料理映像における繰り返し動作のスポッティング手法

浜田 玲子 † 佐藤 真一 ‡ 坂井 修一 †† 田中 英彦 ††
† 東京大学 工学系研究科, ‡ 国立情報学研究所
†† 東京大学 情報理工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1
Tel: 03-5841-7413
{reiko|sakai|tanaka}@mtl.t.u-tokyo.ac.jp, satoh@nii.ac.jp

一般に、映像の索引付けの単位としてはしばしばショットが利用されるが、より正確な索引付けにはさらに小さい重要部分のスポッティングが必要となる。料理映像においては、様々な動きの中から調理動作を認識することが重要であるが、「切る・混ぜる」など、重要な動作の多くは基本の動作要素の繰り返しで実現されることが多い。我々は、このような「繰り返し」に注目し、繰り返し動作を検出することで、料理映像における重要映像の抽出を行なう。本手法では、対象認識や追跡といった複雑な処理を回避し、単純な信号処理によって動作検出を実現した。また、本手法を利用した様々な応用について検討する。

料理映像, 映像の索引付け, 動作検出, 重要シーン検出, 映像要約

Detection of Repetitious Motions in Cooking Video

Reiko Hamada† Shin'ichi Satoh‡ Shuichi Sakai†† Hidehiko Tanaka††
†Graduate School of Engineering,The University of Tokyo
‡National Institute of Informatics
††Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
Tel: +81-3-5841-7413
{reiko|sakai|tanaka}@mtl.t.u-tokyo.ac.jp, satoh@nii.ac.jp

Generally, shots are finest segments for video indexing. However, it is necessary to extract smaller important segments to index the video more precisely. For cooking video indexing, recognition of cooking motions is important. And, these cooking motions tend to repeat same actions some times, such as cutting, mixing or tossing something. In this paper, we are aiming at extraction of important scene through detection of such repetitious motions. And applications using our method are discussed.

Cooking Video, Video Indexing, Motion Detection, Video Abstraction

1　はじめに

映像技術の進歩に伴い、テレビやビデオ、 WWW など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。そこで近年は、これら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に関する研究が盛んに進められている。ここで我々は、料理映像に着目した映像解析および索引付けなどの研究を行なっている [4]。

料理映像には、多くの場合付随するテキスト教材が存在するが、映像にはテキストでは表現しきれない様々な情報を含んでおり、特に料理手順の理解のためには視覚情報が非常に有効である。料理映像に索引付けを行なうことにより、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構築など、様々な実用的なアプリケーションへの応用が可能である。今後は、家庭内への計算機の進出に伴い、このような索引付けされた料理映像や料理レシピの検索に対する需要は高まっていくものと考えられる。さらに、将来は電子化された料理レシピを利用したスマートキッチンや、自動料理システムなどへの応用も考えられる。

一般に映像の索引付けを行なう際には、まず映像を細かい処理単位に分割する。しばしば用いられるのはカメラの切り替え点であるカットで区切られたショットに分割する方法である。カットは画像的な不連続点であるので、信号処理的に高い精度で検出することができる。次に、画像、音声、テキスト（字幕）などの解析を行ない、各ショットの内容を推測して索引をつけるのが一般的である [1, 2]。

しかし、料理映像には一般的な映像と異なる様々な特徴がある。そのため、映像をショットなどの映像単位に分割し、これらに索引をふるという従来の索引付け手法では、料理映像の構成に対応することができない。料理映像の特徴とその問題点、またこれに対する手法についての詳細は次章で説明する。

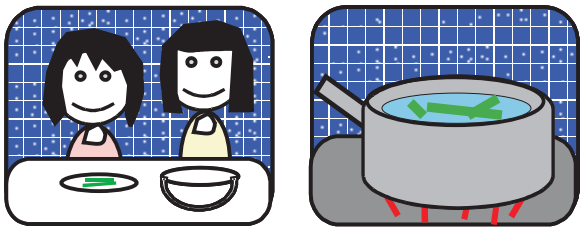
2　料理映像の特徴

本章では、料理映像の特徴、およびそれに適した索引づけ手法について検討する。

図 1に示す通り、料理映像におけるショットは大きく (1) 人物ショット、 (2) 手元ショットに分類される。

人物ショットはスタジオのほぼ全体が映され、調理人やその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、人物ショットから料理に関して視覚的な知見を得ることはできない。

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、視覚的にも重要である。しかし、ショットの中にさらに構造があり、動作準備から一連の複数動作、そして動作後の料理の様子などから成る。調理において中心となる重要な映像を含む一方で、動作と動作の間などは比較的冗長である。



(a) 人物ショット　　(b) 手元ショット

図 1: 料理映像におけるショット分類

ここで料理映像の構成を図 2に示す。図に示す通り、料理映像においては人物ショットと手元ショットがほぼ交互に出現する。そして重要と考えられる手元ショットの中に、さらに重要な部分と比較的冗長な部分が含まれる。

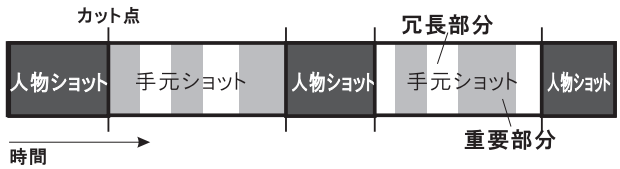


図 2: 料理番組における映像構成の例

このような構成の料理映像において、一般的な手法のように映像をショットに分割してそれぞれに索引をつけたとしても、人物ショットは映像の観点からほとんど重要性がない。また手元ショットに関しても、その中に含まれる構造を考慮せずに重要な映像と冗長な映像を分離することはできない。従って、このような索引付け方法は料理映像に対しては適切ではない。

ここで、手元ショット内の重要部分には、カットのような明らかな画像変化があるわけではない。そこで、映像を分割して各部分に索引を付けるのではなく、重要な部分を解析して直接抜き出し、これを索引とする方が効率がよいと考えられる。

すなわち、従来のショットによる映像の分割は信号処理的な手法であったが、料理映像においては内容解析に基づく分割と索引付けが必要であると考えられる。

次に、料理映像における重要部分について考察する。

料理映像は一種の教材であるが、ほとんどの場合、同じレシピを扱ったテキスト教材が付随している。従って、調理の手順そのものについてはテキストを参照することができる。実際に、料理映像においては、手順のうち自明な部分などは省略されることも多い。しかし、料理映像の価値は、テキストでは表現しきれない視覚的な情報を示すことにあると考えられる。

従って、料理映像の中でも特に重要なのは、比較的複雑な調理動作、および料理や食材の状態に関する映像である。このうち調理動作について、実際の料理映像を参照して検討した結果、調理の中心となる動作の多くは繰り返しの動作であるということがわかった。具体的には、「切る」「あえる」「こねる」「混ぜる」「泡立てる」など、様々な対応する動詞がある。また、動詞からみて繰り返し動作であるとは予測できないような動きでも、実際には繰り返されることがある。これは大事な動作は確認のために数回繰り返すことが多いためと考えられる。図 3に繰り返し動作の例をいくつか示す。本研究では、このような動作の時間方向の周期性に着目した重要部分検出を行なう。

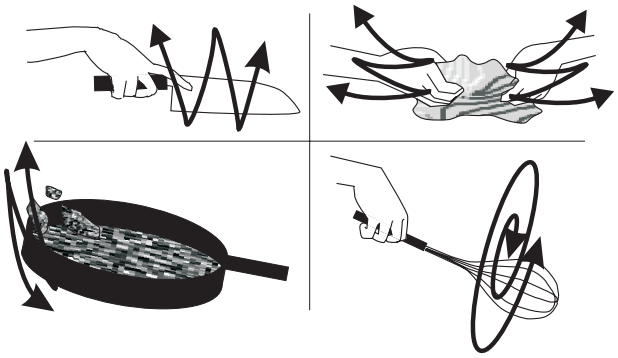


図 3: 調理中の繰り返し動作の例

3　繰り返し動作の検出

本章では、繰り返し動作の検出手法について述べる。

一般に、動作認識手法においては、肌色の検出などによって手領域を認識・追跡し、その軌跡から動作を認識することが多い [6, 5]。しかし、料理映像においては、手元ショットであっても必ずしも手が映るわけではない。そこで、繰り返し動作の検出のためには手元の道具（菜箸など）の振動を認識する必要がある。しかし、このような道具の形状や色には様々なものがあり、その特徴を特定するのは困難である。従って、本研究においては肌色など特定の色を利用する手法は適切ではない。

図 3から、繰り返し動作映像においては、映像の局所領域上を対象物が往復するということがわかる。そこで、本研究では、時間周波数解析によって微少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の検出を行なう。以下にこの手法を説明する。

まず、各フレームを小さなブロックに分割する。図 4に示す通り、各ブロックは 3×3 ピクセルから成る小さな正方形である。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。

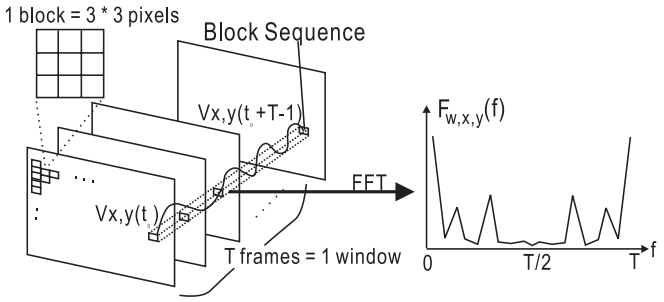


図 4: 映像の分割方法と FFT の適用

図 4に示すように、x, y を一定の位置に固定して t を動かすと、$V_{x,y}(t)$ は位置 x, y における、ブロックの平均輝度値の時間変化を示す。特に、繰り返し動作によってそのブロック上を対象が往復している場合、$V_{x,y}(t)$ は周期的な値の変化を見せるはずである。このように $V_{x,y}(t)$ が周期的な変化を見せるブロック列が複数ある場合、その時間帯には映像中に周期的な振動があると考えられる。

そこで、画像中のすべての x, y におけるブロック列 $V_{x,y}(t)$ にそれぞれ FFT(Fast Fourier Transform) を適用し、その周期性を調べることにした。FFT を適用する時間方向の範囲は、大きさ（フレーム数）$T = 2^n$ の時間窓内とする。この窓を t 方向に T_{step} フレーム単位で移動していくことにより、各時点での振動の有無を調べる。以下に FFT の式を示す。

$$F_{w,x,y}(f) = \left| \frac{1}{\sqrt{T}} \sum_{t=t_0}^{t_0+T-1} V_{x,y}(t) W^{ft} \right|^2 \qquad (1)$$

式 1において、$W^{ft} = e^{-j\frac{2\pi}{T}}$、また $F_{w,x,y}(f) = F(f)$ は周波数 f におけるパワーである。w は窓番号であり、FFT を適用している時間帯を示す。

$V_{x,y}(t)$ 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 $f_0 \leq f < f_0 + N$ とする（図 5）。

まず、範囲内のパワーの総和は式 2で示すように定義される。

$$Power_{w,x,y} = \sum_{f=f_0}^{f_0+N-1} F(f) \qquad (2)$$

次に、範囲内での $F(f)$ の最大値を計算する。また、最大値を与える周波数を f_p とする (式 3)。

$$F(f_p) = \max_{f_0 \leq f < f_0+N} F(f) \qquad (3)$$

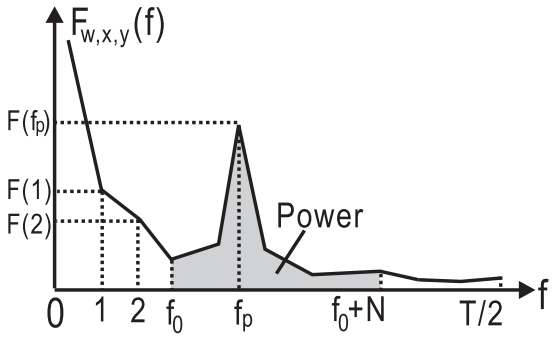


図 5: FFT グラフ

ここで、$F(f_p)$ は最大値ではあるが、ピークといえるとは限らない。そこで $F(f_p)$ 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 4)。

$$F_{peak}(w,x,y) = \frac{F(f_p) \times (N-1)}{\sum_{f=f_0, f\neq f_p}^{f_0+N-1} F(f)} \tag{4}$$

また、低周波数におけるパワー $F(1)$ や $F(2)$ よりも $F(f_p)$ が小さい場合は、本来周期的ではないゆっくりした動作に含まれる繰り返しの断片を検出してしまう可能性があるため、これを排除する。そこで、その指標として $F(f_p)$ の $F(1)$ と $F(2)$ に対する比である R_1, R_2 をそれぞれ用意する。R_1, R_2 が十分に大きいとき、$F(f_p)$ は低周波帯のパワーより十分大きいことがわかる。

また、動きの周期性が曖昧な場合、グラフ全体のエネルギーは大きくなるが、ピークがあまり明確にならない。そこでピークの鋭さの指標 R_{sharp} を式 5のように定義する。

$$R_{sharp} = \frac{F(f_p) \times 4}{\sum_{f=f_p-2, f\neq f_p}^{f_p+2} F(f)} \tag{5}$$

以上までに説明した $Power$, F_{peak}, f_p, R_1, R_2, R_{sharp} の 6 つのパラメータは、FFT のパワーおよびピークの明確さの指標となっている。本研究ではこれらを利用して、映像中の動きの周期性の有無を調べる。

具体的には、2 点以上のブロックにおいて、以上の 6 パラメータがいずれも閾値以上の値を持つとき、その時の窓 w において繰り返し動作の検出を行なう。例えば $w = w_0$ のときに振動が検出された場合、その料理映像のうち、フレーム範囲 $T_{step} \times w_0 \leq t < T_{step} \times w_0 + T$ において繰り返し動作が存在することになる。

なお、このような条件を満たすブロックが 1 カ所のみである場合は雑音である可能性が高いため検出しない。また、フレームの上下左右の端は画像的に不安定で、雑音の振動を検出することがしばしば起こった。そのため、映像の端から 15 ピクセル（5 ブロック）分は考慮にいれないこととした。

4 評価実験

4.1 実験条件

前章で述べた方法に従い、評価実験を行なった。実験には、1 つの料理番組から取得した料理映像 16 レシピ分（合計約 69 分）を利用した。映像の特性を表 1に示す。

表 1: 実験データの特性

時間（合計）	69.1 分
レシピ数	16
ファイル形式	mpeg2
画像サイズ	360 × 240 pixels
フレームレート	15 frm/s

繰り返し動作の検出のためには、時間窓中で数回以上は動作が繰り返されている必要がある。そこで、窓サイズ T は経験的に 32 フレーム（約 2 秒）、また窓の移動ステップ T_{step} は 16 フレーム（約 1 秒）とした。また、対象とする周波数帯は $f_0 = 3, N = 12$ とする。周波数解析に関するパラメータをまとめて表 2 (a) に示す。また、前節で述べた通り、ある時点での窓 w において 2 つ以上のブロックで表 2 (b) に示す 6 つのパラメータが閾値より大きい場合、繰り返し動作を検出する。

表 2: 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

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

(b) 周期性解析パラメータ

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

4.2 結果と考察

実験結果の評価においては、対象の料理映像から、振動動作部分の映像を人手で抜き出し、これを正解とした。そして自動解析結果と人手による正解を照合することで、手法の評価を行なった。

評価実験の結果を表 3に示す。なお、人手による結果を 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 とすると、再現率は Ans_C/Ans_H、適合率は Ans_C/Ans_M である。

表 3に示す通り、本手法では誤検出が少なく、90% 以上の適合率で繰り返し動作を検出することがわかった。一方、検

表 3: 実験結果

Ans_H	Ans_M	Ans_C	再現率	適合率
33	26	24	72.7%	92.3%

出漏れが比較的多く、再現率は 73% 程度である。本手法による成功例、誤検出そして検出漏れの具体的な例を図 6に示す。

まず、図 6 (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要領を要する重要な動作である。また (b) のような、肌色に近い色の食材を扱う映像の場合も正確に検出することができた。

次に、図 6 (c),(d) に誤検出の例を示す。 (c) 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ているため、これを振動として誤検出した。また、 (d) では、画面上端に映る手が調味料を振っており、これを検出した。これは本来は規則的な繰り返し動作であるが、手動で正解を作成する際に、手の映っている領域が小さく、動作自体が目立たなかったため、正解に含められなかった。

最後に、図 6 (e),(f) に検出漏れの例を示す。 (e) は魚に小麦粉をまぶし、これを軽くはたく動作である。検出漏れの原因として、動き自体が小さく単発的であり、さらに手と食材についた小麦粉の色により、輝度値の変化が少なかったことが考えられる。 (f) はネギをゆっくりと炒る動作である。 (f) 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。

人手で正解を作成する際にも、特に図 6 (f) 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場面が多かった。実験の結果では、人間から見ても振動があいまいなものほど検出率は悪い。

本章における評価実験の結果が実用的であるかどうかは、本手法を適用するアプリケーションに依存するが、検出の傾向が比較的人間の感覚に近いため、多くのアプリケーションに適用可能であると考えられる。具体的なアプリケーションの例については、次章で述べる。

5 本手法の応用

本手法の実用性を確認するため、簡単な料理映像の要約アプリケーションを作成した。一般的に、要約された映像は見づらいという研究結果も報告されているが [3]、この原因は音声が断続的に途切れるためであるという。料理映像では視覚的な情報から動作や手順を知ることができるため、音声がなくても理解することができる。そのため、料理は映像要約が効果的な対象であると考えられる。

(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) 誤検出: 調味料を振る

(e) 検出漏れ: 小麦粉をはたく (f) 検出漏れ: ねぎを炒る

図 6: 本手法による繰り返し動作検出結果の例

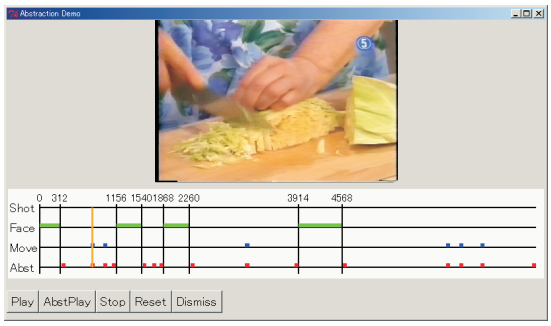


図 7: 料理映像要約アプリケーションの例

図 7に、料理映像要約のアプリケーションの例を示す。この例においては、要約から人物ショットを除くためにまずカット検出を行なう。そして、それぞれのショットから顔領域を検出することで、人物ショットと手元ショットに自動分類する [7]。

そして、各手元ショットにおいて、本手法によって振動が検出された部分、およびショットの最初と最後を 2 秒ずつ拾って要約とする。その手元ショットに振動のない場合は、ショットの真ん中をとる。なお、音は無音とする。

本アプリケーションでは、カット検出や振動動作検出の誤りなどによって、要約に余分や不足の映像が含まれることもあり得る。しかし、利用者は無意識のうちに要約に含まれて

いない映像を類推しながら観賞していると考えられる。そのため要約に多少の誤りがあっても、料理の内容はほぼ理解できることがわかった。

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。ある程度調理に熟練した利用者であれば、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、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、一つあたり数十秒〜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、本手法は現状の精度で様々な応用が考えられる。

6 今後の課題

今後の課題として、まず繰り返し動作の検出精度の向上が挙げられる。本手法はゆっくりした動作に比較的弱いため、ゆっくりした動きを検出した場合のみ動的に周波数解析窓の幅をのばしたり、画像方向の動作の広がりを参考にするなどの工夫が考えられる。

さらに、より精度の良い料理映像の索引付けのため、繰り返しではないが重要な動作の検出を検討する。このような動作の頻度は比較的少ないが、特に手順に特有の動作の中には繰り返しではないものがある。これらの動作には共通の特徴が少ないため、冗長な動作、例えば「机に皿を並べる」などの動作との区別は簡単ではない。しかし、重要な動作であれば映像の中心に比較的長時間映されるため、動作の重心の位置や移動速度などの利用を検討中である。

さらに、料理や食材の状態を示す映像も抽出する予定である。「揚げる」「盛り付ける」などの動作に関しては、動作自体よりも料理や素材の状態や外見が重要である。料理映像においては、調理動作前後の料理や食材の状態を、ほぼ静止して示す部分がある。この様な映像を抽出し、上で述べた重要動作の抽出を合わせると、調理動作と料理の様子から成る重要映像をほぼすべて抽出することができる。

応用としては、以上のように様々な重要部分を抽出することで、前節で述べた映像要約手法をより洗練することができる。また、映像とテキスト教材を統合してより細やかな索引付けを行ない、データベース化する。これにより、レシピ単位だけではなく、手順、動作、また料理の状態を示す映像など、細かい単位での検索・参照が可能となる。

さらに、将来はこれらの結果を実際の台所環境へ適用し、例えば何らかのセンサにより調理の進行度合いを知ることで、現在の手順の仕上がり予想図や、次の手順内容などを提示するスマートキッチンなどの応用が考えられる。

7 まとめ

我々は、料理映像の索引付けについて検討している。従来のショット単位の索引づけでは、料理映像の構造には対応することができないため、ショット内の重要部分を切り出すことで意味的な索引付けを行なう。

本稿では、代表的な重要動作として繰り返し動作に着目し、その自動検出手法を提案した。また、評価実験を通して本手法の有効性を示した。

本手法の応用としては、料理映像の要約アプリケーションを実装し、紹介した。今後の課題として、精度の向上、その他の重要部分の抽出、またこれらの結果を利用したテキストとの対応付けや検索などの応用の検討があげられる。

参考文献

[1] A. G. Hauptmann and H. D. Wactlar, “Indexing and Search of Multimodal Information,” Proc. ICASSP’97, April 1997.

[2] Y. Ariki, “Multimedia Technologies for Structuring and Retrieval of TV News,” News Generation Computing, Vol.18, No.4, pp.341-358, 2000.

[3] M. Christel, M. Smith, C.Taylor, and D. Winkler, “Evolving Video Skims into Useful Multimedia Abstractions,” Proc. of 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, Nov 2000.

[5] 西村 拓一, 向井 理朗, 野崎 俊輔, 岡 隆一, “白黒動画像からの形状特徴を用いたジェスチャのスポッティング認識システム,” 信学論 (D-II), Vol.J81-D-II, No. 8, pp.1812-1821, Aug 1998.

[6] 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, July 1997.

[7] 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づけ,” 第 62 回情処学全大, No.6R-9, Vol.3, pp.31-32, Mar 2001.