

表 1: 発話タイプの分類 (下線部は認識するためのパターンを示す)

[作業:大]	ex. さ, では, ステーキにかかります。
[作業:中]	ex. じゃあ炒めていきましょう。
[作業:小] (いずれのパターンにもマッチしないもの)	ex. なすはヘタを取ります。 きゅうりは半分に切って、
[料理状態] (自動詞)	ex. ニンジンの水分がなくなりました。 ごぼうにアクがあるので、
[留意事項]	ex. 芯は切らないで下さい。
[代替可]	ex. 青ねぎでも結構です。
[食品・道具提示]	ex. 今日はこのハンドミキサーを使います。
[程度]	ex. 今度は 3 分です。
[雑談]	ex. こんにちは。

を意味のある単位に分割するセグメンテーションは映像の構造化の第一歩として捉えられており、分割されたセグメントは検索のインデックスや要約の単位として利用される。

一般に映像のセグメンテーションはショットを単位として行なわれる。ショットとは単一のカメラによって切れ目なしに撮影されたものである。しかし、ショットは要約の単位としては小さい可能性があり、意味的につながりのあるショットの集合であるシーンで分割するのが適当であると考えられる。

そこで、以前我々が提案した、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手法 [5] を用いてセグメンテーションを行なう。この手法では、節を単位として、格フレームや談話素性、背景画像を利用して各節のトピック (下ごしらえ、炒める、盛り付けなど 8 個) を隠れマルコフモデルを用いて推定する。本研究では、このトピック推定結果を用いて、トピックが変わったところでセグメンテーションを行なう。トピック推定では、ノイズを軽減するために作業に関する発話のみを利用している。そこで、作業以外の発話のトピックについては、談話構造解析結果を利用して推定する。図 2 の例では、文「片栗粉をまぶすと、ふっくら仕上がります」は前文と主題連鎖の関係が解析されているので、この文のトピックは前文のトピック「下ごしらえ」と同一とし、この文と次の文の間でセグメンテーションを行なう。

ただし、トピックが「下ごしらえ」の場合、「食材 A の下ごしらえ」→「食材 B の下ごしらえ」のよう

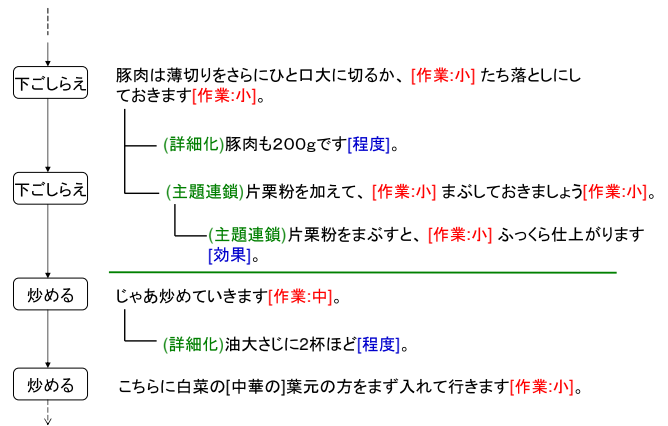


図 2: セグメンテーション

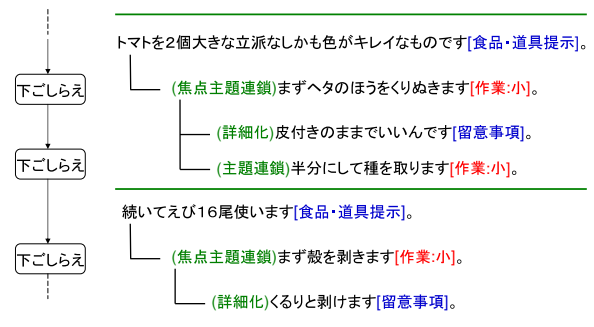


図 3: セグメンテーション (トピックが下ごしらえの場合)

に、複数の食材からなることがあるので、図 3 に示すように、さらに談話構造の部分木でセグメンテーションを行なう。

3.2 代表画像の抽出

次に、各セグメントの代表画像を抽出する。教示映像の場合、顔ショットと手元ショットからなり、手元ショットは行なれている作業に関する情報量が多いが、顔ショットにはあまり情報量がない。そこで、セグメントに最も近い手元ショットを探し、その手元ショットから代表画像を抽出する。手元ショットにおいて物体が認識されたかどうかで以下のように場合分けする。物体認識は加藤ら [2] の手法を用いて行なった。

● 物体が認識された場合

図 4 に示すように、ショット内で、その物体の領域の面積が最も大きい画像を代表画像とする。

● 物体が認識されなかった場合

物体がなるべくアップになっている画像を代表画像とする。アップかどうかの判定は、エッジ率 (エッジが検出された画素/全画素) を計算することにより行ない、エッジ率が最も小さい画像を代表画像とした。





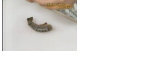
	画素数	トマトを2個大きな立派なしかも色がキレイなものです。
	7673	
	16875	半分にして種を取ります。
	1046	これをひと口大に切ります。
		続いてえび16尾使います。

図 4: 代表画像の抽出

4 重要発話の抽出と整形

4.1 重要発話の抽出

作業教示発話から要約に出力する重要発話の抽出を行なう。映像の要約において、ユーザはまずは料理の手順について知りたいであろうから、作業に関する発話を抽出し、それらを提示する。

文末の節の発話タイプが作業である文を抽出対象とし、そうでない場合は抽出を行なわない。以下の例では、文 (1) は抽出するが、文 (2) は抽出しない。

- (1) 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]
- (2) 煮てから [作業:小] 切っても [作業:小] 構いません [代替可]。

また、以下に示すように、教示者はしばしば同じ内容の発話をすることがある。

- (3) 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。

これは映像を見る人にとっては何度も説明され、わかりやすいが、要約にこれら両方の発話を含めると、見にくくなってしまふ恐れがある。したがって、先述した談話構造解析で繰り返しの発話と認識された場合²、後の文は要約に含めないようにする。

4.2 重要発話の整形

生成された要約が見やすくなるように、4.1 節で抽出した文において以下にあげる整形を行なう。

- 発話タイプが作業でない節を削除

以下の文では、節「なかなか味の馴染みが悪いんで、」の発話タイプが料理状態なので削除される。

- (4) なかなか味の馴染みが悪いんで、 [料理状態] しっかりと混ぜていきます。 [作業:中]

²用言の原形とその格要素が一致するかどうかで認識している。

- 発話タイプが作業である各節において、「接続詞」「副詞」などの項を削除

(5) そして火にかけます (接続詞)。

(6) しっかり水気を切りましょう (副詞)。

- 提題助詞句における提題助詞の格助詞による置換
提題助詞句において、用言の格解析結果に基づき、提題助詞を格助詞で置換する。以下の例では、用言「切る」の格解析結果で格助詞「を」と解釈されたので、要約を生成する際には「きゅうりを」と変換する。

(7) きゅうりは半分に切ります。

- 省略の補完

省略解析で認識された格要素を補完する。以下の文では、省略解析結果「アクを」が補完される。

(8) アクがあるので、 [料理状態] 少し (アクを) 取ります。 [作業:小]

ただし、トピックが下ごしらえで、省略解析結果が物体認識結果と同一の場合は、省略された要素を補完しなくても明らかであるので、省略の補完を行なわない。

- 節末の用言の整形

節末の用言を原形にすることで整形を行なう。

例) 半分に切ります → 半分に切る

以上述べた処理によって、以下のような整形が行なわれる。

- (9) 少し火を弱めて、煮込んで行きます。 → 火を弱めて煮込む
- (10) アクがあったら、ちょっととります。 → アクをとる

5 要約の生成

以上の処理により抽出した、代表画像、重要発話を並べることにより要約とする。トピックが「下ごしらえ」で、物体認識結果がある場合はそれを表示する。また、紹介されている料理名をクロードキャプションから自動抽出し、要約の一番上に出力する。以下の例に示すとおり、料理名はクロードキャプションで“「」”でくられていることが多いというヒューリスティックを利用し、料理名を抽出する。

- (11) 今日は「トマトとえびのスパゲティ」です。



図 5: 生成した要約の例とその誤り例

表 2: 文抽出の精度

適合率	再現率	F
73 / 76 (0.961)	73 / 81 (0.901)	0.930

生成された要約では、代表画像をクリックすることにより、その時刻からの映像をクロズドキャプション付きで再生することができる。

6 評価

NTV「キューピー 3分クッキング」3番組から要約を生成し、評価を行なった。生成した要約の例を図5に示す。トピック推定の精度は節単位で82.9%であった。誤り例としては、図5における上から4つ目のセグメントにおける「量る」がある。この元の文「事前に量っておいてください」のトピックは実際は下ごしらえであり、その上のセグメントにマージされるべきである。図1や図5の出力例からわかるように、少し誤りがあるものの、映像のおおまかな構造が捉えられていることがわかる。

物体認識の精度はショット単位で72.7%であった。誤り例としては図5の上から3セグメント目の「豆腐」がある。これは実際には調味料の説明をしている。

次に、文抽出の再現率、適合率を表2に示す。文抽出の精度は十分に高く、発話タイプの認識が高い精度で行なえていることがわかる。再現率を下げた誤り例としては以下の文がある。

(12) こしょうが入ります。

この文の発話タイプは料理状態と解析されるため、抽出されない。これを「こしょうを入れる」と解釈するのは今後の課題である。また、以下の文のように、用言が省略されている場合がある。

(13) オリーブオイル大さじ2杯。

³クロズドキャプションの誤りによる不正解2例を除いた。

表 3: 文整形の精度

正解	不正解	精度
41	30	0.577

表 4: 文整形の誤り原因

誤り原因	文数
用言省略解析誤り	24
名詞省略解析誤り	3
格解析誤り	2

この例の発話タイプは程度と解析されるため、抽出されない。この文では用言「入れる」が省略されており、格フレームなどを利用することにより、用言の省略の補完を行なう予定である。

次に、文整形の精度を表3に示す³。精度は57.7%であり、その誤り原因を表4に示す。原因の多くは省略解析の誤りによる間違った補完が行なわれたためであり、省略解析の精度向上は今後の課題である。

7 関連研究

三浦らは、我々と同じ料理映像を対象とし、要約を自動生成している[3]。料理番組において、料理動作および料理や素材の状態を示す部分が重要であるとしており、中でも動きの激しい部分に注目し、画面中の各点の速度場であるオプティカルフローを利用して動きベクトルを検出している。この研究での要約は代表画像のみからなり、言語情報は利用されていない。

8 おわりに

本稿では、作業教示映像、特に料理映像を対象とし、言語情報と映像情報を統合することにより、映像を自動要約する手法を提案した。本手法では、トップダウンに映像の構造を捉えることは十分にできていると考えられるが、省略解析をはじめとしたボトムアップな解析には誤りが多く、その改善は今後の課題である。

参考文献

- [1] Daisuke Kawahara, Ryohei Sasano, and Sadao Kurohashi. Toward text understanding: Integrating relevance-tagged corpus and automatically constructed case frames. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1833–1836, 2004.
- [2] 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, 3 2005.
- [3] 三浦宏一, 浜田玲子, 井出一郎, 坂井修一, 田中英彦. 動きに基づく料理映像の自動要約. 情報処理学会コンピュータビジョンとイメージメディア研究会論文誌, Vol. 44, No. SIG9, pp. 21–29, 2002.
- [4] 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, 11 2004.
- [5] 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. 言語処理学会 第12回年次大会, 3 2006.