Computer-based Stroke Extraction in Historical Manuscripts

Rainer Herzog, Bernd Neumann, Arved Solth

Technical Report: Report FBI-HH-B-296/10

Universität Hamburg, Department Informatik
Arbeitsbereich Kognitive Systeme

December 2010

Zusammenfassung

In diesem Bericht wird ein Verfahren zur Extraktion von Strichen in handgeschriebenen Zeichen verschiedener Schriftsysteme vorgestellt. Das Erkennen einzelner Striche als Bestandteile von Zeichen kann für verschiedene Ziele der Manuskriptanalyse, insbesondere der Analyse historischer Manuskripte, von Bedeutung sein, z.B. als Basis für den Vergleich ähnlicher Zeichen, zur Identifizierung von Schreibern, oder zur Zeichenerkennung. Das vorgestellte Verfahren beruht auf der Constrained Delaunay Triangulierung (CDT), die für die Zerlegung ebener Formen vorgeschlagen wurde. Angewandt auf handschriftliche Zeichen markiert das Verfahren anhand von Kontureigenschaften mögliche Anfangs- und Endpunkte sowie Kreuzungen von Strichen. Damit stehen Kontursegmente zur Verfügung, aus denen vollständige Striche gebildet werden können. Es werden experimentelle Ergebnisse für chinesische, amharische und tamilische Zeichen vorgestellt.

Computer-based Stroke Extraction in Historical Manuscripts

Rainer Herzog, Bernd Neumann, Arved Solth

Abstract

Recovering individual strokes in historical manuscripts can provide a valuable basis for various goals of manuscript analysis, e.g. retrieving similar allographs, comparing the handwriting of scribes, or recognising characters. Here we report on stroke analysis using the Constrained Delaunay Triangulation (CDT), previously proposed for shape decomposition in image analysis. Applied to handwritten graphemes, this method marks possible start, end points and intersections of strokes based on local contour properties, thus providing stroke segments from which complete strokes can be formed by concatenation. Results are shown for Chinese, Amharic and Tamil characters.

1. Introduction

Research into historical manuscripts often leads to questions in which the analysis of visual features can be helpful. Writing style, layout and shape of characters may provide significant clues about the cultural origin of a manuscript, relationships to other manuscripts, or even the identity of scribes. For example, as pointed out by Richter[1], visual features may constitute an important criterion for reassembling unearthed fragments of early Chinese manuscripts. He showed that angles between strokes of a Chinese character in different parts of a manuscript can differ systematically and hence indicate their origin from different scribes.

To derive a sound judgement from visual features, several complex and laborious tasks may have to be performed. Let us consider the task of scribe identification, in which an unidentified piece of writing must be assigned to one of two scribes A and B, whose writing is known from available samples. Firstly, one has to determine features suitable for distinguishing between the two scribes. Within limits, the human eye is quite apt at discerning regularities and differences between comparable shapes. But not all significant features are equally salient, and hence useful features such as the relationship between stroke lengths may be overlooked. Also, some features may be less suitable than others because of high variability within the samples of a single scribe. Secondly, once features have been selected for comparison, a preferably large number of occurrences must be extracted from both the known samples and the unidentified piece of writing in order to ascertain the discriminating value of a particular feature through objective statistical criteria.

Note that in this example, as well as other tasks such as layout analysis, it is not important to recognise individual characters or graphemes of the respective writing system. Rather it suffices to identify and characterise reoccurring patterns, be they meaningful or not. The approach is similar to forensic handwriting analysis, where

[1] Richter 2006

writing style and not contents is the issue. In fact, the relevance of forensic methods for the analysis of historical manuscripts has been emphasised by paleographers[2].

This report is about computer support for manuscript analysis. It is part of the research group "Manuscript Cultures of Asia and Africa"[3], in which a number of scholars of in various humanities fields are investigatinge, among other things, the outward appearance of manuscripts in relation to their cultural function. These scholars are providing the palaeographical, archaeological and cultural background knowledge that is indispensable for analysing historical manuscripts, while the research team of the authors of this paper is focusing on technical aspects.

Judging from the scenario of scribe identification described above, one can conclude that computer support would be beneficial for several subtasks:

(i) Computing shape features
(ii) Discovering distinguishing features
(iii) Retrieving similar patterns from image databases
(iv) Providing objective criteria for scribe identification

Basic methods and techniques for such subtasks are available from several subfields in Computer Science, in particular, Pattern Recognition, Computer Vision and Artificial Intelligence. Most of these fields have developed out of a long history of research into computer-based handwriting recognition. Optical Character Recognition (OCR) of handwriting, also called off-line character recognition, is today in widespread industrial use, however mainly for constrained applications such as business forms or postal address reading[4]. On-line character recognition is based on the temporal sequence of stylus positions that are sensed with special writing equipment. On-line handwriting analysis contains additional temporal information, and thus is at an advantage compared to off-line analysis, but obviously it cannot be applied to historical manuscripts. Our approach of making stroke analysis a methodological part of off-line analysis can be seen as an attempt to reconstruct valuable on-line information.

Unfortunately, existing handwriting-analysis technology is not adequate for historical manuscripts – they pose new challenges. Firstly, analysing historical manuscripts typically requires techniques for reading handwriting that is less constrained with regard to expected layout and contents than, say, postal address reading. Furthermore, the writing material has often been corrupted by age, thus requiring more sophisticated methods for coping with incomplete and noisy data. Finally, the whole process may be aggravated by a lack of knowledge about historical writing conventions. These more difficult requirements are only partly offset by a less stringent need for segmenting text into characters, recognising characters and forming meaningful words, as has been pointed out above.

This report is about subtask (i), computing shape features, and in particular about the step of extracting strokes as an important basis for feature computation. Inspired by the

---

[2] Dalton et al. 2007

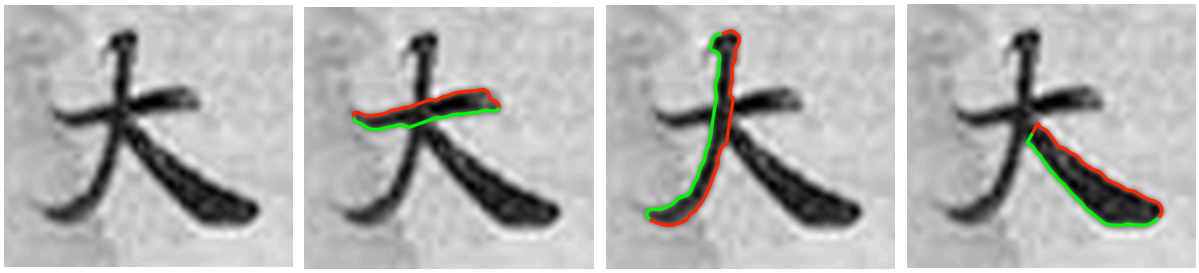
[3] This work has been supported by DFG NE279/10-1 as part of the DFG Research Group 963 "Manuscript Cultures in Asia and Africa".

[4] Cheriet et al. 2009

pioneering work of Richter[5] and in view of the significance of strokes in Chinese characters, we have primarily used Chinese historical manuscripts as examples, but we will also present results with handwriting in other writing systems.

Strokes are commonly defined as the trace of a writing instrument from the point where it first contacts the writing surface to the point where it leaves the writing surface. Extracting strokes can therefore be viewed as a task of partially recovering temporal information. Other definitions constrain a stroke to being approximately straight or smoothly curved, according to a fluent motion. A more precise definition, based on a cognitive model of writing motion, is the Delta LogNormal model[6], but it does not necessarily apply to handwriting with historical tools. Our approach is therefore based on visible evidence for strokes, such as beginnings, endings and direction discontinuities.

Fig. 1 illustrates the stroke extraction for a Chinese character performed with our tools[7]. The image shown in Fig. 1a is taken as input, and the closed contours in the images of Figs. 1b–1d represent extracted strokes as output. Here, the writing direction has also been tentatively reconstructed based on conventions known for modern Chinese writing. The green part of a contour is to the left, the red part to the right, with regard to the direction of writing.



a b c d

Fig. 1: Image of Chinese character (a) and superimposed extracted contours (b - d). The colours indicate tentative writing directions (green at right and red at left).

Our approach is based on a special kind of triangulation of the region occupied by the character called Constrained Delaunay Triangulation (CDT). CDT was first proposed for engineering applications[8] and more recently also applied to printed Chinese characters[9]. Parallel to our research, the analysis of handwritten characters using the CDT has also been investigated following an algorithmic procedure different from ours[10].
The remainder of this contribution is structured as follows. In Section 2, we briefly describe the pre-processing used to extract the contour of a character or graph from the

[5] Richter 2006

[6] Guerfali and Plamondon 1995

[7] Solth et al. 2009

[8] Seidel 1988

[9] Zou and Young 2001

[10] Nel et al. 2009

image. In Section 3, we present the Constrained Delaunay Triangulation in detail, and show how it provides valuable information for analysing written characters. To obtain stroke contours, the ribbon-like components have to be connected in a plausible way, reflecting the natural course of writing and conventions. This is described in Section 4. Experimental results for Chinese and Tamil writing are presented in Section 5. In Section 6, we discuss related work. We then conclude with a summary and a description of our next steps in manuscript analysis based on extracted strokes.

2. Contour extraction

In this section, we shortly describe the main pre-processing steps leading to the polygonal contour representation that is then used as input for stroke extraction. In contrast to commonly used binarisation procedures, we employ watershed binarisation with subpixel accuracy[11] to avoid artefacts arising from coarse pixel structures.

We assume here that the input image has been determined manually or by a preceding segmentation step. Ideally, the image will contain a character or grapheme extracted from the manuscript, but perfect segmentation is not a vital precondition for binarisation and stroke extraction. The idea of the subpixel (or "exact") watershed transform is to determine boundary lines in a continuous version of the image, reconstructed from the discrete image by spline interpolation.

The first step is to transform the image into a boundary image using a boundary indicator function such as the gradient magnitude. Gradient magnitudes can only be determined for half the sampling rate compared to the sampling rate of the grey-value image[12]. To this end, the original image is smoothed with a discrete Gaussian derivative filter at the scale $\sigma=2$. The boundary image can then be viewed as a hilly relief with peaks and ridges at places of large grey-value changes in the original image.

In a second step, the boundary image is transformed into a continuous 3D surface using spline interpolation. Because of the differentiability requirements for the watershed computations, it is useful to choose 5th degree splines.

The next step is to extract watershed lines. These are the locations in the (continuous) relief of the boundary image relief, where a hypothetical water drop could run down to more than one regional minimum. Mathematically, watershed lines connect a local maximum with a saddle point, and the computational procedure is based on determining maxima and saddle points from the spline interpolation. Watershed lines have the nice property that they always form closed boundaries and do not cause artefacts at junctions.

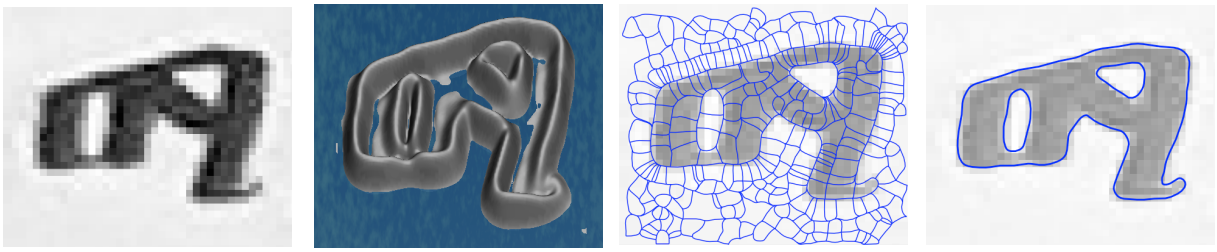
For handwriting images, the result is usually an over-segmentation with many meaningless edges caused by noise or irrelevant grey-value variations. The final step is therefore a merging step, in which edges are removed that separate regions with similar grey-values, or that delineate isolated small regions. Here, thresholds have to be set with care. Polygonal contour representations of the remaining components are obtained by

11 Meine and Koethe 2005

12 Koethe 2003

sampling the contour lines with a density appropriate for the subsequent triangulation (Section 3).

Fig. 2 illustrates this procedure for a noisy low-resolution image of a character in an Amharic manuscript. Note the realistic contours (d) obtained by the subpixel binarisation procedure in spite of the coarse original (a).



a b c d

Fig. 2: (a) Coarse image of an Amharic character, (b) smoothly interpolated boundary image viewed as a hilly relief, (c) over-segmentation by watershed lines, (d) final segmentation after removing insignificant edges

3. Constrained Delaunay Triangulation

In this section, we will describe Constrained Delaunay Triangulation (CDT), which we use to decompose the contour of a grapheme (or character) for stroke extraction. It is applied to the polygonal contour representation obtained from the segmentation procedure described above.

CDT is a variant of Delaunay Triangulation (DT), which is widely used in Computer Graphics and Engineering for surface modelling. Given a set of points, DT connects pairs of points in such a way that the circumference of each resulting triangle does not contain any other point. It is optimal in the sense that it maximises minimal angles, thus avoiding slim triangles and computationally difficult solutions.

CDT is an adaptation of DT that is used to deal with points enclosed by a polygonal boundary[13]. To avoid triangles outside the boundary, the circumference condition is modified as follows:

> Three points form a triangle if the circumference of the triangle only contains points not *visible* from the three points. A point P is defined as visible from a point Q if the line segment PQ does not cross any polygon edge.

For our purposes, CDT is performed only with points representing the contour of the grapheme; Fig. 3 illustrates the result. Note that three types of triangles can be distinguished according to the number of chords they contain (a chord is an edge that does not coincide with a polygon edge): *junction* triangles (in green) with three chords, *sleeve* triangles (in blue) with two chords, and *terminal* triangles (in red) with one chord.

These triangle types provide useful information for decomposing a polygon into stroke-like components. One can observe the tendency that isolated smooth lines are filled by

[13] Seidel 1988

sleeve triangles, crossings and sharp corners are marked by junction triangles, and line endings by terminal triangles.

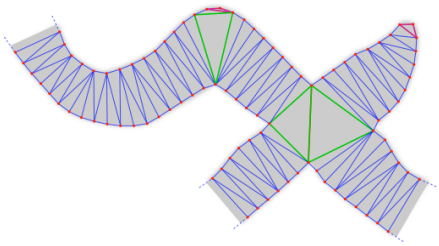


Fig. 3: Result of a CDT consisting of junction triangles (green), sleeve triangles (blue), and terminal triangles (red).

The figure also shows that the occurrence of junction triangles at line corners depends on the curvature. Exact conditions can be derived with help of Fig. 4. It shows the contours of a stroke drawn with an idealised circular stylus with radius S. The cross section of the stylus is shown as a shaded circle on the left. The stroke begins with a straight section, followed by a curve with radius R and angle α, and continues with another straight section. The centre of the stroke is indicated by the heavy dotted line. The shaded triangle and all other triangles generated in the curve are sleeve triangles since the circumference of these triangles lies within the contour circle with radius R and hence does not contain additional contour points. Fig. 4b shows a curved line drawn with R = S. This is the maximal curvature achievable with an idealised circular stylus. Regardless of the angle α, this situation will give rise to junction triangles, because the circumference of a triangle as shown will contain additional contour points.

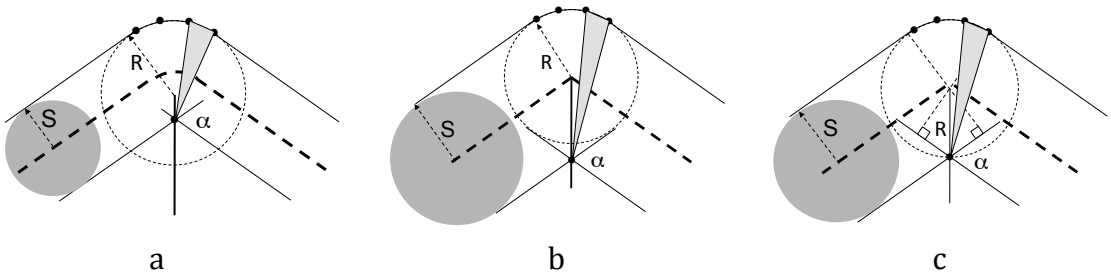


Fig. 4: Examples of curved lines with angle α and outer contour radius R, drawn with a circular stylus, radius S.

(a) No junction triangles for this combination of S, R and α
(b) The triangle shown violates the Delaunay condition. There will always be junction triangles for R = S.
(c) Boundary case with R (1 + cos α/2) = 2S

Fig. 4c illustrates the boundary case: A curved line with angle α and drawn with a stylus of radius S will generate a junction triangle, if the radius R of the outer contour satisfies the following inequality:

$$R < \frac{2S}{1+\cos\alpha/2} \qquad (1)$$

Hence junctions will only occur for sharp bends with an outer contour radius less than the stylus diameter. Note that in all junction cases, the inner contour shows a corner, i.e. a discontinuous direction change. Furthermore, all centerline corners will give rise to junctions. From the perspective of handwriting kinetics, centerline corners are locations

where the writing velocity has come to a standstill, which conforms well with the conceptual notion of a stroke.

In summary, besides marking crossings or forks, junction triangles are useful indicators of line corners, at least under favourable conditions, and may thus mark hidden stroke endings. However, there are possible reasons for exceptions. For one, consecutive strokes in fluent writing may have soft transitions[14], so a corner may become a curve not marked with a junction triangle. Second, small distortions may give rise to spurious junction triangles as shown in Fig. 5. These may, however, be easily recognised because of the small size of the distortion and no significant directional change along the stroke.

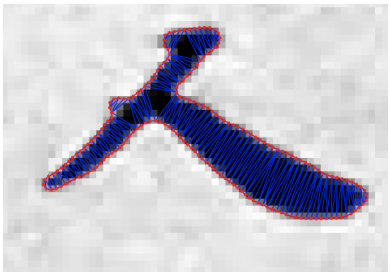
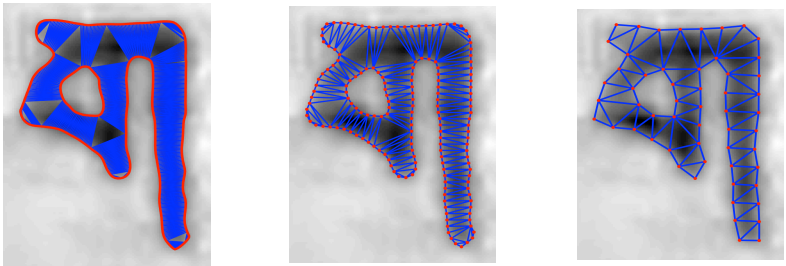


Fig. 5: Small distortions with a contour radius less than half the stroke width may cause junction triangles.

The distance between contour points also plays a part. With close contour points, CDT approximates a procedure for finding the *skeleton* of a shape. In its most popular definition[15], a skeleton consists of the centres of circles fitted to opposing tangential contour pieces – this is exactly the limiting case of sleeve triangles for densely spaced boundary points. Hence for obtaining a triangulation following the skeleton of a shape, dense boundary points are desirable.



a b c

Fig. 6: Triangulations with different spacings of boundary points. Junction triangles at the upper left and right corners and at the lower stroke endings are generated in (a) and (b), but not in (c). The spacing of boundary points must be chosen carefully, dependant on the details that are to be captured.

However, one must also keep computational efficiency in mind if many graphemes are to be processed. For this reason, the chosen spacing of boundary points should be as wide as possible without sacrificing important contour details. Fig. 6 shows triangulations of a Tibetan character with different spacings of boundary points.

[14] Wienecke 2003

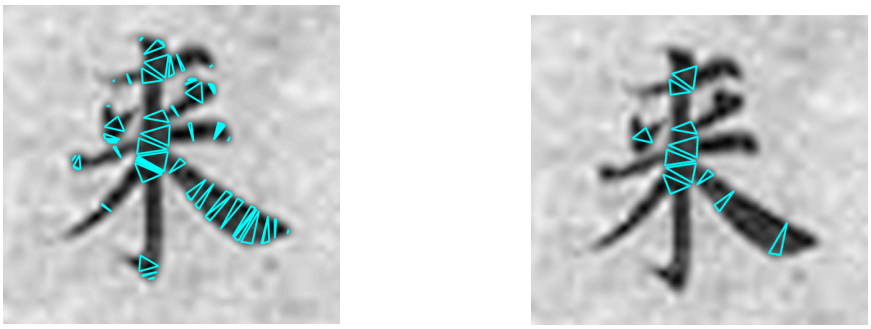
[15] Blum and Nagel 1978

4. Determining strokes

In this section we will describe how strokes can be extracted from a triangulated grapheme. We have implemented a greedy algorithm[16], in which stroke segments are merged at junctions in a single pass without exploring alternatives. This may lead to errors because of premature local decisions in ambiguous situations. The basic idea is to merge stroke segments at junctions if they form a smooth stroke. Stroke segments are defined as areas between either junctions or endpoints and can be easily obtained through triangulation. We distinguish between three kinds of stroke segments: *isolated stroke segments* with no junction at either end, *terminal stroke segments* with one junction, and *connecting stroke segments* between two junctions. Each stroke segment has well-defined properties which can be derived from the triangulated grapheme:

- two lateral contours defined by the contour points between endpoints or junction triangles,
- a medial axis defined by the triangle centres,
- two endpoints defined by the endpoints of the medial axis.

Before merging stroke segments at junctions, some pre-processing of the junctions must be performed to eliminate spurious junctions with an edge length below a threshold, and to merge junctions at apparent stroke crossings. Short stroke segments between two junctions are taken to indicate a crossing and are marked as such. They may become part of more than one stroke. Fig. 7 shows the junctions before and after junction pre-processing.



a b

Fig. 7: Junctions (a) before and (b) after pre-processing

Stroke segment merging proceeds as follows. First, all isolated stroke segments are immediately classified as strokes. Then all junctions are visited and adjoining stroke segments with sufficient directional continuity are merged. If one of the segments is marked as a crossing, it may also participate in a second merging at the same junction, forming the second stroke of the crossing.

After merging pairs with directional continuity, stroke segments are considered for which no merging partner could be found at a junction. Let the unmerged end of the

[16] An algorithm is called greedy if locally optimal decisions are taken irrevocably in the hope of achieving a global optimum.

stroke segment be called A and the other end B. A segment is processed according to the following rules:

1. If the segment is connected to a junction at B but not merged, then the segment is a stroke between A and B.
2. If the segment is already merged with another segment at B, then A is a stroke end.
3. If the segment has a free end at B and its length exceeds its width, then the segment is a stroke.
4. If the segment has a free end at B and its length does not exceed its width, then the segment is a distortion and is merged with one of the other segments at the junction.

The strokes resulting from merging the segments of the grapheme in Fig. 7 are shown in Fig. 8. The writing directions have been reconstructed based on modern conventions of Chinese handwriting and marked by green contours to the right and red contours to the left.

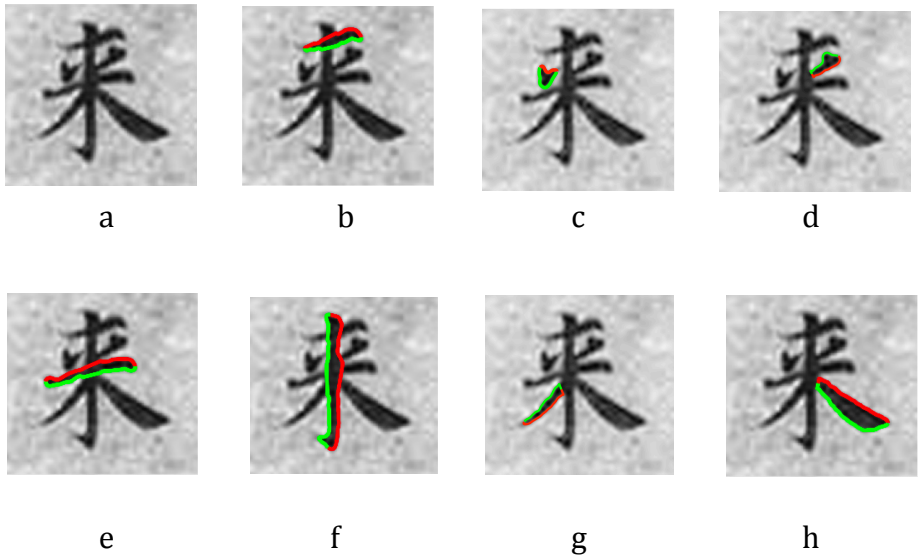


Fig. 8: Image of Chinese character (a) and superimposed extracted contours (b–h). The colours indicate tentative writing directions (green at right and red at left).

5. Experimental results

Our stroke extraction procedure has been evaluated with 339 characters of a section of the historical Chinese manuscript shown in Fig. 9. Ground truth, i.e. information about the correct results, has been provided by one of the authors. The digital images have a resolution of about 60 x 60 pixels per character. The recognition rate is plotted in Fig. 10 as a function of character complexity, measured by the number of strokes. The mean recognition rate for all characters is 72%, for characters with up to 8 strokes 86%. The zero-rate entries 19 and 22–26 are due to a complete lack of such characters in the dataset.

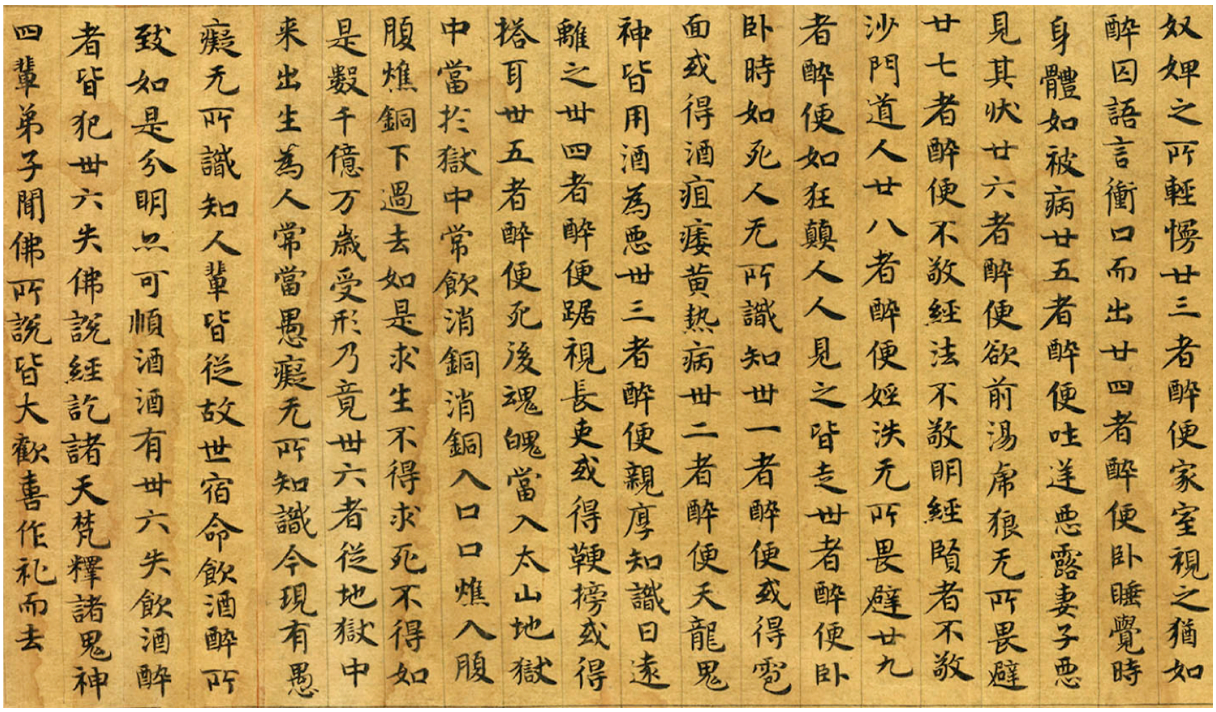


Fig. 9: Section of the *Fo shuo Tiwei jing* 佛說提謂經 (British Library Or.8210/S.2051) used for testing junction-based stroke extraction

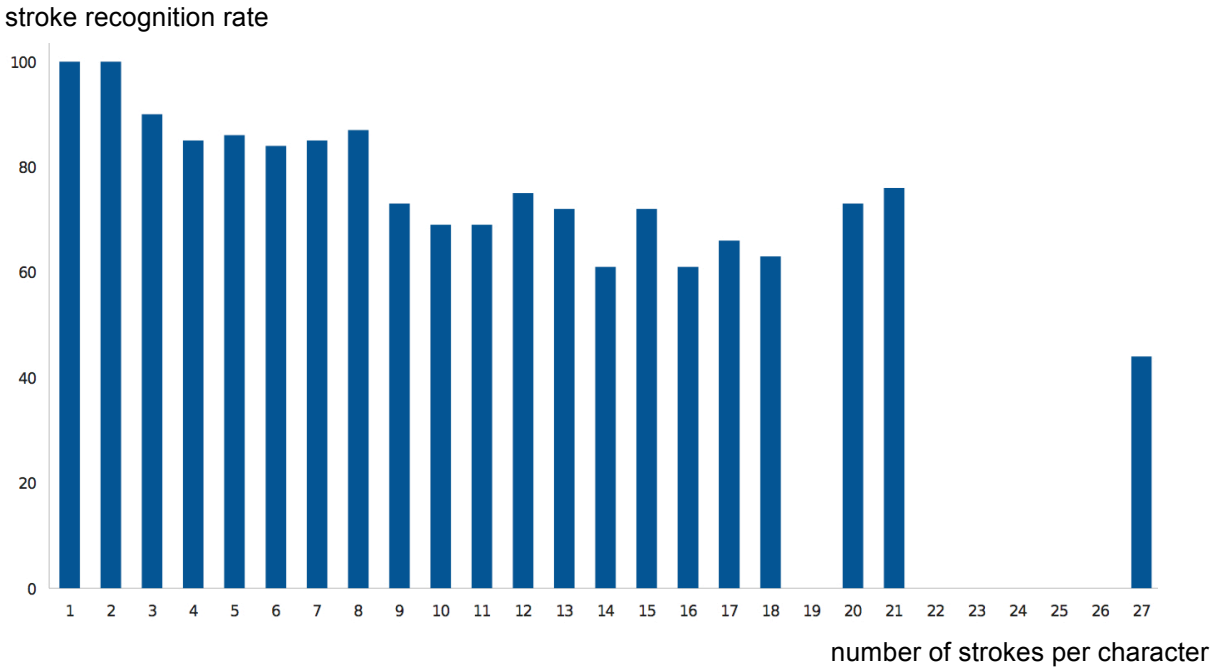
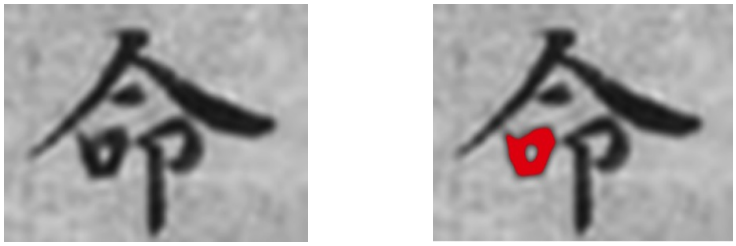


Fig. 10: Stroke recognition rates for characters with increasing complexity

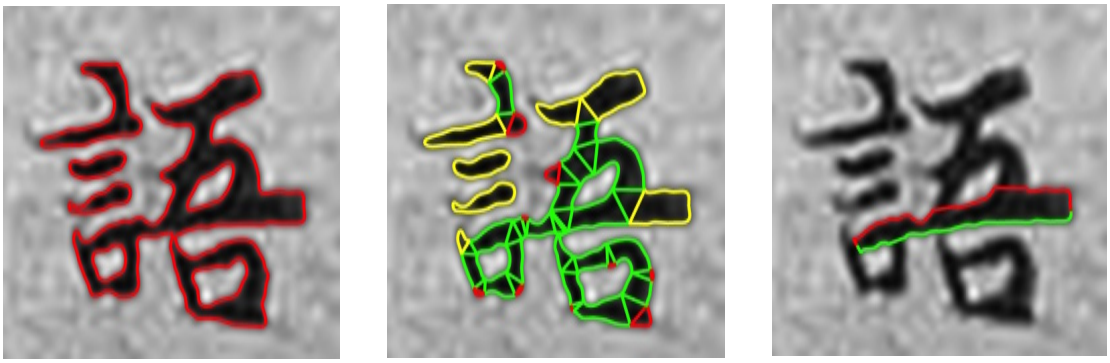
Recognition failure had mainly two reasons. First, junction triangles were not generated when two distinct strokes formed a smooth corner, as shown in Fig. 11. Here, the rectangular shape shown in red has been drawn with three strokes (the upper horizontal and the right vertical section are drawn as a single stroke), but the strokes could not be separated.

Second, stroke segments connected due to imperfect handwriting and subsequent faulty segmentation could not be identified as separate strokes, as shown for an example in Fig. 12. Knowledge about the writing system must be applied to avoid such mistakes.



a b

Fig. 11: Failure of recognising three strokes due to rounded corners. (a) Original character, (b) triangulated part of the character without junction triangles

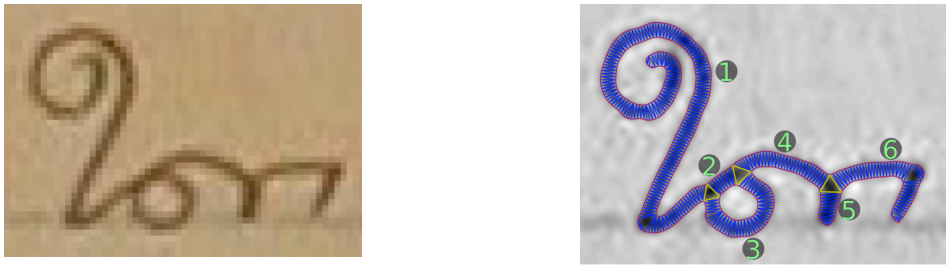


a b c

Fig. 12: Failure of recognising two separate strokes because of a segmentation fault. (a) Segmentation, (b) stroke segments after triangulation, (c) faulty merged stroke

We also applied our approach to a number of Tamil paper manuscripts from the last centuries. In contrast to Chinese characters, syllables are largely written without lifting the pen from the paper. Comparable to Latin script, few parts are passed twice with the writing instrument. Fig. 13 shows some results.

We have followed essentially the same procedure for reconstructing strokes as described above, however using a different merging strategy to better cope with ambiguities in Tamil syllables. Instead of taking irrevocable ("greedy") decisions at junctions, we save alternatives until the end and then submit them to a final rating.



a b

c d e

Fig. 13: Results of applying stroke recognition to a specimen of writing of the Tamil syllable *lai*. (a) Original, (b) triangulation resulting in six strokes segments 1–6, (c)–(e) highest-ranking reconstructions, (c) in terms of a single stroke with Segments 2 and 5 passed twice, (d) in terms of two strokes, in which Segment 3 is separate and Segment 5 is passed twice, and (e) in terms of two strokes in which Segment 2 is passed twice and Segment 6 is separate.

The following criteria for merging stroke segments are applied:

1. Each syllable should be composed of as few strokes as possible.

2. As the main direction of writing is from left to right, the starting position of each stroke should be at its left end.

3. Strokes should begin and end at terminal triangles.

4. At junction triangles, merged stroke segments should have directional continuity.

5. Each stroke segment must be passed at least once, but preferably not more than once.

6. Stroke segments that are passed more than once should be rather straight and short.

7. For the reconstruction of a syllable with more than one stroke, the space between the end of one stroke and the start of the next should be small.

6. Related work

Computer-based analysis of historical manuscripts is a highly interdisciplinary field, with contributions from Pattern Recognition (a subfield of Computer Science), Digital Libraries, Palaeography, and Forensics as major sources.

In Pattern Recognition, historical handwriting analysis is an emerging field that has seen an increasing number of contributions at related conferences, in particular the ICDAR (International Conference on Document Analysis and Recognition), ICFHR (International Conference on Frontiers in Handwriting Recognition) and DAS (International Workshop on Document Analysis Systems), as well as in related journals such as IJDAR (International Journal on Document Analysis and Recognition), ICPRAI (International Journal of Pattern Recognition and Artificial Intelligence) and TPAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence). Most contributions combine existing methods in specific ways to meet the challenges of historical documents. As a recent example[17], Bar-Yosef and co-workers present an approach to writer identification in

[17] Bar-Yosef et al. 2009

historical Hebrew calligraphy documents, comprising an elaborate segmentation step, letter extraction based on medial-axis models, and writer identification using Linear Discriminant Analysis. A similar approach has been reported for Chinese calligraphic handwriting[18], with character models represented by point configurations. Our approach differs by choosing strokes as an intermediate representation; thus it is applicable to a variety of writing systems.

The idea of recovering temporal information and strokes from historical handwriting for improved off-line analysis was first proposed by Doermann and Rosenfeld[19]. They analyse Latin handwriting regarding endpoint and intersection types, relative width and other features to come up with stroke hypotheses and a tentative temporal order. In the dissertation of Wienecke[20], stroke analysis is refined by combining visual features with a biomechanical handwriting model. Invoking such a model would also be useful for our work, but it is difficult to obtain for historical writing tools. Elbaati and co-workers investigate stroke order recovery for off-line Arabic handwriting recognition[21]. Here, strokes are extracted by skeletonisation and placed into a temporal order by maximising a quality criterion with a Genetic Algorithm. Within strokes, a velocity profile is generated based on a biomechanical writing model.

Stroke extraction for Chinese handwriting recognition has been developed by several researchers[22]. All approaches are based on skeletonised characters and in this respect are indirectly related to our triangulation approach, which can also be used to identify skeletons. However, since our aim is handwriting characterisation and not recognition, we prefer stroke representations based on complete characters and not skeletons.

Our long-term goal of providing a toolkit for the analysis of historical manuscripts is shared by several other groups. Moalla and co-workers report about a first approach for analysing the writing style in medieval Latin texts "for the service of palaeography science"[23]. Aiolli and Cuila report on their "System for Paleographic Inspection (SPI)", which has been developed over several years and provides a software suite for characterising and comparing letters in medieval Latin manuscripts[24]. Stokes gives a thoughtful analysis of the requirements for digital palaeographical support and a review of past attempts[25]. He points out that palaeographers hesitate to accept automatic methods for scribe identification, and prefer an interactive, stepwise approach. This differs from forensics, where black-box systems are preferred for scribe identification[26].

---

[18] Zhuang et al. 2004

[19] Doermann and Rosenfeld 1993

[20] Wienecke 2003

[21] Elbaati et al. 2009

[22] Wu 2000, Larmagnac and Dinet 2000, Lin and Tang 2002, Ban et al. 2003 and others

[23] Moalla et al. 2006

[24] Aiolli and Cuila 2009

[25] Stokes 2009

[26] Franke and Srihari 2008

An interesting example for applying a black-box system, originally developed for forensics, to historical documents is reported by Ball and co-workers[27]. They have solved a dispute about two pages of a handwritten satirical newspaper dated 1846, which was suspected to have been authored by Herman Melville, the well-known author of *Moby Dick*. By analysing these pages and some letters known to be handwritten by Melville with the forensic examination system CEDAR-FOX, the authorship of Melville could be established with high certainty.

7. Conclusions and Outlook

We have presented our approach and initial results for stroke analysis of graphemes in historical manuscripts. Strokes are an important basis for characterising handwriting in the context of various tasks, as for example comparing documents with regard to their cultural origin or verifying the identity of scribes. Our approach aims at performing stroke analysis with as little knowledge of the characters and their meaning as possible, in order to provide a generally useful tool for manuscript research. However, as the experiments with Chinese characters have shown, results can be improved in further processing stages by exploiting prior knowledge about the conventions and peculiarities of a particular writing system.

The next step of our work will be the development of a similarity measure based on stroke properties. For Chinese characters, a relational structure consisting of the strokes of a character and their relative positions will be used for comparison, allowing for partial matches. Preliminary work on Tamil handwriting indicates that comparisons could be based on stroke segments forming bends and loops.

Regarding the long-term goal of developing a toolbox for computer-supported manuscript analysis, the current work can only be considered a small step, with more time and resources still required.

References

F. Aiolli, A. Ciula: A case study on the System for Paleographic Inspections (SPI): challenges and new developments. In: Computational Intelligence and Bioengineering, Vol. 196 Frontiers in Artificial Intelligence and Applications, IOS Press, 2009, 53-66.

G.R. Ball, R. Stittmeyer, S.N. Srihari: Writer Verification in Historical Documents. Proc. Document Recognition and Retrieval XVII San Jose, CA, January 2010, SPIE

T. Ban, C. Zhang, W. Shu, Z. Kou: A Novel Approach in Off-line Handwritten Chinese Character Stroke Segmentation. In: Proc. Fifth Int. Conf. Computational Intelligence and Multimedia Applications (ICCIMA03), 2003, 314-318

I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, U. Ehrlich: Adaptive shape prior for recognition and variational segmentation of degraded historical characters. Pattern Recognition 42, 2009, 3348-3354.

M. Cheriet, M. Yacoubi, H. Fujisawa, D. Lopresti, G. Lorette: Handwriting recognition research: Twenty years of achievement... and beyond. Pattern Recognition 42, 2009, 3131-3135

[27] Ball et al. 2010

H. Blum, R. N. Nagel: Shape description using weighted symmetric axis features. Pattern Recognition, 10(3): 167-180, 1978.

J. Dalton, T. Davis, S. van Schaik: Beyond Anonymity: Paleographic Analyses of the Dunhuang Manuscripts. Journal of the International Association of Tibetan Studies, No. 3, 2007, 1-23

D.S. Doermann, A. Rosenfeld: Recovery of temporal information from static images of handwriting. International Journal of Computer Vision, Volume 15 , Issue 1-2 (June 1995), 143-164.

A. Elbaati, H. Boubaker, M. Kherallah, A. Ennaji, H. El Abed, A.M. Alimi: Arabic Handwriting Recognition Using Restored Stroke Chronology, In. Proc. 10th Int. Conf. Document Analysis and Recognition (ICDAR 2009), 2009, 411-415.

K. Franke, S.N. Srihari, Computational Forensics: An Overview. In: Proc. Second International Workshop on Computational Forensics, Washington D.C., 2008, Springer LNCS 5158, 1-10.

W. Guerfali, R. Plamondon: The Delta LogNormal theory for the generation and modeling of cursive characters. In: Proc. Third Int. Conf. on Document Analysis and Recognition (ICDAR'95) - Volume 1, 1995, 495-498

U. Koethe: Edge and Junction Detection with an Improved Structure Tensor, in: B. Michaelis, G. Krell (eds.): Pattern Recognition, Proc. of 25th DAGM Symposium, Magdeburg 2003, LNCS 2781, Springer, 2003, 25-32.

J.-P. Larmagnac, E. Dinet: A stroke extraction method for off-line recognition of Chinese characters. In: Proc. 7th Conf. Document Recognition and Retrieval, 2000, 32-41.

F. Lin and X. Tang: Off-line handwritten chinese character stroke extraction. In: Proc. 16th Int. Conf. on Pattern Recognition (ICPR02), Volume 3, IEEE Computer Society, 2002, 30249.

H. Meine, U. Koethe: Image Segmentation with the Exact Watershed Transform. In: J.J. Villanueva (Ed.): VIIP 2005, Proc. 5th IASTED International Conference on Visualization, Imaging, and Image Processing, pp. 400-405, ACTA Press, 2005

I. Moalla, F. Lebourgeois, H. Emptoz, A.M. Alimi: Image Analysis for Paleography Inspection. In: Proc. 2nd Conf. Document Image Analysis for Libraries (DIAL 06), 2006, 303-311.

E.-M. Nel, J.A. Preez, B.M. Herbst: A Pseudo-skeletonization Algorithm for Static Handwritten Scripts. Int. J. Document Analysis and Recognition (IJDAR) 12, 2009, 47-62

M. Richter: Tentative Criteria for Discerning Individual Hands in the Guodian Manuscript. In: W. Xing (ed.) Rethinking Confucianism: Selected Papers from the Third International Conference on Excavated Chinese Manuscripts, Mount Holyoke College (April 2004, San Antonio), Trinity University, 2006, 132-147

R. Seidel: Constrained Delaunay Triangulations and Voronoi Diagrams with Obstacles. Technical Report 260, Inst. for Information Processing, Graz, Austria, 1988.

A. Solth, B. Neumann, P. Stelldinger: Strichextraktion und -analyse handschriftlicher chinesischer Zeichen. Report FBI-HH-B-291/09, Department of Informatics, University of Hamburg, 2009.

P.A. Stokes: Computer-aided Palaeography, Present and Future. In: M. Rehbein et al., (eds.), Codicology and Palaeography in the Digital Age, Schriften des Instituts für Dokumentologie und Editorik - Band 2, Book on Demand GmbH, Norderstedt, 2009.

W.-H. Wu: Off-line Chinese Character Recognition Based on Stroke Features. Dissertation, National Central University Taiwan, 2000.

M. Wienecke: Videobasierte Handschrifterkennung. Dissertation, Techn. Fak. Univ. Bielefeld, Oktober 3003.

Zhuang, Y., Zhang, X., Wu, J., Lu, X.: Retrieval of Chinese calligraphic character image. In: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, 2004, Part I, 17–24.

J.J. Zou, H. Yan: Skeletonization of Ribbon-Like Shapes Based on Regularity and Singularity Analysis. IEEE Trans. Systems, Man, and Cybernetics - Part B: Cybernetics. Vol. 31, No. 3, 401-407, 2001