

文書筆記過程の分析に関わる墨跡の濃淡変化箇所推定手法の性能評価

ゴー チュイリン¹ 中尾 泰士¹

¹ 北九州市立大学

{goh, nakaoy}@kitakyu-u.ac.jp

概要

墨によって書かれた文書は、墨の濃淡情報を用いることで、その文書が作成された筆記過程を分析できる可能性がある。われわれは、墨の濃淡情報を時系列データとして分析し、墨が徐々に薄くなって、再び濃くなった箇所を筆に墨を含めなおした点として推定する手法を考えた。この手法による分析結果を人の眼で判断した正解と比較検討することで、われわれの手法の性能評価を行なった。

1 はじめに

墨と筆を利用して作成された文書（もんじょ）においては、筆に含まれる墨の量によって墨跡に濃淡変化が見られる。われわれは、文書や書簡等の分析において従来行われてきた、主として記載内容を読み取る研究とは異なり、墨跡の画像的分析により、書き手の文書作成プロセスを明らかにすることを目指している。

毛筆で書かれた文字を情報学的視点から分析した研究としては、例えば、くずし字を文字認識する試み [1] や、古文書デジタルアーカイブから文字検索を行うための画像処理 [2]、文字認識を経ずに画像処理によってのみ文字列検索を行うワードスポットティング研究 [3], [4] などが挙げられる。また、毛筆習字の自動添削システムに向けた毛筆文字の特徴量算出に関する研究 [5] など関連研究として挙げられるだろう。

しかし、われわれのような視点での先行研究は見当たらず、本研究は文書研究の分野において新たな視点による文理融合の萌芽的な役割を果たすと考えている。

通常、書き始めにおいては墨をたっぷり含ませた筆で文字を書き出し、筆に含まれる墨の量の減少につれ、書かれる文字がかすれていく。そのため、

書き手は文字が読みにくくなる前に筆に墨を含めなおすというプロセスを経るのが一般である。この場合、毎回筆に含まれる墨の量が一定ならば、同じ墨の消費量（≒文字量）で規則的に墨の含みなおしが発生することが予想される。

しかし、たとえば、記述の途中で次に書き継ぐべき文言をすぐには思い浮かばず、手が止まって時間が経過した場合、書き手は書き継ぐ前に改めて筆に墨を含めなおすという動作をする可能性がある。また、この部分が重要だと書き手が考える箇所を書くにあたって、（墨の量からは必ずしも必要ではないのに）筆に墨を含めなおす場合もあるだろう。

このように、墨の濃淡情報の分析は、それを書いた書き手の筆記過程における意識の流れをあらわしている可能性がある。

われわれは、画像として手に入る墨跡文書や書簡等を以上のような観点から分析する手法を開発することを通じ、その成果を本来の専門分野（古文書解析、書簡の往来を通じた政治史や文壇史など）の研究者に提供できるようなフレームワークを構築する試みを行っている。本論では、その試みについて現時点での性能評価を行う。

2 解析の流れ

2.1 準備

毛筆で縦書きの文書画像を入手する。ここでは主に早稲田大学図書館の古典籍総合データベース [6] を利用した。

まず、いくつかの前処理を行う。例えば、一つの文書が複数の画像に分割されている場合はそれらを結合する。また、額装などの不要な部分は取り除くなどの処理を行う。さらに、画像の解像度はまちなため、適当な画像サイズに縮小する。画像の高さを H としたとき、 $H < 1000$ ピクセルになるよう

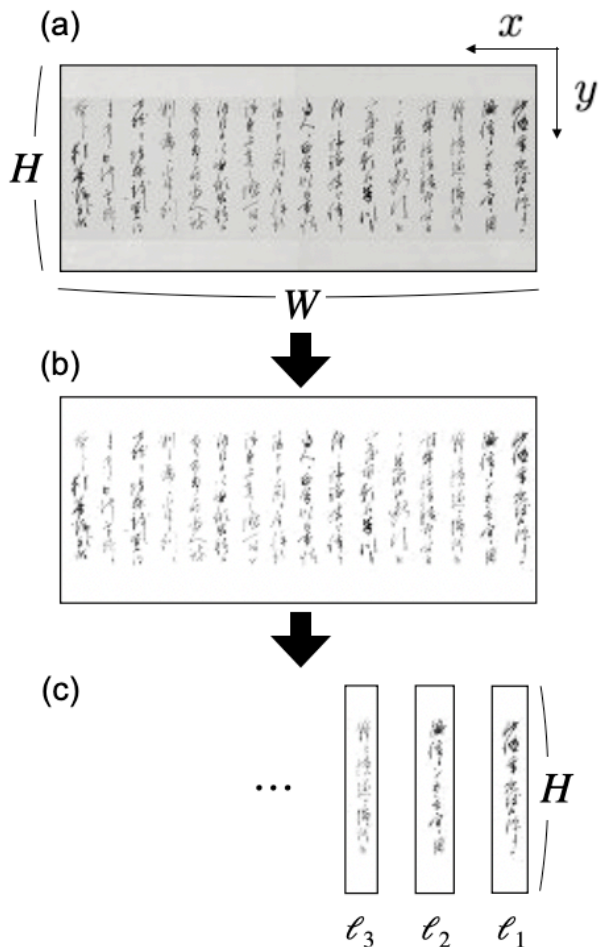


図1 (a) 毛筆で書かれた文書画像を (b) 前処理したあと、(c) 1行ずつに分割する。

に縮小した。

さらに、グレースケール画像に変換する (図1(a))。この時の画像サイズを縦 H 、幅 W ピクセルとしたとき、 $H \times W$ ピクセルの色値はそれぞれ $0 \sim 255$ の値をとるが、大津の二値化 [7] を利用して、画像のコントラストを上げ、文字情報の部分を保持したまま背景を均一化することを試みる (図1(b))。

縦書き文書は、いくつかの「行」からなるが、文書画像を「行」に分割する処理を行う。

われわれは墨の濃さに興味があるため、これ以降は、各点のピクセル値を 255 から引いた値をその点のピクセル値として用いることにする。すなわち、元画像のピクセルが白 (255) の場合は 0 、黒 (0) の場合は 255 が新たなピクセル値となる。

その上で、処理をする文書画像のピクセル値を $p(x, y)$ 、 $0 \leq x \leq L$ 、 $0 \leq y \leq H$ とする。文書は縦書きで右側から書き始められるという点に留意し、通常画像座標とは異なり、右端を $x = 0$ 、左端を $x = L$

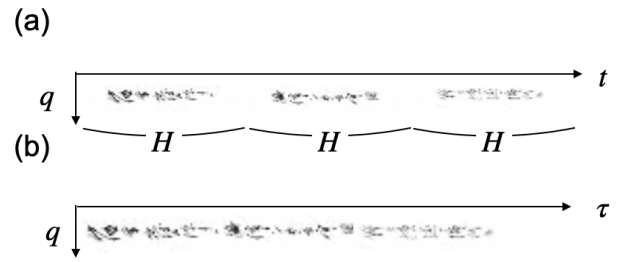


図2 (a) 行に分割した短冊状の画像を縦に連結し、(b) 文字のない余白部分を切り詰めて筆記過程の分析用の画像を作成する。

とすることに注意。また、このピクセル値を y 方向に積分した値

$$P(x) = \sum_{y=0}^H p(x, y) \quad (1)$$

を考えると、 $P(x)$ のグラフは文書の各行に対応した波パターンを描く (谷が行間、山が行の中心付近を表す)。ある x の値を中心として、ある程度の範囲 $x - \varepsilon \leq x \leq x + \varepsilon$ で $P(x) = 0$ となるようなところが行の分割位置となる。そうして、短冊状の画像に分割する。 n 番目の短冊の幅を l_n ($n = 1, 2, \dots, N$) とする (図1(c))。

この短冊状の画像を順番に縦に結合し、細長い 1 本の画像を作成する。このとき、 l_n は一般に同じ値にならないが、この後の処理の都合上、 $l \equiv \max \{l_n | n = 1, 2, \dots, N\}$ にそろえ、これに足りない幅の短冊には「白」色ピクセルを必要に応じて補完する。

いま、 $l \times NH$ のサイズの細長い画像が得られたが、これが、筆記された時間経過をあらわす文字情報となる¹⁾。この画像のピクセル値を $p'(q, t)$ としよう。ここで、 $0 \leq q \leq l$ 、 $0 \leq t \leq NH$ である (図2(a))。

この画像中には、文字が書かれていない余白が存在するため、この余白を切り詰める作業を行う。そのために、

$$P'(t) = \sum_{q=0}^l p'(q, t) \quad (2)$$

を計算し、 $P'(t) = 0$ となるような部分を削除して画像を切り詰めていく。切り詰めた後の画像に対して、改めて τ の時間軸を割り当てる (図2(b))。

1) まれに行間に追加の文字が書かれるような場合があるが、ここでは無視しておく。

2.2 墨を含めなおした箇所の推定

これで、文字が書かれている部分のみを切り出した画像が用意できた。あらためて、画像のピクセル値を $p'(q, \tau)$, $0 \leq q \leq \ell$, $0 \leq \tau \leq \tau_{\max}$ とする。ここで、 τ_{\max} は切り詰めた画像の末尾をあらわす。そして、

$$D(\tau) = \frac{1}{N(\tau)} \sum_{q=0}^{\ell} p'(q, \tau) \quad (3)$$

を計算する。ここで、 $N(\tau)$ は τ の点において、0 (白) ではない値を持つピクセル数をあらわす。この値は τ における平均の文字の濃さを表す量である。

この $D(\tau)$ の値は、墨の濃淡に応じて一般に大小の増減を (不規則に) くりかえす数値データになる (図 3)。

このデータ・パターンから、どこで墨を含めなおされているか (小→大となる点) を抽出する。そのために 2 つの異なる期間をもつ移動平均を用いる。短期移動平均を $M_S(\tau)$, それよりも長期の移動平均を $M_L(\tau)$ とし、

$$M_S(\tau) \equiv \frac{1}{\Delta\tau_S} \sum_{\tau'=\tau}^{\tau+\Delta\tau_S} D(\tau'), \quad (4)$$

$$M_L(\tau) \equiv \frac{1}{\Delta\tau_L} \sum_{\tau'=\tau}^{\tau+\Delta\tau_L} D(\tau') \quad (5)$$

によってそれぞれ定義する。それぞれを計算し、長期の移動平均線 $M_L(\tau)$ を短期の移動平均線 $M_S(\tau)$ が下から上に抜けるように交差したときを墨を含めなおされた時点と推測することにする (図 3)。

この交差点は、株式市場などにおける時系列データにおいて「ゴールデン・クロス」(以下、GC) と呼ばれるもので、相場のトレンドが上昇に移る転換点として知られているものである。

本論では、短期移動平均の計算を $\Delta\tau_S = 50$, 長期移動平均を $\Delta\tau_L = 400$ とした。これは、分析対象にした元画像の 1 行の高さがおよそ 1000 ピクセル程度になるようにしたことによる。また、GC を考えるにあたっては、各移動平均線を Savitzky-Golay filter[8] を使ってスムージングしたものを使った。そのほか、GC の誤検知を避ける目的で、短期移動平均線と長期移動平均線が一定程度以上離れた状態

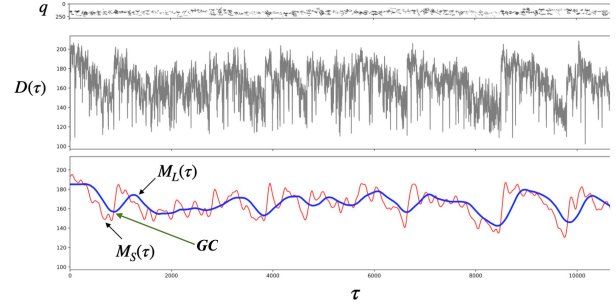


図 3 $D(\tau)$ の例. その下に、対応する短期移動平均 $M_S(\tau)$ (細線) と長期移動平均 $M_L(\tau)$ (太線), GC として判断した例を示す. この例は「山県有朋書簡: 桂太郎宛」(請求記号: チ 06 03917 0022 000) を用いた.

から GC を迎えた場合のみをカウントし、それ以外の場合は、GC としてはカウントしないようにした。

このようにして得られた GC の時点を文書の書き手が墨を含めなおした時点と仮定して、 $\tau \rightarrow t \rightarrow (x, y)$ と逆変換して、元の文書画像における位置を同定する。図 4 に分析事例を示す。

3 性能評価

われわれが用いた手法がどの程度正しい判別を行なったのか、その性能を評価する。12 通の文書についての分析結果の一覧を表 1 に示す。各文書について、人の目で判断した墨の含めなおし箇所の個数 (表 1 中では「正解箇所」と、われわれの手法が判断した墨の含めなおし箇所の個数 (表 1 中では「予測箇所」), 両者が一致する箇所²⁾の個数を「正解数」の列に示す。また、それらから求めた

$$\text{正解率} \equiv \frac{\text{正解数}}{\text{予測箇所}}, \quad (6)$$

$$\text{再現率} \equiv \frac{\text{正解数}}{\text{正解箇所}}, \quad (7)$$

についてもそれぞれ表 1 に示す。

表 1 を見れば、正解率はおしなべて高い (0.75 以上) ことがわかる。「山縣 1」の正解率は他と比べて極端に低い値となったが、これは元の文書自体が墨の濃淡変化が極めて少ないものであることが原因ではないかと推察している。再現率についても概ね高い (0.75 以上) が、「伊藤 1」「伊藤 6」で低い値を示す。「伊藤 1」の文書は長く、濃淡の変化がはっきりしない箇所が多くある特徴がある。また、「伊藤 6」については、こまめに墨を含めなおしているよ

2) 1 文字程度のずれは正解と判断している。

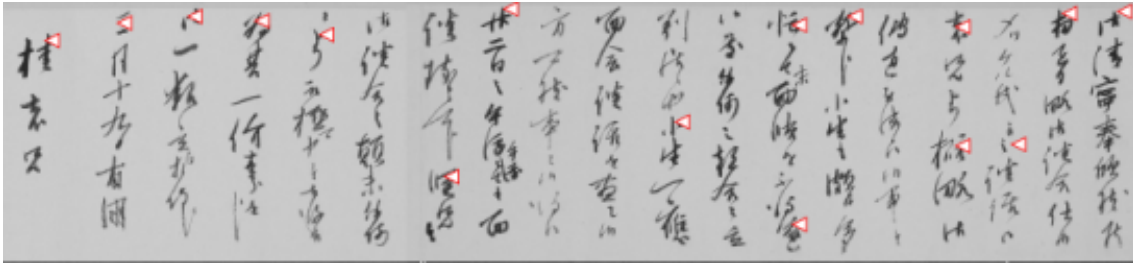


図4 墨を含めなおしたと判定した箇所には◀マークをつけた。

表1 12通の文書を用いた性能評価. 人の目で判断した墨の含めなおし箇所と本論の手法による判断の比較表. 表中の「文字数」は文書の長さを表すための概数であり, 正確な文字数ではない. 「伊藤 n 」($n = 1, 2, \dots, 6$)が伊藤博文によるもの, 「山縣 n 」($n = 1, 2, \dots, 6$)が山縣有朋によるものである. それぞれの文書の出典情報については, 「国」が国立国会図書館永続的識別子, 「早」が早稲田大学古典籍データベース請求記号である.

文書ラベル	出典情報	文字数	正解箇所	予測箇所	正解数	正解率	再現率	備考
伊藤1	国 info:ndljp/pid/9973821	2120	132	92	82	0.89	0.62	再現率低い
伊藤2	早 チ 06 03917 0002 0002	912	66	67	56	0.84	0.85	
伊藤3	早 チ 06 03917 0003 0003	198	24	22	20	0.91	0.83	
伊藤4	早 チ 06 03917 0003 0004	154	13	14	12	0.86	0.92	
伊藤5	早 チ 06 03917 0002 0003	175	21	17	17	1.00	0.81	
伊藤6	早 チ 06 03890 0150 0001	120	19	14	13	0.93	0.68	再現率低い
山縣1	国 info:ndljp/pid/9979758	760	53	72	46	0.64	0.87	正解率低い
山縣2	早 チ 06 03890 0112 0002	266	22	24	21	0.88	0.96	
山縣3	早 チ 06 03917 0018 0007	368	25	25	19	0.76	0.76	
山縣4	早 チ 06 03917 0021 0004	948	125	99	94	0.95	0.75	
山縣5	早 チ 06 03917 0022 0001	133	14	16	13	0.81	0.93	
山縣6	早 チ 06 03917 0030 0004	368	20	18	16	0.89	0.80	

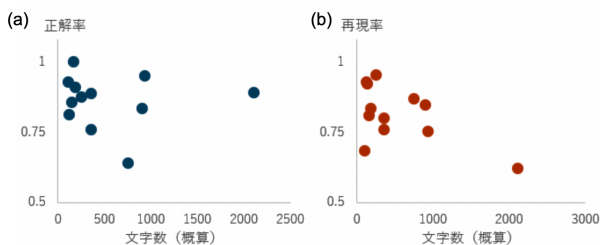


図5 文書の長さとの(a)正解率, (b)再現率の相関.

うで, 移動平均をとると濃淡の変化がはっきりしなくなるという特徴がある.

正解率と再現率について, 文書の長さとの関係を見てみる. 図5は, 各文書のおおよその文字数(表1中の「文字数」)に対する, それぞれ(a)正解率と(b)再現率の散布図である. 相関係数を計算すると, 文字数と正解率の相関係数は -0.07 , 文字数と再現率の相関係数は -0.58 となる. このことから, 文書の長さとの正解率には相関が見られないが, 文書の長さとの再現率の間には負の相関があることがわかる.

4 おわりに

本論では, 墨で書かれた文書の筆記過程の分析に関して, 墨跡の濃淡が変化する箇所を推定する手法

について簡易的な性能評価を行なった. その結果をみると, 正解率は比較的高い数値を示すが, 元画像の性質(濃淡変化が少ないなど)によっては低い数値を示した. また, 再現率は, 文書の長さ(文字数の多さ)に対して負の相関を示す結果が得られた.

本論で用いた, 短期と長期の移動平均の区間(期間)は, それぞれ50と400に固定しているが, 分析する元画像の性質によって異なる区間(期間)にした方が精度が高くなる可能性がある.

われわれの一連の処理の流れにおいて, たとえば文書画像を行へ分割する部分について, 現状では行間にはっきりと縦の空白がある場合のみ認識できる. もし, 行が傾いていたり, 2つの行が少し重なったりしている場合には行の分割ができないため, この点についても検討が必要である.

また, 筆に墨が含まれおされる点(GC)の検出において, 誤検知を避けるため, 長期平均線と短期平均線が一定程度以上離れていることを条件としたが, このパラメータも試行錯誤的に決めたものである. このパラメータについて最適化できる方法を考えたい.

謝辞

本研究については、2024年度公立大学法人北九州市立大学特別研究推進費の助成を受けている。

参考文献

- [1] 北本朝展, カラーヌワットタリン, ボーバー・イリザミケル. Kaggle くずし字認識 —世界規模の人文系コンペ開催への挑戦—. 人工知能, Vol. 35, No. 3, pp. 366–376, 2020.
- [2] 白井啓一郎, 末代誠仁, 井上聡, 久留島典子, 馬場基, 渡辺晃宏, 中川正樹. 古文書字形検索のための画像処理. 研究報告人文科学とコンピュータ (CH), Vol. 2013, No. 7, pp. 1–6, 2013.
- [3] 寺沢憲吾, 長崎健, 川嶋稔夫. 古文書画像を対象にしたワードスポッティング. 画像の認識・理解シンポジウム (MIRU)2005 講演論文集, pp. 522–529, 2005.
- [4] 寺沢憲吾, 長崎健, 川嶋稔夫. 固有空間法と dtw による古文書ワードスポッティング. 電子情報通信学会論文誌, Vol. J89-D, No. 8, pp. 1829–1839, 2006.
- [5] 松尾賢一, 土本良樹. 毛筆文字の特徴量算出ツールの試作. 奈良工業高等専門学校研究紀要, Vol. 43, pp. 39–44, 2007.
- [6] 早稲田大学図書館. 古典籍総合データベース, (2024-12 閲覧). <https://www.wul.waseda.ac.jp/kotenseki/index.html>.
- [7] Nobuyuki Otsu. A threshold selection method from gray-level histograms. **IEEE Transactions on Systems, Man, and Cybernetics**, Vol. 9, No. 1, pp. 62–66, 1979.
- [8] A. Savitzky and M. Golay. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, Vol. 36, No. 8, pp. 1627–1639, 1964.