# Diversity-Performance Relationship in a Handwriting Recognition System based on Bit-plane Decomposition

S.Chindaro, K. Sirlantzis, M. C. Fairhurst and S. Hoque
Department of Electronics
University of Kent
Canterbury CT2 7NT, United Kingdom
*E-mail: S.Chindaro@kent.ac.uk*

## Abstract

*The success gained by applying bit-plane decomposition methods to handwriting recognition have been demonstrated in our previous work [1,2]. In this paper we address the relationship between the diversity and the improvements obtained by applying multiple combinations of various layers. These layers are obtained by applying a method based on an n-tuple based classification system, namely, the Random Decomposition Technique proposed in [1]. We investigate 5 combination methods and 9 diversity measures using data extracted from the NIST[16] database. Results presented in this paper support the use of the bit-plane decomposition approach as a diversification method. Strong correlation was found between both the accuracy and the improvements and the diversity measures in the majority of the combination methods investigated.*

## 1. Introduction

The automatic recognition of handwriting is still one of the most challenging areas in pattern recognition, with profound implications for the machine vision field. Although many different methods have been reported and some have shown very high performance, none has been able to achieve the accuracy and speed of human readers, which is the ultimate target. So there is ample scope for improvement in this well-researched problem.

The potential advantage of using multiple experts in a unified structure in addressing the problem of recognition of handwritten numerals has been demonstrated in previous work [1-4]. The last decade has witnessed extensive research on the problem of combining classification data supplied by various experts, with the aim of improving the generalisation and hence the overall performance of the system. The aim is to improve on the performance achieved by the best member of the pool. The approach is based on the fundamental assumption that more successful classifiers can be built by combining a pool of classifiers which make different but complementary decisions. Because of the requirement that members of the pool in any fusion strategy should produce uncorrelated errors for the combination to be useful, the issue of addressing how different (or diverse) the pool members are has become important in designing successful multiple classifier systems.

The advantages of applying bit-plane decomposition to n-tuple based methods of handwriting recognition have been demonstrated in our previous work [1, 2]. All n-tuple based systems are susceptible to huge memory space requirements, and the problem was explicitly addressed by invoking the principle of bit-plane decomposition. In this paper we address the relationship between the diversity, accuracy and the improvements obtained in multiple combinations of various layers originating from the Random Decomposition Technique [RDT], which is based on the Scanned n-tuple approach [1]. The proposed method enables significant savings in memory requirements compared to the original sn-tuple-based recognition system without degradation in performance. In fact, the observed performance of the recognition system was improved by combining the individual layers.

By decomposition of the original chain codes into different layers, we are striving to introduce in our approach some form of useful diversity, which can be utilized by the different ensemble methods to produce better performance. Despite the different notions that exist on the concept of diversity, there is a consistent approach to the measures that are used by various authors to describe it. Here we use some of them to show that diversity is an important aspect introduced by the layer decomposition method in character recognition. We also show that there is strong correlation between the improvements and the absolute error rates of the layer

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

combination methods used, and the diversity which ensues.

Our paper is organized as follows: in Section 2 a description of the basic sn-tuple is described as well as the RDT. The diversity measures used in this work are described in Section 3. The experimental set-up is described in Section 4, followed by the results and a discussion of the results in Section 5. In Section 6 the conclusions drawn from the investigations are presented.

## 2. The Transformation Strategies

The transformation method used is based on our previous work in [1]. For the sake of completeness a brief description of this approach is described in this section.

### 2.1. The Scanning n-tuple (sn-tuple) classifier

The scanning n-tuple classifier is an n-tuple based classifier. It has been introduced by Lucas *et al* [5] and is shown to have achieved very high recognition rates while retaining many of the benefits of the simple n-tuple algorithm. In an sn-tuple system, each sn-tuple defines a set of relative offsets between its input points which then scans over a 1-D representation of the character image. This uni-dimensional model of the character image is obtained by tracing the contour edges of the image and representing the path by Freeman chain-codes [6]. In the case of multiple contours, all strings are mapped to a single string by concatenation after discarding the positional information. As different characters produce contour strings of widely varying lengths, all these chains are proportionately expanded to a predefined fixed length. Details of the sn-tuple classification algorithm (including pseudo-code) can be found in [5].

### 2.2. Random Decomposition Technique

In this approach the Freeman direction codes are represented in binary. (It is also possible to use other forms of binary notation, for example, Gray coding). Since there are 8 possible distinct direction codes, 3-bit binary numbers are sufficient to represent them. In the Random Decomposition technique (which is based on Random Subspace Method [7]), bits for decomposed layers are chosen arbitrarily from the Freeman direction-codes. Since the same bits must always be chosen from a given contour position, an array of randomly selected numbers from the set {0, 1, and 2} is generated identifying the bit to be sampled from the corresponding chain element. An arbitrary number of templates can be generated; hence the random transformation approach can create many different binary layers which can be used individually, or be incorporated into various combination schemes (Figure. 1).

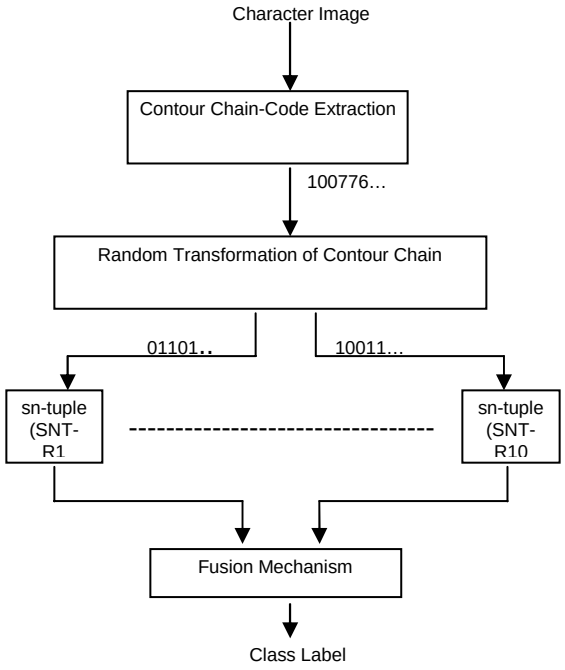


**Figure. 1.** Schematic of the Random Decomposition Method**.**

## 3. The Diversity Measures

Even though there has been no agreed set of standards for quantifying diversity or dependence, a number of measures acquired from the idea of entropy and correlation of individual classifier outputs have been used [8,11,18]. The diversity measures described here are based on oracle outputs. These measures can be split into two categories; pairwise and non-pairwise, and are similar to those used by Kuncheva et al [8]. Given the fact that the focus of this investigation was on measuring the correlation between accuracy, improvements and diversity measures less emphasis is placed on the absolute values and the limits of the diversity measures.

### 3.1. Pairwise Measures

Measures in this category are calculated for the different pairings of classifiers in the pool. These measures are then averaged to give the overall measure for the pool of classifiers. The four measures applied in this study are :

- Yule's Q statistic [9]
- the correlation coefficient [10]

- the disagreement measure [7]
- the double fault measure [11]

For pairwise measures the following definitions are used:

$C_{11}$ : both classifiers correct

$C_{00}$ : both classifiers wrong

$C_{10}, C_{01}$ : one classifier correct, and one wrong

$T = C_{11} + C_{00} + C_{10} + C_{01}$ is the total number of samples (or decisions). In all cases the measurement statistic is averaged over all possible pairs.

**Yule's Q statistic**

Yule's measure of association [9], which is also called Yule's *Q* statistic, is calculated using:

$$Q = \frac{C_{11}C_{00} - C_{10}C_{01}}{C_{11}C_{00} + C_{01}C_{10}} \quad (1)$$

The higher the value of *Q*, the stronger the correlation between classifier outputs. The value of *Q* range from -1 to 1. If classifiers are statistically independent the value of *Q* is zero. The negative values give an indication of the extent the same outputs tend not to occur together.

**The Correlation Coefficient**

The correlation coefficient statistic [10] has the form:

$$\rho = \frac{C_{11}C_{00}C_{10}C_{01}}{\sqrt{(C_{11}+C_{10})(C_{01}+C_{00})(C_{11}+C_{01})(C_{10}+C_{00})}} \quad (2)$$

The lower the value of $\rho$ the higher the diversity.

**The disagreement measure**

This measure gives the proportion of the number of occasions when two classifiers are in disagreement over the total number of decisions [7].

$$D = \frac{C_{10} + C_{01}}{T} \quad (3)$$

The higher the value, the higher the diversity.

**The double-fault measure**

The double fault measure [11] is a ratio of cases where both classifiers make a wrong decision over the total number of decisions.

$$DF = \frac{C_{00}}{T} \quad (4)$$

The lower this measure, the higher the diversity.

## 3.2. Non-pairwise Measures

Five non-pairwise measures were investigated in this study. These are calculated considering all the outputs of all the classifiers in the pool simultaneously.These are:

- the entropy measure [12]
- Kohavi-Wolpert variance [13]
- Interrater agreement [14]
- General diversity [15]
- Coincidence failure diversity [15]

**Entropy Measure**

The entropy measure [12] is given by:

$$E = \frac{1}{C}\sum_{J=1}^{C}\frac{1}{M - [M/2]}\min\{m(z_j), M - m(z_j)\} \quad (5)$$

where *C* is the number of cases in a labeled dataset $\mathbf{z}_j$, *M* = number of classifiers. $m(\mathbf{z}_j)$ is the number of classifiers that correctly classify $\mathbf{z}_j$. Diversity is indicated by values between 0 (no diversity) and 1(highest possible diversity).

**Kohavi-Wolpert variance**

The Kohavi-Wolpert (KW) [13] measure is given by:

$$KW = \frac{1}{CM^2}\sum_{j=1}^{C} m(z_j)(M - m(z_j)) \quad (6)$$

**Interrater Agreement Measurement ( $\kappa$ )**

The average individual classification accuracy is defined by:

$$\varpi = \frac{1}{CM} C\sum_{j=1}^{C}\sum_{i=1}^{M} y_{j,i} \quad (7)$$

and $\kappa$ [14] is given by:

$$\kappa = 1 - \frac{\frac{1}{M}\sum_{j=1}^{c} m(z_j)(M - m(z_j))}{C(M-1)\varpi(1-\varpi)} \quad (8)$$

**Generalised diversity**

If $K$ is a random variable expressing the proportion of classifiers from $C$, that fail on a randomly drawn sample **x**, we can denote by $p_i$ the probability $K = i/C$. Let $p(i)$ denote the probability that $i$ randomly chosen classifiers will fail on a randomly chosen sample **x**. According to [15] the maximum diversity occurs when a failure of one of these classifiers is accompanied by non-failure of the other classifier. The probability of both classifiers failing in this particular instance is $p(2) = 0$. When there is always a simultaneous failure of at least two classifiers the diversity is minimum. We define $p(1)$ and $p(2)$ using the following:

$$p(1) = \sum_{i=1}^{M} \frac{i}{M} p_i \quad \text{and}$$

$$p(2) = \sum_{i=1}^{M} \frac{i}{M} \frac{(i-1)}{(M-1)} p_i \quad (9)$$

Using the above definitions, the generalised diversity measure GD [15] is defined as

$$GD = 1 - \frac{p(2)}{p(1)} \quad (10)$$

Diversity is indicated by values ranging from 0 (minimum diversity) and 1(maximum diversity).

**Coincident failure diversity**

This measure is derived from the GD measure. Coincidence failure diversity (*CFD*) [15] is given by :

$$CFD = \begin{cases} 0, & p_0 = 1 \\ \frac{1}{1-p_0} \sum_{i=1}^{M} \frac{M-i}{M-1} p_i, & p_0 < 1 \end{cases} \quad (11)$$

When all classifiers are always correct or when all classifiers simultaneously either give the correct decision or the wrong decision the *CFD* measure has a minimum value of 0, and a maximum value of 1 when all wrong decisions are unique.

## 4. Experimental Set-Up

The experiments were conducted on characters extracted from the NIST database [16]. This database has pre-defined disjoint training and test sample sets. A 10-class database was extracted. The training dataset consisted of 200 samples per class. 75 samples per class were used for testing. Five fusion strategies were used in the investigations. These are the *product*, *sum, max, median* and *majority voting*. Details of these can be found in [17]. All classifier and combination results used in the investigations include performance averaged over 5-fold cross validation experiments.

Individual layer error rates were first obtained using the methods described in Section 2. The 10 random layers were placed in different pools comprising 5 layers. All possible combinations of pools of 5 were generated, resulting in 252 different layer pools. The 9 diversity measures were calculated for each of the pool of layers. The six classifier combination strategies were applied to each of the groups. Two sets of intermediate tables were created. One incorporated the error rates (1) obtained after the combinations, and the other, the improvements (2) over the best result of each group (see Table 1 below).

**Table 1**. Data organization for the experimental set-up

| Layer Pool | Error Rates (1) or Improvements(2) | | | | | 9 Diversity Measures |
|---|---|---|---|---|---|---|
| | Prod | Sm | Max | Med | Maj | |
| 1 | - | - | - | - | - | ---------- |
| 2 | - | - | - | - | - | ---------- |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 252 | - | - | - | - | - | ---------- |

The pools of layers were then partitioned into randomly selected groups of 10. 10 such groups were generated. The correlation coefficients were calculated for each group for both of the cases (error rates and improvements). The results presented henceforth are thus statistical averages over 10-fold cross validation experiments.

## 5. Results and Discussion

The correlation coefficients are shown in Tables 3 and 4 below. A number of observations can be made from these results. The first is that some of the measures used exhibited strong correlation with the product, sum, median and majority rule. The correlation coefficients between the error results from the combination methods and the

diversity measures (Table 3) are above 0.8 in the cases of the *product*, *sum*, *median* and *majority rule*, and the DF, GD and CF measures, and about 0.7 for the *max* rule. Of particular note are the high values exhibited in relation to the DF measure, which is an indication of accuracy (equation 4). Significant correlation values of above 0.6 were observed with regard to the same combination methods and the D, E and KW measures. The same trend was observed in the case of the correlation coefficients between the improvement over the best individual group member result by the combination methods, and the diversity measures (Table 4). This is in contrast to the results obtained by Kuncheva et al [8] (with the highest correlation coefficient of 0.38, between the majority rule and the CF measure) on the same measures, even though the dataset and the empirical set-up of the experiment were different.

**Table 2.** Correlation coefficients between the error results from the combination methods and the diversity measures**.**

| | Q | ρ | D/ KW | D F | E | $\kappa$ | GD/ CF |
|---|---|---|---|---|---|---|---|
| **Pd** | 0.51 | 0.47 | **-0.63** | **0.87** | **-0.62** | 0.45 | **-0.86** |
| **Sm** | 0.51 | 0.47 | **-0.63** | **0.87** | **-0.62** | 0.45 | **-0.86** |
| **Mx** | 0.27 | 0.22 | -0.38 | **0.70** | -0.37 | 0.20 | **-0.69** |
| **Md** | 0.41 | 0.36 | **-0.53** | **0.81** | **-0.52** | 0.35 | **-0.80** |
| **Mj** | 0.48 | 0.43 | **-0.61** | **0.90** | **-0.61** | 0.41 | **-0.89** |

**Table 3.** Correlation coefficients between the improvement over the best individual group member result by the combination methods, and the diversity measures**.**

| | Q | ρ | D / KW | D F | E | $\kappa$ | GD/ CF |
|---|---|---|---|---|---|---|---|
| **Pd** | -0.35 | -0.30 | 0.47 | **-0.79** | 0.48 | -0.28 | **0.76** |
| **Sm** | -0.35 | -0.30 | 0.47 | **-0.79** | 0.48 | -0.29 | **0.76** |
| **Mx** | -0.27 | -0.21 | 0.40 | **-0.80** | 0.40 | -0.19 | **0.76** |
| **Md** | -0.35 | -0.30 | 0.46 | **-0.76** | 0.47 | -0.29 | **0.75** |
| **Mj** | -0.45 | -0.40 | 0.57 | **-0.85** | 0.58 | -0.39 | **0.84** |

The Q-statistic showed significantly strong correlation with regard to 4 combination rules in relation to the error rates, but only with the majority rule in relation to the improvements. Insignificant correlation was observed with regard to $\kappa$ in the case of improvements and diversity measures (Table 2), but some correlation was observed regarding classifier performances. This was also the case with the ρ measure. It is important to observe that the majority of the combination rules investigated exhibited strong correlation with the majority of the diversity measures, in contrast with those reported elsewhere.

It is also evident from the Tables that D, E and KW measures exhibited the same correlation coefficients, further reinforcing the observations of Kuncheva *et al* [8]. These three exhibited less strong correlation compared to that shown in relationship to the DF, GD and CF measures. However, the figures are significantly higher compared to those quoted as maxima in [8] (0.38) and [19] (0.3). The same correlation coefficients were also obtained with regard to the GD and CF measures. This reinforces the notion that, even though researchers might disagree on the explicit definition of diversity, the ideas converge with respect to its measurement, which points the way to some standardized benchmarks for the measurement of diversity.

## 6. Conclusion

We set out to determine whether the introduction of layering in the RDT variation of the basic sn-tuple method introduced useful diversity for the ensemble methods in handwriting recognition tasks. We did this by exploring whether there was any correlation between this diversity and the improvements which ensued and, also, between the diversity and the accuracy of the ensembles. We achieved the latter by computing the correlation between the error rates of the different ensemble methods and the diversity measures using data extracted from the NIST database.

Our results have supported the use of the layering method as a diversity inducement method based on the diversity measures investigated. Strong correlation was found between both the accuracy and the improvements, and the diversity measures in the majority of the ensembles investigated. We intend to expand our investigations into the diversity amongst combinations of different layers by pooling layers from the Random Method with the other bit-plane decomposition methods from our previous work, namely the Directional and Ordered Methods [1].

## Acknowledgement

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) in carrying out this research.

## References


[1]S. Hoque, K. Sirlantzis and M.C. Fairhurst, "A new chain-code quantisation approach enabling high performance handwriting recognition based on Multi-Classifier Schemes," *Proc. 7th International Conference on Document Analysis and Recognition*, *ICDAR 2003*, Vol. II, August 2003, pp. 834-838,

[2] S. Hoque, K. Sirlantzis and M.C. Fairhurst, "Bit-plane de-composition and the scanning n-tuple classifier." *Proc. Of the 8th International Workshop on Frontiers in Handwriting Recognition*, (IWFHR-8), August 2002, pp. 207-211.

[3] K. Sirlantzis ,M.C. Fairhurst and S. Hoque, "Genetic Algorithms for Multi-classifier System Configuration: A Case Study in Character Recognition," *Proc. MCS 2001*, J.Kittler and F. Roli (Eds), LNCS 2096, 2001. pp. 99-108.

[4] A.F. Rahman and M.C. Fairhurst, "An Evaluation of Multi-Expert Configurations for Recognition of Handwritten Numerals," *Pattern Recognition, 31(9),* 1998. pp. 1255-1273.

[5] S. Lucas and A. Amiri, "Recognition of Chain-Coded Handwritten Character Images with Scanning n-tuple Method," *Electronic Letters, 31(24)*, November, 1995, pp. 2088-2089.

[6] H. Freeman, "Computer Processing of Line-Drawing Images," *ACM Computing Surveys*, 6(1), March, 1974, pp. 57-98.

[7] T. Ho, "The random space method for construction of decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:8, 1998, pp. 832-844.

[8] L. I. Kuncheva and C.J. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, 51, 2003, pp. 181-207.

[9]G. Yule, "On the association of attributes in statistics," *Phil. Tansaction*, A, 194, 1900, pp257-319.

[10] P. Sneath and R. Sokal, "Numerical Taxonomy," W.H. Freeman and Company, 1973.

[11] G. Giacinto and and F. Roli, "Design of effective neural network ensembles for image classification processes," Image *and Vision Computing Journal*, 19:9/10, 2001 pp699-707.

[12] P. Cunningham and J. Carney, "Diversity versus quality in classification ensembles based on feature selection," Technical Report TCD-CS-200-02, Department of Computer Science, Trinity College Dublin, 2000.

[13] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," *in L. Saitta (Editor), Machine Learning Proceedings,* 13th international Conference, 1996, pp. 275-83.

[14] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization," *Machine Learning*, 40:2, 2000, pp. 139-157.

[15] D. Patridge and W. J. Krazanowski, "Software Diversity: Practical Statistics for its measurement and exploitation." *Information and Software Technology*, Vol 39, 1997, pp. 707-717.

[16] "NIST Scientific and Technical Databases," http://www.nist.gov/srd/optical.htm

[17] J. Kittler, M. Hatef , R P. W. Duin and J Matas "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 20, Issue 3, March, 1998, pp. 226 – 239.

[18] K. Tumer and J. Ghosh, "Error correlation and reduction in ensemble classifiers" *Connection Science*, 8:3/4, London:Springer-Verlag, 1996, pp. 127-161.

[19] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information fusion*, 3(2), 2002, pp. 155-148.