動きに基づく料理映像の自動要約手法

浜田　玲子[†]　三浦　宏一[††]　井手　一郎[†††]　佐藤　真一[†††]　坂井　修一[††]
田中　英彦[††]

† 東京大学大学院 工学系研究科
†† 東京大学大学院 情報理工学系研究科
〒 113–8656 東京都文京区 7–3–1
††† 国立情報学研究所
〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{reiko,miura,sakai,tanaka}@mtl.t.u-tokyo.ac.jp, ††{ide,satoh}@nii.ac.jp

要約　近年、マルチメディア情報の扱いの重要性が増すにつれ、テレビ映像の自動要約に関する研究が盛んに行なわれつつある。本稿では、料理映像を対象にした自動要約手法を検討する。料理映像においては、調理動作および素材や料理の状態を示す部分が特に重要である。そこで、画面全体の動きの激しさからこれらを検出する手法を提案する。また、調理動作の中でも特に重要なものとして繰り返し動作に着目し、その自動検出手法について述べる。また、各々の手法について、評価実験によりその有効性を示した。さらに、提案手法によって抽出された重要部分から料理映像要約を生成するアプリケーションを実装した。その結果、要約映像は十分に調理手順の内容を保ちつつ、もとの映像の 1/8 から 1/10 に短縮できた。

A Motion Based Automatic Abstraction of Cooking Videos

Reiko HAMADA[†], Koichi MIURA[††], Ichiro IDE[†††], Shin'ichi SATOH[†††], Shuichi SAKAI[††], and Hidehiko TANAKA[††]

† Graduate School of Engineering, The University of Tokyo
†† Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
††† National Institute of Informatics
2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †{reiko,miura,sakai,tanaka}@mtl.t.u-tokyo.ac.jp, ††{ide,satoh}@nii.ac.jp

Abstract　Reflecting the increasing importance of handing multimedia data, many studies are made on automatic abstraction of TV broadcast video. In this paper, we propose a method to abstract cooking videos. Important segments in a cooking video are cooking motions and appearances of foods. So, the method to extract these important segments referring to intensity of motion in the image is proposed. In addition, based on the observation that especially important cooking motions tend to be repetitious, a method to detect them is proposed. Effectiveness of both methods are shown through evaluation experiments. We also developed a cooking video abstraction system that assembles important segments detected by the proposed methods. The resultant abstracted videos were about 1/8 to 1/10 of the original videos in time, maintaining the understandability of cooking procedures.

1. はじめに

映像技術の進歩に伴い、種々のメディアを通じて様々な映像が発信され、大量に蓄積されつつある。そこで近年、これらのマルチメディアデータを有効に活用するための映像の索引付けや検索、自動要約などに関する研究が盛んに進められている。我々は、様々な映像の中でも生活に密着した料理映像に着目し、

映像の解析および索引付けなどの研究を行なっている。本稿では、特に料理映像の自動要約を目的とした映像解析手法を提案する。

これまで、ニュースやドキュメンタリ映像などを対象とした自動要約に関する様々な研究 [1] がなされているが、要約された映像は見づらいとの報告もある [3]。これは、要約映像において音声が断続的に途切れ、映像との同期も失われるためであるといわれている。しかし、料理映像では音声がなくても視覚的な情報から動作や手順を容易に理解できるため、自動要約が効果を発揮すると考えられる。

料理映像は一般に雑談など冗長な部分を比較的多く含み、閲覧に時間がかかる。そのため一般にレシピ選びや実際の調理の際には、テキスト形式のレシピを閲覧する方が簡便である。しかし、映像はテキストでは表現しきれない様々な重要な情報を含み、特に調理手順の理解のためには視覚情報が非常に有効である。そこで、調理手順の重要な部分を集めた要約映像を作成することにより、短い閲覧時間で調理の様子を視覚的・直感的に知ることができるようになる。さらにこのような料理の要約映像を集めたディジタルレシピライブラリの作成なども考えられる。

また本稿で提案する映像解析手法は、自動要約だけではなく、映像の索引付けにも有効である。料理映像に索引を付けることにより、映像の要約や検索、テキスト教材と統合することによるマルチメディアデータの生成など、様々な実用的なアプリケーションへの応用が期待される。今後は、家庭内への計算機の進出に伴い、このような要約あるいは索引付けされた料理映像や料理レシピに対する需要が高まっていくものと考えられる [4]。

本稿では、まず 2 章で一般的な映像と異なる料理映像の特徴と、調理手順を理解するうえで重要と考えられる部分の映像特徴について述べる。次に 3 章でそのような特徴に基づく重要映像抽出手法の提案と評価を行なう。続く 4 章で、調理動作を含む映像の中でもとりわけ重要である繰り返し動作の検出手法の提案と評価を行なう。最後に 5 章でこれらの手法を組み合わせて実装した自動要約アプリケーションを紹介する。

2. 料理映像の特徴

本章では、一般的な映像と異なる料理映像の特徴と重要部分の定義、および手順を理解する上で重要と考えられる映像の定義を述べる。

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。

料理映像におけるショットは、図 2に示すような (a) 人物ショット、(b) 手元ショットに大きく分類される。

人物ショットは台所のほぼ全体が映され、調理人や助手が調理について説明していることが多い。しかし、手元や食材は部分的に小さく映るのみであり、これから調理に関する視覚的な知見を得ることは難しい。

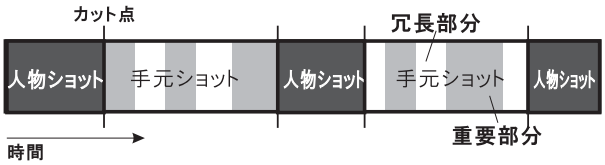
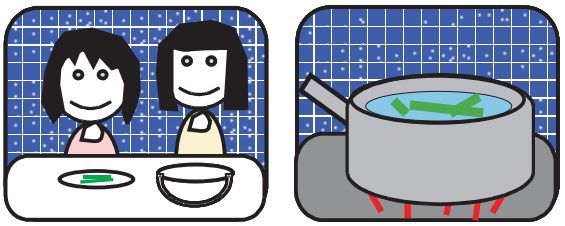


図 1　料理番組における映像構成の例



(a) 人物ショット　　(b) 手元ショット

図 2　料理映像におけるショット分類

一方、手元ショットでは材料を調理する手元や道具が大映しにされ、視覚的に重要な情報を含む。しかし、手元ショットの中には調理中の動作や動作後の料理の様子など、調理を理解する上で重要な映像を含む一方で、動作と動作の間などの比較的冗長な部分も含まれる。

2.2 重要部分の定義

このような構成の料理映像を要約する際、まず視覚的情報の少ない人物ショットは省略する。さらに手元ショットの中から重要部分を抽出する必要がある。ここで料理映像を要約する際に特に必要なのは、(1) テキストでは表現しきれない重要な視覚的情報を要約映像に含むようにすることと、(2) 調理手順の流れを知るのに必要な情報を失わないようにすることである。

(1) の視覚的情報には、大きくわけて 2 種類の映像がある。一つは (a) 調理動作の様子を示すものである。動作の要領、細かいコツなどは、実際に目で見ないと分からないことが多い。もう一つは、調理後の素材の色、盛り付け具合いなど、(b) 料理や食材の状態を示すものである。料理映像には、このような素材などの状態を示すために静止してしばらく様子を写し出す部分がある。また、これらを要約に含めることで、動作と進行に応じた料理の状態を示すことができ、(2) の条件も同時に満たすことができると考えられる。

そこで本稿では、料理映像から (a) 重要な調理動作部分と、(b) 料理や食材の状態を示す部分を抽出し、要約を生成することを考える。

表 1　料理映像の重要部分と動きの特徴

重要部分	動きの特徴
(a) 調理動作	大きい (激しい)
(b) 料理や食材の状態	ほぼ静止

次に、これらの重要部分における映像の動きの特徴を表 1に示す。表 1 に従って、我々はまず画面全体の動きに注目し、映像の中で特に動きの激しい部分を (a) 調理動作として、また静止している部分を (b) 料理や食材の状態として抽出する手法を

提案する。この手法の具体的な内容は 3 章で述べる。

さらに、調理動作には様々なものがあり、特に重要な動作を抽出するためにはより詳しい動きの解析が必要となる。実際の料理映像を参照して検討した結果、調理の中心となる重要な動作の多くは図 3に示すような繰り返し動作であることがわかった。そこで、動作の時間方向の周期性に着目した繰り返し動作の検出手法を提案する。この手法の具体的な内容は 4 章で述べる。

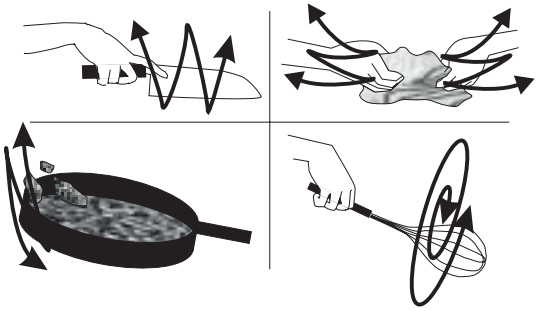


図 3　繰り返し動作の例

3.　重要部分抽出

本章では、画面全体の動きに注目し、調理手順を理解する上で重要である (a) 調理動作部分と (b) 状態部分を抽出する手法を提案、評価する。

3.1　動きに基づく重要部分抽出

本研究では、映像中から動きを検出する手法としてオプティカルフローを利用する。オプティカルフローを検出する手法は数多く提案されているが、現時点では、(1) 映像全体の大まかな動きに注目することが目的で、厳密な解析は必要ない、(2) 大量の画像を処理するためできるだけ単純な手法を用いたい、などの理由から、基礎的な手法である Horn らの手法 [6] を基に実装した。

動きに基づく重要部分抽出の手順を次に示す。

（1） カット検出を行なう。

（2） 各ショットを人物ショットと手元ショットに分類し、人物ショットを除く。

（3） 手元ショットのオプティカルフローを検出する。

（4） 各フレームごとに、全画素のオプティカルフローベクトルの大きさの和をとる (S とする)。

（5） ノイズの影響を軽減するため、10 フレームごとに S の平均をとる ($\overline{S}$ とする)。

なお、カット検出は DCT クラスタリングを利用した手法 [2]、またショット分類は肌色の統計情報を利用して顔領域を検出する手法 [5] を用いて実現した。

実際の料理映像における $\overline{S}$ の時間変化を図 4に示す。

このように変化する $\overline{S}$ に基づいて、重要部分である (a) 調理動作部分と (b) 状態部分を抽出する。

ここで、S のショット内の平均を S_{ave}、また S_{move}, S_1, S_2 をそれぞれ検出に用いる閾値とする。

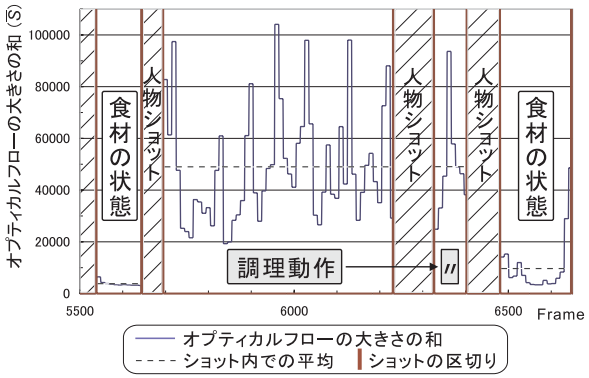


図 4　フレーム毎のオプティカルフローの大きさの和 ($\overline{S}$) の時間変化

まず、$S_{ave} \geq S_{move}$ を満たすショットの中で、$\overline{S} > \alpha S_{ave}$ を満たす部分を調理動作部分として抽出する (α:定数)。これは、全体的に動きの激しいショットのなかでも特に大きな動きを示す部分を調理動作として抽出することを意味する。

次に、$\overline{S} < S_1$ を T フレーム以上満たす部分、あるいは、$S_{ave} < S_2$ をみたすショットの中で、$\overline{S} < S_2$ を満たす部分を料理や食材の状態を示す静止部分として抽出する。前者は動きの少ない映像が連続する部分、また後者は全体的に動きの少ないショットの中で特に動きのない部分を状態として抽出することを意味する。

3.2　カメラワークによる動きの除去

前節の手法では、画面全体に大きな動きが生じるためにカメラワークも調理動作として誤検出してしまう。そこで、すでに検出されたオプティカルフローを利用してカメラワークを検出して、重要部分から除去する。ここでは、カメラワークの中でも特に誤検出の原因となる平行移動 (パン) の検出について検討した。具体的な手順を次に示す。

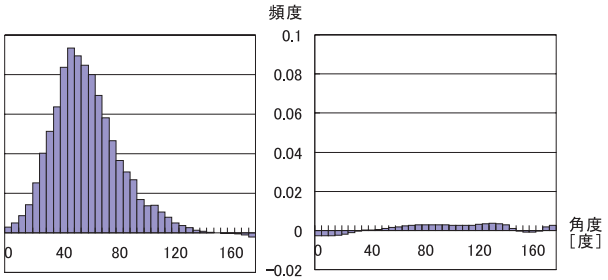
（1） 1 フレーム中の全画素において、オプティカルフローベクトルの向き (角度) を計算する。ベクトルの大きさで重み付けをし、角度の分布をとる。

（2） 一連の動きと見なせる範囲のフレームについて、角度分布の平均をとる。

以上により、パンを含む動きの場合には、角度分布は図 5(a) のようにある程度の大きさの際立ったピークを 1 つもち、カメラワークがなく、調理動作のみの場合には図 5(b) のように明確なピークがないことが観測された。したがってこれを利用し、角度分布のピークの値 (頻度)F_p がある適当な閾値 F_{th} 以上であり、かつピークが 1 つのみであるものをパンとして検出することとした。

3.3　評 価 実 験

ここまでに述べた手法に基づき、料理映像からの動作部分・状態部分の検出実験を行なった。本実験では、ショット分類は手動で行ない、3.1節で述べた手法に基づき動作部分と状態部分を検出する。検出された動き部分からは、3.2 節の手法に基づきカメラワーク（パン）部分を除外した。約 40 分間、6 レシ



(a) カメラワークあり　　(b) カメラワークなし

図 5　オプティカルフローの角度分布

表 2　実験に用いた閾値

$S_{move} = S_{state2} = 10,000$
$S_{state1} = 7,000$
$\alpha = 1.0$
$T = 90$ (3 seconds)
$F_{th} = 0.025$

ピ分の料理映像に対して実験を行ない、各パラメータについては適切な閾値を設定した。具体的な値を表 2に示す。

表 3に重要部分検出実験の結果を示す。目視による結果を Ans_H、自動解析による結果を Ans_M、両者が一致した答を Ans_C 、再現率は Ans_C/Ans_H、適合率は Ans_C/Ans_M とする。なお、目視による重要部分検出においても動作の始まりと終わりはあいまいであるため、フレーム単位での厳密な区間を決定することは困難である。そこで、今回は目視により特定された各重要部分を含む（ただし明らかに重要部分以外と思われる区間は含まない）区間が検出できれば正解とした。

表 3　重要部分検出結果

検出部分	Ans_H	Ans_M	Ans_C	再現率	適合率
調理動作	119	127	117	98%	92%
状態	46	41	39	85%	95%

表 3に示すように、単純な手法により、調理動作および食材の状態を示す静止部分を高精度で検出できることが示された。調理動作の誤検出と状態部分の検出漏れの主な原因は、調理に関係のない動きを誤検出したことであった。

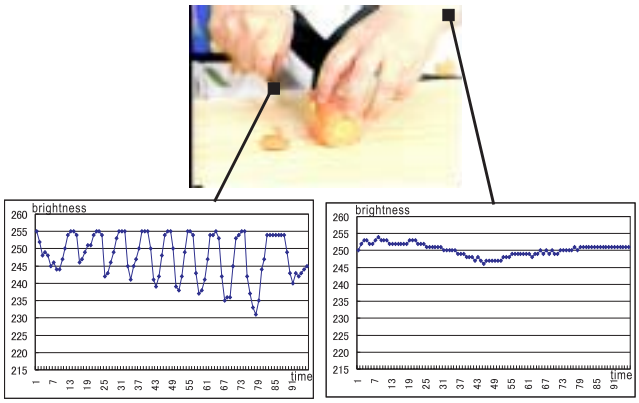
調理動作の検出漏れの原因は、動作が小さすぎたこと、また静止部分の誤検出の原因は重要でない（状態を見せているわけではない）のに画面が静止していたことであった。

4.　重要動作検出

前章では動作部分と状態部分を重要部分とみなして抽出したが、より効果的な要約映像を生成するためには、各々の重要部分の中から特に重要な部分を抽出する必要がある。そこで、本章では、実際の調理動作を参照した結果、調理動作の中でも特に重要な動作の一つであり、かつ画像特徴の明確な「繰り返し動作」の検出手法を提案し、評価する。

4.1　繰り返し動作の検出

繰り返し動作の映像においては、映像の局所領域上を対象物が往復する。そのため、図 6に示すように、繰り返し動作の周辺における輝度値は周期的な変化を示す。そこで、本研究では時間周波数解析によって局所領域の輝度値の時間変化を解析し、その周期性の有無から繰り返し動作を検出する。以下にこの手法を説明する。



(a) 繰り返し動作周辺　　(b) 背景

図 6　局所領域における輝度値の時間変化

まず、各フレームを 3×3 ピクセルから成るブロックに分割する。ここで、各ブロックに含まれるピクセルの平均輝度値を $V_{x,y}(t)$ とする。なお、x, y は画像におけるブロックの空間座標、また t はそのブロックが属するフレームの時間座標である。

次に、画像中のすべての x, y における $V_{x,y}(t)$ にそれぞれ FFT を適用し、その周期性を調べる。FFT を適用する時間方向の範囲は、フレーム数 $T = 2^n$ の時間窓内とする。

$V_{x,y}(t)$ に明確な周期性がある場合、結果の FFT グラフにはある周波数で明確なピークができると考えられる。このようなピークを検出するため、FFT グラフに関するいくつかの統計量を利用する。その際に、人間の繰り返し動作の早さから、考慮する周波数帯を限定した。その範囲を $f_0 \leqq f < f_0 + N$ とする（図 7）。

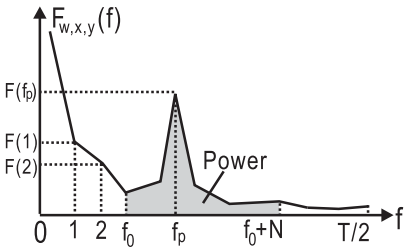


図 7　FFT グラフ

まず、範囲内のパワーの総和を $Power_{w,x,y}$ とする。次に、範囲内での $F(f)$ の最大値を与える周波数を f_p とする。さらに $F(f_p)$ がグラフにおいてどの程度突出しているのかを知る

ため、$F_{peak}(f_p)$ なる指標を定義した。これは、$F(f_p)$ を除く $F(f)$ の平均値と $F(f_p)$ との比である (式 1)。

$$F_{peak}(w,x,y) = \frac{F(f_p) \times (N-1)}{\sum_{f=f_0, f \neq f_p}^{f_0+N-1} F(f)} \quad (1)$$

また、低周波数におけるパワー $F(1)$ や $F(2)$ よりも $F(f_p)$ が十分に大きいかどうかを知るため、$F(f_p)$ の $F(1)$ と $F(2)$ に対する比である R_1, R_2 をそれぞれ定義する。

最後に、ピークの鋭さの指標 R_{sharp} を式2のように定義する。

$$R_{sharp} = \frac{F(f_p) \times 4}{\sum_{f=f_p-2, f \neq f_p}^{f_p+2} F(f)} \quad (2)$$

2 点以上のブロックにおいて、以上までに説明した $Power$, F_{peak}, f_p, R_1, R_2, R_{sharp} の 6 パラメータがいずれも閾値以上の値をもつとき、その時の窓 w において繰り返し動作を検出する。

4.2 評価実験

前節で述べた手法に従って、料理映像からの繰り返し動作検出実験を行なった。実験の対象とした料理映像は、約 70 分間、16 レシピ分の映像である。なお、各パラメータおよび閾値は表 4に示す通りに適切な値を設定した。表 5に実験の結果を示す。目視による結果を Ans_H、自動解析による結果を Ans_M、両者が一致した答を Ans_C、再現率は Ans_C/Ans_H、適合率は Ans_C/Ans_M とする。対象とする料理映像から、繰り返し動作部分を目視で検出し、これを正解とした。

表 4　評価実験における各パラメータの値および閾値
(a) 周波数解析パラメータ　　(b) 周期性解析パラメータ

$T = 32frame$
$T_{step} = 16frame$
$f_0 = 3$
$N = 12$

$Power > 500$	$F_{peak} > 50$
$f_p > 5$	$R_{sharp} > 3$
$R_1 > 3$	$R_2 > 3$

表 5　繰り返し動作検出結果

Ans_H	Ans_M	Ans_C	再現率	適合率
62	59	50	81%	85%

表 5に示すように、本手法では誤検出が少なく、約 85%の適合率で繰り返し動作を検出できることが示された。一方、検出漏れは比較的多く、再現率は 80%程度である。本手法による成功例、誤検出や検出漏れの例を図 8に示す。

まず、図 8(a),(b) は典型的な成功例である。いずれも十分に高速かつ規則的な動作であり、また調理手順の中でもコツのいる重要な動作である。次に、図 8(c) に誤検出の例を示す。これには動作は含まれていないが、画面右上のフライパンの上に置かれた菜箸が規則的に揺れるため、これを振動として誤検出した。最後に、図 8(d) に検出漏れの例を示す。これはネギをゆっくりと炒る動作である。この例に限らず、一般に鍋などで素材にゆっくりと火を通す際の動作は、動きが遅く、そのため規則性も厳密ではないため、検出漏れが多かった。目視で正解を検出する際にも、特に図 8(d) のようなあいまいな動きの場合は繰り返し動作か否か迷うものが多く、実験においてもこのようなものほど検出が困難であった。

(a) 成功例:キャベツの千切り　　(b) 成功例:鶏肉をもむ

(c) 誤検出:肉を焼く　　(d) 検出漏れ：ねぎを炒る

図 8　繰り返し動作検出の例

5. 料理映像の自動要約

3 章および 4 章の手法により重要動作部分を抽出し、これを利用した自動要約アプリケーションを作成した。図 9に、アプリケーション画面の例を示す。

図 9　料理映像要約アプリケーション画面の例

各手元ショットにおいて、4 章の手法によって繰り返し動作が検出された部分の最初の 2 秒、および 3 章の手法によって抽出された静止部分の最後の 2 秒を拾って要約とした。また、繰り返し動作も静止部分も抽出されなかった手元ショットからは、3 章の手法によって一般的な調理動作が抽出された場合、その最初の 2 秒を要約に含むこととした。なお、ここではショット分類は手動で行なった。

本アプリケーションによる要約結果の例を図 10に示す。なお、各フレームは要約に含まれる映像セグメントを表わす。図

10において、黒い縁のものが繰り返し動作部分、灰色の縁のものがその他の動作部分、白い縁のものが状態部分である。

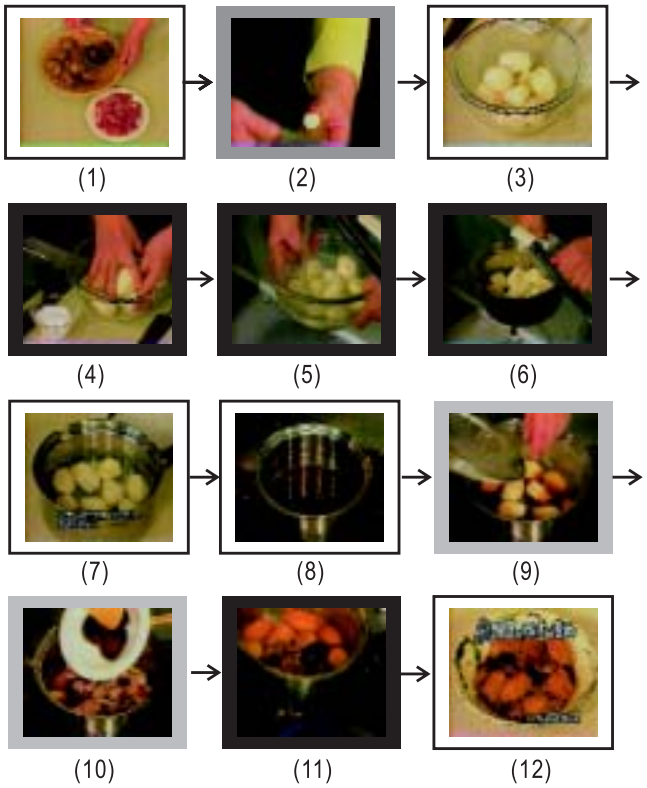


図 10　料理映像から要約された映像セグメント (白い縁：状態映像、黒い縁：繰り返し動作映像、灰色の縁：一般的な動き映像)

図 10において、繰り返し動作 (4)〜(6) は、「里芋を塩でもみ、ぬめりをとって洗い流す」映像である。これらの映像には、調理の手順を伝えるとともに「ぬめりをとる」「洗い流す」といった単語だけでは表現しきれない調理動作に関する重要な視覚的情報が含まれている。(11) も同様に「なべを揺すって味をからませる」という繰り返し動作で、このレシピにおけるコツの部分であり、動きの強さ、早さなど豊富な視覚的情報を含んでいる。次に、繰り返しではない調理動作のうち、(2) は「皮をむく」、(9), (10) は素材を鍋に「入れる」動作である。いずれも、テキストから容易に動きを推測できる動作であるが、要約に含めることで、より調理手順を分かりやすくしている。最後に、(1), (3), (7), (8), (12) は、状態を示す静止部分である。(12) の盛り付けの映像をはじめとして、これらの状態を示す映像には視覚的に重要な情報が含まれるうえ、要約映像における手順の進行を明確にしている。

以上で示したような手法で作成された要約映像は、元の映像とくらべて 1/8 から 1/10 に短縮され、なおかつ調理手順を理解するのに重要な視覚的情報および手順が含まれており、本要約手法の有効性が定性的に示された。

また、ユーザの調理に関する背景知識などに応じてより短い要約で十分な場合には、視覚的に重要な繰り返し動作を残し、その他の動作は必要に応じて省略することが考えられる。この場合、手順の流れに関する情報が部分的に欠落するが、状態部分の映像が適切に抽出されていれば、内容はほぼ理解できるものと考えられる。

このような料理映像の自動要約が実現すれば、これを大量に作成し、要約料理映像データベースを構築することが考えられる。ある程度調理に熟練したユーザであれば、要約映像からそのレシピのおおまかな手順やかかる手間などを知ることができ、元の映像を見るよりも短い時間でなおかつテキストレシピを読むよりも雰囲気をつかみやすいと考えられる。家庭でのレシピ選びなどに利用すれば、一本あたり数十秒〜数分に縮められた映像を閲覧することで、直感的にレシピを選択できるようになる。このように、本手法の様々な応用が考えられる。

6. ま と め

本稿では、動きに基づく料理映像の自動要約手法について提案した。

料理映像においては、画面全体の動きの激しい調理動作部分、および素材や料理の状態を示す状態部分が重要であることに着目し、オプティカルフローによりこれらの重要部分を検出する手法を提案、評価実験によりその有効性を示した。

また、調理動作の中でも特に重要な動作として繰り返し動作に着目し、その自動検出手法について述べ、評価実験によりその有効性を示した。

さらに、両手法を適用した料理映像の自動要約アプリケーションを実装した。その結果、要約映像は十分に調理手順の内容を保ちつつ、元の映像の 1/8 から 1/10 の時間に短縮できた。

今後の課題としては、より柔軟な自動要約アプリケーションを実現するために、動作部分と状態部分以外の重要部分として、字幕の出現する部分を検出したり、要約率を可変にすることなどが考えられる。そのためには、繰り返し動作か否かだけでなく、より細かな動作の分類による重要度の設定が課題となる。

文　献

[1] M. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding techniques”, *Proc. IEEE CVPR’97 Computer Vision and Pattern Recognition*, 1997.

[2] Y. Ariki and Y. Saito, “Extraction of TV news articles based on scene cut detection using DCT clustering”, *Proc. Intl. Conf. on Image Processing*, Vol.3, pp.847-850, 1996.

[3] M. Christel, M. Smith, C.Taylor, and D. Winkler, “Evolving video skims into useful multimedia abstractions,” *Proc. ACM CHI’98 Conference on Human Factors in Computing System*, April, 1998.

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

[5] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. PAMI*, Vol. 18, No. 7, pp.780-785, 1997.

[6] B. K. P. Horn and B. Schunck: “Determining optical flow”, *Artif. Intel.*, Vol.17, pp.185-203, 1981.