動きに基づく料理映像の自動要約

三浦宏一[†] 浜田玲子[†] 井手一郎[††]
坂井修一[†] 田中英彦[†]

近年，マルチメディア情報を有効に活用する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さかんに行われつつある．本論文では，料理映像を対象にした自動要約手法を提案し，検討する．我々は料理映像要約の目的を，調理の全体的な流れを視覚的・直感的に理解するのに十分な映像を作成することとしている．要約映像を作成する際には，映像の重要部分を抽出することが必要となるが，料理映像においては，調理動作および料理や食材の状態を示す部分が特に重要である．これらは画像全体の動きの激しさと関連があることから，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およびカメラワーク（パン）を除去する手法を提案し，評価実験によりその有効性を示した．さらに，この手法によって抽出された重要部分と，調理動作の中でも特に重要な繰返し動作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．放送局の異なる複数の料理番組に提案手法を適用し，要約映像を自動生成した結果，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 1/12 の時間に短縮できた．また，自動要約した映像の一部を，番組制作者によって作成された要約映像と比較することにより，本手法の有効性を確認した．

Motion Based Automatic Abstraction of Cooking Videos

Koichi Miura,[†] Reiko Hamada,[†] Ichiro Ide,[††] Shuichi Sakai[†]
and Hidehiko Tanaka[†]

Reflecting the increasing importance of handling multimedia data efficiently, many studies are made on automatic abstraction of television broadcast video. In this paper, we propose a method to abstract cooking videos. We define cooking video abstraction as shrinking videos maintaining sufficient understandability of general cooking procedures visually and intuitively. To abstract a video, important sub-shot segments need to be extracted from the original video. Important segments in a cooking video are considered as cooking motions and appearances of foods, since visual information that represents essential cooking operation is exceptionally important. These segments have typical motion-related features. Thus, a method to extract such important segments referring to the intensity of motion in the image is proposed. Effectiveness of the method is shown through evaluation experiments. We also implemented an abstracted cooking video browser that assembles important segments detected by the proposed methods and repetitious motions that is especially important among cooking motions. The resultant abstracted videos were about 1/8 to 1/12 of the original videos in time, maintaining the understandability of cooking procedures. And the validity of the abstraction method was checked by comparing some automatic abstracted videos with abstracted videos provided from the broadcaster.

1. はじめに

情報通信技術の発達にともない，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に映像の索引付けや構造化などに関する研究がさかんに進められている．

しかしながら映像には様々な種類のものがあり，各々の映像的特徴も視聴者の視聴目的も様々である．そのため汎用性の高い映像解析技術を用いて，内容に深く立ち入って解析するのは困難であり，高度な内容解析を行うためには，対象とする映像の種類を限定して対象固有の特徴を考慮する必要がある．

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，映像の意味的構造解析や索引付けなどの研究を行っている[1)~3)]．これらの研究では，対象を料理映像に限定し，対象に固有の知識を最大限に活かす

† 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†† 国立情報学研究所
National Institute of Informatics

ことで，比較的簡単な要素技術を用いながらも各々の処理の精度を確保し，実用的なシステムの構築を目標としている．また，料理はきわめて日常的な行為でありながら，豊富な知識と熟練を必要とする複雑な作業であるため，従来から調理支援の需要は存在してきた．今後，家庭内へ計算機が進出するのにともない，このように解析・索引付けされた料理映像や料理レシピの検索の需要は高まっていくと考えられる．

本論文では，一連の研究の中でも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．料理映像は教材映像の一種であり，多くの視聴者は，実用的な教材としての利用を目的としている．その一方で，雑談などの冗長な部分も多く，閲覧にはある程度の時間を要する．そのため日常的な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ことが多い．しかし，映像にはテキストでは表現しきれない様々な重要な視覚的情報が含まれており，特に調理手順の理解のためには非常に重要である．

そこで，本研究ではこのようなレシピ選びや実際の調理の際に，テキスト形式のレシピを閲覧するよりも視覚的情報を多く含み，なおかつ同様の簡便さで閲覧できる要約映像を作成することを目的とする．個人の技量や好みも考慮すべきではあるが[4),5)]，まずは調理手順の重要な部分を集め，短い閲覧時間で手順の全体的な流れを視覚的・直感的に十分に理解できる映像を作成する．つまり，元の映像と併用するのではなく，要約映像単独でも手順の概要が理解可能な精度を目標とする．

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がなされているが[6),7)]，それらの映像は比較的冗長性が低く，多くは元の映像を全体にわたって閲覧することを目的として，要約映像を利用する傾向が強い．一方，本研究と同様の目的で要約映像を作成するものは，スポーツ映像などを対象にした研究[4),5)]に見られる．アメリカンフットボール映像を題材とした要約映像作成の研究[5)]では，映像の意味内容の抽出をハイライト部を記述した外部データベースと映像の対応付け問題に還元して考察している．この手法では，対応付けしやすい外部情報源がある場合には有効な要約映像を作成できると考えられるが，料理映像では，映像内容と直接対応した記述は容易に利用できず，この手法を適用することは難しい．

また，これまでの研究において要約された映像は見にくいとの報告もある[6)]．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．そこで，要約映像を作成する際には，音声部分が不自然にならないように考慮して映像を切り出したり[7)]，制作者側で作成される番組予告などの要約映像では音声は別のものに吹き替えられたりする．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という特徴があるため，本研究で作成する要約映像では音声の連続性などは考慮しない．そのため，画像特徴に沿った柔軟なショット構成が可能となり，精度の良い要約映像が作成できると考えられる．

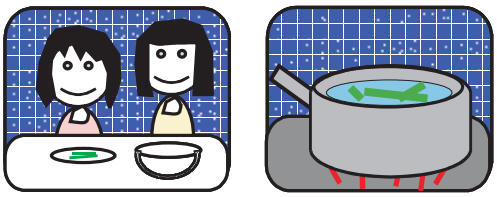
以下，2 章では本研究で対象とする料理映像の特徴と重要部分について述べ，3 章でその重要部分の検出手法について提案する．4 章では，検出した重要部分から要約映像を作成し，評価・考察を行う．最後に 5 章でまとめと今後の課題について述べる．

2. 料理映像の特徴

本章では，料理映像の特徴と，要約と密接に関わる料理映像における重要部分の定義について述べる．

2.1 料理映像の構成

図 1 に示すように，料理映像のショットは大きく図 1 (a) 人物ショット，図 1 (b) 手元ショットの 2 つに分類でき，**図 2** に示すように各々が交互に出現する．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，映像から調理に関して視覚的な知見を得ることは難しい．一方，手元ショットは材料やそれを調理する手元が大きく映され，視覚的に重要な情報を含む．しかし図 2 にも示すように，手元ショットの中にも，動作と動作の



(a) 人物ショット (b) 手元ショット

(a) Face shot (b) Hand shot

図 1 料理映像におけるショット分類
Fig. 1 Shot categories in cooking video.

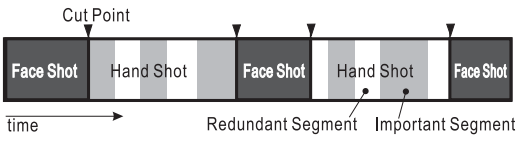


図 2 料理映像のショット構成例
Fig. 2 Structure of cooking video.

間など比較的冗長な部分もある．

また，これらのショットは主に固定カメラで撮影され，カメラワークは，他の種類の映像（スポーツ映像，ドラマ映像など）に比べごくわずかしか用いられないため，次節で述べる重要部分と映像の動きの関係が成り立つ．

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，まず調理のための視覚的情報に乏しい人物ショットを除外する．さらに残された手元ショットの中からも冗長な部分を除外する必要がある．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含むことと，(2) 調理手順の流れを知るのに必要な情報を失わないことである．

(1) の視覚的情報には，大きく分けて 2 種類の映像がある．1 つは (a) 調理動作の様子を示すものである．これは，動作の要領や細かいコツなどは，実際に目で見ないと分からないことが多いからである．もう 1 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映し出す部分がある．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せると考えられる．そこで本研究では，料理映像から (a) 重要な調理動作部分（「調理動作部分」）と，(b) 料理や食材の状態を示す部分（「状態部分」）を抽出し，要約を生成することを考える．これらの重要部分における画像中の動きには，以下のような特徴がある．

(a) 調理動作部分：大きい（激しい）

(b) 状態部分：ほぼ静止

さらに，調理動作には様々なものがあるが，より効果的な要約映像を作成するために，一般的な調理動作の中から特に重要な動作を抽出して特別に扱う．実際の料理映像を参照して検討した結果，特に重要な動作に比較的共通する性質として，**図 3** に示すような繰返し動作があることが観察された．重要な動作には様々なものがあるが，本論文では，調理動作の中でも特に繰返し動作を取り上げることとする．そこで，動作の時間方向の周期性に着目した検出手法[3]を用いて繰返し動作を検出し，要約映像作成に利用することとした．

3. 重要部分検出

本章では，画像中の動きに注目し，調理手順を理解するうえで重要である，(a) 調理動作部分と，(b) 状態部分を検出する手法を提案する．また，(a) のうち特に重要な繰返し動作の検出手法について簡単に紹介する．

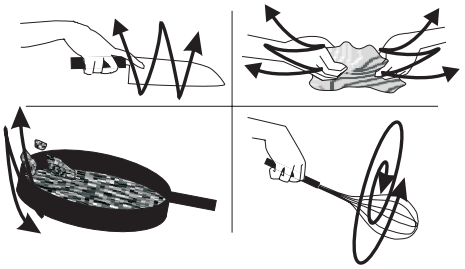


図 3 繰返し動作の例
Fig. 3 Examples of repetitious motions.

3.1 動きに基づく重要部分検出

本研究では，画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[8]，ここでは，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こと，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ことなどから，Horn らの手法[9]を用いた．

動きに基づく重要部分検出の手順を次に示す．

(1) カットを検出

(2) 各ショットを人物ショットと手元ショットに分類し，人物ショットを除外

(3) 手元ショット中のオプティカルフローを検出

(4) 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を積算（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 を平均（$\overline{S}$ 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[10]，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し，分類する手法[2]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) 状態部分を検出する．

ここで，S のショット内平均を S_{ave}，また S_{move}，S_{state1}，S_{state2} を (a), (b) の検出に用いる閾値とする．

まず，$S_{ave} \geq S_{move}$ を満たすショットの中で，$\overline{S} > \alpha S_{ave}$ を満たす区間を調理動作部分として検出する（α：定数）．ただし，60 フレーム（2 秒）以内で隣接する区間については連続する動作と見なし，1 つの区間として検出する．これは，全体的に動きの激しいショットの中でも特に大きな動きを示す部分を調理動作として検出することを意味する．

次に，$\overline{S} < S_{state1}$ が T フレーム以上継続する区

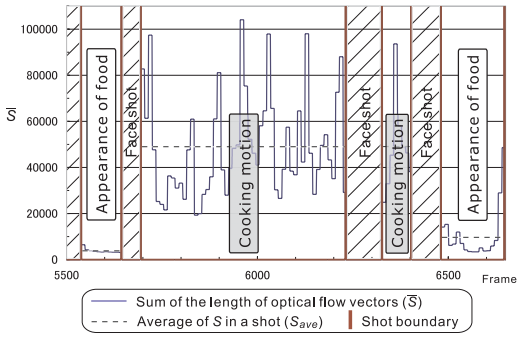
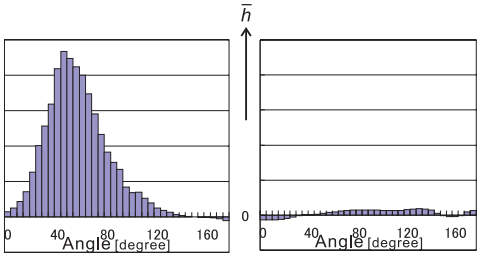


図 4 フレームごとのオプティカルフローの大きさ（$\overline{S}$）
Fig. 4 Temporal transition of the sum of the length of optical flow vectors ($\overline{S}$) in each frame.



(a) パンを含む場合 (b) パンを含まない場合
(a) With *panning*. (b) Without *panning*.

図 5 オプティカルフローの角度分布
Fig. 5 Angle histogram of optical flow.

間，あるいは，$S_{ave} < S_{state2}$ を満たすショットの中で，$\overline{S} < S_{state2}$ を満たす区間を料理や食材の状態を示す静止部分として検出する．前者は動きの少ない画像が連続する部分，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検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，画像全体に大きな動きが生じるために，カメラワークについても調理動作の重要部分として誤検出してしまう．そこでカメラワークを検出し，調理動作として検出された重要部分から除外する必要がある．

料理映像中のカメラワークは，大きくパン（画像が一定方向に平行移動）☆とズーム（対象にズームアップ・ダウン）の 2 種類に分けられる．パンは調理動作や料理や食材の状態を見せるのが目的ではなく，対象から対象へカメラを移動させる途中に現れるものである．したがって，画像中に視覚的知見が得られる重要な事象は映っていないことが多く，重要部分検出の誤検出の主な原因となる．一方ズームは，一般的に画像の中心に視聴者に見せたい重要な事象が映る傾向があることから重要部分に含むこととし，特に区別して検出する必要はない．

そこで，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パンを検出する．カメラワークを検出する研究には過去にも様々なものがあるが，すでに前節で検出されているオプティカルフローを用いた，単純な手法で検出するために，以下のような手順で行った．

(1) フレーム f 中の全ピクセル $p(i,j)$ において，オプティカルフローベクトルの向き（$0 \leq \theta(i,j) < 2\pi$）を求める．ベクトルの大きさ $v(i,j)$ で重み付けをし，角度の分布をとる（angle histogram）．この際，角度分布を $H_f = \{ h_f(\Theta) \mid 0 \leq \Theta < \pi \}$ とし，$\pi \leq \theta(i,j) < 2\pi$ の向きのベクトルに対しては，$\Theta = \theta(i,j) - \pi$ とし，負の重み $-v(i,j)$ を持たせる．

$$h_f(\Theta) = \frac{1}{S} \sum_i \sum_j \delta_\Theta(i,j) \cdot v(i,j) \quad (1)$$

ただし，

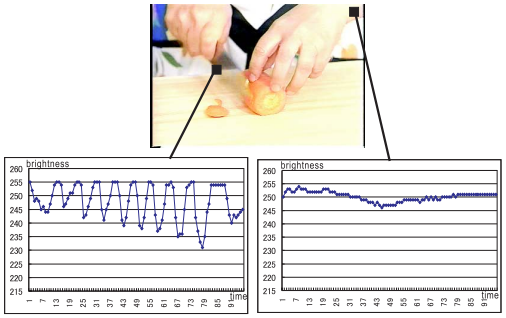
$$\delta_\Theta(i,j) = \begin{cases} 1 & (if\ \Theta = \theta(i,j)) \\ -1 & (if\ \Theta = \theta(i,j) - \pi) \\ 0 & (otherwise) \end{cases}$$

(2) 一連の動きと見なせる範囲のフレーム（f_1〜f_2）について平均した角度分布 $\overline{H} = \{ \overline{h}(\Theta) \mid 0 \leq \Theta < \pi \}$ をとる．一連の動きは，$\overline{S}$ を基に判断する．(1) で $\pi \leq \theta < 2\pi$ の向きのオプティカルフローベクトルに対し負の重みを持たせたことにより，逆方向の動きは打ち消し合うので，雑音（ランダム性を仮定）や動作が角度分布中に占める大きさは小さくなり，パンを検出できる．

$$\overline{h}(\Theta) = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} h_f(\Theta) \quad (2)$$

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 つ持ち，そうでない動きの場合には図 5 (b) のように明確なピークがないことが観測された．このような特徴を利用し，角度分布のピーク値 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し，重要部分から除外することとした．

☆ ここでは，パン（左右方向），チルト（上下方向）などを区別せず，画像が一定方向に平行移動する場合をすべて含めてパンと呼ぶこととする．



(a) 繰返し動作周辺 (b) 背景

(a) Repetitious motion. (b) Background.

図 6 局所領域における輝度値の時間変化
Fig. 6 Temporal transition of brightness in small regions.

3.3 繰返し動作の検出

ここまで述べた手法は，動作部分と状態部分を重要部分と見なして検出するが，より効果的な要約映像を作成するためには，各々の重要部分の中からさらに重要な部分を抽出する必要がある．そこで本研究では，文献 3) の手法を用い，調理動作の中でも特に重要な動作の 1 つである繰返し動作を検出し，要約映像作成の際に用いることとした．

繰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，繰返し動作の周辺における輝度値は周期的な変化を示す．文献 3) の手法では，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返し動作を検出している．以下にこの手法を簡単に説明する．

まず，各フレームを 3×3 ピクセルからなるブロックに分割し，各ブロックに含まれるピクセルの平均輝度値を求める．次に，画像中のすべてのブロックにおける平均輝度値に，それぞれ一定フレーム数の時間範囲で FFT を適用し，周期性を調べる．明確な周期性がある場合，結果の FFT グラフに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，人間の繰返し動作の速さから，考慮する周波数の範囲を $f_0 \leq f < f_0 + N$ と限定する．FFT グラフの例を**図 7** に示す．$F(f)$ は周波数 f におけるパワーである．このグラフから，範囲内での $F(f)$ の最大値を与える周波数 f_p，$F(f_p)$ がグラフにおいてどの程度突出しているのかの指標，範囲内のパワーの総和などのパラメータを定義し，これらの値を参照して繰返し動作を検出する．

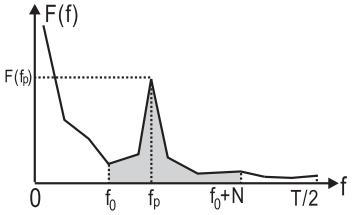


図 7 FFT グラフ
Fig. 7 FFT graph.

表 1 実験データの特性
Table 1 Property of the videos.

Total time	40 minutes
Number of recipes	6
Format	Motion JPEG（transformed into 24 bit color bitmap）
Resolution	320×240 pixels
Frame rate	30 frames/second

3.4 重要部分検出実験

以上に述べた手順に基づき，料理映像から調理動作部分，料理や食材の状態に関する部分を検出する実験を行った．

まず，予備実験としてカット検出とショット分類の実験を，約 100 分間の特定の料理番組の映像（計 600 ショット）を対象として行った．その結果，カット検出については再現率 94.8%，適合率 98.3%の精度が得られた．またショット分類については，人物ショットについては再現率 87.6%，適合率 88.5%，手元ショットについては再現率 89.9%，適合率 89.1%と，要約作成上問題ない程度の性能で自動化できることが示された．

次に，重要部分検出実験を行った．予備実験によりカット検出，ショット分類ともに高い精度が得られることを確認したので，本実験では，動きに基づいた重要部分検出手法単独での評価をするため，カット検出，ショット分類は理想的に行われたものとし，3.1 節で述べた手法に基づいて動作部分と状態部分を検出した．また，3.2 節で述べた手法に基づいてカメラワーク（パン）を検出し，動作部分から除外した．

実験には，ある特定の番組からキャプチャした 6 レシピ分（約 40 分間）の料理映像を用いた．**表 1** に実験に用いた映像の特性を示す．また，**表 2** に本実験で用いた閾値を示す．これらの閾値は予備的な実験に基づいて決定した．

表 3 に重要部分検出実験の結果を示す．単純な手法により，調理動作および料理や食材の状態に関する重要部分を高精度で検出できたことが分かる．

本実験では，目視により検出したものを正解とし，

表 2 実験に用いた閾値
Table 2 Thresholds.

S_{move}	10,000
S_{state1}	7,000
S_{state2}	10,000
α	1.0
T	90
F_{th}	0.025

表 3 重要部分検出結果
Table 3 Result of important segment detection.

重要部分	N_C	N_M	N_O	再現率	適合率
調理動作	117	10	2	98%	92%
状態	39	2	7	85%	95%

正答数を N_C，誤検出数を N_M，検出もれの数を N_O，再現率は $N_C/(N_C+N_O)$，適合率は $N_C/(N_C+N_M)$ 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厳密に定義できないため，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．また本手法は，要約映像作成のための重要部分検出手法であるため，検出された区間に重要部分を包含していることが重要である．そこで本実験では，区間の開始点と終了点に関して，目視と自動検出との間において 3 秒程度の誤差範囲を許容することとし，正解区間が検出できれば正答とした．

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検出してしまったことによるものであった．調理動作の検出漏れの原因は動作が小さすぎたこと，また状態部分の誤検出の原因は重要でない（映像の制作者が状態を見せようとしているのではない）にもかかわらず画像が静止していたことであったが，いずれの場合も稀であった．

また，カメラワーク（パン）を検出したことにより，カメラワーク検出をしない場合に比べて，誤検出の約 40%を削減（17 から 10 に減少）することができた．

4. 料理映像の自動要約

3 章の動きに基づく重要部分検出手法，および繰返し動作検出手法により料理映像の重要部分を抽出し，これを利用した自動要約アプリケーションを作成した．

4.1 要約映像の作成

各手元ショットにおいて，まず，繰返し動作が検出されたショットに対しては繰返し動作部分の先頭を抽出した．また，繰返し動作が検出されないショットや繰返し動作部分から十分に（10 秒以上）離れている部分に関しては，動きに基づく重要部分検出手法によって検出された調理動作部分の先頭を抽出した．さらに，

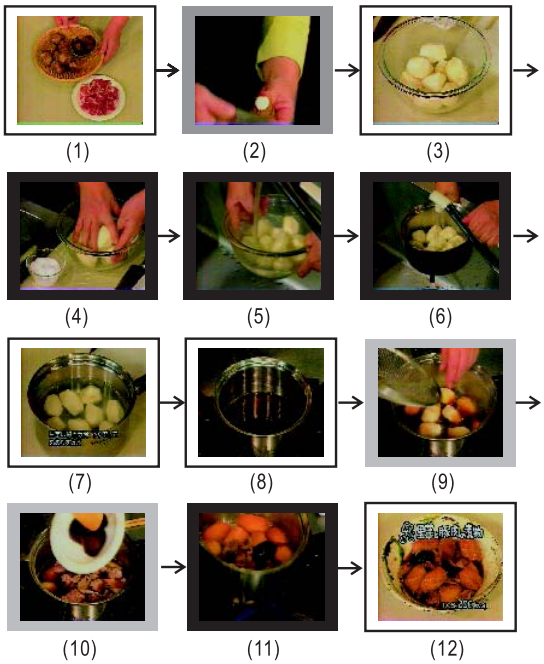


図 8 料理映像から要約された映像セグメント（黒縁：繰返し動作映像，灰縁：一般的な動作映像，白縁：状態映像）
Fig. 8 Video segments abstracted from a cooking video (Black frame: Repetitious motion, Gray frame: Cooking motion, White frame: Appearanece of food).

動きに基づく重要部分検出手法によって検出された料理や食材の状態部分の末尾を抽出した．これらを時系列に沿って各部分 2 秒間の映像を結合し，要約映像を生成した．なお，動きに基づく重要部分検出手法単独による要約映像の評価を行うため，3.4 節と同様に，カット検出，ショット分類は理想的に行われたものとした．

本アプリケーションによる要約の一例を**図 8** に示す．各フレームは要約に含まれる映像セグメントの代表フレームである．図 8 において，黒縁のものが繰返し動作部分，灰縁のものがその他の動作部分，白縁のものが状態部分である．

図 8 において，繰返し動作 (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

表 4 映像要約生成における抽出セグメント数
Table 4 Number of extracted segments in abstraction.

	レシピ数	繰返し	一般動作	状態	平均要約率
番組 1	4	23	23	18	約 1/11
番組 2	2	8	23	7	約 1/10
番組 3	3	8	70	2	約 1/ 9
全体	9	39	116	27	約 1/10

表 5 抽出セグメント数による要約映像の比較
Table 5 Comparison in number of extracted segments.

	Seg_H	Seg_M	Seg_C	再現率	適合率
レシピ 1	12	20	11	92%	55%
レシピ 2	13	24	13	100%	54%
レシピ 3	11	16	10	91%	63%
全体	36	60	34	94%	57%

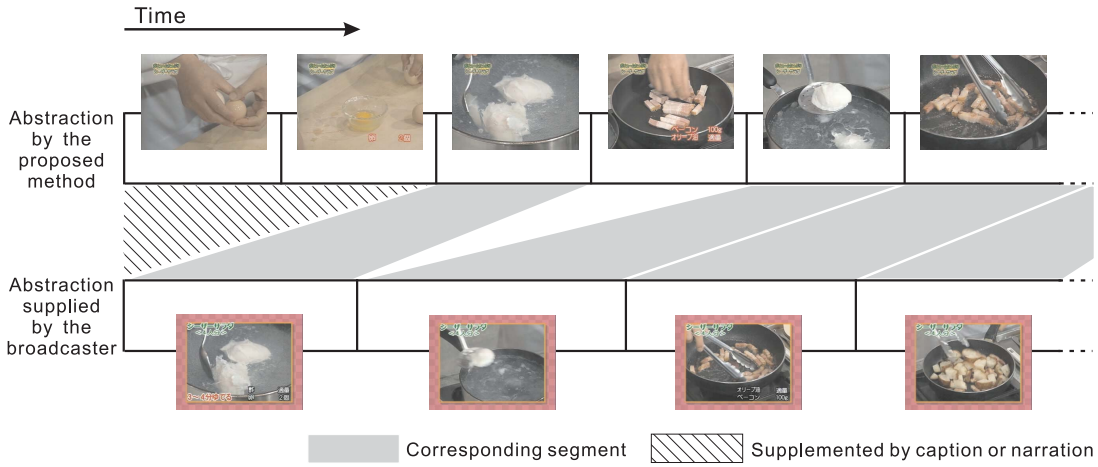


図 9 要約映像の比較
Fig. 9 Comparison of abstracted cooking videos.

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．

放送局の異なる複数の料理番組 9 レシピに対し，以上の手法を適用し要約映像を作成した際に，抽出されたセグメント数を番組ごとにまとめたものを**表 4** に示す．作成された要約映像は，時間的に見ると，元の映像と比べて 1/8 から 1/12 に短縮され，なおかつ表 4 のように番組ごとに傾向の差はあるものの，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．

4.2 要約映像の評価

表 4 に示した番組 3 については，番組の最後に「本日のおさらい」という形で，番組制作者の用意した要約映像が存在する．そこで，番組 3 の 3 レシピ分を対象とし，(1) 番組制作者による要約映像と，(2) 提案手法による自動要約映像を比較した．

まず，要約映像の長さは，(1) では 40 秒と固定長であるのに対し，(2) では，3 レシピ分の平均が約 53 秒であり，後者の方が少し長かった．要約映像の時間については現時点では特に考慮していないので，今後，各々のセグメントに最適な時間を設定することや，ユーザの熟練度に応じて変化させることなどの改良をしていくことが必要である．

次に，映像内容を比較した結果を**表 5** に示す．提案手法による自動要約映像は音声を考慮していないので，視覚的内容のみを比べている．(1) 中のセグメント数を Seg_H，(2) 中のセグメント数を Seg_M，また，両者が一致するセグメント数を Seg_C，再現率は Seg_C/Seg_H，適合率は Seg_C/Seg_M とする．また，(2) 中には，同じ視覚的内容を表すセグメントが重複している場合があるが，比較の際には，それらをまとめて 1 つのセグメントと見なした．この比較の例を**図 9** に示す．

表 5 から，再現率が高いことが分かる．つまり提案手法によって，番組制作者が要約に含めたものの大部分を抽出できている．実際，抽出できなかったセグメントは 2 つのみであった．

一方，適合率はあまり高くないが，その 1 つの原因として，ここで比較した要約映像は，一度番組を見た視聴者が「おさらい」という形で見るためのものであり，本研究で目的とする調理の全体的な流れを視覚的・

表 6 抽出セグメント数による要約映像の比較
（字幕およびナレーションによるセグメントを含む）
Table 6 Comparison of extracted segments (including telop and narration as segments).

	Seg_H	Seg_M	Seg_C	再現率	適合率
レシピ 1	17	20	16	94%	80%
レシピ 2	18	24	18	100%	75%
レシピ 3	12	16	11	92%	69%
全体	47	60	45	96%	75%

直感的に理解するのに十分である映像とは若干目的の異なるものであることがあげられる．

また，提案手法で過剰に抽出した部分の多くは，食材を見せている部分や「切る」・「入れる」といった基本的動作の部分であり，重要度は低いものの，調理の全体的な流れを丁寧に視覚的に表現するためには必要な部分であった．同時にこれらの部分のいくつかは，番組制作者による要約映像において映像セグメントとしては現れないが，新たに挿入された字幕やナレーションによる説明で補われていた部分でもあった．

そこで，**表 6** に人手により字幕やナレーションによる説明部分も 1 セグメントとして数えた結果を示す．これにより，番組制作者による要約映像における字幕やナレーションまで含めると，比較的良い適合率が得られることが示唆された．

5. おわりに

本研究では，動きに基づく料理映像の自動要約手法を提案した．我々は料理映像要約の目的を，調理の全体的な流れを視覚的・直感的に理解するのに十分である映像を作成することとし，対象に固有の特徴を考慮することで，意味的内容に立ち入った精度の高い映像要約システムの構築を目指した．

料理映像においては，画像全体の動きの激しい「調理動作部分」と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し，評価実験によりその有効性を示した．

また，局所領域の輝度値の時間的周期性に着目した手法[3)]を用いて，調理動作の中でも特に重要な動作の 1 つとして繰返し動作を検出し，これら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 1/12 の時間に短縮できた．さらに，本手法により自動生成した要約映像の一部を，番組制作者によって用意された要約映像と比較することにより，提案する自動要約手法の有効性を確認した．

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．家庭でのレシピ選びなどに利用すれば，1 本あたり数十秒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．

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，要約率を可変にしたりすることなどが考えられる．そのためには，繰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．

謝辞 本研究の一部は，科学研究費補助金（基盤研究（B）(2)）「料理映像を題材とするマルチメディア統合システムの提案とその応用」（課題番号：14380173）の支援を受けて行われた．また，本研究に関して有益な助言をいただいた国立情報学研究所の佐藤真一助教授に感謝いたします．

参考文献

1) Hamada, R., Ide, I., Sakai, S. and Tanaka, H.: Associating cooking video with related textbook, *Proc. ACM Multimedia 2000 Workshops*, pp.237–241 (2000).
2) 三浦宏一，浜田玲子，井手一郎，坂井修一，田中英彦：料理映像の構造解析による手順との対応付け，第 62 回情報処理学会全国大会，Vol.3, pp.31–32 (2001).
3) 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，PRMU2001-29 (2001).
4) 益満 健，越後富夫：映像重要度を用いたパーソナライズ要約映像作成手法，信学論（D-II），Vol.J84-D-II, No.8, pp.1848–1855 (2001).
5) 河合吉彦，馬場口登，北橋忠宏：個人適応を指向したスポーツ要約映像の生成法，信学技報，PRMU2000-171 (2001).
6) Christel, M., Smith, M., Taylor, C. and Winkler, D.: Evolving video skims into useful multimedia abstractions, *Proc. ACM Conf. Human Factors in Computing Systems '98*, pp.171–178 (1998).
7) Lienhart, R., Pfeiffer, S. and Effelsberg, W.: Video abstracting, *Comm. ACM*, Vol.40, No.12, pp.55–62 (1997).
8) Beauchemin, S.S. and Barron, J.L.: The computation of optical flow, *ACM Computing Surveys*, Vol.27, No.3, pp.433–467 (1995).
9) Horn, B. and Schunck, B.: Determining optical flow, *Artif. Intell.*, Vol.17, pp.185–203 (1981).
10) 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 (1994).

(平成 14 年 9 月 4 日受付)
(平成 15 年 3 月 28 日採録)

(担当編集委員 加藤 晃市)

三浦 宏一

平成 13 年東京大学工学部電子情報工学科卒業．平成 15 年同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．

浜田 玲子（正会員）

平成 10 年東京大学工学部電子情報工学科卒業．平成 12 年同大学院工学系研究科電気工学専攻修士課程修了．平成 15 年同専攻博士課程修了．博士（工学）．現在同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．平成 14 年本会第 63 回全国大会奨励賞受賞．電子情報通信学会会員．

井手 一郎（正会員）

平成 6 年東京大学工学部電子工学科卒業．平成 8 年同大学院工学系研究科情報工学専攻修士課程修了．平成 12 年同研究科電気工学専攻博士課程修了．博士（工学）．同年より国立情報学研究所助手．平成 14 年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，映像理解，統合メディア処理に興味を持っている．平成 8 年本会第 51 回全国大会奨励賞受賞．電子情報通信学会，人工知能学会，IEEE Computer Society，ACM 各会員．

坂井 修一（正会員）

昭和 56 年東京大学理学部情報科学科卒業．昭和 61 年同大学院工学系研究科情報工学専門課程修了．工学博士．同年工業技術院電子技術総合研究所入所．この間平成 3 年〜4 年，米国マサチューセッツ工科大学招聘研究員，平成 5 年〜8 年 RWC 超並列アーキテクチャ研究室室長．平成 8 年〜10 年筑波大学電子・情報工学系助教授．平成 10 年東京大学大学院工学系研究科助教授，平成 13 年より同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，マルチメディア等の研究に従事．平成 2 年本会論文賞，平成 3 年日本 IBM 科学賞，平成 7 年市村学術賞，ICCD Outstanding Paper Award 等受賞．電子情報通信学会，人工知能学会，IEEE，ACM 各会員．

田中 英彦（正会員）

昭和 40 年東京大学工学部電子工学科卒業．昭和 45 年同大学院工学系研究科博士課程修了．工学博士．同年同大学工学部講師．昭和 46 年同助教授．昭和 62 年同教授．平成 13 年より同大学院情報理工学系研究科教授・研究科長．この間昭和 53 年〜54 年米国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」．本会フェロー．電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IEEE，ACM 各会員．