2C1-02

テキスト情報からの制約に基づく料理画像中の物体検出

Object detection from cooking video by restriction from acompanying text information

高野 求[*1] 三浦 宏一[*1] 浜田 玲子[*1] 井手 一郎[*2] 坂井 修一[*1] 田中 英彦[*1]
Motomu Takano Koichi Miura Reiko Hamada Ichiro Ide Shuichi Sakai Hidehiko Tanaka

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]国立情報学研究所
National Institute of Informatics

We are trying to associate television cooking videos with accompanying textbooks. Although we have developed a method for the association based on general image features, if the existence of cooking materials in videos could be detected automatically, the association should be more accurate. In this work, we propose a method to efficiently detect cooking materials from cooking videos by image processing with the help of restrictions from text information; lists of cooking materials in textbooks and time stamps at which names of materials appear in the closed-caption text are used as restrictions that disambiguate the image recognition results. An evaluation experiment showed that cooking materials can be accurately detected by image processing using restrictions from materials that appear in the textbooks, and that information from closed-caption texts can be helpful.

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、マルチメディアデータ（映像）を構成する画像、音声、テキストの解析技術については個別に研究されてきたが、単一メディアによる解析の限界が認識され、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た。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の知的構造化を目指しており、その研究の一環として料理映像を対象として教材テキストとの対応付け [1][2] を目指している。

この研究では、メディア間の対応付けのための様々な手がかりが必要とされている。例えばテキスト教材と映像中の調理動作の対応付けにおいて、動作と映像中の物体との間の制約を用いることにより、対応付けの性能向上が期待される。そこで本研究では、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）などのテキストメディアから得られる制約により、料理画像中から、対応付けの手がかりとなり得る物体（特に料理素材）を画像処理により効率よく検出する手法を提案する。ここでは、対象を料理番組に限定することにより、テキスト情報からの制約及び対象に関する知識を最大限に活かした実用的なシステムの構築を目指す。

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、一般に困難であることが知られている。

画像と音声の解析による映像の索引付けの研究 [3] では、スポーツ映像中の重要なイベントの前後における音声的特徴（アナウンス中の特定の語の出現や観客の歓声など）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 [4] では、クローズドキャプション中の人物名と画像中の顔の共起性に基づいて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。

連絡先: 高野 求, 東京大学大学院情報理工学系研究科, 東京都文京区本郷 7 丁目 3 番地 1 号, (tel)03-5841-7413, (fax)03-5800-6922, (e-mail)motom@mtl.t.u-tokyo.ac.jp

本研究では、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メディアから得られる情報を手がかりに、画像情報の曖昧性を解消し、画像処理量を削減する。

3. 料理映像の特徴

一般に、画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。本研究で扱う料理番組中のショットは、図 1 に例示するような、(1) フリップショット、(2) 人物ショット、(3) 手元ショットに分類できる。このうち、フリップショットや人物ショットには料理映像において検出すべき重要な物体は映っていないか、映っていても小さく、情報量に乏しい。そこで本研究では手元ショットに着目し、料理映像において特に重要な物体であると考えられる料理素材の検出を目指す。

(1) フリップショット　(2) 人物ショット　(3) 手元ショット

図 1: ショットの分類

4. テキスト情報からの制約に基づく料理画像中の物体検出

4.1 提案手法の概要

本研究において提案する料理素材検出システムの概要を図 2 に示す。

はじめに、画像をショットに分割し、手元ショットを抽出する。次に、抽出された手元ショットから簡単な画像処理により素材を検出する。素材検出のために、各素材ごとに検出器を用意し、素材の画像的特徴に関する知識を記述しておく。各素材検出器は、入力画像中に各素材が存在する確からしさ（確信度）を出力し、複数の検出器による検出結果を総合して、各ショットにおける素材の存在を判断する。

素材検出器による処理の際に、各画像に対して全素材の存在を仮定して各検出器を適用すると計算量も増え、また誤検出

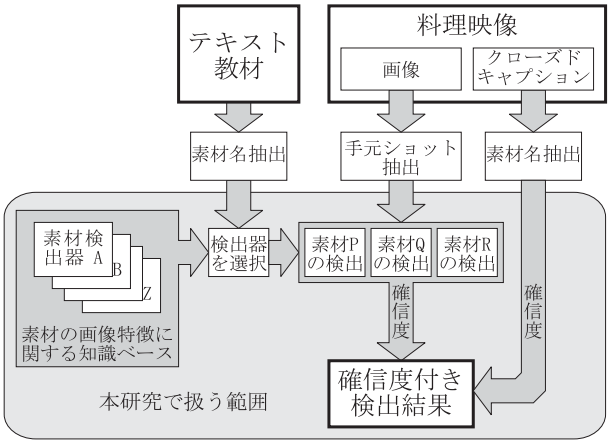


図 2: テキスト情報からの制約に基づく物体検出手法の概要

の可能性も高くなる。そこで本研究では、テキスト教材から得られる情報を利用して、その料理に使用される素材の検出器のみを適用することで、これらの問題を解消する。

また、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とに相関があると仮定して、クローズドキャプション中の素材名の出現により画像のみによる確信度を補うことを図る。

なお、提案手法のうち、カット検出及びショット分類に関してはすでに研究がなされている [5] ため、以下の実験では、手元ショットの画像が得られているものとして、評価を行なう。

4.2 画像特徴による物体検出

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。画像特徴としては色情報を利用する。素材領域は、検出器が知識としてもつ色分布に対する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抽出する。複数の領域が検出されたときは、最大の確信度をもつものを選択する。

確信度の算出には、検出された領域の面積に基づく指標を用いる。素材 M の検出器のもつ色分布とのマハラノビス距離が閾値 d_{th} 以下の色の領域 $R_{M1}, R_{M2}, R_{M3}, \ldots$ が抽出されたときの、指標 s_M を次式で定義する。

$$s_M = \max_{i=1,2,3,\ldots} \frac{\text{領域 } R_{Mi}\text{の画素数}}{\text{フレーム中の画素数}}$$

ここで、フレーム中に映るときの面積は素材ごとに偏っている可能性がある。この影響を排除するため、素材 M の映っている学習用画像に対する s_M の平均値 $\bar{s}_M$ を導入し、素材 M に対する確信度 c_{iM} を次式で定義する。

$$c_{iM} = \min\left(\frac{s_M}{\bar{s}_M}, 1\right)$$

このように定義した確信度を実際の料理レシピに適用した例を図 3 に示す。

素材検出器は確信度 c_{iM} を閾値 c_{th} と比較し、$c_{iM} \geq c_{th}$ の場合にフレーム内にその素材が存在すると判断する。ショット内の過半数のフレームで検出されると、そのショットにその素材が映っていると判断し、最終的な出力とする。

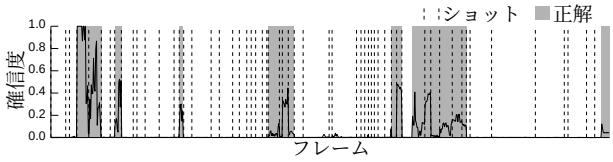


図 3: 素材の存在の確信度の分布の例

4.3 テキスト教材からの制約による検出対象の限定

素材検出器のみでは、画像特徴の類似している素材を区別できないことがある。そこで、料理番組に付随するテキスト教材から得られる情報を手がかりにして曖昧性を解消する。

料理番組のテキスト教材には、各レシピで使用する素材の一覧が掲載されており、料理番組中に登場する素材は基本的にこの一覧の中に含まれるもののみである。そこで提案手法では、全ての検出器を適用するのではなく、この一覧にある素材の検出器のみを適用する。これにより、その料理で使用されない素材の誤検出を排除できる。また、余分な画像処理のための計算量を減らすこともできる。

このようなテキスト教材からの制約による検出性能向上の例を図 4 に示す。ピーマン検出器にキュウリの映っている画像を入力すると高い確信度を出力し誤検出となってしまうが、「このレシピではピーマンは使用しない」というテキスト教材からの制約の利用により誤検出を防ぐことができる。

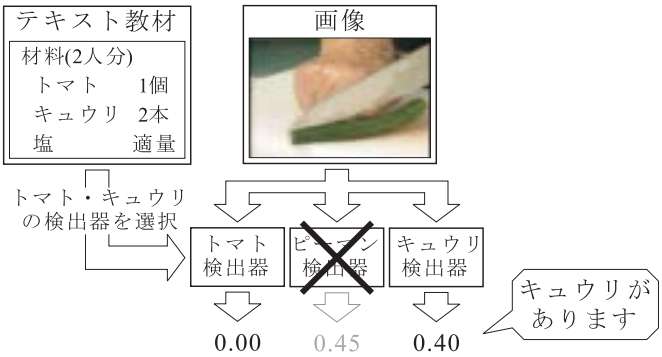


図 4: テキスト教材からの制約による曖昧性の解消の例

4.4 クローズドキャプションからの制約による検出の補助

テキスト教材の素材一覧表を用いても、1 つのレシピ内で画像特徴の類似した複数種類の素材が出現する場合は、曖昧性が解消できないため、クローズドキャプションから得られる情報を用いて検出性能の向上を図る。

主音声（クローズドキャプション）への素材名の出現と、画像への素材の出現には相関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する。素材検出器により、ショットに実際に映っている素材と画像特徴の類似した複数の素材が検出されたときに、ここで導入する確信度が検出器による確信度を補助することで検出性能が高まることを期待する。

その例を図 5 に示す。キュウリとインゲンを使用するレシピ中にインゲンが映っている画像があると、キュウリ・インゲン検出器はともに高い確信度を出力するが、付近のクローズドキャプション中にインゲンが出現していると、画像に映っているのはインゲンである可能性が高いと判断できる。

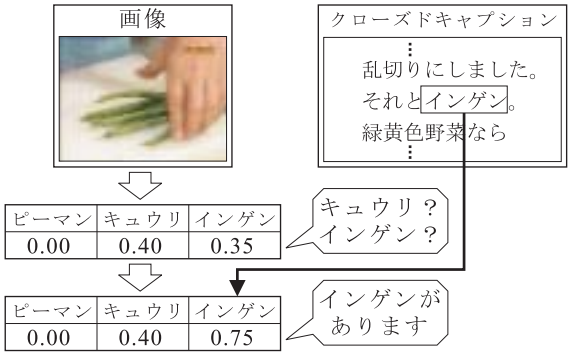


図 5: クローズドキャプションによる精度向上の例

ここでは、手元ショットのクローズドキャプション中に素材名が出現したときはそのショット中に、人物ショットのクローズドキャプション中に素材名が出現したときはその前後の手元ショット中に、素材が出現している可能性が高いと仮定し、これに基づいて確信度を定義する。

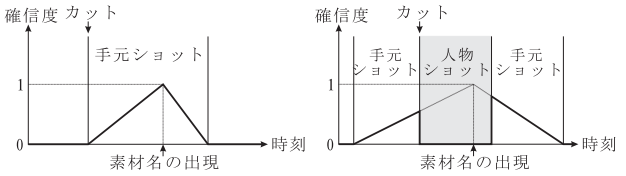
S を手元ショット、S より前にある一番近い手元ショットの直後のショットから、S より後にある一番近い手元ショットの直前のショットまでの区間を S' とする。S' 中のフレーム n のクローズドキャプションに素材 M の素材名が出現したとき、S 中のフレーム f における、この出現に起因する素材 M の確信度 $p_M(f,n)$ を以下の式で定義する。ここで、S の先頭のフレームをフレーム h、末尾のフレームをフレーム t とする。

$$p_M(f,n) = \begin{cases} \dfrac{f-h+1}{n-h+1} & f < n \text{ のとき} \\ \dfrac{t-f+1}{t-n+1} & f \geq n \text{ のとき} \end{cases}$$

S' 中のフレーム $n_1, n_2, n_3, \ldots$ のクローズドキャプション中に素材 M の素材名が出現したときのフレーム f における確信度 $c_{tM}(f)$ を以下の式で定義する。

$$c_{tM}(f) = \max\{p_M(f,n_1), p_M(f,n_2), p_M(f,n_3), \ldots\}$$

このように定義される確信度分布の例を図 6 に示す。また、実際の料理レシピに適用した例を図 7 に示す。なお、ここでは、フリップショットや人物ショットを省いて表示している。



(1) 素材名が手元ショット中に出現したとき　(2) 素材名が人物ショット中に出現したとき

図 6: クローズドキャプション中の素材名の出現による確信度

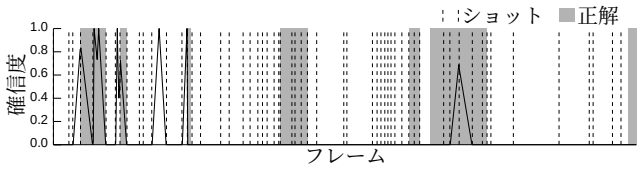


図 7: クローズドキャプション中の素材名の出現による確信度の分布の例

5. 素材検出実験

5.1 画像特徴による検出とテキスト教材からの制約による検出精度の向上

4.2 で述べた素材検出器を実際に作成し、評価のために検出実験を行った。また、4.3 で述べたテキスト教材からの制約を導入し、その効果を検証した。

(1) 実験条件

作成した検出器の一覧を表 1 に示す。素材の色分布情報は、異なる 3 番組から各素材ごとに 5〜20 画像程度を使用し、画像中から素材領域を人手で切り出して収集した。また、確信度の算出に必要な、素材領域の平均的な面積は、異なる 2 番組から各素材ごとに 5〜40 画像程度を使用し、作成した素材検出器を実際に適用して求めた。なお、素材領域抽出のためのマハラノビス距離の閾値は $d_{th} = 2$、素材の存在の判断のための検出器の確信度の閾値は $c_{th} = 0.014$ とした。

表 1: 作成した素材検出器

ニンジン	レモン	カボチャ（果肉）
赤唐辛子	トマト	赤ピーマン
キュウリ（表皮）	ピーマン	小松菜
キュウリ（果肉）	サヤインゲン	

(2) 実験結果

実際の料理画像 1 番組 8 レシピ中の手元ショット（検出器の学習用画像を含まない 232 ショット、3,643 フレーム）に対し、素材検出器を適用して検出実験を行った。結果を表 2 に示す。正検出、検出漏れ、誤検出の各項目は、全ての素材について合計したものである。

また、加熱による料理素材の変色が素材検出に与える影響を排除するため、素材の過熱後の検出漏れを無視した場合の再現率を算出した。ここでは、各素材の加熱を開始したショットまでを検出対象とした。結果を表 3 に示す。

以上のように、素材検出器のみでは類似した色の素材を区別できないため、誤検出が多発する。提案手法では、テキスト教材からの制約の利用によってこの問題を改善する。ここではその効果を検証するため、上記の実験で使用した各レシピについて、素材一覧に出現する素材の検出器のみを適用した場合の適合率を算出した。結果を表 4 に示す。テキスト教材からの制約の利用により、適合率が 25%から 78%へと、大幅に向上して、提案手法が効果をあげていることがわかる。

表 2: 素材検出による検出精度（テキスト教材からの制約なし）

正検出	検出漏れ	誤検出	再現率	適合率
160	54	480	74.8%	25.0%

表 3: 素材検出による検出精度（テキスト教材からの制約なし，加熱前）

正検出	検出漏れ	再現率
52	9	85%

表 4: テキスト教材からの制約を利用したときの検出精度

正検出	検出漏れ	誤検出	再現率	適合率
160	54	44	74.8%	78.4%

5.2 クローズドキャプションからの制約による検出精度の向上

4.4 で述べたクローズドキャプションからの制約を導入し、その効果を検証した。

(1) 実験条件

5.1 の、素材検出器とテキスト教材からの制約による検出実験で、複数の素材が検出されたショットのうち、正検出と誤検出が混在していた 26 ショットに対し、クローズドキャプションによる確信度を導入し、その効果を検証した。

ここで、ショット S において素材 M が検出されたとき、ショット S の素材 M に対する、素材検出器による確信度 $c_M(S)$ とクローズドキャプションにより補強された確信度 $c'_M(S)$ を以下の式で定義する。

$$c_M(S) = \frac{1}{n}\sum_{f\in S} c_{iM}(f)$$

$$c'_M(S) = \frac{1}{n}\sum_{f\in S} \{c_{iM}(f) + w c_{tM}(f)\}$$

ただし、$c_{iM}(f)$、$c_{tM}(f)$ はそれぞれ、フレーム f の素材 M に対する、検出器による確信度、及びクローズドキャプションによる確信度である。また、n はショット S 内のフレームの総数、w はクローズドキャプションによる確信度の重みを表す定数である。ここでは、$w = 1$ とした。

(2) 実験結果

5.1 の素材検出器とテキスト教材を用いた素材検出実験で、正検出と誤検出の両方があったショットに対し、クローズドキャプションによる確信度を適用し、確信度が補強される様子を調べた。

各ショットについて、正検出に対する確信度と誤検出に対する補強前の確信度（素材検出器による確信度）と補強後の確信度の平均を求めた。結果を表 5 に示す。正検出に対する確信度と誤検出に対する確信度の差は、クローズドキャプションからの制約の利用により、0.16 から 0.27 へと大きくなり、正検出した素材に対する確信度が、クローズドキャプションによる確信度により補強されていることがわかる。

表 5: クローズドキャプションによる確信度の補強

	件数	補強前	補強後
正検出	31	0.36	0.53
誤検出	29	0.20	0.26

5.3 考察

表 2 をみると、素材検出器のみでは適合率が非常に低いことがわかる。一方、再現率は比較的高い値を示している。特に、検出対象を加熱前の素材に限定した表 3 の結果は高い検出性能を示している。

素材検出器のみでは誤検出が多いが、テキスト教材からの制約の導入により適合率は大幅に改善され、提案手法の有効性が示された。本実験では、素材検出器は 11 種類と少ないが、提案手法を実用的なシステムに適用する際には多くの素材検出器を使用する必要があり、テキスト教材からの制約により検出器を限定する意義はいっそう大きくなると考えられる。

また、クローズドキャプションによる確信度の導入により、正検出に対する確信度と、誤検出に対する確信度の差が開き、正検出に対する確信度が補強され、クローズドキャプションによる確信度の補強の有効性が示唆された。

6. おわりに

本稿では、テキスト情報からの制約を用いて料理画像中から料理素材の検出を行う手法を提案して評価した。まず、料理画像からの素材検出手法を提案し、素材検出器やクローズドキャプションによる確信度、テキスト教材からの制約について検討・評価実験を行い、素材検出器とテキスト教材により高い精度で検出できること、クローズドキャプションの利用により正検出の確信度が補強されることを示した。

今後の課題としては、素材検出の精度をさらに向上させるために、色情報以外の画像特徴の導入が挙げられる。また、ショット内に映っている素材数の推定ができるようになれば、確信度順に素材を選択することにより、より適切な検出結果が得られるようになると考えられる。

謝辞

本研究の一部は、科学研究費補助金（基盤研究（B）(2)）「料理映像を題材とするマルチメディア統合システムの構築とその応用」（課題番号: 14380173）による。

また、本稿に掲載した画像の一部は、電子情報通信学会パターン認識・メディア理解研究会 VDBWG により公開されている評価用映像メディアデータベース中の「ランクアップ Cooking」中から抜き出した。

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”, 電子情報通信学会論文誌（D-II）, vol.J85-D-II, no.1, pp.79-89, Jan. 2002.

[2] Reiko Hamada, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka: “Associating cooking video with related textbook”, Proc. ACM Multimedia 2000 Workshops, pp.237-241, Nov. 2000.

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

[4] 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.

[5] 三浦宏一, 高野求, 浜田玲子, 井手一郎, 坂井修一, 田中英彦: “料理映像の構造解析による調理手順との対応付け”, 電子情報通信学会論文誌（D-II）, vol.J86-D-II, 2003 掲載予定.