A Regression Approach to Combination of Decisions by Multiple Character Recognition Algorithms

Tin Kam Ho, Jonathan J. Hull, Sargur N. Srihari

Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo
226 Bell Hall, Buffalo, NY 14260, USA

ABSTRACT

A regression method is proposed to combine decisions of multiple character recognition algorithms. The method computes a weighted sum of the rank scores produced by the individual classifiers and derive a consensus ranking. The weights are estimated by a logistic regression analysis. Two experiments are discussed where the method was applied to recognize degraded machine-printed characters and handwritten digits. The results show that the combination outperforms each of the individual classifiers.

1. INTRODUCTION

Studies on character recognition have been focused on developing effective feature descriptions and classifiers. This has resulted in many character recognition algorithms useful for both constrained and unconstrained domains. It has been observed that existing recognition algorithms often have uncorrelated errors [3], and it is suggested that a robust recognition procedure could be obtained by taking advantages of their individual strengths [9].

Recent studies on multiple classifier systems have revealed that substantial improvement in classification correctness can be obtained by combining decisions of independent classifiers [7]. Useful methods that have been explored include the highest rank method, the Borda count method, and the logistic regression method [8]. In this paper, we discuss the logistic regression technique and how it can be applied to combine decisions of multiple character recognition algorithms.

2. REPRESENTING DECISIONS AS A RANKING OF THE CLASS SET

We assume that each recognition algorithm outputs its decisions in the form of *a ranking* of the character class set with respect to an input image. The combination method operates on such rankings and is therefore independent of the details of the internal structure of each individual algorithm. The method is described in Section 3. An example of combining class rankings is given in Section 4.

Recognition algorithms that output a single decision are considered as special cases, where the rank scale has only two distinct levels. Our combination method applies to these cases as well, though better performance is expected in combination of complete rankings. An example of combining single decisions is given in Section 5.

3. LOGISTIC REGRESSION ANALYSIS

The proposed combination method computes a consensus ranking of a class set according to a weighted sum of the rank scores produced by the individual classifiers.

Assume that m classifiers $C_1, C_2, ..., C_m$ are applied to rank a set of character classes with respect to an input image. As a result, a vector $\mathbf{x} = < x_1, x_2, ..., x_m >$ is computed for each class, where x_i is the rank score assigned

to that class by classifier C_i. For convenience in discussion, we assume that x_i for a class c has the largest value if c is the top choice by classifier C_i. For example, in a 62-class character recognition problem, x_1 for class 'B' has value 62 if 'B' is the top choice by classifier C_1.

A combination function of the following form is proposed:

$$f(\mathbf{x}) \;=\; w_1x_1 \;+\; w_2x_2 \;+\; ... \;+\; w_mx_m,$$

where $w_1, w_2, ..., w_m$ are constant weights associated to rank scores $x_1, x_2, ..., x_m$. For a given image, a value of $f(\mathbf{x})$ is computed for each class. The class set can then be reordered by values of $f(\mathbf{x})$. Such a consensus ranking is the output of the combination algorithm. The value of $f(\mathbf{x})$ reflects the strength of confidence of the algorithm in that particular class. A threshold on its value can be experimentally determined such that classes with values of $f(\mathbf{x})$ less than the threshold can be rejected.

Logistic regression analysis is applied to estimate the weights $w_1, w_2, ..., w_m$. The estimation is based on observations of classifier performance on a training set for which all the true classes are known. The analysis uses a binary response variable Y to distinguish the true class of each image from the other classes.

Suppose that for each image in the training set, m rankings of n classes are produced by the m classifiers, and hence a set of values of the vector $\mathbf{x}$ are observed. A binary response Y is then associated with each $\mathbf{x}$, where $Y = 1$ if that class is the true class of the given input, and $Y = 0$ otherwise. An observation tuple $< Y, \mathbf{x} >$ is thus obtained for each class with respect to each image in the training set.

Values of $\mathbf{x}$ fall in an m-dimensional space, and each component value x_i ranges between 1 and n. Using a training set of p images, $n \times p$ observations of $< Y, \mathbf{x} >$ are obtained. At each level of $\mathbf{x}$, the frequency that the class is the true class of a given image ($Y = 1$) is then computed. These frequencies are used to estimate the expected value of Y at a computed level of $\mathbf{x}$ for each class for an unseen image.

For each class, the value of Y has two possible outcomes with respect to a given image. Y therefore has a Bernoulli distribution with mean

$$E(Y) \;=\; 1 \times P(Y = 1) + 0 \times P(Y = 0) = P(Y = 1).$$

Denote the probability $P(Y = 1)$ as $\pi(\mathbf{x})$, where $\mathbf{x} \;=\; (x_1, x_2, ..., x_m)$. Hence $E(Y) = \pi(\mathbf{x})$ which must fall between 0 and 1.

Intuitively we expect that the likelihood of a class being the true class, i.e., $P(Y = 1)$, is larger when it is ranked higher by the classifiers. The relationship between the ranks and the tendency towards being a true class is expected to be monotonic, but not necessarily linear. In fact, we expect $\pi(\mathbf{x}) \to 0$ when component values of $\mathbf{x}$ are small, and $\pi(\mathbf{x}) \to 1$ when component values of $\mathbf{x}$ are large. A logistic distribution is therefore appropriate to model the relationship between $E(Y)$ and $\mathbf{x}$ [1] [5]. Figure 1 shows the graphs of a logistic function.

For the case with one independent variable, the function has the form

$$\pi(x) \;=\; \frac{\exp\ (\alpha \;+\; \beta x)}{1 \;+\; \exp\ (\alpha \;+\; \beta x)},$$

where α, β are model parameters.

Using this function as a model, the odds of having response 1 ($P(Y = 1)$) are

$$\frac{\pi(x)}{1 \;-\; \pi(x)} \;=\; exp\ (\alpha \;+\; \beta x).$$

The log odds is therefore linearly related to the independent variable x

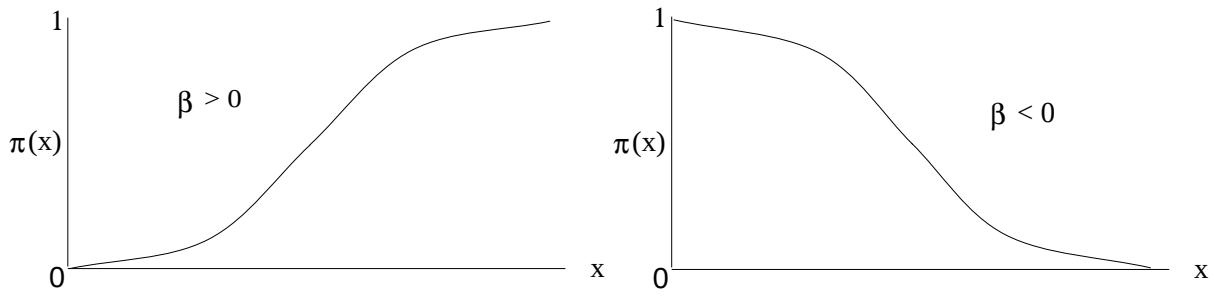


Figure 1: Logistic functions.

$$log\ \frac{\pi(x)}{1\ -\ \pi(x)}\ =\ (\alpha\ +\ \beta x),$$

which can be generalized to the multivariate case with m independent variables as

$$log\ \frac{\pi(\mathbf{x})}{1\ -\ \pi(\mathbf{x})}\ =\ (\alpha\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ ...\ +\ \beta_m x_m).$$

A transformation $L(\mathbf{x})\ =\ log\ \frac{\pi(\mathbf{x})}{1\ -\ \pi(\mathbf{x})}$ is defined. L is referred to as the *log-odds*, or the *logit*, and is linearly related to $\mathbf{x}$.

The logit transformation links the problem to general linear regression analysis. Methods based on maximum likelihood or weighted least-squares can be used to estimate the model parameters $\alpha, \beta_1, \beta_2, ..., \beta_m$ [1] [5]. The relative magnitudes of the estimated parameters indicate the relative significances of the classifiers in their marginal contribution to the logit. Hence the parameters can be used as weights for the rank scores given by each classifier.

In a test run, for each image, the logit for each class is predicted by the estimated model. If a class is ranked at x_1,x_2,...,x_m by classifiers C_1,C_2,...,C_m respectively, its probability of being the true class is given by

$$\begin{aligned}\pi(\mathbf{x})\ &=\ \frac{\exp\ \ (\alpha\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ ...\ +\ \beta_m x_m)}{1\ +\ \exp\ \ (\alpha\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ ...\ +\ \beta_m x_m)}\\ &=\ (1\ +\ \exp^{-1}(\alpha\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ ...\ +\ \beta_m x_m))^{-1}\end{aligned}$$

where α and $\beta_i\ \ (i = 1, ..., m)$ are estimated model parameters.

If only a ranking of the class set is needed, the classes can simply be sorted by the predicted logits in descending order. The class with the largest logit is then considered as the most likely class to which the input image belongs. This process is illustrated in Figure 2.

4. EXAMPLE 1: COMBINATION OF CLASSIFIERS THAT PRODUCE FULL RANKINGS

A multiple classifier system is used to recognize machine printed characters. The system is intended to recognize degraded character images in multiple font styles. The characters are in 62 classes, including the upper and lower case English alphabet and the 10 Arabic numerals. Sample images are collected from live mail, scanned

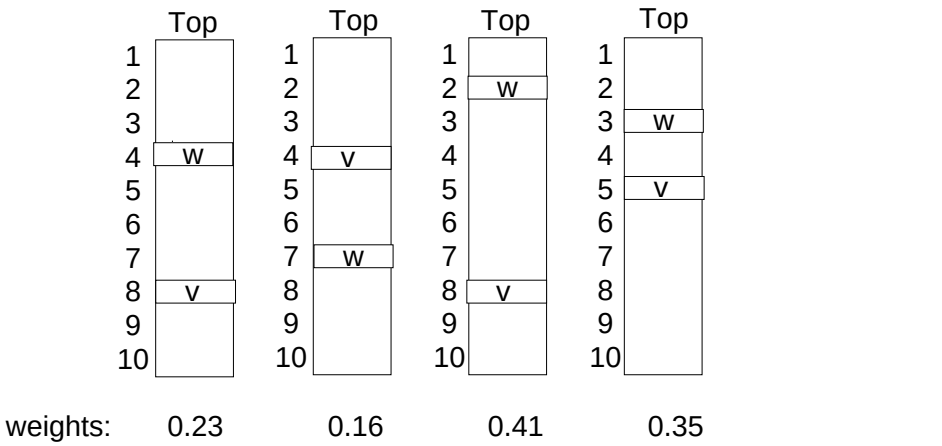


Figure 2: Class set reordering by the logistic regression method.

on a postal OCR at a resolution of 212 pixels per inch. They are then binarized and normalized to 24×24 in size. The quality of the images is highly unstable, and is in many cases very much degraded. Examples of the images in the collection are shown in Figure 3.

Due to similarities in shape after size normalization, several groups of classes are merged into single classes. These include the groups {o,0,O}, {1,l,I,i,j}, {c,C}, {p,P}, {s,S}, {u,U}, {v,V}, {w,W}, {x,X}, and {z,Z}. Therefore, only 48 distinct classes are considered.

Six classifiers are used to recognize the characters. These classifiers use two sets of features in three classification methods.

The features are encoded in binary feature vectors. The first feature vector contains the binary pixel values of the normalized image and has 576 dimensions. The second vector contains feature values obtained by convolving the image with a set of 32 feature templates given in [2]. Those templates are intended for detection of simple structural features like end points and points on edges. For each of the 32 templates, the 24×24 convolution response map is partitioned into 3×3 cells. The response in a cell is set to 1 if there is any occurrence of that feature in that cell. The resultant vector has $32 \times 3 \times 3 = 288$ dimensions.

Each of the two feature vectors is input to three different classifiers, including a Bayesian classifier with feature



Figure 3: Example images in the test set.

independence assumption [6], and two nearest-neighbor classifiers. The Bayesian classifier produces a ranking of the class set according to the posterior probabilities. In the nearest-neighbor classifiers, a Hamming distance is computed between each input vector and each prototype vector. The two classifiers differ in the way the classes are ranked. In the first classifier, the minimum distance between the input vector and all prototypes of a particular class is taken as the distance of the input to that class. The classes are then ranked by these minimum distances. This classifier is referred to as the modified nearest-neighbor classifier. In the second classifier, the distance from an input to a class is computed by averaging the distances from the input to the two closest prototypes for that class. The classes are then ranked in ascending order by these averaged distances. This classifier is referred to as the modified 2-nearest-neighbor classifier. Table 1 summarizes the six classifiers used. A set of 19151 samples is used to train each of the six classifiers.

Table 1: A set of classifiers for character recognition.

Classifier	*Features*	*Classifier Design*
PBC	pixel values	Bayesian with independence assumption
PNC	pixel values	modified nearest-neighbor
P2N	pixel values	modified 2-nearest-neighbor
BBC	Baird features	Bayesian with independence assumption
BNC	Baird features	modified nearest-neighbor
B2N	Baird features	modified 2-nearest-neighbor

The six classifiers are applied to a set of 8000 sample images which are distinct from those used in classifier training. A logistic regression analysis is performed using the rankings of the 48 classes given by the six classifiers.

For each of the 8000 samples, an observation vector of the form $< Y, R_{PBC}, R_{PNC}, R_{P2N}, R_{BBC}, R_{BNC}, R_{B2N} >$ is obtained for each class, where Y is 1 if that class is the true class for that image and 0 otherwise, and R_C is the rank assigned to that class by classifier C for that image. The ranks are represented by a descending number, so that R_C is 48 if that class is ranked at the top by classifier C, and R_C is 1 if it is considered the least similar to the input image by C. There are 48 such vectors for each training image. To simplify the analysis, only the top 10 decisions from each classifier are considered, that is, the vectors are used in the analysis only if any of the R_C's is larger than 38.

Nine different models are attempted in the analysis. Table 2 summarizes the results of an analysis using the SAS procedure CATMOD [13]. For the pixel features, the results indicate that the decisions of both PNC and PBC

are significant when only these two classifiers are used (model 1). However, when P2N is introduced (model 2), PNC becomes insignificant. Since its weight estimate becomes close to zero, it has almost no influence in promoting the rank of any class, and can therefore be ignored. The results for the Baird feature based classifiers are similar (models 3,4). Model 9 shows that when all six classifiers are used, PNC and BNC become insignificant (estimated magnitude of the parameter is small and the standard error is large) and the combination should be based on the four other classifiers.

Table 2: Analysis of maximum likelihood estimates in logistic regression.

Model	Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
1	INTERCEPT	1	-23.7605	0.3833	3842.63	0.0000
	PNC	2	0.2951	0.00818	1302.55	0.0000
	PBC	3	0.1861	0.00625	887.27	0.0000
2	INTERCEPT	1	-25.0556	0.4121	3695.84	0.0000
	PNC	2	-0.0107	0.0215	0.25	0.6180
	PBC	3	0.1730	0.00615	790.01	0.0000
	P2N	4	0.3466	0.0238	212.39	0.0000
3	INTERCEPT	1	-24.9299	0.4149	3610.43	0.0000
	BNC	2	0.3425	0.00920	1384.96	0.0000
	BBC	3	0.1633	0.00612	712.72	0.0000
4	INTERCEPT	1	-26.9303	0.4614	3405.92	0.0000
	BNC	2	-0.0100	0.0205	0.24	0.6274
	BBC	3	0.1448	0.00593	597.07	0.0000
	B2N	4	0.4136	0.0235	308.93	0.0000
5	INTERCEPT	1	-21.8316	0.3559	3762.15	0.0000
	PBC	2	0.2277	0.00692	1082.12	0.0000
	BBC	3	0.2107	0.00684	948.43	0.0000
6	INTERCEPT	1	-25.4009	0.4204	3650.28	0.0000
	PNC	2	0.2186	0.00766	814.00	0.0000
	BNC	3	0.2974	0.00910	1068.49	0.0000
7	INTERCEPT	1	-27.5226	0.4652	3499.64	0.0000
	P2N	2	0.2275	0.00845	724.48	0.0000
	B2N	3	0.3337	0.0104	1028.18	0.0000
8	INTERCEPT	1	-26.1740	0.4099	4078.14	0.0000
	PBC	2	0.1085	0.00614	312.51	0.0000
	BBC	3	0.0836	0.00574	212.28	0.0000
	PNC	4	0.1405	0.00764	338.62	0.0000
	BNC	5	0.2076	0.00896	536.83	0.0000
9	INTERCEPT	1	-27.7886	0.4537	3751.44	0.0000
	PBC	2	0.0996	0.00608	268.56	0.0000
	BBC	3	0.0775	0.00563	189.54	0.0000
	PNC	4	-0.0579	0.0220	6.93	0.0085
	BNC	5	-0.0232	0.0212	1.20	0.2741
	P2N	6	0.2130	0.0244	76.40	0.0000
	B2N	7	0.2648	0.0243	119.12	0.0000

A set of 12000 samples are used to test the performance of these models. Table 3 shows the performance of each of the six classifiers and their combinations by the regression method with parameters given in Table 2. The parameter is set to zero if it is determined to be insignificant.

The fact that the combination (PBC,P2N) performs better than either PBC or P2N individually indicates that even using the same feature set, different classifier designs give independent decisions that can be combined to achieve a higher performance level, though the improvement is not as remarkable as a combination of classifiers that use different feature sets. This suggests that different information contained in the feature vectors is utilized in each classification method. Such information is effectively used in a multiple classifier system.

From Table 3, we can observe improvements to the top choice correct rates in each combination over the individual classifiers. The most significant improvement is obtained by combining four classifiers (PBC, BBC, P2N,

Table 3: Performance of character classifiers and their combinations on test set.

Model	Classifier(s)	Correct Rate (%) at Top N Choices 1	2	3	4	5	10
	PBC	79.3	87.8	91.2	92.8	94.3	97.6
	PNC	85.3	91.3	93.3	94.7	95.4	97.7
	P2N	85.8	91.9	93.9	94.9	95.7	97.9
	BBC	79.1	87.5	90.7	92.4	93.9	97.1
	BNC	84.3	90.8	93.3	94.6	95.4	97.7
	B2N	85.4	91.7	93.9	95.1	95.7	98.0
	Logistic Regression						
1	PBC,PNC	85.4	92.2	94.5	95.7	96.7	98.5
2	PBC,P2N	86.3	92.8	94.7	95.9	96.8	98.6
3	BBC,BNC	85.3	91.8	94.1	95.3	96.1	98.3
4	BBC,B2N	86.0	92.3	94.5	95.7	96.4	98.4
5	PBC,BBC	81.3	89.8	92.7	94.5	95.8	98.3
6	PNC,BNC	86.7	92.6	94.6	95.6	96.2	98.4
7	P2N,B2N	86.9	92.9	95.0	95.9	96.5	98.5
8	PBC,BBC,PNC,BNC	88.1	93.8	95.7	96.6	97.2	98.8
9	PBC,BBC,P2N,B2N	88.8	94.1	95.7	96.7	97.4	98.9

B2N), which gives a net increase of 3% in the top choice correct rate over the best individual classifier (P2N). Corresponding improvements are also observed in larger neighborhoods.

5. EXAMPLE 2: COMBINATION OF CLASSIFIERS THAT PRODUCE HARD DECISIONS

This section describes an application of the logistic model to decision combination in handwritten digit recognition. In this application, each of the classifiers outputs a single decision on the class identity of a given image instead of a full ranking of the class set.

The inputs are binary images of handwritten digits extracted from live mail scanned at 300 ppi and normalized to 16×16 pixels in size. Four classifiers are used to recognize the digits. The first is a second order polynomial classifier (referred to as P). It uses pairs of pixel values as features [14]. The second classifier uses a mixed approach (referred to as M) based on the methods described in [12]. The third classifier is based on stroke analysis (referred to as S) [10]. The fourth classifier is based on analysis of chain codes that represent character contours (referred to as C) [4].

In this application, only one class decision is output by each classifier. A ranking is thus a binary score. For example, if a classifier decides that an input digit is an '8', then the class '8' receives a score '1' and all other classes receive a '0'.

A logistic model was estimated using 6432 images as a training set. Table 4 summarizes the results of a logistic regression analysis on this set of data by the SAS procedure CATMOD. The results indicate that all the four classifiers have significant contribution to correct classification, since each of them receives a substantial weight at similar levels of estimated standard errors.

Table 4: Analysis of maximum likelihood estimates in logistic regression.

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-4.7185	0.2086	511.90	0.0000
P	2	2.7592	0.1701	263.15	0.0000
M	3	1.5955	0.1644	94.21	0.0000
S	4	3.2707	0.1727	358.74	0.0000
C	5	1.4672	0.1659	78.24	0.0000

The four classifiers and the estimated logistic model were applied to a set of 2711 testing images, distinct from

all those used in classifier training or model estimation. Table 5 summarizes the individual performances of the four classifiers and their combination using the estimated model. Different thresholds are applied to the computed logit. The last three rows in Table 5 show the effects on the reject rate and error rate if only the decisions above the thresholds are accepted.

Table 5: Performance of digit classifiers and their combination on test set.

	% Correct	% Error	% Reject
Polynomial	93.6	5.6	0.8
Mixed	84.2	9.0	6.8
Stroke	89.4	9.8	0.8
Chaincode	86.0	2.6	11.4
Logistic Model	95.65	4.35	0.00
with threshold t_1	94.87	2.51	2.62
with threshold t_2	94.21	1.66	4.13
with threshold t_3	73.48	0.04	26.48

6. REMARKS

The success of logistic regression analysis in this context depends largely on the validity of the linearity assumption. This can be examined by graphical methods including plots of the empirical logits and residuals [11]. In our experiments with example 1, we found that a linear model is appropriate for at least top ten decisions in each ranking. Taking more than ten choices is not recommended, as it makes the estimation more costly and worsens the combined performance because of a lack of fit by the linear model.

The Chi-square value computed for each model parameter can be used to evaluate the parameter significance in an ordinary logistic regression analysis. However, caution has to be taken in this application. Because there is only one true class for each image, $Y = 1$ for one class implies $Y = 0$ for all other classes with respect to that image. In other words, values of the response variable Y are related for each image but independent across different images. This may lead to the problem of *overdispersion* or underestimates of the standard errors (Chapter 3, [5]). Nevertheless, the weight estimates are unaffected. As our concern is to obtain a useful combination scheme, we suggest that the significance of a model be evaluated experimentally by observing its performance on a test set. Though, relative significance of the classifiers in a combination can be told by the relative magnitude of the parameter estimates.

Other methods for weight estimation include two-class discriminant analysis and neural network approaches. Logistic regression is more robust to the forms of the distributions of $\mathbf{x}$ and is thus suitable for this application. A neural network can be constructed to estimate the weights by least-squares. However, difficulties in estimation could be encountered if more weight variables are introduced using multiple layers.

7. CONCLUSIONS

A method was proposed for combining decisions by multiple character recognition algorithms. The method re-ranks a set of classes with respect to an input image by computing a weighted sum of the rank scores assigned to each class by the individual classifiers. The weights are estimated by logistic regression analysis using observations of the classifier performances on a training set.

Two examples were described where the method was applied to recognize degraded machine-printed characters and unconstrained handwritten digits. Experimental results show that a substantial increase in correct rate can be achieved by combining a set of independent classifiers. The contribution of each classifier in a combination can be evaluated and insignificant classifiers can be identified and removed.

8. ACKNOWLEDGEMENTS

The support of the Technology Resource Department of the United States Postal Service is gratefully acknowledged. The authors also wish to thank Stephen W.K. Lam and Dar-Shyang Lee for their contribution of code and data in the experimentation, and Ralph Ames for his help in developing the character database.

References

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, 1990.

[2] H.S. Baird, H.P. Graf, L.D. Jackel, W.E. Hubbard, "A VLSI Architecture For Binary Image Classification", in *From Pixels to Features* , J.C. Simon (editor), North Holland, 1989, 275-286.

[3] R. Bradford, T. Nartker, "Error Correlation in Contemporary OCR Systems", *Proceedings of the First International Conference on Document Analysis and Recognition*, Saint-Malo, France, 1991, 516-523.

[4] A. Commike, J.J. Hull, "Rule Learning for Syntactic Pattern Recognition", *Proceedings of Fourth United States Postal Service Advanced Technology Conference*, November 1990, 621-633.

[5] D.R. Cox, E.J. Snell, *Analysis of Binary Data*, 2nd. ed., Chapman and Hall, 1989.

[6] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Addison-Wesley, New York, 1973.

[7] T.K. Ho, J.J. Hull, and S.N. Srihari, "Combination of Structural Classifiers", *Pre-Proceedings of the IAPR Syntactic and Structural Pattern Recognition Workshop*, New Jersey, June 13-15, 1990, 123-136.

[8] T.K. Ho, J.J. Hull, and S.N. Srihari, "Word Recognition with Multi-Level Contextual Knowledge", *Proceedings of the First International Conference on Document Analysis and Recognition*, Saint-Malo, France, 1991, 905-915.

[9] J.J. Hull, A. Commike, T.K. Ho, "Multiple Algorithms for Handwritten Character Recognition", *Proceedings of the International Workshop on Frontiers in Handwriting Recognition*, Montreal, April 2-3, 1990, 117-124.

[10] C.C.L. Kuan, S.N. Srihari, "A Stroke-Based Approach to Handwritten Numeral Recognition", *Proceedings of Third United States Postal Service Advanced Technology Conference*, 1988, 1033-1041.

[11] J.M. Landwehr, D. Pregibon, A.C. Shoemaker, "Graphical Methods for Assessing Logistic Regression Models", *Journal of the American Statistical Association*, **79**, 385, March 1984, 61-71.

[12] T. Pavlidis, F. Ali, "A Hierarchical Syntactic Shape Analyzer", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**, 1, January 1979, 2-9.

[13] SAS Institute Inc., *SAS/STAT User's Guide*, Version 6, Fourth Edition, SAS Institute Inc., 1989.

[14] J. Schuermann, "A Multifont Word Recognition System for Postal Address Reading", *IEEE Transactions on Computers*, **C-27**, 8, August 1978, 721-732.