6R-09

料理映像の構造解析による手順との対応づけ *

三浦 宏一[†], 浜田 玲子[‡], 井手 一郎[††], 坂井 修一[‡], 田中 英彦[‡]

[†,‡]{miura,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp, [††]ide@nii.ac.jp

[†]東京大学工学部 [‡]東京大学大学院工学系研究科 [††]国立情報学研究所

1 はじめに

近年のマルチメディアデータの増大に伴い、その解析がますます重要となりつつある。そのため複数メディアを統合的に処理する手法が注目されている。

我々は、このような統合メディア処理手法の研究の一環として、料理映像を題材とした研究を行っている [1]。料理テキスト教材は映像よりも内容解析が容易である一方、料理映像にはテキスト教材では表現しきれない有用な情報が含まれており、これらを統合することによって互いの情報を補完することができる。そこで我々は、料理映像とテキスト教材中の手順の対応づけを目指している。これにより、テキストと映像をリンクさせた、新たな構造化されたマルチメディアデータの生成が可能となる。

本稿では、そのような統合システムにおける映像処理、及び対応づけ処理について検討する。対象を料理映像に限定していることから、対象分野に特有の知識を活用することで、比較的簡単な処理により、高精度な結果を期待する。

2 関連研究

映像と外部テキストを対応づける研究として、ニュース映像のテロップ中の名詞と電子新聞記事の構造情報を利用して類似度を計算し、ニュース映像と新聞記事を対応づける研究 [2] が行われている。しかしこの研究では映像内容は考慮されていない。また、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付け手法 [3] は、シナリオとドラマ映像中の様々な情報を用いた対応づけを行っている。しかし、ドラマでは映像の順序とシナリオの順序とがほとんど一致するのに対し、料理番組ではしばしばテキスト中の手順と映像中の手順が入れ替わる点で、本研究と本質的に異なる。そのため、本研究では DP マッチングのような時系列の一致を利用した手法ではなく、複数メディアからの情報を効果的に統合し、対応づけを行う必要がある。

3 映像の構造解析による手順との対応づけ

3.1 提案手法の概要

対応づけシステムの全体像及び本稿で扱う部分の構成を図 1に示す。図 1に示す通り、本システムでは、映像の構造解析とテキスト教材の構造解析を並行して行い、その結果を利用して、両者の対応づけを行う。本稿では、テキスト教材の構造解析については既存手法 [1] を利用し、映像解析部と統合処理部の検討を行う。

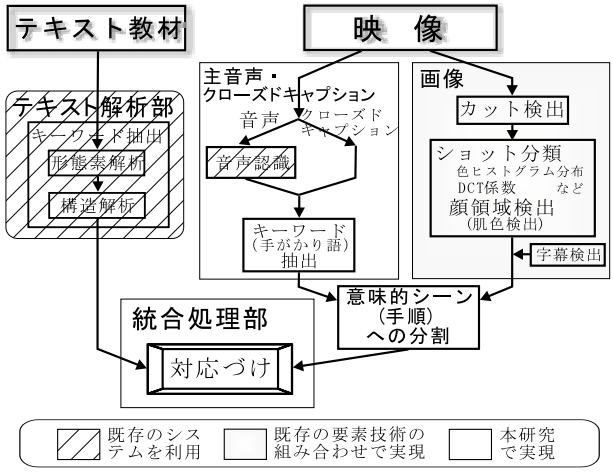

図 1: 料理映像とテキスト教材の統合システムの構成

映像の構造解析は、画像と音声の両方から進める。画像処理に関しては様々な要素技術が研究されており、本研究では、それら既存の手法を効率良く組み合わせることにより高精度な処理を目指す。また音声処理に関しては、音声認識は行わず、主音声の書き下しであるクローズドキャプションを利用し、これにテキスト処理を施す。そしてこれらの解析結果から、映像の構造を抽出する。

最後にそれぞれの解析結果を利用して、映像とテキスト教材の対応づけを行う。

3.2 映像の構造解析

映像の構造解析の目的は、映像の意味的シーンへの分割である。ここで映像の意味的シーンは、テキストにおける料理手順とほぼ対応する。本手法では、まず映像を機械的にショット単位に分割するが、ショットは意味的シーンとしては短すぎることが多い。そこで、同じ手順に含まれると推測されるショットを統合し、意味的シーンを再構成する。

---

* "Association with preparation steps by structural analysis of cooking video"

Koichi Miura[†], Reiko Hamada[‡], Ichiro Ide[††],
Shuichi Sakai[‡], Hidehiko Tanaka[‡]
[†]Faculty of Engineering, The University of Tokyo
[‡]Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[††]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

3.2.1 画像解析

料理映像のショットは、以下のように分類できる。

- A: 人物ショット
  - ▷ A1: 調理台から全身が映っているショット
  - ▷ A2: 上半身のアップショット
- B: 手元ショット
- C: 静止ショット

実際の料理映像のショット構成の例を図 2に示す。

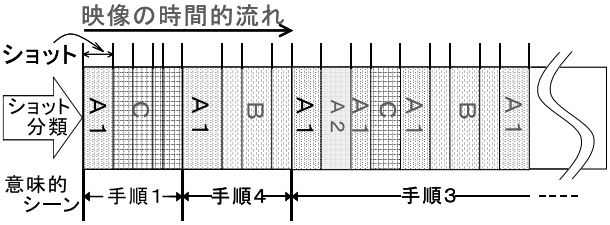

図 2: 料理映像のショット構成

ここで、手順の区切りの直後のショットに着目すると、その 90%以上が A の人物ショット、特に A1 であった。従って、映像を意味的シーンに分割する上で、ショット分類、特に人物ショットの検出が重要であると考えられる。

3.2.2 クローズドキャプションの利用

音声部分から、手順の区切りにある手がかり語を取り出した結果を表 1に示す。

表 1: 手順の区切りの手がかりとなる語

接続	例) では	使用例) では〜を加えます。
指示	例) これを	使用例) これを焼いていきます。
間	例) 〜の間に	使用例) その間に〜を切ります。
場所	例) ここに	使用例) ここに〜があります。
条件	例) 〜たら	使用例) 〜になったら移します。

手順の区切りの言葉は、「では」「まず」などの接続詞が主であると予想される。実際、手順の区切りの 4 割程度の部分にこのような接続詞がみられた。しかし、場合によっては接続詞を使わないこともあり、表 1に示した接続詞以外の手がかり語も考慮すると、手順の区切りの 8 割程度にこれらの言葉がみられた。

3.3 予備実験:人物ショットの検出

3.3.1 実験手順

予備実験として、ショット分類において重要である人物ショットの自動検出を行い、その性能を評価した。人物ショットは画像中の顔領域を以下の手順で抽出することにより検出を行う。

(i) 色情報 (修正 HSV 表色系) を用いて肌色領域を抽出
(ii) 検出された領域から一定の条件 (面積、位置など) により顔領域を決定

料理映像の 4 レシピ分 (計 233 ショット) について、上記の手法を用いて人物ショットの検出を行い、さらに A1 と A2 に分類した。なお、A1 と A2 は顔領域を抽出した後、その面積によって分類した。

3.3.2 実験結果

実験結果を表 2に示す。

表 2: 人物ショットの分類結果

ショットの種類	正解	正検出	誤検出	検出洩れ	再現率	適合率
人物ショット (A1)	81	79	5	2	98%	94%
人物ショット (A2)	43	38	4	5	88%	90%
その他 (B, C)	109	103	4	6	94%	96%

誤検出の主な原因は、壁や肉など、肌色に近い領域を顔として検出したことである。また検出洩れの主な原因は、顔の向きにより肌色領域が小さくなったことや、顔が背景にとけこんだことである。

顔領域の検出洩れの多くは、A1 に見られるが、A1 にはほとんどの場合複数の人物が存在するため、ショット分類の精度にはほとんど影響がみられなかった。本手法は意味的シーンの抽出が目的であるため、表 2の結果は実用的な精度であると考えられる。

3.4 対応づけ

映像を意味的シーンに分割した後には、テキスト教材との対応づけを行う。対応づけは主に、シーン中のクローズドキャプションから材料名などの名詞とそれに対する動詞をキーワードとして抽出し、テキスト教材中のキーワードと照合することにより行う。また、映像の時間的順序も考慮することによって、より高度な対応づけが期待できる。

4 おわりに

本稿では、料理番組における映像とテキスト教材の対応づけを実現するための映像処理及び対応づけ手法を検討した。また、映像処理部分の予備実験として、人物ショットの検出実験を行い、簡単な手法により高精度の結果が得られることを示した。今後は、映像処理部分の改善及び対応づけ手法の詳細を検討する。また、将来には、解析結果を利用した索引づけやデータベース作成など、様々な応用が考えられる。

参考文献

[1] R. Hamada, I. Ide, S. Sakai, H. Tanaka: "Associating Cooking Video with Related Textbook", Proc. ACM Multimedia 2000, pp.237-241, Nov. 2000.

[2] 渡辺靖彦, 岡田至弘, 角田達彦, 長尾真: "TV ニュースと新聞記事の対応づけ", 人工知能学会誌, Vol.12, No.6, pp.921-927, Nov. 1997.

[3] 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付け手法の一提案", 電子情報通信学会論文誌, Vol.J79-D-II, No.5, pp.747-755, May 1996.