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

柴田知秀† 黒橋禎夫†

本稿では，教示映像，特に料理映像を対象とし，言語情報と映像情報を用いてトピックを推定する手法について述べる．トピック推定は，隠れ状態がトピックにあたり，言語，画像，音声情報の様々な素性が観測される隠れマルコフモデルで行う．素性は，発話を汎化した格フレームや背景画像，手がかり表現や語連鎖，無音といった談話素性からなる．2 つの料理番組で実験を行ったところ，提案手法の有効性を示すことができた．

Unsupervised Topic Identification by Integrating Linguistic and Visual Information Based on Hidden Markov Models

Tomohide Shibata† and Sadao Kurohashi†

This paper presents an unsupervised topic identification method integrating linguistic and visual information based on Hidden Markov Models (HMMs). We focus on instruction videos, especially cooking videos. We employ HMMs for topic identification, wherein a state corresponds to a topic and various features including linguistic, visual and audio information are observed. The features consist of case frame, background image and discourse features, such as cue phrase, word chaining and silence. Our experiments on two kinds of cooking programs show the effectiveness of our proposed method.

1. はじめに

近年の計算機・ネットワーク環境の発展により，大量の映像が配信・蓄積されるようになってきた．蓄積された映像を高度に利用するには，映像を意味的なまとまりに分割し，各セグメントにおいて，何に関する映像であるか，何が映っているか，何が行われているかなどといった情報を付与（インデキシング）する必要がある．これは現在のところほとんど人手で行われており，大規模映像を扱うには自動でインデキシングする技術が必要となる．

このような背景のもと，映像解析の分野では，統計的な手法である隠れマルコフモデル（HMM）を用いてショットの解析が行われている[1)~3)]．これらの研究では，まず映像をカメラの切り替わりであるショットに分割し，ショットから色情報やカメラワークといった特徴量を抽出する．そして，HMM でショットの分類を行っている．たとえば，野球映像では，ピッチングショットやランナのアップのショット，観客が映っているショットなどに分類している．これらの研究では高精度を達成するために，あらかじめ正解データを用意しておき，そこからモデルの学習を行っており，他のドメインへの移行が困難である．また，登場人物の発話やナレーション，スポーツ映像における実況などといった映像中の言語情報もショット解析に有用であると考えられるが，多くの研究では利用されていない．

一方，自然言語処理の分野では，近年，Barzilay らが HMM を用いて，生コーパスからトピック遷移モデルを構築する手法を提案している[4)]．このモデルでは，あるトピックではどのような語が出現するか，また，トピックがどのように遷移するかを，生コーパスから HMM で学習している．このモデルではタグ付きコーパスを必要としないため，あるドメインにおける大量のテキストからそのドメインでのトピック遷移モデルを学習することができる．

本研究では，言語情報と映像情報を統合した，HMM によるトピック推定手法を提案する．本研究では再利用価値の高い教示映像を対象とし，スポーツ，料理，DIY などのジャンルのうち，料理映像に焦点をあてる．**図 1** は HMM でトピックの推定を行っている様子を示しており，この例では，順に，「下ごしらえ」，

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

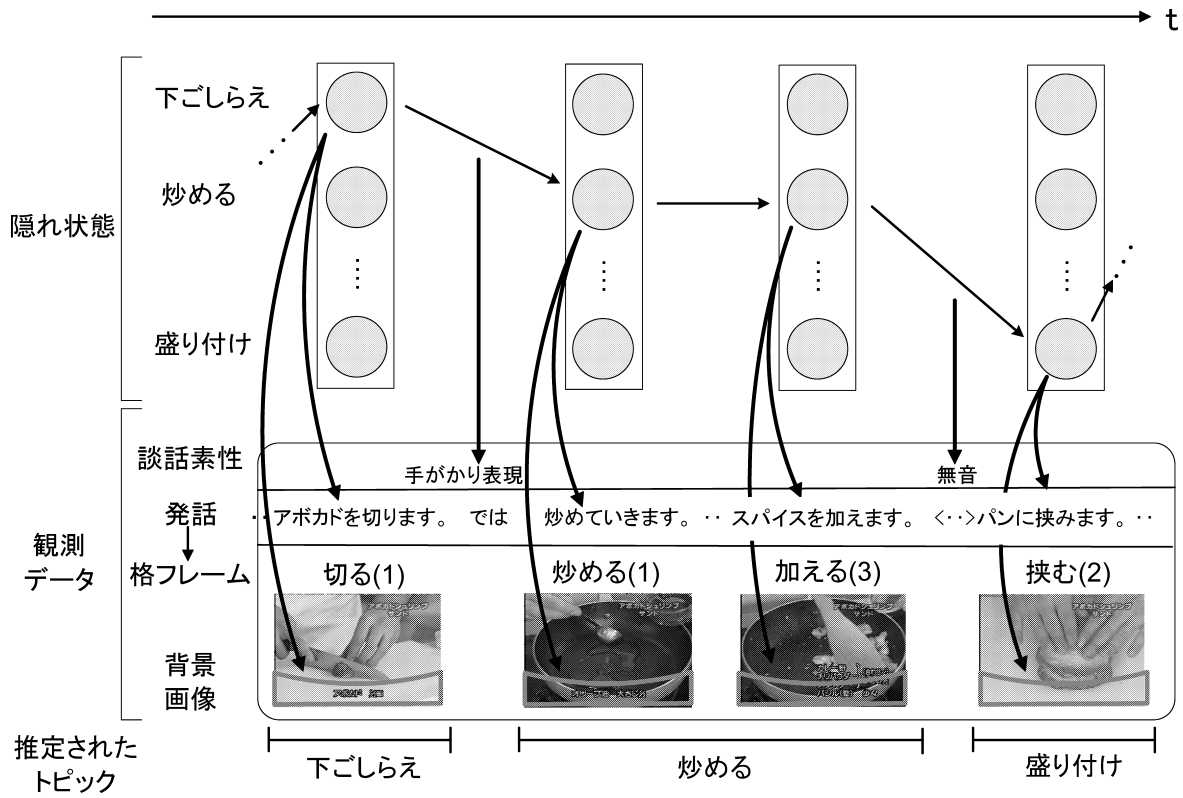


図 1 隠れマルコフモデルによるトピック推定
Fig. 1 Topic identification with Hidden Markov Models.

「炒める」,「盛り付け」とトピックのラベリングを行う．推定されたトピックは，検索のインデキシングや映像の要約生成に利用することができる．トピックをインデキシングに含めることにより,「魚の下ごしらえ」といったクエリで検索することができ，また，トピックが変わったところでセグメンテーションし，各セグメントのラベルにトピックを付与することにより，要約を生成することができる．たとえば,「トマトのスパゲティ」という料理の場合,「下ごしらえ」→「茹でる」→「炒める」→「盛り付け」といった要約を生成することができる．

本研究では，Barziraly らの研究を参考にし，HMM を用いてトピックの推定を行う．この HMM では，隠れ状態が，下ごしらえ，炒める，盛り付けなどといったトピックに相当し，隠れ状態から言語・画像・音声情報の様々な素性が観測される．本研究では隠れ状態であるトピックとして，以下の 8 種類を考える☆．

下ごしらえ，蒸す，ゆでる，揚げる，煮る，炒める，盛り付け，焼く

ここで，これらのトピックは具体的な作業 1 つを指すのではなく，いくつかの作業の集合からなる抽象的な概念として定義する．すなわち,「下ごしらえ」は「野菜を切る」「皮を剥く」「種をとる」などといった作業,「煮る」は「調味料を入れる」「落とし蓋をする」「魚を鍋に並べる」「強火にする」といった作業の集合として考える．

Barzilay らのモデルでは，ドメイン固有の語の分布を生テキストから学習することができるが，低頻度の語についてはどのトピックで出現するかの学習が十分に行えず，データスパースネスの影響を受けてしまう可能性がある．テキストを意味まとまりに分割するテキストセグメンテーション手法や，文または節の関係を明らかにする談話構造解析においては，手がかり表現や語の連鎖などの談話素性を利用し，文章の構造を解析している．我々のモデルでは，ドメイン固有の語の分布とトピック遷移を学習する Barzilay らのモデルに，ドメインに依存しない談話素性を取り入れることにより精度を向上させる．

さらに，言語情報だけでなく画像情報・音声情報も

☆ これらのトピックは「1 から始める料理の基本」 http://www.recipe.nestle.co.jp/from1/sitemap.htm を参考にして設定した．

利用することによりロバストにトピック推定を行う．画像の情報としては，背景の色情報を利用することができる．たとえば，「炒める」「煮る」といった作業はガスレンジ台で行われるため，背景が黒色であることや，「下ごしらえ」「盛り付け」などの作業はまな板の上で行われるため，背景が白色であるといった情報を手がかりとすることができる．また，音声情報としては，トピックが変化するときに無音がおかれることが多く，これを利用する．

2. 関連研究

Barzilay らは，地震や経済といった 5 つのドメインを対象として，HMM を用いて生コーパスからトピックの遷移モデルを構築している[4]．トピック（地震ドメインでは，震度の情報，被害の情報など）が遷移しながら，トピック固有の言語モデル（bigram）を出力し，それを基に文が生成されるというモデルをとっている．この手法ではまず，あるドメインのテキストにおいて，単語 bigram を素性としたコサイン類似度を計算することにより，文のクラスタリングを行う．そして，状態 s_i の言語モデル $p_{s_i}(w'|w)$ と状態 s_i から状態 s_j への遷移確率 $p(s_j|s_i)$ をクラスタリング結果から計算する．その後，Baum-Welch アルゴリズムで尤度が収束するまでパラメータの再計算を行い，パラメータを推定する．また構築されたモデルを，文の並び順の決定と要約の生成の 2 つのタスクに利用し，精度が向上することを示している．Barzilay らは書き言葉を対象としているのに対し，我々はノイズを多く含んだ話し言葉を対象としているため，Barzilay らの手法をそのまま適用するとあまり精度が高くならない可能性がある．したがって本研究では，格フレームや発話タイプ認識などを利用することにより，発話の汎化やノイズの軽減を行っている．また，我々はマルチモーダルな素性やドメインに依存しない談話素性を利用しており，この点でも Barzilay らの研究と異なる．

また，これまで新聞記事や対話などを対象としてトピックセグメンテーションの研究が数多くなされてきた．Hearst は，同一語の繰返しによる語の結束性に基づいたテキストセグメンテーションの手法 TextTiling を提案している[5]．Galley らは，多人数の会議録を対象として，テキストセグメンテーションを行っている[6]．語彙連鎖や手がかり表現などの言語情報や，無音や話者交替などの音声情報を利用し，話題の句切れ目かそうでないかを決定木で学習している．これらの研究は，テキストを意味のあるまとまりに分割することを目的としており，分割されたテキストのトピックを推定することはできない．

Marcu は，RST（Rhetorical Structure Theory）の枠組みで談話構造タグ付きコーパスを作成し，機械学習に基づく談話構造解析を行っている[7]．RST ではテキストを木構造でとらえており，このモデルは論説文などを対象として局所的な修飾関係をとらえることには適しているが，トピックの遷移といった文章全体の大きな構造をとらえることには適していない．

また映像解析の分野において，HMM を用いてショット解析の研究が活発に行われている．Chang らは HMM を用いてショットの解析を行い，野球放送からハイライトシーンを抽出している[1]．また，Nguyen らは，野球放送からハイライトシーンを抽出することを目的として，ビデオデータをインデキシングするための統計的なフレームワークを提案している[2]．各フレームの主成分分析による特徴量，テクスチャ情報であるフラクタル特徴量，移動物体の情報を反映した差分特徴量の 3 つの特徴量を用いて，マルチストリーム HMM でモデル化している．Phung らは，教育ビデオを対象として，階層的 HMM を用いてトピックの遷移を学習している[3]．

映像解析では言語情報を利用した研究は少ないが，いくつかの研究において，映像中の言語情報を利用している．Jasinschi らはベイジアンネットワークに基づき，音声，画像，テキスト情報を用いてテレビ番組のセグメンテーションとインデキシングを行っている[8]．色・エッジ・形状といった画像情報や，ピッチや零交差数といった音声情報に加えて，テキスト情報としては，天気，スポーツ，災害といったカテゴリそれぞれに対してあらかじめキーワードを用意しておき，各カテゴリのキーワードを含む割合を素性として利用している．また，Babaguchi らは，アメリカンフットボールの映像を対象とし，映像から人物・アクション・イベントを抽出し，アノテーションを行っている[9]．アクション・イベントを抽出する際に，各プレイ（“RUN”，“PASS”，“KICKOFF” など）を認識するためのキーワード，たとえば，“RUN” の場合，“UP TO place”，“PASS” の場合，“MAKES CATCH” などをあらかじめ人手で用意しておき，クローズドキャプション中の単語がキーワードにマッチしたら，そのキーワードのプレイに決定している．これらの研究では，言語情報としてキーワードしか用いられておらず，またドメイン固有のキーワードを人手で用意する必要がある．これらの研究に対して，本研究では，ドメイン固有のキーワードを人手で用意する必要がなく，また，格フレームによる発話の汎化，発話タイプの認識，談話素

性といった豊富な言語情報を利用している．

3. 利用する特徴量

まず，トピック推定に利用する特徴量について述べる．利用する特徴量は言語，画像，音声情報からなり，**表 1** にまとめる．

言語情報として，教示者による発話を利用する．教示発話の場合，作業に関する発話が中心であるが，コツや注意，雑談などといった発話も含まれる．トピックの推定には，「切る」「火をつける」「のせる」などといった作業に関する発話が有用であり，その他の発話はノイズとなってしまう可能性がある．そこで，発話の中から作業に関する発話だけを抽出する．

Barzilay らは素性として単語バイグラムを利用しているが，これではデータスパースネスの影響を受けてしまう可能性があるので，本研究では，発話の汎化を行う．最も単純な方法としては，たとえば，「塩を入れる」という作業に関する発話を「入れる」とするように，発話を用言に汎化してしまうことであるが，これでは，「塩を入れる」や「砂糖を鍋に入れる」，「包丁を入れる」がすべて「入れる」となってしまう．本研究では用言の原形ではなく，格・省略解析時に対応付けられた格フレームに汎化することを行う．

しかし，料理ドメインに限定しても作業に関する用言は多様であり，依然としてデータスパースネスの影響を受けてしまい，高精度にトピックを推定することができない可能性がある．そこで，画像の情報として，背景の色情報を利用することを考える．たとえば，「炒める」「煮る」といったトピックのときは作業がガスレンジ台で行われるため，背景が黒色であることや，「下ごしらえ」「盛り付け」などのトピックのときは作業がまな板の上で行われるため，背景が白色であるといった情報を手がかりとすることができる．

またこれらに加えて，トピックが変化したことを示す手がかり表現や無音，トピックが同一であることを示す語連鎖や用言の一致などを利用する．

以下，各特徴量について順に詳しく述べる．

3.1 言語情報

テキストとして料理番組に付随するクローズドキャプション☆を利用する．クローズドキャプションの例を**図 2** に示す．クローズドキャプションに対して JUMAN [10]・KNP [11] で形態素・構文・格・省略解析を行い，その結果に対して以下の処理を行い，言語情報を抽出する．

☆ 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．

表 1 トピック推定に用いる特徴量
Table 1 Features for topic identification.

	特徴量	ドメイン依存	ドメイン非依存
言語	格フレーム	発話を汎化	
	手がかり表現		談話素性（トピック変化）
	語連鎖		談話素性（トピック同一）
	用言の一致		談話素性（トピック同一）
画像	背景画像	画面下部の RGB	
音声	無音		談話素性（トピック変化）

3.1.1 作業に関する発話の抽出と格フレームへの対応付けによる汎化と曖昧性解消

作業に関する発話を，格・省略解析時に対応付けられた格フレームの形で抽出する．これは，用言の曖昧性解消と汎化を行うためである．たとえば，「塩を入れる」や「砂糖を鍋に入れる」は「入れる (1)」に，「包丁を入れる」は「入れる (2)」に対応付けられる☆☆．この処理を以下で詳細に述べる．

発話タイプ認識：先に述べたとおり，コツや注意，雑談といった発話はトピック抽出の際のノイズとなってしまう可能性があるため，作業に関する発話のみを抽出する．作業に関する発話を抽出するために，節を基本単位として発話タイプの認識を行う．発話タイプの分類は柴田ら [12] の研究のものを採用した．**表 2** に発話タイプを分類したものを示す．

まず文を節に分割し，節末の表層パターンを用いて発話タイプの認識を行う．発話タイプのうち，[作業：大]，[作業：中]，[留意事項]，[代替可]，[食品・道具提示]，[雑談]，[効果] は節末のパターンを記述することで認識することができる．たとえば，[作業：中] については「〜ましょう」「〜ていきましょう」など，[代替可] については「〜しても結構です/構いません/よい」など，[留意事項] については「〜できます」「〜しやすいです」「〜を目安にしてください」など，[雑談] については「こんにちは」「〜をお楽しみ下さい」などのパターンを計約 500 種類用意した．[料理状態]，[作業：小] については，他のドメインへの移植性を考え，「自動詞 + “なる”」を [料理状態]，その他を [作業：小] とする一般的なルールを用いた．

作業の抽出：発話タイプが作業（[作業：大]，[作業：中]，[作業：小] の 3 種類）と認識された節を抽出する．以下の (1) の例では，「むく」と「切る」を抽出し，(2) の例では，「さらす」のみを抽出する．

(1) にんじんは皮を むき [作業：小] 半分の長さに 切ります [作業：小].

☆☆ 括弧内の数字は格フレームの ID を示す．

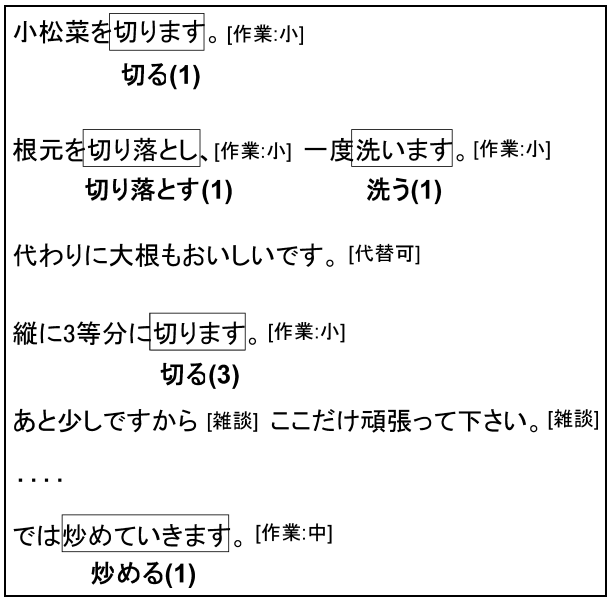


図 2 クローズドキャプションの例と作業の抽出・汎化（[ ] で括られたものは発話タイプ，四角で囲まれた文字は抽出される発話，発話の下の太字は格フレームを示す）
Fig. 2 An example of closed captions and action extraction/generalization. (A phrase sandwiched by a square bracket represents an utterance type and a word surrounded by a rectangle represents an extracted utterance refereeing to an action. A bold word represents a case frame assigned to a verb).

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

[**作業：大**]	[**留意事項**]
・ステーキの材料にかかります.	・最初に肉をパラパラに炒めることがポイントです.
[**作業：中**]	[**代替可**]
・強火で油を温めましょう.	・青ねぎでも結構です.
・じゃあ炒めていきましょう.	[**食品・道具提示**]
[**作業：小**]	・材料は，牛ひき肉，百五十グラムです.
・お鍋にお水を入れます.	[**雑談**]
[**料理状態**]	・暑くなってきましたね.
・ニンジンの水分がなくなりました.	[**効果**]
	・そうすると最初の方はアクが出てきます.

(2) ごぼうにはアクがあるので [料理状態] 酢水に さらします [作業：小].

発話タイプが作業であっても，今行われている作業への言及ではない場合があるので，以下の 2 つの例外処理を設ける．まず，発話末尾の節の発話タイプが作業でない文からは作業発話の抽出を行わない．以下の例では，発話末尾の節の発話タイプが [代替可] であるため，発話タイプが作業であるが「煮る」と「切る」は抽出しない．

(3) 煮てから [作業：小] 切っても [作業：小] 構いません [代替可].

また，理由節・条件節は前後のトピックに言及している場合があるので，作業発話の抽出を行わない．(4) の例では，条件節である「切る」は抽出せず，「炒める」のみ抽出し，(5) の例では，理由節である「揚げる」の節は抽出せず，「入れる」のみを抽出する．

(4) 切りましたら [作業：小] 炒めていきます [作業：中].
(5) プチトマトは油で揚げるので [作業：小]，切り込みを 入れます．[作業：小]

なお，節間の関係は，節末が「〜たら/なら/けば」ならば条件節，「〜ので/から/ため」ならば理由節とするなどの表層パターンで認識する．

表 3 格フレームの例
Table 3 Examples of the automatically constructed case frame.

用言	格	用例
切る (1)	ガ	【主体】
	ヲ	豚肉，大根，こんにゃく，···
	ニ	正方形，楕形，三角形，···
切る (2)	ガ	【主体】
	ヲ	水気，水分，汁気，···
	ノ	なす，豆腐，肉，···
入れる (1)	ガ	【主体】
	ヲ	塩，油，野菜，···
	ニ	鍋，ボール，容器，フライパン，···
入れる (2)	ガ	【主体】
	ヲ	包丁 ···
	ニ	魚，腹，付け根，···

表 4 トピックの変化を示す手がかり表現
Table 4 Cue phrases for topic change.

では，次は，次に，そうしたら，そうしましたら， それでは，それから，その間に，〜している間に， 最初は，今度は，最後は，続いて，まずは，早速， 次の料理は，さあでは，さっでは

格フレームへの対応付けによる曖昧性解消：
一般に用言は複数の意味を持つ．たとえば，「入れる」という用言は，「塩を入れる」，「包丁を入れる」において異なる意味を持ち，これらは異なるトピックで現れる．そこで，用言の原形ではなく，格・省略解析時に対応付けられた格フレームを言語情報の基本単位とする．用言格フレームは，自動収集した料理 Web テキスト 1200 万文☆から河原らの手法[13] で自動構築した．自動構築した格フレームの例を**表 3** に示す．

3.1.2 手がかり表現

Grosz ら[14] などが指摘しているように，トピックの変化を示す手がかり表現がある．本研究では，**表 4** にあげる約 20 個を利用した．

3.1.3 語 連 鎖

TextTiling を代表とするテキストセグメンテーション手法では，各文において，その左右に一定幅の窓を設け，左右の窓間における語彙的結束性を計算し，その値が閾値以下である点でテキストを分割している．この語彙的結束性はトピックが同一であることを検出する手がかりとして利用することができる．

隣接する 2 節の作業が同一の食材に対して行われている場合，それらのトピックは同一である可能性が高いと考えられる．以下の例では，「おろす」，「上げる」ともに同一の食材「かぶら」に対して行われているので，これらの 2 節は同一のトピック（この例では「下ごしらえ」）である可能性が高いといえる．

(6) a. かぶら をおろし金でおろしていきます．
b. おろした かぶら をざるに上げます．

しかし，本研究で扱っている話し言葉の場合，省略が頻繁に起こるので，表層的な文字列の照合では語連鎖を検出できないことが多い．そこで，用言と名詞の省略解析結果も利用し，語連鎖の検出を行う．用言，名詞の省略解析が行われた例をそれぞれ，(7)，(8) に示す☆☆．用言，名詞の省略解析はそれぞれ Kawahara ら[15]，Sasano ら[16] の手法により行った．

(7) a. キャベツ を切ります．
b. 一度 [キャベツ を] 洗います．

(8) a. にんじん を大体 4 cm くらい切ります．
b. [にんじん の] 皮をぐるっとむきます．

3.1.4 用言の一致

隣接する 2 節の用言が一致する場合，それらのトピックは同一である可能性が高いと考えられる．格要素が同じ場合と異なる場合があるが，いずれの場合も同一のトピックであると考えられる．

(9) a. ごぼうにはアクがあるので酢水に さらします．
b. まずはこうして酢水に さらします．

(10) a. とうがらしを 入れて下さい．
b. 鶏手羽を 入れます．

3.2 画 像 情 報

現在の画像処理技術では，人手による強い作り込みなどを行わなければ，映像中から何が映っているのか，またはどのような動作が行われているかといった情報を抽出することは難しい．したがって，浜田ら[17] の研究を参考にし，比較的安定して情報を抽出することができる背景画像に着目する．教示映像の場合，手元のショットの背景画像は画面の下部にあたるため，**図 3** に示すように，各発話の開始時刻における画面下部の RGB の重心を特徴量とする．

3.3 音 声 情 報

料理映像では，教示者による発話，包丁の音やガスレンジの火の音などの調理音があり，包丁の音，ガスレンジの火の音などを識別できればトピック推定に利

☆ Web ページにおいて，大匙，卵，油，味，野菜，皮などといった料理ドメインで頻出する単語の出現する割合がそのページで出現する名詞の 5%以上であるものを料理 Web ページとして取得した．

☆☆ [ ] は省略が補われたことを示す．

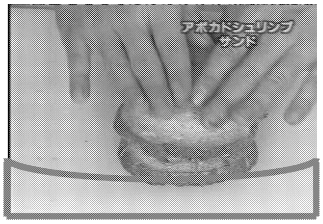


図 3 背景画像の例
Fig. 3 An example of background image.

用することができる．須場は，料理音を，連続性音（焼く音，炒める音など），インパルス性音（切る音，たたく音など），ランダム性音（混ぜる音など）に分類し，それらの検出を行っている[18]．連続性音に関しては精度約 80%で検出できているが，インパルス性音については精度約 30%，ランダム性音については特徴をとらえるのが困難であったと報告しており，調理音を識別することは容易ではない．

本研究では調理音の識別といった処理を行わず，比較的扱いやすい無音区間を利用する．Galley ら[6] などが指摘しているように，トピックが変化するときに間がおかれることが多く，無音がトピックの変化を検出する手がかりとして利用することができる．本研究では，音声の振幅が閾値以下である部分が 1 秒以上続くときを無音とした．なお，音声は，量子化ビット 16 ビット，サンプリングレート 16 kHz とし，無音区間を検出する閾値を予備実験により，振幅レベル 3000 に設定した．

4. HMM によるトピック推定

隠れ状態がトピックにあたり，3 章で説明した種々の特徴量が出力シンボルとして観測される HMM でトピックの推定を行う（図 1）．このモデルでは，発話を汎化した格フレームを基本単位とし，格フレームと背景画像は隠れ状態から出力され，トピックが同一/変化を示す談話素性（手がかり表現，語連鎖，用言の一致，無音）は隠れ状態を遷移するときに出力される．

4.1 パラメータ

HMM のパラメータを以下にあげる．

- 初期状態確率 π_i：状態 s_i から映像が始まる確率．
- 状態遷移確率 a_{ij}：状態 s_i から状態 s_j への遷移確率であり，トピックの遷移確率にあたる．
- 出力シンボル確率密度 $b_{ij}(o_t)$：状態 s_i から状態 s_j から遷移するときにシンボル o_t が出力される確率密度であり，次式で与えられる．

$$b_{ij}(o_t) = b_j(cf_k) \cdot b_j(R,G,B) \cdot b_{ij}(\textit{discourse features}) \quad (1)$$

- 格フレーム $b_j(cf_k)$：状態 s_j から格フレーム cf_k が出力される確率．
- 背景画像 $b_j(R,G,B)$：状態 s_j から背景画像の色情報 (R,G,B) が出力される確率密度．出力を連続量として扱い，平均 (R_j,G_j,B_j)，分散 σ_j の正規分布で確率密度を与える．

$$b_j(R,G,B) = \frac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left\{-\frac{(R-R_j)^2+(G-G_j)^2+(B-B_j)^2}{2\sigma_j^2}\right\} \quad (2)$$

- 手がかり表現，語連鎖，用言の一致，無音 $b_{ij}(\textit{discourse features})$：状態 s_i から状態 s_j に遷移するときに各特徴量が出力される確率．状態 s_i と s_j に依存するのではなく，隣接する状態が同一であるか異なるかに依存すると近似し，隣接する状態が同一の場合（$i=j$）は p_{same}，異なる場合（$i \neq j$）は p_{diff} とする．たとえば，手がかり表現（cue phrase）が出力される確率は以下の式で与えられる．

$$b_{ij}(\textit{cue phrase}) = \begin{cases} p_{same}(\textit{cue phrase}) & (i=j) \\ p_{diff}(\textit{cue phrase}) & (i \neq j) \end{cases} \quad (3)$$

4.2 パラメータの学習

これらのパラメータを，教師なし学習手法である Baum-Welch アルゴリズムで学習する[19]．一般に HMM で高い精度を達成するには，正解ラベル付きデータを作成するか，ドメイン固有の知識に基づく適切な初期パラメータの設定などが必要となる．しかし，これらを採用すると，他のドメインへの移行が困難となってしまう．

そこで，言語手がかりを用いて隠れ状態の直接的推定を行う．すなわち，節の発話タイプが [作業：大] または [作業：中] で，用言の原形がトピック名と一致する場合，節のトピックを直接推定する．この処理が行えるのは，[作業：大] または [作業：中] はひとまとまりの作業を始める宣言であるためである．たとえば，図 1 の節「では炒めていきます.」は，発話タイプが [作業：中] で，用言の原形がトピック名「炒める」なので，この節のトピックを「炒める」と推定する．

上記の方法で直接的推定が行われた節においてはそのトピックに固定しながら，HMM のパラメータの学習を行う．そして，パラメータを学習した後，Viterbi

アルゴリズムを用いてトピックの推定を行う．

なお，トピックの固定を行わない完全教師なし学習の場合は，学習された結果を見て，各クラスタがどのトピック名に対応するかの判断をする必要がある．その方法よりも，言語表現によって直接トピックを推定したものを加えることにより，学習されたクラスタに自動的にトピック名がつき，また，精度が向上する本手法の方がよいと考えた．

5. 実験と考察

5.1 デ ー タ

提案手法の有効性を確かめるために，2 つの料理番組，NHK の「きょうの料理」と NTV の「キューピー 3 分クッキング」の映像を用いて実験を行った．利用したデータの特徴を**表 5** に示す．なお，クローズドキャプションの各発話にはタイムスタンプが付随されており，タイムスタンプ時に発話と背景画像が出力されると考えることにより，クローズドキャプションと映像の同期をとった．4.2 節で述べた，言語手がかりにより隠れ状態が推定された節の割合は，「きょうの料理」において，3.87%（525/13564），「キューピー 3 分クッキング」において，5.74%（107/1865）であった．

5.2 実験と考察

2 番組においてモデルを学習し，学習されたモデルをそれぞれの番組 5 日分に適用して実験を行い，トピックが正しいかどうかを節単位で評価した．**表 6** に実験結果を示す．言語情報と画像情報の両方を利用することにより，言語情報または画像情報の一方を利用するよりも精度が向上していることが分かる（ただし，画像情報のみを利用する場合も，言語表現により隠れ状態を直接推定した結果を利用している）．また，それらに加えて種々の談話素性も利用することにより精度が向上した．無音の情報を入れても精度が上がらなかったのは，クローズドキャプションと実際の映像には若干の時刻のずれがあるためだと考えられる．この問題には，Huang ら[20] が行っているようにクローズドキャプションと映像の正確なアライメントをとるか，音声認識技術の向上にともない，クローズドキャプションではなく音声認識結果を言語情報として利用することにより対処する予定である．

画像情報を加えることにより成功した例を**図 4** に示す．言語情報だけの場合は，「炒める」という発話からトピックを「炒める」と解析しているが，実際はこの発話は後に行う作業について述べている．背景が白色の情報を利用することにより，トピックが下ごしらえと正しく解析することができている．

表 5 実験に用いた料理映像
Table 5 Characteristics of two cooking programs we used for our experiments.

番組	きょうの料理	キューピー 3 分クッキング
データ量	200 日分	70 日分
1 日あたりの放送時間	25 分	10 分
1 日あたりの発話数	249.4 発話	183.4 発話

談話素性を加えることにより成功した例を以下に示す☆．

(11) a. その間にみそ汁の ひき肉 を鍋に入れます．
b. （ひき肉 のうまみを抽出するので）水から煮立てます．
c. では ぶりを盛りつけます．

格フレームと背景画像だけを利用した場合，a を「煮る」，b と c を「盛り付け」と解析してしまう．これは，b と c の背景画像が a と b よりも類似しており，発話 b と発話 c が同じトピックである確率が高いと解析されたためである．格フレームと背景画像に加えて談話素性を利用した場合，a と b を「煮る」，c を「盛り付け」と正しく解析することができている．これは a と b 間における「ひき肉」の語連鎖と，b と c 間における手がかり表現「では」が考慮されているためである．すなわち，語連鎖がある場合はトピックが同一がである確率が高くなり，手がかり表現がある場合はトピックが変化する確率が高くなることが学習され，それらが考慮されて正しく解析されている．

また，3.1 節で述べた，発話タイプ認識による作業発話の抽出，格フレーム選択による汎化と曖昧性解消といった言語処理が妥当であるかを示すために実験を行った．我々の手法と，格フレームを利用せずに用言の表記を利用する手法（**w/o cf**），発話タイプの認識を行わずにすべての発話をトピック推定に利用する手法（**w/o utype**），Barzilay らと同じように言語モデルを生成する手法（**bigram**）を比較した．実験結果を**表 7** に示す．我々の行った言語処理による手法の精度が最も高く，マクネマー検定[21] を行ったところ，有意水準 5%で有意差があった（**proposed** vs **w/o cf**：$P = 0.0219$，**proposed** vs **w/o utype**：$P = 0.00671$，**proposed** vs **bigram**：$P = 1.23e^{-8}$）．よって，発話タイプ認識，格フレーム選択などの言語処理が妥当であることが示された．

以下にトピック推定の主な誤り原因を示す．

☆ 3.1.1 項で述べたとおり，理由節は前後のトピックに言及している場合があるので，文 b の理由節「ひき肉のうまみを抽出するので」はトピック推定に利用しない．

表 6 トピック推定の実験結果
Table 6 Experimental result of topic identification.

用いる特徴量				精度	
格フレーム	背景画像	談話素性	無音	きょうの料理	キューピー 3 分クッキング
✓				61.7%	66.4%
	✓			56.8%	72.9%
✓	✓			69.9%	77.1%
✓	✓	✓		**70.5%**	**82.9%**
✓	✓	✓	✓	70.5%	82.9%

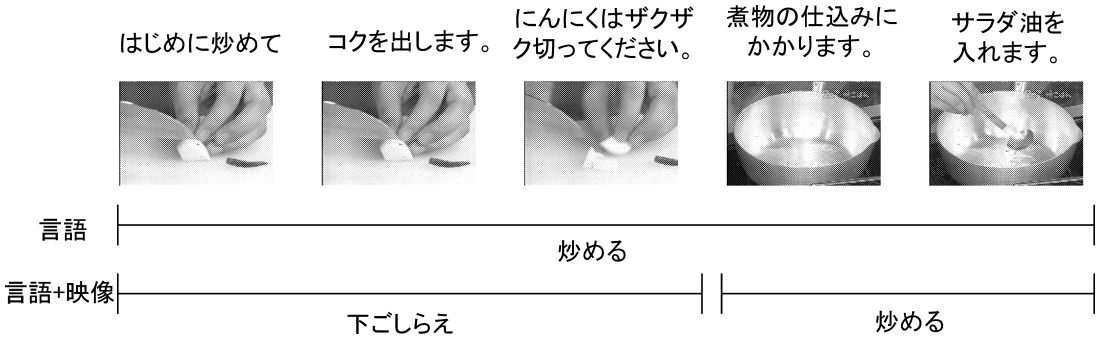


図 4 映像情報を加えることにより成功した例
Fig. 4 An improved example by adding visual information.

表 7 様々な手法の比較実験結果
Table 7 Results of the experiment that compares our method to three methods.

手法	精度	
	きょうの料理	キューピー 3 分クッキング
proposed	**61.7%**	**66.4%**
w/o cf	57.1%	60.0%
w/o utype	61.7%	62.1%
bigram	54.7%	59.3%

自動構築された格フレームの誤り 自動構築された格フレームに誤りがあるものがあった．たとえば，「野菜を切る」と「油を切る」がどちらも「切る (1)」の格フレームとなってしまっている．HMM のパラメータの学習では，格フレーム「切る (1)」はトピック「下ごしらえ」での出力シンボル確率が高いと学習されているので，トピック「揚げる」で出現する発話「油を切る」を「下ごしらえ」と解析してしまう．この問題には，格フレームのクラスタリングの際の閾値を改善することにより対処する予定である．

格解析・省略解析の誤り 格解析・省略解析の失敗により，誤った格フレームに対応付けられていたものがあった．以下の文において，「汁を切る」が，格フレーム「切る (3)」({ 肉，鶏肉，パン，…} を {<数量>} に切る）の格フレームに対応付けられていた．

(12) では小豆の汁を切ります．

これは，汁の意味素には，<液体（その他）>と<料理>があり，<料理>と { 肉，鶏肉 } などの類似度が高いことが原因である．この問題には，名詞の曖昧性解消を行ったうえで格解析・省略解析を行うことにより対処する予定である．

顔カットの利用 教示映像の場合，教示者の手元のショットが中心であるが，教示者のバストショットも存在する．発話のときのショットがバストショットの場合，そこからは手元で行われている作業の背景画像の情報を得られず，トピック推定の手がかりを得ることができない．

現在の顔認識技術では放送映像に対して適用すると誤りが多く，本研究では採用しなかったが，顔認識技術の向上にともない，顔ショットの認識を行い，発話時のショットが顔ショットと解析された場合，発話から最も近い手元のショットを背景画像として利用することができると考えられる．また，顔のショットはトピックが変化するときによく挿入されるので，トピックが異なることを示す手がかりとして利用することができると考えられる．

6. おわりに

本稿では，言語情報と映像情報を統合し，HMM を用いてトピックの推定を行った．HMM では，隠れ状態がトピックにあたり，言語，画像，音声情報の様々な素性が観測される．2 つの料理番組で実験を行ったところ，提案手法の有効性を示すことができた．

本稿では画像情報として背景画像のみを利用したが，今後，顔画像認識や物体認識結果[22)]を素性として取り入れる予定である．

参 考 文 献

1) Chang, P., Han, M. and Gong, Y.: Extract highlights from baseball game video with hidden markov models, *Proc. International Conference on Image Processing 2002* (*ICIP2002*), pp.609–612 (2002).
2) Nguyen, H.B., Shinoda, K. and Furui, S.: Robust highlight extraction using multi-stream Hidden Markov Models for baseball video, *Proc. International Conference on Image Processing 2005* (*ICIP2005*), pp.173–176 (2005).
3) Phung, D.Q., Duong, T.V., Bui, H.H. and Venkatesh, S.: Topic Transition Detection Using Hierarchical Hidden Markov and Semi-Markov Models, *Proc. ACM International Conference on Multimedia* (*ACM-MM05*), pp.6–11 (2005).
4) Barzilay, R. and Lee, L.: Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization, *Proc. NAACL/HLT*, pp.113–120 (2004).
5) Hearst, M.: TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, Vol.23, No.1, pp.33–64 (1997).
6) Galley, M., McKeown, K., Fosler-Lussier, E. and Jing, H.: Discourse Segmentation of Multi-Party Conversation, *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pp.562–569 (2003).
7) Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach, *Computational Linguistics*, Vol.26, No.3, pp.395–448 (2000).
8) Jasinschi, R., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J. and Li, D.: Integrated multimedia processing for topic segmentation and classification, *Proc. IEEE International Conference on Image Processing* (*ICIP2003*), pp.366–369 (2001).
9) Babaguchi, N. and Nitta, N.: Intermodal Collaboration: A Strategy for Semantic Content Analysis For Broadcasted Sports Video, *Proc. IEEE International Conference on Image Processing* (*ICIP2003*), pp.13–16 (2003).
10) 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).
11) 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).
12) 柴田知秀，黒橋禎夫：料理教示発話の理解と作業構造の自動抽出，情報処理学会自然言語処理研究会，No.2004-NL-164, pp.117–122 (2004).
13) Kawahara, D. and Kurohashi, S.: Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis, *Proc. 19th COLING* (*COLING02*), pp.425–431 (2002).
14) Grosz, B.J. and Sidner, C.L.: Attention, intentions, and the structure of discourse, *Computational Linguistic*, Vol.12, pp.175–204 (1986).
15) 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).
16) 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*, pp.1201–1207 (2004).
17) 浜田玲子，井出一郎，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，Vol.J85-D-II, No.1, pp.79–89 (2002).
18) 須場康貴：料理映像の索引付けのための音響解析手法，研究生報告書，東京大学大学院 (2002).
19) Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P.: *The HTK Book* (*for HTK Version 3.4*), Cambridge University Engineering Department (2006).
20) Huang, C.-W., Hsu, W. and Chang, S.-F.: Automatic Closed Caption Alignment Based on Speech Recognition Transcripts, Technical report, Columbia ADVENT (2003).
21) Gillick, L. and Cox, S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *Proc. International Conference on Acoustics Speech and Signal Processing* (*ICASSP*), pp.532–535 (1989).
22) 加藤紀雄，柴田知秀，黒橋禎夫：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，言語処理学会 第 11 回年次大会 (2005).

(平成 18 年 5 月 17 日受付)
(平成 19 年 3 月 1 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．



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．