画像・言語情報の統合的利用による映像データの自動的インデクシングの試み

井手 一郎, 田中 英彦
{ide, tanaka}@mtl.t.u-tokyo.ac.jp
東京大学大学院 工学系研究科*

1 はじめに

1.1 研究の背景

昨今の放送媒体の多様化やチャンネル数の増加に伴い、日々放送される映像の量は増大している。しかし、これらの情報への索引付け (インデクシング) は、一部の番組に対する簡単なものが新聞の TV 欄や TV 雑誌に掲載されている程度である。情報の再利用や映像の検索の需要という点から、本格的な検索ができるようにデータベース化されていることが最も期待されるニュース番組にいたっては、内容のリアルタイム性も手伝って、このような索引付けはほとんど行なわれていない。

従来から新聞記事のデータベース化に関しては様々な研究が行なわれ、すでに検索サービスがパソコン通信などで提供され、商用的にも実用段階にある。しかし、TV ニュースへの索引付けは、画像や音声認識の困難さにより、あまり自動化が進んでいないのが現状である。無論、CMU の Informedia プロジェクト [Info] をはじめ、様々な試みが行なわれているが、言語情報を統計的に処理してキーワードを抽出したり、話題の最初の字幕をそのまま索引として利用したりするにとどまっている。これらの手法は、画像内容とのきめ細かな対応付けが行なわれておらず、また言語情報の利用も本格的とは言いがたい。

1.2 研究の目的と概要

そこで本研究では、画像処理は必要最小限の特徴量抽出にとどめる一方で、映像内容に対してきめ細かな索引付けを行なうために、具体的な内容の認識に言語情報を積極的に利用することを考える。そして、このように画像情報と言語情報を統合的に利用することによる自動的な索引付けの有効性を示すことを目指す。

具体的には、動画像を簡単な画像的特徴量に基づいて分類し、各分類に応じて異なる索引付けを、字幕を中心とした言語情報から抽出する。このような索引付けを行なうことにより、ニュースにしばしば登場する典型的なショットの中から特定の条件に合致するものを検索する際に、前後の余分なショットを省き、実際にその行為を行なっているもののみを提示し、かつ各分類毎に異なる検索内容に応じて付与するキーワードにより、きめ細か検索を可能にする。

本稿では、まずこのような処理を実現するシステムを提案し、実現に必要な要素技術の紹介および各技術の性能評価を行なって問題点を明らかにし、最後にそれらをまとめたシステム全体の実現性を探る。

1.3 動画像の構成と用語の定義

一般に動画像は図 1に示すように階層的に構成されているとみなせる。この構成には、「フレーム」「ショット」「(画像的) シーン」「番組」という純粋に画像的な階層と、その上に重なり合うような、「意味的シーン」「話題」「番組」という意味的な階層の 2 通りがある。

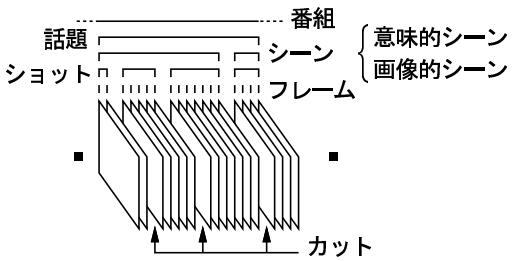

図 1: 動画像の構成と用語の定義

また、「映像」とは、動画像や音声を含めて放送されるあらゆる情報の集合を指すこととする。

2 ショット分類に応じた索引付け機構

2.1 索引付け機構

図 2に、本稿で提案するシステムの概略図を示す。図中色が薄くなっている部分および破線で囲まれている部分は、現時点では実装していない。

2.1.1 画像処理

以下に、画像処理の流れを示す。

1. **動画像の電子化:** 計算機で動画像の処理を可能に
2. **カットの検出:** 画像的に連続した最小のまとまり (ショット) の把握

---

*"An Attempt to Automatically Index Motion Images by Integrated Processing of Image and Language Information"
Ichiro Ide, Hidehiko Tanaka
University of Tokyo, Graduate School of Engineering
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

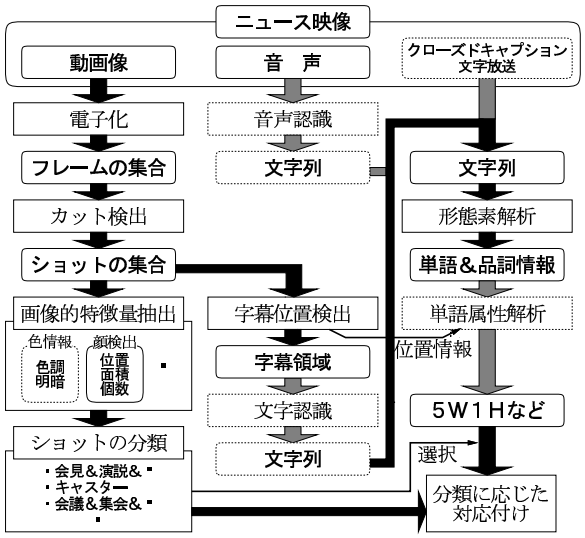

図 2: ショット分類に応じた索引付け機構

3. **画像的特徴量の抽出**: 色調, 明暗などの色情報や、画像中の顔領域の位置, 大きさ, 個数など、画像のみから得られる特徴量を抽出
4. **ショットの分類**: 得られた画像的特徴量の組合せからショットを典型的なショットに分類

ここで重要なのは、典型的なショットへの分類は画像的特徴量のみにより自動的に行なうことである。本稿では、典型的なショットとして次のものを取り上げる。

- **[会見ショット] 会見&演説&報告&インタビュー**
  1, 2 人の人物が中心部に比較的大きく映っている

上記の処理以外に、字幕を抽出するために、1., 2. の後に次の処理を行なう。

3. **字幕の検出**: 字幕の存在と位置を検出
4. **文字認識**: 検出した字幕を認識し、言語処理に利用

2.1.2 言語処理

以下に、言語処理の流れを示す。これらに基づき、各ショット分類に応じて必要なキーワードの候補を用意する。

1. **形態素解析**: JUMAN[JUMAN] を利用
2. **単語の属性解析**: キーワードの抽出に利用

2.1.3 情報の統合: 対応付け

2.1.1 で画像的特徴量に基づいて分類したショット毎に、**2.1.2** で属性解析を行なったキーワード候補の中から必要な属性と一致するものとの対応付けを行なう。

3 画像処理

3.1 画像の電子化

計算機により画像処理を行なうためには、まず始めにアナログの動画像を、ビデオキャプチャボードを利用して電子化し、フレームという静止画像の集合にする必要がある。以下の実験で使用する動画像を電子化した際の諸条件を表 1 に示す。日本や米国の TV 放送で採用されている NTSC 形式の動画像は約 30fps であるが、必要なディスク領域および計算機の処理能力に限界があり、本研究には 5fps で十分と判断した。

項目	条件
キャプチャボード	SUN Video Card
大きさ	横 320 × 縦 240
色数	14,777,216 色 (24bit)
保存データ形式	UYVY 非圧縮 [SunVideo]
標本化レート	5 frame/秒 (=fps)
番組名	NHK ニュース 9
放映日時	平成 9 年 1 月 9 日
総時間数	約 30 分 (8,930frame)

表 1: 動画像電子化の際の諸条件

3.2 カット検出によるショット分割

次に、フレームをショットという連続した画像の集合にまとめる必要がある。そのために、ショット間の境界であるカットを検出するが、本研究では、比較的雑音に対する耐性があり高性能と言われている、分割χ^2検定による検出法 [長坂 92] を採用した。

本手法は、画像を (本研究では、縦 4×横 4=16) 分割し、分割された各領域内において、次のフレームの同一領域との色ヒストグラムの変化をχ^2 検定法により定量化し、閾値を越える領域が一定数 (本研究では 8(/16)=半分) に達すると、カットが発生したとみなす。(非分割)χ^2検定による手法と較べて、局所的な変動に影響されにくいという特徴がある。

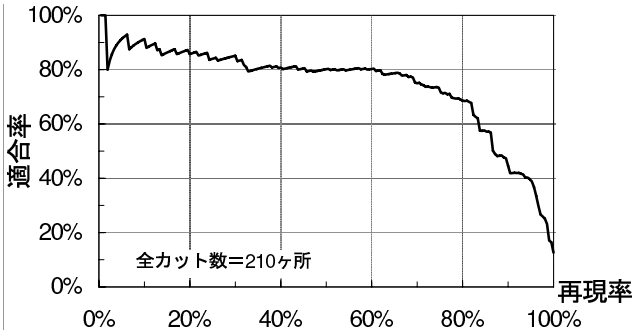

図 3: χ^2検定によるカット検出法の評価

図 3に、分割χ^2検定によるカット検出法を表 1の動画像に適用した結果の再現率と適合率の関係を示す。この結果を見ると、再現率が 60%以下で十分とすれば、80%以上の適合率が出ている。しかし、後段への影響を防ぐためには、この段階でより正確に検出する必要がある。誤検出カットの前後では、フラッシュが画像全体を照らしたり、画像が急激に変化する、という現象が観察された。カット検出でこれ以上の性能向上は見込みにくく、現時

点では画像的に似たショットの集合である画像的シーンを検出するアルゴリズムの採用により、誤検出カットの前後を再びつなぎ合わせることを考えている。

3.3 顔領域の検出とショット分類の例

3.3.1 顔領域の検出

画像中の顔領域を検出する研究は比較的古くから行なわれており、複雑なモデルを用いた手法から単純なパターンマッチングにいたるまで、様々な手法が提案されている。しかし、複雑な手法は計算量が膨大になり、本研究のように多数の画像を解析するような場合は、なるべく計算量が少ない手法を利用したい。

そこで、本研究では色情報に基づいて肌色領域を検出し [佐々木 91]、その領域に外接する矩形の辺の長さの縦横比が一定の範囲内 (0.7〜2.0) のものを検出する手法を採用する。肌色領域の検出の際には、人間の色彩感覚に近い HSV (H:色相, S:彩度, V:明度) 色座標系のうち、明るさに無関係な H と S のみを利用している。

3.3.2 ショット分類の例

2.1.2 で述べたように、本稿ではショット分類の例として、1,2 人の人物 (顔) が画像の中心的領域に存在するものを会見ショットとして分類することを試みる。具体的には、各ショットの第 1 フレームについて、図 4に網かけで示すような範囲に重心が含まれる肌色領域が 1, 2 個のものを検出する。

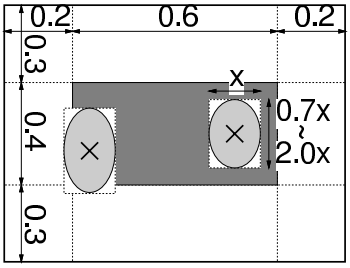

図 4: 「会見ショット」の検出条件

このような条件に基づいた検出実験の結果を表 2に示す。なお、会見ショットは全部で 47 ショット存在する。

検出成功		39	
検出洩れ		8	再現率 83%
誤検出	顔でない領域の誤検出	6	
	図 4の検出条件の問題	8	適合率 64%

表 2: 会見ショット検出実験の結果

このように、単純な画像的特徴量を利用しているにもかかわらず、かなりの性能が得られた。検出洩れは、画像が暗すぎたり図 4の条件が若干不適切であることが原因であった。誤検出の後者の原因は、検出条件を検討すれば除去できそうである。

3.4 字幕の検出

TV 放送には表 3に示すような様々な形態の言語情報が付随している。これらの言語情報を計算機上で処理するためには、電子化する必要がある。そのため、表 3の分類のうち、音声の場合は音声認識を、画像の場合は文字認識を行なう必要がある。現時点では、前者は依然実用段階にはないと考えて利用しないが、後者に関しては、1) オフライン文字認識が実用化されていること、2)TV 放送で用いられている字幕は均整のとれた活字体であること、により利用することを考えている。

形態	例
音声	主音声, 副音声
画像	字幕
電子	文字放送, クローズドキャプション

表 3: TV 放送に付随する言語情報

しかし、文字認識を行なう前に字幕の存在とその位置を特定する必要がある。特に後者は、番組に応じてある程度決まっている画面構成に関する知識により付加的な情報も得られる [渡辺 96] ため、重要である。

字幕はカメラで撮影された画像の上に CG を利用してスーパーインポーズしているため、輪郭のエッジが非常に強く立っている。そこで本研究では、画像を横 4×縦 3=12 に分割し、分割された各領域内の各画素におけるラプラシアンと呼ばれる値 [画像 HB] により、エッジの強さを定量化し、閾値を越える画素の割合が一定値を越えると、字幕が存在すると考える。図 5に、本字幕検出法を表 1の動画像に適用した結果の再現率と適合率の関係を示す。この結果を見ると、適合率が最大で 40%と、誤検出が多すぎて実用に耐えない性能であることが分かる。実験結果の数値が上位の領域を見ると、誤検出の原因は、砂浜の拡大, 木の梢, 縞模様の洋服など、細かな模様になっている場所の無数にあるエッジが検出されていた。今後、このような細かな模様のエッジがショット内ではあまり変動しないのに対して、字幕がショットの途中で登場・消滅する際に急激なエッジの変化が見られることを利用することを考えている。

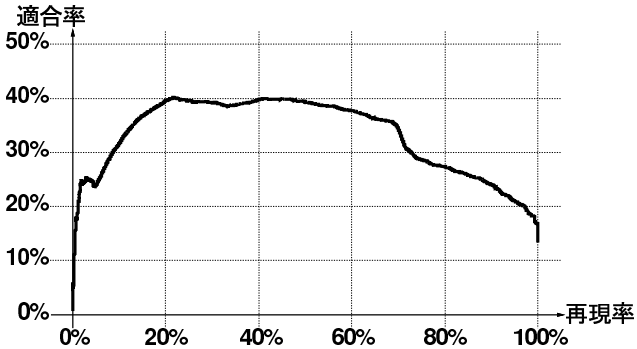

図 5: 字幕位置検出法の評価

4　言語処理

4.1　字幕の分類

表 4に、TV ニュースに登場する字幕の分類と、表 1の動画像中の 130 項目の字幕の各分類毎の割合を示す。

#	分類	割合
1	映像の撮影された場所や時間を示すもの	37%
2	人間を指すもの (組織+氏名+敬称など)	20%
3	抽象的事態の説明 (冒頭のタイトルなど)	12%
4	発言内容そのもの (外国語を翻訳したものなど)	9%
5	その他番組製作に関する情報 (報告, 中継など)	9%
6	映像の具体的な状況説明	8%
7	発言内容を要約したもの	5%

表 4: TV ニュースにおける字幕の分類 (一部 [角田 96])

このうち 1, 2 はそのまま、3, 6, 7 は簡単な抽出処理を経てキーワードになり得る。また 5 は同時に画面中に存在する他の字幕の扱いを制御するのに利用できる。

このような字幕の各分類の特徴に関しては [渡辺 96] に詳しい解析例がある。

4.2　字幕に基づくキーワードの抽出

会見ショットを検索する際には、その会見に関する以下のようなキーワードに基づくと考えられる。

1. **場所情報:** 地名, 会議名など ← 表 4の#1 より
2. **時間的情報:** 年月日, 時刻など ← #1 より
3. **会見者:** 人名, 肩書, 組織名など ← #2 より
4. **会見内容** ← #3, #6, #7 より

ここで、どのようにして字幕に登場する単語に意味的な属性付けを行ない、表 4のような分類を行なうかが問題である。JUMAN の解析結果からは、品詞および若干の文法的属性が得られるだけであり、ここで必要とするような意味的な属性は得られない。例えば、JUMAN は「小沢党首」を「固有名詞+普通名詞」と解釈するが、これだけでは、この語が人名か地名か分からない。

また、分類語彙表 [分類] や EDR 電子化辞書などの辞書は概念的な分類が中心なため、意味的な用法に基づいた属性付けには使いにくい。上記の例の「党首」は分類語彙表では、「1.243:長」に分類されるが、この中には「飢鬼大将」という語もあり、「固有名詞+『1.243:長』」が人名であると規則化するのは無理な上に、「大統領」や「容疑者」といった単語は別の項目に分類されている。

そこで、ここで必要な情報を得るためには、普通名詞がどのような固有名詞と一緒に用いられるかなどを記述した辞書が必要になる。本研究では言語処理を研究することが主目的ではないので、当面は字幕に登場した単語の意味的な属性を記述した辞書を自作して使用するが、このような情報を記述した辞書の登場が期待される。

5　おわりに

本稿では、画像情報単独では得られないキーワードに言語情報を、言語情報単独では得られないショットの分類に画像情報を利用することにより、従来ならば膨大な知識ベースを用意しなければならない処理を、可能な限り放送されている情報のみを用いて行なう手法を提案し、特に画像処理に関して基礎的な解析手法の評価を行なった。言語処理に関しては、実際のニュース番組に登場する字幕の解析を行ない、キーワード抽出を行なうために、意味的属性を記述した辞書の必要性を示した。

今後は、画像処理に関しては、全般的な性能の向上と、扱える画像的特徴量 (明暗や色調など) とそれを用いて分類できるショットの種類 (顔領域が画像中に 3 箇所以上ある場合は会議とみなすなど) を増やす。一方言語処理に関しては、字幕に登場する単語の意味的属性を記述した辞書および解析規則作成を行ない、システム全体を動作させることを目指す。

謝辞

本研究の構想は京都大学工学部の角田達彦助手による懇切丁寧な御指導に負うところが大きく、この場を借りて深謝する。

また、画像処理の基礎的事項に関して教えて下さった上坂英樹氏、自然言語処理に関する有益な示唆を与えて下さった永松健司氏、実装に協力して下さった山本晃司氏に感謝する。

なお、自然言語解析には日本語形態素解析システム JUMAN 3.0β を使用させていただいた。

参考文献

[角田 96] 角田 達彦, 大石 巧, 渡辺 靖彦, 長尾 眞; 「キャプションと記事テキストの最長一致文字列照合による報道番組と新聞記事との対応づけの自動化」; 情報処理学会技術研究報告 96-NL-115, Vol.96, No.88, pp.17-24, Sep. 1996.

[渡辺 96] 渡辺 靖彦, 岡田 至弘, 長尾 眞; 「TV ニュースで用いられるテロップの意味解析」; 情報処理学会技術研究報告 96-NL-116, Vol.96, No.89, pp.107-114, Nov. 1996.

[長坂 92] 長坂 晃朗, 田中 譲; 「カラービデオ映像における自動索引付け法と物体探索法」; 情報処理学会論文誌 Vol.33, No.4, pp.543-550, Apr. 1992.

[佐々木 91] 佐々木 努, 赤松 茂, 末永 康仁; 「顔画像認識のための色情報を用いた顔の位置合わせ法」電子情報通信学会技術研究報告 IE91-2, pp.9-15, Apr. 1991.

[画像 HB] 高木 幹雄, 下田 陽久; 「画像解析ハンドブック」; 東京大学出版会, Jan. 1991.

[分類] 国立国語研究所; 「国立国語研究所言語処理データ集 5 分類語彙表 [フロッピー版]」; 秀英出版, Dec. 1993.

[JUMAN] 松本 裕治, 黒橋 禎夫, 宇津呂 武仁, 妙木 裕, 長尾 眞; 「日本語形態素解析システム JUMAN 使用説明書 version 3.0 Beta」; Jul. 1996.

[SunVideo] "SunVideo User's Guide", pp.125-128; Sun Microsystems, Inc., 1994.

[Info] "The Informedia Project"; http://www.informedia.cs.cmu.edu/.