マルチメディア料理レシピ作成のための料理レシピテキストと料理番組映像との対応付け

道満　恵介[†a)]（学生員）　　カイ　承穎[†*]
高橋　友和[††]（正員）
井手　一郎[†b)]（正員：シニア会員）
村瀬　洋[†]（正員：フェロー）

Association of Cooking Recipe Texts and Cook Shows for the Generation of Multimedia Cooking Recipes

Keisuke DOMAN[†a)], *Student Member*,
Cheng-Ying KUAI[†*], *Nonmember*,
Tomokazu TAKAHASHI[††], *Member*,
Ichiro IDE[†b)], *Senior Member*, and Hiroshi MURASE[†], *Fellow*

[†] 名古屋大学大学院情報科学研究科，名古屋市
Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan
[††] 岐阜聖徳学園大学経済情報学部，岐阜市
Faculty of Economics and Information, Gifu Shotoku Gakuen University, Gifu-shi, 500–8288 Japan
[*] 現在，ブラザー工業株式会社
a) E-mail: kdoman@murase.m.is.nagoya-u.ac.jp
b) E-mail: ide@is.nagoya-u.ac.jp

あらまし　料理レシピテキストと料理番組映像との対応付け手法を提案する．提案手法により，調理動作映像データベースを自動構築し，テキスト主体の任意の料理レシピをマルチメディア化できることが期待される．

キーワード　料理支援，料理レシピ，調理動作映像データベース，マルチメディア化，動作解析

1. まえがき

現在，膨大な量の料理レシピが Web 上で公開されている．しかし，それらの多くはテキスト主体であり，レシピ中の料理用語（特に調理動作）を正確に理解することは難しい．これに対して，図 1 に示すように各調理動作に対する解説映像を補足して料理レシピをマルチメディア化できれば，従来のテキスト主体の料理レシピでは困難な “調理動作の視覚的な理解” が可能となる．ただし，料理レシピ中に出現する素材や調理動作は様々であり，それら全てに対応する大量の解説映像を人手で用意するのは多大なコストがかかる．

そこで我々は，料理レシピテキストと料理番組映像との対応付け手法，マルチメディア料理レシピ及びそれを閲覧するためのインタフェースに関する検討を行ってきた [1]〜[3]．これらの検討をもとに本論文では，マルチメディア料理レシピ作成のための料理レシピテキストと料理番組映像との対応付け手法を提案する．

関連研究として，浜田らは “Cooking Navi” システムを提案している [4]．このシステムも，料理レシピテキストと料理番組映像との対応付けを行うが，一連の調理動作をまとめた手順単位での対応付けしかできない．また，柴田らは，隠れマルコフモデル（HMM）を用いて料理番組映像中のショットを解析・分類する手法を提案している [5]．しかし，この手法では，画像情報として背景の平均色のみを利用しており，調理動作の画像特徴に直接注目していない．これらの手法では，いずれも調理動作単位での細かな対応付けまでは考慮しておらず，個々の調理動作に対する解説映像の抽出は困難である．一方，本研究では，調理動作の画像特徴を解析し，料理レシピテキストと料理番組映像とを調理動作単位で対応づけることを考える．これを様々な料理番組映像に適用することで，様々な素材，調理動作に対応する解説映像からなる調理動作映像データ

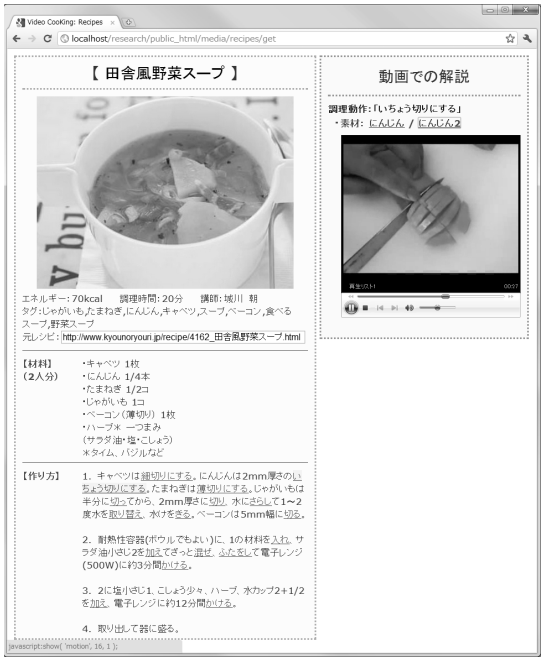


図 1　マルチメディア料理レシピの例（左側：テキスト主体の料理レシピ(注1)，右側：マルチメディア化により補足された調理動作の解説映像）．

Fig. 1　Example of a multimedia cooking recipe (left side: the original cooking recipe(注1), right side: supplementary video clips corresponding to each cooking operation).

(注1)：株式会社 NHK エデュケーショナル：“みんなのきょうの料理,” http://www.kyounoryouri.jp/

ベースを自動構築する．これにより，Web 上の料理レシピ等，直接対応する料理番組映像が存在しない任意の料理レシピに対してもマルチメディア化が可能になる．

2. 提案手法：料理レシピテキストと料理番組映像との対応付け

提案手法における処理の流れを図 2 に示す．一般に，料理番組は，料理レシピテキストに対応して各調理動作を解説する映像，番組内の発話をテキストに書き下したクローズドキャプション (CC) からなる．提案手法では，これらをもとに料理レシピテキストと料理番組映像とを対応づける．以降，各処理について詳述する．

2.1 テキスト処理部：タグ候補の抽出

テキスト処理部では，料理レシピテキストと CC との対応付けを行い，調理動作の解説映像に付与するタグ候補を抽出する．なお，同一の調理動作であっても調理の対象素材により細かな動作が異なることを考慮し，(素材，調理動作）という組でタグを表現する．

2.1.1 料理レシピテキストの解析

以下の処理により，料理レシピテキストからタグを抽出する．まず,「材料」及び「作り方」に関する記述に対して形態素解析を施す．その後，日本語の文法を考慮して，次の規則で「作り方」に出現する素材（名詞）と調理動作（動詞）の組を形成する．

（1） 動詞（連用形）＋助動詞「た」: 直後の動詞までに出現する各名詞と動詞の組を形成する．

（2） 上記以外：直前の動詞までに出現する各名詞と動詞の組を形成する．

これにより，例えば (1)「細切りにしたキャベツ」，(2)「キャベツを細切りにする」に対してはいずれも（キャベツ，細切りにする）という組が形成される．ここで,「名詞（＋ {を/に}）＋する」のような複合表現については，全体で一つのサ変動詞として扱う．手順番号の参照がある場合は，参照先に出現する全名詞を再帰的に抽出し，各名詞と動詞の組を形成する．なお,「材料」に出現しない名詞を含む組は処理対象外とする．

2.1.2 CC の解析

以下の処理により，CC からタグを抽出する．まず，料理レシピに出現した素材と調理動作のみを CC から抽出する．その後，次の規則で，素材（名詞）と調理動作（動詞）の組を作成する．

（1） 動詞（連用形）＋助動詞「た」: 組を形成しない．

（2） 上記以外：直前の動詞までに出現する各名詞と動詞の組を形成する．

(1) は，映像内で現在進行中の調理動作に対応するタグのみを抽出することを意図している．そのため，例えば「細切りにしたキャベツ」のような過去形の表現は，CC の解析処理においては抽出対象としない．

2.1.3 料理レシピテキストと CC の照合

料理レシピテキスト及び CC から抽出されたタグを照合し，タグが一致する箇所を抽出する．このとき，有対動詞（例：「揚げる」⇔「揚がる」）や，同じ動作を意味する動詞（例：「薄切りにする」⇔「薄切りする」）については，シソーラスを作成し利用することで，同じ調理動作とみなす．その後，抽出された各タグを調理動作の解説映像に付与するためのタグ候補とし，CC の発話時刻を付与する．

2.2 画像処理部：調理動作映像の抽出・分類

図 3 に示すように，一般に，料理番組は「人物ショット」と「手元ショット」が交互に出現する．本研究では，料理の状態や調理動作が大きく撮影される「手元ショット」に注目し，この中の映像区間を以下のようなシーンに分類することを考える．

- 繰返し動作：ある動作を複数回繰り返すシーン
  - 集中型：フレーム中の特定範囲で周期的な画素値の変化が観測されるシーン（例：切る）
  - 分散型：フレーム中の広範囲で周期的な画素値の変化が観測されるシーン（例：炒める，混ぜる）

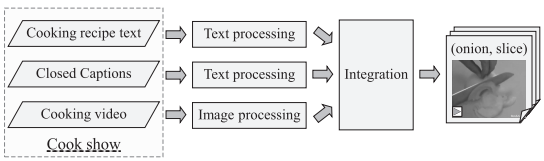


図 2 提案手法における処理の流れ
Fig. 2 Process flow of the proposed method.

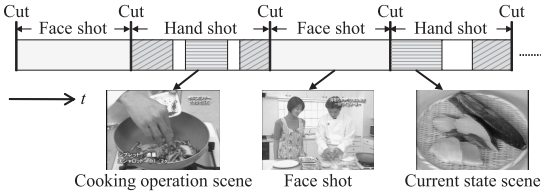
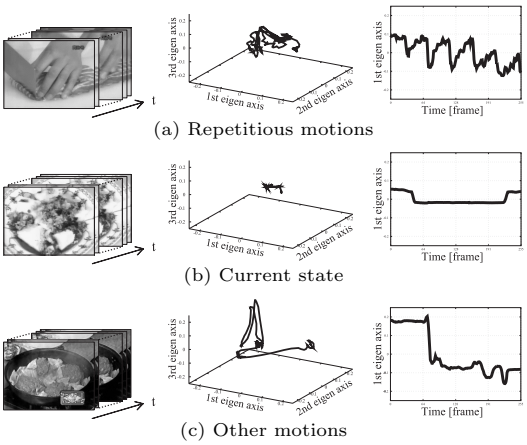


図 3 一般的な料理番組の構造
Fig. 3 General structure of a cook show.



(a) Repetitious motions

(b) Current state

(c) Other motions

図 4　動作解析結果の例（左列：入力映像区間，中列：固有空間上の軌跡，右列：第 1 固有軸上の軌跡）

Fig. 4　Example of the motion analysis results (left: an input video clip, center: its trajectory on the eigenspace, right: the trajectory projected onto the first eigenaxis).

- 状態提示：食材の状態を提示する，または大きな動きを含まないシーン（例：煮る，茹でる,「食材」）
- その他の動作：上記以外（例：盛る，調味する）

なお，本処理は，テキスト処理部（**2. 1**）で抽出されたタグ候補が一つ以上存在する手元ショットに対してのみ適用する．以降，シーン分類処理について詳述する．

2. 2. 1　手元ショットのシーン分類

料理番組映像は，照明条件が安定して良好なスタジオで撮影され，複雑な映像効果を伴うカット変化は少ない．三浦らは，DCT クラスタリングを利用したカット検出，肌色の統計情報を利用した人物の顔領域抽出に基づくショット分類により，料理番組映像からの高精度な手元ショット抽出を実現している [6]．本研究でも，この手法を手元ショット抽出に適用できると考えている．手元ショットを抽出した後，ある窓幅の連続する映像区間内のフレームから固有空間法を作成し，各フレームを固有空間上に投影する．図 4 に示すように，固有空間上に描かれる軌跡はシーンの種類ごとに特徴的である．そこで，軌跡の特徴を最もよく表現する第 1 固有軸上の軌跡をシーン分類のための動作特徴として利用する．具体的には，以下の条件に従い，各映像区間を 3 種類の動作カテゴリーに分類する．

$$\begin{cases} \text{繰返し動作} & \text{if } m \geq \theta_m \\ \text{状態提示} & \text{if } m < \theta_m \text{ and } \Delta_r \leq \theta_{\Delta_r} \\ \text{その他の動作} & \text{otherwise} \end{cases} \tag{1}$$

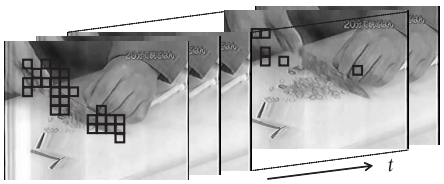
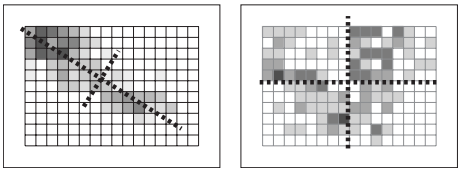


図 5　周波数解析結果の例（矩形領域：繰返し領域）

Fig. 5　Example of the frequency analyses (rectangle region: repetitious region).



(a) Converged　(b) Distributed

図 6　フレーム内の繰返し回数の分布に対する PCA の結果例（点線：固有軸，色の濃さ：繰返しの多さ）

Fig. 6　Example of the results of PCA applied to the distribution of repetition counts (dashed lines: eigenaxes, darker color: larger repetition count).

ここで，m は軌跡中のピークの数，Δr は軌跡中の最小値と最大値の差，θ_m 及び $\theta_{\Delta r}$ は m 及び Δr に対するしきい値である．最終的に，映像区間をずらしながら上記処理を繰り返し，同じ動作カテゴリーに分類された一連の映像区間をまとめて一つのシーンとみなす．

2. 2. 2　繰返し動作シーンの分類

「繰返し動作」シーンのみを「集中型」と「分散型」に細分する．まず，シーン内の各フレームをブロックに分割する．次に，図 5 に示すように，平均輝度値変化の繰返し回数をブロックごとに計数する [7]．ここで，画像的に不安定なフレームの上下左右の端は，処理対象から除外する．その後，繰返し回数の累積分布に対して主成分分析 (PCA) を施す．このとき，図 6 に示すように,「集中型」と「分散型」の間には各軸周りの分散に明確な差異が確認できる．この点に着目し，入力シーンを次の条件に従い細分する．

$$\begin{cases} \text{集中型} & \text{if } \lambda_1/\lambda_2 \geq \theta_\lambda \\ \text{分散型} & \text{otherwise} \end{cases} \tag{2}$$

ここで，λ_1，λ_2 はそれぞれ第 1，第 2 固有値，θ_λ は各軸周りの分散に対するしきい値である．

2. 3　統合処理部：調理動作映像のタグ付け

テキスト処理部（**2. 1**）で抽出されたタグ候補と CC の時刻情報に基づいて，画像処理部（**2. 2**）で分類さ

れたシーンに対するタグ付けを行う．ここで，一つのシーンに対して同じ動作カテゴリーのタグ候補が複数存在する場合には，それらを全て対応づける．なお，調理動作が属する動作カテゴリーは既知とする．

以上の処理を様々な料理番組に対して適用することで，(素材，調理動作）のタグが付いた調理動作の解説映像を収集し，調理映像データベースを自動構築する．

2.4 評価実験

手元ショットを複数の動作カテゴリーに分類して対応付けを行う提案手法の有効性を以下のように評価した．本実験では，動作カテゴリーの情報を利用せず，手元ショットに対するタグ候補をそのまま各シーンに対応づける手法を比較手法とした．

2.4.1 実験方法

評価用素材として，8 本の料理レシピとそれに対応する料理番組映像（320 × 240 pixel，30 fps，計 75 分）を用いた．料理レシピテキストと CC の形態素解析には MeCab(注2)を利用した．料理レシピテキストと CC の照合では，11 種類の動詞からなるシソーラスを作成し利用した．料理番組映像に対するカット検出及び手元ショット抽出，手元ショット中の各シーンに対する正解ラベルの付与は人手で行った．「繰返し動作」シーンの分類におけるブロックサイズ，窓幅はそれぞれ経験的に 16 × 16 pixel，256 フレーム（約 8 秒）とした．以上により，104 の手元ショットを得た．更に，タグ候補が存在する 67 ショットに対するシーン分類処理（**2.2**）の結果，129 シーンが得られた．これらに対する各手法による対応付け精度を調べた．

2.4.2 実験結果

比較手法では，188 組の対応付けが得られ，38.3%が素材，調理動作ともに正しく，50.0%が調理動作のみ正しかった．提案手法では，135 組（素材，調理動作それぞれ 33 種類）の対応付けが得られ，52.6%が素材，調理動作ともに正しく，68.1%が調理動作のみ正しかった．これにより，提案手法の有効性を確認した．

2.4.3 考察

提案手法による主な誤対応は，シーン分類の失敗に起因していた．予備実験におけるシーン分類率は 78.1%であり，カメラワークを伴うシーンの誤分類が多かった．そのため，カメラワークの検出・除去により，対応付け精度の向上が期待できる．

提案手法の各処理部には改良の余地はあるものの，マルチメディア料理レシピにおいて，タグづけられた複数の解説映像を候補として提示することはできる．また，ユーザはそれらの候補を取捨選択して利用できる．取捨選択の負荷は，提示された各解説映像のサムネイルを表示する等，閲覧インタフェースの工夫で軽減できる．そのため，提案手法による対応付け精度は，料理支援のために十分実用的であると考える．

3. むすび

本論文では，料理レシピのマルチメディア化に利用する調理動作映像データベースの自動構築を目的とし，料理レシピテキストと料理番組映像との対応付け手法を提案した．評価実験により，料理支援のための十分な実用性を確認した．今後の課題として，提案手法における各処理部の性能向上，及びマルチメディア料理レシピを閲覧するための最適なインタフェースの検討が挙げられる．

謝辞 本研究の一部は，文部科学省科学研究費補助金 (21013022) による．本研究では，画像処理に MIST ライブラリ(注3)を利用した．

文献

[1] カイ承穎，高橋友和，井手一郎，村瀬 洋，“動作解析による料理レシピと料理番組映像の対応付け，” 第 5 回ディジタルコンテンツシンポジウム講演予稿集，no.1-4, pp.1–5, June 2009.

[2] 道満恵介，カイ承穎，高橋友和，井手一郎，村瀬 洋，“調理動作に注目した料理レシピのマルチメディア化の提案，” 情処学全大，vol.5, pp.189–190, March 2010.

[3] K. Doman, C.Y. Kuai, T. Takahashi, I. Ide, and H. Murase, “Video CooKing: Towards the synthesis of multimedia cooking recipes,” Advances in Multimedia Modeling, Lect. Notes Comput. Sci., vol.6524, pp.135–145, Springer, Jan. 2011.

[4] R. Hamada, J. Okabe, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, “Cooking Navi: Assistant for daily cooking in kitchen,” Proc. 13th ACM Int. Multimedia Conf., pp.371–374, Nov. 2005.

[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，” 情処学論，vol.48, no.6, pp.2129–2139, June 2007.

[6] 三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦，“動きに基づく料理映像の自動要約，” 情処学論，vol.44, no.SIG9, pp.21–29, July 2003.

[7] R. Hamada, S. Satoh, and S. Sakai, “Detection of important segments in cooking videos,” Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, pp.118–123, Dec. 2001.

（平成 22 年 9 月 23 日受付，23 年 2 月 2 日再受付）

(注2)：京都大学：“日本語形態素解析器 MeCab,” http://mecab.sourceforge.net/

(注3)：名古屋大学：“Media Integration Standard Toolkit: MIST,” http://mist.murase.m.is.nagoya-u.ac.jp/