1F4-02

放送映像からの人物相関グラフの構築

Construction of a Human Correlation Graph from Broadcasted Video

小笠原 崇[*1] Takashi OGASAWARA
高橋 友和[*1] Tomokazu TAKAHASHI
井手 一郎[*1*2] Ichiro IDE
村瀬 洋[*1] Hiroshi MURASE

[*1]名古屋大学大学院情報科学研究科 Graduate School of Information Science, Nagoya University
[*2]国立情報学研究所 National Institute of Informatics

In order to extract human correlation from real world data, text information has been the sole source so far. This means the correlation with a person who does not appear explicitly in the text has been disregarded. Therefore, we propose a method to analyze the pattern of the appearance of a person face image in broadcasted video for the extraction of the correlation including these people. In this work, it was confirmed from experiments that we can extract correlations that were not able to be extracted only by text information from image information. In addition, it was also confirmed by comparison with manual extraction, that automatic extraction from both text and image information was appropriate.

1. はじめに

1.1 研究の背景と目的

近年ハードウェアの進歩などにより、大容量の映像資源を蓄積し利用することが可能となっている。それに伴い、それら大容量の映像資源の有効な活用手段の一つとして、映像から何らかの知識を自動的に抽出するというアプローチがある。特に放送映像は、視聴者に何らかの情報を伝えるべく撮影・編集されたものであるため、それ自体に意図や意味が含まれていると考えるのが自然であり、そこから得られる知識には有用性があると考えられる。

本発表では、放送映像に登場する人物に着目し、それらの人物同士が形成する相関関係を抽出することにより、1 に示すような人物相関グラフを構築する手法を検討する。

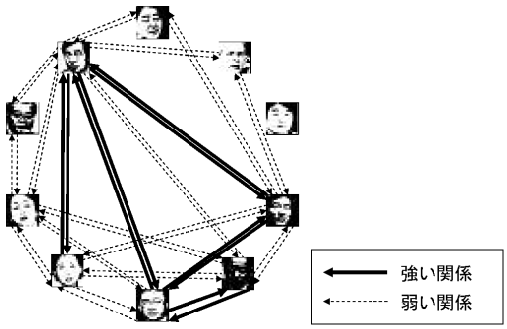

図 1: 人物相関グラフの例

1.2 用語の整理

ここで、映像に関する用語を整理しておく。

本研究で扱う放送映像とは、動画像、音声、文字放送字幕の集合体である。ここで言う文字放送字幕とは、本来聴覚障害者のために放送に埋め込まれた、音声を書き下したテキストデータである。クローズドキャプションとも呼ばれ、近年多くの番組に付与されるようになっている。

連絡先: 小笠原 崇，名古屋大学大学院情報科学研究科，愛知県名古屋市千種区不老町，(TEL)052-789-3310，(E-mail)toga@murase.m.is.nagoya-u.ac.jp

また、映像は画像的には以下のような構成になっている。

- フレーム：動画像の最小構成単位である静止画像
- ショット：画像的に連続するフレーム群
- シーン：意味的に連続するショット群。例えばニュースにおいてはトピック（話題）に相当
- カット：ショット間の不連続点

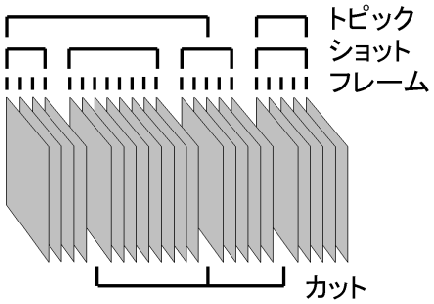

図 2: 映像の構成

2. 放送映像からの人物相関抽出

2.1 映像を用いる意義

情報資源から実社会における人間関係を自動的に抽出する研究は、今までにも広く行われてきた。しかしながらそれらの研究では、人名やそれに準ずる何らかのラベルが、個人を識別するものとして明示的に与えられている必要があった。つまり逆に言えば、明示的にラベルが与えられていない人物に関しては、全く抽出の対象としていなかったのである。これに対し映像では、クローズドキャプションなどのテキスト情報に加え、画像情報も用いることができる。そこで本研究では、画像情報を用いることにより、テキスト等により明示的に言及されない人物も含めた人物相関グラフを構築することを目指す。

ここで、放送映像においてテキスト等により明示的に言及されない人物が重要であるのか、という疑問が生じる。この疑問に対しては、以下の例を示したい。

ある日のニュースにおいて、政治家 A が取り上げられるとする。その際 A を撮影した映像に、秘書 B（音声やクローズドキャプションなどニュース原稿をもとにしたテキストでは通常 B は言及されない）が映っていた。数年後に B が政治家になりニュースに取り上げられたとしても、通常の手法では、B について明示的に言及されていないので、テキスト中で A と B の間の関係は抽出できない。

このように、ある時点でテキストでは言及されないような人物も、人物相関グラフを構築するうえでは重要で、そのような関係を抽出できる点が映像中の画像情報を用いることの利点と考えられる。

しかしながら、画像は解析そのものがテキストに比べ難しく、抽出した情報の正確性に欠けるという欠点もある。そこで本発表では、画像情報を用いる手法がどの程度実践可能で、テキスト情報に基づいた抽出を補完しうるかを調べた結果を報告する。具体的にはニュース映像を素材として、テキスト情報を用いた抽出と画像情報を用いた抽出の結果を比較する。

2.2 画像情報を用いた抽出

画像情報を用いた抽出には、放送映像の構成をふまえた処理が必要となる。ここでは素材としてニュース映像を用いるため、ニュース映像の構成に則した処理を行う。

トピック内の時系列的な共起関係

ニュース映像において一つのトピック（話題）は、図 3 のように、何度か現れるキャスタショットの前後でグループ分けすることができる。最初はそのトピックを紹介するキャスタショットから始まり、続いてそのトピックの核となる人物が登場する（グループ 1）。その後、キャスタショットをはさむごとにそれぞれに続く形でグループ 2、3 が現れる。これらの人物は、そのトピックに直接的な関係はないものの、トピックを補足・展開する役割を果たす人物である。

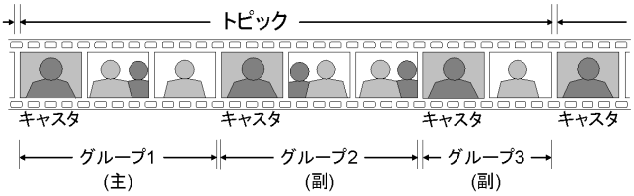

図 3: ニュース映像におけるトピック内グループの構成

このように、グループ 1 とそれ以外のグループにはそれぞれ異なった役割がある。これをふまえ、ここではグループ 1 を ‘主グループ’、それ以外のグループを ‘副グループ’ と呼び、トピック内で共起する人物の組がそれぞれどちらに属するかによって、その共起関係を分類することにする。ただし、ニュース映像において主として映したい人物は画面中央に大きく映るように撮影されるため、あるショットにおいて複数の人物が映っているときは、最も画面中央に近い人物を ‘主たる人物’ とみなして、そのショットを代表させる。

ショット内の物理的な共起関係

上記のトピック内の関係では、各ショットにおいて複数の人物が出現していても、主たる人物以外（以下、‘従たる人物’ とする）は無視していた。トピック内の関係を考える際には従たる人物について考慮する必要がなかったが、同一のショットに映っていた、すなわち同じ物理的空間に存在していたという事実は、本来極めて重要な関係と考えられる。よって、トピック内の関係とは別に、ショット内の共起性による関係を定義する。

関係の強さ

以上の ‘トピック内の共起関係’ と ‘ショット内の共起関係’ という 2 つの観点から見た関係を考慮して、ニュース映像内の 2 人物の組の関係の強さを求めていく。関係は、2 人物の各々に注目し、自分と相手がトピック内・ショット内でどのような関係で共起しているかにより分類し、その関係の強さを重みとして計算する。以下の実験では表 1 のように、直感的に関係の強いと思われる順に 5 から 1 の重みを単純に割り振った。

表 1: 2 人物の共起関係

関係	自分	相手	重み
トピック内の関係	主グループ	主グループ	5
	主グループ	副グループ	4
	副グループ	主グループ	3
	副グループ	副グループ	2
ショット内の関係	2 人物は同一ショット内で共起		1

処理の流れ

以下のような流れで、画像情報を用いた人物相関グラフ構築を行う。入力される ‘ニュース映像群’ は予めトピックに分割されているものである。

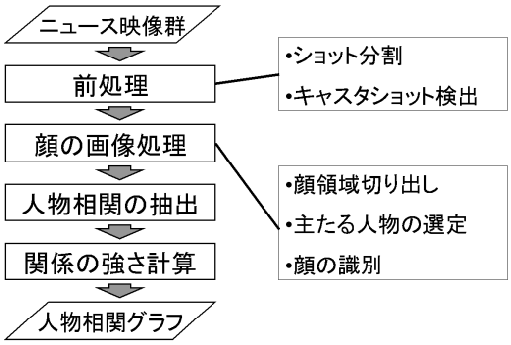

図 4: 処理の流れ

‘ショット分割’ は、時系列的に連続する 2 フレームの類似度比較にてカットを検出することで行ない、‘キャスタショット検出’ は、そのショットの先頭フレームと、各トピックの先頭フレーム（一般にトピックはキャスタショットから始まるため）との類似度を比較することで行なった。両処理とも画像特徴量として RGB 色ヒストグラムを用い、類似度の評価尺度としては 2 フレームの画像の特徴量ベクトルの角度の余弦を用いた。

‘顔領域の切り出し’ は、各ショットから抜き出した 1 フレームに対して [Lienhart 02] のオブジェクト検出手法を適用し、‘主たる人物の選定’ は、各フレーム内で切り出した顔領域のうち、最も画面中央に近いものを主たる人物とし、残りを従たる人物とした。

‘顔の識別’ については、その顔が誰の顔であるかという認識ではなく、どの顔とどの顔が同一のものかという、言わば ‘名寄せ’ ならぬ ‘顔寄せ’ ができればよい。そのための顔判別には固有顔を用いた手法 [Turk 91] などが適用できるものの、その精度は不十分であるため、後のテキスト情報を用いた抽出との比較のしやすさから、以下の実験では人手にて行っている。

3. 実験と考察

3.1 テキストからの抽出との比較

テキスト情報からのみの抽出、画像情報からのみの抽出の両手法（以下、‘テキスト抽出手法’ および ‘画像抽出手法’ と呼ぶ）を、同一のニュース映像に対して適用し、テキスト情報と画像情報とで抽出できる人物相関にどの程度違いが現れるかを調べた。

テキスト抽出手法としては、クローズドキャプションを用い、出現する人物名のトピック内での共起頻度を数えた。

用いた映像は、2001 年 3 月から 2002 年 1 月に実際に放送された計 17 日分のニュース映像のうち、後の評価・考察のしやすさを考え、政治を扱ったトピック 18 件とした。トピック長は短くて 1 分 30 秒程度、長いものは 8 分程度で、18 件の合計はおおよそ 70 分である。

トピック分割は [Ide 04] で提案されている手法等で自動化は可能であるが、本実験では人手にて切り出した。

実験結果

テキスト抽出手法では 32 人、画像抽出手法では 68 人（名前がわからない者 32 人を含む）の人物が検出できた。このうち、前者でのみ抽出できた人物は 5 人、後者でのみ抽出できた人物は 41 人（名前がわからない者 32 人を含む）である（図 5）。また、人物間の関係は、テキストで 312 エッジ、画像で 1,414 エッジ抽出できた。

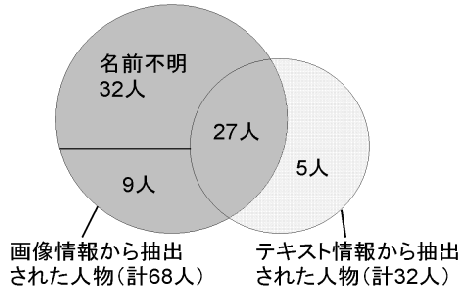

図 5: 画像・テキストから抽出された人物

考察

実験結果を見ると、図 5 のとおり、同一のニュース映像から抽出できた人物の数は、画像・テキスト双方で得られたものが 27 人であるのに対し、テキスト情報のみから得られたものが 5 人、画像情報のみから得られたものが 9 人と、画像による抽出数の方が多かった。さらに、2.1 の例で触れたとおり、ある時点での無名人（名前不明人）の潜在的な重要性をふまえると、名前が知られている人物の数より多い 32 人も得られたことは、興味深い。

以上のことから、放送映像がいかに画像にて暗に（すなわち、テキストには明示しないで）人物を紹介しているかを示すものであるかがわかる。これにより、人物相関を抽出するにあたって画像情報が必要であることが示唆された。

しかしながら、テキスト情報でのみ抽出された人物も存在する。これらの人物が画像情報からは得られなかった原因としては、次の 2 つのものが考えられる。

原因 1 画像抽出手法におけるショット分類や顔領域切り出しの精度の問題

原因 2 そもそも画像には現れない人物の存在

原因 1 に関しては、画像情報はその解析自体が難しいため、避けられない結果である。しかし、この問題については各処理手法の改良により、ある程度の改善は見込める。

原因 2 は、誰かが発言の中で少し触れた程度の人物で、かつ顔を画像として示さなくても視聴者がわかるような有名人に対して起こる。画像に全く出現していない以上、画像のみから抽出することはできない。これは、画像情報による人物抽出および、それに伴う人物相関の抽出が、テキスト情報を完全に代替することはできないことを示す。

3.2 人手による抽出との比較

前述の比較実験では、テキスト手法と画像手法とで、検出できる人物や抽出できる関係に差異があることが確かめられた。そこで次に、各手法による人物相関抽出がどの程度妥当であるかを調べるために、人手による人物相関抽出と両手法による結果とを比較する。

人手による抽出は以下の方法で行った。

1. 被験者を 2 グループに分け、一方には音声なしの（動画像のみの）ニュース映像、もう一方には音声ありのニュース映像（通常の映像）を見せる。
2. ニュースを見た印象として人物 A と関係が強いと思われる人物を、関係が強いと思われる順に 5 人挙げてもらう。
3. 各被験者の挙げた人物 1 位を 5 点、2 位を 4 点 … 5 位を 1 点として集計する。
4. 音声なし、音声ありのそれぞれについて、合計獲得点の多い順に人物を順位付ける。

用いた映像は政治を扱った 3 トピック（合計 18 分程度）で、人物 A は ‘中谷元’ という政治家とした。被験者は、本研究および ‘中谷元’ を知らない大学生 14 名（音声なし 7 名、音声あり 7 名）であり、各被験者には予めニュース映像中に現れる人物の顔と名前を載せたリストを渡しておいた。

評価実験は、人手による人物相関抽出を行うことから、3.1 に比べ小さな規模となっている。

実験結果

各手法で検出できた人物は、テキスト抽出手法で 5 人、画像抽出手法では 10 人（名前不明者 3 人含む）であり、人手による抽出では、‘音声あり’ で 10 人、‘音声なし’ で 11 人であった。

各手法において、抽出した関係の強い順に並べると表 2 のようになった。

表 2: 評価実験結果

テキスト 抽出手法	画像 抽出手法	人手 （音声あり）	人手 （音声なし）
小泉純一郎	菅直人	安部晋三	小泉純一郎
菅直人	小泉純一郎	小泉純一郎	菅直人
安部晋三	**田中真紀子**	山岡賢次	安部晋三
加藤紘一	（名前不明 1）	菅直人	加藤紘一
町村信孝	（名前不明 2）	**田中真紀子**	**田中真紀子**
	安部晋三	辻元清美	福田康夫
	町村信孝	福田康夫	山岡賢次
	市田忠義	児玉健次	町村信孝
	岡田かつや	加藤紘一	岡田かつや
	（名前不明 3）	町村信孝	市田忠義
			辻元清美

考察

実験結果を見ると、3.1 同様、画像抽出手法がテキスト抽出手法では抽出できなかった人物を抽出できていることがわかる。ここで特に注目したいのは、人手による抽出で上位に名前が挙がっている ‘田中真紀子’（表中太字）が、テキスト抽出手法では抽出されず、画像抽出手法でのみ抽出されていることである。このことは、無名時代まで遡らないまでも、現時点での人物相関を抽出するうえで重要な人物であっても、テキストに現れず画像のみで現れる場合があることを示す結果であると言える。

次に、各手法の情報源の対応から、テキスト抽出手法と ‘人手（音声あり）’、画像抽出手法と ‘人手（音声なし）’ をそれぞれ比較してみる。すると、図 6,7 のとおり、両者とも完全な包含関係になっていることがわかる。このことは、テキスト・画像両手法とも、人手で必要と判断されている人物をいくらか取りこぼしてはいるものの、人手で抽出されていない人物は抽出していない、すなわち、人の感覚と全くずれた見当違いな人物は抽出していないことを示すと言える。

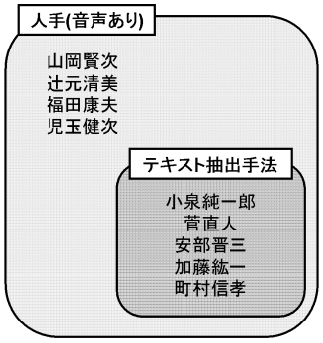

図 6: テキスト抽出手法と ‘人手（音声あり）’ の比較

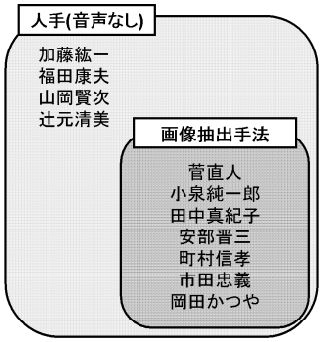

図 7: 画像抽出手法と ‘人手（音声なし）’ の比較

以上のことから、放送映像からの人物相関の抽出において、テキスト情報と画像情報は補完的な関係にあり、また、そのそれぞれは人間の感覚に沿った抽出を行いうる処理であることが確認された。これによって、テキストと画像の両方を使うことで、より充実した人物相関グラフが構築できるであろうことが示唆されたと言える。

4. おわりに

本発表では、放送映像に登場する人物同士が形成する相関関係を自動的に抽出し、人物相関グラフを構築する手法を提案した。放送映像において、文字情報字幕テキストと呼ばれるテキスト情報を容易に用いることができるようになっているなか、扱いにくい映像中の画像情報を用いることの意義を、実験を通して確認した。

本実験では ‘トピックの分割’、‘顔の識別’ については人手にて行ったが、‘トピックの分割’ については [Ide 04] にて自動化が可能であるし、‘顔の識別’ に関しても、[Turk 91] など数多くある顔認識手法を用いることにより自動化されうる。今後はこれらも含めた全ての処理の自動化を目指すとともに、テキスト情報と画像情報の連携や、映像中の位置関係と人物相関の関連を学習した重みの最適化などを行うことにより、さらなる改善を図る。

謝辞

研究に必要な数多くのデータを提供してくださった情報・システム研究機構国立情報学研究所に感謝する。研究に際し数多くのご助言をいただいた名古屋大学 友部博教氏に感謝する。本研究の一部は、文部科学省科学研究費補助金および 21 世紀 COE 研究費による。

参考文献

[Garfield 64] Garfield, E., Sher, I., and Torpie, R.: The use of citation data in writing the history of science, *Technical report, Philadelphia Institute of Scientific Infomation* (1964)

[Ide 04] Ide, I., Mo, H., Katayama, N., and Satoh, S.: Topic threading for structuring a large-scale news video archive, *Image and Video Retrieval -Third Intl. Conf. CIVR2004, Dublin, Ireland, Proceedings- P. Enser, Y. Kompatsiaris, N.E. O'Connor, A.F. Smeaton, A.W.M. Smeulders eds., Lecture Notes in Computer Science*, Vol. 3115, pp. 123–131 (2004)

[Kautz 97] Kautz, H., Selman, B., and Shah, M.: The Hidden Web, *AI Magazine*, Vol. 18, No. 2, pp. 27–35 (1997)

[Lienhart 02] Lienhart, R. and Maydt, J.: An extended set of Haar-like features for rapid object detection, *Proc. IEEE ICIP 2002*, Vol. 1, pp. 900–903 (2002)

[Turk 91] Turk, M. and Pentland, A.: Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71–86 (1991)

[安田 97] 安田 雪：社会ネットワーク分析 -何が行為を決定するか-, 新曜社 (1997)

[井手 99] 井手 一郎, 山本 晃司, 浜田 玲子, 田中 英彦：ショット分類に基づく映像への自動的索引付け手法, 電子情報通信学会論文誌 (D-II), Vol. J82-D-II, No. 10, pp. 1543–1551 (1999)

[井手 01] 井手 一郎, 佐藤 真一：人物関係に基づくニュース映像の検索と閲覧, 電子情報通信学会パターン認識とメディア理解研究会技報 PRMU2001-48 (2001)

[松尾 05] 松尾 豊, 友部 博教, 橋田 浩一, 中島 秀之, 石塚 満：Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol. 20, No. 1, E (2005)