A Statistical Combined Classifier and its Application to Region and Image Classification

Steven J. Simske
Imaging Systems Laboratory
HP Laboratories Palo Alto
HPL-2005-179
October 10, 2005*

archiving, zoning analysis, image classification, classifier, binary classification, normal, combined classifiers

A new method for combining classifiers is introduced for two problem types. (1) Archiving and re-purposing are automated using zoning analysis that performs segmentation (region boundary definition), classification (region typing) and bit-depth determination. For performance throughput reasons, zoning analysis is often performed on a low-resolution (e.g. 50-100 ppi) representation of the document. At these resolutions, heuristic metrics for classification are required. Reported here are metrics for distinguishing photos and color drawings, and a novel classification technique based solely on the statistics of each heuristic metric. The statistical technique allows ready combination of multiple binary classifiers, and provides a lower classification error than simple voting or metric-confidence techniques. This technique permits new metrics to improve the overall classification. The benefit of this technique on archival optimization is shown. (2) The classification of documents with sparse text, and video analysis, relies on accurate image classification. We herein present a method for binary classification that accommodates any number of individual classifiers. Each individual classifier is defined by the critical point between its two means, and its relative weighting is inversely proportional to its expected error rate. Using 10 simple image analysis metrics, we distinguish a set of "natural" and "city" scenes, providing a "semantically meaningful" classification. The optimal combination of 5 of these 10 classifiers provides 85.8% accuracy on a small (120 image) feasibility corpus. When this feasibility corpus is then split into half training and half testing images, the mean accuracy of the optimum set of classifiers was 81.7%. Accuracy as high as 90% was obtained for the test set when training percentage was increased. These results demonstrate that an accurate classifier can be constructed from a large pool of simple classifiers through the use of the statistical ("Normal") classification method described herein.

* Internal Accession Date Only
This technical report combines material to be presented at ICIP 2005 in Genoa, Italy, and DocEng 2005 in Bristol, UK
Approved for External Publication

A STATISTICAL COMBINED CLASSIFIER AND ITS APPLICATION TO REGION AND IMAGE CLASSIFICATION

Steven J. Simske

Hewlett-Packard Laboratories
3404 E. Harmony Road, Mailstop 85, Fort Collins, CO 80528 USA
Steven.Simske@hp.com

ABSTRACT

A new method for combining classifiers is introduced for two problem types. (1) Archiving and re-purposing are automated using zoning analysis that performs segmentation (region boundary definition), classification (region typing) and bit-depth determination. For performance throughput reasons, zoning analysis is often performed on a low-resolution (e.g. 50-100 ppi) representation of the document. At these resolutions, heuristic metrics for classification are required. Reported here are metrics for distinguishing photos and color drawings, and a novel classification technique based solely on the statistics of each heuristic metric. The statistical technique allows ready combination of multiple binary classifiers, and provides a lower classification error than simple voting or metric-confidence techniques. This technique permits new metrics to improve the overall classification. The benefit of this technique on archival optimization is shown. (2) The classification of documents with sparse text, and video analysis, relies on accurate image classification. We herein present a method for binary classification that accommodates any number of individual classifiers. Each individual classifier is defined by the critical point between its two means, and its relative weighting is inversely proportional to its expected error rate. Using 10 simple image analysis metrics, we distinguish a set of "natural" and "city" scenes, providing a "semantically meaningful" classification. The optimal combination of 5 of these 10 classifiers provides 85.8% accuracy on a small (120 image) feasibility corpus. When this feasibility corpus is then split into half training and half testing images, the mean accuracy of the optimum set of classifiers was 81.7%. Accuracy as high as 90% was obtained for the test set when training percentage was increased. These results demonstrate that an accurate classifier can be constructed from a large pool of simple classifiers through the use of the statistical ("Normal") classification method described herein.

Keywords

Archiving, Zoning Analysis, Image Classification, Classifier, Binary Classification, Normal, Combined Classifiers.

This Technical report combines material to be presented at ICIP 2005 in Genoa, Italy, and DocEng 2005 in Bristol, UK.

1. PHOTO/DRAWING CLASSIFICATION PROBLEM

During automated document scanning, segmented regions must be differentially classified (as, for example, text, drawing, photo and table regions) to ensure they are stored with appropriate resolution and bit depth, and undergo appropriate processing (sharpening, color palette selection, etc.) during their capture—this allows re-purposing of the

documents while keeping file size as small as possible.

Statistical models for classification generally distinguish text (and tabular) regions from image (photo and drawing) regions readily; for example, using projection profile information or size heuristics [2-3]. Differentially classifying photos from drawings is important because drawings benefit from sharpening, use a reduced color palette not exceeding 8-bits (one-third the normal 24-bits), and require higher resolution for repurposing than photos. Thus, distinguishing these two types of regions can lead to significant reduction in archiving overhead. Because halftone patterns are not detectable at the low (50-100 ppi) resolutions used for typical dedicated zoning engines, other metrics for differentiating these regions types must be used.

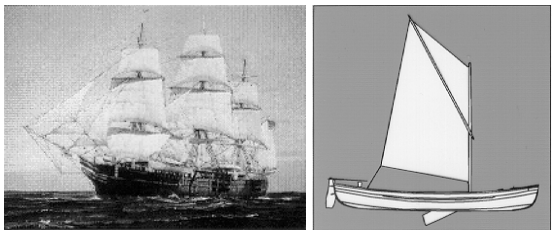

Figure 1. Sample photo region (left) and drawing region (right). The originals were in color.

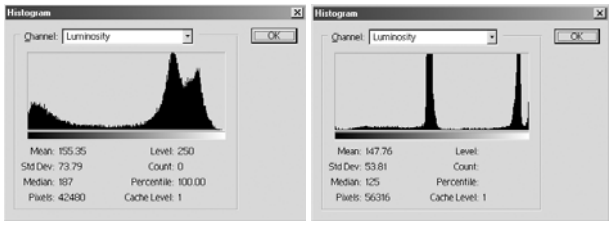

Figure 2. Histograms [1] for photo (left) and drawing (right) as shown in Figure 1.

Group	**Pct2Pk**	**Pct0.5**	**Bimod**
Photo (n=50)	0.26 ± 0.11	0.27 ± 0.11	1.0 ± 0.12
Drawing (50)	0.68 ± 0.15	0.14 ± 0.09	1.3 ± 0.25
Photo (37)	0.22 ± 0.15	0.29 ± 0.15	1.0 ± 0.15
Drawing (15)	0.73 ± 0.15	0.10 ± 0.06	1.5 ± 0.26

Table 1. Mean (μ) ± standard deviation (σ) for the metrics for photos and drawings in **training** and *test* sets.

Three metrics useful for photo/drawing classification are: (1) Percent of the histogram range in the largest two peaks (**Pct2Pk**). For this metric, the two peaks, if present, containing the largest number of pixels, are computed. Peaks are defined as containing > 0.5% of the total number of pixels and being consecutive bins between two minima (inflection points, themselves limited to 10% of the histogram range). For the photo (left) in Figure 1, the histogram in Figure 2 contains three large peaks, the largest two of which contain 30% of the pixels in the region. This value is 90% for the drawing region shown (right) in Figures 1-2. (2) Percent of histogram bins with > 0.5% of the pixels (**Pct0.5**). For the photo shown, this value is ~60%. In contrast, the **Pct0.5** value is ~10% for the drawing. (3) Bimodality of nearest-neighbor pixel differences (**Bimod**). Because photos are continuous tone and drawings comprise areas of uniform color separated by lines and curves, the drawing regions are expected to have a more bimodal appearance in comparing the differences between nearby pixels. Difference histograms for pixels 1, 2 and 3 places apart (at 75 ppi) are computed and compared to difference histograms for random pairs of pixels in the region. The ratios are summed to yield the final **Bimod** value.

For the training set of 50 photos and 50 drawings, and the test set of 37 photos and 15 drawings (all scanned at 75 ppi, all taken from the same 50 scanned pages), the mean and standard deviations of the values for each of the three metrics is given in Table 1. Statistically (t-test), the ranking of these metrics for distinguishing photos from drawings are **Pct2Pk > Bimod > Pct0.5.**

2. IMAGE CLASSIFICATION PROBLEM

The purpose of this research was to determine if a reasonably accurate binary classifier for images could be constructed from a set of simple classifiers through the use of a statistical method for classifier combination. A set of 120 images was assigned to 2 equal size classes: one of "natural" scenes, the other of "city" scenes.

These images were purposely rather diverse as relative, but not absolute, classification accuracy was of interest.

3. CLASSIFIER DESIGN AND APPLICATION TO REGION CLASSIFICATION

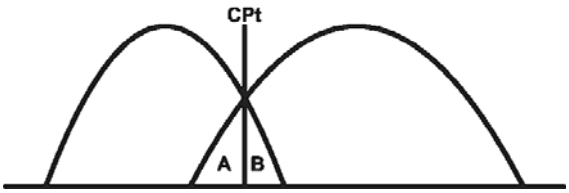

Figure 3. A=area below the critical point (CPt) of the population with the higher mean, and B=area below the critical point of the population with lower mean (B). When the population area is normalized to 1.0, A = B.

All classification involves the generation of a decision boundary [4]. Since each of the metrics described above is used as a binary threshold classifier, the decision boundary is a critical point (or threshold), designated **CPt**, falling between the means of the two curves—above **CPt** one class is assigned and below **CPt** the other class is assigned to each sample.

Figure 3 demonstrates the critical point **CPt** (vertical line segment between **A** and **B**), along with the "error areas" **A** and **B**. Thus, **CPt** is defined as the point between the two means that is equidistant from the two means in multiples of their standard deviations (the two populations need not have equal variance). That is, **CPt** is selected so that the distance from the mean of each population to **CPt** is the same in terms of standard deviations of the means. The number of standard deviations from each mean, $\boldsymbol{\sigma_{CPt}}$, is determined by:

$$\boldsymbol{\sigma_{CPt}} = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2) \qquad \textbf{Equation 1}$$

where μ_1 and μ_2 are the population means, and σ_1 and σ_2 are the population standard deviations. After computing $\boldsymbol{\sigma_{CPt}}$ (which is not the midpoint between the two means unless $\sigma_1 = \sigma_2$), the normalized area A (which is also normalized area B) in Figure 1 is determined directly from a Table of Normal curve areas ("z" value), and designated $\boldsymbol{\alpha(\sigma_{CPt})}$. With **CPt** determined, the number of correctly classified values in the training set can then be ascertained directly (Table 2).

	Pct2Pk	Pct0.5	Bimod
CPt	0.4428	0.2107	1.097
$\boldsymbol{\sigma_{CPt}}$	1.612	0.720	0.973
$\boldsymbol{\alpha(\sigma_{CPt})}$	0.0535	0.2358	0.1652
Photo (50)	46 (92%)	38 (76%)	39 (78%)
Drawing (50)	47 (94%)	41 (82%)	42 (84%)

Table 2. Normal Method data, training set.

In Table 2, the predicted accuracy is given by the value 1-$\boldsymbol{\alpha(\sigma_{CPt})}$. For **Pct2Pk**, **Pct0.5** and **Bimod**, these values are 94.6%, 76.4% and 83.5%, respectively, which are similar to the observed values of 93%, 79%, and 81% determined with reference to **CPt**. These results rely on modeling the binary classification as the problem of finding a decision boundary that is a threshold. Next the multiple classifiers need to be combined. However, the difference in metric accuracy is a problem (Table 3).

Case	Pct2Pk	Pct0.5	Bimod	p
1	R (.9465)	R (.8348)	R (.7642)	.6038
2	R (.9465)	R (.8348)	W (.2358)	.1863
3	R (.9465)	W (.1652)	R (.7642)	.1195
4	R (.9465)	W (.1652)	W (.2358)	.0369
5	W (.0535)	R (.8348)	R (.7642)	.0341
6	W (.0535)	R (.8348)	W (.2358)	.0105
7	W (.0535)	W (.1652)	R (.7642)	.0068
8	W (.0535)	W (.1652)	W (.2358)	.0021

Table 3. Truth table of probabilities for combinations of the three classifiers making the right (R) or wrong (W) classification. The overall probabilities (p) sum to 1.0.

Using the **Pct2Pk** classifier by itself (sum cases 1-4) leads to the correct classification 94.65% of the time. Using the best two out of three (sum cases 1-3 and 5) is only 94.37% accurate, as is the use of **Pct2Pk** unless the other two disagree (also the sum of cases 1-3 and 5). These are the most obvious "voting" schemes, and the results for the training data matches these predictions. Using the **Pct2Pk** classifier by

itself (Table 1, col. 2) gives 93% accuracy on the training data, while the other strategies each give 90% accuracy on the training data.

The Normal Method for classification provides the means to assign relative weights to each of the individual binary (threshold) classifiers. For 2 or more binary classifiers using the Normal Method, the metric weighting is defined to be inversely proportional to the predicted classification error rate; that is, proportional to 1/ $\boldsymbol{\alpha(\sigma_{CPt})}$. Thus, $\boldsymbol{\alpha(\sigma_{CPt})}$ x $\mathbf{W_{Mi}}$ is a constant, and so the metric weighting for the i'th classifier, $\mathbf{W_{Mi}}$, is determined from:

$$\mathbf{W_{Mi}} = [\boldsymbol{\alpha(\sigma_{CPt})}_i \text{ x } (\Sigma_{i=1..N}\{1/\boldsymbol{\alpha(\sigma_{CPt})}_i\})]^{-1}$$

Equation 2

where i=1...N and N is the number of binary classifiers to be used together. Using Equation 2, the weightings for the engines are as shown in Table 4.

	Pct2Pk	**Pct0.5**	**Bimod**
$\mathbf{W_M}$	0.6449	0.1463	0.2088

Table 4. Weightings for each classifier based on the training data, the Normal Method and Equation 2.

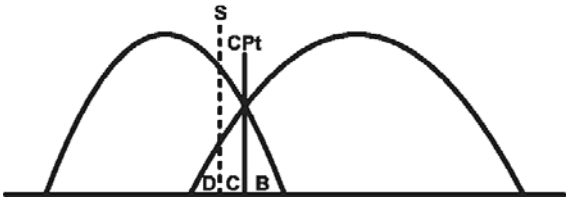

Figure 4. Assigning $\mathbf{W_C}$ for a sample S (dotted line) classified as belonging to the population with lower mean. The ratio of the areas C/(D+C) is $\mathbf{W_C}$ (Normal Method), which is then multiplied by the metric weighting, $\mathbf{W_M}$, to give the classification weight for this metric and sample (S). B and CPt are as in Figure 3.

Since the metric weight for **Pct2Pk** in Table 4 is greater than 0.5, it will outperform any possible voting scheme using the other metrics, as shown in Table 3 and in [6]. However, the Normal Method provides the means to using all three classifiers on any sample by assigning each sample a "sample confidence weight", $\mathbf{W_C}$ (Figure 4).

The value for $\mathbf{W_C}$ is determined by marking the sample value relative to **CPt**, and then determining the residual area under the normal curve for the *non-selected class*. An example will help elucidate this technique. Suppose for the **Pct2Pk** classifier, the sample value is 0.446. Since this value is greater than **CPt(Pct2Pk)**, the classification is "Drawing" rather than photo. Now, 0.446 is (0.446-0.259)/0.114 standard deviations from the Photo mean of 0.259. This is 1.640 standard deviations, for which the z-value is 0.0505. Thus, referring to Figure 4, D = 0.0505, C = (0.0535-0.0505) = 0.0030, and $\mathbf{W_C}$ = (0.0030/0.0535) = 0.0561. Multiplying $\mathbf{W_C}$ by $\mathbf{W_M}$ yields a weighted classification of $\mathbf{W_C W_M}$ = 0.0561x0.6449 = 0.0362 for Drawing for the **Pct2Pk** classifier.

The use of $\mathbf{W_C W_M}$ for each classifier thus provides a means for: (1) the *ability to use multiple binary classifiers simultaneously*, even when one classifier has a majority of the metric weighting (e.g. Table 4); and (2) a *means to rank the samples by distance from the decision boundary*.

To illustrate (1), let us further suppose that the **Pct0.5** and **Bimod** values for the example above are 0.113 and 0.80, respectively (**Pct2Pk** = 0.446). For **Pct0.5**, (0.287-0.113)/0.106 = 1.642 standard deviations from the Photo mean, for which the z-value is 0.0503 and thus $\mathbf{W_C}$ = ((0.2358-0.0503)/0.2358) = 0.7867. Multiplying by $\mathbf{W_M}$ (0.1463), the weighted Drawing classification is 0.1151 for the **Pct0.5** classifier. For the **Bimod** classifier, the 0.80 value yields a Photo classification, and so the value for $\mathbf{W_C}$ is determined from the Drawing mean: (1.34-0.80)/0.25 = 2.16 standard deviations, for which the z-value is 0.0154. Thence, $\mathbf{W_C}$ = ((0.1652-0.0154)/0.1652) = 0.9068. Multiplying by $\mathbf{W_M}$ (0.2088), the weighted Photo classification is 0.1893 for the **Bimod** classifier.

Summing these weighted classifications, the 0.1893 for Photo > (0.0362+0.1151) for Drawing, and so the region is (correctly, as it turns out) classified as Photo, even though **Pct2Pk**, the most accurate classifier, and **Pct0.5** classified the region as Drawing. Using this

technique (dubbed the Weighted Normal Method); the same 93% accuracy was obtained for the training set, although two regions were classified differently than using **Pct2Pk** by itself (Table 5). The $\mathbf{W_C W_M}$ classification strategy provided the best overall accuracy (94.2%) for the test set of 52 regions. For the **Pct2Pk**, **Pct0.5** and **Bimod**, classifiers, the accuracies on the test set were 90.4%, 76.9% and 71.2%, respectively.

	$\mathbf{W_C W_M}$ classification
Photo (50)	46 (92%)—two changed from **Pct2Pk**
Drawing (50)	47 (94%)—none changed from **Pct2Pk**

Table 5. Results, training set, Weighted Normal Method.

4. ARCHIVING OPTIMIZATION

Using the weighted classification results allows the ranking of samples based on their distance from the decision boundary. The value for the example above is "Drawing, 0.1893-0.1513" or "Drawing, 0.0380". Another sample with values of 0.434, 0.160 and 0.98 for **Pct2Pk**, **Pct0.5** and **Bimod**, respectively, has a bias of "Photo, 0.1129". These relative values can be used to rank regions by how likely, statistically, they are to be mistyped. Regions within a range of relative values can then be preserved at sufficient resolution and bit depth to allow full repurposing. For example, photos may be preserved at 200 ppi, 24 bit, and color drawings at 300 ppi, 8-bit paletted color. For regions near the decision boundary, a full-ppi, full-bit depth representation (300 ppi and 24-bit, which increases file size by either 225% or 300%) can be used to ensure that the effects of mistyping will be avoided for these "borderline" classifications. This provides a balance between the need for reduction in file size on one end and the need to preserve region re-purposability on the other. If we use the full representation for the 20% of regions nearest the border between classes, and the (correct) classification (all of the error cases for the $\mathbf{W_C W_M}$ approach fell within this band) for all of the other test cases, archiving the test set requires 94.9 MB. Correctly classifying each region, by contrast, requires 71.9 MB. However, saving all regions at the full representation requires 176 MB (In general, the storage savings will be 100%-the range in which all the errors occur). All of the regions left as "unclassified" can be repurposed with a user interface (UI) based tool [7], if desired, or left at full resolution and full bit depth.

5. DISCUSSION AND CONCLUSIONS, REGION CLASSIFICATION

Several heuristic metrics for classifying photos and drawings are presented. The **Pct2Pk** metric provides > 90% accuracy in classification at low resolution (75 ppi). The other two (**Pct0.5** and **Bimod**) are less accurate, and in voting combination do not provide any improvement over using **Pct2Pk** by itself. The novel Weighted Normal Method for classification provides the means to combine multiple binary classifiers of highly varying accuracy, and allows the identification of "uncertain" cases so that they can be archived at full resolution and bit depth. The method described herein requires no data scaling, no kernel definition, and is easily updated when new training data is obtained—it can thus be deployed in environments where user feedback can be used to improve future workflow performance.

Though the training and test data are shown assumed to be Gaussian in distribution for the model of Figures 3-4, similar results can be obtained by applying the $\mathbf{W_C W_M}$ techniques in a classification that "trusts" all of the data more (similar techniques are taken in determining the kernel to use in kernel methods [4]). In this simplified technique, the decision boundary is set based on minimizing the errors in the training set, and so the boundary is selected to provide the fewest overall errors. On the training set, these boundaries were 0.477, 0.243 and 1.155 for the **Pct2Pk**, **Pct0.5** and **Bimod**, classifiers, respectively, with (implicitly)

improved training accuracies of 96%, 82% and 89%. On the test data, these classifiers had 92.3%, 71.2% and 82.7% accuracy, similar to the Normal Method results. This is likely due to the use of a smaller test than training set, but provides a potential alternative method to define **CPt** for certain classes. Extending this simplified technique, the $\mathbf{W_C W_M}$ strategy can be deployed (Equation 2) for relative weighting (.6306, .1401 and .2293 for the **Pct2Pk**, **Pct0.5** and **Bimod** classifiers, respectively) and using the relative location of sample values in the "error values" of the *non-selected class* to compute the values for $\mathbf{W_C}$. That is, if there are four error values and the sample value falls in the middle of these four, then $\mathbf{W_C} = 0.5$ for this value. Deploying this strategy to the test data resulted in 90.4% accuracy (less than that of the non-simplified method, but again with sparse data sets). Further investigation is necessary to explore other reasonable means of assigning $\mathbf{W_C}$ and $\mathbf{W_M}$ values and to compare to other [4-5] classification methods.

A practical use of this classification technique is now considered. Using this classification scheme together with the full-resolution, full-bit depth approach to regions not readily assigned to either group offers an advantage to archiving. Predicted storage savings for a very large corpus [8] are nearly 50%, while photo and drawing regions will be archived with desired resolution and bit depth. The "failure" rate can be set at any desired level of statistical certainty using the method described herein.

6. CLASSIFIER DESIGN AND APPLICATION TO IMAGE CLASSIFICATION

To extend the work on this classifier, software (GOSSIP, or Gaussian Ordering System for Statistical Image Processing) was generated and a new problem domain (image classification) approached. Ten simple metrics for image comparison were implemented—none were expected to perform as well as the best methods observed for this type of classification problem [9]. They are (1) image entropy, (2) standard deviation of the image histogram, (3) the percent of the image classified as edges (pixels with nearest neighbor variance above a threshold), (4) mean edge value (sharper edges have higher values), (5) mean nearest neighbor variance across the image, (6) mean region size after segmentation with the edge pixels, (7) mean variance within these regions, (8) mean image saturation, (9) mean region size after segmentation by unsaturated pixels, and (10) the mean variance within these regions. These are all very simple metrics, which require no image model, no a priori information about the images, and moreover were calculated on relatively small (60 each) class sizes. The original images were all JPEG images, 2048 x 1536 pixels in size, taken with a digital camera (HP R707).

Metric	Predicted Accuracy	Observed Accuracy	Weight
1. Entropy	0.581	0.558	0.0956
2. ImageHistStd	0.512	0.658	0.0821
3. PctEdges	0.518	0.558	0.0832
4. MeanEdge	0.614	0.625	0.1039
5. MeanPixVar	0.506	0.533	0.0811
6. MeanRegSize	0.625	0.700	0.1070
7. MeanRegVar	0.593	0.667	0.0984
8. MeanSat	0.717	0.783	0.1418
9. MeanSatRegSize	0.613	0.683	0.1037
10. MeanSatRegVar	0.612	0.700	0.1034

Table 6. Simple image metrics used and their relevant statistics for the full set of 120 (60 "natural, "60 "city" images). Weight is computed using Equation 2.

The results for these 10 metrics are shown in Table 6. The mean image saturation was the "best" of these simple classifiers, providing 94/120 correct classifications. Due to the non-Gaussian distribution of its data, however, its predicted accuracy was 0.717, or 71.7%.

Combination Set {from Table 1}	Observed Accuracy
1. Best 2 simple classifiers {6,8}	0.783
2. Best 3 simple classifiers {6,8,10}	0.792
3. Best 4 simple classifiers {6,8,9,10}	0.750
4. Best 5 simple classifiers {6,7,8,9,10}	0.817

5. Best combination {1,4,6,8,10}	0.858
6. Best 3 + {1}, or {1,6,8,10}	0.775
7. Best 3 + {4}, or {4,6,8,10}	0.825
8. Best + {4}, or {4,8}	0.783

Table 7. Some combined classifiers and their accuracies on the entire 120-image corpus (compare to the accuracies of single simple classifiers as shown in Table 6).

Next, $2^N - 1$ (where N=10, the number of simple classifiers), or in this instance 1023, distinct classifiers were created by using these ten metrics in all possible combinations. Table *6* shows the results for 10 of these—that is, the 10 simple classifiers themselves. The best combination was the set (represented in {}) of simple classifiers {1,4,6,8,10}, even though {1,4} have lower accuracy than, for example {2,7,9}. Results for some of these combined classifiers are given in Table *7*.

Table *7* shows that the single best simple classifier {8} provides an accuracy of 78.3%. In combination with any other simple classifier, the same accuracy is obtained. Only after three classifiers are combined does the accuracy rise to 79.2% (e.g. for {6,8,10}, or {4,8,10}). When a fourth classifier is added, the peak accuracy rises to 82.5% (for {4,6,8,10}). Five classifiers happen to provide the highest overall accuracy at 85.8% for the combination of {1,4,6,8,10}. Any combinations of six or more classifiers reduces the accuracy from this set of five. The combination of {1,4,6,8,10} reduces the error rate of the best single, simple classifier by 34.6%.

Although our corpus was small, we divided it into two equal-sized sets for training followed by testing. We used two groups of size 30 for training and testing, and then reversed them. Averaging the two sets, we found that the best single classifier (mean image saturation) performed well in testing (at 78.3%, the same as when the entire corpus is considered for the training set). Deploying the best classifier combination for the training data to the test data improved the classification accuracy to 81.7%. Interestingly, the best classifier combinations for these smaller sets each included classifiers 4 and 8, emphasizing the utility of edge-based and saturation-classifiers for this particular domain of classification. The "optimal" combination from the feasibility testing (the combination of {1,4,6,8,10}) also provided 81.7% mean accuracy in testing, though it is worth noting that sparser training data in these runs altered the statistics of the Normal classifier. During testing, some classification combinations provided accuracy as high as 90%.

7. DISCUSSION, IMAGE CLASSIFICATION

The primary goal of this research was to demonstrate that combinations of simple classifiers can provide a reasonable binary image classifier. However, because one simple classifier (mean image saturation) was considerably more accurate than the other 9 simple classifiers, the results also demonstrate that a reasonably accurate classifier can be further improved in combination with relatively inaccurate classifiers. Four such classifiers—with accuracies ranging from only 55.8% to 70%–in combination with the highest accuracy classifier (mean image saturation) reduced the classification error rate by 34.6%. The resulting classifier has a respectable "feasible" accuracy (85.8%) for such a small corpus (120 documents). Experiments with half of the corpus used for training and half for testing ("unseen") demonstrated that even sparse training sets (30 images per classification) can be used with the method for classifier combination described herein. The error rate during testing was reduced by 15.4% in comparison to the single best classifier.

This paper demonstrated that a large population of simple classifiers can be used to provide an effective domain-specific ("semantic") classifier. The ten image metrics chosen here were not expected to have excellent

discriminative power, and with the exception of (8) mean image saturation, this was the case. In spite of their limitations, a classifier was derived with 85.8% accuracy, even for the limited (120 image) corpus studied.

The optimal classifiers are not simply weighted combinations of the best classifiers. Rather, the optimal combinations are affected by the solution space and, likely, the relative utility of the different classifiers to help "cover" the overall solution space. For example, it is interesting that the optimal set, {1,4,6,8,10} represented five quite distinct metrics. Simple classifier pairs with higher expected correlation (e.g. {6,7} and {8,9}) did not both belong to this optimal set, in spite of the higher accuracies of {7,9} in comparison to {1,4}. This may be analogous to the enhanced performance of some genetic algorithms when both the fittest and least fit strings are propagated to the next generation.

To explore this, the percent of classification decisions shared in common by classifiers {1}, {4}, {7} and {9} with the best three classifiers {6,8,10} were computed. Not surprisingly, for {1}, {4}, {7} and {9} when compared to {6,8,10}, the percentage of classifications in agreement were 55.2%, 61.9%, 68.3% and 70%, respectively. {6,8,10} themselves averaged only 61.7% of their classifications in common, far lower than their mean classification accuracy of 72.8%. Thus, we believe that lack of correlation amongst classifiers may be beneficial in the optimal set of classifiers.

An interesting finding was that the metrics consistently provided greater observed accuracy (Table 6) than predicted accuracy (9 out of 10 times, with mean observed accuracy of 0.647 substantially higher than the mean predicted accuracy of 0.589). This is likely a result of the Gaussian fitting of non-normal data.

One previous study [9] reported on "city vs. landscape" classification. They found color histograms, color coherence vector DCT coefficient, edge direction histogram, and edge direction coherence vectors to have high discriminative power in a weighted k-NN classifier, and obtained an accuracy of 93.9% on a 2716 image corpus. Our future work in this area will focus on comparing the Normal method to k-NN, SVM [4], boosting [10] and other classification methods directly. This requires extension from binary to N-classes. While our method shares common ground with boosting [10], it differs in that it involves no iterative weighting of the training sample, simple coefficient adjustment when new ground truth data is obtained, a linear discrimination boundary, and no reliance on optimization theory.

8. ACKNOWLEDGMENTS

The author thanks Jason Aronoff and Dalong Li for helpful feedback on this work.

8. REFERENCES

[1] Adobe Photoshop 6.0, 1989-2000 Adobe Systems Inc.

[2] Revankar, S.V. and Fan, Z. “Image segmentation system”, U.S. Patent 5,767,978, January 21, 1997.

[3] Wahl, F.M., Wong, K.Y. and Casey, R.G. “Block segmentation and text extraction in mixed/image documents,” Computer Vision Graphics and Image Processing, Vol. 2, pp.375-390, 1982.

[4] Schölkopf, B. and Smola, A.J. “Learning with Kernels”, The MIT Press, Cambridge MA, 2002.

[5] Namboodiri, A.M. and Jain, A.K. “Online handwritten script recognition,” IEEE Trans Pattern Analysis Machine Intell, Vol. 26, no. 1, pp. 124-130, 2004.

[6] Lin, X., Yacoub, S., Burns, J. and Simske, S. “Performance analysis of pattern classifier combination by plurality voting,” Pattern Recog Letters, Vol. 24, pp. 1959-1969, 2003.

[7] Simske, S.J. and Sturgill, M. “A ground-truthing engine for proofsetting, publishing, re-purposing and quality assurance,” Proc. DocEng2003, pp. 150-152, 2003.
[8] Simske, S.J. and Lin, X. "Creating digital libraries: content generation and re-mastering," Proc. DIAL '04, pp. 33-45, 2004.
[9] Vailaya, A., Jain, A. and Zhang, H.J. "On image classification: city vs. landscape", Proc. IEEE Workshop Content-Based Access Image Video Lib., pp. 3-8, 1998.
[10] Freund, Y. and Schapire, R. "A decision theoretic generalization of on-line learning and an application to boosting", J. Comp. Syst. Sciences 55, pp. 119-139, 1997.