

文字認識と異種情報の対応関係に基づいた ニュース放送からの情報抽出

佐藤 俊雄[†] 金出 武雄^{††}

本論文では、映像にスーパーインポーズされている文字の認識結果とクローズドキャプションの対応関係を利用するビデオ情報の解析手法を検討する。人間もしばしば経験する異種情報による情報伝達の効果を、計算機によるビデオ情報の解析に適用し、その信頼性を向上させることを狙いとする。その1つとして、映像中の文字について誤認識された場合に、別の情報であるクローズドキャプションを利用して修正する効果を確認する。また、この文字認識結果とクローズドキャプションの両方に含まれている語を重要な情報と判断し、キーワードとして抽出する。さらに、このキーワードが含まれている映像のフレームとクローズドキャプションの文を選択し、ビデオ情報の要約を作成する方法を提案する。ニュース放送を題材にして、映像中の文字認識の性能を評価するとともに、クローズドキャプションという異種情報との対応関係による情報抽出について有効性を検証する。

Contents Extraction from News Video by Character Recognition and Associating of Multimodal Information

TOSHIO SATO[†] and TAKEO KANADE^{††}

This paper describes methods to recognize characters in video frames and methods to analyze video data associating multimodal information such as character recognition results and closed captions. We introduce two methods to analyze news video data based on its properties for viewers. First, an error correction method for the character recognition using closed captions is explained. Also, we extract a word included in both video frames and closed captions as a key word. Frames in the video data and sentences of the closed caption that include key words are extracted as key frames and key sentences, respectively, to make summaries of the video data. Experimental results for the character recognition, association between recognition results and closed captions, and extraction of key frames and key sentences are explained using seven news videos.

1. はじめに

ビデオ情報^{*}には、映像や音声などの複数の表現形態が含まれている。この複数の表現を用いた情報伝達では、個々の表現が伝える情報だけでなく、その対応関係が特別な役割を果たす場合がある。たとえば、ニュースのビデオ情報では、視聴者が、音声で聞き逃した情報を映像中の顔や文字で回復したり、同一内容を同時に複数の表現で伝えられたときに強い印象を受けるといった効果がある。これらは人間に対する働きかけとして編集者が意図したものであるが、本論文では、この対応関係を利用して計算機によるビデオ情報の解

析の信頼性を向上させることを検討していく。

ニュースは、再利用のニーズが高いことと、構造が比較的単純であることから、ビデオデータベースの研究の題材として注目されており^{1)~4)}、ビデオ情報における異種情報の対応関係の研究も検討されている。文献 4) では、クローズドキャプション^{**}における文のカテゴリと、検出できる顔の数による映像のカテゴリを求め、その間の対応づけによるビデオ情報の分解および表現方法を提案している。この研究では、情報の内容に基づいた対応づけが試みられているが、対応関係の種類が少なく応用が限定されるという問題が存在

[†] 株式会社東芝情報・社会システム社
Information and Industrial Systems & Services Company, Toshiba Corporation

^{††} カーネギーメロン大学ロボティクス研究所
The Robotics Institute, Carnegie Mellon University

^{*} 本論文では、“ビデオ情報”を、放送などから入手できる、映像、音声、クローズドキャプションなどの複数の表現が含まれている情報の意味で用いる。また、“映像”を動画像の意味で用いる。

^{**} クローズドキャプションは、米国の放送のほとんどに含まれているテキスト情報で、ニュースのビデオ情報の場合、音声のナレーション情報とはほぼ同一である。



図1 ニュース映像の文字の例。(a) 映像中の文字を含む1フレーム。(b) 拡大画像 (204 × 14 画素)。(c) 単純しきい値処理による二値化画像。(d) 従来の方法による文字抽出結果

Fig. 1 Characters in news videos. (a) A frame including superimposed captions. (b) Magnified image (204 × 14 pixel). (c) Binary image by simple thresholding. (d) Conventional results for character segmentation.

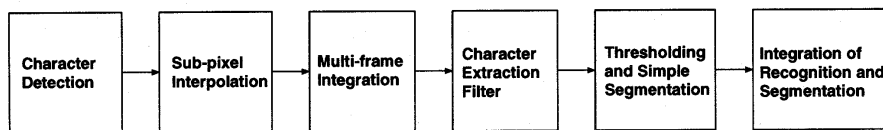


図2 映像の文字認識処理の手順
Fig. 2 Flow of character extraction for videos.

する。この主な原因は、映像側から抽出できるカテゴリの数が限られているためと考えられる。

映像のカテゴリ分けについては、静止画像データベースにおける検索の研究⁵⁾が参考になるが、報告されている有用なシステムのほとんどは、データベースに含まれる画像の種類や検索の目的を限定しており、カテゴリ分けのモデルがもともと構築しやすいものである。一方、ニュースではすべての事象が撮影される可能性があり、そのモデルを確立することは容易ではない。たとえば、静止画像データベースの検索に多く用いられるカラー画像のヒストグラムを特徴量と考えると、ニュース映像を分類することは困難に見える。

この問題を解決するアプローチとして、映像の中から特定の対象を抽出した後に、その部分領域のカテゴリを求めて映像全体のカテゴリとする方法が研究されている。そのような対象として、人物の顔やスーパーインポーズされている文字がある。文献 6) では、ニュース映像に含まれる顔を認識し、名前のカテゴリ分けという問題に単純化することで、画像から多くの分類を獲得している。同様に、映像中に含まれる文字の認識の研究も報告されている^{7)~9)}が、ほとんどが文字認識手法だけの議論であり、ビデオ情報における他の情報との対応関係は検討されていない。

我々も、すでにニュースを題材にして映像中の文字の認識によるインデックス獲得方法について報告している^{10),11)}が、本論文では、この文字認識結果とクローズドキャプションとの対応関係を利用したビデオ情報の解析手法を検討する。

1つは、映像にスーパーインポーズされている文字の認識について、クローズドキャプションを使って認識率を改善する方法である。また、映像とクローズドキャプションの両方に出現する語は、局所的に出現頻度の高い重要な情報と考えて、キーワードとして抽出する手法を検討する。さらに、このキーワードを含む情報として、映像におけるフレームとクローズドキャプションにおける文を抽出し、ビデオ情報の要約を作成することを提案する。

本論文では、2章で映像の文字認識方法の概略を説明し、3章では文字認識結果を示す。4章では、文字認識結果とクローズドキャプションの対応関係に基づいた、ビデオ情報からの情報抽出について説明する。

2. 映像の文字認識方法^{10),11)}

認識対象とする文字は、図 1 に示すようなニュース映像にスーパーインポーズされている英数字である。この種の文字サイズは 10 × 10 画素以下と解像度が低い。また、背景が文字に近い輝度値をとる場合も多く、従来の二値化による抽出では各文字を正確に切り出すことが困難である。

我々は、映像にスーパーインポーズされている文字を認識するうえでの課題は、低い解像度の克服と、複雑背景からの文字抽出の2点であると考え、図 2 に示す手順に従って、線形補間による解像度増大、複数フレームを用いたコントラスト向上、文字抽出フィルタ、および文字のセグメンテーションと認識の統合処理を適用して、文字認識率を改善することを検討して

きた。本章ではこれら手法の概略を説明する。

なお、図1とは異なるデザインとして、輝度の低い文字が輝度の高い背景とともにスーパーインポーズされる例も存在する。このような場合には、文字列を検出した後に文字の輝度を検知してネガポジ反転処理を適応的に追加する必要があるが、今回使用する映像に該当する文字が含まれていないので、本報告では検討しない。

2.1 文字列の検出

映像の文字認識を現実的な時間内に終了させるために、文字列を含むフレームと領域をあらかじめ検出して、処理の対象となるデータ量を減らしておく。映像における文字領域抽出は文献[12]で報告されており、本報告でもこの手法を採用する。

はじめに、各フレームに対して単純なマスク処理による横方向微分と二値化処理により垂直エッジ成分を抽出する。さらに、この結果に対して、左隣の1画素と右方向8画素を含めた10画素の中で5画素以上が抽出画素であれば、注目画素を抽出画素とし、横方向に融合させる。ラベリング処理により抽出画素の連結成分を求めた後、その外接長方形について、(1)幅が70画素以上、(2)長方形内に占める抽出画素の面積が45%以上、(3)高さに対する幅の大きさが0.75以上という3条件が、連続するフレームで満足されるものを選ぶ。

2.2 線形補間による解像度向上

文字の解像度が十分でない各フレームの画像を線形補間により4倍の解像度に変換する。具体的には、4倍の解像度の画像に対して横および縦方向の4画素おきに元画像の値を代入し、残りの画素には近傍に存在する元画像の複数の輝度値に距離で重みをつけた値を代入する。

2.3 複数のフレームを用いたコントラストの増大

ニュース映像では、背景は動きをともなう場合があるが、スーパーインポーズされた文字は一定の輝度値を保ちながら数十フレーム程度同じ位置に表示されている。文字が輝度の大きな白色で表示されている場合には、それぞれの画素について、文字の出現しているフレーム内で最も小さい値に置き換えることで、図3のようにコントラストの高い画像をつくり出すことができる。輝度の最小値を探す開始および終了フレームは、文字列検出処理で得られる結果を使用する。

線形補間による解像度向上と複数フレームによるコントラスト改善を施した結果を図4に示す。1文字のサイズが約30×40画素と十分な大きさとなり、かつ滑らかな文字の外形を保っていることが分かる。一方、

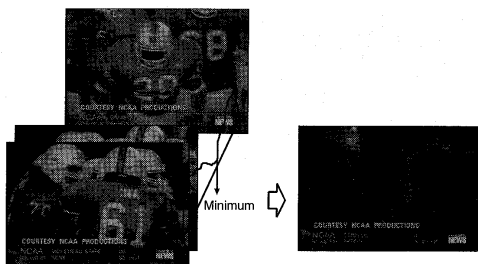


図3 複数フレームを用いたコントラスト改善

Fig. 3 Image enhancement using multi-frame integration.

この例では背景の動きが少なく、文字とのコントラストは改善されていない。

2.4 文字抽出フィルタ

二値化処理を中心とした従来の手法では抽出困難である対象から文字を抽出するために、文字の構造の特徴に基づいたマッチドフィルタを設計する。文字は線分で構成されていることに注目して、縦(90°)、横(0°)、斜め(135°および45°)の4種類の線検出マッチドフィルタを適用し、その出力を合成して文字を抽出することを考える。マッチドフィルタは、図5に示すサンプルについて、4方向の線分に対応する304画素(90°)、122画素(0°)、189画素(135°)、89画素(45°)をそれぞれ選び、選んだ画素と近傍画素の平均から図6のように求める。フィルタのサイズは、1本の文字の線だけが含まれる範囲として、それぞれ15×3、3×7、9×7、9×7画素と定め、各フィルタ内の値は平均がゼロとなるように正規化する。このサンプルで求めたフィルタを他のすべての画像に適用する。抽出処理では、4種類のフィルタとの相関のうち正の値をとるものだけを加算し、しきい値処理により最終的な二値画像出力を得る。

文字抽出フィルタの出力結果を図7に示す。各方向のフィルタの出力である(a)から(d)では、それぞれ対応する線の成分が抽出され、最終的な出力である(f)では、図4では除去できていない複雑背景からの文字抽出が実現されている。また、フィルタの設計に用いていない右側の10文字についても、背景とのコントラストが低い状態であるにもかかわらず抽出できていることが分かる。

2.5 文字分割候補の抽出

複雑背景から文字を抽出した二値画像に対して、縦方向の射影データを用いて、文字の分割候補の左端および右端を検出することができる。この結果、図8のように過分割してしまう場合があるが、この分割の中には正しい文字セグメントが含まれている。次節では、

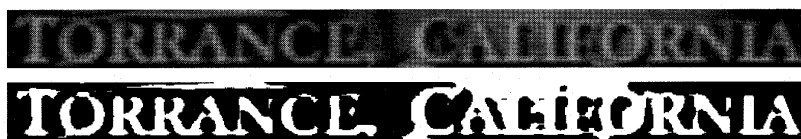


図4 画素間補間と複数フレームによるコントラスト増大の結果 (上) と、その二値化画像 (下)。画像サイズは 813 × 56 画素

Fig. 4 Result of sub-pixel interpolation and multi-frame integration (upper) and its binary image (lower). The image size is 813 × 56 pixel.

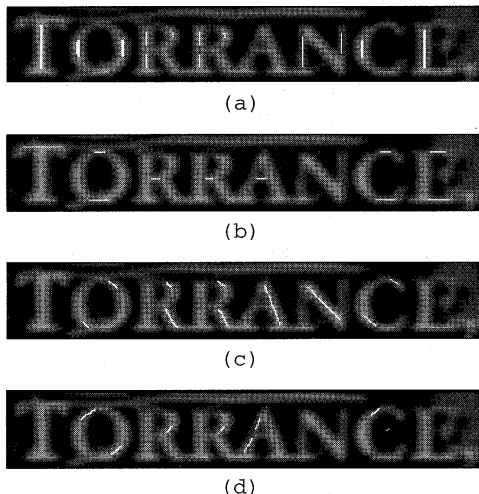


図5 文字抽出フィルタを作成するサンプル画素 (白画素). (a) 90°. (b) 0°. (c) 135°. (d) 45°

Fig. 5 Learning data of filters (white pixels). (a) 90°. (b) 0°. (c) 135°. (d) 45°.

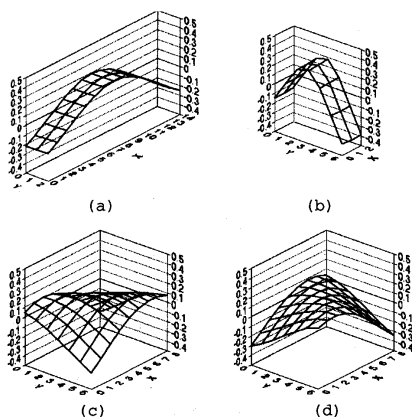


図6 文字抽出フィルタ. (a) 90°. (b) 0°. (c) 135°. (d) 45°

Fig. 6 Character extraction filters. (a) 90°. (b) 0°. (c) 135°. (d) 45°.

文字認識の類似度に基づいてこの分割候補から文字セグメントを選択する方法について述べる。

2.6 文字のセグメンテーションと認識の統合

2.6.1 文字認識

文字の認識は、あらかじめ作成した辞書データとの類似度¹³⁾に基づいて行う。少なくとも1つの文字分割候補を含んでいるセグメント内の二値画像は、19 × 28画素のサイズへ正規化された後、位置ずれによる影響を小さくするために、近傍の画素を計数することではけた濃淡画像に変換される。その後、輝度を正規化した画像データ $n(x, y)$ に対して辞書データ $ref_c(x, y)$ との類似度 m_c を求める。

$$m_c = \frac{1000 \sum n(x, y) \cdot ref_c(x, y)}{\sqrt{\sum (n(x, y))^2} \sqrt{\sum (ref_c(x, y))^2}}$$

最も大きな類似度 m_c を持つカテゴリ c を第1候補の認識結果として選び、さらに第2および第3候補の結果も求める。標準パターン $ref_c(x, y)$ は、ゴシック体文字の印刷物から各文字について約10サンプルを平均して作成する。なお、今回利用した映像では大文字のみが使用されていたので、カテゴリ c は“A”から“Z”の26種類に限定している。

2.6.2 認識結果による文字分割候補の選択

図8のような文字分割候補から認識結果を求める分割範囲 (文字セグメントと呼ぶ) の組合せを複数定め、語の中の平均類似度が最大になるものを選ぶことを考える。類似したアプローチとして、語の辞書データをはじめから利用して文字分割結果と認識結果を同時に求める報告もあるが^{14), 15)}、ここでは簡単に類似度だけで文字の分割と認識結果を決定する方法を検討する。語の辞書データによる認識結果の修正は、次章で説明する文字列照合において実施する。

文字分割候補のすべての組合せを調べることは効率的でないので、語の左端から一部分のセグメントだけを順に評価することで計算コストを低減する。すなわち、図9に示すように、分割候補を複数組み合わせた2文字分の文字セグメントに対して類似度の平均が最大となる組を求め、1番目の文字セグメントを正し

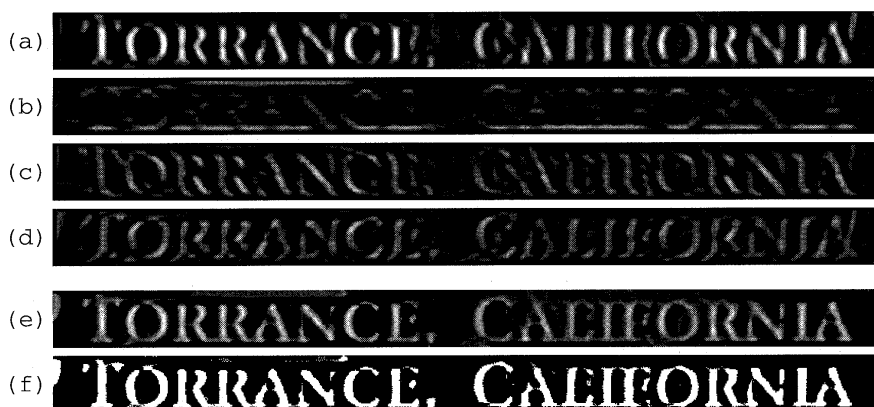


図7 文字抽出フィルタの結果. (a) 90°. (b) 0°. (c) 135°. (d) 45°. (e) 4フィルタの統合結果. (f) 二値画像

Fig. 7 Results of character extraction filters. (a) 90°. (b) 0°. (c) 135°. (d) 45°. (e) Integrated result of four filters. (f) Binary image.



図8 垂直方向の射影データによる文字分割候補の検出結果 (白線)

Fig. 8 Result of detection for segmentation candidates using vertical projection (white lines).

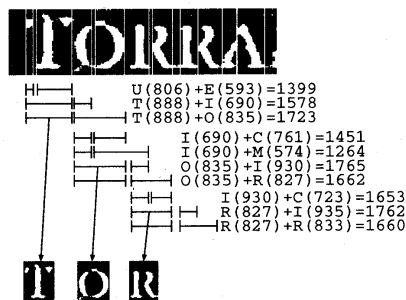


図9 文字セグメントの選択処理

Fig. 9 Character segmentation process.

い分割結果とする処理を順に繰り返す。ここで、評価する類似度は第1候補の結果であり、類似度が500以下の文字セグメントは評価しない。また、文字セグメントの幅に上限値を設定することで選択の数を限定する。各文字セグメントにおける縦方向の位置は、先に決められる横方向の分割の範囲に対して二値画像を横方向に射影して決定する。

図9の例では、3文字目の“R”において過分割した領域(“T”との類似度が高い)の類似度が高くなるが、2番目の文字セグメントの類似度が低いので、1番目の文字セグメントを誤って選択しない。今回のデータは大文字のアルファベットだけという単純な文字の構

成で、しかもその1文字はたかだか2つに過分割される場合がほとんどであるので、このような方法で誤った分割を避けながら計算コストを下げるができる。

図8で検出した文字の分割候補に対して、この方法を適用した結果を図10に示す。左の語の最後のセグメントにコンマが含まれている結果を除いて、正しい文字セグメントが選択できている。

文字セグメントを評価する際には最も大きい類似度とともに2番目および3番目に大きい類似度を求めており、最終的に選択されたセグメントに対して第3候補までの類似度とカテゴリをそのまま文字認識の結果として出力する。

3. 映像の文字認識結果

前章で説明した映像の文字認識方法を、30分のニュース放送7本を用いて評価した。映像はMPEGデータとして352×242画素のサイズで記録されており、このなかには、文字を含むフレーム群が256カ所、375の文字列領域(行に対応する)、965語が含まれている。

はじめに、文字検出処理について、文字列が含まれているフレームの検出と、フレーム内に含まれる文字領域の検出性能を調べた。表1に示すように、文字を含むフレームの検出率は91.8%で、語に対応する部分領域については76%を検出できている。



図 10 文字セグメントの選択の結果

Fig. 10 Segmentation results.

表 1 文字検出の結果

Table 1 Results of character detection.

	全数	正検出数 (検出率)	誤検出数
フレーム	256	235 (91.8%)	26
文字領域	375	336 (89.6%)	35
語	965	733 (76.0%)	—

表 2 文字の認識結果 (全文字数 5076)

Table 2 Results of character recognition (total 5076 characters).

	従来法	提案する方法
正解文字数	2360	4237
認識率	46.5%	83.5%

この正しく文字検出できた対象について、前章で説明した、線形補間による解像度増大、複数フレームを用いたコントラスト増大、文字抽出フィルタ、文字のセグメンテーションと認識の統合の処理群を施し、文字認識性能を評価する。30 分の番組に対する処理時間は、ワークステーション (MIPS R4400 200 MHz) を用いて約 120 分であった。

表 2 に、我々のアプローチと従来方法による文字の認識結果を示す。従来方法とは、元画像を一定しきい値で二値化し、縦方向の射影の端で文字を検出して認識した結果 (すなわち、解像度の向上、複数フレームによるコントラスト増大、文字抽出フィルタ、文字のセグメンテーションと認識の統合処理を除いた文字認識結果) で、文字認識は同じ方法を用いている。認識率は約 1.8 倍向上しており、我々のアプローチが映像の文字認識に有効であることが分かる。

4. ビデオ情報における異種情報の対応の利用

4.1 映像中の文字とクローズドキャプションの対応

ニュースでは同一の話題を異なる形態の情報で伝えているので、その情報間には関連が存在するはずである。たとえば、前章で用いた 7 番組について映像中に目視確認できる 965 語とクローズドキャプションに含まれる 23,649 語の関係を調べてみると、424 語が共通に用いられており、明らかに対応関係が存在する。

本章では、このような対応関係をビデオ情報の解析に利用する例として、映像中の文字認識結果をクローズドキャプションの情報を用いて改善する方法と、両方の表現に共通して存在する語に基づいて、ビデオ情

報から重要な情報を抽出する方法を検討する。

4.2 文字認識における他の情報の利用

ビデオデータベースの検索でキーワード入力及要求されることを考えると、映像中の文字認識では文字の認識率よりも語の認識率が重要な意味を持つ。我々の結果では文字の認識率は 83.5% であったが、語を構成する文字の認識結果の第 1 候補がすべて正しい場合は全 733 語のうち 400 語で、語の認識率としては 54.6% しか得られていない。一方、第 3 候補までに正解が含まれている結果を調べると、正しい語は 498 語 (67.9%) まで増えるので、適切な修正処理により認識率を改善できると考えられる。一般的に、どのような手法を用いても 100% の文字認識を達成するのは不可能なので、語の認識率を改善するために、このような後処理が必要となる。

ここでは、各文字の第 3 候補までの認識結果を類似度で重みづけし、クローズドキャプションの語を集めた辞書 (23,649 語から重複を除いた 4,824 語; 以下、CC 辞書と呼ぶ) の中で最も近い語を選ぶことで認識率を向上させる手法を考える。また、一般的な辞書として、*Oxford Advanced Learner's Dictionary* (69,517 語; 以下、*Oxford* 辞書と呼ぶ) を併用し、それぞれの効果を評価する。

これら辞書に含まれる語と認識結果を比較する基準として、以下に示すような編集距離を考える。認識結果の語を表す文字列を a とし、 $a(i)$ を a の i 番目の文字を表すシンボルとする。各文字は第 3 候補までの認識結果が得られるので、任意の文字のシンボル p に対して、 p_j^c を j 番目に類似度の大きな認識結果の文字カテゴリ、 p_j^s をそのときの類似度と定義する。また、 b を比較する辞書の語とし、 $b(i)$ を b の i 番目の文字とする。 $a(i..)$ と $b(i..)$ を、それぞれ a と b の i 番目の文字から始まる部分的な文字列とすると、 a と b の編集距離 $d(a, b)$ を次のように定義する^{16), 17)}。

$$d(a, b) = \min \begin{cases} 1 + d(a(2..), b), \\ 1 + d(a, b(2..)), \\ c(a(1), b(1)) + d(a(2..), b(2..)) \end{cases} \quad (1)$$

この $d(a(2..), b)$, $d(a, b(2..))$ および $d(a(2..), b(2..))$ については、式 (1) と同様に部分的な文字列の距離を

表3 語の認識率 (全体の語の数: 検出できた733語)
Table 3 Word recognition rate (total: detected 733 words).

	正解 (正解率) words(rate)	誤認識 を修正	正解認識 を誤修正
修正処理無	400 (54.6%)	—	—
修正処理 (CC)	455 (62.1%)	162	107
修正処理 (Ox)	442 (60.3%)	131	89
修正処理 (CC+Ox)	514 (70.1%)	157	43

CC: CC 辞書, Ox: Oxford 辞書

用いて定義され, これを繰り返し適用することで文字列の距離が求められる. $c(p, q)$ は第3候補までの認識結果を有する文字 p と辞書の文字 q の間のコスト関数で, 次のように定義する.

$$c(p, q) = \begin{cases} 1 & (\forall_i p_i^c \neq q) \\ 1 - \frac{p_i^c}{p_1^c} & (p_i^c = q) \end{cases} \quad (2)$$

a と b の正規化した距離 $\hat{d}(a, b)$ は以下に示すように定義できる.

$$\hat{d}(a, b) = \frac{d(a, b)}{\max(\text{len}(a), \text{len}(b))} \quad (3)$$

ここで $\text{len}(a)$ は文字列 a の文字数を表す. この定義に従えば, a と b が同じ場合に距離が最小の0であり, a と b が1文字も同じでない場合に最大距離1になる.

認識結果はCC辞書とOxford辞書の全単語について総当たりで照合される. 1番組に含まれる103語に対するCC辞書(4,824語)およびOxford辞書(69,517語)との照合時間は, ワークステーション(MIPS R4400 200MHz)を用いてそれぞれ1分14秒と21分57秒であった. この処理時間は, 語における一部の文字の認識結果を用いて照合辞書を選別する方法¹⁴⁾などにより改善できると考える.

認識候補に対して2種類の辞書データから最も近い語を選択した結果を表3に示す. 一部の文字認識に間違いがあった157語について正しく修正され, 語の認識率は70.1%に向上している. 一方, すべての文字が正しく認識されているにもかかわらず, 辞書に含まれていないために別の語が選択された例が固有名詞を中心に43語存在する. これについては, すべての文字が一定値以上の類似度をとる語では認識結果をそのまま用いるという判断を加えることで改善されと考える.

また, 表3から, 2種類の辞書を用いることで特に誤修正が減っていることが分かる. これはCC辞書とOxford辞書と合わせて用いることで必要な語彙全体が網羅できているためと考えられる. 表4に示すよう

表4 映像中の語に対する辞書データの語彙 (* 含まれている語)
Table 4 Vocabularies of dictionary corresponding to superimposed captions (* included).

画像中の語	CC 辞書	Oxford 辞書
WASHINGTON	*	*
GEORGIA		*
AUGUSTA	*	
LYNNE	*	*
IAN		*
KAWAKITA	*	
GINGRITCH	*	
CNN	*	
NBA	*	
WHITEWATER	*	
COURTESY		*
MIDNIGHT		*

表5 ニュース映像の語の認識結果
Table 5 Word recognition results for news videos.

全語数	検出語数	認識語数	CC 辞書にない語
965	733	514	261

に, 映像にスーパーインポーズされている語の中で, 有名な場所("WASHINGTON", "GEORGIA"), 典型的な西洋人の名前("IAN", "LYNNE"), 一般的な語("MIDNIGHT", "COURTESY")はOxford辞書に含まれ, 比較的小さな街の名称("AUGUSTA"), 西洋では奇異な人名("KAWAKITA"), 最近使われるようになった事件("WHITEWATER"), 組織名("CNN", "NBA")は, CC辞書に含まれており, 2種類の辞書の語彙が異なっている.

以上の結果から, クローズドキャプションは一般的な知識に含まれていない語について参照可能な情報であり, その利用により, 通常の後処理では回復が容易でない固有名詞に対して認識性能を向上させているといえることができる.

4.3 ビデオ情報からの重要な情報の抽出

前節の修正までを含めた最終的な映像中の語の認識結果を表5に示す. 認識できた語のうち約半数である261語はクローズドキャプションに含まれておらず, 映像の内容に基づいた新しい検索インデックスとして利用できる. 一方, 残りの約半数はクローズドキャプションにも含まれている語であるが, 画像中の文字とクローズドキャプションの両方で同時に表現されていれば, 複数表現にわたって局所的に出現頻度の高い語として, テキスト情報におけるTFIDFによる方法¹²⁾と同様に重要な語と見なすことができる.

この考えに基づいて, 同一トピックの異種情報の中で共通に用いられている語をキーワードとして抽出する. ここでトピックとは, 1つの話題に対応したビデ

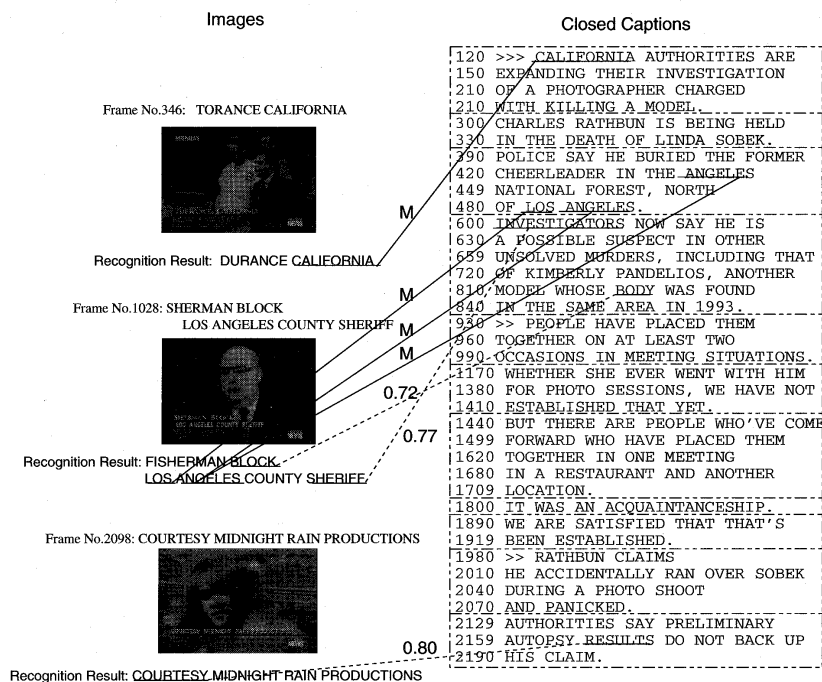


図 11 映像中の語とクローズドキャプションの対応づけ

Fig. 11 Correspondance between recognized words in images and closed captions.

オ情報の単位を指し、実験で利用したニュース放送ではクローズドキャプションの記号“>>>”で区切られる範囲として容易に求められる。映像のトピックの範囲については、対応するフレーム番号から限定できる。

図 11 に示すトピックの例では、記号“M”で示した 4 語*が両方の表現に含まれており、キーワードとして選ぶことができる。映像にスーパーインポーズされる文字には、地名、人名や肩書き、画像の著作権表記の情報が含まれており、たとえば、地名と人名はニュース報道の 5W1H に対応している重要な情報である。図 11 の例では、両方の表現に存在する語として 2 つの地名が選ばれており、トピックの発生場所に関するキーワードとして抽出されている。

さらに、キーワードを含む映像中のフレーム（キーフレーム）と、キーワードを含むクローズドキャプションの文（キーセンテンス）とを重要な情報として抽出することを考える。図 11 では、このキーワードを含む 2 枚のキーフレームとキーセンテンス 2 文を抽出できている。キーセンテンスからは、カリフォルニアで起きた殺人事件で犯人の余罪が調査されている事実

を、2 枚のキーフレームで示される登場人物や状況とともに理解することができる。本章の最後では、このように抽出した情報を用いて、ビデオ情報の要約を作成することを検討する。

4.4 類義語による抽出情報の拡大

映像中の文字情報のうち肩書などの一部の情報には、クローズドキャプションであえて異なる語を用いて表現されているものがある。図 11 の例では、映像中の“SHERIFF”という語はクローズドキャプションに含まれていないが、同じ捜査側の意味を表す“POLICE”、“INVESTIGATORS”、“AUTHORITIES”という語が含まれている。このような意味的な関連まで考えると、映像中の語とクローズドキャプションの対応づけが増え、新たなキーワードを加えることができる。

意味的な関連を調べるために、シソーラスを用いて共通の上位語の深さをそれぞれ関連性を調べる語の深さの和で正規化した類似度¹⁸⁾を算出する。シソーラスとして約 9 万 2 千語を網羅する上位下位シソーラス WordNet 1.5¹⁹⁾を用いた。図 12 の例に従えば、“SHERIFF”のノードの深さを n_i 、“INVESTIGATOR”のノードの深さを n_j 、共通のノードの深さを

* ここでは“LOS ANGELES”は 2 語として扱う。

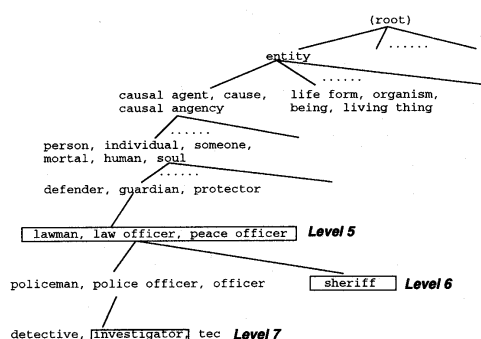


図 12 シソーラスにおける語の構造

Fig. 12 Structure of thesaurus.

表 6 認識した語とクローズドキャプションの対応

Table 6 Correspondence between recognized words and closed captions.

	対応の総数	正しい対応 (割合)
語の完全一致	565	488 (86%)
類義語の一致	104	45 (43%)

n_c としたときに、以下の計算で 2 語の意味の類似度 s を 0.77 と求めることができる。

$$s = \frac{n_c \times 2}{n_i + n_j} \quad (4)$$

式 (4) に従えば、語が一致する場合に類似度は 1 となる。また、このシソーラスは語形変化に対応しているので、数や格が異なる語の間についても類似度 1 として対応づけが可能になる。

関連性を調べるクローズドキャプションの語は、Link Grammar²⁰⁾を用いた構文解析により主語と判断された語だけを選択する。これは、前後に接続される品詞の情報を持つ辞書を用いて、辞書に含まれている語を中心に最適の接続の組合せを見つける手順を、語間の接続は交差しないという英文の性質に基づいて簡略化して実行する構文解析方法である。

たとえば、0.7 以上の類似度をとる対応づけを選べば、図 11 では破線で示す 3 カ所の対応関係を加えることができる。“BLOCK”と“MODEL”，および“COURTESY”と“RESULTS”の対応づけは間違っているが、“SHERIFF”と“INVESTIGATORS”の対応は妥当である。

表 6 に、語が一致する場合 (類似度 1.0) と類似度 0.7 以上の類義語による対応づけについて、人間が正しいと判断できる数を示す。この判断は、図 11 の例で説明したように、ニュースのトピックで使われている意味において対応づけが正しいことを基準とした。ここで、文字認識で誤った語は対応関係の評価から

除いている。語が一致する場合の正しい対応づけの数は 86% と高く、類義語に対する対応づけも約 4 割が正しいことが分かる。

類似した意味の対応づけまで拡大して、キーフレームとキーセンテンスを選び、要約の情報を増やすことも可能である。図 11 の例では、1 つのキーフレームと 2 つのキーセンテンスを加えることができる。3 枚目のキーフレームと最後のキーセンテンスは、正しい対応づけで選ばれている情報ではないが、同時に選ばれた第 4 文は、捜査側が殺人事件の拡大の範囲の詳細を説明する重要な情報で、この文の主語である“INVESTIGATORS”と映像中の“SHERIFF”との正しい対応づけにより抽出されている。

本手法では、類似度により抽出する情報の量を制御できるという特徴がある。この特徴を利用することで、たとえば、ユーザの要求に合わせて提供する情報量を変化させる情報フィルタリングの機能を、ビデオ情報の内容に基づいて実現可能になる。

4.5 ビデオ情報の要約の作成

図 11 から対応づけがされたキーフレームとキーセンテンスを抽出すると、図 13 のような要約を作成することができる。一方、従来の方法³⁾により要約を作成すると、たとえば図 14 のような結果が得られる。図 14 において、キーフレームは各フレームについて次フレームとの RGB ヒストグラムとの差を求めることで検出し、キーセンテンスについては TFIDF 値が高い語を含む 4 文を抽出している⁴⁾。この TFIDF 値は、1 トピックに含まれる各々の語の頻度について、十分大量に収集されたニュース番組のクローズドキャプションにおける頻度との割合として求める。図 14 の枠で囲んだ語は高い TFIDF 値を持つキーワードである。このキーフレームの抽出結果は検出しきい値の設定により違ってくる。

クローズドキャプション情報に注目してみると、従来法では、固有名詞などの珍しい語を含む文が画像とは独立に選択されているのに対し、我々の方法では、画像内容との共通性に基づいて情報が抽出されるので、画像との関連を理解しやすい要約が作成できていることが分かる。特に、編集者が強調したいという意図に基づいて重要な情報を両方の表現に含めている場合には、本方式で作成された要約は有用になると考えられる。

5. ま と め

本論文では、ビデオ情報から重要な情報を抽出する目的で、映像にスーパーインポーズされた文字を認識

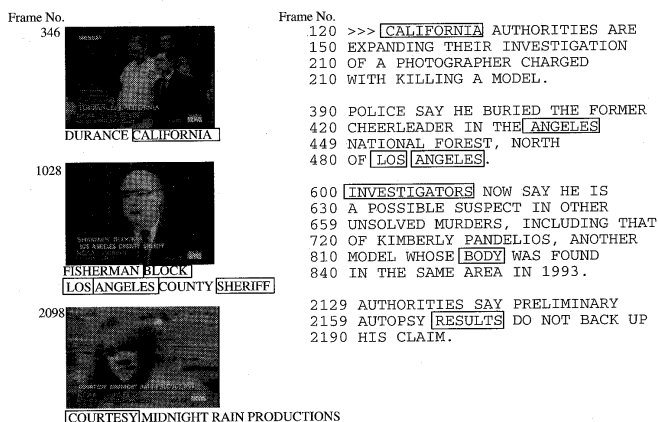


図 13 要約の作成例

Fig. 13 Summary of news topic.

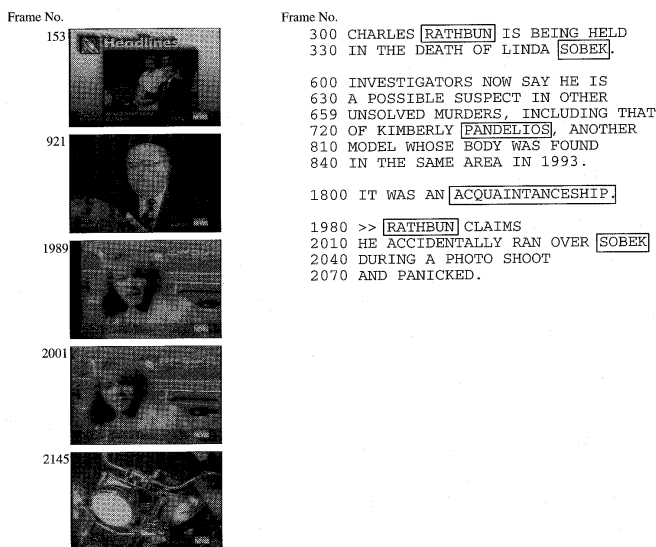


図 14 従来法による要約の作成例

Fig. 14 Summary of news topic using conventional method.

し、クローズドキャプションとの対応を求める方法について述べた。

映像の文字認識結果に対して、ビデオ情報に含まれる別の情報であるクローズドキャプションを利用することで語の認識率を向上させることができた。これは、音声と映像中の文字との相互作用により理解がしやすくなるというビデオ情報の人間への働きかけを想起させる方法である。

また、文字認識結果とクローズドキャプションの対応関係からキーワードを抽出する方法を検討し、その

キーワードを介してキーフレームとキーセンテンスを選択する要約作成方法を提案した。これは、音声と映像中の文字で同じ情報が提示されたときに人間が強い印象を受けるという作用を、計算機によるビデオ情報からの情報抽出に利用した方法である。今後、この要約の作成方法について、ユーザスタディにより有効性を確認し、新たな応用を検討していきたい。

謝辞 有意義な議論をしていただいた筑波大学中村裕一先生および学術情報センター佐藤真一先生に感謝いたします。

参 考 文 献

- 1) Zhang, H.J., Gong, Y., Smoliar, S. and Tan, S.Y.: Automatic Parsing of News Video, *Proc. IEEE International Conference on Multimedia Computing and Systems*, pp.45-54 (1994).
- 2) Brown, M.G., Foote, J.T., Jones, G.J.F., Jones, K.S. and Young, S.J.: Automatic Content-Based Retrieval of Broadcast News, *Proc. ACM Multimedia*, pp.35-43 (1995).
- 3) Wactlar, H.D., Kanade, T., Smith, M. and Stevens, S.: Intelligent Access to Digital Video - The Informedia Project, *IEEE Computer*, Vol.29, pp.46-52 (1996).
- 4) Nakamura, Y. and Kanade, T.: Semantic Analysis for Video Contents Extraction - Spotting by Association in News Video, *Proc. ACM Multimedia*, pp.393-401 (1997).
- 5) 加藤俊一, 栗田多喜夫: 画像の内容検索—電子美術館への応用, 情報処理, Vol.33, No.5, pp.46-52 (1992).
- 6) Satoh, S. and Kanade, T.: NAME-IT - Association of Face and Name in Video, *Proc. IEEE CVPR*, pp.368-373 (1997).
- 7) Kurakake, S., Kuwano, H. and Odaka, K.: Recognition and Visual Feature Matching of Text Region in Video for Conceptual Indexing, *Proc. SPIE Storage and Retrieval in Image and Video Databases*, 3022, pp.368-379 (1997).
- 8) 佐藤 隆, 新倉康巨, 谷口行信, 阿久津明人, 外村佳伸, 浜田 洋: MPEG 符合映像からの高速テロップ領域検出法, 電子情報通信学会論文誌, Vol.J81-D-II, No.8, pp.1847-1855 (1998).
- 9) Lienhart, R. and Stuber, F.: Automatic Text Recognition in Digital Videos, *Proc. SPIE Image and Video Processing IV*, 2666, pp.180-188 (1996).
- 10) Sato, T., Kanade, T., Hughes, E.K. and Smith, M.A.: Video OCR for Digital News Archives, *Proc. IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp.52-60 (1998).
- 11) Sato, T., Kanade, T., Hughes, E.K., Smith, M.A. and Satoh, S.: Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions, *Multimedia Systems*, Vol.7, pp.385-395 (1999).
- 12) Smith, M.A. and Kanade, T.: Video Skimming and Characterization through the Combination of Image and Language Understanding Technique, *Proc. IEEE CVPR*, pp.775-781 (1997).
- 13) 飯島泰蔵: パターン認識理論, 森北出版, 東京 (1989).
- 14) Zhao, S.X. and Srihari, S.N.: A Word Recognition Algorithm for Machine-printed Word Images of Multiple Fonts and Varying Qualities, *Proc. 3rd International Conference on Document Analysis and Recognition*, pp.351-354 (1995).
- 15) Fang, C. and Hull, J.J.: A Hypotheses Testing Approach to Word Recognition Using an A* Search Algorithm, *Proc. 3rd International Conference on Document Analysis and Recognition*, pp.360-363 (1995).
- 16) Hall, P.A.V. and Dowling, G.R.: Approximate String Matching, *ACM Computing Surveys*, Vol.12, pp.381-402 (1980).
- 17) Wagner, R.A. and Fischer, M.J.: The String-To-String Correction Problem, *J. ACM*, Vol.21, pp.168-173 (1974).
- 18) 長尾 真 (編): 自然言語処理, 岩波書店, 東京 (1996).
- 19) Fellbaum, C.: *WordNet*, MIT Press, Cambridge, MA (1998).
- 20) Sleator, D.K. and Temperley, D.: Parsing English with a Link Grammar, CS Technical Report CMU-CS-91-196, CMU, Pittsburgh, PA (1991).

(平成 11 年 4 月 1 日受付)

(平成 11 年 10 月 7 日採録)



佐藤 俊雄 (正会員)

昭和 60 年名古屋工業大学計測工学科卒業。昭和 62 年同大学院生産システム工学専攻修士課程修了。同年(株)東芝入社。平成 8 年から 10 年にかけて米国カーネギーメロン大学客員研究員。現在柳町情報・社会システム工場にて、画像処理技術の開発に従事。IEEE, 電子情報通信学会各会員。



金出 武雄

昭和 48 年京都大学大学院博士課程修了。同年同大学情報工学科助手。昭和 51 年同助教授。昭和 55 年米国カーネギーメロン大学計算機科学科高等研究員。昭和 60 年同教授。現在同 U.A. and Helen Whitaker 全学教授。同大ロボティクス研究所所長。工学博士。計算機視覚, ロボットの腕, 自律走行車, VLSI センサに関する研究に従事。米国工学アカデミー外国特別会員。IEEE Fellow, AAAI Fellow, ACM Fellow.