AUTOMATED WRITER IDENTIFICATION FOR SYRIAC SCRIBES

Emma Dalton[1] and Nicholas Howe[2]

Departments of [1]Engineering and [2]Computer Science, Smith College, Northampton, MA

Background and Motivation

The Syriac Language

Syriac (Fig. 1) is an ancient dialect of Aramaic in which over 10,000 manuscripts were written. These manuscripts include some of the ealiest translations of the Bible, and provide a critical link in the preservation of the writings of Aristotle. Syriac Christianity was also one of the most far-reaching branches of the religion, making contact with early Islam, India and even China [1].

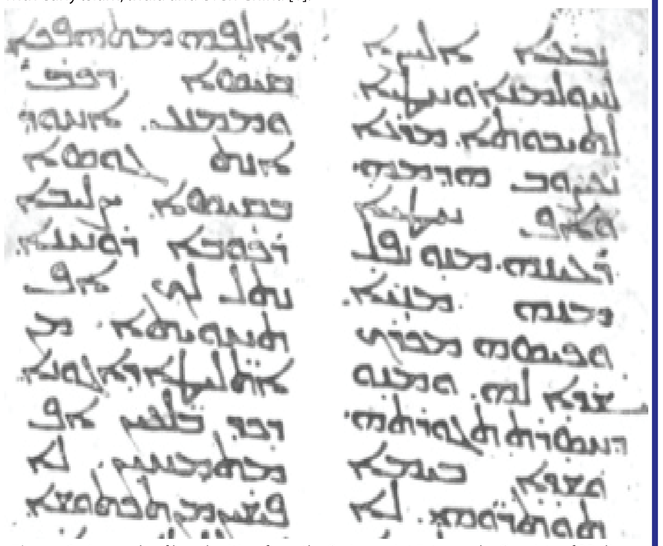


Figure 1: *A sample of handwriting from the Syriac script Estrangelo. Consists of 23 distinct letters used for text-dependent identification.*

Despite all of these compelling aspects, the study of the corpus of Syriac manuscripts is stunted, while the study of the Christian religion has focused mainly on Latin texts. In recent years there has been a shift in attitude, and now there are several universities offering instruction in the Syriac language, and the Vatican has opened its manuscript collection for general academic use.

The Problem of Authorship

While there are many Syriac manuscripts to study, and some are easily attributed to a particular scribe because they include a colophon recording the copyist and date, the majority of preserved documents (95 percent) do not include this important information [1].

The goal of this project is to create a system which can take one document and digitally compare it to other preprocessed manuscripts, to determine possible handwriting matches, with a score to describe how close the match is. This will assist scholars of Syraic by providing them with connections between documents through authorship, which may speed up the process of study considerably.

Manuscripts Used

While many Syriac manuscripts exist, relatively few are available in the digital format needed for computerized identification. The documents used for this project were provided by Brigham Young University in collaboration with the Vatican. From the commercially available CD-ROM, 4 pages from each of 19 documents in the Syriac script Estrangelo were used to create and test a possible scribe identification system.

References

[1] M. Penn. "Draft NEH Proposal", Unpublished, 2007.

[2] E. Learned-Miller, "Data driven image models through continuous joint alignment" IEEE Transactions on Pattern Analysis and Machine Intelligence vol:28, no:2, pp. 236-250 Feb. 2006.

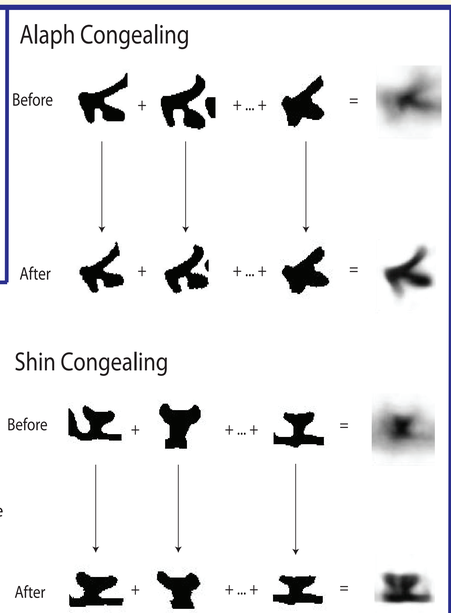
[3] N. Bulacu, and L. Schomaker, "Text-Independent Writer Identifcation and Verifcation Using Textural and Allographic Features", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.29 no.4, pp.701-717, April 2007

[4] I. Bar-Yosef, I. Beckman, K. Kedem, and I. Dinstein, "Binarization, character extraction, and writer identifcation of historical Hebrew calligraphy documents" Int. J. Doc. Anal. Recognit. vol:9, no:2, pp. 89-99, Apr. 2007

Methods

Text-Dependent Methods

These methods require the user to have prior knowledge of the language used. Many letter samples are combined using Congealing [2], to create more "average" characters. The transformations used to get to these more general characters are then used to compare letter samples, and therefore pages of text to one another. This can be done with both whole letters and letter pieces.



Whole Letters

- select character samples from pages of text
- process all letters using congealing
- use congealing transforms to compare individual letters (Fig. 2)

The whole letter method compares individual letters on relative scaling, along with their relative tilt. By comparing letters to the computed "average" letters, the method can easily be adapted to any letter set.

Figure 2: *Comparison of letter samples of alaph and shin before and after congealing.*

Letter Parts and Letter Concavities

- isolate letter parts using the letter skeleton
- isolate concavities from the convex hull by subtracting the entire letter (Fig. 3)
- independently congeal each across the whole letter set

These methods specifically target author specific traits, such as long tails or extra curves in particular letters. This specificity takes extra time to compute.

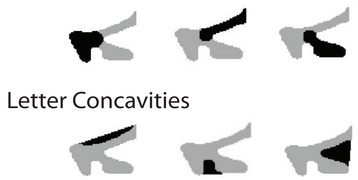


Figure 3: *Examples of Letter Parts and letter concavities isolated from one letter alaph.*

Text-Independent Methods

Derive statistics from large blocks of text. This requires no prior knowledge of the script and is very fast, but also very general. All of the methods used here are derived from Schomaker, et al [3]

The Run-Length Feature

- measures background lengths
- both vertical and horizontal (Fig. 4)
- 2 histograms used to compare text pages (Fig. 5)

Figure 4: *Schematic showing examples of run-lengths on background both vertical and horizontal.*

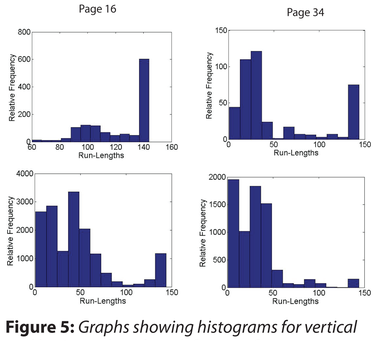
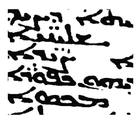


Figure 5: *Graphs showing histograms for vertical and horizontal run length features for two pages.*

The Contour Direction feature

- measures the general tilt of the text
- uses letter contour
- calculates angle φ (Fig. 6)
- histogram compares text pages (Fig. 7).



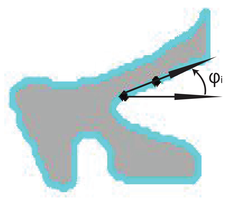


Figure 6: *Schematic showing how Contour Hinge feature is extracted. Angle φ is calculated from the horizontal.*

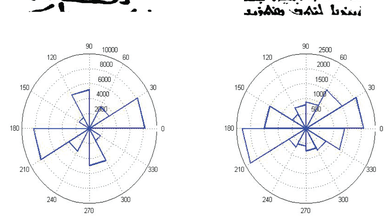
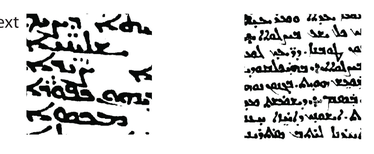


Figure 7: *Comparison of two segments of text using the contour hinge feature. Radial histograms show two pages of text to be dissimilar.*

The Contour Hinge feature

- measures the general "roundness" of the text
- uses letter contour
- calculates two angles, φ1 and φ2 (Fig. 8)
- 2-dimensional histogram compares text pages (Fig. 9).



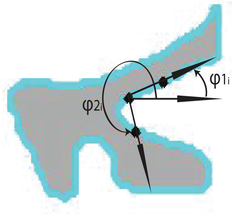


Figure 8: *Schematic showing how Contour Hinge feature is extracted. Two angles, φ1 and φ2 are calculated from the horizontal from each point on the letter contour.*

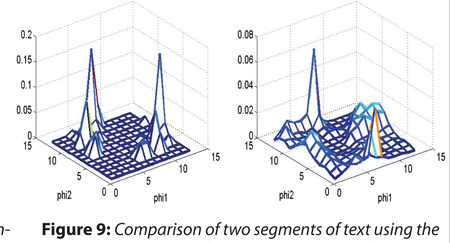


Figure 9: *Comparison of two segments of text using the contour hinge feature. Surface plots show the 2-dimensional histogram of φ1 and φ2. The two plots are significantly dissimilar to provide author identification.*

Results and Discussion

Table 1 shows the results for in-document recognition for each method used. In absence of multiple documents by a sufficient number of authors, in-document recognition is used as a measurement of writer identification. Using a combination of text-dependent and text-independent methods with the dataset used for this project is able to achieve 100% in-document recognition.

Table 1: Percent In-Document Recognition for All Methods Used

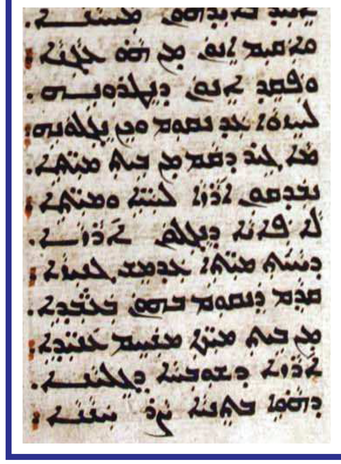
Category	Method Name	Percent In-Document Recognition
Text-Independent	Contour Direction PDF	16
	Contour Hinge PDF	70
	Horizontal Run-Length PDF	16
	Vertical Run-Length PDF	68
	Contour Hinge + Vertical Run-Length	76
	All Text-Independent Methods	66
Text-Dependent	Whole Letters	80
	Letter Parts	79
	Convex Hulls	88
	All Text-Dependent Methods	98.7
Combined	Text-Dependent and -Independent	100

Overall, text-dependent methods performed better than text-independent ones, and the convex hull feature performs best individually with 88% in-document recognition. Of text-independent methods, the contour hinge feature perfomed best, with a 70% recognition rate.

For the combined methods, only the contour hinge and vertical run-length were used from the text-independent methods. Using all of the methods available decreased the accuracy of the system considerably.

Conclusion and Future Work

In an experimental setting, the system developed is capable of working at 100% accuacy for in-document recognition. This is considered highly successful in this context. However, to be a successful system for use by Syriac scholars, the system must be expanded and adapted to compare new page information with a more complete dataset.



Other future work might include:

- Expansion of the current set of documents
- Exploration of more writer-identification methods
- Verification using another language, possibly the Hebrew dataset used by Bar-Yosef et al [4]
- Expansion into the other Syriac scripts, Serto and East Syriac (Fig. 10).

Figure 10: *A sample of writing in the East Syriac script. This is generally blockier than Estrangelo, while Serto is thinner and more cursive.*

Acknowledgements

Thank you to Professor Nicholas Howe for his continued support, and Professor Michael Penn of the Mount Holyoke Religion Department for providing guidance and information about the Syriac language and Syriac studies issues.