テキストからの制約に基づく料理画像中の物体検出 *

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

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

[†]東京大学工学部 [‡]東京大学大学院情報理工学系研究科

[††]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、増加しつづけるマルチメディアデータを効率良く利用するための解析技術が重要になりつつある。従来、各メディア単独の解析技術については盛んに研究されてきたが、それらの限界が認識され、複数メディアを統合的に解析する手法が注目されるようになっている。我々は、このような統合メディア処理により、既存の比較的単純な要素技術を統合的に利用した映像解析手法の確立を目指しており、その一環として料理映像を対象とした知的構造化の研究 [1] を行っている。

本研究ではその要素技術として、比較的解析の容易なテキスト教材やクローズドキャプション（文字放送字幕）などのテキスト情報からの制約により、料理画像中から知的構造化の手がかりとなり得る物体（特に料理素材）を効率的に検出する手法を提案する。本研究では、対象を料理番組に限定することにより、テキスト教材から得られる情報及び対象に固有の知識を最大限に活かした実用的な手法を提案する。

2 関連研究

画像処理のみによる一般的な物体検出は、従来数多く研究され、困難であることが知られている。

スポーツ映像を対象とした画像と音声の解析による索引付けの研究 [3] では、映像中の重要なイベントの前後における音響的特徴（特定の語の発話や観客の歓声など）の検出により画像処理部を起動することで、効率的にイベントを検出している。

また、ニュース映像中の人物の顔と人物名を対応付ける Name-It システム [2] では、画像中の顔の特徴とクローズドキャプション中の人物名の共起性に基く対応付けを行うことで、登場人物候補の曖昧性を解消している。

* "Object detection from cooking video by restriction from accompanying text"
Motomu Takano[†], Koichi Miura[‡], Reiko Hamada[††], Ichiro Ide[‡‡], Shuichi Sakai[‡], Hidehiko Tanaka[‡]
[†]Faculty of Engineering, [‡]Graduate School of Infomation Science and Technology, [††]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 料理画像中の物体検出

3.1 料理映像の特徴

映像は多数のフレームからなり、画像的に連続なフレームの集まりをショットと呼ぶ。本研究で扱う料理映像中のショットは、図 1 に示す 3 種類に分類できる。こ

(1) 人物ショット (2) フリップショット (3) 手元ショット

図 1: ショットの分類

のうち（1）、（2）には調理手順の理解において重要となる事象は映っていないか、映っていても小さく、視覚的情報に乏しい。そこで本研究では（3）に着目し、料理映像において主体的役割を果たす料理素材の検出を目指す。

ここで、ショット分割や料理映像を対象としたショット分類に関しては既に研究されている [4] ため、以下では、手元ショットは予め抽出されているものとする。

3.2 提案手法の概要

本研究で提案する素材検出手法の概要を図 2 に示す。

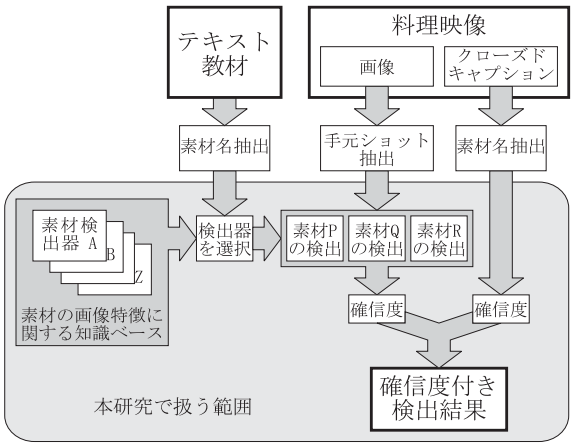

図 2: 料理素材検出手法の概要

素材検出のために、手元ショットに対して各検出器を適用する。検出器は素材ごとに用意し、予め素材の画像特徴に関する知識を記述しておき、画像中に素材が存在する確信度を出力する。

ところが、各ショットに対して全素材の検出器を適用すると計算量も誤検出数も増える。そこで提案手法では、テキスト教材やクローズドキャプションなどのテキストから得られる制約を利用して以下で述べるようにこれらの問題に対処する。

3.3 画像特徴による素材検出

素材検出器は、予め定義された素材の画像特徴に関する知識に基づき、手元ショット中から条件を満たす画像特徴をもつ領域を抽出し、その領域が目的の素材である確からしさ（確信度）を出力する。画像特徴には様々なものがあるが、現時点では料理素材を最も明確に特徴付ける色情報を利用する。素材領域の抽出は、検出器が知識としてもつ色分布に対してマハラノビス距離が閾値 d_{th} 以下である色の画素を取り出すことにより行う。複数の領域が検出されたときは、最大の確信度をもつ領域を選択する。

図 3 に料理素材の色分布の例を示す。

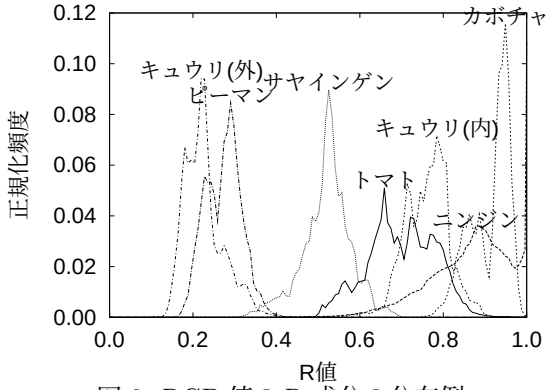

図 3: RGB 値の R 成分の分布例

抽出された領域 D の確信度 c_D を以下の式で定義する。

$$c_D = k \cdot \frac{\text{領域 } D \text{ の面積（画素数）}}{\text{画面全体の面積（画素数）}}$$

ただし k は正規化定数であり、学習用画像に検出器を適用したときに確信度の平均が 1 になるように定める。あるレシピの手元ショット中の 753 画像に対して、$d_{th} = 2.0$、素材の有無を判断するための確信度の閾値 $c_{th} = 0.1$ としてニンジンの検出実験を行ったところ、再現率 49%、適合率 11%という結果を得た。

3.4 テキストからの制約による検出精度向上

3.4.1 テキスト教材からの制約

3.3 のニンジン検出実験で適合率が低かったように、素材検出器だけでは画像特徴が似ている素材を区別できないため、料理番組に付随するテキスト教材の制約により検出精度を高める。制約としては、テキスト教材中の各レシピに掲載されている素材の一覧を利用する。映像中の素材は基本的にこの一覧に含まれるもののみであるため、一覧中の素材の検出器のみを適用して絞り込むことで検出精度の向上が期待できる。

3.4.2 クローズドキャプションからの制約

一方、クローズドキャプションが存在する番組に対しては、そこから得られる情報によりさらに検出対称を絞り込める可能性がある。素材の画像への出現と、素材名のクローズドキャプションへの出現にはある程度強い共起性があると考えられる。そこで、素材検出器による確信度とは別に、素材名がクローズドキャプションに出現した時刻付近で大きくなるような確信度を導入し、両者を総合することで精度の向上を目指す。素材名の抽出は [1] で作成した辞書を用い、マッチングにより行う。クローズドキャプションによる精度向上の例を図 4 に示す。

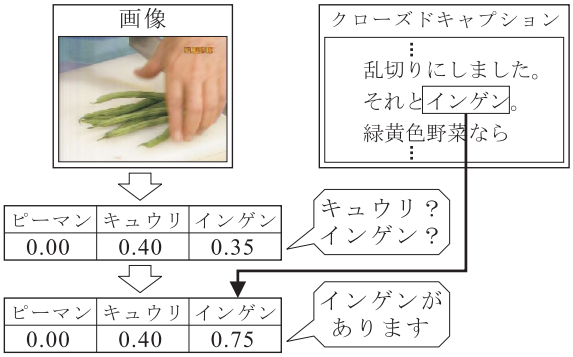

図 4: クローズドキャプションからの制約による曖昧性解消

4 おわりに

本稿では、料理映像中から、付随するテキスト教材やクローズドキャプションなどのテキストからの制約を利用して効率的に料理素材を検出する手法を提案した。今後は、より多くの画像特徴の導入による検出器や確信度の改良のほか、実際に知的構造化や索引付けに適用し、提案手法の効果を検証する。

参考文献

[1] 浜田玲子，井手一郎，坂井修一，田中英彦：“料理テキスト教材における調理手順の構造化”，信学論（D-II），vol.J85-D-II, no.1, pp.79-89, Jan. 2002.

[2] Shin'ichi Satoh, Yuichi Nakamura, Takeo Kanade: “Name-It: Naming and detecting faces in news videos”, IEEE Multimedia, vol.6, no.1, pp.22-35, Jan.-Mar. 1999.

[3] Yuh-Lin Chang, Wenjun Zeng, Ibrahim Kamel, Rafael Alonso: “Integrated image and speech analysis for content-based video indexing”, Proc. IEEE Multimedia 1996, pp.306-313, June 1996.

[4] 三浦宏一：“料理映像の構造解析による手順との対応づけ”，卒業論文，東京大学工学部電子情報工学科，Mar. 2001.