July 2017

Multi-Character Field Recognition for Arabic and Chinese Handwriting

Daniel Lopresti
Lehigh University, lopresti@cse.lehigh.edu

George Nagy
Rensselaer Polytechnic Institute, nagy@ecse.rpi.edu

Sharad C. Seth
University of Nebraska-Lincoln, seth@cse.unl.edu

Xiaoli Zhang
Rensselaer Polytechnic, zhangxl@rpi.edu

Summit on Arabic and Chinese Handwriting (SACH), pp. 93-100, 2006.

Multi-Character Field Recognition for Arabic and Chinese Handwriting

Daniel Lopresti
Lehigh University
Bethlehem, PA 18015
lopresti@cse.lehigh.edu

George Nagy
Rensselaer Polytechnic
Institute DocLab
Troy, NY 12180
nagy@ecse.rpi.edu

Sharad Seth
University of Nebraska
Lincoln, NE 68588
seth@cse.unl.ed

Xiaoli Zhang
Rensselaer Polytechnic
Institute DocLab
Troy, NY 12180
zhangxl@rpi.edu

Abstract

Two methods, Symbolic Indirect Correlation (SIC) and Style Constrained Classification (SCC), are proposed for recognizing handwritten Arabic and Chinese words and phrases. SIC reassembles variable-length segments of an unknown query that match similar segments of labeled reference words. Recognition is based on the correspondence between the order of the feature vectors and of the lexical transcript in both the query and the references. SIC implicitly incorporates language context in the form of letter n-grams. SCC is based on the notion that the style (distortion or noise) of a character is a good predictor of the distortions arising in other characters, even of a different class, from the same source. It is adaptive in the sense that with a long-enough field, its accuracy converges to that of a style-specific classifier trained on the writer of the unknown query. Neither SIC nor SCC requires the query words to appear among the references.

1 Introduction

From the perspective of character recognition, Arabic and Chinese are at the opposite ends of the spectrum. The former has a small alphabet with word-position dependent allographs, is quasi-cursive, and has "diacritics", ascenders and descenders. The latter has an indefinitely large number of classes (of which only the first ~20,000 have been coded), essentially word-level symbols (many with a radical-based substructure), and fixed-pitch block characters. Arabic strokes can be approximated by arcs of circles, while most Chinese strokes are straight, with a ~1:7 range in width (like brush strokes), and a flourish at the end. Unlike Arabic, Chinese does not have deliberate loops.

They also exhibit some commonalities. Both have been incorporated in the scripts used by other languages: Arabic in Urdu and Persian, Han in Japanese and Hangul, among many others. Both have traditional roots and forms dating back several thousand years, preserved in a large body of classical manuscripts, and have undergone considerable and diverse modifications in each host language and region of the world. Nevertheless, both scripts have preserved sufficient uniformity to link cultures which can no longer understand each other's speech. Their classical forms are prized and cultivated in calligraphy, which combines visual and language arts. Neither script has upper and lower case.

Industrial strength Arabic and Chinese OCR products must also be able to recognize Latin characters, "Arabic" (Indian) numerals, and Western punctuation. This introduces additional complexity, more because of the need to handle diverse, intermingled reading orders and output codes than because of the increased number of classes.

Many thousands of papers (the very first of which, coincidentally, is [1]) have been written on Chinese character recognition. By the time of our first survey [2] much of the research was reported in Chinese and Japanese publications. In our second survey [3] we found little new research in the West. Recent research collaboration with Professor C-L Liu at the Pattern Recognition Laboratory of the Chinese Academy of Science (CASIA), visits with Professor X. Ding at Tsinghua University, and a tour of Hanwang High Technology in Beijing acquainted us with the largest concentrations of character research activity in the world and some of China's thriving OCR industry.

Research on Arabic character recognition (actually on Farsi) began in the late sixties. Scattered projects, mainly by speakers of Arabic in the West, increased until the turn of the millenium, when research began to grow exponentially. Nearly one thousand reports have already been published, mostly in English and French. Nevertheless, work on Arabic OCR lags far behind Chinese OCR because of the lack of monolithic government and market support, and of large, publicly available databases. For a recent survey of the state-of-the-art in offline Arabic handwriting recognition please see [4].

Our team has over 100 years of sustained experience in character recognition, with over one hundred research publications on this topic to our credit. In addition to intra-departmental access to native speakers and writers of both Arabic and Chinese with a background in pattern recognition and signal and image processing, we have forged strong professional bonds with many of the leading researchers in both areas. We are convinced that extrapolating successful methods from Western

(including Russian) OCR is insufficient for either Arabic or Chinese because, ideally and optimally, every glyph of an entire document must be considered simultaneously before a label is assigned to any one of them. In practice, this notion translates to field classification, where glyphs that are difficult to recognize in isolation (or that cannot be isolated/segmented) are recognized in conjunction with several others.

Because of the wide range of different problems exhibited by the two scripts, we believe that tackling both simultaneously is a valuable strategy for research that will bring benefits not only to character recognition on other scripts (like those derived from Sanskrit), but also to the wider field of pattern recognition. Below, we outline how we propose to apply field classifiers that have already proved successful on easier tasks to Arabic and Chinese documents.

We bring two orthogonal ideas to the table: Symbolic Indirect Correlation (SIC) and Style Constrained Classification (SCC). The former recognizes unknown sequences of features (possibly spanning several characters) by finding and reassembling its constituent subsequences in the feature sequence representation of labeled reference text. The unknown word(s) need not be represented in the reference set, only their lexical constituents (i.e., symbol polygrams). Style-based classification, on the other hand, has been applied to distorted but segmented patterns. It maximizes the posterior probability of the field's feature vector of same-source words or phrases given the transcript, under the constraint of source or style specific statistical dependence between all the features of the field. Over the last decade, we have developed (and published) the necessary mathematical apparatus for field classification based on both SIC and SCC.

As is customary in many-class problems, we will use a hierarchical approach to reduce the number of candidate classes to which we apply the full power of our advanced methods. We believe that top-50 classification with less than 1% error on a candidate list of several thousand Arabic words or Chinese characters is within the state of the art, and that field classification can differentiate similar candidates in this reduced list.

We are not aware of any adequate handwritten test data with full context in either Arabic or Chinese. Our proposals for the essential characteristics of such a database were presented at SDIUT05 [5]. Even though the appropriate test data is not yet available, it is still possible to initiate this research on the currently available isolated word and character collections

2 Arabic Character Recognition

Symbolic indirect correlation (SIC) is a general approach to recognition of text that cannot be reliably segmented into characters, as is the case with most offline and online handwriting in non-hieroglyphic scripts.

SIC recognition is based on local matches between unsegmented patterns at both the feature and lexical levels. At the feature level, the unknown pattern is compared to a known (reference) string of features and the results are captured in the form of a match graph. Another matching process is used to find polygram co-occurrences between the lexical transcripts of the reference string and every class to be recognized. In a second-level matching, the order of feature co-occurrences is compared to the order of polygram co-occurrence in the lexical transcript of each class and the unknown pattern is given the label of the best matching lexical class.

SIC offers distinct advantages over prevailing approaches. It avoids the usual integrated segmentation-by-recognition loop. Unlike other whole-word recognition methods, SIC does not need feature-level samples of the words to be recognized. Finally, unlike methods based on Hidden Markov Models, it does not require estimation of an enormous number of parameters by a fragile bootstrap process. Furthermore, SIC can compensate for noisy features or inaccurate feature matching by increasing the length of the reference set.

We introduced SIC in [6,7] with a representation based on ordered bipartite graphs and established its advantages through simulations with a significant amount of noise. Later investigations showed that in the presence of excessive noise, the sub-graph isomorphism based approach to the second-level matching requires an unreasonably large reference set [8,9]. A maximum-likelihood approach [10] avoids this computational bottleneck in the second-level matching. Since this method seems promising for Arabic recognition, we describe in some detail how we build candidate solutions to the query; interested readers will find full technical details in [10,11].

The second level matching assigns the labels of the best-fitting segments in the reference set to each matching segment in the query. The assignment is constrained by the order of the (possibly overlapping) matches. The probability of each candidate solution to the query is computed as follows.

With a large enough reference set, the feature matches between the query and the "words" in the reference string cover most, if not all, of any query word. Further,

a feature match may or may not occur between the query and any given reference word or phrase; the same is true also for a lexical match (bigrams or higher polygrams) between the pair. Thus, when a candidate solution to the query is built by assigning a polygram to each matched feature segment in the query, one of four possible conditions applies to the assignment with respect to every reference word. The assigned polygram:

(1) occurs in the reference word and there is a segment match (*valid match*),

(2) does not occur in the reference word but there is a segment match (*spurious match*),

(3) occurs but there is no segment match (*missed match*), and

(4) does not occur and there is no segment match (*correct reject*).

These conditional probabilities can be estimated by matching the reference words against each other. Then, they are used to estimate the likelihood of each candidate solution and the solution with the maximum likelihood is chosen.

While our work on SIC so far has been restricted to English handwriting, we believe that it would apply well to Arabic handwriting because of the many common characteristics shared by the two. These include linear order of writing, strong baseline, and three well-defined zones (ascender, descender, and median). Other unique features of Arabic writing also speak favorably for SIC:

- Connection of adjacent letters is prescribed by rigorous rules in Arabic. The resulting connected components at the sub-word level (PAWs) may themselves be connected by hasty writers. The segment-free recognition of SIC has been demonstrated to work on cursive English writing.
- Different shapes of letters at the beginning, middle, and end of a word require only that sufficient instances of each kind be included in the reference set.
- Occasionally, Arabic writing breaks from the usual right-to-left order by placing two successive characters one on top of the other. If this happens with some consistency in the writing, a feature-level match of the compound character in the unknown word and the reference string would be correlated in SIC with the corresponding bigram in the second-level matching. Similar considerations apply to the recognition of letters that are sometimes written out of sequence by Arabic writers.

In order to substantiate these claims, we have recently initiated work on applying SIC to recognize offline handwriting using a sample of images from the database of handwritten Arabic town names [12]. At this point, we have only completed the first-level matching at the feature and lexical levels. The features were adopted from English handwriting with minor variations; we expect substantial improvement with feature sets specifically developed for Arabic writing, such as the one reported in [13].

In our preliminary explorations, we selected a reference set of eight town names (numbered 2, 7, 8, 29, 45, 48, 51 and 52), transcribed by writer ae07. We chose four other town names (numbered 1, 3, 14, and 42) by the same writer as query words. The latter had good bigram coverage by the reference set. We used the Smith-Waterman algorithm [14] to find the local alignments (matches) for feature-level matching. The algorithm uses a flexible cost function that allows for mismatches, insertions, and deletions and finds the optimal sequence of such steps needed to match the two subsequences. For a given cost function, it finds the strongest matches starting at every position of the query string against the reference string. False matches abound at shorter segments hence match-score thresholds are set to minimize the likelihood of a false match. Empirically, this is found to filter out most of the matches corresponding to unigrams and character fragments. The same algorithm was adapted to lexical matching of the transcripts of the query and reference strings.

Two examples of the lexical and feature match graphs are shown in Figures 1 and 2 to convey a sense of how the SIC approach might apply to Arabic handwriting recognition. The same query word, ae07_014, is matched with the reference word ae07_002 in Figure 1 and with the reference word ae07_045 in Figure 2. In each figure, part (a) shows the lexical match graph and part (b) shows the feature match graph. The lexical matching is carried out for exact bigram and higher-order n-gram matches and, typically, results in one or two matches. The feature matching typically yields many more edges for the same pair of words, even for a threshold value that is high enough to minimize single-character matches. The strength of a match is shown as a positive-integer weight of the corresponding edge in the graph; it is indicative of the extent of the matching segments in the feature domain. In the examples we have chosen to show only the top-three candidate edges in each case. For visualization, the strength of matches is denoted by the thickness of lines and line colors (magenta, gray, and brown in the decreasing order).

Figure 1 (a) shows that there is only one lexical match in this example: the bigram (aaA laB) at position 5 in the query word matches with the same bigram at position 5 in the reference word, where the positions are counted from right-to-left in accordance with the Arabic

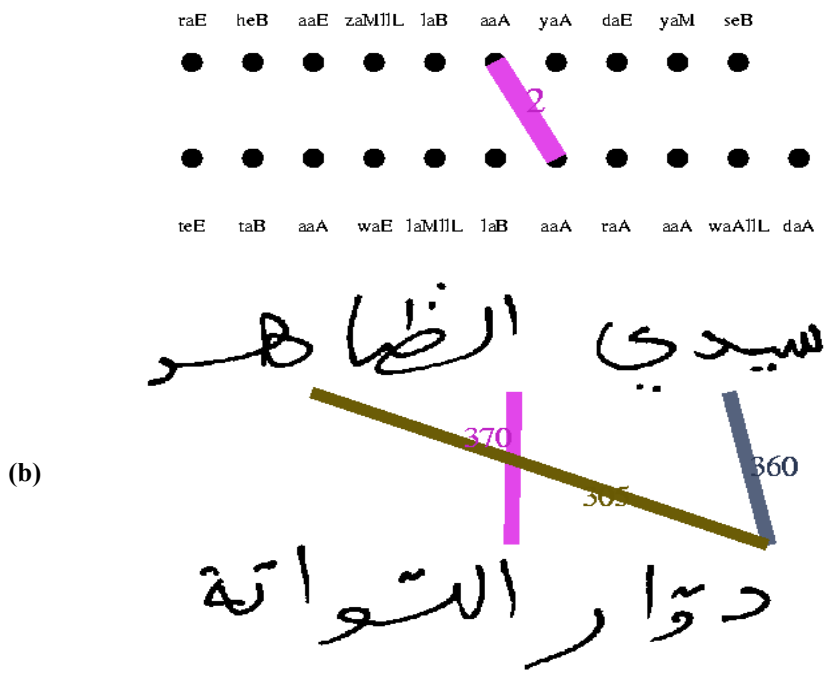

Figure 1: *Lexical and feature graphs of the query word ae07_014 and the reference word ae07_002.*

Figure 2: *Lexical and feature graphs of the query word ae07_014 and the reference word ae07_045*

writing convention. In Figure 1 (b), this lexical match is correctly identified by the strongest match, of strength 370, in the feature graph. However, the feature graph also includes two spurious edges, of strength 365 and 360 respectively, that do not have corresponding edges in the lexical graph.

In the second example, Figure 2(a) shows two lexical matches: a 4-gram at the beginning of the query word matching at position 5 in the reference word and a trigram matching at the end of both the words. Both of these are also found in the feature graph, in Figure 2(b), among the top-3 matches. However, the strongest edge, of strength 925, corresponds to the trigram and the next strongest edge, of strength 545, corresponds to the 4-gram. The third edge, of strength 250, is spurious. We note that because of both signal noise and variability in the character-widths, the strength of a correctly matched edge in the feature graph is only weakly correlated with the strength of its corresponding edge in the lexical graph.

Even though the feature set used in our examples is not particularly well adapted to Arabic, the feature matching process correctly picks the lexical matches in many cases. However, there are many spurious matches as well. Our second-level matching process, described in [10] is shown to be robust against a large number of spurious matches but at the expense of increased computation time. Therefore, we plan to explore a post-processing approach to eliminate some bad matches. The basic idea here is to use non-sequential features to screen the 2-D regions identified by every feature match. Hull [15] employs a similar idea in another context, to select candidates for whole-word recognition, even when they are printed in different fonts. Consider, for example, the feature graph in Figure 2(b) showing the top-3 matches. The image region in the query word corresponding to the weakest edge, which has the strength of 250, has a flat stroke with a diacritic mark above it while that corresponding to the reference word has diacritic marks below the stroke. It should be possible to reject this match based on 2-D features that summarize the dominant directions of each black pixel in different sub-regions of the two images.

3 Chinese Character Recognition

Almost any method can recognize neat handwritten Chinese with better than 85% accuracy, and newsprint at over 95%. Making allowance for the usual 2:1 reject/error trade-off, this implies that we must concentrate on 35% of the handwritten material and 15% of the print. These figures are based on the characters used in the People's Republic of China, which are somewhat harder to recognize than the characters used in Taiwan, Japan, and Korea because the simplifications fifty years ago removed many "redundant" strokes.

Printed Chinese characters are usually fixed-pitch, without ascenders or descenders, and all the characters fit into the same size, horizontally aligned, bounding boxes. Whether reading order is left-to-right or top-to-bottom is easily determined. Segmentation is, however, a major source of error in handwriting. Rushed writers connect and even overlap characters, do not adhere to a clear baseline, and cannot squeeze complex characters into bounding boxes that are ample for simpler characters. (However, some text produced by expert calligraphers is nearly indistinguishable from print, as exemplified in the Proceedings of the conferences on Computers and Chinese Input/Output Systems in the early 1970's.)

Both handwriting and print exhibit pairs (occasionally even triples) of characters with almost identical shapes but different meanings. (Some researchers deliberately exclude such confusion pairs from reported error statistics.) Human readers resolve such ambiguities through broad context. A far more restricted set of language constraints is also used in Chinese OCR. Dictionary (lexicon) look-up cannot be applied in the same way as in Western languages, but the extreme skew of the distribution of unigrams and of two- and three-character sequences can be readily exploited. We note, in particular, that the number of Chinese family and given names, where mistakes cannot be tolerated, is less than in most Western nations. Foreign names may be transliterated or printed in their native script.

We discussed a new approach to segmentation-free character recognition in the section on Arabic. Here we present style-constrained field classification, which is the only recourse when there is insufficient linguistic context. When we cannot read a letter, we look for easier-to-recognize instances of the same shape. Other instances of an unknown character may be easier to classify because there is less (or different) noise, or they are segmented better, or because there is more language context. Adaptive algorithms that benefit from typeface and writer intra-class consistency of this kind have been known for decades [16,17,18,19] but found their way into commercial systems only recently [20]. The scope of adaptation is typically a page: it is assumed that each page is written by a single person, or printed in a limited set of typefaces. A set of reliable prototypes is collected in a first pass, and the remaining problematic characters are recognized in one or more subsequent passes.

Some easily confused characters where adaptation can help are shown Table 1. (We use printed examples for ease of interpretation by readers who cannot read Chinese. The handwritten version of these characters are, of course, even more ambiguous.) However, adaptation works well only with long fields, where there are several samples of each class. In Chinese, much longer fields are needed than in alphabetic languages.

Table 1: *Two Han confusion pairs (ri/yue and dao/diao) in seven fonts. The left column and its transliterations are the font names. On the right are a few of the 7x6/2=21 possible different-font confusion pairs, on which conventional singlet classifier is likely to make errors.*

Chinese Font			Font Confusion
方正姚体	(fang zheng yao ti)	日 曰	
方正舒体	(fang zheng shu ti)	日 曰	日曰 / 日曰
华文细黑	(hua wen xi hei)	日 曰	日曰 / 日曰
华文行楷	(hua wen xing kai)	日 曰	日曰 / 日曰
华文中宋	(hua wen zhong song)	日 曰	
新宋体	(xin song ti)	日 曰	
幼圆	(you yuan)	日 曰	
		(ri yue)	
方正姚体	(fang zheng yao ti)	刀 刁	
方正舒体	(fang zheng shu ti)	刀 刁	
华文细黑	(hua wen xi hei)	刀 刁	刀刁 / 刀刁
华文行楷	(hua wen xing kai)	刀 刁	刀刁 / 刀刁
华文中宋	(hua wen zhong song)	刀 刁	
新宋体	(xin song ti)	刀 刁	
幼圆	(you yuan)	刀 刁	
		(dao diao)	

Table 2: *Scenarios faced by a singlet classifier (above) and by a pair classifier (below), on two pairs of similar characters. When it is known that both characters are from the same style (font or writer), the confusions are more easily resolved by a style-constrained classifier that has additional information from another character in the same style.*

	Font1/ Font2	Font2/ Font1
Singlets:	日/日	曰/日
	刁/刀	刁/刀
Pairs:	子日/日子	子曰/日子
	刁钻/刀钻	刁钻/刀钻

We have recently demonstrated a much less intuitive aspect of local shape consistency that we call inter-class style. The underlying idea is simply that knowing how an individual writes a **g** or a **p** may help us predict how she may write a **q**. In fact, the shape of every class provides some information about every other class. In a statistical framework, we say that the features of one class are style-conditionally dependent on the features of another class. Abandoning the customary independence assumption leads to a more complex mathematical framework.

Nevertheless, the optimal maximum a posteriori (MAP) classifier can be formulated neatly [21,22,23]. In the last three years we demonstrated significant gains in accuracy through style-constrained field classification on both printed and hand-printed digits. Now propose to do the same for Chinese and Arabic. Table 2 suggests why pairs of Chinese characters are easier to classify than individual characters.

We note that printers, copiers, scanners and cameras also introduce useable style constraints. Human readers resort to field classification when necessary. Like humans, machines must also be enabled to do field classification, only when needed, because it is expensive. The number of field classes increases exponentially with field length. Whereas with numerals we used field lengths up to 5, for Chinese we propose to apply style constraints only to selected pairs and triples.

For the sake of completeness, we note that font recognition is an inefficient form of style-constrained classification. It generally requires separate features for font and class identification, is confused when fonts share some shapes, wastes statistical evidence on identifying the font when only class labels are wanted, and cannot accommodate font miscegenation.

4 Interaction

All OCR systems benefit from some human help, typically at the beginning or the end of the process. Scanning is almost always checked, because even current scanners occasionally bungle digitization. At the beginning, the operator may label some unusual characters, select a language model, or provide some general format information. He or she may also occasionally assist page segmentation. At the end, low-confidence labels are verified or corrected. When there are too many errors, the entire page may be keyed in instead of corrected.

Current OCR systems do not make the most efficient use of the operator, perhaps because such work is often outsourced and offshored. For urgent and critical applications, the operator may well be the end user. Workers with other primary missions are not likely to tolerate the repetitive routine of data entry personnel. The interaction must therefore take place wherever and whenever it is most effective and, above all, it should not be wasted. The software should attempt every task. The operator must, however, have the opportunity of correcting the result whenever necessary. Whether a particular error needs to be corrected requires the kind of judgment that at present machines lack.

We have made the case for a personal, mobile, multilingual support system at SDIUT 2004 [5], on interactive table interpretation at DAS06 [24], and on large scale document entry at DIAL06 [25]. We have argued that every operator action should result in some change in the configuration of the system that decreases the likelihood of the same situation occurring again. In other words, the system must improve with use. The defense rests.

References

[1] R. G. Casey and G. Nagy, "Recognition of Printed Chinese Characters," *IEEE Transactions on Electronic Computers,* vol. 15, pp. 91-101, February 1966.

[2] R. G. Casey and G. Nagy, "Chinese Character Recognition: A Twenty-five-year Perspective," in *Proc. International Conference on Pattern Recognition*, Rome, Italy, pp. 1023-1026, 1988.

[3] J. Kanai, Y. Liu, and G. Nagy, "An OCR-oriented Overview of Ideographic Writing Systems," in *Handbook of Character Recognition and Document Image Analysis*, H. Bunke and P. S. P. Wang, Eds.: World Scientific, 1997, pp. 285-304.

[4] L. M. Lorigo and V. Govindaraju, "Offline Arabic Handwriting Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 712-724, May 2006.

[5] D. Lopresti and G. Nagy, "Mobile Interactive Support System for Time-Critical Document Exploitation," in *Proc. Symposium on Document Image Understanding*, College Park, MD, pp. 111-119, 2005.

[6] G. Nagy, S. C. Seth, S. K. Mehta, and Y. Lin, "Indirect Symbolic Correlation Approach to Unsegmented Text Recognition," in *Proc. Conference on Computer Vision and Pattern Recognition Workshop on Document Image Analysis and Retrieval (DIAR'03)*, Madison, WI, p. 22, 2003.

[7] G. Nagy, D. P. Lopresti, M. Krishnamoorthy, Y. Lin, S. Seth, and S. Mehta, "A Nonparametric classifier for unsegmented text," in *Proc. SPIE*, San Jose, pp. 102-108, 2004.

[8] A. Joshi and G. Nagy, "Online Handwriting Recognition Using Time-Order of Lexical and Signal Co-Occurrences," in *Proc. 12th Biennial Conference of the International Graphonomics Society*, Salerno, Italy, pp. 201-205, 2005.

[9] D. Lopresti, A. Joshi, and G. Nagy, "Match Graph Generation for Symbolic Indirect Correlation," in *Proc. SPIE-IS&T Symposium on*

Document Recognition and Retrieval, Vol. 6067-06, San Jose, CA, 2006.

[10] A. Joshi, G. Nagy, D. Lopresti, and S. Seth, "A Maximum-Likelihood Approach to Symbolic Indirect Correlation," in *Proc. International Conference on Pattern Recognition*, Hong Kong, China, 2006.

[11] A. Joshi, "Symbolic Indirect Correlation Classifier," Rensselaer Polytechnic Institute,ECSE Department, Troy, NY, Ph. D. Thesis 2006.

[12] V. Märgner and M. Pechwitz.*IFN/ENIT-database: Database of Handwritten Arabic Words*. [Online].Available: http://www.ifnenit.com/index.htm.

[13] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling," in *Proc. Int. Conference on Document Analysis and Recognition, ICDAR*, pp. 893-897, 2005.

[14] T. F. Smith and M. S. Waterman, "Identification of common molecular sequences," *Journal of Molecular Biology,* vol. 147, pp. 195-197, 1981.

[15] J. J. Hull, "Incorporating Language Syntax in Visual Text Recognition with a Statistical Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 1251-1256, December 1996.

[16] G. Nagy and G. L. Shelton, "Self-Corrective Character Recognition System," *IEEE Transactions on Information Theory,* vol. 12, pp. 215-222, April 1966.

[17] H. S. Baird and G. Nagy, "A Self-correcting 100-font Classifier," in *Proc. SPIE Conference on Document Recognition and Retrieval*, San Jose, CA, pp. 106-115, 1994.

[18] Y. Xu and G. Nagy, "Prototype Extraction and Adaptive OCR," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, pp. 1280-1296, December 1999.

[19] T. K. Ho and G. Nagy, "OCR with no shape training," in *Proc. International Conference on Pattern Recognition*, Barcelona, Spain, pp. 27-30, 2000.

[20] I. Marosi and L. Tóth, "OCR Voting Methods for Recognizing Low Contrast Printed Documents," in *Proc. 2nd IEEE International Conference on Document Image Analysis for Libraries, DIAL 2006*, Lyon, France, pp. 108-115, 2006.

[21] S. Veeramachaneni and G. Nagy, "Adaptive classifiers for multisource OCR," *International Journal of Document Analysis and Recognition,* vol. 6, pp. 154-166, March 2004.

[22] S. Veeramachaneni and G. Nagy, "Style context with second order statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 14-22, January 2005.

[23] P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 88-98, January 2005.

[24] D. W. Embley, D. Lopresti, and G. Nagy, "Notes on Contemporary Table Recognition," in *Proc. Document Analysis Systems VII, 7th International Workshop, DAS 2006*, Nelson, New Zealand, pp. 164-175, 2006.

[25] G. Nagy and D. Lopresti, "Interactive Document Processing and Digital Libraries," in *Proc. 2nd IEEE International Conference on Document Image Analysis for Libraries, DIAL 2006*, Lyon, France, p. 8 pages, 2006.