

計算幾何学を用いた 毛筆文字画像からの ストローク抽出

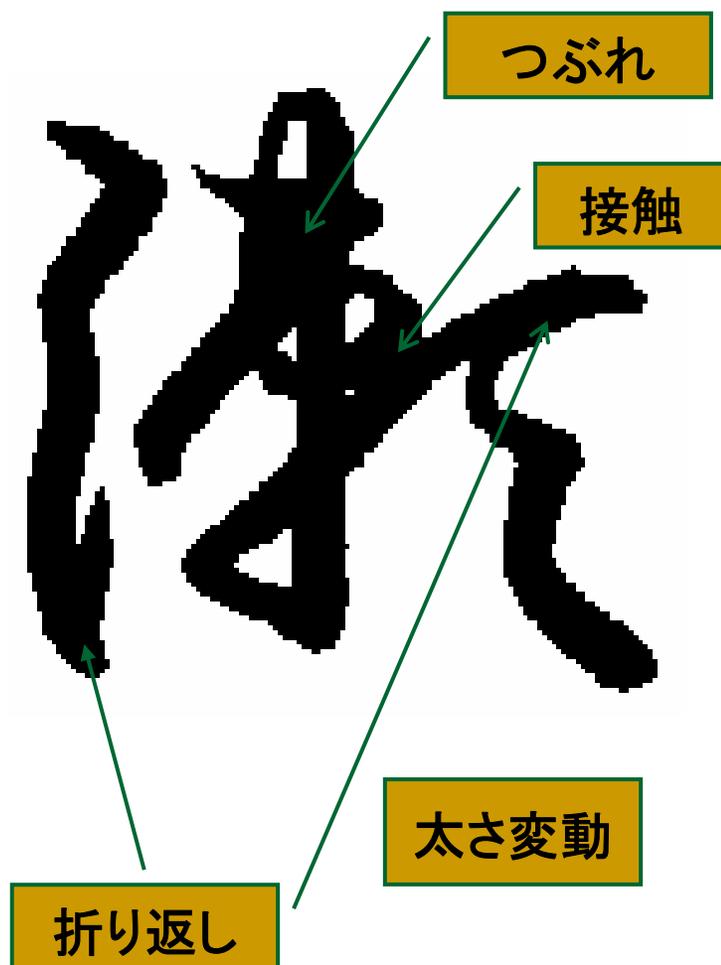
大阪市立大学創造都市研究科 安倍 広多
京都大学東南アジア研究センター 柴山 守

研究の目的

- 毛筆文字画像からのストローク抽出
 - 応用例
 - 毛筆文字のオフライン文字認識
 - (オンライン文字認識のための) ストローク辞書データ自動生成
 - フォントデータ自動生成
 - 有効な手法が存在していない
- 
- 毛筆文字画像からストロークを抽出する新しい方法を提案・実装・(予備的)評価

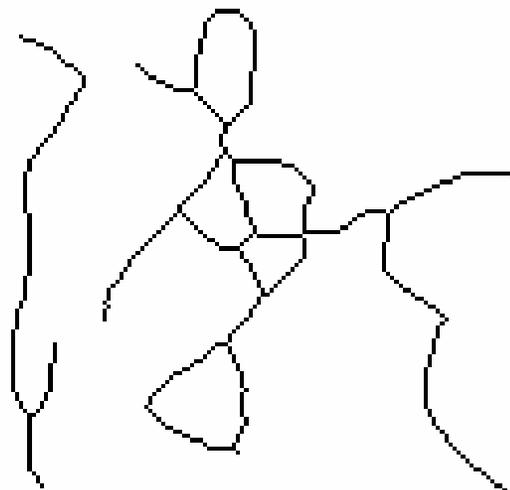
淋
浦
激

毛筆文字からのストローク抽出が 困難な理由



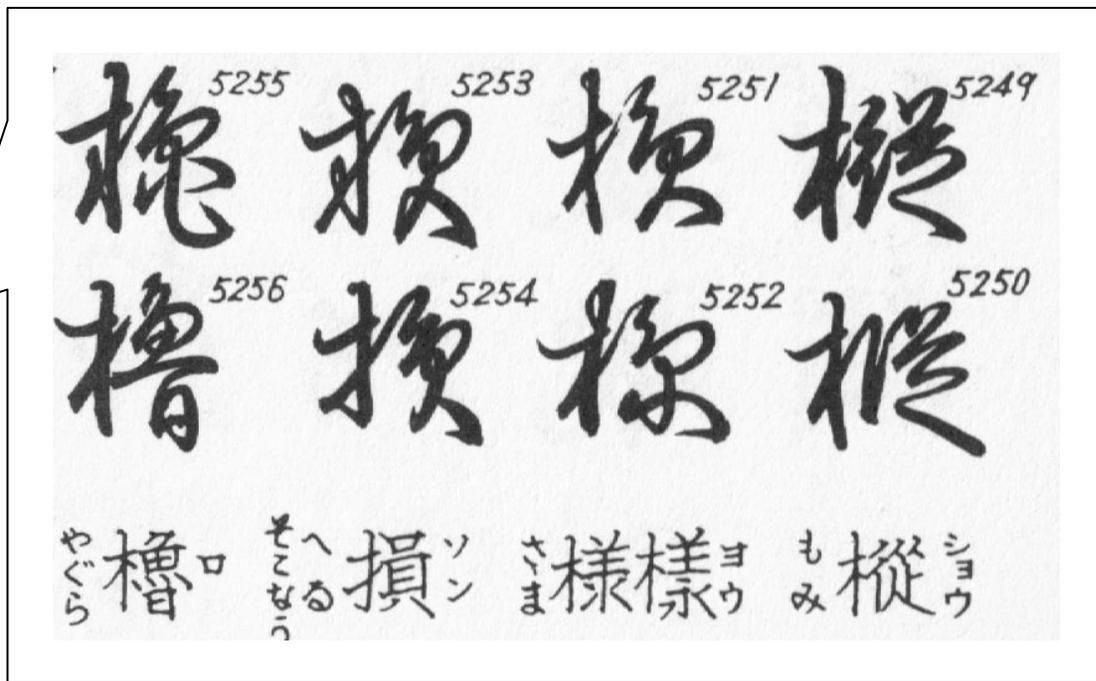
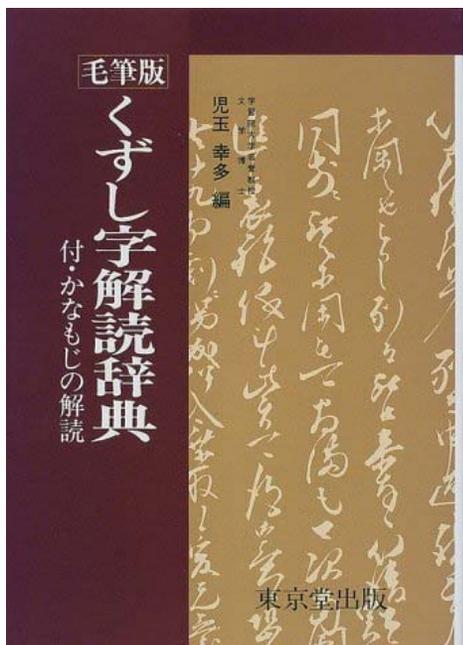
- カスレ
 - 今回は対応見送り

- 細線化すると...



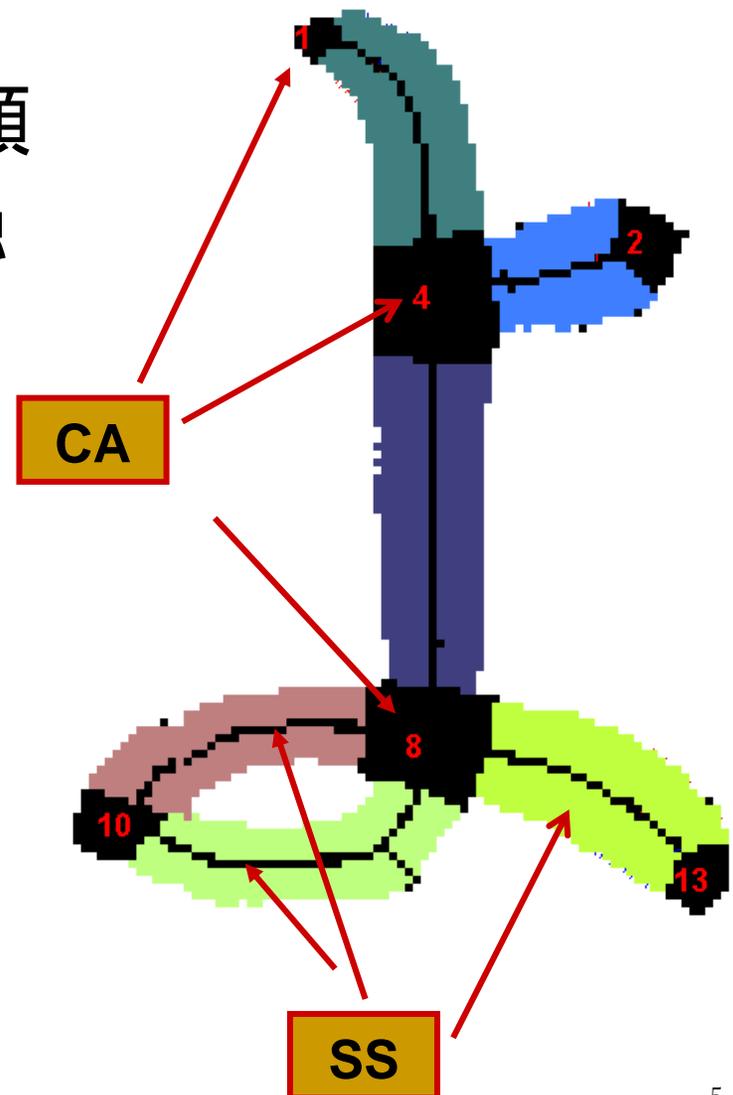
今回対象とする毛筆文字画像

- 「毛筆版くずし字解読辞典」東京堂出版
 - スキャン・2値化したもの (150×150ピクセル程度)



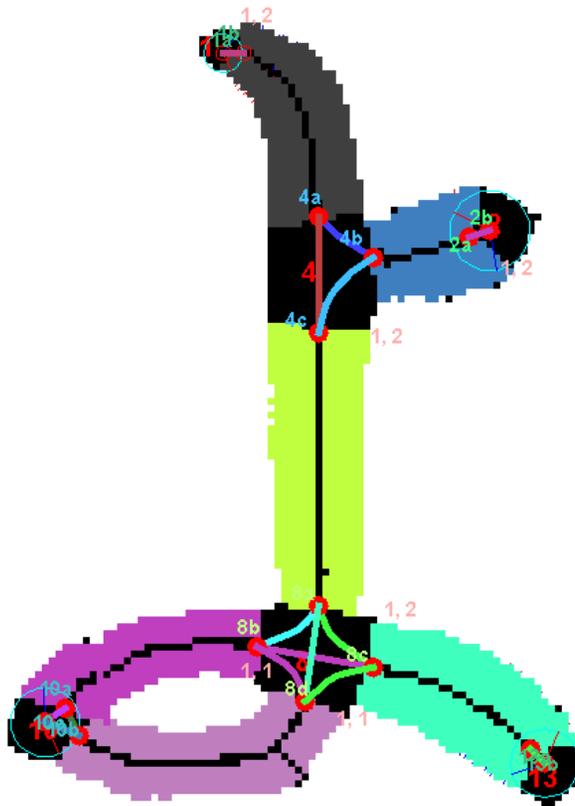
アルゴリズムの流れ(1)

- 黒ピクセルを2種類に分類
 1. 他のストロークと交差・接触していない部分
 - 明らかなストローク領域
 - **SS (Sub Stroke)**
 2. その他の領域
 - 接触・交差を含むややこしい部分
 - **CA (Complicated Area)**

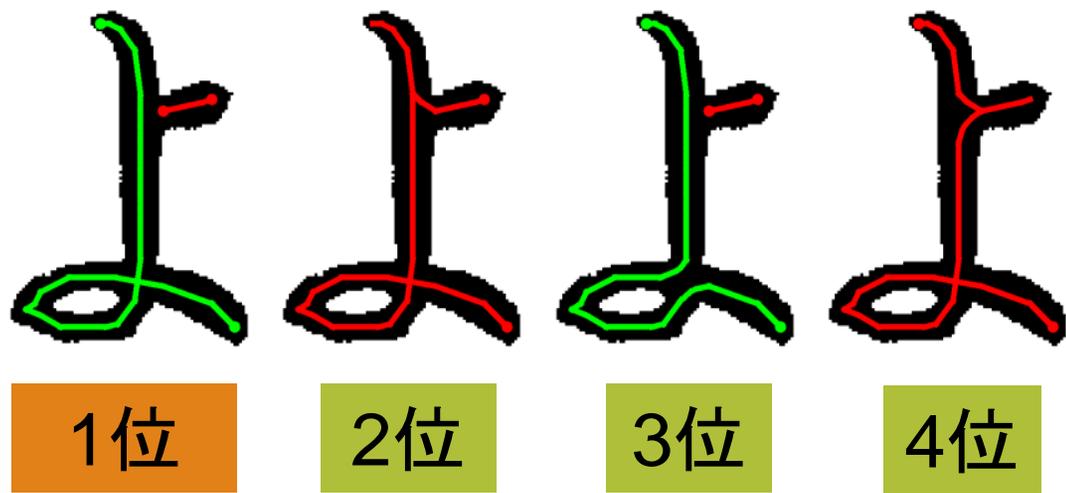


アルゴリズムの流れ(2)

- CA内部の筆の動きを推定
- ストローク候補を生成
評価値順に並べる



ストローク抽出結果

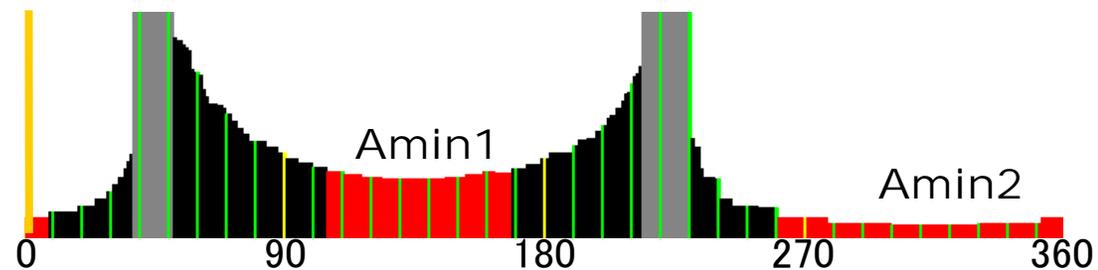
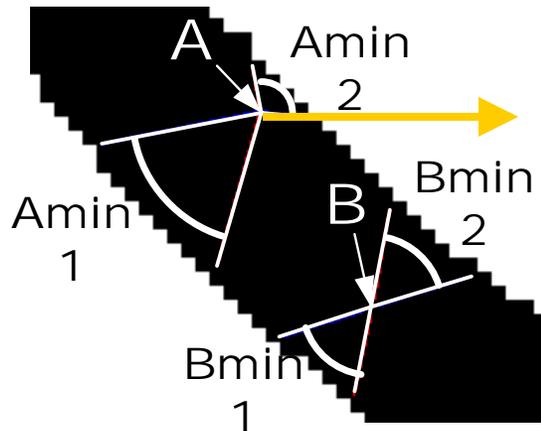


アルゴリズムの詳細

SSの抽出(1)

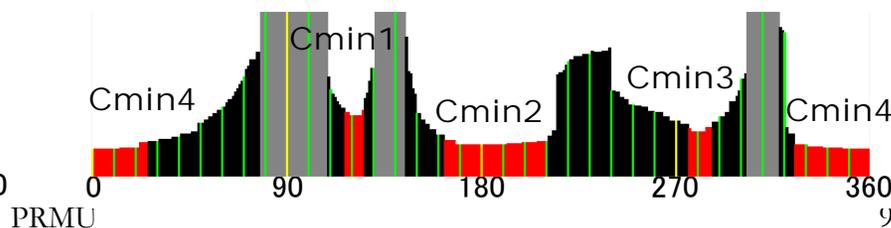
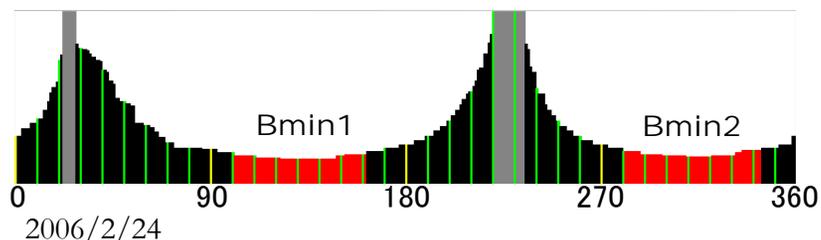
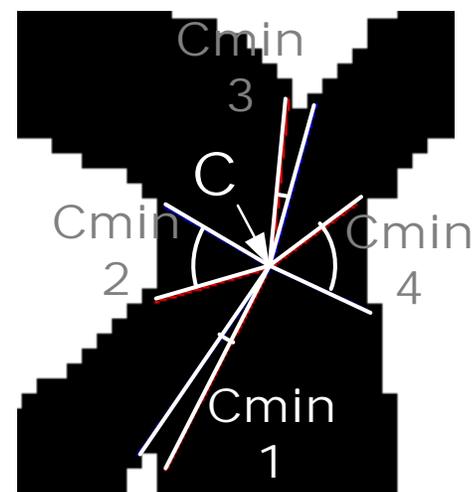
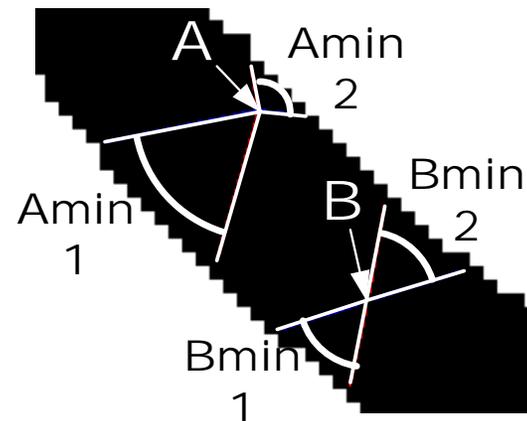
■ 角度距離グラフ

- 黒ピクセルから周囲360度方向をスキャン
 - レーダーのように
- 白ピクセルまでの距離をグラフ化

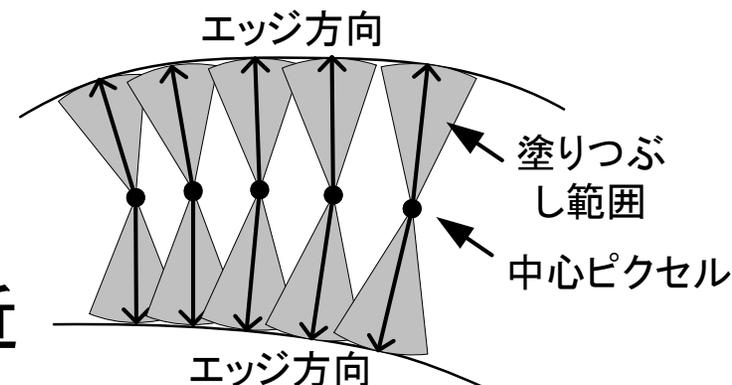


SSの抽出(2)

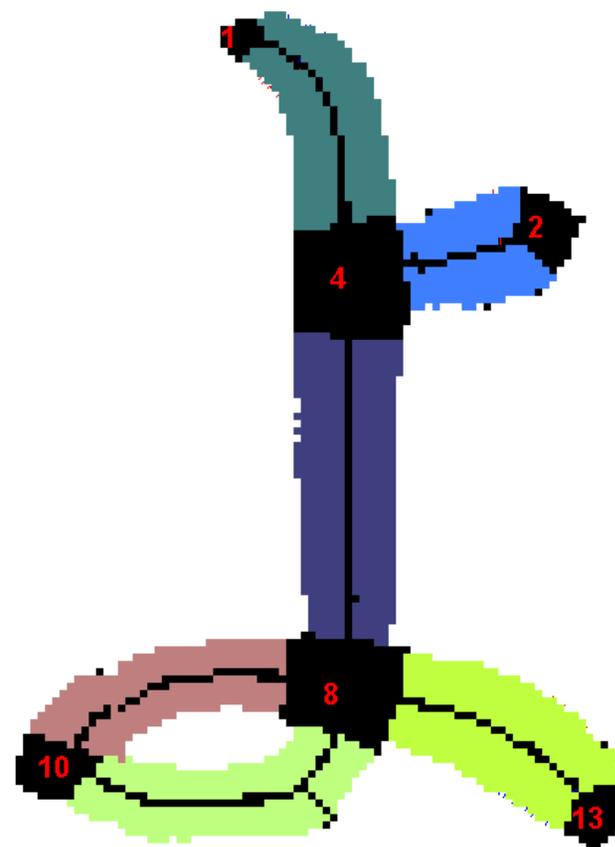
- 角度距離グラフを手がかりに筆の中心が(ほぼ確実に)通ったピクセル(中心ピクセル)を判定
 - 最も値が小さい2つの極小値がほぼ等しい
 - 2つの極小値の角度がほぼ180度
 - 極小値の近傍が、ほぼ「まっすぐ」
 - 交点内部ではない



SSの抽出(3)

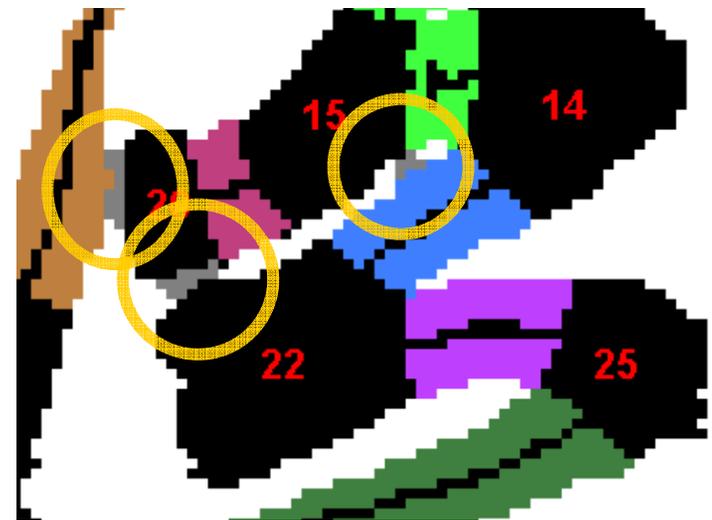
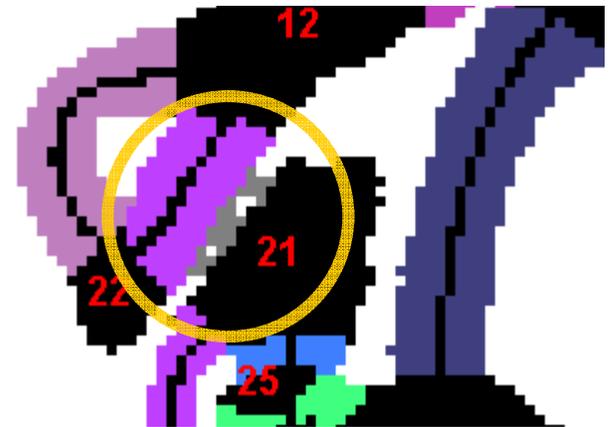


- 中心ピクセルの集合を互いに近いもの同士でグループ化
 - それぞれのSSに分離
- 端から端に並べる
- SSの中心線を筆が通ったときに塗りつぶす周囲の領域を判定
 - 右図の色つき部分
 - **塗り残した領域がCA**
- SSとCAの接続関係はグラフで表現



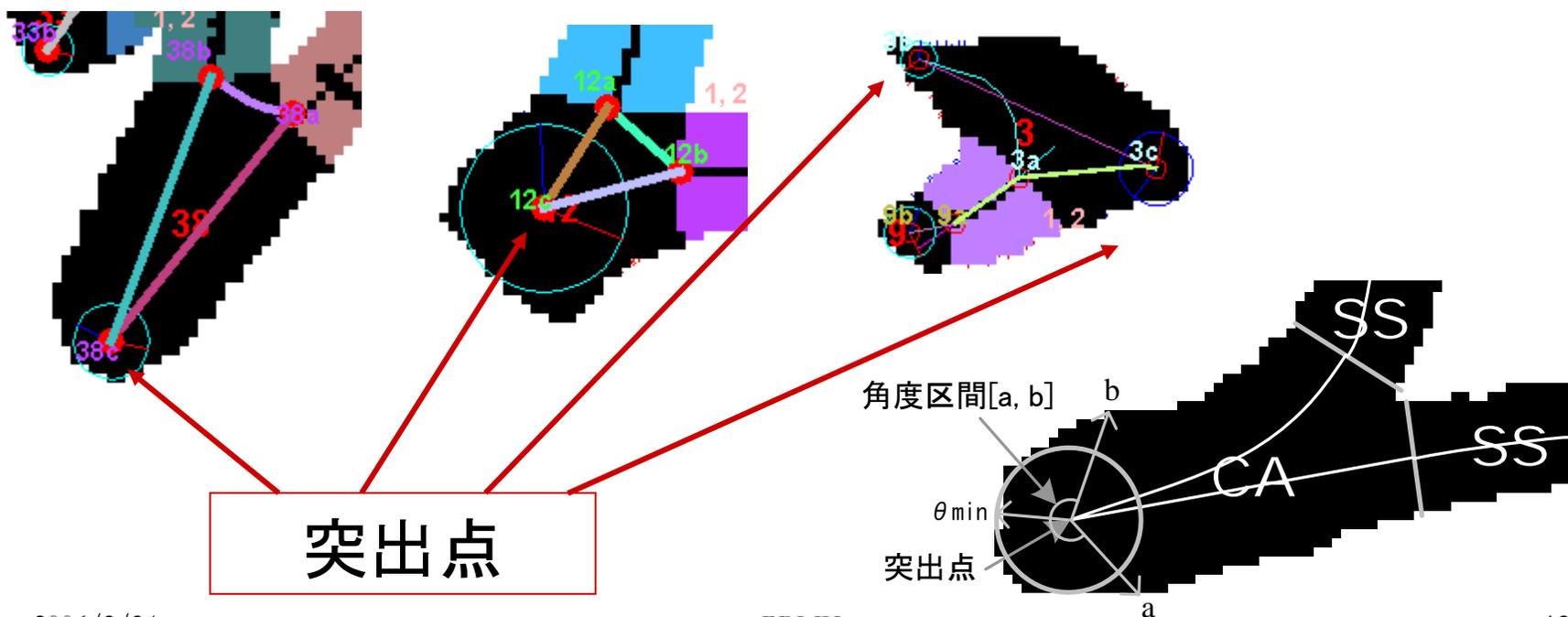
接触点除去

- 接触点で偽のSSが生成されることがあるため、これを除去
- 条件
 - SSの幅が十分細い
 - 端点間の距離が十分短い
 - SSが「くびれている」
 - 角度距離グラフで判定
- ここで取り除ききれなかった接触点は、ストローク列挙時に考慮



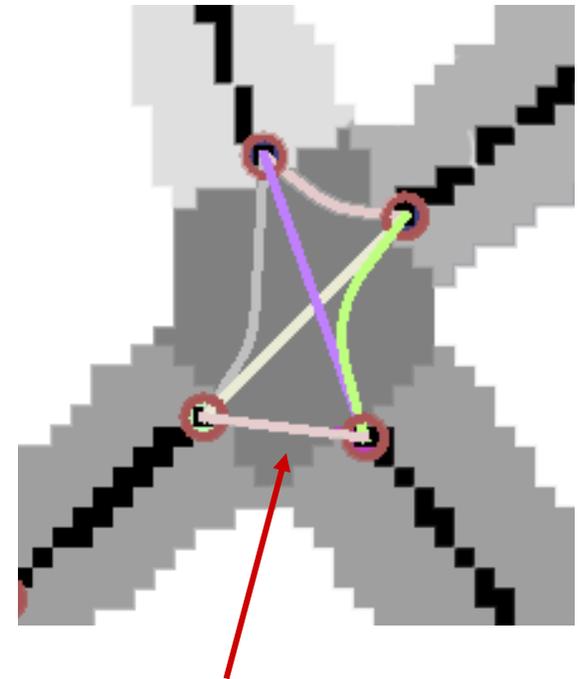
CA内の突出点抽出

- CAの「出っ張っている部分」
 - ストローク推定の上で重要
- 角度距離グラフで判定



CA内部の経路推定(1)

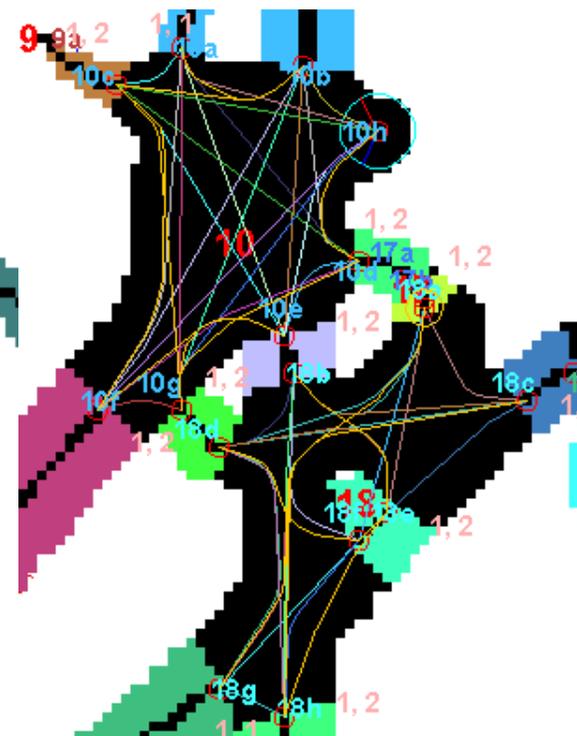
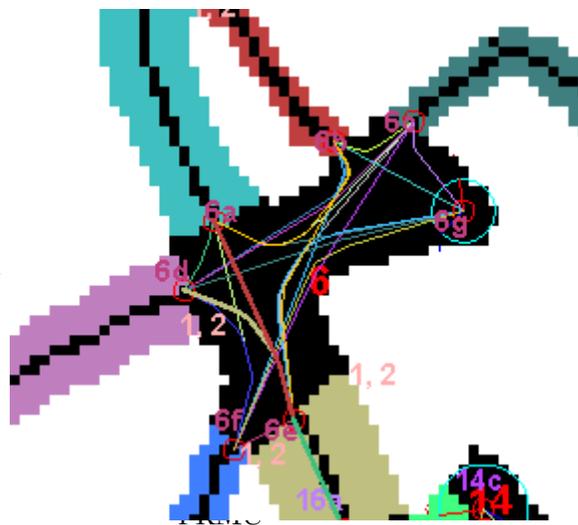
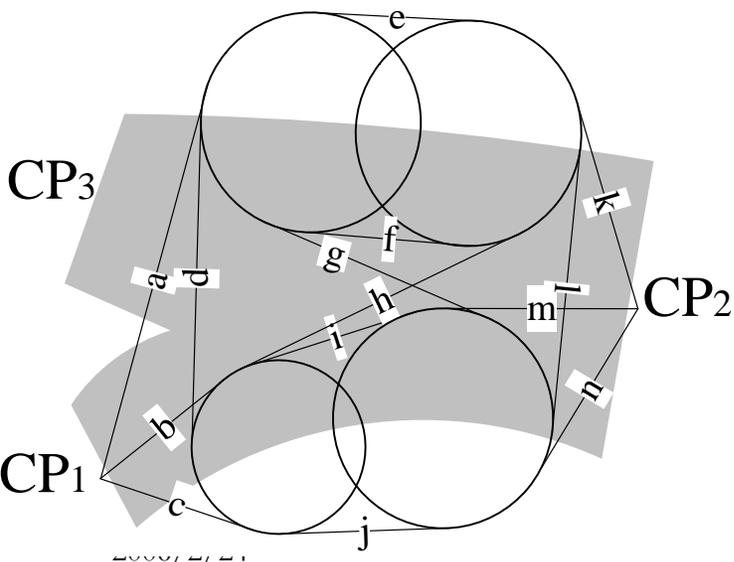
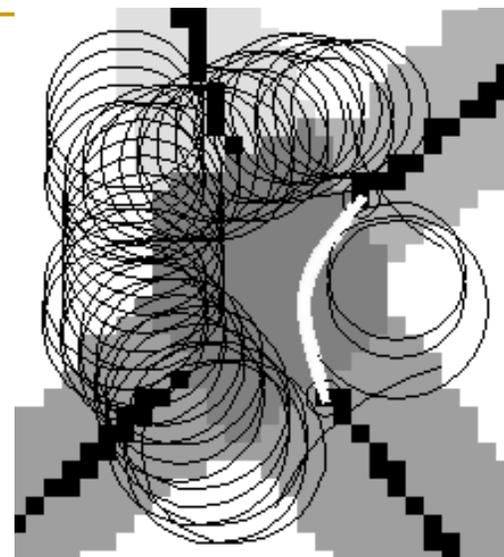
- CA内部での筆の動きを推定
 - SSと突出点の全てのペア間で
 - CA内パス
- 観察
 - CAの縁から, 筆の半径 r 以内の領域は筆の中心が通ることではない



CA内パス

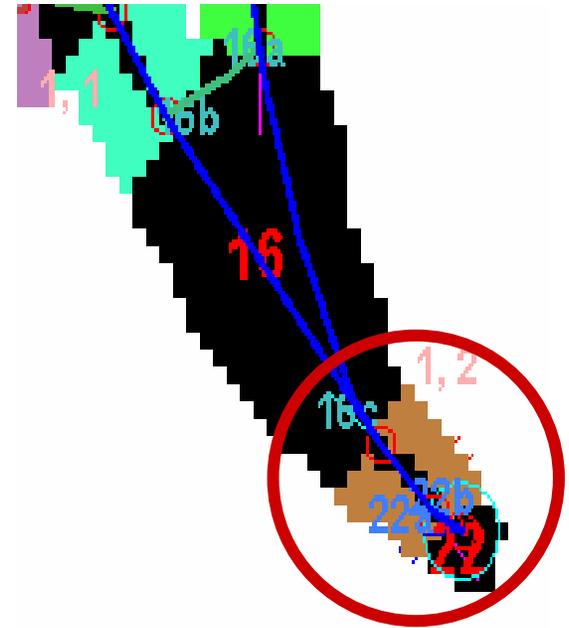
CA内部の経路推定(2)

- CAの縁を中心に円を配置
 - 円内を通らない経路を求める
- 全ての円の間で接線を求める
 - 他の円で遮られないもの
- 端点から端点まで全経路を探索



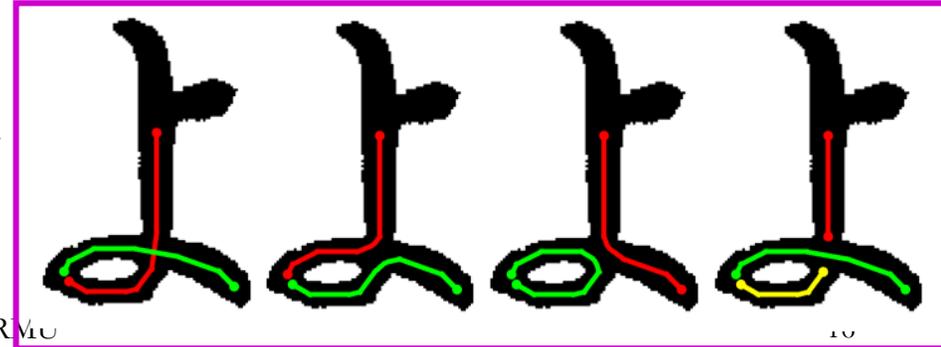
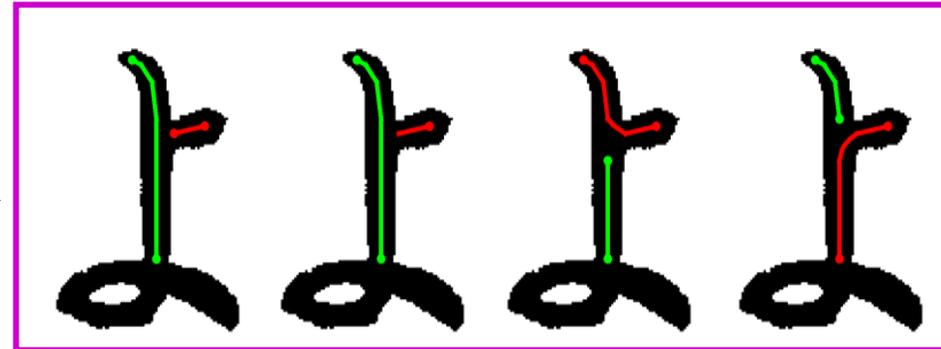
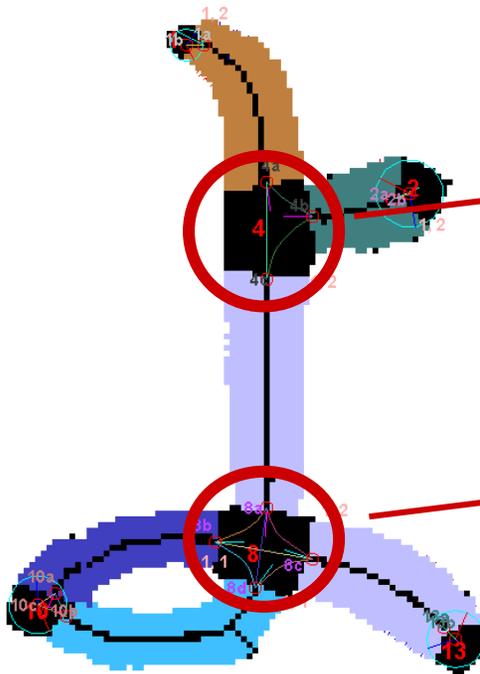
ストロークの列挙(1)

- CAを通るストロークを列挙
- 考慮すべきこと:
 - SSを筆が通過する回数**
 - 0 ... SSが偽(接触点)
 - 1 ... 通常のストローク
 - 2 ... 折り返しや交点部分
- 各SSに最小筆通過数, 最大筆通過数を決定
 - 基本的には最小, 最大筆通過数 = 1
 - ごく短いSS: 最小筆通過数 = 0
 - 比較的真っ直ぐなSS: 最大筆通過数 = 2



ストロークの列挙(2)

- CAに接続する全SSの筆通過数の全組み合わせを列挙
- さらに, 筆通過数に適合するCA内パスの全組み合わせを列挙
- 選んだCA内パスと周囲のSSと接続

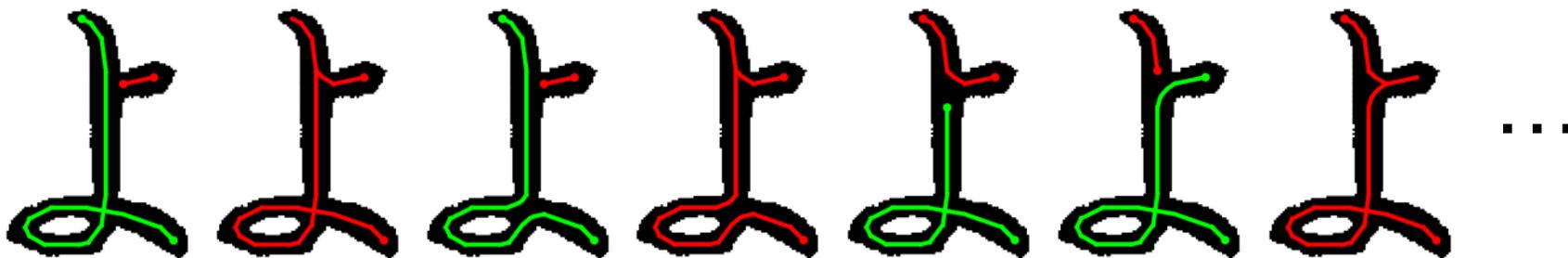


ストロークの接続(1)

- 隣接CAの候補同士を接続



接続



ストロークの接続(2)

- 接続結果は膨大な数になる
- 毎回接続した後, ありそうにないストローク集合を破棄することで計算量を抑制
 - 各ストローク集合を評価
 - 評価値の良い上位100候補だけを残す

ストローク集合の評価関数

■ありそうにないストローク集合に大きな値を与える

■不自然な個所にペナルティを加算

■ ストロークの評価

- CA内パス曲がり角度
- S字カーブ
- ループ
- T字路
- 太さの差
- 不自然な折り返し
- 折り返し長

- 折り返し入口
- 折り返し内の太さ
- 折り返し数
- 折り返し角度

■ ストローク集合の評価

- ストローク数
- ストローク交差部分の長さ

評価

抽出例

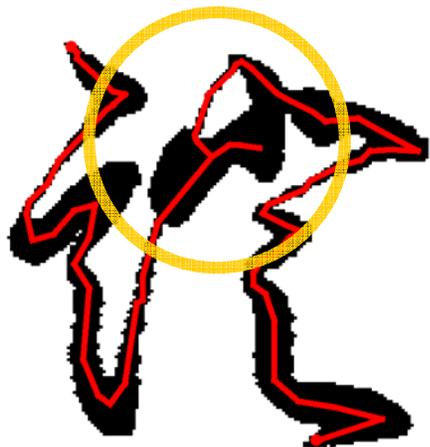
強社強好強強強

好強強強強強強

強強強強強強強

強強強強強強強

失敗の要因



突出点の見落とし



筆通過数=3



折り返しのコスト過大



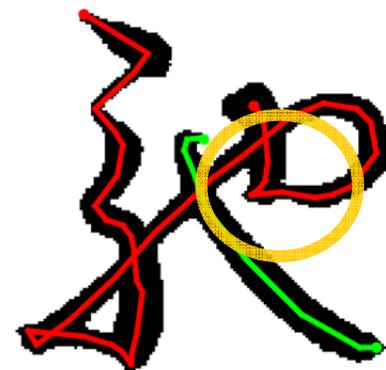
太さの差を過大評価

2006/2/24



正解候補を
途中で破棄

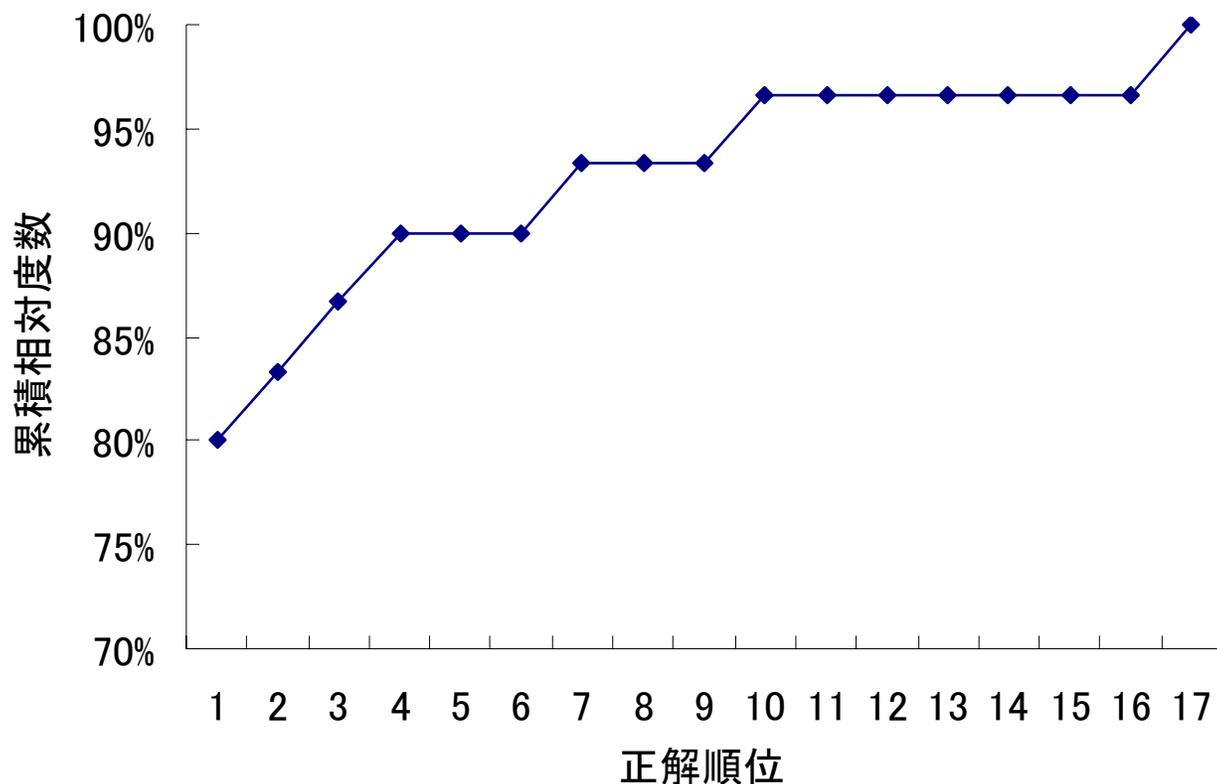
PRMU



別のストロークを
接続

ストローク正解率(参考)

- 「毛筆版くずし字解読辞典」から抽出した40文字
- 正解順位の累積相対度数



まとめ

- 毛筆文字からストロークを抽出する方法を提案
 - 角度距離グラフによる明らかなストローク部分(SS)と、そうでない部分(CA)の分離
 - 計算幾何学的手法によるCA内部のパス推定
 - 折り返しを考慮したストロークの列挙
 - ストローク評価関数
- 今後の課題
 - 評価関数の改良(「自然な文字の書き方」の知識を追加)
 - より精密な評価
 - ストロークの向き(書いた方向)・筆順(ストローク間の順序)判定
 - カスレへの対応
- (時間があれば)デモ