n-gram 解析を用いた画像中のパターン抽出

1D－5

内田 淳　梅村 恭司

豊橋技術科学大学 情報工学系

1 はじめに

インターネット等により、情報化が進む世界において我々は全く知らない文字等を目にする機会があるかもしれない．そして、そこから意味のある固まりを見つけることは非常に困難であるが有益である．そこで我々は画像の中のパターンを分析することにより意味のある固まりを図形として抽出する方法を考える．ここではなるべくフォント情報を持たずに検索することを理想とする．それによって、フォントの有無に関わらずあらゆる文字の検出が期待できる．

本稿ではコンピュータが作り出す画像イメージに対して n-gram 解析の手法を用い、それによって意味のある固まりを取り出す方法を提案する．また、作成したプログラムについてその検出結果を示し，考察を行う．そして最後に，今後の課題と共にまとめる．

2 アプローチ

一般的に文書とはアルファベット、平仮名、漢字等何らかの文字の組合せで構成される．したがってこれらの文字は文書中に何度も現れ、また、繰り返し現れる文字の組合せは意味がある固まりだと予測することもできる．今、我々は文字をただの図形、つまり点の集合と考えることにした．この場合も文書中の文字列と同様に繰り返し現れる点のパターンは意味のある固まりと考えられる．またこの方法ではフォント情報が必要なく，おおよその大きさを与えるだけでよい．そのため日本語，英語といった枠にとらわれずあらゆる文字及び図形に対応が可能だと考えた．

3 図形における n-gram

一般的に n-gram とは文書中の n 個の部分文字列である．つまり文書中にこの n-gram が何度現れるかを調べ、一定回数以上現れた n-gram を意味のある固まりとして取り出すことが本来の n-gram 解析として行なわれている [1][2][3][4]．しかし、我々の手法では文字を図形として扱うため本来の n-gram とは少し異なり、図形を文字列に変換したものを n-gram とする．本実験において我々は対象となる画像として bitmap 形式という二値画像を扱った．二値画像は白い部分が 0、黒い部分が 1 に割り当てられているため、n-gram は n 個の 0,1 の組合せである．つまり、n-gram の内容が n 個の 0,1 の組合せに変わっただけで基本的な考え方は本来のものと同じである．

4 図形の文字列化

本稿において、n-gram は 0,1 の組合せであることは述べたが、それはある画素を始点とするある大きさの図形を 0,1 からなる文字列に変換したものである．つまり，着目している画素から右及び下に広がる n×n 画素からなる図形をある経路に従って 0,1 の文字列に変換し、この文字列をその画素のデータとして割り当てることになり，あとは，自然言語で用いられる n-gram 抽出を行う．

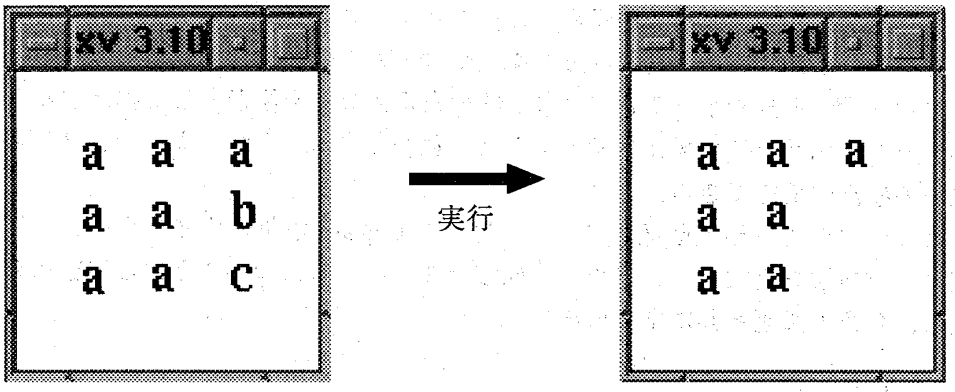
Extracting Pattern in Image Using N-gram Analysis

Jun Uchida

Department of Information and Computer Sciences, Toyohashi University of Technology

5 実験結果

以下に作成したプログラムによる抽出結果を示す.



6 考察

実際に作成したプログラムによってフォント情報をほとんど持たずに繰り返し現われる図形を抽出することができたが，実行の際，切り出す図形の大きさ，また，出現頻度，一致の長さ，黒い部分の割合に対する閾値は我々が与えるものであり，これらの設定は抽出結果に大きく影響した．これは図形によってその特徴が異なるためであり，よってその特徴に適した条件値を設定する必要がある．しかしながら，現段階では繰り返し実験することによってその値を選ばならず，大変難しい問題である．

7 まとめ

m-gram 解析を適用することにより，繰り返し現われる図形を抽出する手法を提案し，実際にフォント情報をほとんど持たずに意味のある図形を抽出するプログラムを作成した．しかしながら，各条件値は抽出結果に大きく影響することが分かった．よって，様々な種類の画像で実験することにより最適な条件値を求める等，より実用的にしていくことが今後の課題である．

参考文献

[1] 森 信介, 長尾 眞,「n-gram 統計によるコーパスからの未知語抽出」, 情報処理学会研究報告書,NL108-2,pp7-12(1995)

[2] 下畑 さより, 杉尾 俊之,「隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出」, 情報処理学会研究報告書,NL114-3,pp13-18(1996)

[3] 國吉 芳夫, 中西 正和,「ギャップのある n-gram による言回し抽出」, 情報処理学会研究報告書,NL117-6,pp37-44(1997)

[4] 吉川 裕之, 貴島 寿朗, 梅村 恭司,「n-gram 解析を用いたプログラム中の非定型パターン欠損の検出」, 情報処理学会研究報告書,PRO15-2,pp9-15 (1997)