

料理映像の特徴を利用した要約手法の検討

三浦 宏一[†] 浜田 玲子^{††} 井手 一郎^{†††} 坂井 修一[†] 田中 英彦[†]

[†] 東京大学大学院 情報理工学系研究科 〒113-0033 東京都文京区本郷 7-3-1

^{††} 東京大学大学院 工学系研究科 〒113-8656 東京都文京区本郷 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: ^{†,††}{miura,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp, ^{†††}ide@nii.ac.jp

あらまし 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、索引付けや知的構造化の研究を行っている。本稿では、特に料理映像の自動要約手法の検討を行う。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素材の状態を示す部分が特に重要である。これらは画面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を提案し、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要約映像を作成し、手法の有効性を確認した。

キーワード 料理映像、映像要約、動き検出、知的構造化、重要部分抽出

Abstraction Method Using Characteristics of Cooking Videos

Koichi MIURA[†], Reiko HAMADA^{††}, Ichiro IDE^{†††}, Shuichi SAKAI[†], and Hidehiko TANAKA[†]

[†] Graduate School of Information Science and Technology, The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

^{††} Graduate School of Engineering, 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: ^{†,††}{miura,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp, ^{†††}ide@nii.ac.jp

Abstract We propose a method to abstract cooking videos. We define cooking video abstraction as making videos sufficient to understand general cooking procedures visually and intuitively. To make an abstracted video, important subshot segments need to be extracted from the original video. Important segments in a cooking video are cooking motions and appearances of foods, since visual information that represents essential cooking operation is exceptionally important. These segments have typical motion-related features. Therefore, the proposed method refers to motion in the image detected as an optical flow. An experiment to detect important segments from motion-based structural analysis showed high detection rate and promising results. We assembled the important segments, and the resultant abstracted video showed the effectiveness of the proposed method.

Key words cooking video, video abstraction, motion detection, video structure analysis, important segment detection

1. はじめに

近年の情報通信技術の進歩に伴い、テレビ放送やインターネットなどを通じて提供されるマルチメディアデータは増大の一途をたどっている。このような大量のデータを収集・整理し、

効率良く保存・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おり、これに関する様々な研究が盛んに進められている。

しかしながら映像には様々な種類があり、種類ごとに映像的特徴も視聴者の見る目的も様々である。そのため汎用性の高い

映像解析技術は解析内容が限定されてしまい、個々の映像内容に沿った解析を行うことは困難である。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する必要がある。

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に着目し、映像やテキスト解析、索引付けなどの研究を行っている[1]～[4]。これらの研究では、対象を料理映像に限定し、対象に関する知識を最大限に活かすことで、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、実用的なシステム構築の実現を目標としている。

また、料理は日常的に行われ、しかも豊富な知識と熟練を必要とする複雑な作業であり、従来から調理支援の需要は存在してきた。さらに今後の家庭内への計算機の進出に伴って、このように解析、索引付けら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ものと考えられる。

本稿では、特に料理映像の自動要約を目的とした手法の検討を行う。料理映像の自動要約が実現すれば、映像による料理レシピの閲覧の他に、効率的な料理映像データベースの検索や閲覧などの応用が考えられる。

以下、2.において様々な映像要約手法を紹介し、料理映像の要約が目指すところについて述べる。3.では料理映像の特徴について述べ、4.において動き検出による映像解析手法について提案する。そして5.で提案手法による重要部分の検出実験とその評価を行い、6.において実験結果を利用した要約映像を作成する。7.では今後の方針を述べ、最後に8.でまとめとする。

2. 映像要約手法

2.1 関連研究

映像要約を自動生成する手法は、これまでも様々な研究が行われている[5]～[10]。

Informedia skimming プロジェクト[5]では、音声の書き下しであるトランスクリプトから TF-IDF 法によってキーワードを抽出し、それに対応する映像部分が重要であるとした。さらに、顔検出やカメラワークの検出、音量の大きさなどから重要シーンを決定して映像要約を作成している。MoCA プロジェクト[7]では、映画を対象として、主要人物のクローズアップ、爆発音や発砲音などから決定される特定のイベントを検出し、これらをつなぎ合わせることで予告編的な要約映像を作成している。これらの手法の多くは、映像のシンタクスを利用した手法であり、より実用的な精度を得るためには、さらに深く映像の意味内容を解釈する必要があると思われる。しかし、重要なシーンなどの意味的な内容を解釈するためには、対象映像を限定した高度な処理が必要であったり、現状では人手による補助が必要になることも多い。

一方、アメリカンフットボール映像を題材とした要約映像生成の研究[13]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。この手法は、利用しやすい外部データベースがある場合には有効であると思われる。

また、コンテンツの意味内容にまで立ち入った高度な情報抽

出処理を行わず、容易に抽出できる単純な特徴量のみを用いることで汎用性を求め、自動的かつ低コストに要約情報を抽出する研究も行われている。オーディオビデオコンテンツの符号化データ上の特性を用いて要約を行う研究[11]では、MPEGで符号化されたデータから動き情報、色配置、またオーディオのレベルを抽出し、閾値処理を行うことによって外部から与えられる要約情報長に応じた要約映像を作成している。さらに、個人の好みを反映した要約映像作成の研究[12]では、ユーザが過去に見たフレームと、早送りによりスキップしたフレームの特徴量からユーザの好みを学習し、それによる重要度を求め、要約映像を作成している。しかしながらこの研究でも人手による注釈データを用いて欠点を補う可能性が述べられているように、実際に精度の良い要約映像を作成するためには、映像の意味解釈が必要である。また汎用性を求めているが、ある程度映像の特性に依存してしまうことは避けられないようである。

我々は、精度よい要約映像を作成するためには、対象とする映像の種類を限定し、効率よく映像の意味内容を抽出する必要があると考える。そこで、料理映像を題材とし、対象に特化した処理を行うことで映像の意味内容にまで立ち入り、実用的で高精度な自動要約システムの実現を目指す。

2.2 料理映像を対象とした要約

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の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有効である。

そこで、我々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含み、なおかつ簡便に閲覧できるインタフェースを提供することを目的とする。個人の技量、好みの問題も考慮すべきことではあるが[12],[13]、まずは調理手順の重要な部分を集め、短い閲覧時間で調理の全体的な流れを視覚的・直感的に充分に理解できる映像を作成する。

これまでの映像要約研究においては、要約された映像は見づらいとの報告もある[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言われている。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],[11]、放送局などで制作される番組予告などの要約映像では音声は別のものが使われたりする。しかし、料理映像では音声がなくと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、本稿では、作成する要約映像には音声を含まないこととする。これにより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。そのような意味においても料理映像は、自動要約が効果的な題材であると考えられる。

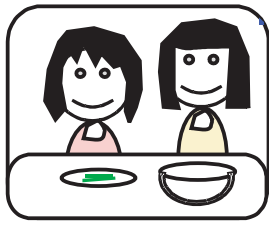


表3 実験に用いた閾値

Table 3 Thresholds.

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

表4 重要部分検出結果

Table 4 Result of important segment detection.

重要部分	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%

5.2 結果と考察

表4に重要部分の抽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。

表4に見られるように、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ている。

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）にもかかわらず画面が静止していたことであったが、いずれの場合も非常に稀であった。

また、カメラワーク（panning）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%（17の誤検出の内の7つ）を削減することができた。

6. 要約映像の作成

5.で行った重要部分抽出実験の結果より、簡単な要約映像を作成した。今回の手法としては、重要部分として検出された部分のうち、調理動作部分に関しては最初の2秒間を、料理や食材の状態部分に関しては最後の2秒間を取り出し、それらを単純に時系列に沿って結合した。

要約結果の例を図6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す。図6において、灰色の縁のものが調理動作部分、白色の縁のものが料理や食材の状態部分である。

図6において、調理動作（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（14）も同様に「なべを揺すって味をからませる」という動作で、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。また、（2）は「皮をむく」（10）、（11）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）、（9）、（12）、（15）は、調理や食材の状態部分である（15）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。

以上に示したような手法で作成された要約映像は、元の映像と比べて1/8から1/10の時間に短縮され、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を示すものであると考えられる。

7. 今後の方針

本章では、映像の内容解析をより深く行うために、現在考えているいくつかの事柄について述べる。

7.1 動きの分類

4.2で提案した手法では、単純に全体的な動きの大きさを利用し、調理動作部分を検出したただけであった。しかし、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。したがって今後は、動きが起こっている画面上の位置、また動き方なども利用して、調理動作を分類し、重要度を決定する。

調理の中心となる動作の多くを含む繰り返し動作[4]を含め、さらなる動作の分類手法を検討する予定である。

7.2 字幕検出

映像の解析において、動き以外で重要な手がかりとなりそうなものに、字幕（open caption）がある。

料理映像における字幕は図7に示すように（a）料理のコツや（b）使用する材料を示し、調理の手順を理解する上での重要な情報を与えている。特に本研究のように音声を用いない要約映像を作成する場合には、字幕は内容理解の大きな助けとなると考えられる。

映像から字幕を検出する研究は、[5],[17]などでなされており、これらを参考に実装する予定である。なお、字幕は料理番組の種類によってその出現の仕方が違うことも考えられるので、その扱いには注意をする必要がある。

8. おわりに

本稿では、料理映像を対象とした自動要約手法の検討を行った。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充分である映像を作成することとし、対象に特化することで、意味的内容に立ち入った精度の高い映



(a)

(b)

図7 料理映像中の字幕（open caption）

Fig. 7 Open captions in cooking video.

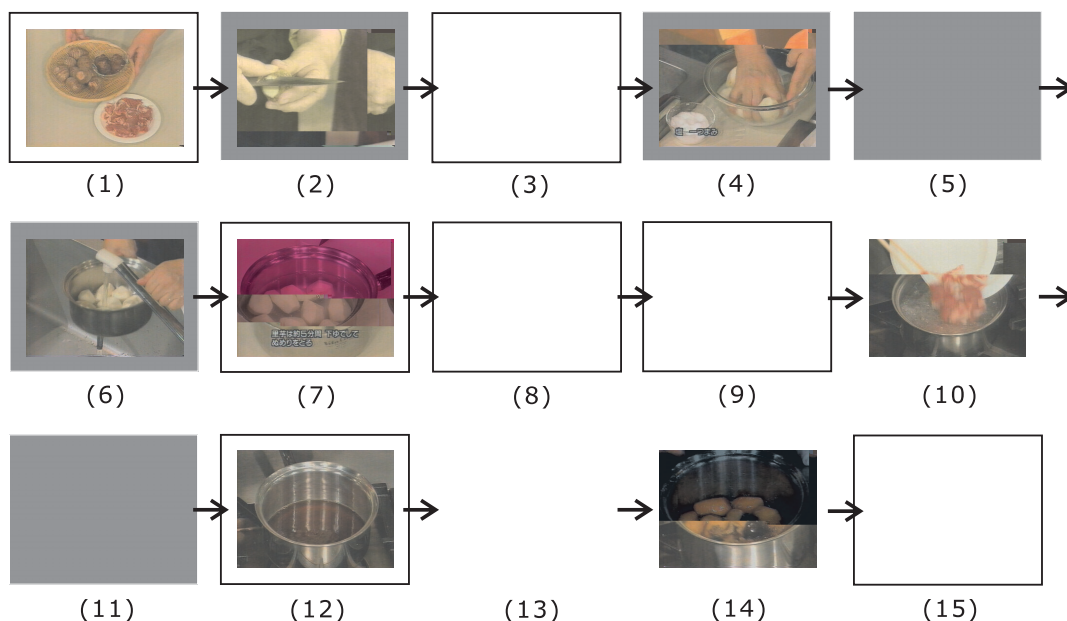


図6 料理映像から要約された映像セグメント

(灰色の縁：調理動作映像，白い縁：料理や食材の状態映像)

Fig. 6 Video segments abstracted from a cooking video

(Gray frame: Cooking motion, White frame: Appearance of food).

像要約システムの構築を目指している。

まず，料理映像の特徴から，視覚的に重要であるのは手元ショットであるとし，さらにその中でも調理動作部分と料理や食材の状態部分が重要であるとした。さらに，これらの重要部分は画面全体の動きの激しさと関連があることに着目し，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，この手法により検出された重要部分を単純につなぎ合わせることで要約映像を作成した結果，要約映像は調理手順の内容を保ちつつ，元の映像の 1/8 から 1/10 の時間に短縮することができた。

今後の課題としては，7. で述べたような手法を用い，さらに深く映像を解析していくことが挙げられる。また，今回は要約映像を作成する際に，重要部分を単純に 2 秒間ずつ抽出しただけであったが，要約手法の詳細についても検討する必要がある。さらに，要約映像自体の評価法も大きな課題である。

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] 浜田玲子, 井手一郎, 坂井修一, 田中英彦, "料理テキスト教材における調理手順の構造化," 信学論 (D-II), vol.J85-D-II, no.1, pp.79-89, Jan. 2002.
- [3] 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 回情処学全大, no.6R-9, vol.3, pp.31-32, Mar. 2001.
- [4] R. Hamada, S. Satoh, S. Sakai, and H. Tanaka, "Detection of important segments in cooking videos," Proc. IEEE Workshop on CBAIVL 2001, pp.118-123, Dec. 2001.
- [5] M. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," Proc. ICCV98, pp.61-70, Jan. 1998.
- [6] M. Christel, M. Smith, C. Taylor, and D. Winkler, "Evolving video skims into useful multimedia abstractions," Proc. CHI'98 Conf. Human Factors in Computing Systems, pp.171-178, April 1998.
- [7] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," Commun. ACM, vol.40, pp.55-62, Dec. 1997.
- [8] 森山 剛, 坂内正夫, "ドラマ映像の心理的内容に基づいた要約映像の生成," 信学論 (D-II), vol.J84-D-II, no.6, pp.1122-1131, June 2001.
- [9] 加藤和也, 吉高淳夫, 平川正人, "文脈を考慮に入れた映画の要約作成," 情処学研報, AVM-36-5, pp.25-30, March 2002.
- [10] H. Sundaram and S. Chang, "Constrained utility maximization for generating visual skims," Proc. IEEE Workshop on CBAIVL 2001, pp.124-131, Dec. 2001.
- [11] 菅野 勝, 中島康之, 柳原広昌, "映像の特徴に応じた AV データからの自動要約抽出方式に関する検討," 情処学研報, AVM-35-5, pp.25-30, Dec. 2001.
- [12] 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, Aug. 2001.
- [13] 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, pp.83-90, Jan. 2001.
- [14] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," ACM Computing Surveys, vol.27, no.3, pp.433-467, 1995.
- [15] B.K.P. Horn and B. Schunck, "Determining optical flow," Artif. Intel., vol.17, pp.185-203, Aug. 1981.
- [16] 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, 1994.
- [17] 新居啓之, 桑野秀豪, 倉掛正治, 杉村利明, "映像中のテロップ表示フレーム検出方法," 信学論 (D-II), vol.J83-D-II, no.6, pp.1477-1486, June 2000.