

医療教育用映像のための自動インデクシングの検討

Automatic indexing for educational video of medical treatment

丹羽 弘充^{*1}

Hiromitsu Niwa

河村 高守^{*1}

Takamori Kawamura

田村 哲嗣^{*2}

Satoshi Tamura

速水 悟^{*2}

Satoru Hayamizu

^{*1} 岐阜大学大学院工学研究科

Graduate School of Engineering, Gifu University

^{*2} 岐阜大学工学部

Faculty of Engineering, Gifu University

This paper proposes a framework of automatic indexing using speech and video for medical treatment education. A language model for speech recognition was constructed to include the words used in the medical treatment field. Speech recognition is conducted to obtain audio metadata. Audio metadata consist of begin and end time, recognition results, and confidence scores. Video metadata, e.g. time information and a score computed using optical-flow analysis are generated. The video metadata are independent from cameraworks. The system integrates the results of audio and video metadata.

1. はじめに

近年、インターネット回線の高速化や、HD DVD やブルーレイディスクなどの普及により、映像や音声などの大容量マルチメディアコンテンツを利用する機会が増えている。しかし、これらのデータの中から効率的に情報を得るのは困難である。そこで、映像・音声の各部分においてどのような内容であるのかをメタデータ(データに関するデータ)として付与することにより、効率的なデータの閲覧を行うことが考えられている[大附 03, 橋本 05]。

また、医療教育(図1)においては、患者に対して行う処置を、素早く正確に行わなければならない。実際に自分が行った処置手順が、指導者と比較して正しいかどうかを後から見直し、改善点を学習する必要がある[露木 06, 河村 07]。さらに、自分が行った処置手順以外にも、多くの受講者が行った処置方法を参考にすることで、より高い理解を得ることができる。近年では全国各地で医療教育の講習会が開かれており、受講者が会場へ赴き、指導者のもとで処置手順を確認しなければならないため、多大な時間をかける必要がある。

上記のような問題を解決するため、本研究では実際に医療教育を目的として撮影されたBLS(Basic Life Support: 一次救命処置)、ACLS(Advanced Cardiovascular Life Support: 二次救命処置)の映像に対して、音声認識、画像処理を行うことにより、メタデータ付与を行う。これにより、ユーザが必要としている内容を視聴するのにかかる時間の短縮や、改善点の発見を容易にすることを目的としている。

2. システム構成

本システムの構成を図2に示す。音声認識結果の情報に、補助的に画像処理の結果を用いることにより、メタデータを作成する。メタデータには、映像中に含まれる人工呼吸や心臓マッサージなど、医療講習会のシナリオに基づいて定義されたシーンの名称、開始時間、終了時間が記録されており、これを利用することにより、効率的なメディアデータの閲覧が可能となる。各処理について以下に述べる。



図1 医療教育指導の風景

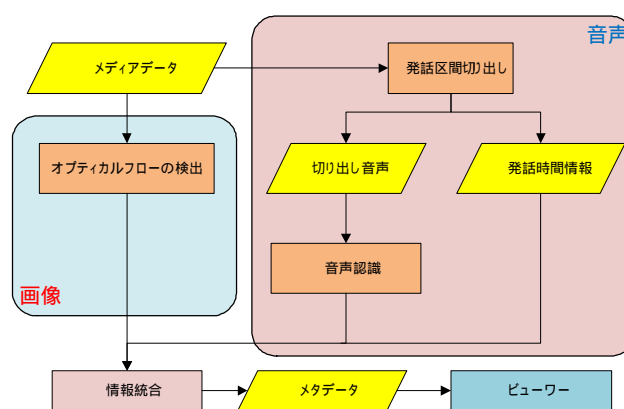


図2 本システムの構成

2.1 音声認識

初めに、音声区間の切り出しを行う。これは、発話以外の雑音による認識率低下を防ぐためであり、出力のパワーが低い部分で切り分けている。切り出された音声に対して、音声認識エンジン Julius[河原 05]を用い、音声認識を行う。音響モデルは、

CSJ(日本語話し言葉コーパス)を用いた, triphone による不特定話者音響モデルを使用する. 言語モデルは単語 N-gram を用い, ACLS, BLS 講習会で撮影された映像の音声から作成した書き起こし文を基に構築する. これは, 医療分野では専門的な単語, 略語が多く含まれており, 一般的なコーパスで作成された言語モデルでは, それらの専門性の高い語句を認識することが困難なためである. 単語辞書も同様に作成するが, このとき, 「AED」と「除細動器」などのように, 言い回しの違いによる問題を抑えるため, 認識用辞書では同様の単語として扱う. 音声認識の結果より, 切り出された各音声データにおける単語の発話時間, 発話内容を得る. さらに, 発話時間情報のデータを用いて切り出した音声の発話時間に修正を加えることにより, メディアデータ全体中の発話時間を得る. ここで, XML データに各単語の信頼度などを同時に記録しておく(図3). 単語信頼度は, 0.0 から 1.0 の範囲で表す値であり, この値が 1.0 に近いほど, その単語に似たスコアをもつ他の競合候補がほとんどなかったことを示し, 0 に近づくほど, その単語と同じ程度のスコアをもつ他の単語候補が多く出現していたことを示す. 図3においては, cmscore が単語信頼度を表している.

```
<?xml version="1.0" encoding="utf-8" ?>
<root>
- <SEGMENT begin="1.95" end="7.58" name="bls2\bls2.0000.wav">
  <WORD id="1" begin="2.24" end="2.62" cmscore="0.96" n-score="-25.01">周囲</WORD>
  <WORD id="2" begin="2.63" end="2.88" cmscore="0.48" n-score="-27.33">が</WORD>
  <WORD id="3" begin="2.89" end="3.21" cmscore="0.53" n-score="-25.42">安全</WORD>
  <WORD id="4" begin="3.22" end="3.51" cmscore="0.28" n-score="-23.85">です</WORD>
```

図3 音声認識の結果

2.2 オプティカルフローの検出

画像特徴量として, オプティカルフロー[Horn 81]を使用する. オプティカルフローとは, 画像間における濃淡パターンを対応付けし, その移動量をベクトル表現したものである. オプティカルフローを用いるにあたって, 結果に影響を与える要素としては, 以下の二つが考えられる.

- 撮影者のカメラワーク
- 被写体の動作

テレビ番組などを考えた場合, 熟練した撮影者が撮影を行うことから, カメラワークの特徴を用いてシーン検出に利用できる特徴を取り出すことができる. 一方, 医療教育映像においては, 撮影技術を持たない人が撮影を行う場合が多い. そのため, 撮影者のカメラワークから特徴を取り出すことは難しい. また, 上記のような理由から, 被写体の動作が隠れてしまう場合があり, 詳細な動作を使用することは困難である.

本研究では, 映像中, 横軸方向に三分割したうち, 中央部分の情報を用い, 縦軸方向の移動量の和を求めている. 心臓マッサージのシーンでは, 図4のように, この値が周期的に変化する. この値は, カメラワークによる影響が少なく, 重要な情報を画面中央近くに置いて撮影する, といったような, ある程度の一般的な撮影知識を持っている撮影者が撮影した映像に対してならば, 特徴的な結果を得ることができる. また, 図5に示すように, 今回の映像では, 心臓マッサージの処置を行っているシーンが大きな割合を占めている. そのため, 振幅 A , 周期 T , 検出部分の幅 W , 高さ H とし,

$$G(t) = \frac{50 \times A \times T}{W \times H} \quad (1)$$

により求めた値を, 時刻 t における画像特徴量とする.

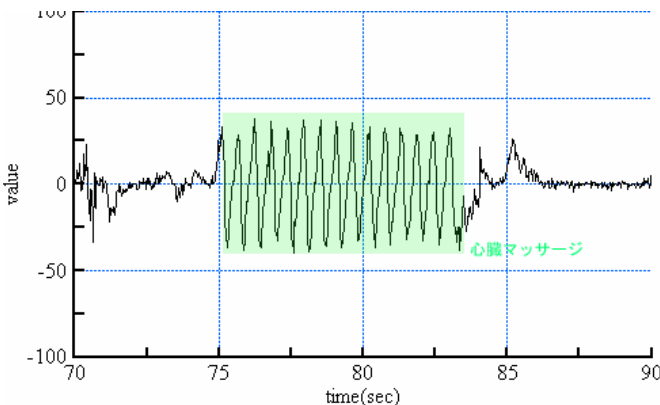


図4 縦軸方向オプティカルフローの和

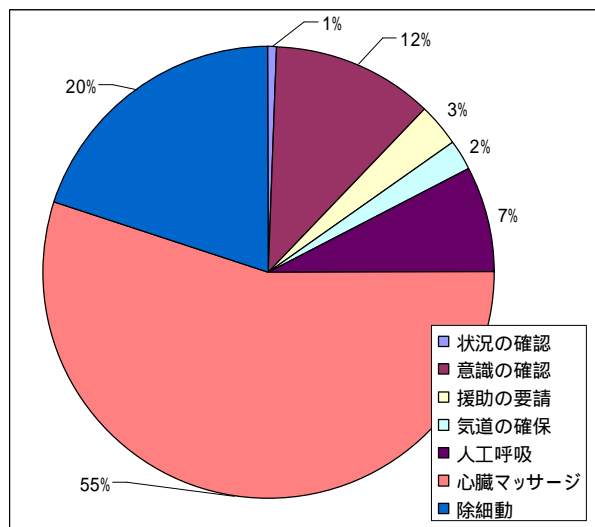


図5 映像中における処置手順の割合

2.3 情報統合

音声認識, 画像処理によって得られたデータをもとに, メディアデータに対するメタデータを作成する. 音声認識結果の単語には, 各シーンに対応するキーワードの定義を行った(表1). これは, 言語モデル, 単語辞書作成時に使用した書き起こし文より, 人手で作成している. 時刻 t において, 画像特徴量 $G(t)$, 音声認識における単語のスコア(cmscore)を $C(t)$ とすると, 統合後のスコア $P(t)$ は,

$$P(t) = \begin{cases} (G(t) + C(t))/2 & \dots \text{if keyword} = \text{Cardiac compression} \\ G(t) - C(t) & \dots \text{otherwise} \end{cases} \quad (2)$$

で求められる(図6). 発話内容が心臓マッサージに関するものであった場合, $P(t)$ は画像と音声の特徴量の平均とする. 心臓マッサージ以外のシーンに関するものであった場合は, 音声特徴量から, 画像特徴量を引いた値を $P(t)$ とした. これにより得られた値 $P(t)$ が閾値よりも高い場合, 時刻 t におけるシーンを決定する. 本研究では, 閾値を 0.5 に設定している.

最終的に得られたシーン情報に対して, 医療教育のシナリオより作成された処置手順の正解データと DP マッチングを行うことにより, 誤りを修正する.

表1 キーワードの例

シーン	キーワード
状況の確認	周囲 安全
意識の確認	もしも 大丈夫
援助の要請	百十九番 ナースコール
気道の確保	気道 見て
人工呼吸	フェイスシールド 循環サイン

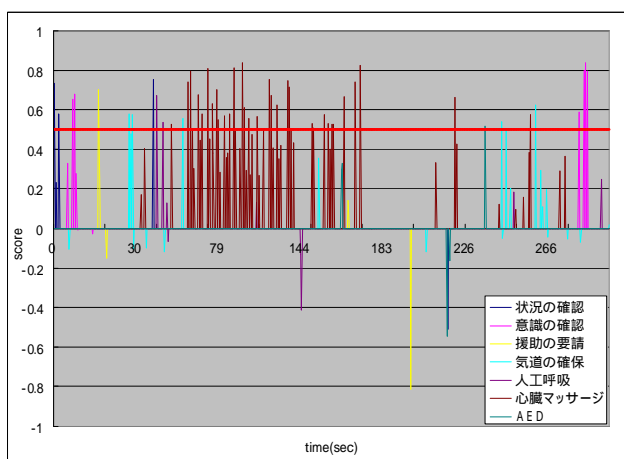


図6 統合後のスコア $P(t)$ の例

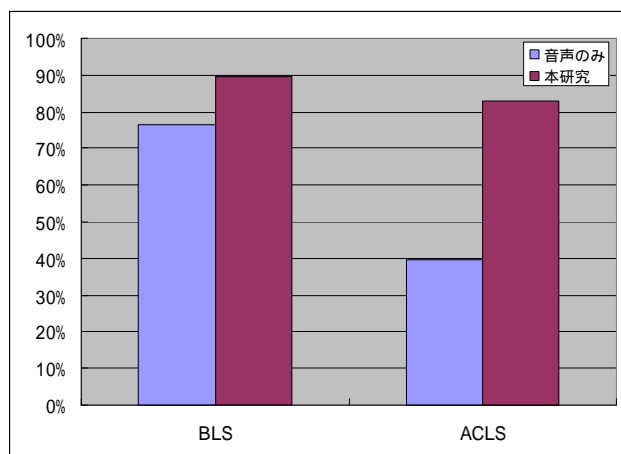


図7 シーン認識結果

3. 評価実験

得られたデータが正しくシーン分割されているか評価を行った。評価には、ACLS, BLS 講習会で主要な発話者にピンマイクを装着した状態で撮影されたものを用いた。使用したデータは 30 分程度であり、これらに対して、正解データとして手動でシーン分割したものを作成し、音声のみから作成したデータと、本研究のデータと比較した。また、画像情報では特定のシーンに対する特徴量を用いているため、画像情報のみでシーン分割を行うことは困難である。シーン分割の正解時間は、正しくシ

ーン分類されている時間の、データ全体時間に対する割合で求めている。音声のみを使用した場合が BLS で 69.6%, ACLS で 39.7%の正解率、画像情報も用いた場合が BLS で 89.5%, ACLS で 82.9%の正解率であり、改善が見られた(図7)。音声認識のみの結果を見た場合、ACLS の認識率が大幅に低くなっている。これは、BLS では一人または二人の発話者が演技を行うのに対し、ACLS 講習会では複数の発話者が存在している場合が多く、音声認識率が低下するためである。

4. まとめ

本研究では、音声情報、画像情報を用いて、医療教育映像の自動インデクシングを行った。音声情報のみの場合と、画像情報も用いた場合を比較した。その結果、画像情報も用いた方がシーン分割の正解率が向上した。

今後の課題は、より多くの画像特徴量を用い、統合方法を改善することで、各シーンに対して精度の高いシーン分割結果を得ることである。音声認識では、医療教育という限られた分野のため、利用できるデータが少なく、信頼性の高い結果を得ることができなかったため、より多くのデータを用いた言語モデル、単語辞書の作成が必要である。また、本システムを完全に自動化することも挙げられる。医療教育の講習会では、受講者が講習会へ赴き、指導者に評価を行ってもらわなければならないが、将来的には、オンラインで自動インデクシングを行うことにより、遠隔地からも講義を受けることができるようになると思われる。

5. 謝辞

岐阜大学大学院医学系研究科、岐阜県生産情報研究所の皆様には医療教育映像の撮影やシステムの作成などを通じ貴重な御助言を頂きました。また、本研究の一部は、文部科学省の岐阜・大垣地域知的クラスター創成事業の支援により行われました。以上の皆様には、感謝の意を表したいと思います。

参考文献

- [大附 03] 大附克年, 別所克人, 水野理, 松尾義博, 松永昭一, 林良彦: 音声認識を用いたマルチメディアコンテンツのインデクシング, 情処研報, SLP, 2003.
- [橋本 05] 橋本幸司, 速水悟: 音声認識を用いた e-Learning システムの評価, 人工知能学会第 19 回全国大会, 2005.
- [露木 06] 露木敏勝, 西岡正行, 三ツ橋史緒子, 岩瀬裕美子, 菅沼太陽: 医療教育における e-Learning の実践, 第 26 回医療情報学会連合大会論文集, 2006.
- [河村 07] 河村高守: 付与情報を利用した教育支援に関する研究, 岐阜大学工学研究科修士論文, 2007.
- [河原 05] 河原達也, 李晃伸: 連続音声認識ソフトウェア Julius, 人工知能学会誌 Vol.20 No.1, 2005.
- [Horn 81] B.K.P.Horn, B.G.Schunck: Determining optical flow, Artificial Intelligence Vol.17 pp.185-204, 1981.