動作解析による料理レシピと料理番組映像の対応付け

蒯 承穎[†]　高橋 友和[‡]　井手 一郎[†*]　村瀬 洋[†]

†名古屋大学大学院情報科学研究科　〒464-8601 愛知県名古屋市千種区不老町 1

‡岐阜聖徳学園大学経済情報学部　〒500-8288 岐阜県岐阜市中鶉 1-38

＊国立情報学研究所　〒101-8430 東京都千代区一ツ橋 2-1-2

E-mail:　†{kuai, ide, murase}@murase.m.is.nagoya-u.ac.jp

‡ttakahashi@gifu.shotoku.ac.jp

あらまし　我々は，調理者を支援する際に調理動作を視覚的に提示することを考え，調理動作映像データベースの自動構築を目指している．本稿では，料理レシピおよび対応する料理番組映像を対象とし，料理レシピ中の調理動作に対応する映像を抽出する手法を提案する．提案手法では，テキスト情報と画像情報を統合し，動作解析を行い，最終的に動作・素材名詞の組と料理映像区間を対応付ける．実験により，調理動作のみでは 70%近く，調理動作・素材の組では 50%以上の対応付け精度が得られた．これにより，調理動作毎に調理映像のショットと対応付けて収集することで，調理動作映像データベースが構築可能であることを確認した．

キーワード　料理レシピ，料理映像，動作解析，調理動作映像データベース

Associating Cooking Recipe with Cooking Video by Motion Analysis

KUAI Cheng Ying[†]　Tomokazu TAKAHASHI[‡]　Ichiro IDE[†*]　and　Hiroshi MURASE[†]

†Graduate School of Information Science，Nagoya University

1 Furo-cho，Chikusa-ku，Nagoya-shi，Aichi，464-8601 Japan

‡Faculty of Economics and Information，Gifu Shotoku Gakuen University

1-38 Nakauzura，Gifu-shi，Gifu，500-8288 Japan

＊National Institute of Informatics　2-1-2 Hitotsubashi，Chiyoda-ku，Tokyo，101-8430 Japan

E-mail:　†{kuai, ide, murase}@murase.m.is.nagoya-u.ac.jp

‡ttakahashi@gifu.shotoku.ac.jp

Abstract　We are considering providing people with visual explanations of cooking motions as a cooking support service, and therefore aiming at building a video database of cooking motions. In this paper, we propose a method for associating cooking recipe text with cooking video by integrating analysis of text and image information. An experiment was performed on cooking video broadcast in Japan with their recipes. The results demonstrated the proposed method can achieve a success rate of approximately 70% in associating motions to video segments, and the capability of building a cooking motion video database was confirmed.

Keyword　cooking recipe，cooking video，motion analysis，cooking motion video database

1. はじめに

料理は日常生活の一環として重要な役割を担うものである．そのため，料理本をはじめとして様々な形の料理支援方法が存在する．近年，インターネットの普及により，一般の人々によって調理された日々の料理レシピが Web 上で多く公開されている［1］．しかしながら，料理初心者が料理レシピのとおりに調理することは難しい．それは料理レシピがテキスト主体であること，レシピ中には料理特有の専門用語が多く現れ，それらに対して詳細な説明が少ないことが原因として挙げられる．一部の料理レシピには調理過程を示す写真が添えられているが，写真のような静止画では調理動作を表現するためには不十分である．一方，ユーザに対して動画や音声によって料理を支援するソフトウェアが開発されている［2］．特に，動画を用いることによって，より分かりやすい形で調理動作を提示することが可能である．しかし，これらの動画や音声は人手によって編集されるため，規模を大きくするには多

くのコストが必要である．そのため，料理レシピに料理映像を自動付与する技術が望まれ，本稿ではこのような調理動作映像データベースの構築を目的とし，料理映像における調理動作に注目した料理レシピと料理映像の対応付け手法を提案する．

料理レシピに料理映像を付与する研究として，テキスト・画像情報を統合した料理映像の対応付けに関する研究が提案されている［4］．この研究では，調理手順の制約条件を考慮することで，テキストブロックと映像シーンを対応付ける．また，画像情報を加え，料理番組に付随するクローズドキャプションの構造解析を行い,作業手順を自動的に構造化する研究もある[5]．これらに対して，本研究では，料理映像中の調理動作に注目し，複雑な手順構造を考慮せず，簡単かつ柔軟な条件設定により，料理映像と対応する料理レシピテキストおよびクローズドキャプションの解析・照合を行う．一方，映像中の各フレーム画像全体の動きを分類するために，フレーム画像を特徴空間上にプロットし，その軌跡を解析する．さらに，特に重要な繰り返し動作として検出された映像に対して，繰り返し動作出現分布の形状により分類を検討する．その後，素材・調理動作の組と映像をショット単位で対応付ける．

以下，2 節で対応付けに用いる料理映像と料理レシピの特徴について述べ，3 節で提案手法の詳細を述べる．4 節では提案手法を用いた対応付け実験の結果から提案手法の有効性を示す．最後に 5 節でむすぶ．

2. 料理映像と料理レシピの特徴

2.1. 料理映像の構成

一般的な料理映像の構成を図 1 に示す．料理映像は複数の人物ショットと手元ショットが交互に出現する．ショットとは，映像が大きく切り替わるフレーム（カット点）によって区切られる映像区間である．人物ショットは図 1 に示すように人物の全身が映っているショットである．これに対して，手元ショットは手元のみに注目したショットである．手元ショットには，料理の状態や調理動作が大きく映されているため，特に重要であると考えられる．そこで，本研究では手元ショットのみに着目する．

2.2. クローズドキャプション(CC)

一般に放送されている料理映像には，図 2（右）で示すような文字放送字幕（CC; Closed Caption）が付随する．これは主に聴覚障害者のために，主音声を書き下したディジタルテキストである．CC の各文には発話された時刻が付与されている．

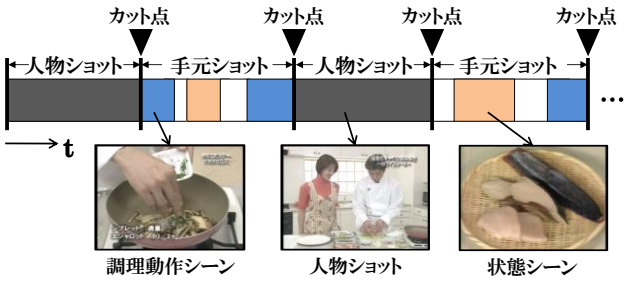


図 1: 料理映像の構成

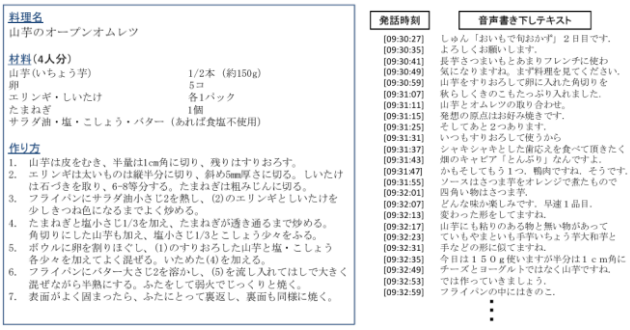


図 2: （左） 料理レシピテキスト，（右） CC

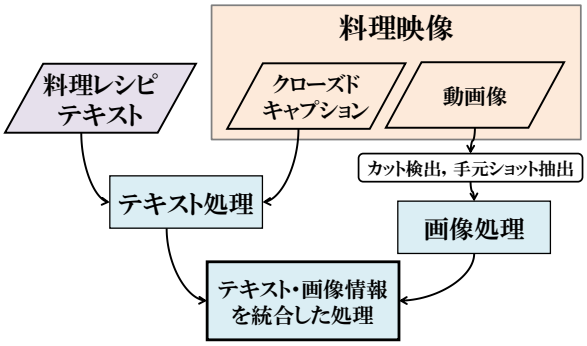


図 3: 提案手法による対応付けの処理手順

2.3. 料理レシピ

本研究では，図 2（左）で示すような料理番組に対応する料理レシピをもう 1 つのテキスト情報として利用する．料理レシピテキストは一般に「材料」と「作り方」の 2 つの項目からなる．「材料」は，各レシピで使用する素材の一覧であり，料理番組中に登場する素材はすべてこの一覧の中に含まれる．「作り方」は調理の手順を示す．

3. 料理レシピテキストと料理映像の対応付け

3.1. 提案手法の概要

提案手法による料理レシピと料理映像の対応付けの流れを図 3 に示す．提案手法は料理レシピおよび CC を解析するテキスト処理と，料理映像を分類する画像処理をそれぞれ行い，それらを統合することで調理動

表 1: 調理動作の分類

	分類		例
1-1	繰り返し動作	集中型	切る
1-2		分散型	炒める，混ぜる
2	状態提示		煮る，茹でる,「食材」
3	その他の動作		盛る，調味する

作の対応付けを行う．テキスト処理部では，料理映像に対応する料理レシピ，CC をそれぞれ解析し，料理の素材，および，調理動作をタグとして抽出する．また，画像処理部では，手元ショットの分類や繰り返し動作の分類により調理動作を抽出する．

本研究では料理映像に含まれる動作の特徴から対応付ける調理動作を以下の 3 種類に大別する．

1. **繰り返し動作:** 周期的な動きが含まれるもの
2. **状態提示:** 大きな動きを含まず，ほぼ静止状態が継続するもの
3. **その他の動作:** 1.，2.に当てはまらないもの

さらに 1.の「繰り返し動作」を映像全体の動きの分布によって，以下の 2 種類に分類する．

1-1. 集中型: 特定の領域のみが周期的に変化

1-2. 分散型: 映像全体が周期的に変化

このような調理動作の分類を具体的な例とともに表 1 に示す．

3.2. 料理レシピと CC の照合によるタグの抽出

はじめに，料理映像中の手元ショットに対応付けるタグを料理レシピと CC を照合することで作成する．まず，料理レシピにおける「材料」と「作り方」の各文に対して形態素解析を施す．以下の実験では形態素解析に MeCab[1]を用いた.そして,料理レシピ文中の「材料」に出現する名詞を素材として抽出する．なお，調味料については除外する．また,「作り方」に出現する動詞を調理動作として抽出する．ここで，調理動詞には,一般の文中では使われない特殊な表現が存在する．このような表現は,「サ変名詞＋（を/に）＋する」のようなサ変動詞や複合動詞に関するものである（例：「落としぶたをする」,「せん切りにする」など）．このような表現に関しては，事前に動詞辞書を構築することで対応する．

次に，素材と調理動作を関連付けた組を作成する．関連付けを行う際に料理レシピの「作り方」の部分を使用する．その際に，料理レシピの「材料」から抽出された素材を用いて,「作り方」に出現する名詞を照合する．一致していれば，その名詞を素材と判断する．「作り方」には手順番号が付与されているので，1 つの手順を 1 文として扱う．そして，以下の条件を用いて料理レシピから素材と調理動作を関連付ける．

1. ある 1 文中に出現する動詞の前にある全ての素材は，その動詞に係る．
2. ある 1 文中に出現する動詞が「連用形＋た（助動詞）」の場合，直後の素材に係る．
3. ある 1 文中に動詞が出現して素材が出現しない場合，その前文に出現する全ての素材がその動詞に係る．
4. ある 1 文中に手順番号が使われる場合，その 1 文中に出現する調理動作に係る素材とともに手順番号の手順に出現する素材も対応付ける．

料理レシピと同様に，料理映像に対応する CC に対して，素材・調理動作の抽出と関連付けを行う．ここで，料理レシピに出現した素材と調理動作のみを CC から抽出し，それらの関連付けを以下のような条件で行う．

1. ある調理動作にはその前の調理動作との間に出現するすべての素材と対応付ける．
2. 間に素材がない場合，前の動作に係るすべての素材と対応付ける．
3. ある 1 文中に出現する動詞が「連用形＋た（助動詞）」の場合，関連付け対象から除外する．

最後に，CC・料理レシピのそれぞれから抽出された素材・調理動作からどちらにも表れるものを調理動作として抽出する．このとき，CC・料理レシピ特有の表現を考慮し，以下の条件を用いる．

1. 有対動詞の統一
   料理レシピと CC では，同じ動作に対して「他動詞」と「自動詞」が混在して使われることがある（例：「揚げる⇔揚がる」）．この場合，同じ調理動作を表すものが異なる動作と認識されてしまうため，有対動詞を統一するための辞書を利用して，照合を行う．
2. 動作表現の不一致
   CC には口語的な表現が多く含まれるため，助詞が省略される場合がある．一方，料理レシピは比較的文語的に書かれているため，正しく照合できないことがある（例：「薄切りにする⇔薄切りする」）．この場合，「する」と「(を/に)＋する」の前の部分のみを利用して照合を行う．

以上で得られる調理動作を料理映像に対するタグとする．また，このタグに対して CC の発話時刻を付与する．なお，ここで抽出された調理動作は表 1 のいずれの動作の種類であるかは既知であるとする．

[1] http://mecab.sourceforge.net/, “日本語形態素解析システム和布蕪”.

3.3. 動作解析による調理動作の分類

3.3.1. 手元ショットの分類

本節では料理番組映像から調理動作を抽出する．まず，各料理映像をカット検出により，ショット単位に分割し，手元ショットを抽出する．カット検出，および，手元ショットの抽出は既存手法［3］を利用する．これ以降の処理はショット単位で行う．

次に，手元ショットを映像内に含まれる動作を解析することで，表 1 に示す 3 種類の調理動作に分類する［6］．これらを分類するため，はじめにショット内の一定長の連続するフレーム画像を用いて固有空間を作成する．そして，各フレーム画像をこの固有空間上に射影する．このとき，各画像は図 4 中で示すように固有空間上で軌跡を描く．この軌跡をショットが含む動作特徴として扱う．図 4 中から各動作の軌跡には明らかな違いが見られる．本稿では，最も大きな特徴を持つ第一固有軸上の軌跡（図 4 右）のみを用いて手元ショットの分類を行う．分類には，軌跡に含まれるピークの数 m，軌跡の最大値と最小値の差 Δr を用いる．ここで，以下の条件を満たす箇所をピークとする．

$$\begin{cases} g(t)-g(t\pm1)\geq\theta_1, \\ g(t+1)-g(t+2)\geq\theta_2, \\ g(t-1)-g(t-2)\geq\theta_2. \end{cases}$$

ここで $g(t)$は時刻 t における第 1 固有軸上の値，θ_1 θ_2はそれぞれピークの強さに対するしきい値を表す．そして，以下の条件を用いて手元ショットを分類する．

$$\begin{cases} \text{繰返し動作} & (m\geq\theta_m) \\ \text{状態提示} & (m<\theta_m \ \ and \ \ \Delta r\leq\theta_{\Delta r}) \\ \text{その他の動作} & (m<\theta_m \ \ and \ \ \Delta r>\theta_{\Delta r}) \end{cases}$$

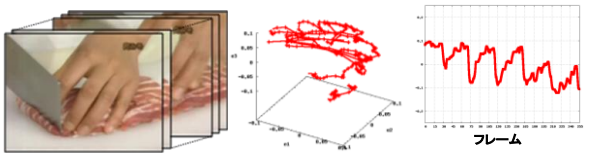
ここで，θ_m はピークの数 m に対するしきい値，$\theta_{\Delta r}$ は軌跡の最大値と最小値の差 Δr に対するしきい値とする．

3.3.2. 「繰り返し動作」の分類

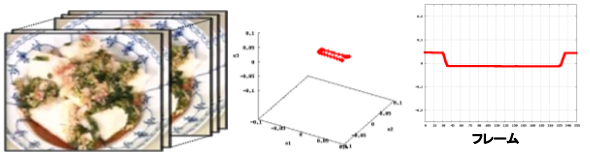
以上で分類された手元ショットのうち，「繰り返し動作」をさらに詳細に分類する［7］．映像中の連続するフレームに対して，フレーム画像中の各局所領域の輝度値の時間変化を周波数解析する．ある周波数で明確なピークが存在する局所領域の位置に関して繰り返し局所領域の出現回数を積算する．図 5 に示すように，映像中の複数の繰り返し局所領域と判定された回数が少ないほど薄く，多いほど濃く表現されている．次に，累積された繰り返し局所領域の分布によって，繰り返し動作を「集中型」（図 6a）と「分散型」（図 6b）に分類する．局所領域をサンプル点として主成分分析を行うことで，2 つの固有値 λ_1，λ_2 を得る．各固有値に対応する固有ベクトルは図 6 における破線の軸と一致する．「集中型」は図 6a に示すようにそれぞれの軸まわりの分散が異なり，反対に「分散型」は図 6b に示すように 2 つの軸まわりの分散が同程度となる．そのため，以下の式によって，繰返し動作を分類する．

$$\begin{cases} \text{集中型} & ((\lambda_1-\lambda_2)\geq\theta_\lambda) \\ \text{分散型} & (\text{その他}) \end{cases}$$

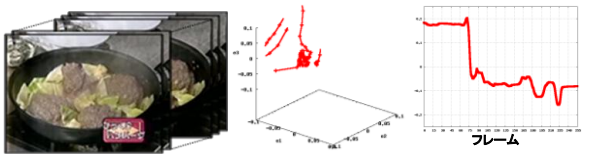
θ_λは各軸まわりの分散に対する閾値とする．



(a) 繰り返し動作



(b) 状態提示



(c) その他動作

図 4: 料理映像（左: 料理番組映像，中: 固有空間上の軌跡，右: 第 1 固有軸上の軌跡）

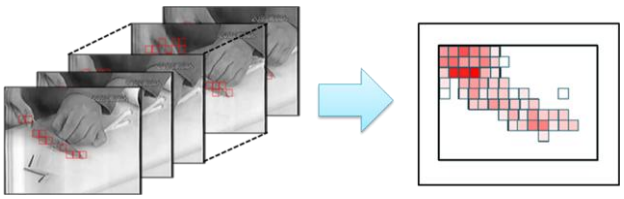
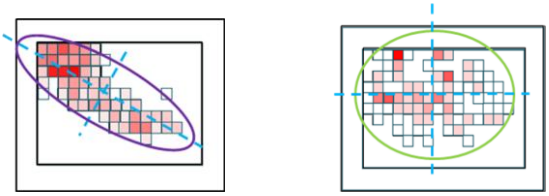


図 5: 周波数解析（濃く示される局所領域ほど，繰り返しの頻度が高い）



(a)集中型　　(b)分散型

図 6: 繰り返し動作の種類（破線は局所領域の分布を主成分分析した際の軸）

3.4. 料理映像に対するタグの対応付け

最後に，以上の結果を統合することで素材・調理動作の組と手元ショットの対応付けを行う．まず，分類した料理映像の時刻を用いて，対応するタグを列挙する．そして，それらのタグの中から手元ショットの分類結果と動作の種類が一致するものを対応付ける．

4. 実験

4.1. 実験条件

提案手法を用いて，料理レシピと料理番組映像の対応付け実験を行った．実験には，NHK「きょうの料理」の料理レシピ 8 本[2]と対応する映像計 75 分を用いた．なお，本実験においてカット検出，手元ショット抽出は人手で行った．また，ショットと対応するタグがない場合，その手元ショットを分類対象から除外した．繰り返し動作の分類に用いる窓幅は，予備実験によって 256 フレーム（約 8 秒）と決定した．評価基準として，目視によって与えた正解に対する対応付け精度を用いた．「動作」のみおよび「動作・素材」の組，2 つの基準で対応付け精度を評価した．「動作・素材」においては，「動作」が正解でなければ，「素材」が正しくても正解として扱わないものとした．

4.2. 実験結果

対応付け精度に関して，「動作」のみでは 68.1%であり，「動作・素材」の場合では 52.6%であった．

4.3. 考察

まず，テキスト処理部について考察する．料理レシピと CC それぞれの解析結果を照合する際には，異なる動作表現で生じた問題が多く見られた．たとえば，（ふる⇔かける），（添える⇔のせる）のような動詞対はほぼ同じ動作を表す．しかし，表現は全く異なり，正しく照合できない．そのため，以上のような料理用語の言い換え表現に関する辞書が必要となる．

次に画像処理部分については，カメラワークを含む手元ショットへの対処が挙げられる．また，「繰り返し動作」と「その他の動作」を高精度に分類するため，最適な閾値を検討する必要がある［6］．

最後に，テキスト情報と画像情報を統合した対応付けについて考察する．正確に対応付けできなかった主な要因として，図 7（左）を示すように映像中の主たる動作ではないものの存在が挙げられる．湯が煮立っているが，映像全体としては鍋の焦げをしゃもじで煮溶かしていた．また，「透き通る」や「しんなりする」など，動作では表現できないものも存在した．図 7（右）に示すように，対応したのは「卵，固まる」であるが，「固まる」は動作を表現する動詞ではないので，映像で表現されないものもあった．

しかし，この結果は，複雑な構造を持つ料理レシピと料理映像の対応付け性能としては低くないと考える．また，本研究の最終目標は調理動作映像データベースの作成であり，実際にユーザが利用する際に，様々な検索候補から正しいと思うものを選択して利用すれば良いと考えるため，この性能でも十分実用的であると考える．

[2] https://www.kyounoryouri.jp/から入手．

図 7：誤った対応付けの例

5. むすび

本稿では，調理動作映像データベースを構築することを目的とし，テキスト情報と画像情報を統合した動作解析による料理レシピと料理映像の対応付け手法を提案した．実際に放送された料理映像を用いて対応付けを行った結果，動作のみでは 68.1%，動作・素材では 52.6%の対応付け精度が得られた． 今後は，調理動作映像データベースを用いた料理支援インタフェースの実現を考えている．

謝辞

共に研究を進めてきた名古屋大学村瀬研究室の諸氏，特に野田雅文氏と道満恵介氏に，心から感謝いたします．研究に必要なデータを提供してくださった国立情報学研究所（「評価用映像メディア DB」[8]）に感謝いたします．本研究では，画像処理に MIST ライブラリ (http://mist.suenaga.m.is.nagoya-u.ac.jp/) を使用しました．

文　献

[1] クックパッド株式会社，“毎日の料理を楽しみに COOKPAD”，http://cookpad.com/

[2] Nintendo，“しゃべる！DS お料理ナビ”，http://www.nintendo.co.jp/ds/a4vj/

[3] 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.21-29，2003.

[4] 浜田玲子，佐藤真一，坂井修一，田中英彦，“料理映像における繰り返し動作のスポッティング手法”，信学技報，MVE2001-29，2001.

[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.2129-2139，2007.

[6] 蒏承穎，高橋友和，井手一郎，村瀬洋，“画像特徴の時間変化に基づく料理映像の分類”，信学総大，A-16-2，2009 .

[7] 蒏承穎，志土地由香，高橋友和，井手一郎，村瀬洋，“料理映像における調理動作の解析”，第 4 回デジタルコンテンツシンポジウム，No.8-2，2008.

[8] 馬場口登，栄藤稔，佐藤真一，安達淳，阿久津明人，有木康雄，越後富夫，柴田正啓，全炳東，中村裕一，美濃導彦，松山隆司：“映像処理評価用映像データベースについて”，信学技報，PRMU2002-30，2002.