Score Aggregation from Multiple Sources and Training in the Context of Lexicon Reduction using Holistic Features

Sriganesh Madhvanath and Venu Govindaraju

Center of Excellence for Document Analysis and Recognition (CEDAR)
Department of Computer Science (CEDAR)
State University of New York at Buffalo
Amherst, New York 14228–2567
Tel: 716 645 6164 (x103)
Fax: 716 645 6176
email: govind cedar.buffalo.edu

Abstract

Holistic methods developed for small, static lexicons are not easily extended to the large and dynamic lexicon scenario owing to word-level feature variability and paucity of training samples.

A methodology of coarse holistic features and heuristic prediction of ideal features from ASCII is proposed to address these issues. The proposed methodology is based on the axiom that real-world examples of handwritten words may be viewed as the ideal exemplar of the word class distorted by the scriptor, stylus, medium and intervening electronic imaging processes.

On a test set of 3,000 handwritten city names, we achieved a 70% reduction in the lexicon size (1,000) with 98.7% accuracy.

1 Introduction

In this paper, we consider the application of the holistic paradigm to recognition scenarios involving large or dynamically generated lexicons. A methodology of coarse features and heuristic prediction of ideal features is proposed as a solution to the problems of word-level feature variability and paucity of training samples, and the applications of lexicon reduction proposed.

The holistic paradigm of matching features [1] extracted from the word at a whole-word level has found successful application in recognition scenarios involving small, static lexicons such as reading legal amounts on checks [2]. The paradigm is not easily extended to large and dynamic lexicons for two chief reasons:

1. *Word-level feature variability*: Every word in the lexicon is a pattern class for the holistic paradigm. Given any set of holistic features, a large degree of intra-class variation may be expected in the offline context wherein writing is unconstrained and images contain noise from the medium and the surroundings. When the classes are numerous or dynamically determined, the separability of word classes in feature space is considerably diminished, and overwhelmed by the intra-class variation. This is in contrast to machine print, wherein intra-class variation is limited by the number of font sizes and types considered.

2. *Paucity of training samples*: A small static lexicon generally allows the collection of training samples of each word class, and the use of statistical and syntactic methods for classification.

   In the large/dynamic lexicon scenario, it is clearly not possible to collect enough samples of each word to represent even the common writing styles. In related domains such as machine print and online recognition, handwriting models have been used to synthesize samples for training. However the variations in binary offline images of unconstrained writing are a result of interactions between writer, medium, instrument and subsequent scanning, binarization and segmentation from the background and surrounding text, and are consequently difficult to capture using models of handwriting. Thus the large intra-class variation in handwritten text is also the reason why realistic simulation of offline handwriting samples is presently not viable.

Real-world examples of a handwritten word may be modeled as distortions of an “ideal” word exemplar of the word class. The ideal exemplar is pure cursive, devoid of baseline skew and character slant, and exhibits evenly spaced reference lines, *i.e.*, all ascenders and descenders are of uniform height, and their extent is equal to the width of the middle zone. This notion of ideal exemplar described conforms to the style adopted for writing instruction imparted in elementary school. The writing style adopted by an individual in adulthood may thus be regarded implicitly as a distortion of the cursive ideal learned as a child.

Recognition of real exemplars of a given word class as distortions of the ideal exemplar is assisted by a reduction in the distance between the former and the latter in holistic feature space. This may be achieved by a combination of techniques designed to counteract or compensate for different forms of distortion. These include normalization, the use of distortion-independent features and extraction algorithms, and training.

For example, the presence of perceptual features such as ascenders and descenders is relatively unaffected by discreteness of writing, their detection can be made invariant with respect to character slant or stroke width, and the use of a variable, rather than a fixed grid for registering them yields feature positions that are relatively independent of scale.

Training refers to the process of constructing models of each of the word classes, by extracting features from a representative training set of real exemplars of each word class which implicitly capture different types of distortion. However, it is difficult in practice to obtain a training set that captures all manifestations of distorting influences (for example, all degrees of skew, noise, stroke width) that may be encountered in a test image.

2 Our Approach

We propose the following hybrid approach to address the external influences that distort the ideal shape of a handwritten word :

- Preprocessing for slant normalization and noise removal [2] are the first steps. (Figure 1).
- Angular or local reference lines for invariance with respect to baseline skew (Figure 2) are estimated [3].
- Perceptual features are visually conspicuous features of the word shape that have been cited in reading studies as being utilized in fluent reading, and include ascenders, descenders and length. Their robustness with respect to writing style makes them suitable for unconstrained writing. A graph-based framework [4] for the representation and matching of perceptual features is adopted.

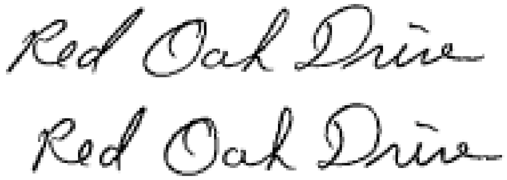

Figure 1: Phrase image: (a) original, (b) following slant normalization

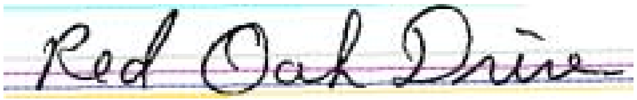

Figure 2: Angular reference lines at estimated baseline skew: the center-line bisects the zone bounded by the baseline and the halfline.

- In the absence of training samples of the word classes and a handwriting model to synthesize them, features of the ideal word exemplar obtained by a process of heuristic prediction Image features are matched with their ideal lexicon counterparts using distortion models and matching schemes appropriate to the representation. In particular, perceptual features (*e.g.*, Figure 3) are matched using a constrained bipartite graph matching algorithm [4] and stroke-based structural descriptions are matched using elastic matching techniques.

3 Lexicon Reduction System

Lexicon reduction refers to the rapid elimination of lexicon entries that are unlikely to match the given image, prior to recognition. The objective of the system is to obtain a score for each lexicon entry. Match scores computed for the different feature categories are combined by a process of score aggregation.

3.1 Image feature extraction

The only features extracted after pre-processing and normalization are local maxima on the upper contour and local minima on the lower contour (Figure 4). The lower minima divide the image vertically into segments, and length of the word is estimated as the number of segments. Candidate ascenders and descenders are identified from among the extrema and their

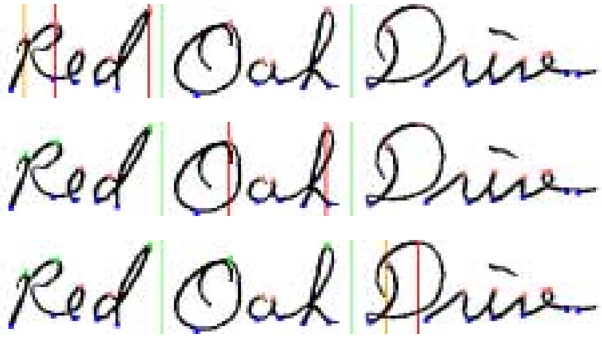

Figure 3: Lexicon-driven matching of ascenders using POSMATCH. Matches with normal and optional lexicon ascenders are shown in different shades.

positions are expressed in terms of Holistic Segment Distance. Each position is of the form $x.y$, where x is the segment number, and y is the offset into the segment computed as a fraction of the width of the segment (Figure 5). This method preserves continuity of position across segment boundaries while affording greater precision than the segment number alone.

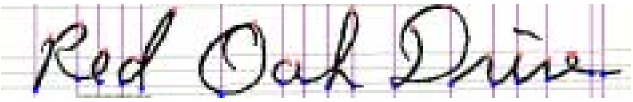

Figure 4: Vertical grid imposed by lower contour minima

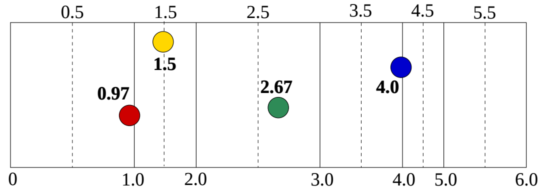

Figure 5: Positions of features specified in terms of Holistic Segment Distance. The distance is linear in between consecutive grid lines. The grid itself is variable and determined by local minima on the lower contours of the handwritten phrase.

The philosophy of "over-extraction" is implemented and more candidates are identified than necessary. Using logistic functions confidences of candidate ascenders and descenders are computed from the normalized heights of the extremum in question, and those of the flanking extrema.

3.2 Matching

In order to match coarse features extracted from the image against lexicon entries, it is necessary to "invert the lexicon", *i.e.*, have available the holistic features corresponding to each lexicon entry. The "ideal" features of the lexicon entries are *predicted* directly from ASCII using a set of heuristic rules. The features extracted from the word image are viewed as distortions of the idealized features captured by the inverted lexicon. Tables 1& 2 show the "predicted" features with frequencies for the uppercase and lowercase characters using our feature extraction modules.

The extracted length of the image is compared with the predicted length of a given lexicon entry and a length score *lmatch* is computed as a function of their difference and their arithmetic mean.

Positional features - ascenders and descenders - are matched separately. Let us denote the *Goodness of Match* score, *Similarity of Position* score, *Similarity of Confidence* score and *Degree of Mismatch* score resulting from the matching of ascenders by *amatch*, *apmatch*, *acmatch* and *aunmatch* respectively, and the corresponding ones for descenders by *dmatch*, *dpmatch*, *dcmatch* and *dunmatch*. When the image has no candidate descenders and the lexicon entry has no predicted descenders, all of the descenders scores are zero. In general, a value of zero for all scores corresponding to a particular feature category indicates that the feature category in not relevant to the matching task.

3.3 Score aggregation

The objective of score aggregation is to compute for each lexicon entry, a composite score from the various match scores obtained from different feature caterories. Score aggregation is performed sequentially at different levels, in a cumulative manner, as indicated by Figure 6.

Confidences of individual feature candidates are computed from image-based measurements such as normalized heights of maxima using logistic functions, as described earlier under the heading of feature extraction.

Feature candidates extracted from the image are matched with predicted features of the given lexicon entry using the bipartite graph matching scheme developed for perceptual features, and various match scores are computed.

A single score is computed for each feature category from the corresponding match scores. This computation is modeled by logistic functions. Equations below provide the *lscore*, *ascore*, and *dscore* respectively.

Chr	**Freq**	**#Minima**				**#Desc**		**#Maxima**				**#Asc**		
		1	2	3	4	0	1	1	2	3	4	0	1	2
a	238	32.4	66.0	1.7	0.0	100.0	0.0	74.4	24.7	0.9	0.0	100.0	0.0	0.0
b	13	76.9	23.1	0.0	0.0	100.0	0.0	38.5	53.8	7.7	0.0	0.0	92.3	7.7
c	57	93.0	7.0	0.0	0.0	100.0	0.0	98.2	1.8	0.0	0.0	100.0	0.0	0.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	
i	169	85.2	14.2	0.6	0.0	100.0	0.0	97.5	2.5	0.0	0.0	99.4	0.6	0.0
j	1	100.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0
k	39	12.8	76.9	10.3	0.0	100.0	0.0	39.0	58.5	2.4	0.0	2.4	97.6	0.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	
w	43	16.3	81.4	2.3	0.0	100.0	0.0	7.3	14.6	75.6	2.4	100.0	0.0	0.0
x	3	66.7	33.3	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
y	35	31.4	62.9	5.7	0.0	2.9	97.1	19.4	77.4	3.2	0.0	100.0	0.0	0.0
z	2	50.0	50.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0

Table 1: Character class frequencies and class-wise frequency distributions of number of minima, maxima, ascenders and descenders in lowercase characters from 340 unconstrained street name images

Chr	**Freq**	**#Minima**				**#Desc**		**#Maxima**				**#Asc**		
		1	2	3	4	0	1	1	2	3	4	1	2	3
A	134	6.7	91.8	0.7	0.7	100.0	0.0	83.3	15.2	0.8	0.8	99.2	0.8	0.0
B	47	19.1	76.6	4.3	0.0	100.0	0.0	70.2	27.7	2.1	0.0	100.0	0.0	0.0
C	59	100.0	0.0	0.0	0.0	100.0	0.0	95.1	4.9	0.0	0.0	100.0	0.0	0.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	
I	3	100.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	75.0	0.0	0.0
J	7	85.7	14.3	0.0	0.0	42.9	57.1	100.0	0.0	0.0	0.0	87.5	0.0	0.0
K	13	15.4	69.2	15.4	0.0	100.0	0.0	25.0	58.3	16.7	0.0	25.0	75.0	0.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	
X	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Y	1	0.0	100.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0
Z	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Character class frequencies and class-wise frequency distributions of number of minima, maxima, ascenders and descenders in uppercase characters from 340 unconstrained street name images

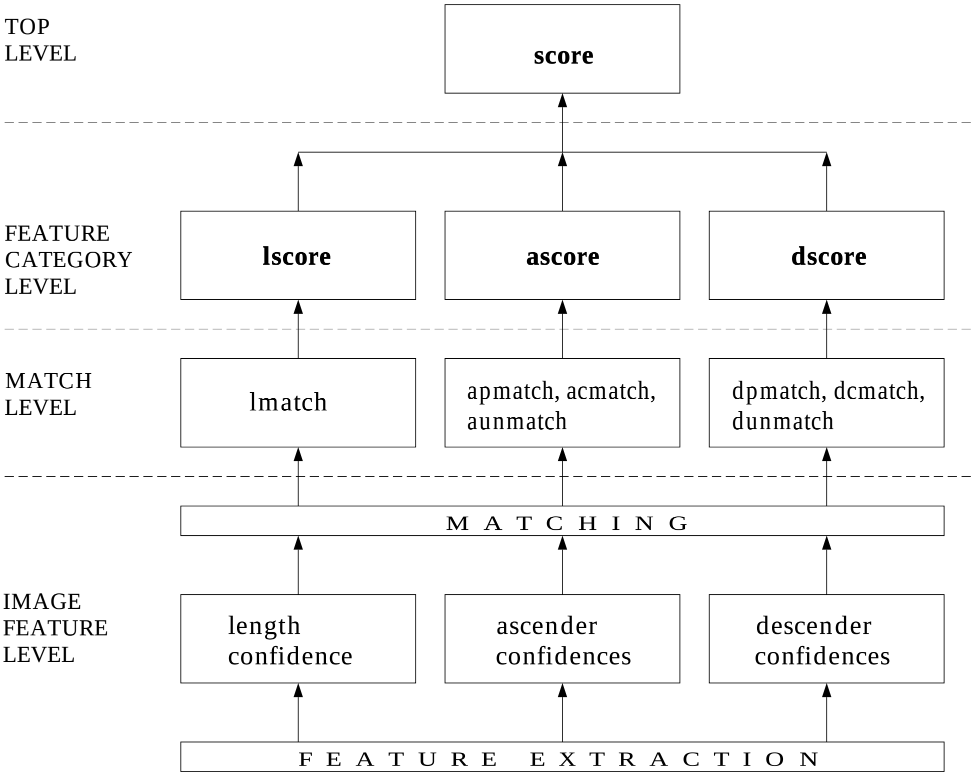

Figure 6: Cumulative score aggregation : scores at each level are computed from scores at the previous level

$$\frac{1}{1+e^{L+L_m.lmatch}}$$

$$\frac{1}{1+e^{A+A_m.amatch+A_p.apmatch+A_c.acmatch+A_u.aunmatch}}$$

$$\frac{1}{1+e^{D+D_m.dmatch+D_p.dpmatch+D_c.dcmatch+D_u.dunmatch}}$$

In practice, due to the correlation between the *Goodness of Match* score on the one hand and the *Similarity of Position* and *Similarity of Confidence* scores on the other, some of the weights are very close to zero, and the corresponding terms may be ignored in the computation.

The overall *score* is computed from the feature category scores *lscore*, *ascore* and *dscore* :

$$score = \frac{1}{1+e^{S+S_L.lscore+S_A.ascore+S_D.dscore}}$$

3.4 Weight estimation

We have used the logistic function for score computation at every level. The logistic function has the advantages of being continuous and bounded in the interval [5], and is particular suited for the modeling of the correlation between one or more measurements and a dependent binary response variable. It may be computed by Logistic Regression on data points each composed of measured values of the variables and the desired binary response. Logistic Regression computes the weights associated with each of the measurements which maximize the likelihood of the binary response given the measurements, and is conveniently available as a procedure in the SAS statistical package. Unlike Bayesian estimation wherein the measurements are often assumed to be independent in order to simplify the computation, Logistic Regression is capable of dealing with correlated measurements [5].

No training is required for the computation of match scores, which reflect the cost incurred in matching the image features with the predicted features of the lexicon, and are a byproduct of the bipartite graph matching procedure for perceptual features [4].

A set of 1400 binary images of citynames extracted from US mail was used for training. Each image in the set is matched with the entries in a small dynamically generated lexicon composed of the ASCII truth of the image along with a small number of other strings randomly extracted from a static list of approximately 1700 citynames.

The length score obtained by each lexicon entry is combined with a boolean flag denoting whether or not the entry is the truth, to form a data point for Logistic Regression. Logistic Regression is then used to estimate the weights L and L_m in the expression for *lscore*.

The weights A, A_m, A_c, A_u for *ascore* and the corresponding weights for *dscore* are estimated by Logistic Regression in an identical manner.

The same set of word images and lexicons is used as for the preceding level. Given an image and the corresponding dynamically generated lexicon, the feature category scores *lscore*, *ascore* and *dscore* are computed. A boolean flag associated with the set of scores denoted whether or not the lexicon entry is the truth of the image. The data points are partitioned into two sets by the relevance or irrelevance of the descender feature category. The weights S, S_L, S_A, S_D, and S', S'_L, $S'_A$are estimated by Logistic Regression on the two partitions.

It is also possible, in theory, to compute the overall score directly from the various match scores for each of the feature categories, and bypass the feature category level of score aggregation. However the training data available is not sufficient for reliable estimation of weights. We have therefore taken the two-step approach of estimating the likelihood of a class being the true class based upon each feature category independently, and combining these scores into one overall score.

3.5 Pruning strategy

A bound on the maximum permissible difference between the estimated length of the image and the predicted length of lexicon entries is used to eliminate a subset of the lexicon. The score aggregation procedure computes for each surviving lexicon entry an overall score. Given the scored lexicon, pruning or reduction involves identification of the entries from the lexicon that are similar to the image, and may be accomplished in several ways.

A simple rank based strategy would be to consider the top N entries of the ranked lexicon to be the reduced lexicon. This strategy has the advantage of yielding reduced lexicons of constant size, for a subsequent classification stage. However when scores are available, it is important to keep in mind that the scores effectively partition the lexicon entries into equivalence classes such that two lexicon entries within an equivalence class have the same score. A reasonable constraint on any pruning scheme is that the reduced lexicons generated by the scheme consist of a whole number of partitions. It is not intuitive that there are two lexicon entries X and Y with the same score such that X is included in the reduced lexicon, while Y is not.

The pruning strategy we have used for evaluation of the system's reduction performance is essentially rank-based, but is careful not to split an equivalence class in determining the reduced lexicon :

> Given a scored and ranked lexicon, and a rank threshold T_r, the bound on length is employed implicitly and scores are computed for the lexicon words with permissible length. These words are ordered by score, and all entries of rank greater than T_r are discarded. Finally, all entries in the same score-equivalence class as the $(T_r + 1)$th entry are also *discarded*, to obtain a reduced lexicon whose size is never greater than the threshold T_r.

The reduction ρ and accuracy α achieved by the system are functions of the rank threshold T_r.

4 Experimental results

1. *Fixed lexicon*
A set of 768 lowercase images of city names from actual mailpieces was used as the test set. Seven of the images were rejected due to poor thresholding. The lexicon was a randomly generated list of 1000 city names. The results are tabulated in the table below.

Total	**Rejects**	**Top 100**	**300**	**500**	**700**
768	7	571	700	747	758
%	0.9	75.0	92.1	98.2	99.6

As apparent from the table, the first 500 in the ranked lexicon contained the truth value in 98.2% of the cases. Thus the system presently achieves 50% reduction in the size of the lexicon with under 2% error.

2. *Dynamic lexicon*
A set of 700 images of citynames at 200 ppi was used for the evaluation of lexicon reduction performance. With each image was associated a dynamically generated lexicon of 1000 words comprised of the truth and 999 others selected randomly from a list of 3000 citynames. The images are unconstrained with respect to writing style, writing medium and instrument, contain noise from other lines in the address block, and all of the artifacts of suboptimal binarization including fragmentation and missing characters.

T_r	α	T_r	α
10	30.7	90	80.7
20	43.9	100	82.8
30	53.7	150	90.9
40	61.1	200	95.8
50	67.3	250	96.9
60	72.5	300	98.1
70	76.9	350	98.4
80	79.5	400	98.4

Table 3: Table 1: Accuracy α (expressed in percent) for different choices of T_r

An image is said to be rejected when either the reduced lexicon is empty, or when the top ranked entry has an associated score of zero. Reasons for rejection include (i) feature extraction failure : erroneous reference lines, greatly over- or under-estimated length, (ii) missing characters and fragmentation in the image, and (iii) poor ascender or descender scores because of uppercase hand-printing or writing styles wherein these features are not significantly above or below the body of the word. 85 out of 700 test images (12.1%) are rejected by the system.

The accuracy of reduction α for different values of T_r in the range [0,400] is tabulated in Table 3, and plotted in Figure 7.

Accuracy figures are computed as fractions of the number of accepted cases, 615. Reduction accuracy is seen to saturate at a value of 98.4% above $T_r = 350$. The implicit restriction on length reduces the size of the lexicon from a 1000 to 287 on the average, but also eliminates the truth from the lexicon in 1.6% of the accepted cases. For unconstrained handwritten word images of poor quality extracted from live mail, the system achieves a reduction of 70% in the size of the lexicon with less than 2% error.

5 Summary

A lexicon reduction system based on perceptual features for unconstrained isolated handwritten word is described. The process of score computation may be seen as one of cumulative aggregation over the image feature, feature match, feature category and overall score levels. At each level except for the feature matching level, scores are computed as logistic functions of scores or measurements at the preceding level. The overall score combines the independently computed contributions of the relevant feature categories.

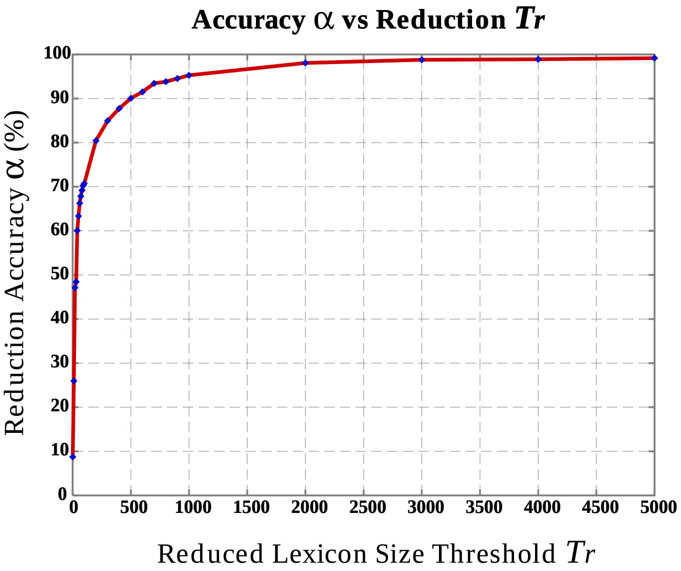

Figure 7: Accuracy of reduction for different choices of threshold T_r

Empirical results are presented for the reduction of dynamically generated lexicons of 1000 words given low-quality images of city names extracted from the US mail. A reduction of 70% in the size of the lexicon is achieved with under 2% error.

References

[1] R.N. Haber L.R. Haber and K.R. Furlin. Word length and word shape as sources of information in reading. *Reading Research Quarterly*, 18:165–189, 1983.

[2] G. Kim and V. Govindaraju. Efficient chain code based image manipulation for handwritten word recognition. In *Proc. of the SPIE symposium on electronic imaging science and technology (Document Recognition III), San Jose, CA*, volume 2660, pages 262–272, February 1996.

[3] S. Madhvanath and V. Govindaraju. Contour-based image processing for holistic handwritten word recognition. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR-97), Ulm, Germany*, August 18-20, 1997.

[4] S. Madhvanath and V. Govindaraju. Holistic lexicon reduction. In *Proceedings of the Third International Workshop on Frontiers in Handwrit-*

ing Recognition, (Buffalo, New York, May 25-27), pages 71–81, 1993.

[5] T.K. Ho. *A Theory of Multiple Classifier Systems and its Application to Visual Word Recognition*. PhD thesis, State University of New York at Buffalo, 1992.