Classifier Combination for Hand-Printed Digit Recognition

Michael Sabourin Amar Mitiche Danny Thomas

Bell-Northern Research *and*
INRS-Télécommunications
Verdun, Quebec, Canada H3E 1H6

George Nagy

Electrial, Computer, and Systems Eng.
Rensselaer Polytechnic Insitute
Troy, New York 12180-3590

Abstract

Independent decisions by two high performance nearest-neighbour hand-printed digit classifiers are combined in a principled manner. Three combination methods are investigated: Bayesian combination, Dempster-Shafer evidential reasoning, and dynamic classifier selection. On a test set of 60,000 handprinted digits, dynamic classifier selection performs slightly better than Bayesian or Dempster-Shafer evidential reasoning, but the lowest error rate is obtained by K-nearest-neighbour combination. Single-parameter classifier combination is used to generate error-reject curves. Essentially error-free classification is obtained at the cost of 4% rejects! The zero-reject error rate decreases from 1.18% for the best single classifier system to 0.67% for the combined classifier.

1 Introduction

In pattern recognition, the optimal decision regarding the identity of an unknown pattern is given by Bayes' Rule. Unfortunately, Bayes' Rule requires complete knowledge of the class probability densities for the entire pattern space. Without such knowledge, the nearest-neighbour rule is generally regarded as the best classifier, with an asymptotic error rate less than twice the Bayes rate. Nearest neighbour (NN) classifiers are now practical [4] due to the availability of pruning algorithms, search optimization, and the advent of computers with sufficient memory and processing speed. As a result, the dominant factor limiting recognition accuracy is no longer the choice of classifier, but rather the choice of features.

The combination of two or more classifiers - based on different features - can compensate for the inadequacies of a monolithic system [13] [16]. Indeed, the error profiles produced by different classifiers can be quite distinct. Errors in the overlapping portion of the error profile are difficult, but not necessarily impossible, to correct. One strategy for combining multiple sources of category information is to integrate the classifier information directly and determine the category with the highest aggregate score. An alternative strategy is to explicitly determine the classifier most likely to produce the correct decision. The source of category information may be in the form of distances, probabilities [15] [9] [14], confusion matrices, or rank ordering [16] [10].

We perform large scale experiments using hand-printed digits written by 500 writers not used during training. The outputs of two *high performance* NN classifiers are combined using Bayesian combination, Dempster-Shafer evidential reasoning, and dynamic classifier selection.

2 Nearest Neighbour Classification

Introduced in 1967 [7], the nearest-neighbour (NN) rule assigns the label of the nearest pattern in the reference set to the unknown pattern. Let $x_i \in \Omega$ be a labelled reference pattern, selected from the reference set $X = \{x_i\}_{i=1}^{M}$, where M is the number of reference patterns, and Ω is the pattern space. Let $y \in \Omega$ be the unknown pattern. Then the nearest neighbour of y, denoted x_{nn}, is

$$x_{nn} = arg\ min_{x \in X}\ \ d(x, y),$$

where $d(\cdot)$ is a distance measure. The label of x_{nn} is assigned to y.

We consider two different feature sets: tangents uniformly sampled on the character contour [2], and Zernike moments [11]. In [4], we observed that approximately half of the NN classification errors using tangents were correctly recognized when Zernike moments were used.

Pruning removes superfluous patterns from the reference set, specifically those patterns which do not affect the NN decision boundaries [1] [8]. In [4], pruning reduced the reference set of 118,000 hand-printed characters by 80% for contour tangents and by 76% for Zernike moments, with a tolerable loss of recognition accuracy of 0.07%.

Search optimization [3] discards patterns which cannot be the nearest neighbour. Let p_i be a reference pattern, a be a fixed anchor point, and x an unknown pattern. By the triangle inequality, $d(p_i, a) \leq d(x, a) + d(p_i, x)$, and $d(p_i, a) \geq d(x, a) - d(p_i, x)$. If another reference pattern, p_j, is nearer to x than p_i, then $d(p_j, x) < d(p_i, x)$ which implies

$$d(p_j, a) \leq d(x, a) + d(p_i, x)$$

and

$$d(p_j, a) \geq d(x, a) - d(p_i, x).$$

Any p_j not satisfying these conditions is discarded.

0-8186-4960-7/93 $3.00 © 1993 IEEE

K	Substitutions	Rejections
1	0.196 %	3.52 %
2	0.060 %	6.98 %
3	0.038 %	10.36 %
4	0.022 %	13.86 %

Table 1: Rejection Criterion.

This optimization method is computationally efficient, since we can precompute $d(p_i, a)$, and quickly eliminate a large portion of the reference set. More substantial search improvements can be realized using several anchor points, determined using a Kohonen associative memory [5] as a vector quantizer. The average number of distance computations is reduced from 24,648 to 375. Search optimization and pruning improved the NN query time by a factor of 80.

Rejection Criterion. Improvements in recognition accuracy are possible at the expense of rejection errors. Certain applications require extremely low substitution errors, but tolerate significant rejection rates. We use the following rule: *If the K-nearest neighbours for both classifiers are of the same class, then the result is accepted, otherwise the pattern is rejected.* For suitable values of K, very low substitution rates can be achieved in the relatively homogeneous portions of feature space isolated by this rejection rule (Table 1).

Screening. The rejection criterion is used to detect characters with high confidence, and unlikely to benefit from classifier combination. It is relatively rare for all classifiers to select the same output category and be in error (this occurs 0.20% of the time). Most classifier combination algorithms have difficulty resolving this type of egregious error. Using the screening rule, 96.48% of the characters are *accepted*, with a substitution error rate of 0.20%. The remaining 3.52% characters are considered "hard-to-recognize", and must undergo further processing.

3 Multiple Classifier Structure

Our data is entered in rectangular boxes which are easily detected using a contour following algorithm. Within each box, the image is smoothed and the connected components of the smoothed image are detected. Features of the segmented character are then computed. The contour tangents are derived by smoothing and sampling the chain-code description of the character. The Zernike moments are computed by determining the inner product of the character bitmap with the Zernike kernel of the form $V_{nm}(\rho, \theta) = R_{nm}(\rho)e^{jm\theta}$ where $R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s(n-s)!\rho^{n-2s}}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!}$. These features form the pattern vectors used for classification. These patterns are then classified using NN. Each classifier output may be any measure produced by the classifier, or a function of these measures. For example, the output may be defined as the actual decision (i.e., the chosen category), a distance measure, a probability measure, or the rank of each category. For our purposes, the classifier output takes the form of (1) the label and distance of the *nearest* reference pattern to the unknown pattern, (2) the label and distance of the K *nearest* reference patterns, or (3) the label and distance to *each* reference pattern. When using commercial character classifiers, the output is usually type (1). Since we are using proprietary classifiers, we will use type (3). The classifier combination logic will either directly specify the combined decision or specify it indirectly by choosing the more credible classifier.

Parameter	Mut. Info (Tangents)	Mut. info (Zernike)
Distance to winner	.0988	.0989
Distance to non-winner	.0950	.0920
Freq. of winner in KNN	.0660	.0587
Category freq. entropy	.0626	.0810
Distance ratio	.0971	.0961

Table 2: Mutual information, $I(\mathcal{C}, \mathcal{X})$, for various parameters, $\mathcal{X}$.

4 Combination Algorithms

Two major theories have dominated the field of distributed evidence processing: the Bayesian theory and the Dempster-Shafer theory [14]. We implemented both, but found more success with approaches based on voting [13], ranking [16] [10] and a novel dynamic classifier selection algorithm.

Dynamic Classifier Selection (DCS). We first select classifier parameters which have the highest mutual information with "correctness". These parameters form a new "meta" pattern space, labelled by a correctness value that indicates which classifier was correct for that (training) meta-pattern. During classification, a NN classifier finds the nearest reference meta-vector, thus selecting the classifier whose output category is adopted.

Parameter Selection. Let $\mathcal{X}$ be a candidate classifier parameter that takes on values $\{x\}$, and let $\mathcal{C}$ be the classifier "success" variable, given by $\mathcal{C} = \delta(t, c)$, where t is the true identity of the unknown, and c is the classifier output category. The mutual information between $\mathcal{X}$ and $\mathcal{C}$ is $I(\mathcal{C}, \mathcal{X}) = H(\mathcal{C}, \mathcal{X}) - H(\mathcal{X})$, where $H(\mathcal{X}) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ is the entropy of $\mathcal{X}$, and $H(\mathcal{C}, \mathcal{X})$ is the joint entropy of $\mathcal{C}$ and $\mathcal{X}$. The mutual information is the average amount of uncertainty in the decision that is resolved by observing parameter value $\mathcal{X}$.

To determine the most informative parameters for each classifier, we use a special set of 60,000 characters. $I(\mathcal{C}, \mathcal{X})$ was computed for assorted classifier parameters (Table 2) on screened training data, hence we have only 3.5% of the 60,000 samples available to generate a set of reference parameter vectors, or 2100 samples. To make the pattern space sufficiently dense, we use only three parameters for each classifier: (1) distance to the winner, (2) distance to the first non-winner,

Set	NN-Tang $L2$-dist	NN-Zern $L1$-dist	Prob. Reas.	Evid. Reas.	DCS
1	98.39	96.89	98.58	98.55	98.79
2	98.56	97.01	98.82	98.67	98.92
3	98.56	97.20	98.88	98.76	98.97
4	98.61	96.97	98.99	98.87	98.88
5	98.60	97.40	98.95	98.91	98.90
avg	98.54	97.09	98.84	98.75	98.89

Table 3: Multiple Classifier Combination.

Set	NN-Tang $L1$-dist	NN-Zern $L1$-dist	Ratio	Rank	Joint $K = 5$
1	98.64	96.89	99.16	99.10	99.15
2	98.86	97.01	99.36	99.40	99.39
3	98.75	97.20	99.24	99.24	99.36
4	98.91	96.97	99.14	99.34	99.37
5	98.96	97.40	99.31	99.35	99.41
avg	98.82	97.09	99.24	99.29	99.34

Table 4: Classifier Combination via Single-Parameters.

and (3) the distance ratio of the first non-winner to the winner.

5 Experimental Results

The NIST-3 database of hand-printed characters [6] contains approximately 250,000 isolated hand-printed digits scanned at 300 dots per inch. We partitioned the NIST database into the following non-overlaping sets: (1) training set, 118,000 digits, (2) special training set for DCS, 60,000 digits, (3) test set, 60,000 digits, and (4) reserved, 12,000 digits. The training set is used as NN references and as a means to generate a probability model of the pattern space (needed for probabilistic and evidential reasoning). The special training set is used for generating DCS parameters, as well as building confusion matrices. The test set comprises samples from 500 writers *not used in training.* Results are given in Table 3.

6 Single Parameter Combination

We then tried a simple single parameter decision rule. Three parameters were considered: (1) $L1$-distance ratio of the non-winner to winner, (2) classifier rank (number of consecutive winners nearest the unknown), and (3) decision of the *joint KNN classifier* (the most frequent category amongst the K nearest neighbours from two classifiers is chosen). For cases (1) and (2), the output of the NN classifier with the larger parameter value is selected.

In order to generate error-reject curves, a parameter based rejection rule is needed. To generate error-reject curves, we set a threshold, α. If the parameter value, ρ, is less that α, then reject. By adjusting the value of ρ, we can easily compute the rejection rate and the error rate (Figure 1).

A sample of character recognition errors is given in Figure 2. (Note that for display purposes, the characters are normalized to a fixed size, sometimes causing fragmentations to disappear.) These errors are attributed to (1) insufficient data in the reference set, (2) unrepresentative training patterns, (3) ambiguous test samples, and (4) character segmentation errors. Building a very large, representative set of reliably labeled data is certainly an urgent research objective.

7 Conclusions

This study suggests that nearest-neighbor classifiers operating on different feature sets can be readily combined to decrease overall error. Our experiments confirm that the run-time of NN classifiers can be dramatically reduced through pruning the training set and optimizing the search. The individual classifiers perform best with $K = 1$, and with $L1$ rather than $L2$ distance metric. Our contour-tangent features are better than Zernike moments, even though they provide no information about the interiors of characters with holes.

The combined classifiers perform better than the monolithic classifier system, even when one of the classifiers is weaker than the other, provided that the weaker classifier provides new category information which is not available to the stronger classifier. The specific method is not as important as the idea of combination. In our experiments, we noted that both probabilistic reasoning and dynamic classifier selection performed equally well (note that DCS used an additional 60,000 training samples to select optimal parameters). Single-parameter classifier selection required no training, and is easily adapted to error-rejection tradeoffs. The joint KNN method ($K = 5$) performed best. The consistency of our results among subsets of the data gives us confidence that we are not observing an artifact.

8 Acknowledgment

This work is supported in part by NSERC grant #OGP0004234.

References

[1] Toussaint, G.T., Bhattacharya, B.K., and Poulsen, R.S., *The application of Voronoi diagrams to nonparametric decision rules,* Computer Science and Statistics: 16th Symposium of the Interface, Altanta, Georgia, 1984.

[2] Sabourin, M., and Mitiche, A., *Optical character recognition by a neural network,* Neural Networks, vol. 5, no. 5, pp. 843-852, 1992.

[3] Ramasubramanian, V., and Paliwal, K.K., *An efficient approximation-elimination algorithm for fast nearest-neighbour search based on a spherical distance coordinate formulation,* Pattern Recognition Letters, Vol. 13, No. 7, pp. 471-480, 1992.

[4] Sabourin, M., Mitiche, A., Thomas, D., and Nagy, G., *NN #1 or Hand-Printed Digit Recognition using Nearest Neighbours,* 2nd Symposium on Document Analysis and Information Retrieval, Las Vegas, pp. 397-410, 1993.

[5] Sabourin, M., and Mitiche, A., *Modeling and classification of shape using a Kohonen associative memory with selective multi-resolution,* Neural Networks, vol. 6, no. 2, pp. 275-283, 1993.

[6] Wilson, C.L., and Garris, M.D., *The NIST-3 Handprinted Character Database* National Instute of Standards and Technology, Advanced Systems Division, 1990.

[7] Cover, T.M., and Hart, P.E., *Nearest-neighbour pattern classification,* IEEE Transactions on Information Theory, pp. 21-27, Jan., 1967.

[8] Hart, P.E., *The condensed nearest-neighbour rule,* IEEE Transactions on Information Theory, pp. 515-516, May, 1968.

[9] Sudkamp, T., *The Consistency of Dempster-Shafer Updating,* International Journal of Approximate Reasoning, vol. 7, pp. 19-44, 1992.

[10] Tubbs, J.D., and Alltop, W.O, *Measures of Confidence Associated with Combining Classification Results,* IEEE Transactions on Systems, Man and Cybernetics, vol. 21, no. 3, pp. 691-693, 1991.

[11] Khotanzad, A., and Hung, Y.H., *Rotation invariant pattern recognition using Zernike moments,* ICPR-9, pp. 326-330, 1988.

[12] Franke, J., and Mandler, E., *A Comparison of Two Approaches for Combining the Votes of Cooperating Classifiers,* ICPR-11, pp. 611-614, 1992.

[13] Rice, S.V., Kanai, J., and Nartker, T.A., *A Report on the Accuracy of OCR Devices,* Technical Report, Information Science Research Institute, University of Nevada, Las Vegas, pp. 1-6, March, 1992.

[14] Mandler, E., and Schurmann, J., *Combining the Classification Results of Independent Classifiers Based on Dempster-Shafer Theory of Evidence,* pp. 381-393, Pattern Recognition and Artificial Intelligence, E.S. Gelsema and L.N. Kanai (eds.), Elseiver Science, North Holland, 1988.

[15] Barnett, J.A., *Computational Methods for a Mathematical Theory of Evidence,* Proc. 7th Intl. Conf. Art. Intel., Vancouver, B.C., pp. 865-875, 1981.

[16] Ho, T.K., Hull, J.J., and Srihari, S.N., *On Multiple Classifier Systems for Pattern Recognition managing evidential reasoning in a hierarchical hypothesis space,* ICPR-11, pp. 84-87, 1992.

[17] Nagy, G., *Candide's Practical Principles of Experimental Pattern Recognition,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, No. 2, pp. 199-200, 1983.

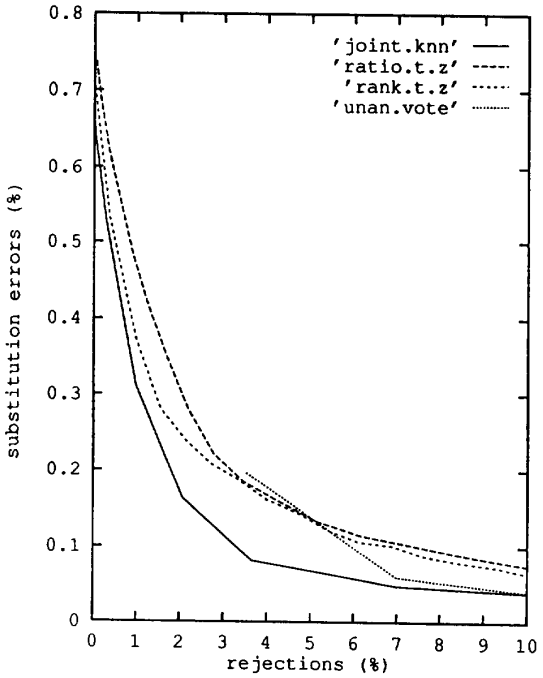

Figure 1. Error-Reject Curves

Figure 2. Misrecognized characters