料理映像の構造解析による調理手順との対応付け

三浦　宏一[†]　高野　求[†]　浜田　玲子[†]　井手　一郎[††]
坂井　修一[†]　田中　英彦[†]

Associating Semantically Structured Cooking Videos with their Preparation Steps

Koichi MIURA[†], Motomu TAKANO[†], Reiko HAMADA[†], Ichiro IDE[††], Shuichi SAKAI[†], and Hidehiko TANAKA[†]

あらまし　近年のマルチメディアデータの増加に伴い，その解析技術の重要性が増しつつあり，またそこに含まれる情報を最大限に利用するため，複数メディアを統合的に処理する手法が注目されている．そこで我々は，様々な映像の中でも生活に密着した料理映像を対象とし，映像とテキスト教材中の手順の対応付け手法を提案する．料理番組などテキスト教材の存在する教養映像においては，テキスト教材は映像よりも閲覧しやすい一方，映像にはテキスト教材では表現しきれない有用な視覚的情報が含まれており，これらを統合することによってより高次な映像の構造化・索引付けが期待される．本論文では，まず，映像の構造を解析し，対応付けの際の映像の単位となる映像ブロックを定義する．また，クローズドキャプションとテキスト教材中の手順を解析し，キーワード抽出による対応付け手法を提案・実装する．さらに評価実験により，提案手法を用いて，映像ブロック単位での対応付け，すなわち映像への索引付けが高精度で行われることを示す．また，対象を限定することで，比較的簡単な要素技術をうまく組み合わせることにより実用的な精度が得られることを示す．

キーワード　料理映像, テキスト教材, 対応付け, 映像ブロック, 映像への索引付け

1. まえがき

近年の情報通信技術の進歩にともない，入手可能なマルチメディアデータは増加の一途をたどっている．このような大量のデータを整理し，効率良く蓄積・検索するため，マルチメディアデータの解析はますます重要な技術となっている．

マルチメディアデータは，一般に画像，音声，テキストからなる．従来これら各メディアの自動解析手法については個別に研究されてきた．例えば画像解析については，カット検出や類似画像検索など，さまざまな研究がなされている．しかし，画像単独での一般的な物体認識が困難なように，画像のみからマルチメディアデータの意味的内容に立ち入って解析することは難しい．また，テキスト解析については，重要文や語句の抽出，要約作成など，高次元の意味的内容が比較的容易に解析できるが，テキストのみでマルチメディアデータの全体像を把握することは容易ではなく，映像と結び付けられることでより効果的に利用することができる．このように，各メディア単独でのマルチメディアデータの自動解析には限界があり，精度の高い意味的内容解析には複数メディアを統合的に処理することが必要である．

我々は，このような統合メディア処理手法の研究の一環として，料理映像を対象とした知的構造化・索引付けに取り組んでいる [1]．料理番組などテキスト教材の存在する映像において，テキスト教材は映像よりも閲覧が容易である一方，映像にはテキスト教材では表現しきれない有用な視覚的情報が多く含まれており，これらを統合的に処理することで互いの情報や弱点を補完し合うことが期待される．このような補完の結果，より高次な映像の構造化・索引付けの実現はもちろんのこと，テキストと映像をリンクさせた形の，利用しやすい新たなマルチメディアデータが生成されることが期待される．さらに，解析結果を利用した映像要約や知識抽出など，知的調理支援につながる応用が考え

[†] 東京大学大学院情報理工学系研究科, 東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033, Japan
[††] 国立情報学研究所, 東京都
National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, Japan

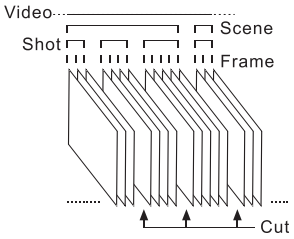


図 1　映像の画像的構造
Fig. 1　Graphical structure of video.

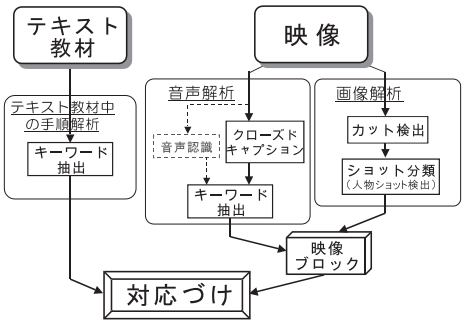


図 2　テキスト教材と映像との対応付け手法
Fig. 2　Method to associate a cooking video with a related textbook.

られる．

本論文では，料理映像の構造を解析し，調理手順単位での映像とテキスト教材中の手順とを対応付ける手法を提案する．一方，対象を料理映像に限定することで，既存の比較的単純な要素技術を用いながらも，これまで汎用的な手法では得られなかった実用的な精度を得ることを目指す．

まず，**2.** で映像とテキスト教材中の手順の対応付け手法の概要を示し，**3.** で対応付けのための料理映像の構造解析手法を提案する．さらに **4.** では，テキスト解析手法について，**5.** では，映像と手順の対応付け手法について提案する．**6.** では，**3.** で提案した映像解析手法及び **5.** で提案した対応付け手法の評価実験の結果を示す．また，その結果をふまえた評価と考察を行う．最後に **7.** でまとめとする．

2. 映像とテキスト教材中の手順の対応付け手法

2.1 映像の構成と用語の定義

本論文では，画像と同期して放送される音声や文字放送などを全て含んだデータの集合体を映像（video）と呼ぶ．映像の中の画像は多数のフレーム（frame）からできており，画像的に連続なフレームの集まりをショット（shot）と呼ぶ．またショットとショットの境（画像の変化点）をカット（cut），さらに意味的なまとまりのあるショットの集合をシーン（scene）と呼ぶ．このような映像の構造を図 1 に示す．

2.2 関 連 研 究

独立した外部テキストを映像と対応付ける研究としては，ニュース映像の字幕（オープンキャプション）と電子新聞の記事の構造情報を利用して類似度を計算し，ニュース映像中のトピックと新聞記事を対応付ける研究 [2] がある．このような研究では，対応付けによって，新聞記事から得られる意味情報をニュース映像の解析に利用できる．しかし，対応付けの際の映像内容解析としては，字幕のテキストを参照するのみであり，具体的な映像内容は考慮されていない．

また，DP マッチングを用いたドラマ映像とシナリオ文書の対応付けに関する研究 [3] においては，複数のメディアから参照できるパターンを抽出し，その対応付けを DP マッチングにより最適化することにより，メディア間の対応付けを行っている．この研究も対応付けにより，映像の構造化・データベース化を行っているが，料理番組ではしばしばテキスト教材中の手順と映像中の手順が入れ替わるのに対し，ドラマでは映像の順序とシナリオの順序とが基本的には入れ替わらない点で，本研究と本質的に異なる性質の対応付け手法である．そのため，本研究では単純に DP マッチングのような時系列に沿った一次元的な同期手法を用いることはできない．そこで，以下で述べるように，各々のメディア中のキーワードを参照して映像の構造に合わせた対応づけを行う．

2.3 提案手法の概要

対応付け手法の構成を図 2 に示す．まず，映像の構造とテキスト教材中の手順を並行して解析する．次に，それぞれの内容を統合的に判断して，映像と手順を対応付ける．

映像の構造解析は，画像と音声の両方から進める．画像については，カット検出によりショット単位に分割し，それぞれのショットを画像内容に応じて分類する．ここで必要な画像処理に関しては様々な要素技術が研究されており，本研究では，それら既存の手法を効率良く組み合わせて利用する．一方，音声については，本研究では音声認識は行わず，主音声の書き下し

であるクローズドキャプション(注 1)を用いて，これにテキスト処理を施す．テキスト処理としては，テキストを形態素解析した後，素材名などの名詞や調理動作を表す動詞などのキーワードを抽出する．そしてこれらの解析結果から，テキスト教材中の手順と対応付けるための映像の構造を抽出する．これについては **3.** で詳述する．

テキスト教材の解析についてもクローズドキャプションと同様に，形態素解析，キーワード抽出などを行う．これについては **4.** で詳述する．

最後にそれぞれの解析結果を利用して，映像と手順を対応付ける．これについては **5.** で詳述する．

3. 料理映像の構造解析

映像と手順を対応付けるために，料理映像の構造を解析する．映像は画像と音声の両面から解析するが，まず，画像解析によって映像の構造を解析し，手順と対応付ける際の単位となる映像ブロックを定義する．また，クローズドキャプションの解析によって対応付けの際のキーワードを抽出する．

3.1 画像解析

3.1.1 映像ブロックの定義

画像解析ではまず，カットを検出し，映像をショット単位に分割する．カット検出手法としては，色ヒストグラムや色コリログラムで画像の色調変化を検出する手法 [4] をはじめ，様々なものが検討されているが，本システムにおいては，DCT クラスタリングを利用する手法 [5] を採用する．料理映像は，スタジオ内の理想的な照明条件下で撮影されるため，高精度のカット検出が期待される．また，このカット検出手法は，カット検出と同時にそれぞれのショットを形成するクラスタの特徴量も得られることから，将来的にはその特徴量を，カット検出後に行うショット分類に利用することも考えられる．

映像の構造解析のため，カット検出の後にショットを分類する．図 3 に示すように，料理映像中のショットは，画像内の構成に基づき大きく（a）人物ショットと（b）手元ショットに分類できる．さらに，(a）人物ショットは（a1）全身ショットと（a2）上半身ショットに分けられる．

(注 1)：クローズドキャプションとは，主に聴覚障害者のために，主音声を書き下したものが文字放送の形で提供されるデジタルテキストであり，米国においては生放送も含め，大多数の番組において文字放送により提供されている．また日本においても，提供される番組数は増えつつあり，その中には料理番組も含まれる．

（a1）全身ショット （a2）上半身ショット （b）手元ショット

（a）人物ショット （b）手元ショット

図 3 料理映像におけるショット分類
Fig. 3 Shot categories in cooking video.

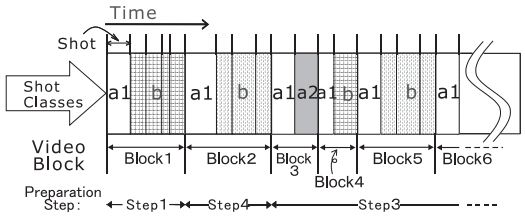


図 4 料理映像のショット構成の例
Fig. 4 Example of a shot based structure in a cooking video.

この分類による実際の料理映像におけるショット構成の例を，図 4 に示す．ここで，手順の区切りの直後のショットに着目すると，その 90%以上が人物ショット（a），その中でも全身ショット（a1）であった(注 2)．

このことより，一つの手順のまとまりとして「映像ブロック」を定義し，これを映像における対応付けの最小単位とする．映像ブロックは，図 4 に示すように，全身ショット（a1）を先頭とし，次の全身ショット（a1）が出現するまでの連続するショットの集合とする．

3.1.2 人物ショット検出

人物ショットの検出は，映像を映像ブロックに分割する際の手がかりとなるため，ショット分類の中でも特に重要である．人物ショットは画像中に人物の顔が映っているため，顔領域を抽出することで検出できる．

顔領域を抽出する手法には様々なものがあるが，ここでは単純に顔領域の存在・位置・大きさが分かれば充分なので，より高度で複雑な目や口などの位置をモデル化した手法は採用せず，

（1） 肌色領域を抽出

（2） 検出された肌色領域から一定の条件により顔領域を決定

という手順で単純かつロバストに抽出し，実用的な精度で人物ショットを検出することとする．

肌色領域の抽出には，それに適しているとされる修

(注 2)：複数の料理番組 38 レシピ 137 手順を調べた結果．

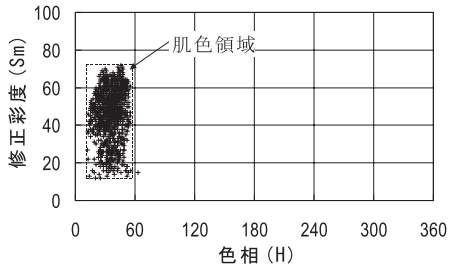


図 5　H-Sm 空間における肌色領域の分布
（実際の肌色領域のサンプリングに基づく）
Fig. 5　Skin colored regions on the H-Sm plane.

正 HSV 表色系 [6]（H：色相, Sm：修正彩度, V：明度）を用いた．

このうち，V については閾値処理を行い，暗い領域を排除するためのみに用いた．肌色の判定は，実際の様々な顔領域の色を参考に，H–Sm 空間中の矩形領域（図 5 [7]）を肌色領域と設定し，各ピクセルにおける H と Sm の値がこの領域に含まれるかどうかで行った．この際，2 値化された画像に対し 3×3 のメディアンフィルタをかけ，雑音を除去した．

肌色領域の抽出は，純粋に色のみを手がかりにしたものであるため，人間の顔だけでなく，手，木のしゃもじ，机など，似た色をもつ対象も検出してしまう．このようなものを排除するため，肌色領域を抽出した後，以下のような条件により顔領域を抽出した．

（1） 抽出された肌色領域に外接する矩形の縦横比 $r = x/y$ について，$r_{min} \leqq r \leqq r_{max}$ を満たすもののみを顔領域として抽出．
— 一般的に人間の顔が映る場合，その形にはある程度の制約があるため．またこの値は顔の向きによって大きく変化することから，本手法では値域をある程度の余裕をもって $r_{min} = 0.38$, $r_{max} = 1.4$ と経験的に設定した．

（2） 画面の端に接しているものは除外
— 料理映像においては，顔領域が画面の端に接するようなカメラワークが用いられることはほぼなく，肌色領域が画面の端に接する場合は，調理中の手などが映っていることが多いため．

（3） 面積が大き過ぎるもの（画面の 1/12 以上），相対的に小さいもの（最大面積の 30%未満）は除外
— 料理映像においては，顔領域は画面上である程度の大きさで映っており，大き過ぎたり小さ過ぎる場合は，顔以外のもの（壁，テーブルなど）を検出してしまっていることが多いため．また，料理映像においてはスタジオセットの構造上，同一ショットに顔領域が複数ある場合は，それらの面積は同程度となる．顔の向きなどによる面積の変化を考慮したうえで，相対的に小さいものも雑音として除外する．これらの閾値は，経験的に設定した．

（4） 少なくとも一部が画像の上半分の領域にある
— 料理映像中のショットの構図上，顔領域が画像の上半分にかからないことはないため．

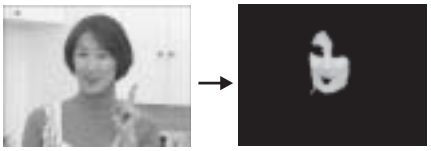


図 6　顔領域の抽出例
Fig. 6　Example of human face extraction.

これらの条件のうち，（1）に関しては，一般的な顔の画像特徴を利用したものである．また，（2）～（4）の条件は料理映像の性質から導き出した条件ではあるが，スタジオセット内で撮影される一般の映像にも適用できる条件である．

図 6 に顔領域の抽出例を示す．なお，（a1）と（a2）は顔領域を抽出した後，その面積を閾値処理（閾値：画面全体の 1/36）することによって分類した．

3.2　音声テキストの解析

本研究では，音声内容の解析のために，音声認識を行う代わりに，主音声の書き下しとして放送局から提供されるクローズドキャプションを利用し，これを音声テキストとして解析する．クローズドキャプションをテキストデータとして取得する際に，出現時刻もあわせて記録し，これを元に映像との同期をとる．

得られたクローズドキャプションテキストに形態素解析を施した後，素材名や調理に関わる動詞など，対応付けの手がかりとなるような語をキーワードとして抽出する．この詳細についてはテキスト教材中の手順の解析とまとめて **4.**で説明する．

4.　テキスト教材中の手順解析

映像とテキスト教材中の手順とを対応付けるためには，映像の構造解析と並行して手順のテキストも解析する必要がある．手順の解析については，クローズドキャプションと同様に形態素解析を施し，素材名を表

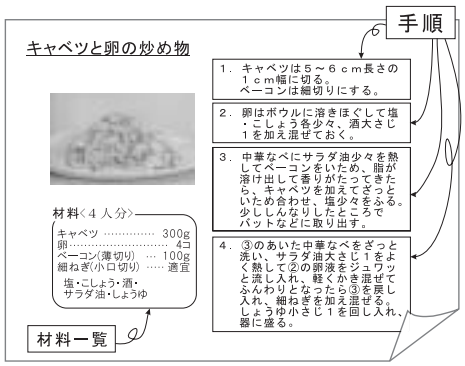


図 7　テキスト教材の例
Fig. 7　Example of a cooking textbook.

す名詞，調理動作を表す動詞などのキーワードを抽出する．

ここで，テキスト教材の例を図 7 に示す．テキスト教材は主に「手順」と「材料一覧」からなり,「材料一覧」は，素材名を抽出する際の重要な手がかりとなる．

クローズドキャプション及びテキスト教材中の手順に対するテキスト解析は，以下のように行った．

- 形態素解析は日本語形態素解析システム JUMAN [8] を用いる．
- 素材名は，テキスト教材中の材料一覧と一致するもののみを抽出する．この際，辞書を用いることで，ひらがな，カタカナ，漢字と表記の異なるものがあっても同一素材名として扱う（以下の動詞についても同様）．
- 形態素解析の結果から動詞を抽出するが,「下さる」「できる」「する（サ変動詞で名詞を伴わない単独のもの)」は料理に関係ないものとみなし，除外する．
- 上記の素材名と動詞のみをキーワードとする．
- 句点，または動詞の後の読点やスペース(注 3)を区切りとして 1 つの文とみなし，同じ文に属すると判断される素材名と動詞は，関連するものとみなす．

テキスト教材の手順を解析する際に，以前の手順番号を参照しており，かつその直後に「の＋名詞」が存在しない場合のみ，参照されている手順に含まれる素材名全てを補うこととした．ただし，補うのは直接参照している手順までとし，参照先の手順がさらに以前の手順を参照していても，それは補わないこととした．表 1 に手順の参照の具体例を示す．

(注 3)：料理映像におけるクローズドキャプションテキスト中では読点が使われることは少なく，スペースが多く使われる．

表 1　以前の手順を参照している手順の例
Table 1　Example of reference to previous steps.

手順 2:[1](手順番号) をすり鉢に入れ，小麦粉，砂糖を加えてすり混ぜる． ⟶ 参照している手順番号 [1] の直後に「の＋名詞」が存在しないため，手順 1 に含まれる素材名すべてをここに補う．
手順 5:[3](手順番号)の(の)鴨肉(名詞) を薄切りにして，器に盛る． ⟶ 参照している手順番号 [3] の直後に「の＋名詞」が存在するので，この場合は素材名を補うことはしない．
手順 4:フライパンで [1](手順番号) をいためる． 手順 5:[3](手順番号)の(の)なべ(名詞) を火にかけて [4](手順番号) を加え，煮る． ⟶ 手順 4 において，手順番号 [1] を参照しているので，手順 1 に含まれる素材名をここに補う． ⟶ 手順 5 は手順番号 [3] と手順番号 [4] を参照しているが，手順番号 [3] の箇所には，直後に「の＋名詞」が存在するので，素材名は補わない．また，手順番号 [4] の箇所に補うのは，手順 4 に含まれる素材名のみである．(手順 1 の素材名までは補わない．)

また，動詞に関しては，映像とテキスト教材での表現の違いを吸収するため，表 2 のような辞書を必要に応じて作成し，辞書中の調理動作と一致する動詞は正規化された動詞に置き換えて解析した．この辞書には，動作を上位概念に置き換えるもの，表記の差を吸収するものなどが含まれている．ここでは，表 2 に示すものを含めて，31 個の調理動作を表す動詞を辞書に収録した．なお，この辞書をさらに実用的なものとして扱うためには，料理辞典や他のレシピを用いて，語彙を増やす必要がある．

このテキスト解析処理は，テキスト教材中の「材料一覧」の語をキーワードとして用いる部分において，料理テキスト教材特有の性質を利用し，処理の単純化を図っているが，例えば，ドメイン毎に主体となるキーワードを記述した辞書を作成すれば，別ドメインのテキストに対しても適用できる手法である．

5.　映像ブロックと手順との対応付け

映像と手順の対応付けは，映像は映像ブロックを単位とし，また，手順はテキスト教材中の一つ一つの手

表 2　動詞辞書の内容
Table 2　Contents of the verb dictionary.

正規化後の動詞	調理動作
切る	千切りにする 薄切りにする ・・・
入れる	流し入れる 加える ・・・
切り目を入れる	切れ目を入れる 切り込みを入れる
味見する	味をみる
洗う	洗い流す
焼く	焼きます（口語）

順を単位とする．

料理映像においては，テキスト教材中の手順と映像中の手順の順序が一致するとは限らない．さらには，1つの手順が映像中では 2ヶ所以上の部分に分かれて出現したり，手順に対応する映像がない，または逆に映像がどの手順にも対応しない，といったことがしばしばある．このようなことをふまえ，対応付けは，抽出されたキーワードを元に，映像ブロックがどの手順と対応しているかを次の手順により判断した．この対応付けの様子を図 8 に示す．また，抽出されたキーワードそれぞれに，以下の式により得点を設定した．これは出現キーワードの珍しさを考慮したものである．

$$\frac{1}{M} \times \frac{1}{N} \quad \left(\begin{array}{l} M : \text{キーワードの出現手順数} \\ N : \text{キーワードの出現映像ブロック数} \end{array} \right)$$

（1） 映像ブロックを 1 つとり，その中に含まれる全キーワードを各手順に含まれるキーワードと照合する．

（2） 映像ブロック中の全キーワードの中で，素材名とそれに関連する動詞の両方が，ある手順と一致した場合は次のように行う．

一致する手順が 1 つのみの場合： 映像ブロックをその手順に対応付ける．

2 つ以上の手順と一致する場合： 映像ブロックは複数の手順に属する可能性があると判断する．映像ブロックが複数の手順に属する可能性があると判断されるのは，この場合のみである．

（3） （2）以外の場合は，キーワードの一致する手順に対してそのキーワードの得点を加えていき，映像ブロックがもつそれぞれの手順に対する得点のうち，最も得点の高い手順にその映像ブロックを対応付ける．得点の一番高い手順が 1 つに決定できない場合は，前後の映像ブロックが対応する手順のうち，得点の高い方の手順に対応付ける．それでも決定できない場合は，

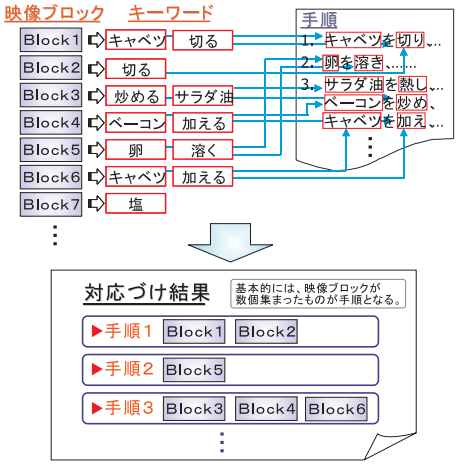


図 8　対応付けの様子
Fig. 8　Example of the association process.

表 3　映像キャプチャの際の条件
Table 3　Conditions of video capturing.

大きさ	横 320× 縦 240
色数	24bit
保存データ形式	Motion JPEG（圧縮率 約 1/3）
標本化レート	10 フレーム/秒

前の映像ブロックに対応付けられたものと同手順に対応付ける．

（4） 映像ブロックが，複数の手順に属する可能性があると判断された場合も，前後の映像ブロックに対応付けられた手順を参照し，前後の手順もしくは前後の手順と連続する手順のみに対応付けが許されるものとする．つまり結果として 1 つの手順のみに属することもある．

（5） 対応する手順が決定可能な映像ブロックから順に，対応付けを行う．

6. 実　　験

6.1 画像処理実験

ここでは，対応付けのための前処理である映像の構造解析手法の中から，画像処理部分のカット検出，及び人物ショット検出の実験を行った．

6.1.1 実 験 条 件

映像キャプチャの際の条件を表 3 に示す．

実際の処理に際しては，各フレームを非圧縮の PPM 形式に変換して利用した．また，この前処理実験には約 100 分間の特定の料理番組の映像（計 600 ショット）を用いた．

表 4　カット検出結果
Table 4　Result of cut detection.

N_C	N_M	N_O	再現率	適合率
568	10	31	94.8%	98.3%

表 5　人物ショットの検出結果
Table 5　Result of human shot detection.

ショットの種類	N_C	N_M	N_O	再現率	適合率
人物ショット（a1）	169	24	25	87%	88%
人物ショット（a2）	68	18	20	77%	79%

表 6　顔領域の検出結果（表中の数字は顔の数を示す）
Table 6　Result of face region detection.

ショットの種類	N_C	N_M	N_O	再現率	適合率
人物ショット（a1）	480	36	162	72%	92%
人物ショット（a2）	75	22	27	74%	77%

6.1.2　カット検出

DCT クラスタリングを基にしたカット検出手法を適用した結果を表 4 に示す．正答数を N_C，誤検出の数を N_M，検出もれの数を N_O とすると，再現率は $N_C/(N_C+N_O)$，適合率は $N_C/(N_C+N_M)$ である．

表 4 より，料理映像のカット検出が高精度で行えることがわかる．検出もれの大部分は，前後のショットがオーバーラップしながら切り替わるカット（ディゾルブ）であった．

このようなカットは，様々な手法を用いても検出が困難であることが知られている．しかし本研究で対象とする料理映像においては，このようなカットの前後のショットは大部分が同じ対象を映しており，このカット点でシーンが変化することはほぼないため，このようなカットの検出もれは致命的ではないと考えられる．

6.1.3　人物ショットの検出

人物ショットの検出結果を表 5 に示す．なお，カット検出は正しく行われたものと仮定し，600 ショットのそれぞれ最初のフレームに対し，顔領域を抽出することにより人物ショットを検出した．また，顔領域の抽出結果を表 6 に示す．

誤検出の主な原因は，肌色領域によって顔領域を抽出しているために，壁や鶏肉などの色が似ている部分を検出してしまったことである．また検出もれの主な原因は，顔の向きにより肌色領域が小さくなってしまったこと，顔に似た色の壁と一体となった領域として検出されてしまったことであった．

顔領域の検出もれの多くは，表 6 からもわかるように，人物ショット（a1）に見られるが，このショットには多くの場合複数の人物が存在するため，そのうちのいずれかを検出することができ，表 5 のようにショット分類の精度にはあまり影響がみられなかった．

表 7　対応付け結果（対応付け手法単独の評価）
Table 7　Result of association (Evaluation of the association method individually).

評価対象	全体数	正解	誤り	その他	成功率
映 → 手	242	203.5	29.5	9	84%
映 ← 手	94	74	20	–	79%
映 ↔ 手	94	59	35	–	62%

6.2　映像と手順の対応付け

最後に，映像と手順の対応付け実験を行った．実験には 20 レシピ分の特定の料理番組の映像（約 150 分）とそれに対応するテキスト教材を用いた．

まず，画像処理部分のカット検出及び人物ショット検出は理想的に行われたと仮定し，対応付け手法単独での性能を評価した．その結果を表 7 に示す．

表 7 における「映 → 手」は，映像ブロックを基準として対応付けを評価したもので，映像ブロックに正しい手順が対応付けられたものを正解としている．また，「映 ← 手」は，手順を基準として対応付けを評価したものであり，手順に対して映像ブロックが不足なく対応付けられたものを正解としている．最後に，「映 ↔ 手」は映像ブロック，手順が双方過不足なく対応付けられたものを正解としている．

映像ブロックは，n 個の手順に属するものもあるので，その時は 1 つの手順に対して正解を $1/n$ として計算した．また，表 7 における「その他」の映像ブロックは，料理全体の説明をしている映像ブロックなど，テキスト教材中の手順のどこにも対応しないものである．本実験では，映像ブロック全てがどの手順にも対応しない場合以外は，そのような映像ブロックも必ずいずれかの手順に対応付けてしまうため，誤りとして扱った．なお，成功率は，（正解数）/（全体数）とした．

このとき，各映像ブロックから抽出された平均キーワード数は，素材名 2.3 個，動詞 5.1 個であり，そのうち対応付けに利用されたものの数は素材名 2.2 個，動詞 2.0 個であった．一方，各手順から抽出された平均キーワード数は，素材名 3.8 個，動詞 5.6 個であり，そのうち対応付けに利用されたものの数は素材名 3.3 個，動詞 3.5 個であった．

本提案手法はテキストの順序にかかわらずに対応付けることができるため，映像の流れとテキスト教材中の手順の順序が複雑に入れ替わっているものや，1 つ

表 8　対応付け結果（システムの総合評価）
Table 8　Result of association (Evaluation of the whole system).

評価対象	全体数	正解	誤り	その他	成功率
映 → 手	222	179.2	31.8	11	81%
映 ← 手	94	62	32	–	66%
映 ↔ 手	94	49	55	–	41%

の手順に対応する映像が 2ヶ所以上に分かれて出現するものなども，多くの場合，正しく対応付けられた．

次に，画像解析部分のカット検出及び人物ショット検出についても **3. 1**で述べた手法により検出し，システムの総合的な性能を評価した．その結果を表 8 に示す．

カット検出及び人物ショット検出における誤りを含むため，表 7 と比べて「映 → 手」における全体数（映像ブロックの数）が異なり，主に映像ブロックの誤検出に伴って，成功率も低下している．

なお，これら全体を通した実験においても，処理時間については，カット検出処理において映像の長さの数倍程度かかるのみであり，その他についても，特に計算量が問題となる処理はなかった．

6. 3　考　　察

まず，対応付け手法単独の結果（表 7）と総合的な結果（表 8）とを比較すると，後者は前者よりも成功率が下がっているが，これは，カット検出とショット分類の誤りを含んでいるからである．このうち，カットは表 4 の実験結果からも分かるように高精度で検出できているため，成功率低下の主な原因は，人物ショット検出であった．

しかし，映像ブロックを基準とした対応付け（「映 → 手」）はわずか 3%の低下にとどまっており，対応付け手法によって人物ショット検出の誤りをある程度補うことができている．つまり，**3. 1. 2**で提案した人物ショット検出手法は，映像ブロックを基準とした対応付けのためにはほぼ十分なの性能でもあったと言える．一方，手順を基準とした対応付け（「映 ← 手」）では 13%低下，さらに双方とも過不足ない対応付け（「映 ↔ 手」）は 21%も低下しており，人物ショット検出手法のさらなる改善が必要である．

次に，対応付け手法単独の失敗の原因を分析すると，手順への対応付けを誤った映像ブロックのうち，3 割程度はそのブロック中に有力なキーワードがなく，前後の誤った映像ブロックを参照してしまったために生じたものであった．他の原因としては，以下のようなものが挙げられるが，全体的にみると，結果的にはキーワード不足が原因となり，誤りが生じている．したがって，素材名・動詞の他に調理器具名なども辞書を作成してキーワードとするなど，キーワードを増やすことが成功率向上のために必要であると考えられる．

- 同じ文に属する素材名と動詞を関連するものとみなしたことにより，本来は関係のない素材名と動詞を誤って関連付けてしまった．
- 指示語や「野菜」「材料」等の抽象的な語の指す具体的な内容を解析していないため，素材名が拾えなかった．

これらの改善策として，以下のようなものが考えられる．

- テキストの構文解析を行い，動詞と名詞の係り受け関係をみたうえで、関連付ける．
- 指示語や抽象的な語の具体的な指示内容を明らかにする．

最後に，提案手法の性能について評価する．提案手法はそもそも映像への意味的な索引付けを目指した，映像ブロックを基準とした対応付け手法である．表 8「映 → 手」の結果より，8 割以上の成功率で対応付けができており，そのような索引付けに成功していることが分かる．

このような索引付けが実現すると，映像を閲覧する際に，対応するテキスト手順を表示させたり，例えば，同じ手順に対応する映像ブロックの中から取捨選択し，要約映像を作成するなどの応用が考えられ，そのような応用のためには取捨選択の過程でより良いブロックを残すなどの工夫ができるため，本手法の性能は実用的であると考える．

一方，手順を基準とした対応付け結果は，カット検出およびショット分類処理を理想化した場合では 8 割程度であるが，総合的な評価では 7 割程度にとどまっている．

手順を基準とした対応付け手法の応用としては，調理支援システムとして，調理の進行状況に応じて手順に対応した映像を提示することや，手順に対応する映像を大量にデータベースに蓄積し，それらを組み合わせることによって映像の存在しない調理手順に対応する映像を生成することなどが考えられる．調理支援システムなどに利用する場合には，手順に対応する映像ブロックが不足なく対応付けられていれば十分であると考えられ，8 割程度という値は実用的な精度であると考えられる．

また，手順に対応する映像生成などの応用においては，さらに，映像ブロック，手順が双方過不足なく対応付けられている必要があるが，これはカット検出およびショット分類処理を理想化した場合でも 6 割程度，総合的な評価では 4 割程度となっているため，このような応用のためには，提案手法を改善する必要がある．

このような手順を基準とした対応付け，つまりテキストへの映像の付与という観点においては，提案手法はその第一段階と位置付けられ，最終的には手順をより細かく構造化し，個々の調理動作へ映像を対応付けることで，よりきめ細かな単位での処理が可能となる．このような対応付けのためには，テキスト教材中の手順の構造解析 [1]，及びさらに深いレベルでの映像の構造解析が必要である．

7. むすび

本論文では，料理映像と，それに付随するテキスト教材中の手順の対応付け手法を提案した．まず，対応付けのための映像の構造解析を検討し，映像と手順を対応付ける手法を提案・実装した．そして評価実験により，提案手法を用いて，映像ブロックを基準とした手順への対応付け，すなわち映像への意味的な索引付けが高精度で行えることを示した．また，対象を限定することで，比較的簡単な要素技術をうまく組み合わせることにより実用的な精度が得られることを示した．

提案手法は，テキストの順序にこだわらないため，映像の流れとテキスト手順の順序が複雑に入れ替わっているものや 1 つの手順に対応する映像が 2ヶ所以上に分かれて出現するものなどにも適用可能である．

本論文では対象を料理映像に限定しているが，提案手法は，(1) キーワードを集めた辞書を作成することで，例えば，組み立て作業などの手順を紹介するマニュアル映像や外部テキストの存在する教育番組などにも応用できるものである．また，これらの映像の構造が料理映像と異なる際にも，(2) 人物ショット検出に替わる映像構造解析手法を導入することによって適用が可能である．

今後の課題としては，対応付けの精度をさらに向上させるために，**6. 3**で述べた改善を行うこと，また，映像の構造解析の個々の手法の精度を向上させることなどが挙げられる．

さらには，映像への索引付けにとどまらず，テキストへの映像の付与などにも対応した，より精密な映像の構造解析，対応付け手法を検討する必要がある．これらが実現すると，より細かい単位での映像への意味的索引付けの実現はもちろん，さらには，対応付け結果を利用した映像要約や，また，テキストと映像をリンクさせた形の利用しやすい新たなマルチメディアデータの生成など，様々な応用が考えられる．今後の家庭内への計算機の進出に伴い，要約映像はレシピ選びの際の簡潔なレシピ閲覧に，構造化された映像は実際の調理時に的確に指示するのに有効な教材として用いられるなど，これらの研究は将来の知的調理支援につながるものになりうると期待される．

謝辞　本論文に掲載した画面画像は，本会パターン認識・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．

文　献

[1] R. Hamada, I. Ide, S. Sakai, and H. Tanaka, “Associating cooking video with related textbook,” Proc. ACM Multimedia 2000 Workshops, pp.237–241, Nov. 2000.

[2] 渡辺靖彦, 岡田至弘, 角田達彦, 長尾　真, “TV ニュースと新聞記事の対応づけ,” 人工知能学誌, vol.12, no.6, pp.921–927, Nov. 1997.

[3] 柳沼良知, 坂内正夫, “DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応付け手法の一提案,” 信学論 (D-II), vol.J79-D-II, no.5, pp.747–755, May 1996.

[4] M. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding,” Proc. Intl. Conf. on Computer Vision, pp.61–70, Jan. 1998.

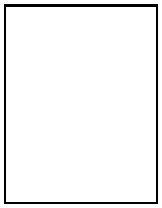
[5] 岩成英一, 有木康雄, “DCT 成分を用いたシーンのクラスタリングとカット検出,” 信学技報, PRU93-119, pp.23–30, Jan. 1994.

[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,” テレビ誌, vol.49, no.6, pp.787–797, June 1995.

[7] 井手一郎, 山本晃司, 浜田玲子, 田中英彦, “ショット分類に基づく映像への自動的索引付け手法,” 信学論 (D-II), vol.J82-D-II, no.10, pp.1543–1551, Oct. 1999.

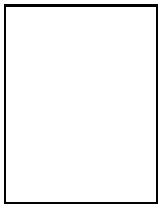
[8] 京都大学大学院情報学研究科知能情報学専攻言語メディア研究室, “日本語形態素解析システム JUMAN 第 3.6 版,” Nov. 1998.

（平成年月日受付，月日再受付）



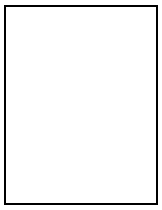
三浦　宏一

平 13 東大・工・電子情報卒．平 15 同大大学院情報理工学系研究科電子情報学専攻修士課程修了．修士（情報理工学）．映像解析，映像要約に関する研究に従事している．



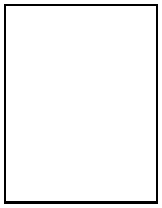
高野　　求

平 15 東大・工・電子情報卒．現在同大大学院情報理工学系研究科電子情報学専攻修士課程在学中．映像解析に興味を持っている．



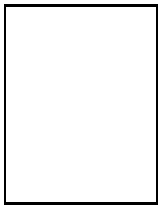
浜田　玲子　（正員）

平 10 東大・工・電子情報卒．平 12 同大大学院工学系研究科電気工学専攻修士課程了．平 15 同専攻博士課程修了．博士（工学）．現在同大大学院情報理工学系研究科リサーチフェロー．自然言語処理，マルチメディア統合処理に興味を持っている．平 14 第 63 回情報処理学会全国大会奨励賞受賞．情報処理学会会員．



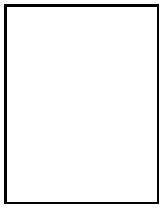
井手　一郎　（正員）

平 6 東大・工・電子卒．平 8 同大大学院工学系研究科情報工学専攻修士課程了．平 12 同研究科電気工学専攻博士課程了．博士（工学）．同年国立情報学研究所助手．平 14 より総合研究大学院大学数物科学研究科助手併任．自然言語処理，統合メディア処理に興味を持っている．平 7 第 51 回情報処理学会全国大会奨励賞受賞．人工知能学会，情報処理学会，IEEE Computer Society，ACM 各会員．



坂井　修一　（正員）

昭 56 東大・理・情報科学卒．昭 61 同大学院工学系研究科情報工学専門課程了．工学博士．同年工業技術院電子技術総合研究所入所．この間平 3〜4，米国マサチューセッツ工科大学招聘研究員，平 5〜8 RWC 超並列アーキテクチャ研究室室長．平 8〜10 筑波大学電子・情報工学系助教授．平 10 東京大学大学院工学系研究科助教授，平 13 より同大大学院情報理工学系研究科教授．計算機システム一般，特にアーキテクチャ，並列処理，スケジューリング問題，マルチメディアなどの研究に従事．平 2 情報処理学会論文賞，平 3 日本 IBM 科学賞，平 7 市村学術賞，平 7 ICCD Outstanding Paper Award など受賞．情報処理学会，人工知能学会，IEEE，ACM 各会員．



田中　英彦　（正員）

昭 40 東大・工・電子工学卒．昭 45 同大学院工学系研究科博士課程了．工学博士．同年同大学工学部講師．昭 46 助教授．昭 62 同教授，平 13 より同大大学院情報理工学系研究科教授，研究科長．この間昭 53〜54 ニューヨーク市立大学客員教授．計算機アーキテクチャ，並列処理，自然言語処理，メディア処理，分散処理，CAD 等の研究に興味を持っている．著書「非ノイマンコンピュータ」,「情報通信システム」，共著書「計算機アーキテクチャ」,「VLSI コンピュータ I,II」,「ソフトウェア指向アーキテクチャ」．情報処理学会，人工知能学会，日本ソフトウェア科学会，IEEE，ACM 各会員．