

新聞記事における人名と写真内の顔領域との自動対応づけ

山田 剛一 杉山 一成 与那嶺 靖典 中川 裕志

1F-06

横浜国立大学 工学部 / 東京大学 情報基盤センター

1 はじめに

あるメディアを知的に扱うということは、そのメディアのコンテンツであるデータそのものを扱うだけでなく、そのメディアのコンテンツの持つ意味内容までを扱うということである。さらにマルチメディアの場合、それを構成する各メディアの持つ意味内容、およびメディア間の意味関係までを含めたものがマルチメディアの意味内容となるため、それらを扱わなければ、マルチメディアを真に知的に扱っていることにならない。

さて、そのような意味内容を記述することによって、知的なマルチメディアコンテンツを構築することが可能となり、知的検索をはじめとした様々な応用が実現可能となるのであるが、その知的なマルチメディアコンテンツを構築する際に、メディアの内容を人間が意味解釈を行って記述していくのでは、人的資源上、限界がある。そこで、そのような意味内容をデータから自動生成しよう、というのが本研究である。

本研究では、マルチメディアコンテンツとしてWeb上で公開されている写真ニュースを扱う。写真ニュースは記事本文と写真画像からなるが、その記事本文中の人名と写真中の顔画像との対応関係を推定するシステムを構築している(図1)。

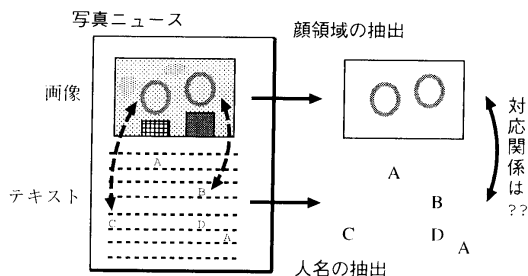


図1: 人名と顔画像領域との対応関係を推定.

我々は以前、同様に写真ニュースを対象として、記事本文と写真の両方に現れる人物の人名群と顔領域群を特定するシステムを構築した[1]。この段階でも、例えば人名から写真を検索するといった応用が可能なのであるが、人名と顔画像領域との一対一の対応関係になっていないため、顔を選択して人物情報を検索するといった応用には利用できなかった。

人名と顔の一対一の対応関係は、記事本文中に「写

真右」などの明示的な表現がある場合にはそれだけで推定が可能である¹が、一般には、顔の識別が必要となる。写真ニュースの場合、現れる人物が不特定なため事前の学習が不可能であるし、もし著名人だけを対象とするにしても、画像の解像度が低く、照明条件も悪く、正面を向いているとも限らない状況であるので、顔の識別は現実的には不可能である。

本研究では顔の識別は行わず、次節で述べる以下の仮説に基づいて、人名と顔の一対一の対応関係を推定する。

仮説 1 ある記事において重要な人物であるほど、記事本文中でも写真中でも重要に扱われる。

2 対応関係の推定方法

本研究のシステムは、我々が以前構築したシステム[1]の発展形である。一対一の対応関係の推定法を述べる前に、このシステムの動作原理を簡単に示しておく。

本研究での最も基本となる考え方は、以下の仮説として表されるものである。

仮説 2 両メディアに共通して現れる人物とそうでない人物とでは、記事本文中の人名の現れ方や、写真中の顔の現れ方に違いがある。

本研究ではこの仮説に基づき、各メディアにおけるその現れ方の違いを学習し、それによって両メディアに共通して現れる人物かどうかを判定する。機械学習システムには、分類モデルとして決定木を使う(5.0を用いている)。

また、「両メディアに共通して現れる人物の人数によって、顔領域や人名周辺の言語表現の特徴に差がある」という事実がある。これは、例えば写真内での顔の位置や大きさを考えれば明らかであろう。そこで本システムでは、特徴を学習する際には人数ごとに場合分けをして学習を行う。推定時には人数は未知であるため、本システムでは「共通して現れる人物は何人である」という人数に関する仮説をまず立てた上で、そのそれぞれの仮説の下に処理を行なう。その仮説は、両メディアに共通して現れる人物が{1人, 2人, 3人以上}の3通りである。システムは、各仮説の下に得られた結果の組み合わせの中から最も確からしいものを選択し、その結果の組からシステム全体としての判断を下す。

システムの構成としては、まずメディアごとに処理内容が異なることから、記事本文を処理する言語モジュールと、写真画像を処理する画像モジュールとがある。これらのモジュールは、機械学習システムが出力する判断の信頼度を重みとして持つ候補群を出力する。さらに、これら言語/画像モジュールの結果から、システム

¹人数が多い場合にはそう単純ではない[2]。

全体としての結果を導く統合モジュールがある。統合モジュールは、言語/画像の各モジュールより出力される候補をまとめ、それらの各モジュールによる重みから全体としての重みを計算し、出力する。

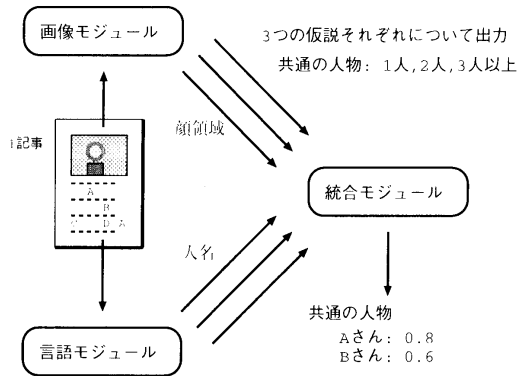


図 2: システム構成の概観.

本研究のシステムでは、機械学習システムが判断した際に出力する重み(信頼度)に基づいて、一対一の対応づけを実現している。具体的には、画像、言語の各モジュールからの候補をそれぞれ重みの高い順に並べ、同一順位の人名と顔領域が対応していると推定する。これは、画面に出ているという特徴の強い人名が、テキストに出ているという特徴の強い顔画像と対応する、という考えに基づいている。この考え方の根底には、前掲した次の仮説がある。

仮説 1 ある記事において重要な人物であるほど、記事本文中でも写真中でも重要に扱われる。

この仮説は直感的には正しい。しかし、新聞記事では、重要な人物の写真が常に手に入るとは限らないといった事情があるため、この仮説はすべての場合において成り立つとは言えない。

3 評価

実際の写真ニュースを解析し、メディアの統合の効果を評価した。その効果を確認するため、言語/画像の各モジュール単体からの情報のみによる解析結果と比較した。

評価に用いたデータは、毎日新聞社が Web で公開している写真ニュース²[3]である。写真の特徴としては、通常キャプションが存在せず、画像のピクセル数は 250 × 200 程度のものである。今回の評価では 1997 年 5 月と 6 月の記事のうち、写真がカラーで、記事本文中に画像中の人物の人名を含む 228 記事を使用した。言語モジュールは 4 fold、画像モジュールは 3 fold の交差検定を行った。

システムが正しい人名と顔領域の組を抽出できるかを評価するため、その再現率/適合率を求めた。

²現在は、写真ニュースは PHOTO ジャーナルと改名されている。

再現率/適合率は次の式で定義した。

$$\text{再現率} = \frac{\text{出力した正しい組の重みの総和}}{\text{実在する組の総数}} \quad (1)$$

$$\text{適合率} = \frac{\text{出力した正しい組の重みの総和}}{\text{出力したすべての組の重みの総和}} \quad (2)$$

各モジュールの評価結果は、表 1 のようになった。

表 1: 顔 - 人名の組の再現率/適合率

出力元	再現率	適合率
言語モジュール	0.70	0.70
画像モジュール	0.53	0.63
統合モジュール and 統合	0.30	0.49
統合モジュール or 統合	0.42	0.47

人名と顔領域の対応をつけるためには、基本的には、言語/画像の双方のモジュールの出力が正しくなくてはならない。よって、一対一の対応づけの精度は、言語/画像の各モジュール単体の精度の積となるはずであり、実際そのような結果になっている。

4 おわりに

今回は、人名と顔の一対一の対応関係を決定づける、各メディアにおける重要度として、両メディアに共通して現れる人物であるという判断の信頼度を用いた。これを、各メディアにおける一般的な重要度算出方法と組み合わせた場合の実験も行ってみたく考えている。

また、現在は「写真右」などと明示的に書かれている場合の例外処理を行っておらず、本稿で提案した方法のみによるシステムとなっている。今後、例外処理も含めた統合的なシステムを構築し、性能評価をしたいと考えている。

謝辞 充実した写真ニュースを Web 上で発信している毎日新聞社に敬意を表すると同時に感謝いたします。

参考文献

- [1] Koichi Yamada, Kazumari Sugiyama, Yasunori Yonamine, and Hiroshi Nakagawa. Identification of coreference between names and faces. *ACL'99: the 37th Annual Meetings of the Association for Computational Linguistics: Workshop: Coreference and Its Applications*, 1999.
- [2] Rohini K. Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, Vol. 28, No. 9, pp. 49-56, 1995.
- [3] 毎日新聞社. 毎日新聞 AULOS 写真ニュース. (現 Mainichi INTERACTIVE <http://www.mainichi.co.jp/>), 1997.