

テンプレートマッチを用いた画像認識による顔文字の検出

情報・通信工学科 学籍番号:1211072 寺田研究室 佐々木 透

1 はじめに

顔文字とは文字列に視覚的情報を持たせる手法の一つであり、自然言語処理では文書の感情判定に用いる研究がある。顔文字の感情判定には必ず顔文字を検出するステップが存在し、文字コードによる検索手法 [1][2] が用いられている。しかし、Unicode は毎年文字が新規追加され、顔文字にも未知の記号が増えていくと文字コードによる検出手法では検出精度が低下する。本研究では、顔文字に使われる文字が未知の記号であったとしても人間には顔文字として認識出来ることに着目し、画像認識を用いた顔文字の検出法を実装した。

2 関連研究

2.1 CAO: A Fully Automatic Emoticon Analysis System[1]

Ptaszynski らは目と口の形状の組み合わせに感情が表れるという理論のもと、入力文書内から顔文字を自動で検出し、感情判定を行うシステムである CAO を実装した。CAO では Web 上に存在する顔文字辞書から感情ラベルがついている顔文字を収集し、10,137 の顔文字から構成されるデータベースを構成して以下のステップで検出した顔文字の感情分類を行っている。

1. 検出した顔文字を DB 上の顔文字と完全一致検索
2. 検索失敗した場合は目と口の組み合わせのみを検索

入力文書内の顔文字の検出のアルゴリズムについては、収集した顔文字に出現する頻度の高い 400 の記号が 3 つ以上含まれる行に対して上の 2 ステップの検索を行っている。

2.2 ツイートに出現する顔文字等の文字と記号に着目した感情分類 [2]

三好らは Twitter について、ツイートに含まれる顔文字を用いて感情抽出、および顔文字の感情分類を利用して、ツイートを肯定 (Positive) と否定 (Negative) に分ける PN 分類を実装している。顔文字の抽出には次の正規表現を用いている。

```
((?!C{3,}).){2,}
```

上記の C は英数字、ひらがな、カタカナ、または漢字の 1 文字を表し、この正規表現は、次の条件を満たす文字列を半角括弧で囲んだ文字列にマッチする。

- C が 3 文字以上連続しない
- 2 文字以上の文字列

3 提案手法

本手法は日本語の文章を入力とし、顔文字に使用される頻度の高い記号が連続している領域は顔文字である可能性が高いとして、次のようなステップにより顔文字を検出する。

1. 顔文字の収集と記号の画像化
2. 入力テキストを画像化
3. マッチ
4. 座標の出力と顔文字領域の検出

3.1 顔文字の収集と記号の画像化

Web 上に存在する顔文字を収集し、収集した顔文字に出現する頻度の高い記号を画像化する。

3.2 入力テキストの画像化

顔文字の検索対象となる日本語の文章を前項の記号の画像化と同様の条件で画像化する。

3.3 座標の出力と顔文字領域の検出

3.2 の画像中に、3.1 で画像化した記号とが含まれる座標をテンプレートマッチ法によって検出する。3 回以上連続している箇所を顔文字領域と定義し、顔文字領域の始点と終端を出力する。

4 実装

4.1 開発環境

本システムにおける画像の描画は Java Standard Edition 7 の Graphics2D クラスを用いており、テンプレートマッチングは Python2.7 と Intel の OpenCV2.4[3] によって実装している。

4.2 描画

Java Graphics2D クラスの drawString() メソッドにより描画する、白い長方形領域にフォントを MSP ゴシック、大きさ 16pt で描画する。入力文字列は文字の上下にパディングとして 1 ピクセルの白色ドットが入るように描画し、テンプレート画像は記号の黒色部分をのまわりに 1 ドットだけパディングピクセルを入れて切り詰めたものとする。

4.3 座標の出力

マッチした座標は次の形式で出力している。

template 番号, x 座標, テンプレート画像の幅

4.4 検出アルゴリズム

上の出力を受けて 3.3 で述べた顔文字領域の始点と終端を次の手順に従って判定している。

1. 出力のリストを, x 座標にしたがってソート
2. (x 座標 + テンプレート幅 + 24 ピクセル) 以内に次のマッチングした座標の有無を判定
3. 2 が True ならカウントを進めて次の座標に対して 2 を実行
4. 2 が False かつ, カウントが 3 未満の場合はカウントを初期化して次の座標に対して 2 を実行
5. 2 が False かつカウントが 3 以上の場合は探索を終了し, 最初のカウントした x 座標を始点, と最後にカウントした (x 座標 + テンプレート幅) を終点として出力

5 評価実験

先行研究の CAO が顔文字の検出に用いた 400 記号からひらがな, カタカナ, 漢字を除いたものをテンプレート画像として実験を行った. テンプレートに存在しない記号が多く含まれ現在広く使われる顔文字として特殊顔文字があげられる. 特殊顔文字 500 と顔文字を含まない 500 の日本語文に対して本手法, CAO が用いた検出手法, 正規表現による検出手法, の 3 つによって適合率, 再現率を算出した.

5.1 実験結果

表 1 適合率と再現率

検出手法	適合率	再現率	F 値
本手法 A	0.905 (418/462)	0.836(418/500)	0.869
本手法 B	0.972 (418/430)	0.836 (418/500)	0.899
CAO	1.00 (390/390)	0.780 (390/500)	0.876
正規表現	0.959 (402/419)	0.816 (408/500)	0.882

本手法は文字コードによる検索と比較して再現率は高いが, 適合率が下がる傾向にあった. また本手法が日本語文字列を False positive と検出したものは頻度順に次の 3 パターンに分けられた.

1. ”!”や”?”など, 同一記号の連続 (23/44)
2. テンプレートが偶然密集する (12/44)
3. 英単語, 数字 (9/44)

2 に関しては現状対処の方法がないが, 1 と 3 に関しては描画ステップに渡す入力文字列に正規表現でフィルタリングをかけることによって文字列の描画前に除去が可能であると考えられる. 表 2 行目の本手法 B 上の 1 と 3 を除いて計算したものである. また以下はテンプレートに含まれない記号を検出できた例の一部である.

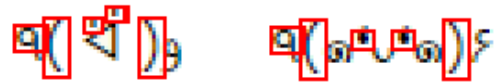


図 1 テンプレートに含まれない特殊文字が検出できた顔文字の例

6 まとめと今後

6.1 本手法の利点

画像認識による顔文字の検出は, 見かけ上顔文字として出現頻度の高い記号であれば文字コードの違いに関わらず検出することが可能であるため, 再現率は高くなる. テキストによる検出手法で新規の顔文字に対応するためには検索に用いるキーを増やすという手間を必要とするが, 画像認識を用いた検索ではそのような負担を軽減することにつながると思われる.

6.2 本手法の欠点

適合率は文字コードによる検出方法よりも低くなる傾向になる. 入力文字が隣接する記号と接近しテンプレートとの類似度が下がるケースが僅かに存在し, 画像認識特有の欠点と言える.

6.3 検索アルゴリズムの改良

今回はテンプレートについて, マッチした座標とその連続回数に基づいて顔文字の検出をしているが, ”” ”と” ”. ” のよう同一の座標に複数のテンプレートが反応して回数のカウントが 2 回進む場合があるのでクラスタリングによって共通化する改善方法が考えられる. 5.3.2 で示した通り誇張表現などで記号を連続したり, 日本語文でも英単語や数字が入るケースがある. そのようなケースは, 同一テンプレートが連続して出力されている場合は無視する, 正規表現によって描画前に英単語, 数字列を除去するなどして適合率を高める方法などが考えられる.

6.4 顔文字の感情分類への応用

顔文字に関する研究は顔文字の感情分類が主流であり, そして顔文字の感情分類は顔文字中に出現する記号パターンをもとに行う場合が多いが, テンプレートマッチによって代替が可能である.

参考文献

[1] Michal Ptaszynski, “AO: A Fully Automatic Emoticon Analysis System Based on Theory of Kinesics”IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 46-59, 2010. 1.

[2] 三好 辰明, 太田 学, “ツイートに出現する顔文字等の文字と記号に着目した感情分類”The 5th Forum on Data Engineering and Information Management. 2013. 3.

[3] OpenCV , <http://opencv.jp/>.