1V-05

料理映像における画像認識手法の検討

浜田 玲子, 坂井 修一, 田中 英彦
{reiko,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、我々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している [5]。現在我々が提案している統合システムを図 1に示す。

本稿では、このようなシステムにおける画像解析部の検討を行なう。本研究では図 1のシステムを前提としているため、テキストの解析結果を映像処理に反映させ、画像中に登場する材料名や道具をあらかじめ予測し、絞り込んだ解析を行なうことが可能である。

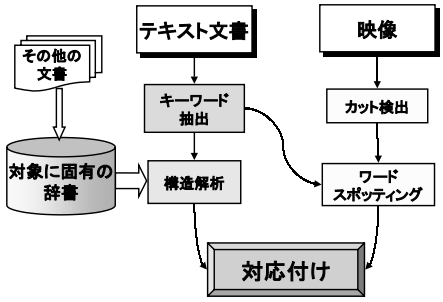

図 1: 料理映像とテキストの統合システム

また、我々の最終的な目標は統合システムの構築であるため、その画像認識部にあたる本研究では、既存の要素技術をできる限り効率良く利用することを検討する。しかし、料理映像という比較的特殊な映像が対象であるため、完全に汎用な要素技術だけですべての必要な認識を実現することは困難が予想される。そのため、既存の技術では不足な部分については、新たな画像認識手法を検討する必要がある。

本論文では、対象となる料理映像の構成を紹介し、その解析において既存の要素技術の適用が可能な部分と、新たな手法が必要となる問題について検討する。

2 料理映像における画像認識

2.1 画像認識の目標

テキストとの対応づけを目指す映像解析において、最終的に必要となるのは意味的なシーンの抽出である。図 2に示す通り、料理映像における意味的なシーンは、対応する料理手順における 1 ステップか、あるいは「切る」「ゆでる」といった調理動作ひとつぶんであると考えられる。しかし、映像の階層構造における最小単位であるショットは、一般的に意味的なシーンの単位としては細か過ぎることが多い。一方で、ほとんどの場合はそのような意味的なシーンの区切りはカットのうちの一つと一致する。

そこで、本研究では、まず映像をショットに分解し、これらの細かすぎるショットを意味的に連続すると考えられるもの同士で再統合することによってシーンの検出を行なう。またその過程で、対応づけに必要となる情報を映像から抽出することを目標とする。

2.2 料理映像の構成

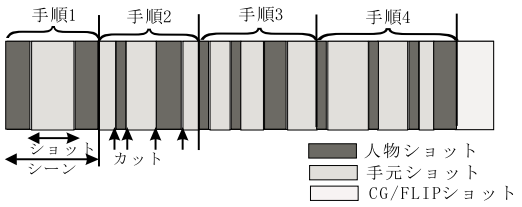

図 2: 料理番組映像の構造

料理映像において検出されるショットは、大きく (1) 手元ショット、(2) 人物ショット、(3)CG/FLIP ショットに分けることができる。図 3に示す通り、手元ショットとは、料理をする手元や料理道具（フライパン、包丁など）、素材などが大映しになっているショット、また人物ショットとは、人物の全身や上半身が映り、料理に関する解説などを中心に行なっているショット、そして CG/FLIP ショットとは、CG あるいはフリップによって文字（場合によっては図や写真）のみが映されている

* "An Study on Image Recognition for Cooking Programs"
Reiko Hamada, Shuichi Sakai, Hidehiko Tanaka
Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ショットである。

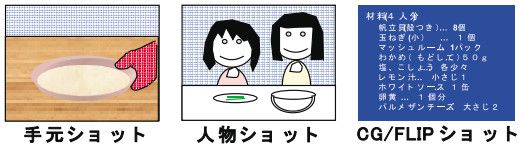

図 3: ショットの分類

次に、料理映像全体の構成を述べる。図 2に示す通り、各手順に対応するシーンの中には複数のショットが含まれる。調理中に関してはほとんどすべての映像が手元ショットか全身ショットに含まれる。一方で CG/FLIP ショットは映像の最後に挿入され、素材や手順などを表示することが多い。本研究では、これらショットの分類や画像の特徴を利用して、図 2に示す手順の切れ目や、どの手順に対応するかの情報を抽出する。

2.3 映像認識手法

カット検出

カット検出はもともと、索引づけや検索における画像処理の中でも、最も一般的かつ重要な要素技術である。本研究でも、映像の区切り検出の下準備としてカット検出を行なう必要がある。カット検出手法としては、色ヒストグラムや色コレログラムで画像の色調変化を検出する手法など様々なものが検討されているが、本システムでは精度の良い DCT クラスタリングを利用するカット検出手法 [1] を導入する。多くの料理番組はスタジオ内の理想的な照明条件下で撮影されるため、高いカット検出率が期待される。

ショット分類

カット後には、前節で述べたようにショットを分類する。我々は現在、ショットの完全自動分類を検討している。まず、CG/FLIP ショットには動きがないため、一定時間以上静止する映像を検出することで自動分類可能である。また、手元ショットと人物ショットの分類については、既存の顔検出手法 [1, 2] を適用することによって、顔の存在する映像を人物ショット、それ以外を手元ショットというように分類が可能であると考えられる。また、手元ショットは人物ショットに比べて動きが大きいことが多いため、これを利用することもできる。

ショットの分類がなされた後には、手元ショットと人物ショットの出現パターンの分析により、シーンカットの情報を得ることが考えられる。現時点では、シーンカットの直前・直後は人物ショットになることが多いということが分かっている。

対象認識

本研究では、映像に対応するテキスト教材から、対象とする映像中に現れる素材や道具を予測することができる。また、料理という限られた範囲が対象であることから、対象となる素材や道具のデータベースを作成することが可能である。その中から特に特徴的な色・形を持つ素材を認識することで、映像内容の情報を得ることができる。また、認識手法を単純にすることにより、ある程度の認識精度を期待することができる。

さらに、現在は、手の動きを利用した認識を検討中である。ジェスチャー認識の分野では、手や人間の身体の様々な動きについての研究がなされている [3, 4]。しかし多くの場合、これらの手法は専用の背景や決まった動きが前提であったり、事前に背景を撮影しておき、後ほど差分をとるような場合が多い。しかし本研究ではジェスチャー認識用の画像ではないため、手の背景は一般的には台所である。このような背景における自然な手の動き（特に調理動作）の認識について、現在は HMM の利用などを含めた効果的な手法を検討中である。

3 まとめ

我々は、料理映像とその補助的な料理テキスト教材の手順の対応づけを目標として、料理映像とテキストの統合システムの実現を目指している。本稿では、そのようなシステムの画像処理部について検討を行なった。

まず料理映像の特徴をいくつか紹介し、その特徴を利用した画像処理に関して、既存手法を利用する処理、あるいは今後必要となる手法に関して検討した。

今後は、ショット分類方法の再考、手動作認識手法の検討などを行ない、具体的かつ効果的な画像処理システムの構築を検討する。

参考文献

[1] 有木 康雄, 杉山 善明, 石川 則之, 寺西 俊裕, 櫻井 光康: “ニュース映像中の記事に対する音声・文字・映像を用いた索引づけと分類”, 信学技報, PRMU96-97, pp.31-38, 1996.

[2] S. Satoh, Y. Nakamura, T. Kanade: “Name-it: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing”, Proc. of IJCAI-97, pp. 1488-1493, 1997.

[3] T. Starner, A. Pentland: “Visual Recognition of American Sign Language using Hidden Markov Models”, *Intl. Workshop on Face and Gesture Recognition,* 1995.

[4] 西村 拓一, 向井 理朗, 野崎 俊輔, 岡 隆一: “白黒動画像からの形状特徴を用いたジェスチャのスポッティング認識システム”, 信学論 (D-II), vol.J81-D-II, No. 8, pp.1812-1821, Aug 1998.

[5] R. Hamada, I. Ide, S. Sakai, H. Tanaka: “Associating Video with Related Documents”, ACM-MM’99, Oct. 1999.