Handwritten Document Image Analysis at Los Alamos: Script, Language, and Writer Identification

Judith Hochberg, Kevin Bowers, Michael Cannon, and Patrick Kelly

Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545
{judithh, tmc, kelly}@lanl.gov
kbowers@eecs.berkeley.edu

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 281 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. Pilot results indicate that a variation of the method may be applicable to writer identification.

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.

To the best of our knowledge, script identification has never been attempted for handwritten documents. Because of the dramatic individual differences in handwriting, we found a feature-based approach to be most successful, in contrast to the template matching we have previously applied to machine printed documents [2-3]. In the spirit of Wilensky et al. [4], each document was characterized by a single feature vector, containing summary statistics taken across the document's black connected components. The documents were then classified using linear discriminant analysis.

The main focus of this work was script identification: the method was 88% accurate in distinguishing among six scripts, including challenging pairs of related (and visually similar) scripts such as Roman/Cyrillic and Chinese/Japanese. We also took a first look at language identification within the Roman script: the method was 85% accurate for English versus German documents. Finally, we report promising pilot results (80% accuracy for a rough implementation) on a variation of our method applied to writer identification from free text.

2. Data

We assembled a corpus of 496 handwritten documents from six scripts: Arabic, Chinese, Cyrillic, Devanagari, Japanese, and Roman. The scripts are illustrated in Figure 1. For the most part, document images were obtained from foreign language speakers we were acquainted with or whom we contacted through the Internet. Over

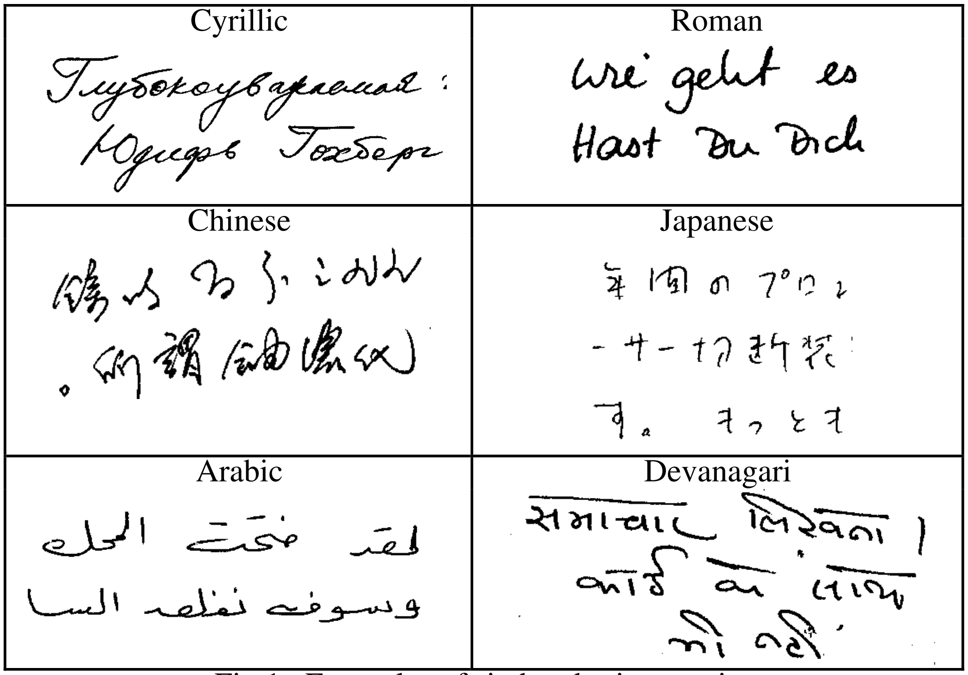

Fig 1. Examples of six handwritten scripts

75% of the documents we collected were 'natural' -- letters, lecture notes, official documents, etc. The remaining documents were written on request. 281 different writers were represented in the corpus.

Around a third of the documents had at least one document quality issue such as ruling lines, line curvature, line skew, character fragmentation, or brevity (fewer than 100 connected components). Character fragmentation and ruling lines were addressed in preprocessing. 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.

3. Script identification

Connected components. The basic element of the analysis was the eight-connected black component. After finding all the components in a document image, unusually small or large components were filtered out in order to remove speckle, ruling lines, and outsize components in general. Some filtering criteria were absolute (e.g., removing components with height or width less than three pixels), and some were relative (e.g., removing components with height or width more than four standard deviations above the document mean).

Features. Once filtering was completed, several features were extracted from the remaining components. To develop the feature set we studied the document images and determined which visual features guided our human script identification. The final set of features was:

- relative Y centroid
- relative X centroid
- number of white holes
- sphericity
- aspect ratio (height/width)

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.

Discrimination. The classification method used a collection of linear discriminant functions. A separate Fisher linear discriminant [5] was trained to separate each possible pair of scripts in the dataset (Arabic vs. Chinese, Arabic vs. Cyrillic, etc.). 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".

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.

Results. The linear discriminant analysis was 88% accurate. Table 1 breaks down these results by script, and also presents the cross-classification matrix. The individual percentages for the different scripts were pleasingly uniform, especially since the amount and quality of data available for the different scripts varied considerably. When documents were misclassified, the errors were sensible: Roman and Cyrillic tended to be confused, and likewise Chinese and Japanese.

Character fragmentation adversely affected classification: 90% of documents with no fragmentation, or only mild fragmentation, were correctly classified, compared to 81% of documents with moderate or severe fragmentation ($F = 5.21$, $p < 0.05$). Ruling lines also appeared to affect classification -- 89% of unruled documents were correctly classified, compared to 81% of ruled documents -- although this difference was just short of statistical significance ($F = 3.18$, $p = 0.07$). Of the 366 documents in the corpus with no or mild fragmentation, and without ruling lines, 91% were classified correctly.

4. Language identification

Method. Of the Roman script documents in the corpus, 107 were in English and 58 in German. Using the same preprocessing, feature selection, and classification techniques described for script identification, we attempted to distinguish between these two groups. We also tried to identify the languages directly, using a single discriminant analysis for the seven-way discrimination among Arabic, Chinese, Cyrillic, Devanagari, Japanese, German, and English.

Results. For the two-way (English vs. German) task, correct identification averaged 85% (84% for English, 86% for German). For the seven-way task, 80% of English and German documents were correctly identified by language. Interestingly, overall *Roman* identification improved to 93% when the two languages were split apart. 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. Classification of the other scripts was not affected by the Roman split.

5. Writer identification

While experimenting with a variation of the method described in sections 3 and 4, we obtained exciting pilot results for writer identification. We present them here, knowing that they are extremely preliminary, in the hopes that they may inspire further research.

Method. Connected components were identified, filtered, and features extracted as described above, producing a five-element vector per component. Then a *k*-means cluster analysis was performed across the

Table 1. Script identification results

Script	% correct	Classified as					
		Arabic	Chinese	Cyrillic	Devanagari	Japanese	Roman
Arabic	89%	51	0	0	3	2	1
Chinese	87%	0	104	0	0	8	8
Cyrillic	88%	1	0	49	2	0	4
Devanagari	88%	0	0	1	22	1	1
Japanese	86%	3	6	0	0	63	1
Roman	91%	2	1	9	0	3	150
Average	**88%**						

entire training set, resulting in 256 clusters, or connected component types. Now each document could be represented as a histogram of cluster occurrences, and documents compared to each other on the basis of their histogram similarity, using a distance metric such as Kullback-Leibler entropy [6].

The pilot study analyzed the 282 documents in our corpus whose writers had contributed more than one document to the corpus. Using the histogram comparison method, we determined how many of these documents were most similar to another document by the same writer -- a one-nearest-neighbor classifier.

Related work by Wilensky et al. on Chinese writer identification from free text also used a nearest-neighbor algorithm, representing each document by a feature vector similar to the ones we used for script and language identification [4].

Results. As shown in Table 2, roughly 80% of documents by multi-document writers were closest to another document by the same writer. This was particularly impressive given that (at the time) the corpus contained 466 documents, representing 260 different writers. In other words, the odds of picking another document by the same writer by chance were extremely low.

In comparison, Wilensky et al.'s accuracy was much higher (98% or better), but their task was much easier since it involved only fifteen writers and 106 documents. In addition, their feature selection (in contrast to ours) was optimized for writer identification.

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 when script and language identification were performed together. It would be worthwhile to explore the language identification task more thoroughly.

Pilot work showed 80% accuracy for writer identification. We believe that these results could be improved with a more judicious selection of features. Since the pilot study was a casual offshoot of our main line of research, the features used to represent documents were chosen for their utility in script identification, not writer identification. In fact, we were careful not to choose features that showed substantial individual variation, such as white component sphericity. A more fully-developed algorithm along these lines could be useful in the intelligence field or in forensics.

References

[1] Spitz, A. L (1977). Determination of the script and language content of document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:235-245.

[2] Hochberg, J., Kelly, P., Thomas, T., Kerns, L. (1997). Automatic script identification from document images using cluster-based templates. *IEEE Transactions on Pattern Analysis and Machine In-*

Table 2. Pilot results for writer identification

Script	# documents by multi-document writers	# correctly matched	% correctly matched
Arabic	14	12	86%
Chinese	99	79	80%
Cyrillic	37	29	78%
Devanagari	19	16	84%
Japanese	3	3	100%
Roman	110	87	79%
total	282	226	80%
total # writers			260

telligence 19176-181.

[3] Patent: *Script identification from images using cluster-based templates* (5,844,991).

[4] Wilensky, G., Crawford, T., Riley, R. (1997). Recognition and characterization of handwritten words. In Doermann, D. (ed.): *Proceedings of the 1997 Symposium on Document Image Understanding Technology*. College Park, MD: University of Maryland Institute for Advanced Computer Studies, pp. 87-98

[5] Duda, T., Hart, P. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, pp. 114-118.

[6] Deco, G. & D. Obradovic (1996). *An Information-Theoretic Approach to Neural Computing*. New York: Springer, p. 10.

Acknowledgments

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36. We would like to thank Steve Dennis, Ron Riley, and Greg Wilensky for their help. 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.