IJDAR (1999) 2: 45–52

International Journal on IJDAR
Document Analysis and Recognition

Script and language identification for handwritten document images

Judith Hochberg[1], **Kevin Bowers**[2], **Michael Cannon**[1], **Patrick Kelly**[1]

[1] Computer Research and Applications Group (CIC-3), Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; e-mail: {judithh,tmc,kelly}@lanl.gov
[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA

Received December 1, 1998 / Revised April 5, 1999

Abstract. A system for automatically identifying the script used in a handwritten document image is described. The system was developed using a 496-document dataset representing six scripts, eight languages, and 279 writers. Documents were characterized by the mean, standard deviation, and skew of five connected component features. A linear discriminant analysis was used to classify new documents, and tested using writer-sensitive cross-validation. Classification accuracy averaged 88% across the six scripts. The same method, applied within the Roman subcorpus, discriminated English and German documents with 85% accuracy.

Key words: Script – Language – Handwriting – Discrimination – Features

1 Introduction

Script and language identification are important parts of the automatic processing of document images in an international environment. A document's script (e.g., Cyrillic or Roman) must be known in order to choose an appropriate optical character recognition (OCR) algorithm. For scripts used by more than one language, knowing the language of a document prior to OCR is also helpful. And language identification is crucial for further processing steps such as routing, indexing, or translation.

For scripts such as Greek, which are used by only one language, script identification accomplishes language identification. For scripts such as Roman, which are used by many languages, it is normally assumed that script identification will take place first, followed by language identification within the script (e.g. [1]). Alternatively, it may be possible to skip script identification as an intermediate step, recognizing languages directly regardless of their script.

Correspondence to: J. Hochberg

In previous work, Los Alamos National Laboratory developed a highly accurate automatic script identification system for machine printed documents [2, 3]. This paper reports on our extension of this research to handwritten documents. While the main focus of the work was script identification, we also took a first look at language identification within the Roman script. We attempted to distinguish German and English in both ways mentioned above: with script identification as an intermediate step, and directly.

Handwritten documents present three challenges for script identification. First, some scripts, particularly Roman and Cyrillic, resemble each other more when handwritten than when printed. Second, handwriting styles are more diverse than printed fonts. Cultural differences, individual differences, and even differences in the way that people write at different times, enlarge the inventory of possible character and word shapes seen in handwritten documents. Third, problems typically addressed in preprocessing, such as ruling lines and character fragmentation due to low contrast, are common in handwritten documents due to the variety of papers and writing instruments used.

The examples in Fig. 1 illustrate the first two, and the most fundamental, of these challenges. The two Russian writers have different writing styles – connected versus discrete characters – as do the German writers. The stylistic differences are so marked that they outweigh the differences in character inventories that are so striking in machine printed Russian and German text.

These challenges made it impossible to successfully apply to handwritten documents the same template matching approach that we used for machine printed documents [2, 3]. The success of that method depended on our ability to identify a core set of templates for each script that were reliable indicators of a document's script identity. Because of the greater variability of handwriting styles, sufficient numbers of reliable templates could not be identified. We therefore took a fresh, feature-based approach in which each document was characterized by a single feature vector. Each vector contained summary statistics taken across the document's black connected

Table 1. Scripts and varieties in dataset

Script (single variety)	Total images	Varieties (if more than 1)	Number of images
Arabic (Arabic)	57		
Chinese	120	original characters	69
		simplified characters	51
Cyrillic (Russian)	56		
Devanagari	25	Hindi	21
		Marathi	4
Japanese	73		
Roman	165	American English	40
		British English	67
		German	58

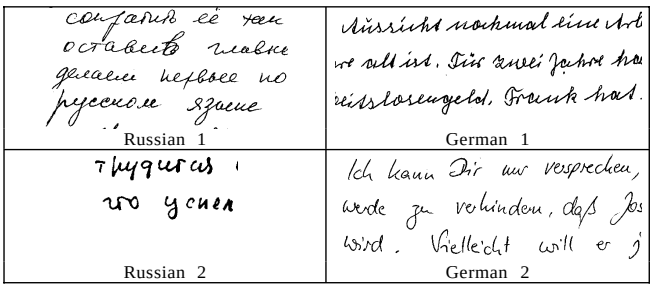


Fig. 1. Fragments of handwritten Russian and German documents

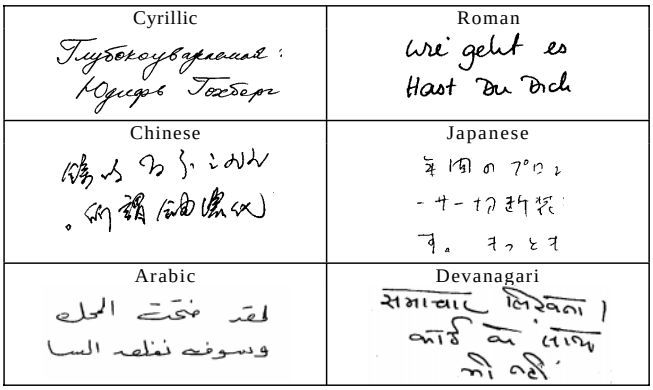


Fig. 2. Examples of six handwritten scripts

components, such as the components' average aspect ratio. The documents were then classified by script using linear discriminant analysis.

The method was 88% accurate in distinguishing among six scripts. We also addressed, in exploratory work, the feasibility of applying the same method to language identification. We found the method 85% accurate within a subcorpus of English and German documents.

2 Data

2.1 Obtaining handwritten documents

We assembled a corpus of 496 handwritten documents from six scripts: Arabic, Chinese, Cyrillic, Devanagari, Japanese, and Roman. English and German were both represented within Roman. The scripts are illustrated in Fig. 2. Table 1 summarizes the number of documents per script and variety.

For the most part, document images were obtained from foreign language speakers we were acquainted with or whom we contacted through the Internet. We asked them to lend us letters from friends and family, or other existing handwritten documents. If they had no documents, we asked them to write down or transcribe a paragraph or two. Over 75% of the documents we collected were 'natural' – letters, lecture notes, official documents, etc. The remaining documents were written on request. 279 different writers were represented in the corpus, and all were native speakers of the languages they wrote in. The preponderance of natural documents and the variety of scripts and writers included make this corpus a valuable resource.

Documents were scanned in using an Agfa scanner equipped with StudioScan II software. They were scanned as line art (black and white rather than grayscale), using a resolution of 200 dpi. After scanning, Adobe Photoshop was used to remove any irregularities: illustrations, doodles, machine printing, postal markings, marginal ruling, cross-outs, foreign characters, and anomalous writing, such as a few sentences written sideways. For each document, or each set of similar documents from a single source, a fixed-format ASCII "info" file was written to encode general information about the image, such as its script, language, and a writer ID code.

2.2 Document quality issues

As mentioned in the Introduction, document quality issues such as ruling lines and character fragmentation due to low contrast are common in handwritten documents. Table 2 summarizes the incidence of these phenomena in our corpus. The ratings "none", "mild", "moderate", and "severe" were subjective judgments, recorded in the info files as each image was scanned.

The issues of character fragmentation and ruling lines were addressed in preprocessing (see Sect. 3.1). We did not attempt to correct for the other phenomena, but simply included all documents in the training and testing process in order to perform a realistic test of the classification method.

Table 2. Document quality problems in corpus

	none	mild	moderate	severe
text line curvature	260	222	14	0
text line skew	360	118	18	0
character fragmentation	196	214	64	22
	yes		no	
ruling lines (horizontal, vertical, or both)	53		443	
short document (≤= 100 components after filtering)	29		467	

3 Method

For each document, we found black connected components (assuming eight-connectedness); removed speckle, ruling lines, and outsize components; extracted five features per component; and calculated the mean, standard deviation, and skew for each component feature across all components in the document. A linear discriminant analysis was trained to identify the script of each image, and a similar procedure was used for language identification. The algorithm was implemented on a Sun workstation using a C++ library for image processing developed at Los Alamos.

The following sections describe these steps in more detail.

3.1 Connected components

The basic element of our analysis was the eight-connected black component. After finding all the components in a document image, we filtered out unusually small or large components in order to remove speckle, ruling lines, and outsize components in general.

The filtering algorithm passed through a document's connected components two times. On the first pass, we removed components that were either small, or long and thin. These criteria were defined as follows:

"Small":

- height of bounding box < 3 pixels OR
- width of bounding box < 3 pixels OR
- total area (bounding box height $*$ bounding box width) ≤ 30 pixels

"Long and thin":

- height of bounding box ≤ 5 pixels and $\frac{\text{height}}{\text{width}}$ of bounding box ≤ 0.1 OR
- width of bounding box ≤ 5 pixels and $\frac{\text{height}}{\text{width}}$ of bounding box ≤ 10

During this pass, we also computed the mean and standard deviation of bounding box height and width measurements for the components retained. These statistics were then used as the basis of a second filtering pass in which we removed unusually large components: those with bounding box height or width more than four standard deviations above the mean.

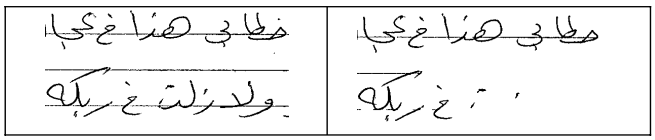


Fig. 3. Fragment of Arabic with ruling lines embedded in text, before filtering (*left*) and after (*right*)

The filtering algorithm effectively removed most speckle and ruling lines. It had difficulty with ruling lines that were deeply embedded in text; this most commonly occurred in Arabic documents. Figure 3 illustrates a typical example of embedded ruling and the outcome of the filtering algorithm applied to it. Filtering caused the erasure of substantial chunks of text in such documents. However, since our goal was script identification and not OCR, this was not a problem as long as the remaining connected components were sufficiently indicative of the document's script.

3.2 Features

Once filtering was completed, several features were extracted from the remaining components. To develop the feature set we first studied the document images and determined which visual features guided our human script identification. This analysis focused on the two most difficult script distinctions: Chinese vs. Japanese and Roman vs. Cyrillic. The first distinction is difficult because Japanese *kanji*, or root characters, are directly based on Chinese characters. The second distinction is difficult because Roman and Cyrillic are genetically related and share many characters.

The features selected are listed in Table 3. Some of them were based on this visual analysis. Others were general properties of connected components widely used in the document image processing community. They are a different selection from those used by [1] for machine printed script identification.

For each of the five connected component features, three document summary statistics were calculated: the mean, standard deviation, and skew. This created a fifteen-element vector for each document.

Working within the framework of the linear discriminant analysis described in the next section, we experimented with reducing the feature set through stepwise subtraction of features. There was no dramatic gain from removing features, so only results using the full feature set are reported.

3.3 Script identification

Our classification method used a collection of linear discriminant functions. A separate Fisher linear discriminant [4] was trained to separate each possible pair of scripts in the dataset (Arabic vs. Chinese, Arabic vs. Cyrillic, etc.). The offset value within each classifier was

Table 3. Connected component features. Component width and height are based on the component's bounding box

Feature	Meaning	Motivation/s for the feature (see Fig. 2)
Relative Y centroid	$\frac{\text{vertical centroid}}{\text{component height}}$	– Roman has more ascending characters than Cyrillic – Japanese components tend to have a higher relative centroid than Chinese
Relative X centroid	$\frac{\text{horizontal centroid}}{\text{component width}}$	– General property
Number of white holes	found by connected component algorithm	– Chinese characters are more complex than Japanese – Cyrillic words tend to be longer (contain more characters) than Roman words due to more connected writing style
Sphericity	$\frac{\text{\# black pixels}}{\text{component parameter}^2}$	– General property
Aspect ratio	$\frac{\text{component height}}{\text{component width}}$	– Cyrillic has fewer ascending characters, hence flatter words, than Roman – Cyrillic words tend to be longer (contain more letters) than Roman words due to more connected writing style

chosen to minimize the percentage of training data misclassified in either script. New documents were classified by applying each individual linear discriminant to the document's feature vector, while keeping track of the results. The document was then assigned to the class receiving the most "votes". If two or more scripts "tied" – in other words, if more than one script received the top number of votes – the classifier tried to resolve the tie by counting votes just among the tied scripts.

The classifier was tested through writer-sensitive cross-validation. For each writer, the classifier was trained on all data except that writer's documents. Then the writer's documents were classified using the trained classifier. We calculated the percentage of documents correctly classified for each script, and averaged these percentages to produce an overall accuracy figure unbiased by the scripts' sample sizes.

We experimented with a number of other classifiers, including a neural network based on the same fifteen document summary features, but found the linear discriminant approach to be both the most robust and the most straightforward.

3.4 Language identification

Our collection of Roman script documents included 107 in English and 58 in German. We used these documents to explore two different models for language identification:

– *Direct*: The six-way script identification test described in the previous section was modified by splitting English and German into separate categories. In other words, a single discriminant analysis was used for the seven-way discrimination among Arabic, Chinese, Cyrillic, Devanagari, Japanese, German, and English. The entire writer-sensitive cross-validation procedure was repeated for this model.
– *Two-step*: A subcorpus of data was formed, containing only the 165 Roman language documents. A single linear discriminant between German and English documents was tested on the subcorpus using writer-sensitive cross-validation. This model assumed that script identification would be used as an intermediate step to identify Roman documents prior to language identification.

4 Results

4.1 Script identification

The linear discriminant analysis, as tested through writer-sensitive cross-validation, was 88% accurate. Table 4 breaks down these results by script, and also presents the cross-classification matrix. The individual percentages for the different scripts were pleasingly uniform – within an eight-percentage-point range – especially keeping in mind the disparate amounts of data available for the different languages. When documents were misclassified, the errors were of the type that anyone familiar with the scripts would predict. Roman and Cyrillic tended to be confused, and likewise Chinese and Japanese. Eight documents were not classified due to a three-way tie vote among the component classifiers. The correct script was among the tied scripts in all eight cases.

Table 4. Script identification results

Script	% correct	Classified as						
		Arabic	Chinese	Cyrillic	Dev.	Jpn.	Roman	tie
Arabic	89%	51	0	0	3	2	1	0
Chinese	83%	0	100	0	0	8	9	3
Cyrillic	88%	0	0	49	2	0	4	1
Devanagari	88%	0	0	1	22	1	1	0
Japanese	86%	2	5	0	0	63	1	2
Roman	91%	0	1	8	0	3	150	2
Average	88%							

Table 5. Feature ranking (according to coefficient magnitude) in two-way discriminant analyses

Rank	Connected component feature	Document summary statistic
	Cyrillic vs. Roman	
1	Y centroid	mean
2	number of white holes	mean
3	number of white holes	standard deviation
4	Y centroid	skew
5	aspect ratio	standard deviation
6	sphericity	standard deviation
7	aspect ratio	mean
8	Y centroid	standard deviation
9	X centroid	mean
10	X centroid	standard deviation
	Chinese v. Japanese	
1	number of white holes	mean
2	sphericity	mean
3	number of white holes	standard deviation
4	X centroid	skew
5	X centroid	standard deviation
6	aspect ratio	standard deviation
7	number of white holes	skew
8	sphericity	standard deviation
9	X centroid	mean
10	sphericity	skew

4.2 Features and document quality issues revisited

Features

Did the role of the different features in the discriminant analysis reflect the importance we ascribed to them during our visual analysis of the documents (recall Table 3)? To address this question, we trained a two-way linear discriminant between Cyrillic and Roman, and one between Chinese and Japanese, using all documents available in the relevant scripts (i.e., not withholding any for testing). Before doing this analysis we unit-normalized each of the fifteen features separately, across all documents, in order to eliminate differences of scale between the features. After training the discriminants, we output and ranked the coefficients assigned to each feature. The magnitude (absolute value) of each coefficient indicated how strongly it affected the classification.

The results mostly confirmed our visual analysis, but provided some surprises. Table 5 lists (in coefficient rank order) the top ten features found by the two discriminants. For Roman vs. Cyrillic, Y centroid and number of holes features topped the list, as expected given Roman's abundance of ascending characters and Cyrillic's greater connectedness. For Chinese vs. Japanese, number of holes features dominated, as expected given the greater complexity of Chinese characters (Japanese uses simple *kana* characters for word endings and function words, in addition to the complex Chinese *kanji* characters used for root morphemes). On the other hand, the importance of sphericity and X centroid features in discriminating Chinese vs. Japanese was unexpected. In a post-hoc analysis, we confirmed that mean sphericity tended to be greater in Japanese documents than in Chinese; this may be a further reflection of the greater complexity of Chinese characters. The skew of the X centroid feature tended to be greater for Chinese, while its standard deviation tended to be greater for Japanese, a difference that is difficult to interpret. The difference in Y centroid between Chinese and Japanese that we observed visually (Table 3) turned out not to play a substantial role in classification.

All five connected component features, and all three document summary statistics (mean, standard deviation, and skew) ranked in the top five for either Chinese vs. Japanese or Cyrillic vs. Roman. This validated our experimental finding that reducing the feature set did not improve classification.

Document quality issues

We used the classification outcomes from the script identification cross-validation test to assess how the quality issues mentioned in Sect. 2.2 affected classification accuracy. The only characteristic to significantly affect classification accuracy was character fragmentation. 90% of documents with no fragmentation, or only mild fragmentation, were correctly classified, compared to 81% of documents with moderate or severe fragmentation. We determined that this difference was statistically significant by performing an analysis of variance, or ANOVA. The test gave an F value, indicating the magnitude of the effect, of 5.21. Given the number of documents in the test, this effect was significant at the 95% confidence level ($p < 0.05$). Ruling lines also appeared to affect classification – 89% of unruled documents were correctly classified, compared to 81% of ruled documents – but this difference was just short of statistical significance ($F = 3.18$, $p = 0.07$).

Table 6. Script identification accuracy for different-sized documents

Quartile	Number of connected components	% correctly classified
1	42-223	85%
2	224-331	93%
3	332-520	90%
4	521-1815	86%

Of the 366 documents in the corpus with no or mild fragmentation, and without ruling lines, 91% were classified correctly.

Surprisingly, and happily, document length did not affect classification. We addressed this question in two ways. First, we divided the documents into four quartiles, based on the number of connected components per document, and compared classification outcomes across the quartiles. As shown in Table 6 below, the results did not show any meaningful pattern. This was confirmed statistically ($F = 1.66$, $p = 0.17$). Second, we examined classification outcomes for the shortest documents in our corpus: those with fewer than 100 connected components. 26 of these 29 documents were correctly classified.

Language identification

Language identification of English versus German was fairly successful, as shown in Table 7. With the direct method – simply splitting English and German to create a seven-way classifier - – 80% of English and German documents were correctly identified by language. Splitting English and German somewhat worsened average classification accuracy for the other five scripts, from 87% to 85%.

Interestingly, splitting Roman into English and German slightly improved Roman *script* identification. 93% of Roman documents were correctly classified as Roman using the seven-way classifier (Table 7), compared to 91% with the six-way classifier (Table 4). It may be that the heterogeneity of the combined Roman group adversely affected the script classifier's performance, so that dividing the group into two smaller, more homogeneous groups helped.

With the two-step method – English/German discrimination following script identification – the language discrimination step averaged 85% accuracy. Assuming 91% accuracy for Roman identification (as in Table 4), this result implied that the entire process of script identification followed by language identification would be 77% accurate for English and German, which is slightly worse than the direct method.

What distinguished English and German handwriting? Overall, German writers had a more uniform writing style than English writers, with relatively little variation among component features within a document (see Fig. 4). Looking at the fifteen document features and their values in the two languages, four of the standard deviation features showed the most significant variation between the two languages ($F \geq 19.5$, $p \leq 0.00002$).

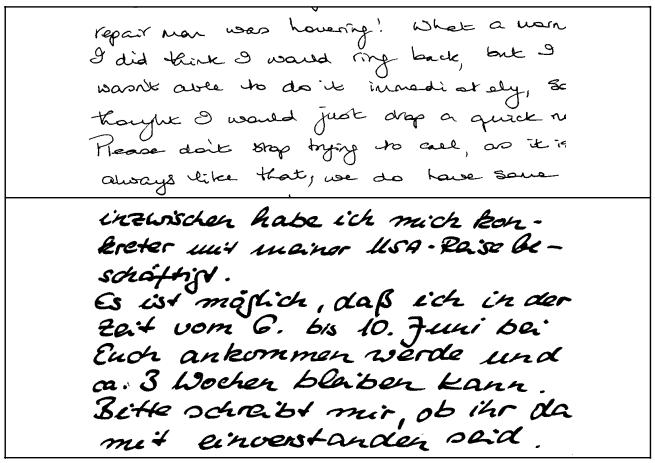


Fig. 4. Fragments of English (*top*) and German (*bottom*) handwriting illustrating greater connected component uniformity of German writing. Standard deviations for the English document were 0.07, 0.10, and 0.34 for relative x centroid, relative y centroid, and aspect ratio, respectively. Standard deviations for the German document were 0.04, 0.06, and 0.18

Compared to English documents, German documents had a lower standard deviation for relative x centroid, relative y centroid, and aspect ratio, implying more uniformity in their writing.

Unexpectedly, the fourth standard deviation feature – the standard deviation of connected component sphericity – showed the opposite pattern, with English writers having a lower standard deviation. We cannot account for this difference.

Since all English and German documents in the collection were written by native speakers of the two languages, we cannot tell whether these differences were *linguistic* or *cultural*. In other words, we cannot tell whether the greater uniformity of German handwriting had anything to do with English versus German *per se*, or if it reflected how handwriting is taught in different countries. One could address this question by looking at how Germans write English, and vice versa.

Arguing against a cultural interpretation, an attempt at a three-way discrimination among American English, British English, and German failed. Correct German identification dropped to 67% (from 86% in the English v. German test); American and British English were each correctly identified only 54% or 55% of the time.

5 Conclusion

The feature set and classifier we developed served to discriminate scripts with 88% accuracy. While not as accurate as script identification for machine printed document images [2, 3], this result exceeded our initial expectations given the variability of handwritten documents. Classification accuracy was higher for documents without fragmented characters and ruling lines. Language identification for English versus German was 85% accurate once Roman identity was known, and 80% accurate

Table 7. Language identification results

Method	Language or script	% correct	Classified as		
			non-Roman (inc. ties)	English	German
Direct	English	78%	8	83	16
	German	81%	4	7	47
	Roman	93%			
Two-step	English	84%	not applicable	90	17
	German	86%	not applicable	8	50

when script and language identification were performed together.

We see three possible extensions of this research. The first would be to take a more thorough look at language identification. Collecting the necessary data would be challenging – ideally, one would want to acquire handwriting from four or five Roman script languages and two or three Cyrillic or Devanagari languages. Many intriguing questions could be addressed: whether the suggestive results we saw with English and German would extend to a larger set of languages, whether direct or two-step language identification would prove best in the long run, and whether observable differences are cultural or linguistic.

The second extension would address a limitation of the linear discriminant analysis: its inability to provide a mechanism for identifying documents written in a script not seen in training – a "script unknown" classification. An attempt to use a classifier based on probability density functions to provide this functionality was unsuccessful. Our impression was that the heterogeneity of our dataset, and the relatively small sample size relative to the 15-dimensional space, hindered our ability to compute representative PDFs for each script.

The third extension would be to use some modification of our current method to identify individual writers by their handwriting, perhaps along the lines of [5]. Recent pilot work by us in this area is reported in [6]. The dataset we have assembled would afford an exciting possibility of doing this in a multilingual environment.

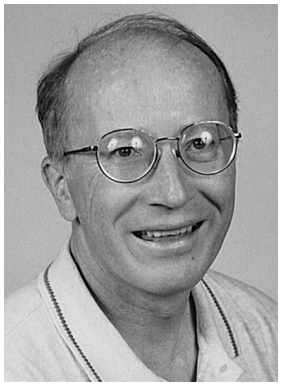
Acknowledgements. Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. Some of these results were previously presented as [6]. We would like to thank Steve Dennis for support, feedback, and data, and Ron Riley and Greg Wilensky for inspiration and data delivery. David Nix performed the neural network analysis. Most of all, we would like to thank the many people from around the world who shared with us their writing and that of their friends and families.

References

1. Spitz, A. L.: Determination of the script and language content of document images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(3), 235–245, 1997
2. Hochberg, J., Kelly, P., Thomas, T., Kerns, L.: Automatic script identification from document images using cluster-based templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), 176–181, 1997
3. Patent: Script identification from images using cluster-based templates (5,844,991)
4. Duda, T., Hart, P.: Pattern Classification and Scene Analysis. New York: John Wiley, 1973, pp. 114–118
5. Wilensky, G., Crawford, T., Riley, R.: Recognition and characterization of handwritten words. In: Doermann, D. (ed.): Proc. of the 1997 Symposium on Document Image Understanding Technology. College Park, MD: University of Maryland Institute for Advanced Computer Studies 1997, pp. 87–98
6. Hochberg, J., K. Bowers, M. Cannon, P. Kelly: Handwritten document image analysis at Los Alamos: Script, language, and writer identification. In Proceedings of the 1999 Symposium on Document Image Understanding Technology. College Park, MD: University of Maryland Institute for Advanced Computer Studies, 1999, pp. 161-5

Judith Hochberg holds a BA from Harvard and a PhD from Stanford, both in linguistics. She has been a staff member at Los Alamos National Laboratory since 1989, and is an Adjunct Professor of Electrical and Computer Engineering at the University of New Mexico in Albuquerque, New Mexico. She is interested in human language in all modalities: speech, text, and document images. Multilingual applications are of particular interest.

Kevin Bowers received the B.S. degree in electrical engineering from Purdue University in 1997, and the M.S. in electrical engineering from the University of California at Berkeley in 1998. He is currently pursing the Ph.D. in electrical engineering, also at UC-Berkeley, under a fellowship from the Hertz Foundation. His research interests include signal processing, optical electronics and electromagnetics. Presently he is a member of the Plasma Theory and Simulation group at UC-Berkeley where he is working on large area plasma sources.



Michael Cannon holds a BS in physics from the University of Washington, and a Doctorate in Electrical Engineering from the University of Utah. His specialty is digital image processing. His professional interests include digital image deblurring, multispectral imaging, and pattern analysis. He is a senior member of the IEEE.

Patrick M. Kelly received a Bachelors degree in computer engineering in 1990, and a Masters degree in electrical engineering in 1992, from the University of New Mexico, Albuquerque, New Mexico. He is currently Team Leader for Image Processing and Analysis at Los Alamos National Laboratory, Los Alamos, New Mexico. His research interests include document image analysis, image understanding, pattern recognition, and clustering techniques. Patrick also teaches short courses about computer programming for Learning Tree International.