副音声利用による映像への索引付け

Video Indexing based on Sub-Audio Content

タンティッケ　　井手一郎　　田中英彦

Tan Tit Keat　　Ichiro Ide　　Hidehiko Tanaka

東京大学大学院　工学系研究科

Graduate School o Engineering, The University o Tokyo

Indexing is necessary or reutilising and extracting specific in ormation rom the enormous volume o video data that is increasing rapidly. Image and text data, by using image and language processing techniques, have been employed in various video-indexing research. However, the conventional indexing techniques are still ar rom ully practical utilisation. In this paper, we propose and assess a new video-indexing technique, by employing the sub-audio data which is originally used to enhance the blinds' understanding o the video content. We also propose a simple method to extract the ollowing 4 basic keywords, "location", "subject", "action" and "object" rom the sub-audio text.

1　はじめに

近年、放送される映像量の増大に伴い、これらの情報の再利用や検索に必要な索引付けの需要が高まっている。従来、映像データへの索引付けは、画像認識・字幕解析や台詞の利用によるものがあった。研究例としては、映像の内容を PDS(Program Description Script) と呼ばれる方式により記述し、ユーザーからの自然言語による要求を自然言語処理によって解析し、PDS の検索を行うことによって、希望する場面の表示を行うものがある [1]。しかし、これは放送局側があらかじめ情報を付与しておくことを前提とした研究であり、PDS の作成はほとんど手作業であり、映像量の増大に追い付くのは困難である。

一方、Carnegie Mellon 大学を中心に行われている Informedia プロジェクトは、デジタルライブラリの構築を目指したプロジェクトであり、この中で、音声, 言語, 画像情報を統合した認識が行われている [2]。まず、ビデオの音声は高精度の音声システムによってテキストに変換され、全文テキストデータとして蓄積される。映像情報は、画像処理技術によって認識され、音声から得られたテキストとの対応付けがなされる。しかし、このシステムは索引付けの際に画像内容にあまり深く踏み込んでいないという問題点がある。

本研究では視覚障害者のための副音声放送を利用し、副音声から必要な情報を抽出する手法と、それを画像情報に対応付けする手法を提案する。日本語番組を対象にして、副音声の利用によって受信側で画像内容に踏み込んだ索引付けを行うことを目指す点で、本研究はこれらの研究とは趣を異にする。

2　副音声利用による索引付け

本研究のシステム構成を図 1に示す。

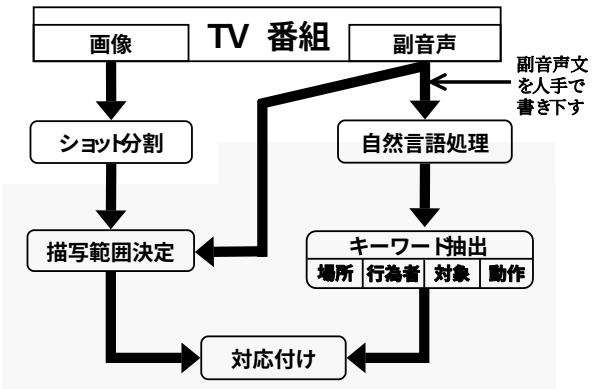

図 1: システム全体の概要

2.1　既存技術を利用した処理

2.1.1　動画像の電子化

動画像をビデオキャプチャボードでフレーム単位に電子化した際の諸条件を表 1にまとめる。

項目	条件
キャプチャボード	SUN Video Card
大きさ	横 320 × 縦 240
色数	16,777,216 色 (24bit)
保存データ形式	UYVY 非圧縮 [3]
標本化レート	5fps (frame per second)

表 1: 動画像電子化の際の諸条件

2.1.2 カット検出

副音声と映像の同期調整を行うために、カットを検出する必要がある。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検定によるカット検出法 [4] を採用した。

2.1.3 音声認識

TV 番組によっては副音声が主音声と混じった状態で放送されることもあるため、単純に音声認識を既存の手法で行うことはできない。また将来はこのような情報が電子的な形で付加されて放送されるもとも考えられる。このため、本研究では副音声の認識は自動化せず、人間が書き下し、完全に行えたものとみなすことにする。

2.1.4 形態素解析

副音声文からの情報抽出を自動化するために、まず形態素解析を行う必要がある。本研究では、形態素解析に日本語形態素解析システム JUMAN version 3.0 β[5] を利用した。

2.2 本手法独自の処理

2.2.1 副音声に対する自然言語処理

日本語形態素解析システム JUMAN は語義の解析までは行わないため、副音声からの情報抽出を自動化するために、さらにいくつかの工夫が必要である。

「人物名詞」: 人物を表す名詞　JUMAN は、「母」,「警察」,「判事」などのように人間を指す語を普通名詞と判断する。2.2.3で行う「行為者」情報の抽出のためには、これらを他の「花」,「車」などのような普通名詞と区別する必要がある。表 2に挙げる 4 本の NTV-4 の「火曜サスペンス劇場」ドラマの副音声から、人間を指し得る普通名詞と判断したものをサンプルとして集め、分類語彙表 [6] の利用により、これらのサンプルと同じ分類にあり、かつ人間を指し得ると判断した全ての普通名詞を、「人物名詞」として収集した。他のドラマに対して実験を行う際には、JUMAN によって普通名詞と出力され、かつ収集した「人物名詞」の中の語と一致するものについて、その形態素情報を「普通名詞」から「人物名詞」に置き換える。また、後段の処理を簡単にするために、JUMAN によって「人名」と出力された語 (JUMAN 固有名詞辞書中にある人名) に対しても、その出力を「人物名詞」に置き換える。

放送年月日	番組名
1997 年 08 月 12 日	犯罪心理分析官 3
1997 年 09 月 02 日	転勤判事
1997 年 10 月 14 日	わが町 IX
1997 年 10 月 28 日	待っている妻

表 2: データ収集を行った TV 番組

「場所名詞」: 場所を表す名詞　JUMAN は、「家」,「部屋」などのように、場所を指す語を普通名詞と判断する。よって、「人物名詞」と同様に、収集した辞書に基づき、このような語の形態素情報を「普通名詞」から「場所名詞」に置き換える。また、後段の処理を簡単にするために、JUMAN によって「地名」と出力された語 (JUMAN 固有名詞辞書中にある地名) に対しても、その出力を「場所名詞」に置き換える。

形態素解析により分割された語の復元　形態素解析によって、本来は 1 つの語がいくつかの形態素に分割されたものについて、正確な対応付けを行うために、表 3に示すように、再び結合して復元する必要がある。結合された語の形態素情報は元の 2 番目の語に合わせる。なお、普通名詞にはサ変名詞も含む。

前後の品詞	普通名詞・人物名詞	
例文	野球 選手	普通名詞 人物名詞
出力	野球選手	人物名詞

表 3: 複合語の復元の例

2.2.2 副音声の描写範囲の決定

本研究では、副音声のないショットを考慮に入れず、副音声のあるショットのみを対象に索引付けを行う。なお、3.1で記す 2 本の実験用のドラマにおいて、全ショット数の 50.4%に副音声が存在する。副音声を利用してショットへ索引付けを行うには、まず各々の副音声がどのショットに、そしてそのショットのどの部分に対応しているか、すなわち副音声の描写範囲を決める必要がある。

単一のショット内に一本の副音声しか含まれない場合は、そのショット全体を描写範囲とする。また、単一のショットに複数の副音声が含まれる場合は、図 2に示すように、最初の副音声の描写範囲をショットの先頭から次の副音声の発声の始点の直前までにし、2 番目の副音声の描写範囲を発声始点からショットの終点までにする。

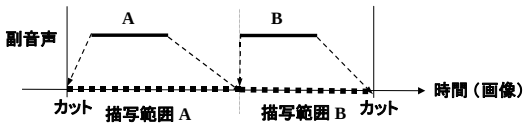

図 2: 描写範囲の決定

また、副音声が 2 つのショットに跨る場合は、その副音声の長さ (何フレームにわたるのか) を調べ、半分以上が後のショットに含まれるなら、図 3のように、その描写範囲を次のショットの先頭から始まることにする。逆に、半分以上が前のショットに含まれる場合は、その描写範囲は前のショットに含まれると考える。

3.2 カット検出

分割χ^2検定法を用いてドラマ 1 に対してカット検出を行ったが、再現率を約 80%に設定すると適合率は約 80%になる。なお、再現率と適合率は各々次のように定義した:

$$\text{再現率 [\%]: } \frac{\text{実験結果の正解数}}{\text{真の正解数}} \times 100$$

$$\text{適合率 [\%]: } \frac{\text{実験結果の正解数}}{\text{実験結果の総数}} \times 100$$

動きの激しいショットにおいて、正確なカットが検出できなかったのが誤りの主因と考えられる。このようにカット検出は依然実用的なレベルにはないため、次の対応付けの実験を行う際は、カット検出をあらかじめ手動で行ったデータを用いる。

3.3 対応付け

副音声から抽出された情報の描写範囲への対応付けの評価を行う。副音声文から正しい情報を抽出でき、かつ、その描写範囲が正確に定められたものを正解とする。よって、真の正解数は各情報における真の描写範囲数となる。実験評価を表 5にまとめた:

項目	場所	行為者	動作・対象
再現率	92.3%	94.1%	93.8%
適合率	92.0%	86.5%	86.1%

表 5: 対応付けの評価

対応付けの誤り率の原因は、副音声の発声タイミングとカットが大きくずれことや、場所名詞や人物名詞の認識の失敗などである。

4 おわりに

本研究では、顔や物体などの画像認識を用いていないため、あらかじめ膨大な知識ベースをもつ必要もなく、TV 放送の受信側における映像への索引付けに利用できると考えられる。また、再利用価値の高い教育番組への応用も期待している。

しかし、全てのショットに対して完全に索引を付けることができないのが現状である。これを改善するには、次の 3 つの方針が考えられる:

1. 画像処理技術を利用し、ショット間の類似度を分析することによって、副音声のあるショットの索引を、近隣の副音声のない画像的に類似したショットに適用することが考えられる。
2. 本手法では、決定された副音声の描写範囲 (画像) の全てが抽出された情報に対応するのではなく、他の場面が映っている部分を含む可能性もある。より正確な描写範囲を決定するには、物体抽出や追跡 (Tracking) 技術が役に立つと思われる。追跡技術を使うことによって、どの「行為者」がいつからいつまでどのような「動作」を行うのかについてより正確に解析できるため、描写範囲の決定をより細かくできる。
3. 本研究では、音声認識は行わなかったが、本システムを実用的なものにするには、自動音声認識が必要である。今後、音声認識技術の向上や副音声のような情報が電子的な形で付加された TV 放送を期待したい。

結論として、副音声の映像への索引付けの有効性が本研究により示された。しかし、現段階では、全てのショットに索引付けすることはできないため、本手法を本格的に実用化するのは困難である。今後の課題として、より複雑な画像処理技術を利用することや、本研究では全く利用しなかった主音声・字幕などの活用によって、より完全な索引付けシステムを作り上げることが考えられる。

参考文献

[1] 柴田 正啓: 「映像の内容記述モデルとその映像構造化への応用」, 信学論, D-II, vol.J78 D-II, no.5, pp.754-764 (1995).

[2] T.Kanade, M.Mauldin, R.Reddy, M.Sirbu, S.Stevens, D.Tygar: “Informedia Digital Video Library”, http://www.informedia.cs.cmu.edu

[3] “SunVideo User’s Guide”, pp.125-128, Sun Microsystems, Inc.(1994).

[4] 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 (1992).

[5] 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」(1996)

[6] 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 [フロッピーディスク]」, 秀英出版 (1993).

[7] 古矢弘, 中井春男, 山崎誠也: 「中学国語総合便覧・全訂増補版」, 正進社, pp.122-134.