.80
L7, L8	7.35	2.58	5.40	4.50	–	0.90

selected at random for training and the rest used for testing, in each of 5 cross-validating experiments.

It is important to note here that exactly the same type of classifier was used in both the proposed systems. Hence, the variability that can be observed with respect to the performance and other characteristics of the individual classifiers and the fusion strategies applied are solely due to the differences in the feature set quantisation processes used and the *information content* of the resulting layers used as training sets. The diversity in information content can be thought to be reflected by the degradation of the image clarity in the sequence of faces in Fig. 3(b) for the Ordered Quantisation method. However, this visual representation of the information content of training feature set does not seem to hold for the Random Quantisation-based images in Fig. 3(c). The latter, although they are visually very similar to the layer corresponding to the least informative bit (i.e., LSB) shown in Fig. 3(b) when considering the individual classifier error rates for the corresponding layers presented in Table 1, indicate that they contain significantly different information. Thus, while Layer f of the Ordered Quantisation (first row) has an error of 96.5%, that of the Random Quantisation achieves 12.5% error, which is similar to the classifiers generated by layer 5 of ordered method. This inconsistency between the visual representation and the information content as measured by the observed recognition performance is interesting in its own right and worthy of further investigation.

From Table 1, which includes the mean error rates of the individual classifiers created by the two proposed methods, it can be easily observed that while the Ordered method results in heterogeneous performances the Random method generates a significantly homogeneous set. Thus, if differences in performance are considered to express diversity in information content of the training sets (since the classifiers used are identical in all other respects), our experimental setting can be claimed to be performed in appropriately varied conditions. Also, it is not difficult to observe in the same table that the most successful of all the individual classifiers produced are among those trained on the most significant bit layers of the Ordered Feature Quantisation, as is expected since the layers (feature subsets) created by the Random Quantisation consist of a mixing of highly informative and less informative bits.

In Table 2 we present the mean error rates and their standard deviations (columns 2 and 3) of subsets of classifiers trained on the Layers shown in column 1, which were used to form combination schemes for the Ordered case (indicated by 'O'). Similar results for the Random Quantisation case (indicated by 'R') are presented in Table 3. The error rate of the most successful classifiers in every subset is included in column 4 of these tables, while the fifth and sixth columns present the error rates achieved by the well-known Sum and Majority Voting decision fusion strategies [1]. The seventh column of the tables contains, for comparisons, the absolute gain (error rate reduction) achieved by the best performing combination rule (in this case, the sum rule) with respect to the most successful among the classifier in the corresponding ensemble (defined in the first column of the tables). It is easily observed that in both the Ordered and the Random Quantisation cases the combination improves the performance in comparison to that achieved by the best individual classifiers. However, the most important observation to be made here is that in neither case does the combination of the two best performing individual classifiers (pairs 06, 07 and R4, R10. for the Ordered and Random Quantisations, respectively, included in the bottom part of the two tables) correspond to the highest gain from the fusion. Rather, the most successful combinations correspond to

Table 3
Performance after fusion (error rates) of random layers only

Layers combined	Error rates (%)					Absolute gain
	Before fusion			After fusion: fusion rule used		
	Mean of pool	Standard deviation	Best of pool	SUM rule	Majority Vote	
Rl–R10	12.68	3.62	11.60	6.70	6.70	4.90
Rl–R4	12.88	3.20	11.90	7.90	9.30	4.00
R4–R7	12.60	3.86	11.90	7.20	8.60	6.30
R7–R10	12.58	3.75	11.60	7.20	9.90	4.40
R4, R10	11.75	4.71	11.60	8.20	–	3.40
R2, R7	13.85	2.98	13.50	10.60	–	2.90

classifier subsets that include the best individual classifiers as well as others with lower accuracy.

A possible hypothesis for the reasons underlying these observations can be formed if we consider them in conjunction with the estimated diversity measures κ for the same ensembles of classifiers. The κ values with the corresponding standard deviations are presented in Tables 4 and 5 for the Ordered and Random Quantisation, respectively. In order to provide an additional perspective we have also included, in column 2, the error reduction achieved as a percentage of the error rate of the best component classifier. A closer examination of the information in these tables reveals that in all cases the most successful combination ensembles are those in which the addition of a number of reasonably performing classifiers to the best performing pair resulted in an increase of the group diversity (reduction of the agreement statistic κ). It is not difficult to observe that the top performing pairs are characterised by high levels of agreement in their output and this is the reason for not exhibiting the best performance when combined.

Table 4
Estimated diversity measures for the ordered layers with the corresponding gain achieved by the combination

Layers combined	Relative gain	Agreement statistic (κ)	Std. of κ
LI–L8	44.44	0.3723	0.0170
LI–L4	−7.82	0.0917	0.0115
L3–L6	9.84	0.5566	0.0334
L4–L8	46.30	0.7988	0.0246
L5–L8	42.59	0.8767	0.0224
L3, L4	−3.26	0.3709	0.0346
L4, L5	0.84	0.6842	0.0343
L6, L7	33.3	0.9263	0.0280
L7, L8	16.67	0.8769	0.0210

Table 5
Estimated diversity measures for the random layers with the corresponding gain achieved by the combination

Layers combined	Relative gain	Agreement statistic (κ)	Std. of κ
R1–R10	42.24	0.8450	0.0285
R1–R4	33.61	0.8454	0.0230
R4–R7	39.50	0.8392	0.0312
R7–R10	37.93	0.8419	0.0294
R4, R10	29.31	0.8497	0.0515
R2, R7	21.48	0.8194	0.0375

Table 6
Comparisons with alternative face recognition algorithms

Classification algorithm	Recognition error rates (%)
Self-organizing map + convolutional network	3.8
Top–bottom hidden Markov model	13.0
Pseudo-2D hidden Markov model	5.0
Eigenfaces (Euclidean distance)	10.0
Best combination of ordered layers	2.9
Best combination of random layers	6.7

On the other hand, the most diverse ensembles are also those including very poor classifiers and as a result their fusion does not cause significant gains in performance. It seems that the evidence produced in this work supports the hypothesis that there exists some kind of (possibly complex) relation between individual classifier accuracy and ensemble diversity (see also [24] for a similar hypothesis obtained through a bias/variance decomposition of the recognition error) which should be investigated further in order to gain insight into the mechanisms governing the creation of multi-classifier systems with guaranteed performance improvements. Finally it is easily realised from Table 6 that the best performing among the proposed multi-classifier systems outperform significantly a number of alternative methods applied to the same database, as reported in [25].

7. Conclusion

We have presented two methods for exploiting the advantages of multiple classifier systems in face recognition and showed that they can both be used with significant success. We have further investigated the issue of diversity within the generated ensembles of classifiers which emerges as an important concept in classifier fusion. This concept has been typically used in a rather vague way to date, and hence we have proposed a formal definition based on statistically independent classifiers and used the κ statistic to quantitatively assess it. Our results indicate that diversity among the classifiers in a combination scheme is not sufficient to guarantee performance improvements. Rather, some type of trade-off seems to be necessary between the participant classifiers' accuracy and ensemble diversity in order to achieve maximum recognition

gains. We believe that this is an issue worthy of further rigorous investigation, mainly in relation to the information content of the feature sets used for classifier training.

Acknowledgement

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC).

References

[1] L. Lam, Classifier Combinations: Implementations And Theoretical Issues, Multiple Classifier Systems Lect. Notes in Comp. Sci. (LNCS-1857), Springer, 2000, pp. 77–86.
[2] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (8) (1998) 832–844.
[3] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[4] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learning 51 (2003) 181–207.
[5] A. Jain, S Bolle, S. Pankanti (Eds.), Biometrics—Personal Identification in Networked Society, Kluwer Academic Publishers, Boston/Dordrecht/London, 1999.
[6] B. Achermann, H. Bunke, Combination of face classifiers for person identification, in: Proceedings of IAPR International Conference on Pattern Recognition, Vienna, Austria, (1996), pp. 416–420.
[7] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Anal. Mach. Intell. 17 (10) (1995) 955–966.
[8] S. Hoque, K. Sirlantzis, M.C. Fairhurst, Intelligient chain-code quantization for multiple classifier based shape recognition, in: Proc. of UKCI-02, Birmingham, UK, (2002), pp. 61–67.
[9] J.W. Schwarz, R.C. Barker, Bit-plane encoding: a technique for source encoding, IEEE Trans. Aerospace Electr. Syst. 2 (4) (1966) 385–392.
[10] M.S. Hoque, M.C. Fairhurst, Face recognition using the moving window classifier, in: Proceedings of 11th British Machine Vision Conference (BMVC2000), vol. 1, Bristol, UK, (2000), pp. 312–321.
[11] J.L. Fleiss, Statistical Methods for Rates and Proportions, Wiley, New York, 1981.
[12] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[13] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P. Duin, In independence good for combining classifiers? in: Proceedings of 15th International Conference on Pattern Recognition (ICPR'2000), Barcelona, Spain, (2000), pp. 168–171.
[14] D. Ruta, B. Gabrys, Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems, in: Proceedings of SOCO 2001, Paisley, Scotland, 2001.
[15] C.A. Shipp, L.I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, Inf. Fusion 3 (2) (2002) 135–148.
[16] S. Hoque, K. Sirlantzis, M.C. Fairhurst, A newnew chain-code quantisation approach enabling high performance handwriting recognition based on multi-classifier schemes, in: Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR 2003), vol. II, 2003.
[17] S. Chindaro, R.M. Guest, M.C. Fairhurst, J.M. Potter, Assessing visuospatial neglect through feature selection and combination from geometric shape drawing performance and sequence analysis, Int. J. Pattern Recognit. Artif. Intell. 18 (7) (2004) 1253–1266.
[18] A. Ross, A. Jain, Information fusion in biometrics, Pattern Recognit. Lett. 24 (2003) 2115–2124.
[19] S. Chindaro, K. Sirlantzis, M.C. Fairhurst, Component based feature space partition and combination in multiple colour spaces for texture classification, in: Proceedings of the IEE Visual Information Engineering Conference (VIE 2005), Glasgow, UK, 2005.
[20] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in Ensemble Feature Selection, Technical Report TCD-CS-2003-44, Computer Science Department, Trinity College Dublin, Dublin, Ireland, 2003.
[21] M.C. Fairhurst, M.S. Hoque, Moving window classifier: approach to off-line image recognition, Electr. Lett. 36 (7) (2000) 628–630.
[22] M.S. Hoque, M.C. Fairhurst, A moving window classifier for off-line character recognition, in: Proceedings of 7th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, (2000), pp. 595–600.
[23] M.S. Hoque, M.C. Fairhurst, An improved learning scheme for the moving window classifier, in: Proceedings of 6th International Conference on Document Analysis and Recognition, Seattle, Washington, USA, (2001), pp. 607–611.
[24] T.G. Dietterich, Ensemble methods in machine learning, in: J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science (LNCS-1857), Springer, 2000, pp. 1–15.
[25] S. Lawrence, C.L. Giles, A. Tsoi, A. Back, Face recognition: a convolutional neural network approach, IEEE Trans. Neural Networks 8 (1) (1997) 98–113.