


FRAMEWORK FOR QUANTITATIVE ANALYSIS OF SCRIPTS

Rajan, Vinodh 
University of St Andrews

Category: Short Paper
Session: 7
Date: 2014-07-11
Time: 11:00:00
Room: 315 - Amphipôle

1. Introduction

1.1 Overview

Scripts are usually seen as simple carriers of languages. Research on scripts until recently has been minimal and niche, except for the field of paleography. Scripts are an important part of the cultural heritage of humanity and its analysis and study requires more research. Fortunately, there is a growing interest in analysis of scripts. Altmann ^[1] published a volume titled “Analyses of Scripts: Properties of Characters and Writing Systems” to explore various properties of writing systems and scripts such as complexity, ornamentality and distinctivity.

Changizi et al ^[2] discuss the various contour configurations of written symbols and their similarity to the environment in which they were produced. They also study the distribution of the configurations of various scripts. They ^[3] further discuss the character complexity and the redundancy of stroke combinations of various writing systems in human history. It is to be noted that analysis in [2] [3] and most methods in [1] were performed manually. Traditionally, analysis and study in paleography have also been done manually. Digital paleographic methods are at present making more inroads into the field. However, applying quantitative analysis on paleographic data is not yet popular and standardized ^[4]. This is partially due to the difficulty of quantifying paleographical features, and partially due to the lack of defined metrics with theoretical and qualitative underpinnings.

1.2 Proposed Framework

We propose here a quantitative analysis framework for scripts that is largely computational and requires minimal user interaction. We do not particularly aim at providing a completely autonomous framework, but rather to aid the user as much as possible, with the ability to manually intervene/override as required. The framework is based on various methods and techniques employed in the field of graphonomics. Our framework is grounded on the principles of handwriting production and handwriting analysis.

We also explore various features used in the related area of gesture design and recognition and its application to the analysis of scripts. Through this, we attempt to find relevant metrics (with qualitative significance) with sufficient evaluation that might be used for glyphs and scripts for various purposes such as classification, visualization etc. The computed quantitative features could serve as descriptors for scripts, and be used for comparing and analyzing scripts. This is especially applicable to the field of paleography, where such quantitative features are much needed.

2. Quantitative Analysis Framework

Our proposed framework consists of the following modules.

2.1 Spline Conversion

The characters of scripts are externally represented as B-splines. B-splines are very efficient in preserving the shape and curvature of glyphic segments. Additionally, they can be manipulated without significant effort. Rather than representing glyphs as pixelated data, converting them into splines eases analysis. This conversion of the glyphic shape of a character can be done automatically

or manually. In a manual process, the user defines each shape of a character directly using a set of B-splines, or explicitly draws the shape, which is then internally converted into B-splines. An automatic conversion of glyphs involves thinning and then its conversion into splines.

2.2 Trajectory Reconstruction

The shape of a character is static and does not contain all information required for analysis. Dynamic information relating to pen movements are not present in the shape. Trajectory reconstruction attempts to recover this temporal information [5]. This kinematic information is essential in defining the character. With paleographic scripts reconstructing dynamic information is necessary as the trajectory is usually unknown. Also by altering the trajectory, the changes in dynamic features can be observed.

The recovery is performed by conducting a global search using a set of heuristics, such as length minimization and curvature minimization [6]. Especially in case of paleographic scripts, the algorithm is able to provide several alternative viable writing trajectories.

2.3 Stroke Segmentation

Characters are best analyzed as sequences of natural strokes. Breaking them down into basic strokes is the optimal way for analyzing written characters. It also enables us to understand the process of handwriting. Stroke segmentation retrieves the structure of a character based on its trajectory. This is performed by segmentation of the character at various important landmark points of the recovered trajectory such as the *extrema of curvature* [7].

2.4 Character Representation

Writing is usually considered to consist of two fundamental types of strokes – up-strokes and down-strokes [8]. This distinction is necessary since both these two types behave differently. It has been proved that down-strokes usually do not show a lot of variation in handwriting compared to up-strokes [9].

A character is internally represented as a set of strokes. This is consistent with the way that the character is internalized and produced by humans. This allows us to derive better features that are more natural and descriptive. Later, it will be possible to apply handwriting modelling to generate alternative scribal variants.

2.5 Feature Extraction

For quantitative analysis, features need to be computed from the characters. These serve to quantify several aspects of the characters. We considered various features used in the field of gesture recognition [10][11] [12][13] and found the features listed below to be relevant to the analysis of characters. We also propose some features in addition to those found in the literature. These features/metrics could additionally serve as descriptors for the scripts. As quantitative features these can be widely used in analysis and/or visualization.

2.5.1 Production Features

The effort that is required to realize and produce a character is an important element of its analysis. It is related to the *number of velocity inversions, number of velocity breaks, number of pen lifts, etc.*, which are computed from the stroke representation of the character. These features are calculated from the temporal information that was re-constructed at the earlier stage. These are relevant as they quantify the dynamic handwriting behavior present in the character.

2.5.2 Geometric Features

Geometric features throw more light on the visual aspect of characters. These are essential for the study of the judged (visual) complexity [14]. The features that relate to visual appearance are *compactness, openness, number of crossings, average curvature, sum of internal angles, bounding area, etc.* Some of these, like *compactness and openness*, are ratios of several parameters such as *length of strokes, distance between first and last points*, while others, like *average curvature*, are derived directly from the glyph structure.

2.5.3 Cognitive Features

Though cognitive features cannot be directly measured, some cognitive features could be interpreted from the geometry of a character. The number of *unique landmark points* required to plan the trajectory is a possible feature that has correlation with the cognitive load of the glyph. An additional measure is the *number of minimal points* required to recreate the character.

2.5.4 Stroke Features

Stroke based features such as the *primary direction* of the glyph, *ratio of upstrokes to downstrokes*, *direction change*, *histogram of inter-stroke angles* etc. are also computed for a character.

3. Prototype Implementation

A prototype of the framework has been implemented in Python with the modules discussed above. We are planning to analyze the development of Indic scripts using the framework. The source code will be released under an open source license when the project reaches maturity. Its repository will also include a complete a set of Indian paleographic scripts. Below, we briefly describe functionality yet to be implemented in the prototype.

From the perspective of paleographic scripts, the available glyphs in the literature are usually very noisy. Scanning and importing them would require several layers of pre-processing and noise-removal. For modern scripts, importing the Bezier curves from the respective fonts could be done directly.

Trajectory reconstruction has been implemented only for single stroke characters. A primitive implementation exists for multi-stroke characters. This needs to be made more rigorous and accurate.

Based on the stroke structure of a glyph several additional features such as *entropy of writing* could be calculated as required. Also, the features are to be used for normalized glyph shapes. The behavior of the features with respect to various scribal variants needs to be analyzed further.

Proper visualization of the various quantitative features would help to better study and understand the characters within a script and also compare several scripts. Such visualization is particularly helpful in studying paleographic scripts and analyzing the changes that took place over time. Various statistical analyses of the features and visualization techniques would be built into the implementation.

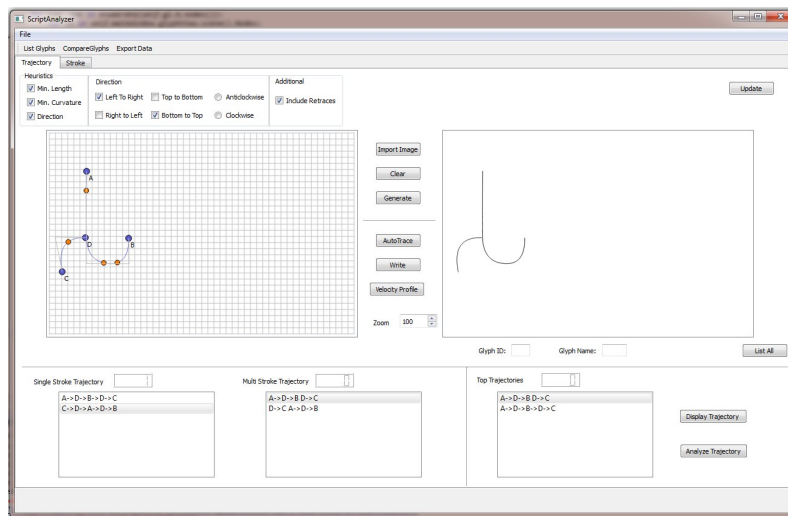


Fig. 1: Spline Conversion and Trajectory Generation

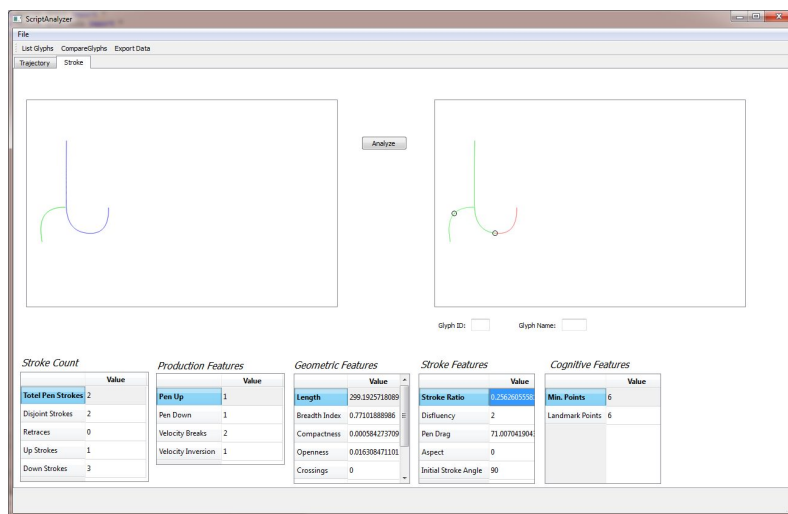


Fig. 2: Stroke Segmentation & Feature Extraction

4. Similar Projects

There are other digital paleographical projects, such as Hand Analyser by Peter Strokes, which work on pixelated images. The current project is more focussed on using *strokes* to derive various features. Integration/Adaptation of those techniques needs to be further looked into.

5. Summary

We have presented a computational framework for quantitative analysis of scripts. The framework requires minimal user interaction, and is based on the principles of handwriting analysis and handwriting production. We also present a prototype implementation of the proposed framework. We believe this framework and its implementation would facilitate more quantitative study on scripts.

References

1. **Altmann, Gabriel, and Fan Fengxiang** (2008), eds. *Analyses of script: properties of characters and writing systems*. Vol. 63. Walter de Gruyter. APA
2. **Changizi, Mark A., et al.** (2006) *The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes*. *The American Naturalist* 167.5: E117-E139.
3. **Changizi, Mark A., and Shinsuke Shimojo.** (2005) *Character complexity and redundancy in writing systems over human history*. *Proceedings of the Royal Society B: Biological Sciences* 272.1560: 267-275.
4. **Stokes, Peter.** (2009) *Computer-aided palaeography, present and future*. *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*. *Schriften des Instituts für Dokumentologie und Editorik*, 2: 309-338.
5. **Doermann, David S., and Azriel Rosenfeld.** (1995) *Recovery of temporal information from static images of handwriting*. *International Journal of Computer Vision* 15.1-2: 143-164.
6. **Jager, Stefan** (1996). *Recovering writing traces in off-line handwriting recognition: Using a global optimization technique*. *Pattern Recognition*,., *Proceedings of the 13th International Conference on*. Vol. 3. IEEE, 1996.
7. **Li, Xiaolin, Marc Parizeau, and Réjean Plamondon** (1998). *Segmentation and reconstruction of on-line handwritten scripts*. *Pattern recognition* 31.6: 675-684.
8. **Noordzij, Gerrit, and Peter Enneson** (2006). *The stroke: Theory of writing*. Hyphen.
9. **Teulings, Hans-Leo, and Lambert RB Schomaker** (1993). *Invariant properties between stroke features in handwriting*. *Acta psychologica* 82.1: 69-88.
10. **Rubine, Dean** (1991). *Specifying gestures by example*. Vol. 25. No. 4. ACM.
11. **Long Jr, A. Chris, et al** (2000). *Visual similarity of pen gestures*. *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM.
12. **Willems, Don, et al** (2009). *Iconic and multi-stroke gesture recognition*. *Pattern Recognition* 42.12: 3303-3312.
13. **Tucha, Oliver, Lara Tucha, and Klaus W. Lange.** (2008) *Graphonomics, automaticity and handwriting assessment*. *Literacy* 42.3: 145-155.
14. **Stenson, Herbert H.** (1966) *The physical factor structure of random forms and their judged complexity*. *Perception & Psychophysics* 1.9: 303-310.