# ADAPTIVE COORDINATION OF MULTIPLE CLASSIFIERS

TIN KAM HO
*Bell Labs, Lucent Technologies*
*700 Mountain Avenue, Murray Hill, NJ 07974, USA*

Parallel use of multiple classifiers has been known to improve classification accuracy. Yet the additional demand on computational resources often prohibits its application to every input. With concerns for both accuracy and speed, we investigate strategies to coordinate and combine alternative classifiers that can adapt to certain conditions of the input. The design of such strategies is based on a detailed analysis of the classifiers' performances on images synthesized using a parameterized defect model. Individual accuracies, comparative strengths, correlation with defect parameters, confusion probabilities, and agreement of decisions are studied to provide support for the strategies.

## 1 Introduction

The advantages of using multiple classifiers in parallel in a document analysis system have been realized in recent years.[3] It has become well known that using combined decisions often results in improved accuracy. However, since using multiple classifiers incurs significant additional costs, it is less than ideal in a production environment demanding high speed. The combination strategies would be more interesting if they can adapt to certain conditions of the input, so that additional classifiers would be invoked only if necessary.

To design adaptive coordination strategies, it is important to analyze the classifiers' performance in response to different conditions of the input. Among the characteristics of the input that could affect classification accuracy, image quality is often believed to be the most important. In practice, however, analyses have been hampered by the lack of a systematic way to measure and describe image quality. In addition, it is difficult to collect sufficient samples to provide a comprehensive and uniform coverage of the range of interesting qualities.

Recently, we attempted to evaluate classifiers using synthetic images generated using a quantitative, parameterized image defect model.[4] The use of synthetic images permits control of image quality and thus allows us to overcome the above mentioned difficulties. More importantly, it permits explicit grouping of images sharing similar accuracy expectations. These groupings enable us to study the trends in accuracies and their correlation with different factors. Such analyses cannot be done on a random partition of samples.

In this study, we extend the evaluation to three well-known and widely

used classifiers, with focus on their comparison and combination. Our goal is to explore the more significant factors affecting accuracy that may lead to useful classifier coordination strategies that can adapt to corresponding input conditions.

## 2 The Data

Training and testing images were created using a parameterized model of document image defects.[1] The model describes 12 different distortion effects, from which we selected three that correspond to the most commonly studied scanning and printing noise. The selected parameters are `e`, standard error of the Gaussian blurring kernel, `t`, binarization threshold, and `s`, variance of pixel sensor sensitivity around a zero mean. The last is a source of the so called "salt-and-pepper" noise across an image.

The rest of the parameters in the model were set to either constants or their default distributions. We fixed the spatial sampling rate in both x and y directions at 300 pixels per inch, and the text size at 10 point. There is zero skew and both x and y scaling factors are 1.0. We let sub-pixel translations in both x and y directions vary randomly and uniformly within the range of [-0.5, 0.5) pixels.

The parameters of concern were set to vary within a range at fixed intervals: `e` (blur) varies from 0.5 to 2.5 (inclusive) by a step of 0.5; `t` (threshold) varies from 0.2 to 0.4 (inclusive) by 0.05; and `s` (sensitivity) varies from 0.0 to 0.1 (inclusive) by 0.025. These ranges were selected so that for most symbols the images neither vanish nor become shapeless black blobs. They are also consistent with the values used in the public–domain *Bell Labs BLidm0 Character Image Database*.[7] The parameter values are identical to those used in our previous study.[4] There are five values for each parameter, and thus 125 triplets. We studied the behavior of the classifiers over this lattice of 125 points in the parameter space. Figure 1 shows some samples in this space.

We included only the 26 upper case letters in the experiment. The intention was to focus solely on classification by shape. Lower case letters, digits, and punctuations were not used because of complications in error estimation caused by case and shape confusions that need to be resolved by context.

Input to the defect model consists of shape prototypes from 40 commonly used typefaces. Taking an ideal prototype (represented by an outline description) of each symbol, the generator creates one image when the value of each parameter is fixed. If `s` is zero, the image is uniquely created. If `s` is positive, pixel sensitivity varies randomly across the image, and the image is subject to a secondary random sampling from a distribution of all those images with



Figure 1: Sample images in the chosen space. From top to bottom, s varies by increments of 0.025: (e,t,s) = (*,*,0.0), ..., (*,*,0.1); from left to right, e varies by increments of 0.5 on a fixed t, then t varies by increments of 0.05 for each set of e: (e,t,s) = (0.5,0.2,*), ..., (2.5,0.2.*), (0.5,0.25,*), ..., (2.5,0.25,*), ..., (0.5,0.4,*), ..., (2.5,0.4,*).

the same variance in pixel-sensitivity. With each parameter triplet (`e, t, s`) we created three sets of samples: TR1, TR2 and TE. Each of TR1, TR2, and TE contains about 130,000 samples (26 symbols $\times$ 40 fonts $\times$ 125 parameter triples). The inexact figure was due to images that vanished under certain defect parameters. TR1 was used for training the classifiers. TR2 was used to observe classification accuracies and to develop the coordination strategies. TE was used to test the coordination strategies.

## 3 Classifier Accuracies in the Defect-Parameter Space

### *3.1 Accuracies as a Function of Parameter Values*

Three well known classifiers were used in this study: nearest-neighbor matching using Hamming distance (referred to as NN), the `C4.5` decision tree classifier,[6] and the Bayes classifier with feature independence assumption. In all cases the raw binary pixel values were used as input features. The images were normalized in size to $16 \times 16$ and each input feature vector has 256 dimensions. The choice of these classifiers is somewhat arbitrary, but they represent typical tradeoffs in accuracy and space/time efficiency (Table 1). [a]

The performance of each classifier was examined in more detail when accuracies were evaluated on samples generated using each triplet of defect parameters. The association of the values of each defect parameters and classifier accuracies was first measured using Spearman's rank correlation coefficient (Table 2). Here the rank correlation coefficient was used for its robustness to input distributions. Values of each parameter were ranked by their magnitude.

[a] Speed was measured on an SGI Challenge XL with 150MHz IP19 processors.

Table 1: Size, speed and accuracy of the classifiers.

| classifier | size (Mbytes) | speed (char/sec) | %correct on TR2 | %correct on TE |
|---|---|---|---|---|
| NN | 34.44 | 1.79 | 98.08 | 98.06 |
| C4.5 | 0.65 | 1193.21 | 95.74 | 95.77 |
| Bayes | 0.09 | 2090.37 | 88.47 | 88.51 |

The strongest correlation was found between classifier accuracy and the value of `e`, followed by correlation between accuracy and the value of `t`. Values of the parameters also affect strongly the gaps between accuracies of various classifiers. That is, when the value of `e` is higher, not only are the accuracies of all classifiers lower, but the weaker classifiers are inferior by a larger margin. Values of `s` alone are insignificantly correlated with the accuracies. To better examine the effects of `e` and `t`, we show accuracies in the (`e`,`t`) space at fixed values of `s` (Figure 2).

From Figure 2, it can be seen that accuracies of all three classifiers decrease continuously and nearly monotonically with values of `e` and `t`. The most interesting change occurs at the highest value of `e` and near the highest values of `t`, where a sudden drop in accuracies of all three classifiers can be observed. In other words, the effect of `t` is strongest only when `e` is at the highest (2.5), where both effects are strongly amplified. This pattern of changes is observed at every value of `s`.

The nearest-neighbor classifier remains superior to the other two classifiers throughout the space. Similarly, the `C4.5` classifier is more accurate than the Bayes classifier at every parameter setting. Their differences are enlarged with increased values of `e` and `t`.

Table 2: Spearman's rank correlation coefficients ($r_s$) of accuracies and defect parameters (* marks the ones that are significant at the level of 0.0001). c: correct rate of the corresponding classifier. `e`, `t`, `s`: values of the blurring, thersholding, and pixel sensitivity parameters, respectively. `et`, `es`, `ts`: products of values of the parameter pairs (`e`, `t`), (`e`, `s`) and (`t`,`s`), respectively. The last two lines show the correlation between the difference in accuracies of the two corresponding classifiers and the parameter values.

| classifier | $r_s(c,$ `e`$)$ | $r_s(c,$ `t`$)$ | $r_s(c,$ `s`$)$ | $r_s(c,$ `et`$)$ | $r_s(c,$ `es`$)$ | $r_s(c,$ `ts`$)$ |
|---|---|---|---|---|---|---|
| NN | -0.824 * | -0.269 | 0.052 | -0.877 * | -0.336 * | -0.040 |
| C4.5 | -0.751 * | -0.309 | -0.078 | -0.787 * | -0.393 * | -0.161 |
| Bayes | -0.728 * | -0.190 | 0.018 | -0.720 * | -0.313 | -0.033 |
| NN - C4.5 | 0.756 * | 0.317 | 0.169 | 0.769 * | 0.457 * | 0.234 |
| C4.5 - Bayes | 0.632 * | 0.159 | -0.148 | 0.621 * | 0.148 | -0.109 |

### *3.2 Accuracies by Class*

Errors made by each of the three classifiers are concentrated on different sets of classes. The classes and corresponding frequencies of errors are listed in Table 3, sorted by the frequencies. The knowledge of most common confusions could be exploited to determine when a decision by a particular classifier should be trusted. Table 4 shows the accuracy by class and the reliability of decisions. The left half of the table shows how often an image belonging to a known true class is correctly classified (accuracy). The right half of the table shows that, when the decision of a classifier is one particular class, how often that decision is correct (reliability). For designing coordination strategies, the reliability measure is more useful since the true class is not known at classification time.

Table 3: Distribution of errors among classes.

| classifier | number of errors by class |
|---|---|
| NN | E(264), L(200), I(178), F(172), T(134), O(133), D(113), J(108), U(104), B(104), R(96), P(90), C(90), Q(88), Z(72), G(67), H(66), Y(64), X(63), V(62), S(62), K(47), M(43), A(34), N(22), W(20). |
| C4.5 | E(392), R(311), B(296), L(285), H(259), F(247), Q(246), T(241), I(232), N(228), U(224), O(223), P(212), M(195), D(194), Z(179), W(179), Y(173), V(171), K(166), S(162), C(157), X(152), J(151), G(147), A(113). |
| Bayes | O(1278), M(1203), E(942), Q(923), B(851), G(836), D(795), T(783), W(770), R(726), H(601), V(527), C(517), N(514), L(433), P(379), S(376), K(368), X(346), Z(321), Y(307), F(301), J(251), I(241), U(237), A(162). |

### *3.3 Frequency of Classifier Agreement*

When accuracy is the primary concern, additional resources may be committed and multiple classifiers can be applied in parallel. The improvement in reliability (measured by the correct rate of the agreed decisions) is observed in this experiment (Table 5 compared to individual accuracies from Table 1). The distribution of agreed decisions over the defect-parameter space is highly skewed (Figure 3). Combined with the observed high reliability, this suggests that images from the regions where the classifiers agree more often are easier from the perspectives of these classifiers. Thus one can obtain a sort order of the parameter triplets using the frequency of agreement by all three classifiers. To show an intuitive relationship with image quality as perceived by human eyes, Figure 4 displays some images from the TR2 set arranged in this order.

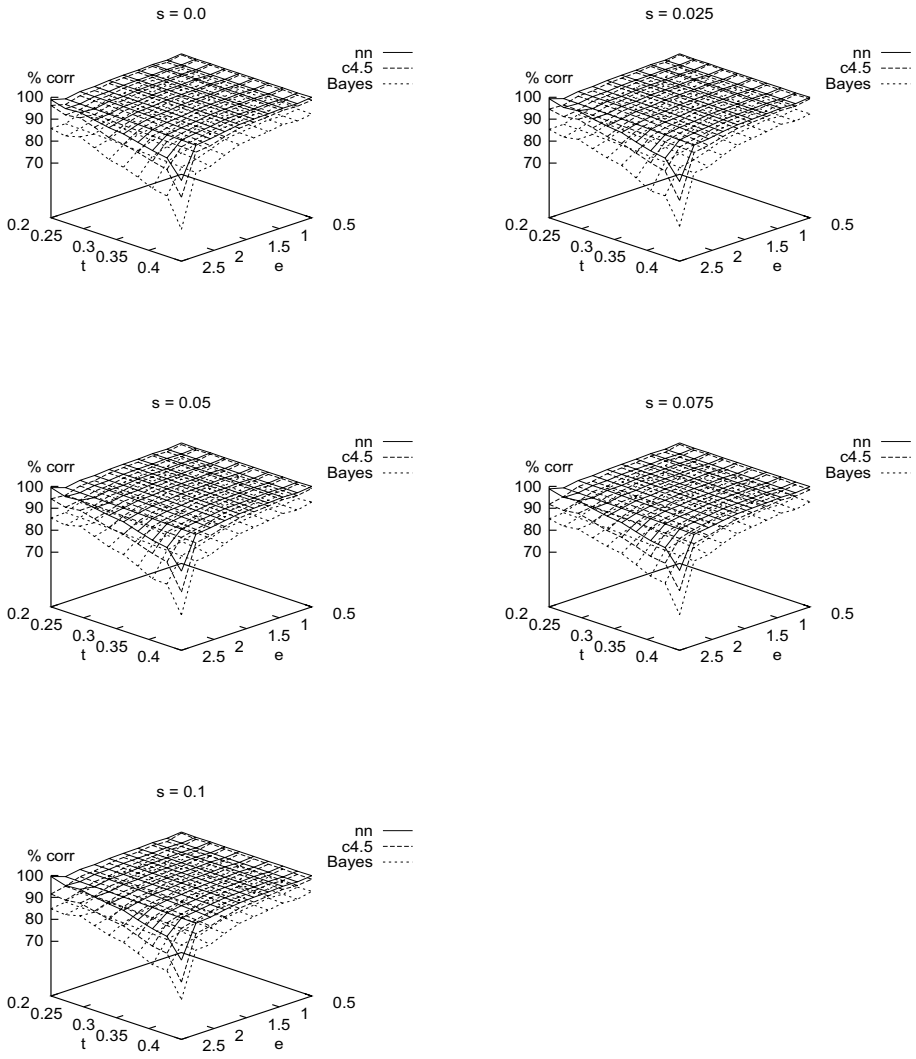



Figure 2: Accuracies of individual classifiers in `(e,t)` space at fixed values of `s`.

Table 4: Accuracy by class and reliability of decisions. For each class: accuracy measures %correct of those images labeled as that class; reliability measures %correct of those images decided as that class.

| truth | NN | C4.5 | Bayes | decision | NN | C4.5 | Bayes |
|---|---|---|---|---|---|---|---|
| A | 99.32 | 97.74 | 96.76 | A | 99.02 | 96.98 | 98.57 |
| B | 97.92 | 94.08 | 82.98 | B | 99.15 | 94.51 | 81.05 |
| C | 98.20 | 96.86 | 89.66 | C | 97.98 | 96.09 | 92.17 |
| D | 97.74 | 96.12 | 84.10 | D | 97.54 | 95.21 | 96.31 |
| E | 94.72 | 92.16 | 81.16 | E | 96.91 | 94.48 | 93.29 |
| F | 96.56 | 95.06 | 93.98 | F | 95.42 | 92.96 | 86.33 |
| G | 98.66 | 97.06 | 83.28 | G | 99.14 | 96.42 | 95.68 |
| H | 98.68 | 94.82 | 87.98 | H | 98.78 | 95.89 | 85.15 |
| I | 96.44 | 95.36 | 95.18 | I | 93.25 | 91.20 | 73.60 |
| J | 97.84 | 96.98 | 94.98 | J | 96.30 | 96.79 | 89.08 |
| K | 99.06 | 96.68 | 92.64 | K | 99.44 | 95.80 | 88.79 |
| L | 96.00 | 94.30 | 91.34 | L | 95.58 | 95.27 | 81.53 |
| M | 99.14 | 96.10 | 75.94 | M | 99.56 | 97.03 | 80.26 |
| N | 99.56 | 95.44 | 89.72 | N | 99.50 | 95.77 | 85.82 |
| O | 97.33 | 95.53 | 74.37 | O | 97.61 | 93.08 | 79.84 |
| P | 98.20 | 95.76 | 92.42 | P | 98.14 | 96.26 | 97.43 |
| Q | 98.24 | 95.08 | 81.54 | Q | 98.75 | 95.67 | 73.12 |
| R | 98.08 | 93.78 | 85.48 | R | 98.95 | 94.40 | 91.80 |
| S | 98.76 | 96.76 | 92.48 | S | 98.50 | 97.29 | 93.47 |
| T | 97.32 | 95.18 | 84.34 | T | 97.61 | 97.70 | 96.57 |
| U | 97.92 | 95.52 | 95.26 | U | 97.82 | 95.42 | 86.35 |
| V | 98.76 | 96.58 | 89.46 | V | 98.41 | 96.14 | 91.08 |
| W | 99.60 | 96.42 | 84.60 | W | 99.80 | 97.43 | 94.10 |
| X | 98.74 | 96.96 | 93.08 | X | 99.52 | 97.55 | 92.40 |
| Y | 98.72 | 96.54 | 93.86 | Y | 98.88 | 97.26 | 89.73 |
| Z | 98.56 | 96.42 | 93.58 | Z | 98.78 | 97.06 | 99.05 |

Table 5: Frequency and reliability of agreed decisions.

| decision agreed by: | (#agreed/#total) | (#correct/#agreed) |
|---|---|---|
| NN and C4.5 | 95.78% | 99.40% |
| NN and Bayes | 88.73% | 99.35% |
| C4.5 and Bayes | 87.90% | 98.93% |
| NN, C4.5 and Bayes | 87.06% | 99.65% |

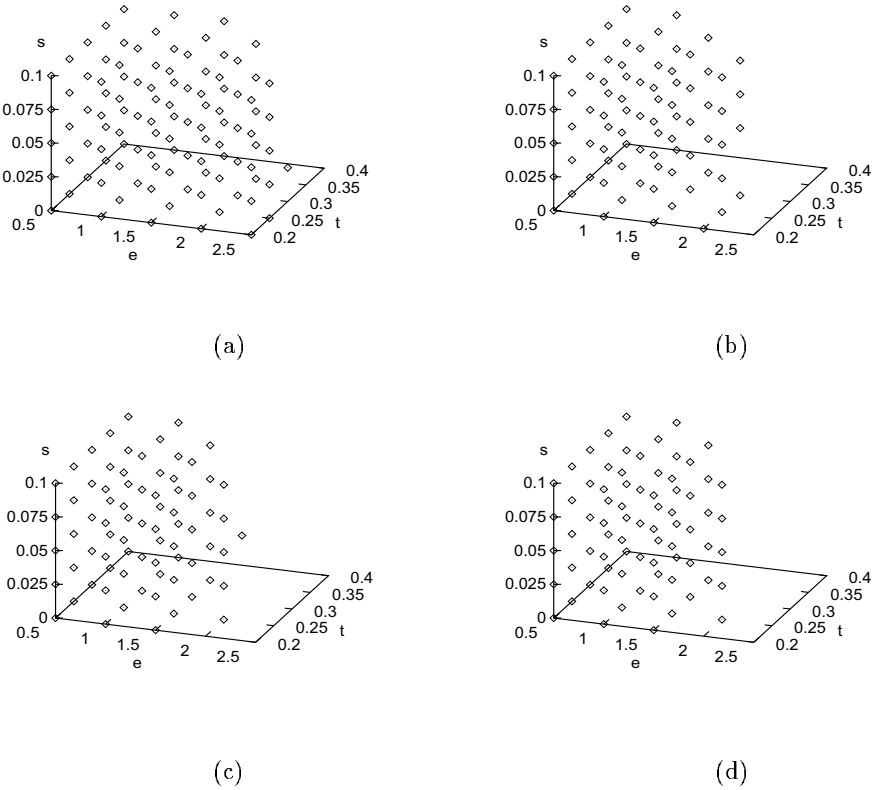



Figure 3: Areas where (a) NN and C4.5, (b) NN and Bayes, (c) C4.5 and Bayes, and (d) all three classifiers agree over 90% of the time.



Figure 4: Sample images at parameter settings sorted by decreasing frequency of agreement of three classifiers (for each symbol: left to right, then top to bottom).

## 4 Coordination Strategies

We have analyzed the classification accuracies from several perspectives: accuracies over the entire parameter space, their correlation with each individual defect parameter, accuracies on different classes, and distribution and reliability of agreed decisions. Many heuristics for classifier coordination can be developed by exploiting these characteristics. Other than these, one can also take into account the estimates of point size and typeface, the frequency of occurrences of each symbol on a document page, and the possibility of resolving class confusions by contextual knowledge.

The choice of coordination strategy is largely determined by the tradeoff between accuracy and efficiency. At an extreme, if accuracy is the only concern, one could use the most conservative (and most expensive) approach, which is to apply all available classifiers in parallel to each input, and accept the decision only if they all agree. But more often, constraints on classification speed would suggest using the fastest classifier on as many images as possible, and deferring the use of the more expensive classifiers until necessary.

We studied classifier coordination strategies driven by one or more of the following three factors: image quality, probability of class confusions, and reliability of agreed decisions.

### *4.1 Dynamic Classifier Selection*

Given the accuracy function of the classifiers in the parameter space, one may wish to find the position of an arbitrary image in the space, and use the least expensive classifier that can achieve a pre-determined level of accuracy for images with corresponding parameter values. Depending on the chosen accuracy level, one may switch to different classifiers at different parameter values. Applying this strategy to the testing set TE, we obtained a smooth tradeoff curve of accuracy and speed (Figure 5).

However, doing this would require estimation of values of the defect parameters for an arbitrary input source. This is a nontrivial task. In reality, when the image is affected by many distortions, it could be impossible to isolate the effect of each. An early attempt to infer the values of defect parameters from a set of input images is discussed by Baird.[2] Such methods have higher probability of success when applied to a large collection of samples from a single source. Typically certain assumptions about the sample source have to be made, such as that images from a printed page are affected by similar distortions.

Comparatively, confusion-driven strategies are more feasible and easier to implement. Here, assuming speed is a primary concern, one would prefer to use the classifiers in this order: Bayes, C4.5, and then NN. One would use the next classifier to verify the decision only if the decided class by the previous classifier is known to be unreliable. For instance, suppose that we trust the decision by C4.5 or Bayes only if the estimated reliability for the decided class is 97.5%, and use NN otherwise. We would achieve an overall correct rate very similar to that of NN (98.03% by selection versus 98.06% by NN only), but NN would be called only 85% of the time. By changing the threshold on decision reliability, continuous tradeoff can be realized (Figure 5). We emphasize that such a tradeoff is not always achievable using a single classifier.

### *4.2 Dynamic Classifier Combination*

Suppose our goal is to maximize accuracy as well as speed, a strategy driven by classifier agreement could be used. Since C4.5 and Bayes are fast, both of them can be applied to an input image. Their decisions are used only if they agree with each other. If they disagree, NN is called and its decision is used. Using this strategy, the overall correct rate on TE would be 97.8%, better than that achievable with either C4.5 or Bayes alone, and NN would be called only 12.13% of the time, resulting in an average speed of 14.48 characters/second, which is an 8 times speed-up from that of using NN alone. A more conservative modification is that when NN is called, its decision is used only if it agrees with either C4.5 or Bayes. In such a case, the speed will remain the same, the

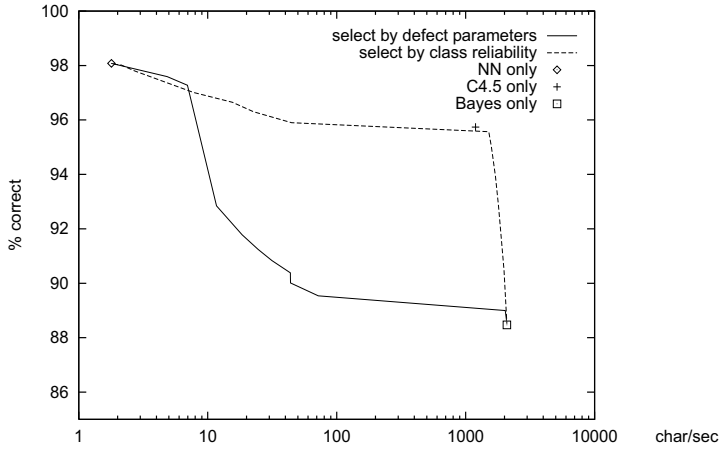



Figure 5: Accuracy and speed tradeoff when classifier is dynamically selected.

reliability of decisions would be promoted to 98.52%, but a 1.68% reject rate would have resulted.

In Section 3.3 we observed that frequency of classifier agreement has an intuitive association with image quality. Following this intuition, we partitioned TR2 into two subsets by whether C4.5 agrees with Bayes, and studied the reliability of decisions in each subset. It can be seen that the unreliable decisions occur much more often when these two classifiers disagree (Table 6).

Based on Table 6, one can design reliability-based coordination strategies similar to those discussed in the previous section. For instance, when C4.5 and Bayes agree with one another, if the decision turns out to be an 'I', it has a 7.25% chance to be wrong. This decision should be further confirmed using the NN classifier. When C4.5 and Bayes disagree and NN is called, if the decision by NN turns out to be an 'F', 'I', 'J', 'L', or 'U', it has a high ($> 25\%$) probability of being wrong, and should better be rejected. Applying these two example strategies to TE, the reliability of a decision is further improved to 98.86% with a 3.01% reject rate.

## 5 Conclusions

Our analysis of classifier accuracies in the defect parameter space indicate that differential treatment for images of different quality is reasonable. The comparison of performance of multiple classifiers in this space suggests rich

Table 6: Reliability of decisions conditioned on whether C4.5 and Bayes agree.

| when C4.5 and Bayes agree | | | | when C4.5 and Bayes disagree | | | |
|---|---|---|---|---|---|---|---|
| decision | NN | C4.5 | Bayes | decision | NN | C4.5 | Bayes |
| A | 99.75 | 99.83 | 99.83 | A | 82.87 | 42.40 | 47.90 |
| B | 99.88 | 99.39 | 99.39 | B | 95.44 | 71.59 | 6.90 |
| C | 99.49 | 99.00 | 99.00 | C | 85.34 | 72.01 | 8.45 |
| D | 99.06 | 99.67 | 99.67 | D | 89.29 | 73.58 | 18.78 |
| E | 98.38 | 99.95 | 99.95 | E | 88.99 | 69.66 | 17.85 |
| F | 97.85 | 96.87 | 96.87 | F | 68.27 | 44.21 | 16.06 |
| G | 99.71 | 99.85 | 99.85 | G | 96.26 | 80.94 | 21.55 |
| H | 99.63 | 99.15 | 99.15 | H | 93.06 | 72.70 | 12.05 |
| I | 95.70 | 92.75 | 92.75 | I | 65.80 | 61.69 | 10.14 |
| J | 99.15 | 98.63 | 98.63 | J | 58.66 | 62.79 | 10.71 |
| K | 99.78 | 99.33 | 99.33 | K | 95.42 | 59.60 | 10.50 |
| L | 98.49 | 97.86 | 97.86 | L | 66.15 | 62.74 | 7.96 |
| M | 99.87 | 99.49 | 99.49 | M | 98.58 | 89.29 | 6.06 |
| N | 99.95 | 99.64 | 99.64 | N | 95.99 | 66.55 | 12.33 |
| O | 99.19 | 99.05 | 99.05 | O | 93.04 | 77.85 | 6.83 |
| P | 99.46 | 99.63 | 99.63 | P | 83.61 | 59.28 | 42.93 |
| Q | 99.87 | 99.08 | 99.08 | Q | 94.28 | 81.46 | 6.70 |
| R | 99.69 | 99.23 | 99.23 | R | 94.96 | 68.90 | 27.08 |
| S | 99.59 | 99.91 | 99.91 | S | 85.53 | 66.32 | 12.12 |
| T | 98.87 | 99.57 | 99.57 | T | 90.43 | 86.59 | 32.99 |
| U | 99.64 | 98.45 | 98.45 | U | 72.46 | 43.48 | 13.61 |
| V | 99.51 | 98.95 | 98.95 | V | 89.81 | 74.03 | 13.88 |
| W | 99.95 | 99.64 | 99.64 | W | 99.01 | 85.47 | 21.38 |
| X | 99.89 | 99.35 | 99.35 | X | 94.35 | 72.62 | 12.41 |
| Y | 99.70 | 99.05 | 99.05 | Y | 88.33 | 70.32 | 14.56 |
| Z | 99.72 | 99.98 | 99.98 | Z | 86.05 | 58.92 | 60.00 |

possibilities for coordination strategies. Such coordination strategies may be adaptive to image quality, probability of confusion between classes, agreement of decisions by selected classifiers, or a mix of these factors. While direct estimation of image quality is difficult, selection strategies driven by decision reliability or classifier agreement are ready to use.

In our study we analyzed only the decisions of the classifiers without considering each classifier's ability to reject. We have not investigated the possibility of retraining the classifiers using selected training data, such as images from a certain region of the defect space, or images of a particular font. Other kinds of adaptive strategies can conceivably be constructed when such possibilities are taken into account.

We have left out the question of whether the images generated with the defect model are similar to those from reality. For this, we refer the readers to a recent study,[5] where experiments on validating and estimating defect models are discussed.

## Acknowledgments

I would like to thank Oscar Agazzi and Henry Baird for many helpful discussions.

## References

1. H.S. Baird, Document Image Defect Models, in H.S. Baird, H. Bunke, K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer–Verlag, 1992, 546-556.
2. H.S. Baird, Calibration of Document Image Defect Models, *Proceedings of the 2nd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 26-28, 1993, 1-16.
3. T.K. Ho, *A Theory of Multiple Classifier Systems And Its Application to Visual Word Recognition*, Doctoral Dissertation, Department of Computer Science, SUNY at Buffalo, 1992.
4. T.K. Ho, H.S. Baird, Evaluation of OCR accuracy using synthetic data, *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 24-26, 1995, 413-422.
5. T. Kanungo, H.S. Baird, R.M. Haralick, Validation and Estimation of Document Degradation Models, *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, April 24-26, 1995, 217-228.

6. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
7. *English Document Database I & II CD–ROM Set*, The Intelligent Systems Laboratory of the Department of Electrical Engineering, University of Washington, Seattle, WA, 1993.