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

柴田知秀[†1] 加藤紀雄[†2,*1] 黒橋禎夫[†1]

近年の計算機・ネットワーク環境の進歩により，膨大な映像アーカイブが蓄積されるようになった．本研究では作業教示映像である料理映像を具体的題材とし，料理映像に現れる食材の物体モデルを自動学習し，それを用いて物体認識を行う手法を提案する．まず，物体がアップになっている画像を抽出し，その画像における注目領域を決定する．次に，画像の周辺の発話から重要な単語をキーワードとして抽出し，注目領域と対応付ける．このような注目領域とキーワードのペアを大量に収集することにより，物体モデルを構築する．物体モデルが構築された後，物体モデルの色情報と談話構造に基づく単語の重要度を考慮することにより，物体認識を行う．2 つの料理番組，計約 96 時間分の映像から物体モデルを構築したところ，約 100 食材の物体モデルが構築でき，その精度は 77.8%であった．また，そのモデルを利用して物体の認識を行ったところ，精度は F 値で 0.727 であった．

Automatic Object Model Acquisition and Object Recognition by Integrating Linguistic and Visual Information

Tomohide Shibata,[†1] Norio Kato[†2,*1] and Sadao Kurohashi[†1]

Recent years have seen the rapid increase of multimedia contents with the continuing advance of information technology. We focus on cooking TV videos, which are instruction videos, and propose a method for acquiring object models of foods and performing object recognition based on the acquired object model. Close-up images are first extracted from image sequences, and an attention region is determined on the close-up image. Then, an important word is extracted as a keyword from utterances around the close-up image, and is made correspond to the close-up image. By collecting a set of close-up image and keyword from a large amount of videos, we can acquire the object model. After that, object recognition is performed based on the acquired object model and discourse structure. We conducted an experiment on two kinds of cooking TV programs. We acquired the object model of around 100 foods and its accuracy was 77.8%. The F measure of object recognition was 0.727.

1. はじめに

近年の計算機・ネットワーク環境の進歩により，放送映像や講義映像，Web コンテンツなど，膨大な映像アーカイブが蓄積されるようになった．検索や要約など，映像アーカイブを高度に利用するためには，その内容理解が必要となる．

実世界の情報把握の中でまず必要なことは，人の発話と視覚情報を統合的に扱い，ある物体への言及が発話にあり，その場にその物体があるということを対応付けて把握することである．発話の中には，その場に存在しない物体への言及もあり，また，その場にも重要でない物体もあるので，発話で言及され，かつ，その場にある物体と，そうでない物体を区別することが重要となる．このような対応付けは映像の内容理解への第 1 歩であり，映像自動インデキシングを高度化することができる．

さらに，そのために必要なこととして，物体の特徴を自然に学習できることがきわめて重要である．ここで，「自然な学習」の意味するところは，人手による余分なラベル付けなどは行わず，与えられた実世界情報アーカイブのみから学習することである．これは，実世界の多様性に対して知識を作り込むことは困難だからである．たとえば，Duygulu ら[1] は，多くの画像に人手でキーワードを付与し，それぞれの画像を領域分割し，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が，このような手法では，画像にキーワードを付与することに大きなコストがかかっ

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo
*1 現在，農林中央金庫
Presently with The Norinchukin Bank

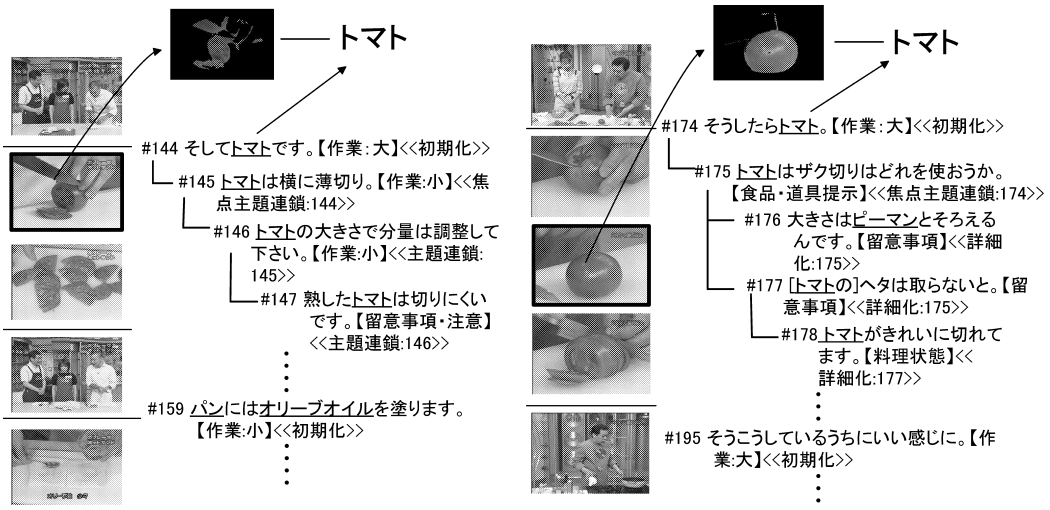


図 1 注目領域とキーワードのペアの収集の概要
Fig. 1 An overview of collecting sets of an attention region and a keyword.

てしまう．また，高野ら[2] は，料理映像中の素材を検出するために，あらかじめ映像中から料理素材が映っている画像を人手で用意し，素材の色分布情報を得ており，コストがかかってしまう．

このような背景のもと，本論文では，人の発話と視覚情報を統合的に扱う最初の試みとして，映像の中でも再利用価値の高い作業教示映像，具体的には料理映像から物体，具体的には食材の色モデルを学習し，それを用いて物体認識を高精度化する手法を提案する．

この物体のモデル学習における重要な仮定は，発話中の重要部分（キーワード）と映像中の重要部分（注目領域）は多くの場合に一致しているということである．この仮定のもとにキーワードと領域を対応付けることにより，物体モデルの自然な学習が可能となる．**図 1** の左側の例では，トマトがアップの画像になったときの発話「そしてトマトです．トマトは横に薄切り.」から，重要部分「トマト」を抽出し，また，アップ画像中から最も注目されている領域を抽出し，その領域とキーワード「トマト」を対応付けることにより，トマトが「赤色」であるという学習を行うことができる．1 つの画像のみから学習を行うと，解析誤りを含む可能性があるので，このようなペアを大量に集めて総合することにより，安定した学習が可能となる．

上記で述べた，発話中での重要部分と映像中の対応付けの段階では，物体に関する知識がないため，誤って領域とキーワードを対応付けてしまう可能性があるので，映像全体から物体モデルを学習した後，映像中の色情報，発話での単語の重要度などを統合して，物体認識を行う．

なお，本研究では，発話はクローズドキャプションを用いる．将来的には音声認識の高度化とともに実際の発話を扱えるようになると考えている．

物体モデル学習の手順の概要を以下に示す．以下では，物体がアップになっている画像を**アップ画像**，アップ画像で視聴者が注目するであろう領域を**注目領域**，1 台のカメラで連続して撮影された映像区間を**ショット**，発話から抽出した重要な単語を**キーワード**と呼ぶ．

(1) エッジ処理を行うことによりアップ画像を抽出する．
(2) 抽出されたアップ画像において，領域分割を行い，画面の中心に近い，領域の画素が領域の重心に密集しているなどの特徴量を考慮することにより，注目領域を決定し，ショットを単位として，1 つの代表的なアップ画像を抽出する．
(3) 抽出された代表的なアップ画像の周辺の発話から，重要な単語をキーワードとして抽出する．画像に最も近い単語をキーワードとして抽出するといった単純な処理ではなく，隠れマルコフモデルに基づくトピック推定[3] によるゴミデータの除去，発話タイプ解析や談話構造解析[4] に基づく単語の重要度計算により，キーワード抽出を行う．
(4) キーワードごとに注目領域の RGB 頻度分布を求め，その最頻値の RGB の平均を物体モデルとする．

大量の映像から物体モデルを学習した後，対象画像と発話が与えられている条件のもとで，各ショットで作業の中心となっている対象物体を決定する問題を，

本研究では物体認識と呼ぶ（詳細は 4 章を参照のこと）．対象画像の周辺の発話で重要度が高く，かつ，対象画像の代表色と物体モデルの色が近いような語を発話から抽出することにより，物体認識を行う．

本手法は映像処理・言語処理ともに料理映像に特化した処理ではないため，ニュースやスポーツなどといったすべての映像ジャンルには適用できないものの，工作や DIY などといった作業教示映像一般に適用できる．また，処理自体をそのまま適用することはできないが，発話中の重要部分（キーワード）と映像中の重要部分（注目領域）は多くの場合に一致しているという考えに基づいた物体のモデル学習は，ロボットやマルチモーダル対話システムなど，他の実世界情報処理においても適用することができると考えている．

2. アップ画像の抽出と注目領域の決定

本章では，画像列からアップ画像を抽出し，その画像において注目領域を決定する手法を説明する．なお本研究では，映像から毎秒切り出した画像列に対して以下の処理を行う．

2.1 ショットへの分割

ある一瞬においてアップになった画像を抽出するのではなく，一定の範囲で映っている物体のアップ画像を抽出するために，映像をショットに分割し，ショットを映像処理の基本単位とする．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．また，カット点において，ニューラルネットワークを用いて顔の検出を行い[5]，顔が検出されたカットは以下の処理の対象から除外する．

2.2 エッジ処理によるアップ画像の抽出

次に，エッジ処理を行うことにより，アップ画像の抽出を行う[⋆1]．3 × 3 の Sobel の一次微分でエッジ処理を行い，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，画面の端に食材が映ることはほぼないと考え，画面の真ん中を中心とする楕円内（**図 2** に示す楕円[⋆2]）だけを考える．エッジ率が閾値（$Th_{edge} = 0.5$）を下回った画像をアップ画像として抽出する．この処理により，食材のアップではない画像や，上記の顔認識処理では認識できなかった顔が映っている画像を除外することができる．

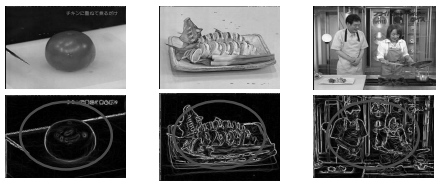


図 2 エッジ処理によるアップ画像の判定
Fig. 2 Close-up image judgement based on the edge detection.

2.3 領域分割

上記の処理で得られたアップ画像において，注目領域を決定するために，以下の手順で領域の分割を行う（**図 3**，**図 4**）．

(1) ショットを基本単位とし，上記のエッジ処理によりアップ画像と判定された画像について，2.2 節で設定した楕円内の画素を RGB3 次元空間に写像し，カラーヒストグラムを求める．ここで，ショット内にある複数のアップ画像について，1 つの RGB3 次元空間に写像することに注意されたい．

(2) 3 × 3 × 3 のメディアンフィルタにより平滑化を行う．

(3) RGB 3 次元空間において，山登り法で画素数の極大点を探索する．ただし，極大値が閾値を下回るものは除外する．図 3 の例では 4 つの極大点が得られている．

(4) 元画像において，各極大点に属する画素のラベリングを行うことにより，領域の分割を行う．図 4 の例では，手の領域，まな板の領域，包丁の領域，にんじんの領域の 4 つの領域が得られている．

2.4 注目領域の決定

注目領域は，なるべく面積の大きい領域，画面の中心に近い領域，領域内の画素が領域の重心に密集している領域などであると考えられる．ここで，「トマト」や「かぼちゃ」のような円に近い食材であれば，領域内の画素が領域の重心に密集しているので，上記の指標を利用することにより，背景の領域と区別できるが，「ねぎ」や「ごぼう」のような細い食材では，領域内の画素が領域の重心に密集していないので，上記の指標では，背景の領域と区別することができない．そこで，上記の指標に加えて，領域が円または長方形にどれくらい近いかという指標も考慮する．

そして，分割された各領域で特徴量を計算し，領域

⋆1 エッジ処理によりアップ画像の抽出が行うには，食材の乗っている器やまな板などが単色であり，そこではエッジが検出されない必要があるが，一般に作業教示映像では，作業の対象となっている物体に注意を引かせるために，物体が乗っている台などは単色である傾向にある．

⋆2 楕円の大きさは予備実験により設定した．

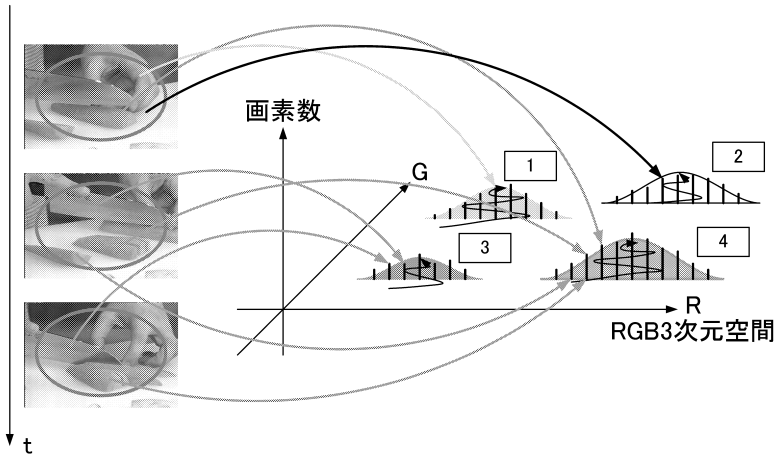


図 3 RGB 空間への写像と山登り法による極大点の探索
Fig. 3 Mapping the pixels to the RGB dimension and searching for the maximum point with the hill-climbing method.

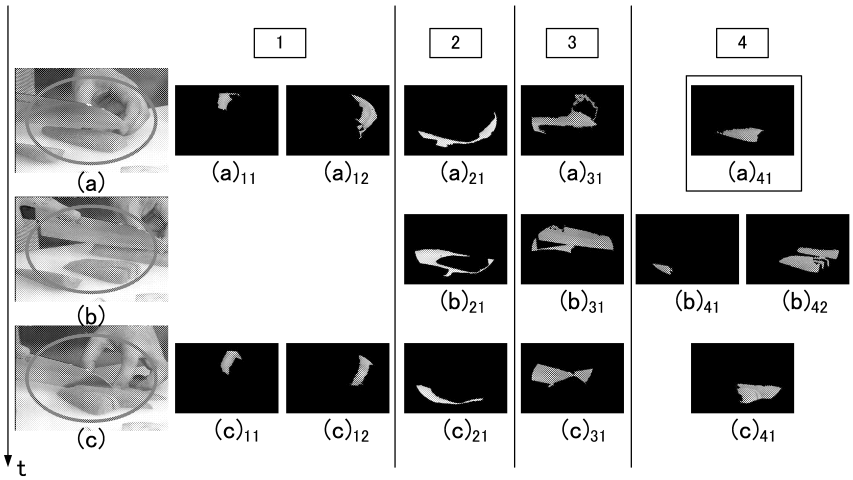


図 4 ラベリングによる領域分割
Fig. 4 Region segmentation with labeling.

の評価値を算出し，ショット単位で最も高い評価値を得た領域を注目領域とする．ここで，誤って手の領域や包丁の領域を注目領域として抽出しないように注意する必要がある．

2.4.1 特徴量の計算

注目領域を決定するために，各領域において**表 1** にあげる特徴量を計算する．

ただし，以下の条件を満たす場合，その領域を注目領域として抽出しない．

- オプティカルフロー
  動きの速い領域は，手や包丁であることが多い．そこで，領域内のオプティカルフローを計算し，ショット内において，領域のオプティカルフロー値の平均が閾値以上の場合，その領域を注目領域としない．本研究ではブロックマッチング法でオプティカルフローの検出を行った．オプティカルフローを検出した例を**図 5** に示す．図 4 の例では，領域 3 においてオプティカルフローの平均値が閾値以上であるので，領域 3 が除外される．
- 上半分率
  料理番組において，手元のショットの場合，斜め上方向から撮影することが多いため，画面の上半分だけに食材の領域がくる可能性は低いと考え[⋆1]，領域の画素のうち，画面の半分より上にある割合が閾値（$Th_{upperratio} = 0.95$）以上の場合，この

⋆1 上半分率が閾値以上である領域をランダムに 30 個調べたところ，そのうち食材の領域であったものは 2 個であった．

表 1 領域の特徴量
Table 1 Features for each region.

S:	面積
G_{dis}:	各画素と領域の重心の距離の平均
C_{dis}:	各画素と画面の中心点の距離の平均
$Circularity$:	どのくらい円に近いかの指標（円形度）であり，周囲長 l，面積 S を用いて以下の式で与えられる．$\frac{4\pi S}{l^2}$ (1) この値は，0 から 1 までの値をとり，円に近いほど 1 に近い値をとる．
$Rectangularity$:	どのくらい長方形に近いかの指標（長方形度）であり，面積 S，外接長方形の面積 S_{rec}（=（短軸）×（長軸））を用いて以下の式で与えられる．$\frac{S}{S_{rec}}$ (2) この値は，0 から 1 までの値をとり，長方形に近いほど 1 に近い値をとる．

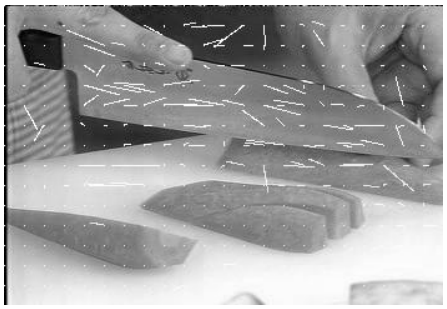


図 5 オプティカルフローの検出
Fig. 5 Opticalflow detection.

領域を除外する．図 4 の例では，$(a)_{11}$，$(c)_{11}$ の領域が除外される．

- 手の領域
  手の領域を注目領域として抽出しないように，松橋ら[6] の手法で，領域の代表色の RGB を修正 HSV に変換し，肌色であるかどうかの判定を行う．予備実験により，(H, S) が，$20 \leq H \leq 35$, $40 \leq S \leq 65$ の範囲内にあるときに肌色領域と見なした．ただし，かつおやごぼうなどのような肌色に近い色の食材の領域も除去してしまう可能性があるので，手は画面の上半分にくることが多いことを考慮し[⋆1]，肌色かつ上半分率が閾値（$Th_{upperratiohand} = 0.8$）以上の場合，この領域を除外する．図 4 の例では，$(a)_{11}$，$(a)_{12}$，$(c)_{11}$ の領域が除外される．
- 接楕円率
  領域の周囲長のうち，閾値（$Th_{boundingratio} = 0.35$）以上の割合が 2.2 節で設定した楕円に接している場合，この領域が食材である可能性は低いと考え，除外する[⋆2]．図 4 の例では，$(c)_{21}$ の領域が除外される．

2.4.2 注目領域の決定

注目領域を決定するために画像 i での領域 k の評価値 $score(R_{k,i})$ を下式で計算する．

$$score(R_{k,i}) = k_S \cdot S + k_{CR} \cdot max(Circularity, Rectangularity) - k_C \cdot C_{dis} - k_G \cdot G_{dis} \quad (3)$$

この評価値は，なるべく面積の大きいもの，円または長方形に近いもの，画面の中心から近いもの，各画素が領域の重心から近いものを抽出するために設定されている．予備実験に基づき，各係数は，$k_S = 0.1$, $k_{CR} = 0.3$, $k_C = 0.3$, $k_G = 0.5$ とした．

1 つの画像から同じ極大点に属する領域が複数得られた場合，最も評価値の高い領域をその画像の代表領域とする．図 4 の画像 (b) において，極大点 4 に属する，$(b)_{41}$ と $(b)_{42}$ の 2 つの領域があるが，評価値の高い $(b)_{42}$ を代表領域とする．

そして，ショット内で領域 k の評価値の総和 $score(R_k)$ をとり，

$$score(R_k) = \sum_{i \in Shot} score(R_k, i) \quad (4)$$

ショット内で最も高い評価値を得た領域を注目領域とする[⋆3]．図 4 では，極大点 4 の領域が注目領域として抽出される．そして，ショット内で最もエッジ率の小さいアップ画像をショットでの代表画像とし，その画像における注目領域を物体モデルの学習データとして採用する．図 4 では最もエッジ率の小さい画像 (a) の $(a)_{41}$ が注目領域として選ばれる．

⋆1 手が映っている画像をランダムに 30 枚収集したところ，手の領域の上半分率の平均は 0.91 であった．
⋆2 接楕円率が閾値以上である領域をランダムに 30 個調べたところ，そのうち食材の領域であったものは 3 個であった．
⋆3 ここで，ショット内の異なる画像に現れる領域が同一の物体の領域であるかどうかの対応付けは，同じ極大点に属するかどうかで判断できることに注意されたい．

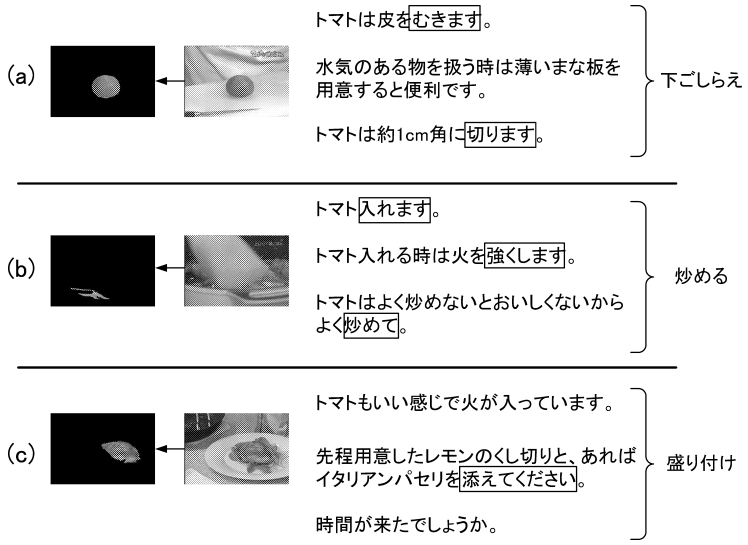


図 6 作業に関する発話と背景画像からのトピック推定例
Fig. 6 Examples of topic identification from utternaces referreing to actions and background image.

3. キーワードの抽出

自然言語処理では，近年の計算機環境の進歩，大規模コーパスの利用により，形態素解析・構文解析といった基本的な解析が高い精度で行えるようになってきた[7]．さらに，述語項構造解析や省略・照応解析などの文脈処理も少しずつ行えるようになってきた．

本章では，自然言語処理を用いて，前章で抽出したショットの代表画像の周辺の発話から重要な単語を抽出し，キーワードとして代表画像と対応付ける手法を説明する．本研究では，発話としてクローズドキャプション*1を用いた．

まず，ショットの代表画像の周辺の発話が，「下ごしらえ」,「炒める」,「盛り付け」などのどのトピックにあたるかの推定を行い，トピックが「下ごしらえ」以外の部分のショットの代表画像を捨てることを行う．この処理により，食材が変形・変色したものを除去することができ，良質な学習データを得ることができる．次に，ショットの代表画像の周辺の発話から重要な単語を選ぶ処理を行う．そのために，まず談話構造解析を行い，話題のまとまりに分割し，ショットの代表画像に近いまとまりの範囲内で，発話のタイプ，談話構造などに基づき，最も重要な単語をキーワードとする．

*1 クローズドキャプションとは，もともと聴覚障害者用に開発された，書き起こしテキストのことである．映画の字幕やニュースのテロップのような画面につねに表示されるオープンキャプションとは異なり，専用の機器で表示することができる．

3.1 隠れマルコフモデルによるトピック推定

前章の画像処理により，**図 6** のようなショットの代表画像が得られる．料理映像の場合，トピックが「炒める」や「盛り付け」の場合，食材が変色・変形し，原型をとどめていないことが多く，物体モデルの学習に適しているとはいえない．そこで，トピックが「下ごしらえ」の部分のみからショットの代表画像を収集し，その他のトピックからは代表画像の収集を行わない．

柴田らは，隠れマルコフモデルに基づき，「切ります」「火を強くします」などといった作業に関する発話と，そのときの背景画像を利用し，その発話が，「下ごしらえ」,「炒める」,「盛り付け」など，8 つのトピックのうちどれにあたるかの推定を行っており[3]，その結果，トピックが「下ごしらえ」と推定された部分から学習データを収集する．図 6 の例では，(a) の「下ごしらえ」の領域から学習データを収集し，(b) の「炒める」や (c) の「盛り付け」から得られた学習データは捨ててしまう．

3.2 談話構造解析

基本的には，ショットの代表画像が得られた時刻に近い発話において，その画像に映っている食材への言及があるが，たとえば，以下のように作業対象の食材以外の食材への言及がある場合がある．

(1) ねぎ でも結構です．

(2) 玉ねぎ と同じくらいの大きさです．

表 2 発話タイプの分類
Table 2 Utterance-type classification.

【作業：大】 ・ステーキの材料 にかかります. **【作業：中】** ・強火で油を温め ましょう. ・じゃあ炒め ていきましょう. **【作業：小】** ・お鍋にお水を入れます. **【料理状態】** ・ニンジンの水分がなくなりました.	**【留意事項】** ・肉をパラパラに炒めること がポイントです. **【代替可】** ・青ねぎ でも結構です. **【食品・道具提示】** ・材料は，牛ひき肉，百五十グラム です. **【雑談】** ・ブランデーのパフォーマンスも お楽しみください.

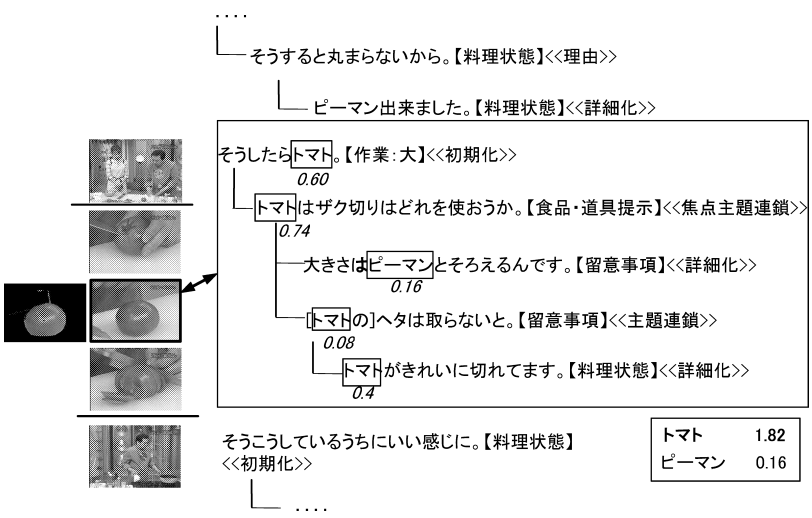


図 7 談話構造解析に基づくキーワードの抽出
Fig. 7 Keyword extraction based on discourse structure analysis.

そこで，一定の範囲内で重要な語をキーワードとして抽出する．一定の範囲を決定するために，クローズドキャプションの談話構造解析を行い[4]，話題のまとまりを認識する．すなわち，任意の 2 文間の関係を明らかにし，密につながる部分（理由，詳細化，主題連鎖などの談話関係を持つ部分）とそうではない部分（初期化）を区別し，そうでない部分を話題の分割点とする．

談話構造解析の概要を以下に示す．

(1) 入力文を形態素解析器 JUMAN[8] で形態素解析した後，構文解析器 KNP[9] で構文・格解析する．

(2) 自動構築した用言[10]・名詞[11] の格フレームを用いて省略の解析を行う．

(3) 節末の表層パターンを用いて各発話のタイプを認識する．発話のタイプは**表 2** にあげた【作業：大】，【料理状態】，【留意事項】など 9 種類を考え，【作業：小】，【料理状態】以外の発話タイプは表中に示したようなパターンで認識を行い，自動詞を【料理状態】，それ以外を【作業：小】とする．

(4) 発話タイプ・省略解析結果を含む語の連鎖の検出・接続詞を中心とした表層ルールを統合することにより，談話構造を解析し，文間の関係を明らかにする．談話構造のモデルとしては，文を基本単位とし，関係する文どうしがある結束関係でリンクされたグラフ構造を考える[12]．

談話構造解析の結果，**図 7** 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（<<>>）は親の文との結束関係を示すものである．

そして，代表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，その談話構造木内で，発話のタイプの解析や談話構造解析結果を考慮し，最も重要な単語をキーワードとして選ぶ．この処理を次節で述べる．

表 3 語の重要度計算の関数
Table 3 Functions for calculating the word importance.

$f_{utype}(W_i)$:	発話タイプが【作業】,【食品・道具提示】,【料理状態】なら 1,【代替可】なら 0.1，それ以外なら 0.3 を返す関数
$depth(W_i)$:	談話構造木の深さを返す関数．多くの研究者が指摘しているように，一般に談話構造の根に近い方が文の重要度が高いと考えられる[14),15)]．
$f_{clause}(W_i)$:	W_i が主節にあれば 1，従属節にあれば 0.5 を返す関数
$f_{topic}(W_i)$:	以下のように，W_i の文節に提題助詞を含む場合 1.5，それ以外なら 1 を返す関数 (3) トマトは 皮をむきます.
$f_{anaphora}(W_i)$:	省略解析結果なら 0.5，それ以外なら 1 を返す関数
$f_{time}(F_{W_i}, F_{image})$:	W_i を含む発話のフレーム数 F_{W_i} と画像のフレーム数 F_{image} を用いて， $f_{time}(F_{W_i}, F_{image}) = \begin{cases} 1 - \frac{\lvert F_{W_i} - F_{image} \rvert}{F_{th}} & (\lvert F_{W_i} - F_{image} \rvert \leq F_{th}) \\ 0 & (\text{otherwise}) \end{cases}$ (5) で与えられ，画像の時刻と発話の時刻が離れるほど小さい値をとる．ここで，$F_{th} = 1{,}000$ とした.

3.3 キーワードの抽出

シソーラス[13)] で食材タグのふられた名詞 W_i に対して，発話タイプ，談話構造解析結果，画像と発話の時間のずれなどに基づき，以下の式でスコア付けを行い，談話構造木で最もスコアの高い名詞をキーワードとして選ぶ.

$$score(W_i) = \sum_{W_i \in Tree} f_{utype}(W_i) \cdot \frac{1}{\sqrt{depth(W_i)}} \cdot f_{clause}(W_i) \cdot f_{topic}(W_i) \cdot f_{anaphora}(W_i) \cdot f_{time}(F_{W_i}, F_{image}) \quad (6)$$

ここで，各関数を**表 3** に示す．各関数が返す値は，予備実験を基に設定した．また，予備実験の結果，画像と 1,000 フレーム（= 33.3 sec.）以上離れた発話からキーワードを抽出しても副作用が生じることが多いことから，$F_{th} = 1{,}000$ とした.

図 7 の例では，談話構造木内でスコアを計数すると，トマトが 1.82 点，ピーマンが 0.16 点となり，最も高いスコアを得たトマトがキーワードとして抽出される.

なお，表記揺れに対応するために，キーワードとして形態素解析器 JUMAN が出力する代表表記を利用する．これにより，たとえば，表記が「じゃがいも」,「じゃが芋」「ジャガイモ」のものを代表表記「じゃが芋」に,「玉ねぎ」,「たまねぎ」,「タマネギ」を代表表記「玉ねぎ」にマージすることができる.

3.4 物体モデルの構築

以上で説明した処理を行うことにより，ショットの代表画像の注目領域とキーワードのペアを得ることができる．実際に得られた注目領域とキーワードのペアの例を**図 8** 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.

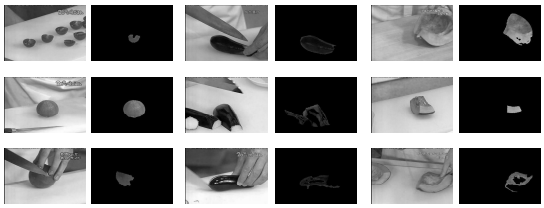


図 8 収集された注目領域とキーワードのペアの例と構築された物体モデル
Fig. 8 Examples of collected sets of an attention region and a keyword and constructed object models.

食材ごとに，注目領域の RGB 頻度分布を求め，その最頻値の RGB の平均を物体モデルとする.

4. 物体認識

構築した物体モデルを利用して，物体の認識を行う．ここで，本研究における物体認識を，「画像列とクローズドキャプションが与えられている条件の下，ショットを単位として，そのショットで作業の対象の中心となっている食材を 1 つ認識する」と定義する．たとえば，鍋に玉ネギが入った状態で，にんじんを入れるという作業を行っているショットでは，作業の対象であるにんじんのみを正解として付与しておき，にんじんが認識できれば正解とする．また，同時に複数の食材が作業の対象となっている場合は，正解として複数の食材を付与しておき，そのうち 1 つを認識できれば正解とする.

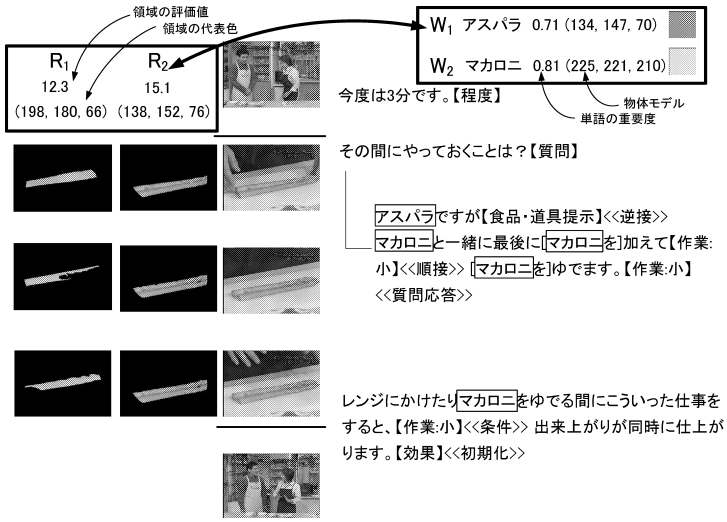


図 9 物体認識
Fig. 9 An overview of object recognition.

物体認識の手順を以下に示す（**図 9**）.

(1) 対象画像を含むショットを単位とし，物体モデルの構築と同じ手順で領域分割を行う．図 9 では，まな板の領域 R_1 とアスパラの領域 R_2 が得られている．

(2) 対象画像を含むショットと，ショットに一番近い談話構造木を対応付ける．物体モデル構築時はショットに一番近い談話構造木に含まれる食材からキーワードを抽出したが，物体認識時は，一番近い談話構造木とその前後の談話構造木に含まれる食材すべてを候補とする．図 9 では，ショットに一番近い談話構造木は，発話「その間にやっておくことは？」と「アスパラですが ···」を含む談話構造木であるが，その談話構造木と，その前の談話構造木（発話「今度は 3 分です.」を含む談話構造木）とその後ろの談話構造木（発話「レンジにかけたりマカロニ ···」を含む談話構造木）に含まれる食材が物体認識の候補となる．

(3) 領域分割された領域 R_k，談話構造木内の各名詞 W_i に対して，式 (4) で与えられる領域の評価値 $score(R_k)$，式 (6) で与えられる名詞 W_i の重要度 $score(W_i)$ を物体モデルの構築時と同じように計算する．そして，領域の代表色と名詞 W_i の物体モデルのユークリッド距離 $distance(R_k, model(W_i))$ を計算し，以下のスコアが最も高くなる領域 R_k と名詞 W_i のペアを見つけ，その値が閾値以上の場合，名詞 W_i を物体認識結果として採用する．このスコアは，領域の評価値がなるべく高く，名詞の重要度がなるべく高く，また，領域の代表色と名詞の物体モデルの色がなるべく近いような領域と名詞のペアを得るために設定した．

$$\underset{R,W}{\operatorname{argmax}} \frac{score(R_k) \cdot score(W_i)}{distance(R_k, model(W_i))} \quad (7)$$

図 9 では，まず，領域 R_1 の評価値が 12.3，代表色が (198, 180, 66)，領域 R_2 のスコアが 15.1，代表色が (138, 152, 76) と計算され，アスパラ（W_1）の重要度が 0.71，マカロニ（W_2）の重要度が 0.81 と計算される．そして，アスパラ（W_1）と領域 R_2 のペアが最も高いスコアを得たので，アスパラが物体認識結果として採用される．

5. 実　　験

NHK の「きょうの料理」，NTV の「キューピー 3 分クッキング」の映像を用いて，物体モデルの学習と物体認識の実験を行った．

5.1 物体モデルの学習

NHK の「きょうの料理」の映像 205 日分（計 85.4 時間），NTV の「キューピー 3 分クッキング」の映像 64 日分（計 10.7 時間）から物体モデルの学習を行った．実験に用いた映像の情報を**表 4** に示す．

物体モデルの学習結果を**表 5** に示す．ここで，各食材の物体モデルが正しいかどうかは，物体モデルとして得られた RGB の値が各食材の色として妥当なものであるかどうかで判定した．表 5 から分かるように，約 100 食材の物体モデルを自動構築することができ，その精度は 77.8%であった．

表 4 実験に用いた映像の情報
Table 4 Information about the videos used for our experiments.

	きょうの料理	キューピー 3 分クッキング
映像数	205	64
総時間	85.4 時間（307,500 秒）	10.7 時間（38,400 秒）
ショット数	28,215	5,698
アップ画像数	95,343	15,098
収集された注目領域とキーワードのペアの数	627	43

表 5 物体モデルの学習の実験結果
Table 5 Experimental result of object-model acquisition.

食材数	正解数	精度（%）
108	84	77.8

表 6 収集された注目領域とキーワードのペアのサンプル数別の精度
Table 6 Object-model acquisition accuracy by the collected set of attention region and keyword.

収集されたサンプル数	食材数	正解数	精度（%）
2 サンプル以上	83	68	81.9
3 サンプル以上	64	56	87.5
5 サンプル以上	37	35	94.6

図 10 注目領域とキーワード「キウイ」のペア
Fig. 10 Collected sets of an attention region and a keyword “kiwi fruit”.

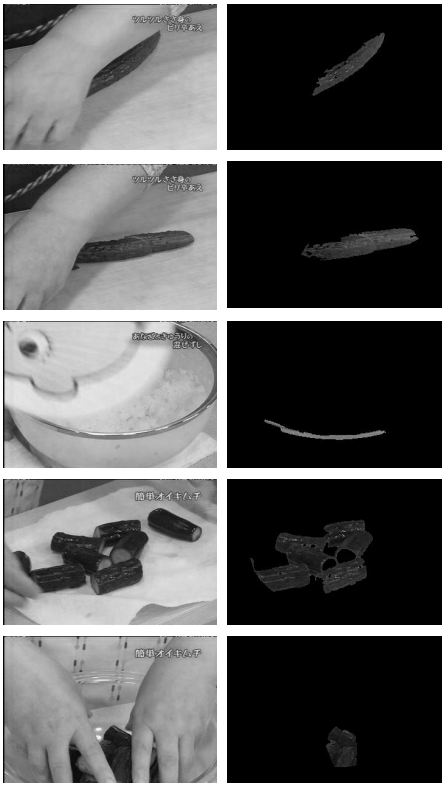


図 11 注目領域とキーワード「キュウリ」のペア
Fig. 11 Collected sets of an attention region and a keyword “cucumber”.

また，物体モデル構築の精度を，各キーワードにおいて収集された注目領域とキーワードのペアのサンプル数別に算出した精度を**表 6** に示す．収集されたサンプル数が多いほど，安定して学習が行えていることが分かる．たとえば，キーワード「キウイ」では注目領域とキーワードのペアが 2 つ収集された（**図 10**）．そのうちの 1 つは正解であったが，もう 1 つは誤って手の領域が注目領域として抽出された．その結果，キウイの物体モデルは誤って構築された．一方，キーワード「キュウリ」では注目領域とキーワードのペアが 5 つ収集された（**図 11**）．そのうちの 1 つは「キュウリ」が映っていない画像であったため失敗であったが，残り 4 つは正解であったため，結果として正しく物体モデルが構築された．以上から，今後映像データが増加すればより安定した学習が行えることが期待できる．

次に，3.1 節で行った隠れマルコフモデルによるトピック推定が物体モデル構築で有効であったかどうかを示すための実験を行った．**表 7** に隠れマルコフモデルによるトピック推定を行った場合と行わなかった場合それぞれの物体モデル構築の精度を示す．トピック推定を行い，トピックが「下ごしらえ」の部分からのみ物体モデルの学習データを得ることにより，精度が 13 ポイント上昇しており，このタスクでトピック推定が有効に働いていることが分かる．「下ごしらえ」，「煮る」，「盛り付け」などのトピックの設定は料理ド

表 7 隠れマルコフモデルに基づくトピック推定による精度向上
Table 7 Accuracy improvement by the HMM-based topic identification.

トピック推定	精度（%）
なし	64.8（70/108）
あり	**77.8（84/108）**

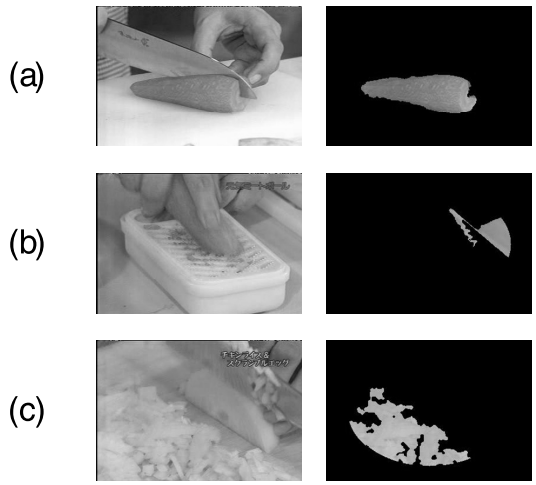


図 12 収集された注目領域とキーワード「にんじん」の例（(a) 成功，(b) 注目領域抽出誤り，(c) キーワードの物体が映っていない）
Fig. 12 Examples of collected sets of an attention region and a keyword “carrot”((a) success, (b) attention region extraction failure, (c) the food does not appear in the image).

メイン固有のものであるが，他のドメインでもいくつかのトピックを設定し，物体が変形・変色する前であるようなトピックからのみ物体モデルの学習データを得ることにより，物体モデル構築の精度を向上させることができると考えられる．

以下に物体モデル構築の誤り原因を示す．収集された注目領域とキーワードのペアは，**図 12** に示すように，(a) 成功，(b) 注目領域抽出誤り，(c) キーワードの物体が映っていない，の 3 種類に分類することができ，(b) 注目領域抽出誤り，(c) キーワードの物体が映っていない例を以下にあげる．

注目領域抽出誤り

- まな板領域と混じってしまう
  白菜などのような白っぽい食材の場合，平滑化を行う際に，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ある（**図 13**）．
- 食材の領域を動領域として除去してしまう
  2.4.1 項で述べたオプティカルフローを計算することにより，手や包丁といった動領域を除去することに成功したが，誤って食材の領域を動領域として除去してしまうことがある（**図 14**）．

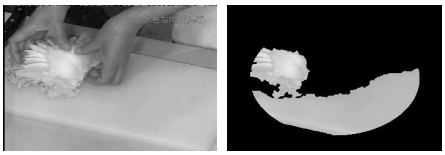


図 13 注目領域抽出誤り例（白菜）
Fig. 13 An example of failure of attention region extraction (“Chinese cabbage”).

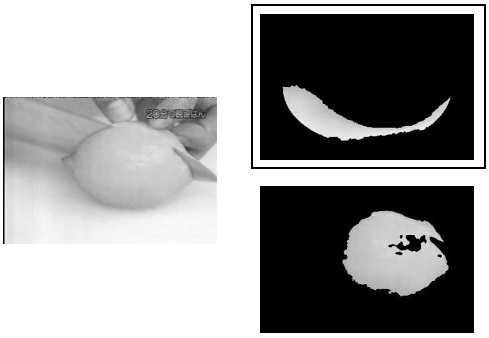


図 14 注目領域抽出誤り例（レモン）
Fig. 14 An example of failure of attention region extraction (“lemon”).

図 15 キーワードの物体が映っていない例（きのこ）
Fig. 15 An example in which the food does not appear in the image (“mushroom”).

キーワードの物体が映っていない

- トピック推定誤り
  隠れマルコフモデルでトピックを推定した結果，下ごしらえ以外のトピックを誤って下ごしらえと推定した場合，キーワードの物体が映っていないことがある（**図 15**）．この問題に対しては，隠れマルコフモデルによるトピック推定の精度を向上させることにより対処する予定である．
- 顔画像認識誤り
  ニューラルネットでの顔認識やエッジ処理による顔画像除去に失敗したものがある．

5.2 物体認識

構築した物体モデルを用いて，「きょうの料理」「キューピー 3 分クッキング」それぞれ 5 番組に対

表 8 物体認識の実験結果
Table 8 Experimental result of object recognition.

	適合率	再現率	F 値
提案手法	100/132 (75.8%)	100/143 (69.9%)	0.727
ベースライン			
色情報のみ	75/132 (56.8%)	75/143 (52.4%)	0.545
最も多く言及された食材	90/138 (64.4%)	90/143 (62.9%)	0.641
言語解析のみ	97/135 (71.9%)	97/143 (67.8%)	0.698

して物体認識の実験を行った[⋆1]．正解は 4 章で述べた物体認識の定義に従ってショットに対して付与し，再現率，適合率，F 値で評価した．なお，評価を行ったのは食材が加熱される前のみとした．

物体認識結果と以下のベースラインの精度を**表 8** に示す．

色情報のみ 対象画像が属するショットに一番近い談話構造木内に出現する食材の中で，その物体モデルの色と，対象画像が属するショットで評価値が最も高い領域の代表色とのユークリッド距離が最も小さい食材を物体認識結果とする．

最も多く言及された食材 色情報は利用せず，画像に一番近い談話構造木で最も多く言及された食材を物体認識結果とする．

言語解析のみ 色情報は利用せず，言語解析で最も重要と判断されたキーワードを物体認識結果とする．

物体認識の精度は F 値で 0.727 であり，色情報のみ，単純に単語の頻度のみ，言語解析のみを利用したベースラインを上回っていることが分かる．このうち，言語解析のみと提案手法の差分が，映像全体から得られた物体モデルを利用した効果である．物体認識のアルゴリズムを説明した図 9 で再度説明すると，物体モデル学習時にはアスパラの領域とキーワード「マカロニ」が対応付けられる（マカロニの方がアスパラよりも重要と判断されているため）が，映像全体から学習されたアスパラとマカロニの物体モデルを利用することによって，正しくアスパラの領域とキーワード「アスパラ」が対応付けられることとなる．

物体認識の誤りの多くは物体モデル構築誤りに起因するものであり，「あさつき」「鰻」などの物体モデルの構築が失敗したため[⋆2]，それらを認識することができなかった．さきほど示したように，より大量の映像を利用できればより物体モデルの学習が安定するので，それにともない物体認識の精度も向上することが期待できる．

また，本研究では，食材の物体モデルを 1 つの RGB 値で表現するといった比較的単純な形態を採用しているが，言語情報も利用することにより，色情報のみでは区別できないような場合でも正しく物体認識を行える．たとえば，対象画像に緑色の物体があり，そのときの発話が「今日はほうれん草を使います．みずなでも結構です.」であった場合を考える．ほうれん草，みずなともにその物体モデルが緑色であり，単に色情報のみを利用すると，どちらであるかの区別がつかないが，この場合，3 章で述べた言語処理を行うと，発話「みずなでも結構です.」の発話タイプが【代替可】であることから，ほうれん草の方が重要度が高いことが分かり，結果としてほうれん草と正しく物体認識できる．

一方で，本研究では各食材でモデルを 1 つとしていることにより，たとえば，「ねぎ」では「白ねぎ」の色が学習されており，このモデルでは「青ねぎ」の認識を行うことができない．この問題には，物体モデルを学習する際の誤りを少なくしたうえで，色情報の分布をクラスタリングすることにより，複数の色モデルを構築することで対処することができると考えられる．

同じように，本研究では表皮と果肉の色が異なる場合，どちらか一方の色が学習されれば正解とした．たとえば，「なす」の場合，皮の色が学習されているため，「なす」が切られて果肉の白色の部分が映った画像で「なす」と認識を行うことができなかった．今後，表皮と果肉や，加熱後の色情報，複数の色からなる食材の色情報などの学習も行う予定である．

6. 関 連 研 究

大量の映像から，人手で正解データを用意することなく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っている[1)]．画像を領域分割し，領域と単語の対応

⋆1 この映像は，物体モデルの学習にも利用している．

⋆2「あさつき」の場合，注目領域とキーワードのペアが 2 つ得られたが，そのうちの 1 つが注目領域抽出誤りであったため，誤ったモデルが構築された．「鰻」の場合は，注目領域とキーワードのペアが得られなかったため，物体モデルの構築が行うことができなかった．この原因の 1 つはトピック推定誤りであり，実際はトピックが「下ごしらえ」である部分が「盛り付け」と推定されたため，実際には鰻が映っているショットから注目領域とキーワードのペアが得られなかった．

付けを EM アルゴリズムを用いて学習している．また，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[16]．本研究では，物体認識時に発話を利用するため，物体認識候補がしぼられるが，これらの研究では，物体認識時に人手によって付与されたキーワードを利用しないので，すべての画像に付与された全キーワードが物体認識候補となるという点で，問題設定が異なる．

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[2]の研究がある．まず，ニンジン，ピーマン，赤唐辛子，サヤインゲンなど 11 個の食材を対象とし，各素材の映っている料理画像中から素材領域を人手で切り出し，領域中の RGB の統計をとることにより，素材の色分布情報を得ている．そして，物体認識対象画像において，色分布情報とのマハラビス距離が閾値以下の面積を抽出し，その面積に基づき素材の確信度を計算し，確信度の最も高い食材を物体認識結果としている．その際に，番組に付随するレシピテキストの材料に表れる食材だけに対象を限定することにより精度向上を達成している．素材の加熱前に対して実験を行ったところ，再現率 74.8%，適合率 78.4%であった．我々の物体モデルの自動学習手法は，この研究における人手での素材領域の切り出しを自動で行っていることにあたる．我々の物体認識の精度は，高野らの研究よりも少し劣るものの，約 100 食材を対象にできていることを考えると，この研究と同等以上の成果をあげられていると考えている．また，高野らはレシピテキストだけでなく，クローズドキャプションを利用し，クローズドキャプションに出現する時刻付近で大きくなるような確信度を統合する手法を試しているが，確信度を補強するにとどまり，物体認識の精度向上には至っていない．これは単純に画像に近い単語を重要としているためであると考えられ，本研究で行ったような言語処理によるキーワードの重要度決定手法が有効であると考えられる．

7. 結　　論

計算機・ネットワーク環境の進歩により，ユビキタスセンサ情報社会，ロボット社会が到来しようとしており，実世界での情報把握・理解の実現がますます重要になってきている．本論文では，大量の映像を実世界情報アーカイブの例として，発話と視覚情報を統合的に扱う枠組みを提案した．具体的には，作業教示映像である料理映像を対象とし，言語情報と映像情報を統合することにより，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を用いて，物体の認識を行う手法を提案した．本手法では，まず，物体がアップになっている画像を抽出し，その画像における注目領域を決定する．次に，画像の周辺の発話から重要な単語をキーワードとして抽出し，注目領域と対応付ける．このような注目領域とキーワードを大量に収集することにより，物体モデルの構築を行った．2 つの料理番組，計 96 時間分の映像から物体モデルを構築したところ，約 100 食材の物体モデルが構築でき，その精度は 77.8%であった．そして，構築された物体モデルの色情報と談話構造に基づく単語の重要度を考慮することにより，物体認識を行った．2 つの料理番組で実験を行ったところ，精度は F 値で 0.727 であった．

今後の課題としては，まず，表皮と果肉の色や，1 つの食材に対して複数の色情報の学習，形状の学習などを行うことがあげられる．そして，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し，映像情報を利用することにより，言語解析の精度を向上させる予定である．

参 考 文 献

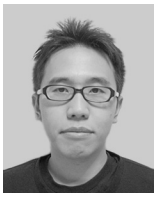
1) Duygulu, P., Barnard, K., de Freitas, N. and Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *European Conference on Computer Vision* (*ECCV*), pp.97–112 (2002).
2) 高野　求，三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦：テキストからの制約に基づく料理画像中の物体検出，情報処理学会第 65 回全国大会，Vol.2, pp.255–256 (2003).
3) 柴田知秀，黒橋禎夫：言語情報と映像情報を統合した隠れマルコフモデルに基づくトピック推定，情報処理学会論文誌，Vol.48, No.6, pp.2129–2139 (2007).
4) Shibata, T., Tachiki, M., Kawahara, D., Okamoto, M., Kurohashi, S. and Nishida, T.: Structural Analysis of Instruction Utterances using Linguistic and Visual Information, *Proc. 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems* (*KES2004*), pp.393–400 (2004).
5) Rowley, H.A., Baluja, S. and Kanade, T.: Neural Network-Based Face Detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.20, No.1, pp.23–38 (1998).
6) 松橋　聡，藤本研司，中村　納，南　敏：顔領域抽出に有効な修正 HSV 表色系の提案，テレビジョン学会誌，Vol.49, No.6, pp.787–797 (1995).
7) 黒橋禎夫：言語コンピューティング，人工知能学

会誌，Vol.22, No.5, pp.711–718 (2007).
8) Kurohashi, S., Nakamura, T., Matsumoto, Y. and Nagao, M.: Improvements of Japanese Morphological Analyzer JUMAN, *Proc. International Workshop on Sharable Natural Language*, pp.22–28 (1994).
9) Kurohashi, S. and Nagao, M.: A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures, *Computational Linguistics*, Vol.20, No.4 (1994).
10) Kawahara, D. and Kurohashi, S.: Zero Pronoun Resolution based on Automatically Constructed Case Frames and Structural Preference of Antecedents, *Proc. 1st International Joint Conference on Natural Language Processing*, pp.334–341 (2004).
11) Sasano, R., Kawahara, D. and Kurohashi, S.: Automatic Construction of Nominal Case Frames and its Application to Indirect Anaphora Resolution, *Proc. 20th International Conference on Computational Linguistics*, No.1201–1207 (2004).
12) Kurohashi, S. and Nagao, M.: Automatic Detection of Discourse Structure by Checking Surface Information in Sentences, *Proc. 15th COLING*, Vol.2, pp.1123–1127 (1994).
13) 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).
14) Ono, K., Sumita, K. and Miike, S.: Abstract generation based on rhetorical structure extraction, *Proc. 15th COLING*, pp.344–348 (1994).
15) Marcu, D.: Discourse trees are good indicators of importance in text, *Advances in Automatic Text Summarization*, Mani, I. and Maybury, M. (Eds.), pp.123–136, The MIT Press (1999).
16) Feng, H. and Chua, T.-S.: A Bootstrapping Approach to Annotating Large Image Collection, *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp.55–62 (2003).

(平成 18 年 11 月 2 日受付)
(平成 19 年 12 月 4 日採録)

柴田　知秀

1979 年生．2002 年東京大学工学部電子情報工学科卒業．2007 年東京大学大学院情報理工学系研究科博士課程修了．博士（情報理工学）．現在，京都大学大学院情報学研究科特任助教．自然言語処理の研究に従事．



加藤　紀雄

1981 年生．2005 年東京大学工学部電子情報工学科卒業．2007 年東京大学大学院情報理工学系研究科修士課程修了．現在，農林中央金庫勤務．

黒橋　禎夫（正会員）

1966 年生．1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了．博士（工学）．2006 年 4 月より京都大学大学院情報学研究科教授．自然言語処理，知識情報処理の研究に従事．