Informatik

Computerverfahren für den Handschriftvergleich von Manuskripten

Rainer Herzog, Arved Solth, Bernd Neumann

Ein zentrales Leitmotiv des Sonderforschungsbereich „Manuskriptkulturen“ ist die Untersuchung von visuellen Manuskriptmerkmalen: Welche Aufschlüsse geben Gestaltungsmerkmale wie Schrift- und Bildanordnung, Zeichenformen und Schreibercharakteristika über die Entstehung und das kulturelle Umfeld von Manuskripten? Zur Beantwortung dieser Fragen werden in zunehmendem Maße Computerverfahren herangezogen, die auf Entwicklungen der Digitalen Paläographie und dem umfangreichen Fundus von Bildverarbeitungsmethoden aus der Informatik basieren. Computerverfahren bieten zwei wichtige Vorteile gegenüber einer „händischen“ Beurteilung: Sie erlauben die Verwendung objektiver, explizit angebbarer Kriterien, und sie ermöglichen die Bearbeitung größerer Datenbestände, so dass Ergebnisse auf eine breitere statistische Basis gestellt werden können.

In diesem Beitrag wird das grundsätzliche Vorgehen bei einem computerbasierten Schreibervergleich von Manuskripten beschrieben. Stammen zwei Manuskripte (oder zwei Mengen von Manuskripten) aus der Hand desselben Schreibers? Diese Frage stellt sich bei historischen Manuskripten häufig, und ihre Beantwortung kann interessante Aufschlüsse über eine zeitliche und räumliche Einordnung geben.

Bild 1 zeigt als Beispiel drei von ca. 2000 Bambusleisten aus der Manuskriptsammlung der Yuelu Akademie in Changsha (Hunan). Obwohl die Provenienz der Manuskripte unklar ist, konnten sie auf das späte 3. Jh. v. Chr. datiert werden und stammen vermutlich aus einem Grab. Die ursprüngliche Bündelung der Bambusleisten ist zerstört, so dass eine Zuordnung rekonstruiert werden muss. Bild 2 zeigt als weiteres Beispiel ein Sanskrit-Manuskript mit Paratexten, bei dem die Identität von Schreibern der einzelnen Textblöcke von Bedeutung ist.

Informatics

Computer Methods for Comparing the Hands of Manuscripts

A central guiding theme of the collaborative reseach project ‘Manuscript Cultures’ is the analysis of visual manuscript features: What can be learned from design aspects, such as text and picture layout, character shapes and writing characteristics, about the origin and the cultural environment of a manuscript? To provide answers for such questions, researchers increasingly employ computer methods based on developments in Digital Palaeography and on the large pool of image processing methods developed in Computer Science. Computer methods offer two important advantages as compared to human judgement: They allow the use of objective, explicitly specified criteria, and they enable the processing of large data volumes, putting results on a broader statistical basis.

In this contribution we describe the basic approach for a computer-based comparison of the hands of manuscripts. Have two manuscripts (or two sets of manuscripts) been written by the same person? This question arises frequently in connection with historical manuscripts, and the answer may provide interesting clues about their temporal and geographical classification.

The first example (fig. 1) shows three of ca. 2000 bamboo strips from the collection of the Yuelu Academy in Changsha (Hunan). Although the provenence of the manuscripts is uncertain, they have been dated to the late 3rd century BC and were presumably unearthed in a tomb. The original binding has been broken, hence one has to reconstruct their correct order. Determing whether strips originate from the same hand may provide an important clue for this task. As a further example, Fig. 2 shows a Sanskrit manuscript with paratexts. Whether the main text and a paratext have been written by several hands may be of importance to manuscript researchers.

Bild 1: Bambusleisten aus der Sammlung der Yuelu Akademie, Changsha. (Aus: Zhu Hanmin 朱漢民 u. Chen Songchang 陳松長, eds. 2013. Yuelu shuyuan cang Qin jian (san) 岳麓書院藏秦簡（叁）. Shanghai cishu chubanshe.)

Fig. 1: Bamboo strips from the collection of the Yuelu Academy, Changsha. (In: Zhu Hanmin 朱漢民 u. Chen Songchang 陳松長, eds. 2013. Yuelu shuyuan cang Qin jian (san) 岳麓書院藏秦簡（叁）. Shanghai cishu chubanshe.)

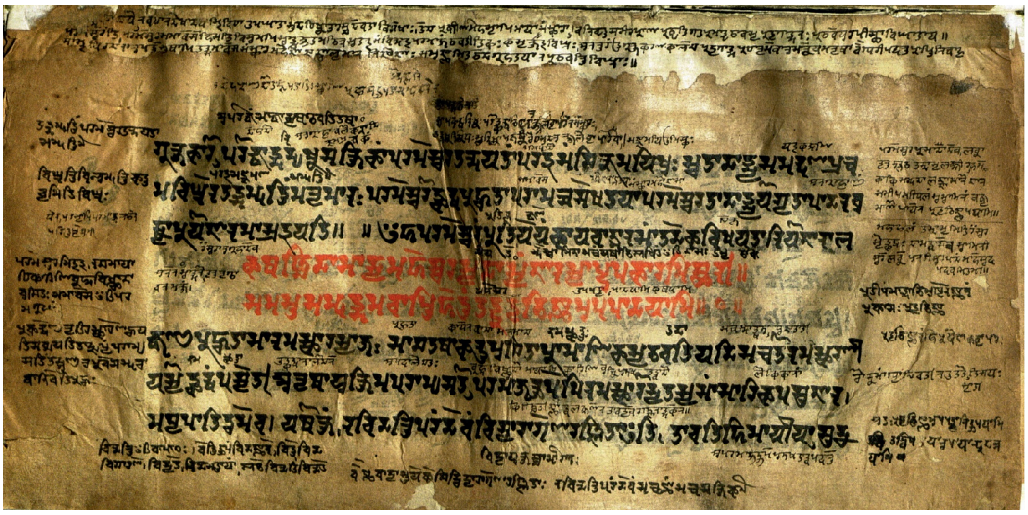


Bild 2: Srinagar Oriental Research Library (Indien), MS 1161, *Īśvarapratyabhijñāvimarśinī*

Fig. 2: Srinagar Oriental Research Library (India), MS 1161 *Īśvarapratyabhijñāvimarśinī*

Ein Computerverfahren zum Schreibervergleich für zwei Manuskripte M1 und M2 gliedert sich in folgende grundsätzliche Schritte:

A Bestimme Schriftzeichen oder Grapheme, die für einen Vergleich von M1 und M2 in Frage kommen.

B Bestimme die statistische Verteilung von Merkmalen an den ausgewählten Graphemen für M1 und M2 und überprüfe damit die Hypothesen von identischen bzw. verschiedenen Schreibern.

A computer method for comparing the hands of two manuscripts M1 and M2 is structured into two basic steps:

A Determine characters or graphemes suitable for a comparison of M1 and M2.

B Determine the statistics of features of the selected graphemes for M1 and M2 and use them to examine the hypotheses of identical or different hands.

Bei Schritt A kommt es im Wesentlichen darauf an, Zeichen oder Schriftzüge („Grapheme“) zu finden, die in beiden Manuskripten genügend häufig auftreten, um einen statistisch signifikanten Vergleich zu ermöglichen. Hier können Computerverfahren gegenüber manuellen Verfahren deutliche Effizienzgewinne bieten. Allerdings muss dazu das nichttriviale Problem gelöst werden, in einem größeren, in Gestalt von Bildern vorliegendem Datenbestand nach gleichartigen, aber unterschiedlich ausgeprägten Graphemen zu suchen. Wir skizzieren im Folgenden die dazu erforderlichen Teilschritte für chinesische Manuskripte, bei denen sich einzeln abgrenzbare Zeichen als Vergleichsobjekte eignen.

The main purpose of Step A is to find characters or writing patterns (‘graphemes’) which occur sufficiently often in both manuscripts and hence provide a base for a statistically significant comparison. For this task, computer methods may provide distinct efficiency gains when compared to a manual search. However, this requires solving a non-trivial problem: The computer program must find graphemes of equal kind but varying appearance in potentially large amounts of—often poor—pictorial data.
In the following, we sketch out the necessary steps for Step A using Chinese manuscripts as examples where characters can be delimited and hence provide convenient objects for comparison.

Segmentierung

Segmentierung dient der Abgrenzung einzelner Zeichen (i) voneinander und (ii) vom Hintergrund. Bild 3 illustriert den ersten Teilschritt am Beispiel eines in mehrere Spalten strukturierten chinesischen Manuskriptes. Zunächst wird das Bild in vertikaler Richtung mit einem Tiefpassfilter geglättet, dann werden die Spaltengrenzen als „Täler“ der Intensitäten des geglätteten Bildes mithilfe einer Wasserscheiden-

Segmentation

The segmentation process has the purpose of separating characters (i) from each other, and (ii) from the background. Fig. 3 illustrates the first task using a column-structured Chinese manuscript as an example. In a first step, the image is smoothed vertically with a low-pass filter. Then the ‘valleys’ of the smoothed image are determined as column boundaries using the watershed segmentation method (details are beyond the

Segmentierung bestimmt. Auf analoge Weise – nach horizontaler Glättung – können die zu einzelnen Zeichen gehörigen Felder abgegrenzt werden.

scope of this article). In an analogous way – after horizontal smoothing – the areas occupying single characters are separated from each other.

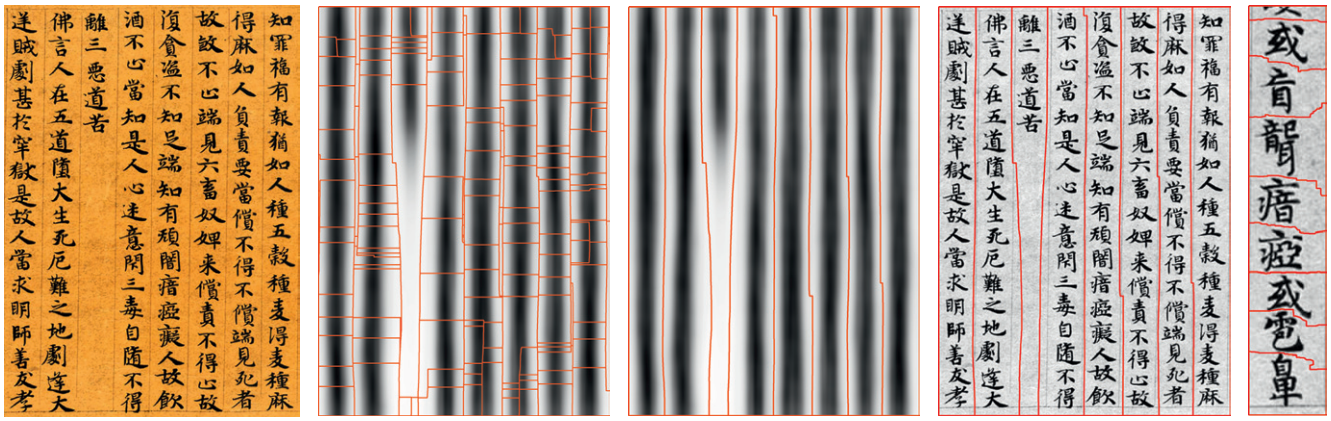


Bild 3: Zeilen- und Zeichensegmentierung durch Wasserscheiden-Segmentierung in vertikal bzw. horizontal geglätteten Bildern.
British Library Board, Or. 8210/S.2051

Fig. 3: Row and character segmentation using the watershed method in vertically or horizontally smoothed images, respectively.
British Library Board, Or. 8210/S.2051

Im zweiten Teilschritt wird die Kontur des Zeichens ermittelt. Dies kann bei schlecht erhaltenen Manuskripten (z.B. bei fleckigem Hintergrund oder beschädigter Schrift) aufwändige Verfahren erfordern.

The second segmentation task is to determine the contour of the character. This may require complex and expensive methods in cases of poorly preserved manuscripts (e.g. if the background is stained or the writing is damaged).

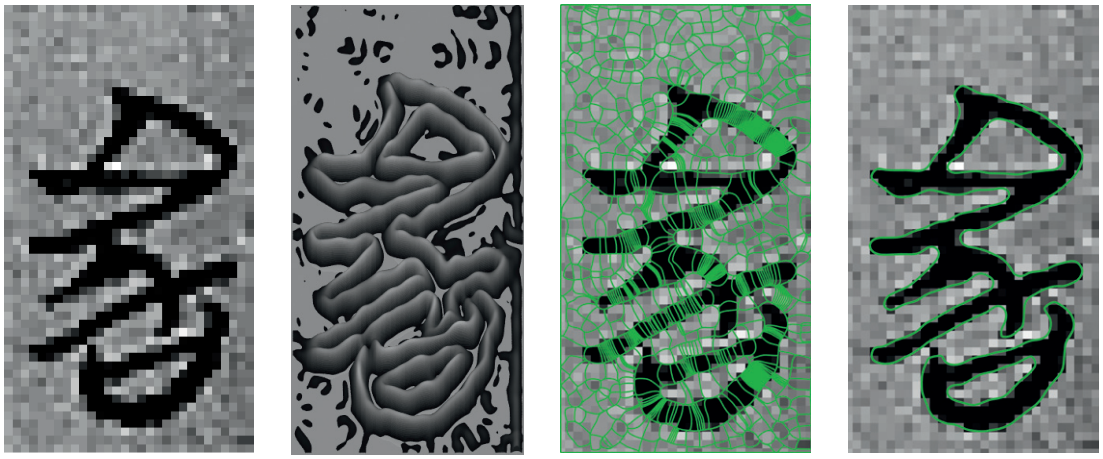


Bild 4: Segmentierung von Konturen mit Subpixelgenauigkeit

Fig. 4: Segmentation of contours with subpixel accuracy

Bild 4 illustriert einen Segmentierungsansatz, der speziell für grob aufgelöste Bilder entwickelt wurde, wie sie von historischen Manuskripten häufig nur zur Verfügung stehen. Dazu wird eine kontinuierliche Funktion (Spline-Funktion fünfter Ordnung) an das diskrete Raster der Bildintensitäten angepasst. Mit dem Wasserscheiden-Verfahren entsteht dann

Fig. 4 illustrates a segmentation procedure which has been specially developed for images with coarse resolution. Unfortunately, this is often the case with historical manuscripts. To obtain smooth boundaries, a continuous function (a 5th-order spline function) is fitted to the discrete image intensities. Using the watershed method and a grid with arbitrary

zunächst ein übersegmentiertes Bild in einer beliebig einstellbaren Subpixelrasterung. Durch sukzessives Verschmelzen von Teilbereichen mit geringem Kontrast gewinnt man dann die endgültigen Zeichenkonturen.

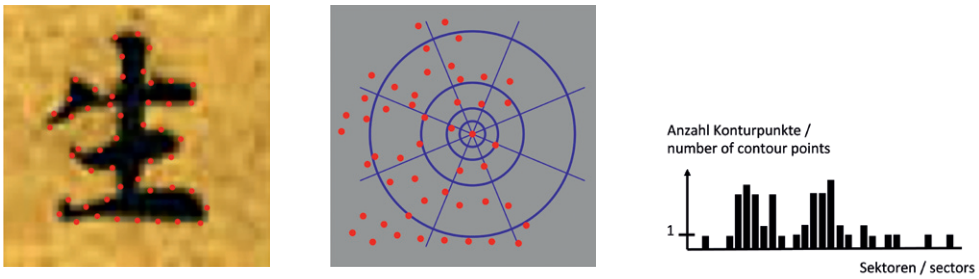
Gleichartige Zeichen finden

Zeichen, die für einen Vergleich in Frage kommen, werden in der Regel von Manuskriptforschern als Ergebnis einer händischen Voruntersuchung vorgeschlagen. Dem Rechner verbleibt die Aufgabe, alle Vorkommen eines vorgegebenen Zeichens in einem möglicherweise großen Datenbestand zu finden. Dabei sind Abweichungen gegenüber der Vorgabe bis zu einem gewissen Grade zu tolerieren. Zum Vergleich eines Zeichens mit der Vorgabe kommen verschiedene Verfahren in Frage, hier wird das Shape-Context-Verfahren verwendet. Dazu wird eine Zeichenkontur durch gleichmäßig verteilte Konturpunkte angenähert, illustriert in Bild 5a mit einer reduzierten Zahl von Punkten. Zu jedem Konturpunkt P wird ein Punktkontext in Gestalt eines Sektorhistogramms bestimmt (Bild 5c), das die Anzahl von Konturpunkten in Sektoren um P herum enthält (Bild 5b).

subpixel granularity, an oversegmented image is generated first and then refined to the final contour by merging subregions with weak contrast.

Finding characters of equal kind

Typically, manuscript researchers can propose characters suitable for a comparison based on the results of a preliminary manual investigation. The computer then has the remaining task to find all occurrences of a given character in a potentially large database, tolerating deviations from the given character up to a certain degree. This task is difficult because the right kind of tolerance is needed: Structural deviations should be less acceptable than deformations. There exist several methods for comparing two shapes, here we will present the Shape-context method. First, the contour of a character is approximated by equally spaced contour points, illustrated in Fig. 5a with a reduced number of points for the sake of clarity. For each point P one determines a point context in terms of a sector histogram (Fig. 5c), which specifies the number of other contour points in sectors around P (Fig. 5b).



a b c

Bilder 5a-c: Im Shape-Context-Verfahren wird ein Konturverlauf durch Sektorhistogramme für alle Konturpunkte beschrieben (s. Text).

Figs. 5a-c: Using the Shape-Context method, a contour is described by sector histograms for all contour points (see text).

Die Punktkontexte eines Zeichens stellen eine angenäherte Formbeschreibung dar und erlauben einen Formvergleich von zwei Zeichen. Dabei werden Punktkontexte zunächst optimal in Korrespondenz gebracht und dann im Hinblick auf ihre Übereinstimmung bewertet. Der Rechner findet also eine Menge von Zeichen, die dem Grad ihrer Übereinstimmung mit der Vorgabe nach geordnet werden können. Bild 6 zeigt das vorläufige Ergebnis einer Suche nach gleichartigen Zeichen für die markierte Vorgabe. Mithilfe eines Schwellenwertes und der Endkontrolle durch einen Manuskriptforscher werden daraus endgültig akzeptierte Zeichen ausgewählt.

The point contexts of a character provide an approximate shape description and allow the shape comparison of two characters. To this end, the best correspondence of point contexts is determined first, and then all pairs of corresponding point contexts are evaluated regarding agreement. This way, the computer determines matching characters which can be ordered by their degree of agreement. Fig. 6 shows the preliminary result of a search for characters matching a given character marked in red. By means of a threshold and the final control by a manuscript researcher, the final set of accepted characters will be determined.

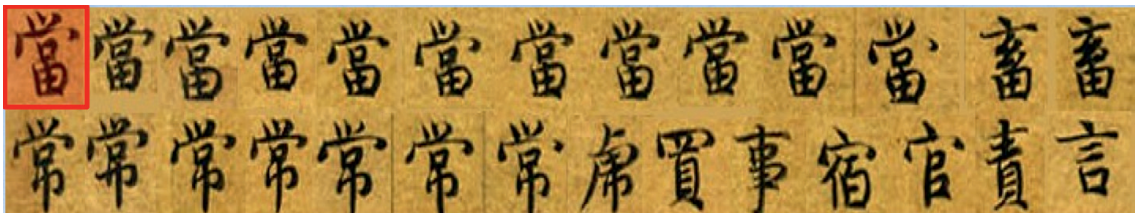


Bild 6: Durch Formvergleich mithilfe des Shape-Context-Verfahrens werden die zu einer Vorgabe ähnlichen Zeichen in Abbildern von Manuskripten ermittelt.

Fig. 6: By performing shape comparison with the Shape-Context method, characters matching a given character are determined in manuscript images.

In Schritt B gilt es nun zu prüfen, ob sich Manuskript M1 anhand von Merkmalen einer in den vorhergehenden Schritten ausgewählten Zeichenmenge von Manuskript M2 unterscheiden lässt. Bei chinesischen Zeichen und zahlreichen weiteren Schriftsystemen kann die Merkmalsgewinnung durch eine genaue Beschreibung der Strichstruktur eines Zeichens wesentlich unterstützt werden. Im Folgenden wird ein Verfahren zur Strichextraktion vorgestellt, das für diesen Zweck entwickelt wurde.

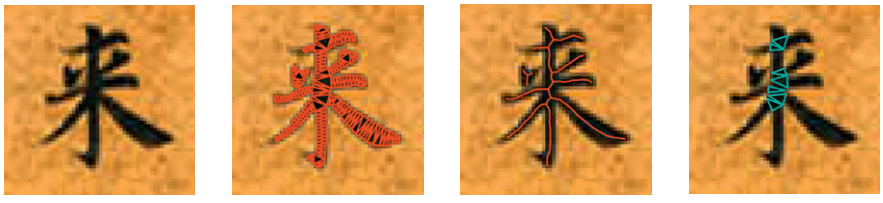
After completion of Step A, it is the goal of Step B to find out whether Manuscript M1 can be distinguished from Manuscript M2 based on shape features of the characters selected in the previous steps. For Chinese characters and several other writing systems, obtaining interesting features from characters can be significantly facilitated by first determining the exact stroke structure of a character. In the following, we present a method for stroke extraction which has been developed for this purpose.

Strichextraktion

Die einzelnen Striche eines Zeichens können nach erfolgter Segmentierung mithilfe eines speziellen Triangulierungsverfahrens (Constrained Delaunay Triangulation) ermittelt werden. Dabei wird die Zeichenkontur durch diskrete Konturpunkte angenähert, die dann innerhalb des Zeichens zu kleinstmöglichen, nicht-überlappenden Dreiecken verbunden werden. Hieraus können das Skelett des Zeichens sowie Verzweigungs- und Kreuzungspunkte gewonnen werden (Bild 7).

Stroke extraction

Individual strokes of a character can be determined after segmentation using a special triangulation procedure (Constrained Delaunay Triangulation). The contour of a character is approximated by discrete contour points which are then connected to form smallest possible, non-overlapping triangles. Triangulation procedures are well known for various applications of Computer Graphics, and triangulations can be computed efficiently. From the triangles, one can determine the skeleton of a character as well as branching and crossing points (Fig. 7).



a b c d

Bild 7: Aus der Triangulierung (b) einer Zeichenkontur (a) lassen sich das Skelett (c) sowie Junction-Dreiecke (d) an Verzweigungen und Kreuzungen ermitteln.

Fig. 7: From a triangulation (b) of a character contour (a) one can obtain the skeleton (c) as well as junction triangles (d) at branching and crossing points.

Die dadurch definierten Teilstriche werden dann in einem weiteren Verarbeitungsschritt zu endgültigen Strichen verschmolzen. Bild 8 zeigt das Ergebnis für ein chinesisches Zeichen.

This way, partial strokes are defined and they are merged into final strokes in a further processing step. Fig. 8 shows results for a Chinese character.

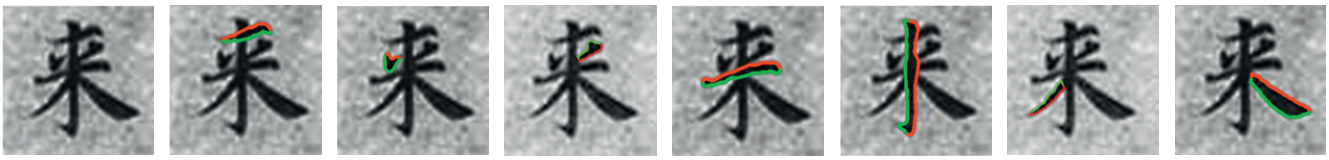


Bild 8: Extrahierte Striche als Basis für eine Merkmalsbestimmung. Die Farben Rot und Grün kodieren die linke bzw. rechte Seite eines Striches in Schreibrichtung.

Fig. 8: Strokes extracted for feature determination. The left and right sides of strokes, viewed in writing direction, are marked in red and green, respectively.

Merkmalsbestimmung

Das Repertoire möglicher Merkmale reicht von augenfälligen Charakteristika, wie etwa Strichlängen und Winkel zwischen Strichorientierungen, bis hin zu nicht so offensichtlichen Maßen, etwa Häufigkeitsverteilungen von Konturgradienten oder Konturkrümmungen. Die Eignung eines Merkmals hängt davon ab, ob seine Werte insgesamt eine deutliche Varianz haben. An dieser Stelle sind Erfahrungen und Beobachtungen von Manuskriptforschern hilfreich, die auf geeignete Merkmale aufmerksam machen können. Merkmalsbestimmung wird in Bild 9 beispielhaft für Winkel zwischen den Strichen chinesischer Zeichen illustriert. Dazu werden rechnerisch Geraden an die Mittellinien der Striche angepasst und Winkel zwischen Geradenpaaren bestimmt, in Bild 9 für Proben von zwei Schreibern illustriert.

Determining features

There exists a large repertoire of possible features, ranging from obvious characteristics, such as stroke lengths or angles between strokes, to less conspicuous measurements, for example histograms of contour gradients or contour curvatures. Whether or not a feature is suitable for scribe comparison, depends on its variability across the manuscripts under examination. For feature selection, the experiences and observations of manuscript researchers are often helpful. In Fig. 9, feature determination is illustrated for the angles between strokes of Chinese characters. Using established computer procedures, straight lines are fitted to the center lines of strokes, then angles between the straigth lines are determined. The figure shows samples of two different scribes (top line and bottom line, respectively).

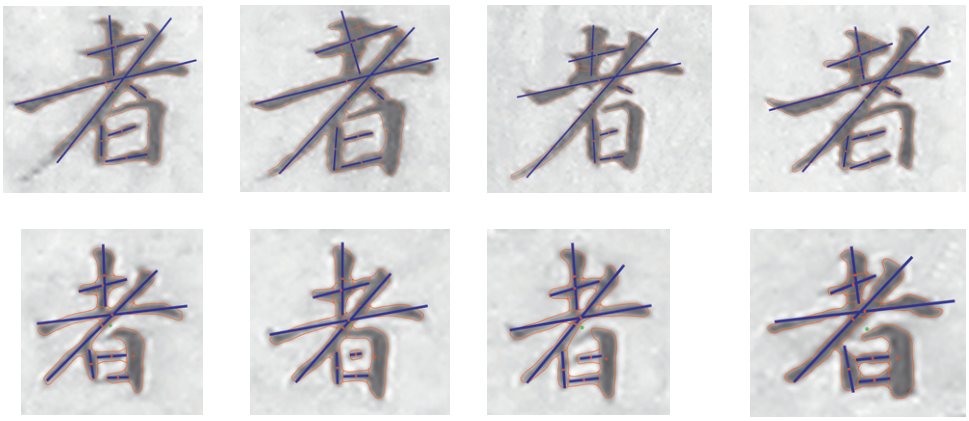


Bild 9: Proben von Schreiber 1 (oben) und Schreiber 2 (unten) mit rechnerisch an die Striche angepassten Geraden. Winkel zwischen Geraden können als Merkmale verwendet werden.

Fig. 9: Samples of Scribe 1 (above) and Scribe 2 (below) with straight lines fitted to the strokes by computer procedures. Angles between the straight lines can be used as features.

Hypothesentest

Aufgrund der variierenden Ausprägungen von Merkmalen in den Manuskripten M1 und M2 kann nun geprüft werden, ob beide Manuskripte von demselben Schreiber (Hypothese H_0) oder von verschiedenen Schreibern (Hypothese H_1) stammen. Intuitiv würde man H_0 den Vorzug geben, wenn die Verteilungen

Hypothesis test

Based on the varying values of features determined for Manuscripts 1 and 2, one can now examine whether both manuscripts originated from the same scribe (Hypothesis H_0) or from different scribes (Hypothesis H_1). Intuitively, one would prefer H_0 if the distributions of feature values for M1

von Merkmalswerten in M1 und M2 weitgehend übereinstimmen. Mit Methoden der Statistischen Entscheidungstheorie können Beurteilungen dieser Art fundiert erfolgen.

Eine wichtige Rolle spielt dabei der Chi-Quadrat-Test, mit dem die Konfidenz berechnet werden kann, ob die Merkmalswerte tatsächlich einer gemeinsamen Verteilung genügen. Man weiß dann also, mit welcher Wahrscheinlichkeit die Annahme gleicher Schreiber korrekt ist.

Bild 10 illustriert den Hypothesentest, wobei die Länge des oberen senkrechten Striches als Merkmal verwendet wird (s. Bild 9). Es wurden drei Gauß-Verteilungen angepasst, an die Merkmale der einzelnen Schreiber sowie an alle Merkmale zusammengenommen. Man sieht, dass sich die Mittelwerte der Strichlängen von Schreiber 1 und 2 deutlich unterscheiden. Der Eindruck verschiedener Verteilungen der beiden Schreiber wird durch einen Likelihood-Test belegt: Die Wahrscheinlichkeit, die Merkmale unter der Hypothese H_1 zu beobachten, ist etwa 20 mal größer als unter der Hypothese H_0.

and M2 are approximately the same. Judgements of this kind can be placed on formal grounds with the help of Statistical Decision Theory.

The Chi-square test plays an important part here. With this test one can compute the confidence in the proposition that feature values obey the same distribution. In consequence, one knows how probable it is that the assumption of identical scribes is indeed correct.

Hypothesis testing is illustrated in Fig. 10, using the length of the top vertical stroke as prime feature (see Fig. 9). Three Gaussian distributions have been fitted, to the feature values of the two scribes individually and to all feature values taken together. One can see that the means of the stroke lengths of the two scribes differ significantly. The impression of different distributions is confirmed by a likelihood test: The probability of observing the feature values of the samples given that H_1 is true, is 20 times greater than the probability of observing the feature values given that H_0 is true.

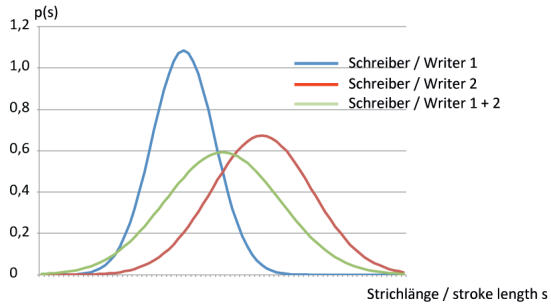


Bild 10: Gauß-Verteilungen, angepasst an die Häufigkeitsverteilungen der Strichlängen des mittleren Querstriches von Schreiber 1, Schreiber 2 und beiden Schreibern zusammen.

Fig. 10: Gaussian distributions fitted to the histograms of the stroke lengths of the central horizontal strokes for Scribe 1, Scribe 2 and both scribes combined.

Zusammenfassung

Wir haben in diesem Beitrag das grundsätzliche Vorgehen bei einem rechnergestützten Handschriftvergleich von Manuskripten beschrieben. Für jede der vorgestellten Methoden sind in der wissenschaftlichen Literatur auch Alternativen vorgeschlagen worden, die je nach Eigenschaften der Manuskripte und der verwendeten Schriftsystems Vorteile bieten können. Auch muss betont werden, dass ein Urteil über die Identität von Schreibern natürlich in der Regel an einer größeren Zahl von visuellen Merkmalen festgemacht werden muss. Zusätzlich müssen auch alle ergänzenden Informationen berücksichtigt werden, die Manuskriptforscher beibringen können, insbesondere aufgrund einer inhaltlichen Auswertung der Manuskripte. Die sich daraus ergebende Tendenz kann als a priori Wahrscheinlichkeit elegant mit dem oben beschriebenen Hypothesentest verbunden werden.

Summary

In this contribution, we have described the basic steps for comparing the hands of manuscripts using computer methods. For each of the methods presented above one can find alternatives in the scientific literature which may offer advantages depending on manuscript properties and the writing system in question. We want to emphasize that a judgement about the identity of scribes must be based on more than a single visual feature, of course. Furthermore, all complementary information must be considered which can be procured by manuscript researchers, in particular from an evaluation of the manuscript contents. The propensities resulting from such information can be elegantly introduced as a prior probability into the hypothesis test described above.