言語情報と映像情報の統合による物体のモデル学習と認識

加藤 紀雄　　柴田 知秀　　黒橋 禎夫
東京大学工学部　東京大学大学院情報理工学系研究科
{kato,shibata,kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

実世界情報、映像情報などの高度な利用のためには、その内容の構造的理解が必要である。我々は、料理番組映像の検索・要約を目的として、その発話の構造解析を行なっている [4] が、映像中の発話を、映像 (画像列) を参照せずに解析するのには限界があり、発話の文脈と映像から得られる現場の文脈の冗長性を利用しながら解析する必要がある。

しかし、現状では映像に何が映っているかを認識することは、強い作り込みを行わない限り相当に難しい。物体認識を行なうには、色・形状・大きさといったその物体の知識が必要であり、この知識をどのようにして得るかが問題となる。画像に人手でキーワードを付与したデータから対応付けを学習し、物体認識を行なう手法もあるが [1]、画像にキーワードを付与するには大きなコストがかかってしまう。そこで、本研究では、映像に対して、言語処理と映像処理を統合的に適用することによって、大量の映像から物体のモデルを教師なし学習し、それを用いて物体認識を行なう手法を提案する。

まず、モデル学習を行ないやすい、物体が大写しになっている画像を抽出し、その画像周辺の発話からキーワードを抽出することにより、画像とキーワードのペアを大量の映像から収集し、そこから物体モデルを構築する。そして、学習した物体モデルと談話構造解析を利用することにより、物体の認識を行なう。

2　関連研究

大量の映像から教師なしで物体モデルを学習するような研究はなく、画像にキーワードを付与した正解データから画像と単語の対応付けを学習している研究が多い。Duygulu らは、複数のキーワードが付与された画像をもとに物体認識を行なっている [1]。画像を領域分割し、領域と単語の対応付けを EM アルゴリズムを用いて学習している。Feng らは、キーワード付与された少量の学習データとキーワードの付与されていない大量の学習データをもとに Bootstrap 手法を用いて物体モデルを学習している [2]。

また、物体認識に関連するものとして、我々と同じ料理ドメインでは高野ら [5] の研究がある。まず物体のモデルを学習するために、物体が大写しになっているような画像を人手で与え、そこから色情報を抽出することにより、物体モデルを獲得する。そして、料理映像中の素材を認識をする際に、番組に付随するテキストやクローズドキャプションからの制約を加えることにより、認識精度を向上させている。物体のモデルを人手で与えている点と、画像とそれに近い発話を単純に対応付けている点で我々の研究と異なっている。

本研究では、人手でキーワードを付与したデータを用いるのではなく、映像中の発話を深く解析することによりキーワードを抽出し、大量の映像から物体モデルを自動学習する。そして、構築した物体モデルを用いて、談話構造解析の結果を参照しながら物体認識を行なう。

3　物体モデルの自動学習

物体モデルの自動学習の第一段階として、まず色情報 (RGB) を学習する。対象とするのは食材とし、包丁、まな板、鍋といった道具は扱わない。

モデルを学習するために、物体が大写しになっているような画像とキーワードのペアを収集する。以下では、大写しになっているような画像をアップ画像、アップ画像で最も焦点のあたっている領域を注目領域と呼ぶ。

3.1　アップ画像とキーワードのペアの収集

単純に動画像列から画像を切り出し、その時刻に近い発話から抜き出した名詞を対応付けてもよい学習データを得ることはできない。そこで、画像処理と、発話を談話構造解析し重要な単語を抽出する処理を行なうことにより、物体が大写しになっているような画像とキーワードのペアだけを収集する。解析の概要を図 1 に示す。

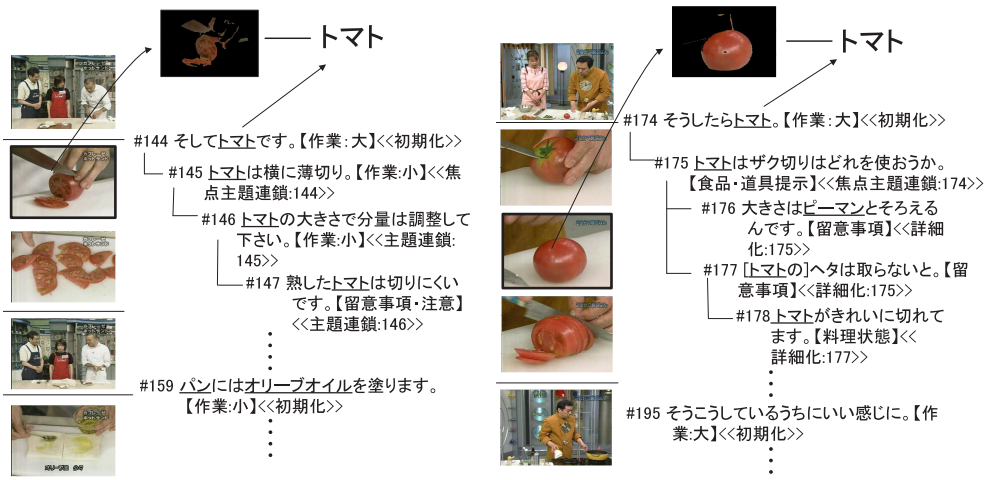

図 1: 注目領域とキーワードのペアの収集の概要

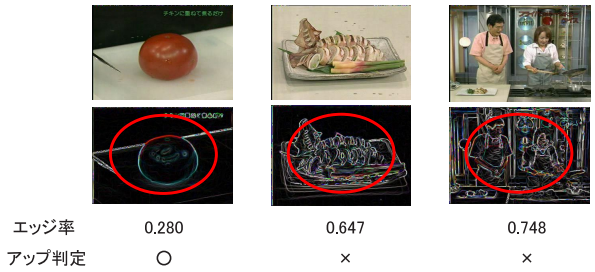

図 2: エッジ処理によるアップの判定

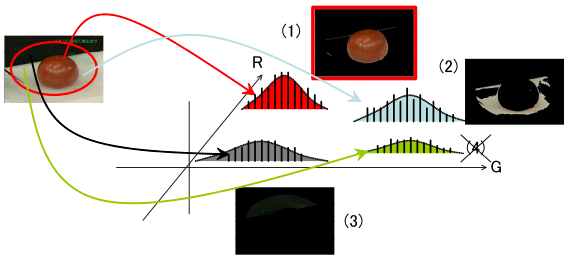

図 3: RGB 空間への写像と極大点の探索

画像とキーワードの対応付けは、ある瞬間の画像とその時 (またはその近く) の発話で行なうのではなく、ショット (単一のカメラから撮影されたフレームの集合) と談話構造木といったある程度広い範囲同志で行なう。それは、省略が多いことや、発話には作業の説明だけでなくコツや雑談などといった様々なタイプがあるため、きちんと対応がとれないためである。例えば、図 1 の右側の例で、トマトの画像に一番近いものを対応付けると、トマトの画像と「ピーマン」を対応付けてしまうことになる。

3.1.1 アップ画像と注目領域の抽出

以下のような画像処理により、物体が大写しになっているような画像を抽出し、そこから注目領域を抽出する。

1. ショットに分割

   隣接する 2 フレームのカラーヒストグラムの差が閾値以上であるところをカット点 (ショットとショットの境界) とし、ショット単位に分割する。

2. エッジ抽出によるアップ画像の抽出

   すべての画像に対して、3*3 の Sobel の一次微分でエッジ抽出を行ない、エッジ率 (エッジ検出された画素/ 全画素) を計算する。ただし、画面中心部を中心とする楕円内 (図 2 の楕円) だけを考える。ショット内で最もエッジ率の小さいものを選び、エッジ率が閾値 (0.5) を下回った場合、その画像をショットを代表するものとする。この処理により、食材が複数映っている画像や、人が映っている画像を除外することができる。

3. RGB 空間への写像と極大点の探索

   エッジ処理によりアップと判定された画像について、各画素を RGB3 次元空間に写像する。その後、3*3 のメディアンフィルタにより平滑化を行い、山登り法で極大点を探索する。極大値が閾値を下回るものは除外する。図 3 では (4) の領域が除外される。

4. 注目領域の抽出

   抽出された極大点のうち、画像の中心と重心との距離 (A)、重心から各点までの距離の分散 (B)、極大値の大きさ (C) を下式のように重み付けし、最もスコアの小さいものを注目領域として選ぶ。

$$A \times 0.6 + B \times 0.4 - C \times 5.0 \qquad (1)$$

   図 3 では、(1) の領域が注目領域として抽出される。

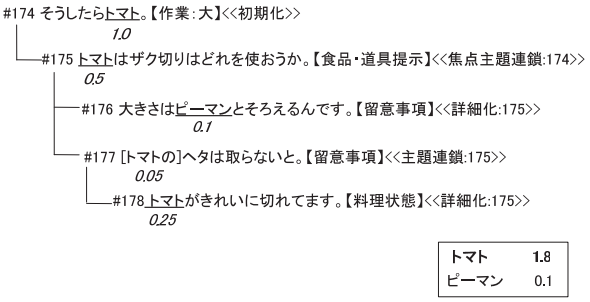

図 4: 談話構造木からキーワードの抽出

3.1.2 キーワードの抽出

映像に付随するクローズドキャプションの談話構造解析を行ない [4]、一つの談話構造木中で、最も重要な単語を選ぶ。談話構造解析の概要を以下に示す。

1. 入力文を JUMAN/KNP で形態素・構文・格解析する。
2. 自動構築した用言・名詞の格フレームを用いて省略の解析を行なう。
3. 節末の表層パターンを用いて発話タイプを認識する。発話のタイプは作業、料理状態、留意事項など 9 種類を考える。
4. 省略解析結果・発話タイプ・語連鎖・表層ルールを統合することにより文間の関係を明らかにする。

談話構造解析の結果、図 4 のような構造が得られる。図において、文中の括弧 ([]) で示されたものは省略要素が補われたものであり、節末の括弧 (【】) は発話のタイプ、括弧 (<<>>) は結束関係、親の節の文番号／節番号を示すものである。シソーラス [3] を用いて食材タグのふられた名詞に対して、る食材は一つで談話構造解析結果に基づき、以下のようなスコア付けを行ない、最もスコアの高いものを選ぶ。

$$Score = \sum_{w_i \in Tree} f_{utype}(w_i) \cdot 1/depth(w_i) \cdot f_{clause}(w_i) \cdot f_{anaphora}(w_i) \quad (2)$$

ここで、$f_{utype}(w_i)$ は発話タイプが<作業>、<食品・道具提示>、<料理状態>なら 1、それ以外なら 0.3 を返す関数、$depth(w_i)$ は木構造での深さを返す関数、$f_{clause}(w_i)$ は、w_i が主節にあれば 1、従属節にあれば 0.5 を返す関数、$f_{anaphora}(w_i)$ は省略解析結果なら 0.5、それ以外なら 1 を返す関数である。このスコア付けは、作業の説明を行なっている発話には重要な食材名がくる可能性が高いことや、談話構造木の最初の方の発話が重要であるなどといったことを反映したものである。

例えば、図 4 で、174 文目の「トマト」は、1.0(作業) × 1 / 1(木の深さ) = 1.0 点、176 文目の「ピーマン」は 0.3(留意事項) × 1 / 3 (木の深さ) = 0.1 点といったスコアが与えられ、談話構造木内でスコアを加算すると、トマトが 1.8 点、ピーマンが 0.1 点となり、トマトが選ばれる。

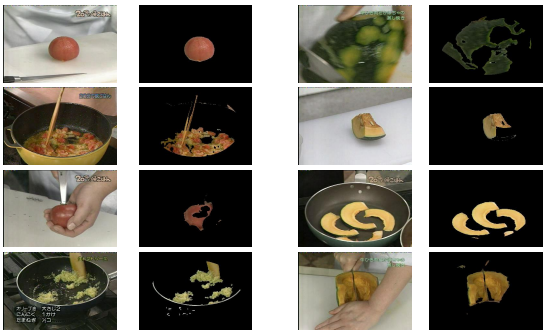

図 5: 収集された注目領域とキーワードのペア例

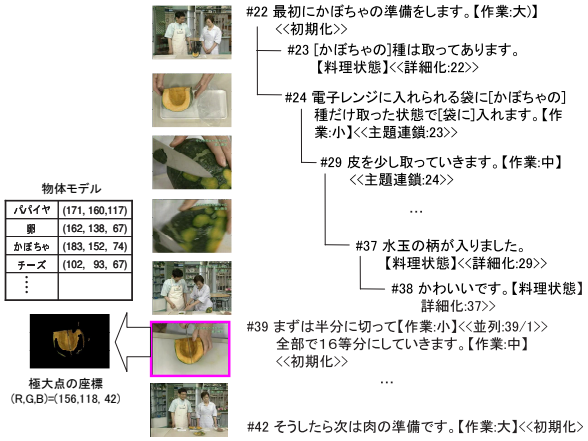

図 6: 物体認識

注目領域とキーワードの対応付け アップ画像が属するショットと、重なる時間が最も長い談話構造木を対応付けることにより、画像から抽出された注目領域と、談話構造木で選ばれたキーワードのペアを収集する。実際に得られた注目領域とキーワードのペアの例を図 5 に示す。図の左の列は原画像、右の列はそこから抽出された注目領域を示す。

3.2 モデルの構築

食材ごとに、注目領域の RGB データを計数し、最も頻度の高い RGB (の平均) を物体モデルとする。

表 1: 注目領域とキーワードの抽出の収集精度

食材名	抽出成功数/正解数/総数	ペア収集精度	注目領域抽出精度	物体モデル	
かぼちゃ	8/9/15	0.6	0.889	(183, 152, 74)	○
じゃがいも	8/8/25	0.32	1.0	(183, 162, 129)	○
れんこん	5/7/11	0.636	0.714	(221, 210, 180)	○
トマト	11/15/44	0.341	0.733	(117, 69, 40)	○
アスパラ	0/2/8	0.25	0.0	(31, 18, 8)	×
白菜	6/10/16	0.625	0.6	(106, 91, 76)	×

表 2: 物体モデルの学習の実験結果

全食材数	正解数	精度 (%)
94	60	63.8

4 物体認識

次に、得られたモデルを用いて、物体の認識を行なう。対象画像に対し、3.1.1 節で述べた処理を行なうことにより注目領域を抽出し、注目領域の極大点と物体モデルのユークリッド距離を計算する。そしてユークリッド距離の逆数に、3.1.2 節で述べた談話構造解析結果によるスコアをかけ、最もスコアの高いものを物体認識結果とする。図 6 の例では、画像情報だけを参照すると、パパイヤ、卵、かぼちゃなどが候補となるが、談話構造解析結果によるスコア付けにより、かぼちゃが選ばれる。

5 実験

NHK の「きょうの料理」の映像約 2 年分を用いて実験を行なった。まず、注目領域とキーワードの抽出の収集精度を表 1 に示す。正解数とは、収集された画像に食材が大写しである画像数、抽出成功数とは、注目領域が正しく抽出された画像数を表す。また、物体モデルの学習結果を表 2 に示す。

白菜などのような白っぽい食材の場合、平滑化を行なう際に食材の領域と背景やまな板が同一になってしまい、注目領域が正しく抽出されないことが多い。また、アスパラなどといった細長い食材の場合も、注目領域を抽出することに失敗してしまうことが多い。これは、注目領域を抽出する際に、物体の重心に密集しているものを優先しているからであり、この問題には、テンプレートマッチングを導入して対処する予定である。

次に、物体認識の実験を 5 番組に対して行なったところ、精度は 50.7%であった。現在は認識対象画像を単独で解析しているため注目領域の抽出で誤っているものが多い。認識する際にフレーム間差分を考慮することにより、注目領域を時系列で追跡する予定である。

6 結論

本稿では、大量の映像から、物体モデルを自動構築し、学習した物体モデルと談話構造解析結果を参照することにより、物体の認識を行なう手法について述べた。

今後は、色情報だけでなく、形状などといった特徴も学習し物体モデルを精密にするとともに、物体認識結果を省略解析・談話構造解析といった言語解析と統合する予定である。

参考文献

[1] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision(ECCV)*, pp. 97–112, 2002.

[2] HuaMin Feng and Tat-Seng Chua. A bootstrapping approach to annotating large image collection. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 55–62, 2003.

[3] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[4] Tomohide Shibata, Masato Tachiki, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida. Structural analysis of instruction utterances using linguistic and visual information. In *Proceedings of Eighth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2004)*, pp. 393–400, 9 2004.

[5] 高野求, 三浦宏一, 浜田玲子, 井手一郎, 坂井修一, 田中英彦. テキストからの制約に基づく料理画像中の物体検出. 情報処理学会第 65 回全国大会, 第 2 巻, pp. 255–256, 3 2003.

[6] 柴田知秀, 黒橋禎夫. 隠れマルコフモデルによるトピックの遷移を捉えた談話構造解析. 言語処理学会 第 11 回年次大会, 3 2005.