料理番組における映像とテキスト情報の対応づけ

4U－7

浜田 玲子, 井手 一郎, 坂井 修一, 田中 英彦
{reiko,ide,sakai,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

近年、テレビやビデオ、WWW などを通してますます大量のマルチメディアデータが発信されるようになり、これらの膨大なデータを収集・整理し、効率の良い利用法を模索するための研究が盛んに進められている。最近では特にニュース番組などテレビ映像の索引づけや分類、スキミングといった技術に関する研究が多く行なわれているが、本研究ではこれらとは異なり、番組の内容に付随したテキスト教材の存在する料理番組に着目し、その統合的な再構成を目指す。

料理番組では多くの場合、番組内で料理方法を実演するとともに、内容をまとめたものを別途テキスト教材や WWW ページなどで公開している。一般に、料理番組などでは教材に記されない多くの情報が映像中に存在するが、放送映像を見ながら調理などをするのは困難であり、実際にはテキスト教材を見ながら行なうことになる。そのため、テキスト教材には記述されない映像中のノウハウを効果的に活用することが難しい。そこで、本研究では将来の台所への計算機の進出を見越し、テキスト教材中の情報に映像からの情報を対応付けることによって、教材に不足している情報を補ったマルチメディア統合データの再構成、さらには扱う対象が限定されていることや手順の不可逆性など固有の特徴を活かした新しいマルチメディア統合技術の提案を目指す。

2　関連研究との比較

はじめに、多くの研究がなされているニュースやドラマと料理番組の違いを述べる。著者らが今回最も着目しているのは、容易に入手可能なテキスト教材が存在する点である。一般に、画像認識により映像の意味的な内容を推測することは非常に困難であるが、このようなテキスト情報は画像や音声に比べ扱いやすく、またテキストの内容を認識処理に反映させることで、より的を絞った処理が可能になる。これまで、同様なテキストの存在する映像を扱った研究としては、ドラマ映像とシナリオの対応づけを行なう研究がある [1]。ここでは、ドラマにおいてシナリオと映像の対応するイベントがほぼ一対一に生起しているため、これらの間隔を非線形に伸縮することで最も最適な対応を求める DP マッチングにより対応づけを行なっている。しかし、料理番組においてはテキスト教材中の手順と異なる順序で番組が進行することが多く、映像とテキスト教材の時系列の順序が対応しないことがあるため、このような厳密な手法は用いることができない。そのため、料理番組の対応づけにおいては映像、音声の内容を解析したり番組の構成を参照するなど、様々なヒントを総合的に利用する必要がある。

また、ニュース番組における索引づけなどの研究 [2] においては、画像、音声など各メディアから独立にヒントを抽出し、それらを時間軸に沿って対応づけるなどの単純な統合技術を用いている。しかし料理番組においては扱う対象が限定される上、手順は基本的に不可逆であり、さらにテキスト教材が利用できるため、各メディア間でのフィードバックを利用した、既存手法にはないより高度な統合手法を実現できる可能性がある。

3　映像とテキストの対応づけ

3.1　対応づけ手法

テキスト教材においては、調理方法はいくつかの手順に分かれており、それぞれに手順番号がふられている。そこで本研究では、最終的には図 1に示すようなテキストベースの教材の各手順と、それらの手順に対応するビデオ映像の対応づけによるマルチメディアデータの再構成を目指している。

そのために、まず料理番組のビデオ映像における手順番号を抽出する。料理番組は様々な映像から構成されているニュース番組と異なり、一般的にはほとんどスタジオ内の映像で構成され、また進行は聞き手と料理人との会話形式で進められることが多い。

次に映像は図 2に示すように画像、音声、字幕からなる。手順や材料は声に出して説明されるため、音声は手順番号を推測するうえで大きなヒントとなる。また画像は、大きく (1) 手元のショット、(2) 人物ショット、(3)CG・フリップショットに分けられる。人物ショットはスタジオのほぼ全体が映されるが、手元のショットでは材料を調理する手元や道具が大映しにされるため、音

* "Associating Video and Text-book in TV Cooking Programs"
Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka
Graduate School of Engineering, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

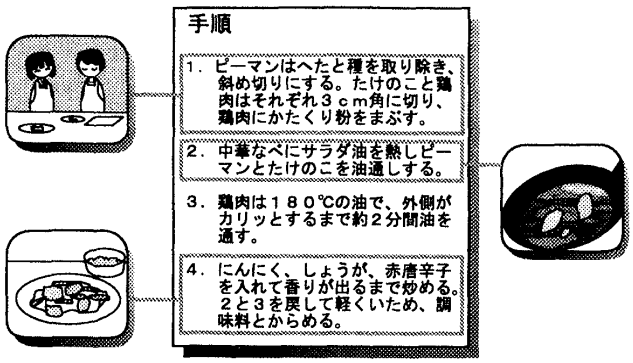

図 1: テキスト教材と映像の調理手順の対応づけ

声、テキストの内容から絞りこんでから対象を画像的に解析することができる。また番組によっては、字幕を利用することもできる。

一方、テキスト教材は、図2に示すように材料の一覧と手順からなる。音声中の単語や字幕の単語をテキスト教材の手順中に現れる単語と比較することで、手順番号を推測することができる。ところが手順中に出現する材料名が材料の一覧の表記としばしば異なることがあり(鶏もも肉→鶏肉など)、このような場合のために手順と材料名をあらかじめ対応付けておく必要がある。

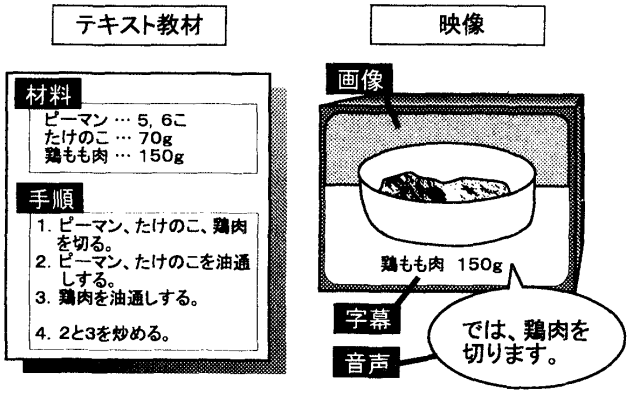

図 2: 料理番組におけるテキスト情報と映像情報

3.2 予備実験

テキスト教材と番組の音声データから映像の各ショットの料理手順番号を推測する簡単な予備実験を以下の通り行なった。なお、テキスト教材は料理番組の WWW ページから抜粋し、音声データは人が書き下した。

1. 各ショットの音声データに含まれる単語 (名詞・動詞・カタカナ語) を抽出し、このうちテキスト教材中にも出現している単語をキーワードとする。
2. それぞれのキーワードがテキスト教材中のどの手順に出現しているかを単語毎に調べ、延べ数で最も多く出現していた手順をそのショットの手順とする。
3. 分類不能だったショットのうち前後のショットが同じ手順に分類されたものは前後と同じ手順に分類する。また CG・フリップショットは手順解析に含めない。

10 分間の料理番組 3 回分 (ショット数 54) に対する実験を行なった。結果は表 1 に示す通り、平均約 6 割の分類に成功し、簡単なアルゴリズムでも音声のヒントからある程度映像を分類することが可能であることが示された。なお、成功率 $= \frac{\text{分類に成功したショット数}}{\text{全ショット数}}$ である。今後、分類アルゴリズムの改善や画像からのヒントを考慮して、分類成功率の向上を目指す。

表 1: 予備実験の結果:手順分類の成功率

番組	1	2	3	平均
成功率	50%	58%	68%	59%

4 まとめ

本稿ではテキスト教材付きの料理番組に着目し、料理番組を扱う意義とその特徴を通して映像とテキスト教材の対応づけ手法を提案した。また、その準備として簡単な予備実験とその結果について報告した。

今後の課題としては、画像からのヒントも利用して映像とテキストの対応づけを行ない、最終的には新たなマルチメディア統合技術手法の提案と構築を目指す。また、様々な料理番組のデータを蓄積することで、材料や嗜好などから希望の料理を抽出したり、料理の計画表を作成可能な料理データベースを作成するなど、計算機の台所への進出を踏まえた応用例を検討していく。

参考文献

[1] 柳沼 良知, 坂内 正夫, “DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応づけ手法の一提案” , 信学論, Vol.J79-D-II, No.5, pp.747-755, May 1996.

[2] A. Haupmann, M. Witbrock, “Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval” Intelligent Multimedia Information Retrieval, Mark T. Maybury, Ed., AAAI Press, pp.213-239, 1997.