一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

HCGシンポジウム2015
HUMAN COMMUNICATION GROUP SYMPOSIUM 2015
HCG2015-D-9-3

調理者の視線運動パターンに基づく調理動作識別手法

井上 裕哉† 平山 高嗣††,† 道満 恵介†††,† 川西 康友† 井手 一郎†
出口 大輔††††,† 村瀬 洋†

† 名古屋大学大学院情報科学研究科
†† 名古屋大学大学院実世界データ循環学リーダー人材養成プログラム
〒 464–8601 愛知県名古屋市千種区不老町
††† 中京大学工学部 〒 470–0393 愛知県豊田市貝津町床立 101
†††† 名古屋大学情報戦略室 〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †inoueh@murase.m.is.nagoya-u.ac.jp, ††{hirayama,kawanishi,ide,murase}@is.nagoya-u.ac.jp

あらまし 本発表では調理者の視線運動パターンの分析によって調理動作を識別する手法を提案する．視線情報が人間の行動を理解するうえで重要であることから，頭部に装着した視線計測装置により視線運動データを取得し，調理動作識別に用いる．視線運動パターンの表現手法には視線情報に着目した行動認識において有用とされる N-gram の頻度ヒストグラムを採用する．従来は前フレームからの相対的な注視点移動のみが N-gram を構成する記号に用いられていた．これに対して本研究では，調理行動には調理対象への注視を持続させる振る舞いが頻繁に表れることから「停留」と，動作に対する集中度を考慮するため「瞬き」を表現する記号を加えた．そして，視線運動パターンが対象の調理動作である尤度を出力する SVR (Support Vector Regression) モデルを学習した．実験の結果，従来手法より 0.168 高い 0.856 の平均 F 値が得られ，提案手法の有効性を確認した．
キーワード 調理動作識別，視線分析，視線運動パターン，停留，瞬き，N-gram，SVR

Cooking Operation Classification Based on Analysis of Eye Movement Patterns

Hiroya INOUE†, Takatsugu HIRAYAMA††,†, Keisuke DOMAN†††,†, Yasutomo KAWANISHI†,
Ichiro IDE†, Daisuke DEGUCHI††††,†, and Hiroshi MURASE†

† Graduate School of Information Science, Nagoya University
†† Graduate Program for Real-World Data Circulation Leaders, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
††† School of Engineering, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota-shi, Aichi, 470–0393 Japan
†††† Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
E-mail: †inoueh@murase.m.is.nagoya-u.ac.jp, ††{hirayama,kawanishi,ide,murase}@is.nagoya-u.ac.jp

Abstract In this presentation, we propose a classification method of cooking operations by analyzing eye movement patterns. Since gaze information is important in understanding human behavior, we obtain it by a head–mounted device, and use it to classify cooking operations. We use the N-gram model known as effective in action recognition that focuses on gaze information. Conventionally, only relative movement from the previous frame was used as symbols for the N-gram. However, since in cooking, users pay attention to cooking ingredients and equipments, we consider fixation as a component of the N-gram. We also consider eye blinks which may reflect concentration. The proposed method estimates the likelihood of the cooking operations by Support Vector Regression (SVR) using frequency histograms of N-grams as explanatory variables. The effectiveness of the proposed method was confirmed through an experiment, which obtained the average F-score of 0.856, 0.168 higher than the conventional method.

Key words Classification of cooking behaviors, gaze analysis, eye movement pattern, fixation, blink, N-gram

1. はじめに

近年，料理教室が人気を博しており，調理技術の向上を目指す人が増えている．これに対して，食材や調理操作ごとに映像データベースを構築し，レシピの各調理手順に対応する映像を付与するなど，情報システムを導入することで初心者が調理操作などを視覚的に理解できるようにする支援が実用化されつつある [1]．このような情報システムによる調理者の支援を考える際には，調理者が「何をしているのか」，あるいは「次に何をしようとしているのか」という調理行動を理解する必要がある．従来研究では，台所上の固定カメラで撮影された調理シーンの映像特徴に基づいて調理動作が認識されている [2]．しかし，画像による認識は照明条件や調理器具などの環境的要因に影響されやすい．そこで本研究では，人間の行動が認知・判断・動作の過程を経てなされると考え，調理状況の視覚的な注意や認知を重要視し，図 1 に示すような調理動作による視線運動の違いを分析する．視線運動は意思や精神状態といったような人間の内部状態を反映し [3]，その視野の周辺には行動に関連する情報が含まれる [4] とされる．また，これらの視線情報を計測するためのウェアラブルセンサは、高性能化，小型化，低廉化により，一般人が日常的に使用できる環境が整いつつある．

視線情報の分析に基づいて調理者の行動を理解できれば，調理者の支援のほか，料理レシピへの画像掲載や調理映像の要約のために見どころを抽出できる．また，熟練者と初心者における視線運動の違いを抽出することで，調理動作のコツだけではなく，認知・判断のコツを形式化できる可能性もある．そこで，本研究ではそのような調理行動理解の前段階として，視線情報と調理動作の関係性を検証するために，視線運動パターンに基づいて調理の基本動作を識別することを目標とする．本研究で着目する視線運動パターンは視線方向の時系列遷移であり，例えば（右，右，下）や（上，左，下）などと表現される．調理の基本動作を含む短い区間での解析が可能であれば，それらを組み合わせた一連の行動で調理内容を把握したり，時間粒度が細かい調理技術の個人差を抽出する手助けになると考えられる．さらに，視線情報を用いた解析を行うことで，画像ベースの手法では困難であったセグメンテーションを行うことができる．そこで我々は視線情報の分析に基づいた調理行動理解に関する研究に取り組んでいる [5]．

関連研究としては Bulling らが視線運動を利用して数種類のデスクワークの識別を [6]，木谷らが頭部運動を利用して「ジャンプ」や「走る」などスポーツ行動の識別を行っている [7]．また大垣らは注意を向ける方向の小さな変化に関係する視線運動と，その大きな変化に関係する頭部運動を反映した 1 人称視点映像の Optical flow とを組み合わせて更に高精度にデスクワークを識別できることを示している [8]．しかし，本研究で対象とする調理動作とデスクワークとでは動作の性質が異なる．一例として，デスクワークでは視点が常に動いているのに対し，調理行動では調理対象への局所的な注視が持続するような振る舞いが頻出する．そこで本研究では，これらの従来手法を調理動作識別に適した特徴量を導入して改良することで，視線運動パターンの分析に基づいて調理の基本動作を識別する．

以下，2. で視線運動の分析に基づく調理動作識別手法について述べ，3. で提案手法の有効性を確認するための実験及びその結果と考察について述べる．最後に 4. で今後の課題について検討し，本報告を結ぶ．

(a) 切る（Cut）例　　(b) 混ぜる（Mix）例

図 1　調理動作による視線運動の違い．赤丸と赤線はそれぞれ，現在の注視点と過去の注視点の軌跡を表す．

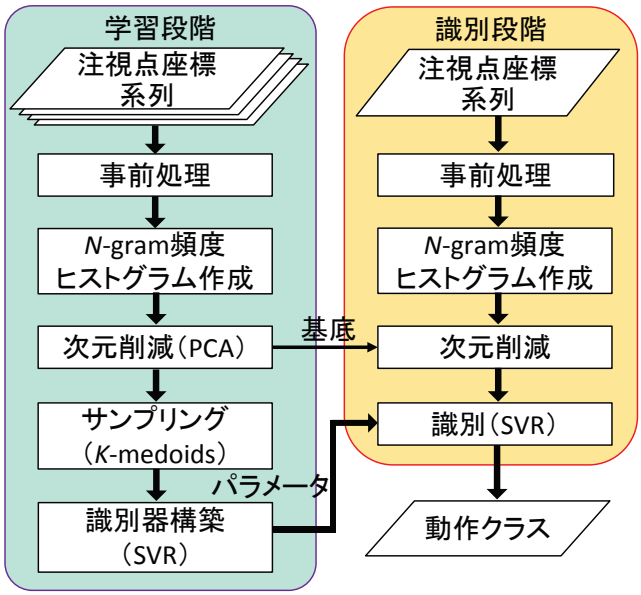

図 2　調理動作の識別手順

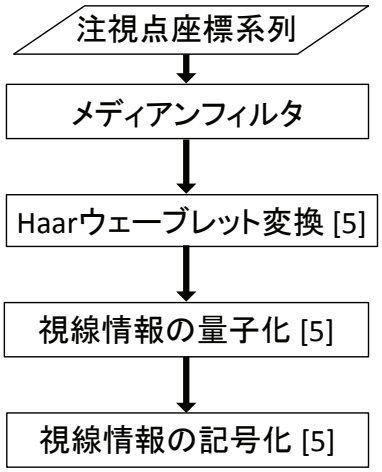

図 3　事前処理の手順

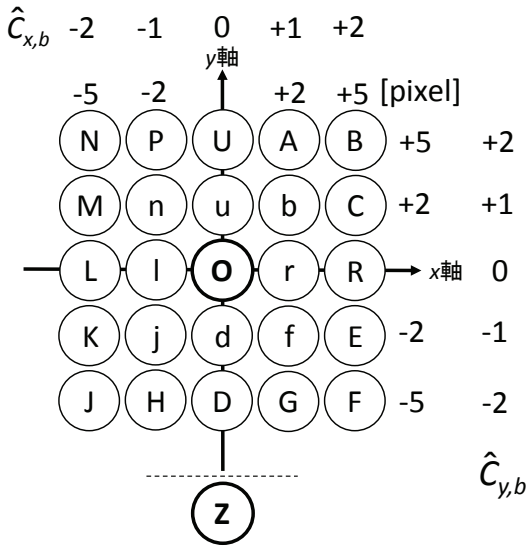

図 4　注視点の移動に応じて付与する記号

2. 視線運動パターンの分析に基づく調理動作識別手法

図 2 に提案手法の処理手順を示す．事前処理で視線情報を記号化し，そこから得られる N-gram のヒストグラムを特徴量として調理動作の学習・識別に用いる．更に図 3 に事前処理の手順を示す．計測した視線座標系列は雑音を含むため，まず x 座標系列，y 座標系列それぞれに対してメディアンフィルタを適用する．次にフレーム毎の座標値をフレーム間の移動量 $C = \{C_{x,b}, C_{y,b}\}_{b=1}^{T}$ に変換するために，文献 [6] と同様に以下の式 (1), (2) に示すスケールパラメータ α を固定した Haar ウェーブレットによる連続ウェーブレット変換 (CWT-SD: Continuous Wavelet Transform for Saccade Detection) を行う．

$$C_{x,b} = \frac{1}{\sqrt{\alpha}} \int \psi\Big(\frac{t-b}{\alpha}\Big) x_t \, dt \tag{1}$$

$$\psi(\beta) = \begin{cases} 1 & (0 \leqq \beta < \frac{1}{2}) \\ -1 & (\frac{1}{2} \leqq \beta < 1) \end{cases} \tag{2}$$

ここで $C_{x,b}$ は b フレーム目における，x 座標の前フレームからの移動量を表し，x_t は t フレーム目において計測された x 座標を表す．また，スケール α は視線計測の標本化周波数に依存する（以下の実験では 100 ms に相当するよう $\alpha = 6$ とした）．これを y 座標についても同様に計算する．更に大小 2 種類（負の閾値を含めれば 4 種類）の閾値 H と L を設けて式 (3) のように 5 段階に量子化を行ったうえで，量子化した x 座標系列，y 座標系列を統合して記号化する．

$$\hat{C}_{x,b} = \begin{cases} 2 & (H \leqq C_{x,b}) \\ 1 & (L < C_{x,b} \leqq H) \\ 0 & (-L < C_{x,b} \leqq L) \\ -1 & (-H < C_{x,b} \leqq -L) \\ -2 & (C_{x,b} \leqq -H) \end{cases} \tag{3}$$

現フレームに付与された記号は前フレームからの相対的な注視点移動を表す．注視点の移動方向と移動量に応じて付与する記号を図 4 に示す．図 4 中の記号において，小文字が小さな視線移動を表し，大文字が大きな視線移動を表す．こうすることで，Cut のような視線運動のスケールが小さな動作と共に，Mix のような視線運動のスケール大きな動作も扱うことができる．調理行動には調理対象への注視を持続させる振る舞いが頻繁に現れるが，文献 [6] で提案されている記号化方法ではそのような停留を考慮していない．そこで文献 [6] の記号に加えて視線方向の遷移がない原点（記号「O」）を設ける．また，調理動作には Cut のような常に高い集中を要する動作と，Mix のようなそれほど集中を要さない動作がある．そこで本研究では集中状態を反映するとされる瞬き [9] を考慮することで更なる識別精度向上を目指すために，瞬き記号「Z」を設け，合計 26 通りの記号のいずれかで表現する．なお，瞬きの検出方法については 3. で後述する．

以上の事前処理によって作成した記号列を視線運動として利用し，これから抽出される特徴が調理動作によって異なると想定する．

本研究ではこの記号列をある一定の時間幅の解析窓で区切り，各々に 1 つの調理動作が含まれるものと考える．解析窓はその幅を文献 [6] に合わせ 900 フレーム（15 秒）とし，1 秒の時間粒度で解析を行うため，60 フレーム単位に移動させる．ただし解析対象の時区間に続く次の 60 フレームの間で動作が変わる場合は，その動作が変わる時点までその解析窓を拡げる．そして解析窓ごとに視線運動 N-gram の頻度ヒストグラムを作成する．文献 [6] では頻度ヒストグラムの値から統計量を抽出したものを特徴量としているが，本研究では頻度ヒストグラム全体を特徴量とし，1-gram から N-gram までのヒストグラムを結合して用いる．ここで，このような頻度ヒストグラムをそのまま特徴ベクトルとすると高次元になってしまうため，主成分分析（PCA: Principle Component Analysis）を用いて次元数を削減する．

最後に調理動作ごとに SVR(注1) (Support Vector Regression) [10] モデルを構築する．この時，対象の調理動作に対しては大きな尤度を出力し，それ以外の調理動作に対しては小さな尤度を出力する（尤度の範囲は-1 から 1 とする）．このモデルを one-against-all の識別器と見なし，識別段階において識別器の尤度が 0 以上を示した時，特徴ベクトルが対象動作を行っているとする．この時，対象動作を行っている標本以外は負事例として扱うため，2 クラスの標本数に偏りが生じる可能

(注1)：本実験では lib-SVM に含まれる epsilon-SVR を使用した．

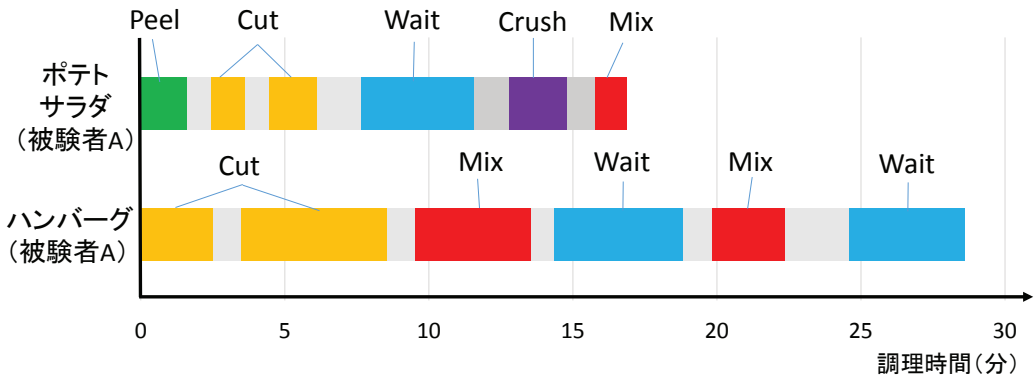

図 5　データセット中の調理動作の例

性が高い．そのため本研究では正事例と負事例の比に応じて，k-medoids クラスタリング [11] を応用して学習する負事例を間引く．この時 k の値は正事例の数とし，セントロイドの最近傍点を負事例として学習に用いる．

3. 実　　験

提案手法による調理の基本動作の識別可能性を検証するために，調理映像 7 本に対して調理動作の識別を行った．これらの調理映像は，4 本がハンバーグの，3 本がポテトサラダの調理過程を撮影したもので，調理動作として「Cut」,「Mix」,「Wait」,「Crush」,「Peel」の 5 種類を含む．各調理動作の定義は文献 [2] に従う．このデータセットに含まれるレシピごとの調理動作の一例を図 5 に示す．

視線計測には NAC 社製 EMR-9 [12] を用いた．測定範囲は水平方向に ±40°，垂直方向に ±20° で，標本化周波数は 60 Hz である．そして水平方向分解能は 0.1°，垂直方向分解能は 1°，一人称視点映像の解像度は 640（H） × 480(V) である．2. で述べたように視線座標系列を視線記号列に変換し，前述の解析窓の幅で視線記号列を分割したところ，標本数は 7,880 となった．それぞれに上記 5 つの動作のいずれかが割り当てられる．このサンプルに対して映像単位での交差検定を行った．評価実験では動作ごとに学習した SVR モデルを用いて，テストデータが対象動作かそれ以外の動作のいずれかという 2 クラス識別を行い，F 値で評価を行った．F 値は正確性と網羅性の総合的な評価に利用される尺度とされ，以下の式で表現される．ここで precision は識別された結果に含まれる正解の割合を表す指標，recall は識別されるべき対象に対して実際に識別できた割合を示す．

$$F = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{4}$$

3.1 事 前 実 験

まず，未定義のパラメータを決定する必要がある．注視点の移動に応じて記号を付与する．図 4 中の記号において小文字と大文字，つまり小さな動きと大きな動きを適切に区別するためには，閾値を適切に設定することが重要である．ここで小さい動きの閾値 L は視線運動の停留状態の定義 [13] に基づき決定する．本実験で使用する EMR-9 では 480 pixel で視野角 40° を表現していることから，1° は 12 pixel となる．そして，停留は 100 ms の間に 1° 以内の範囲に注視点が収まる必要があり，装置の標本化周波数は 60 Hz であることから，1 フレーム当たりの移動量が 2 pixel 以下である必要がある．そのため本研究では閾値 L を 2 pixel とした．大きい動きの閾値 H については，適切な値の仮説が存在しないため，実験的に求めた．上記のデータセットを全て用いて実験したところ，最大の精度を得られた $H = 5$ pixel とすることにした．

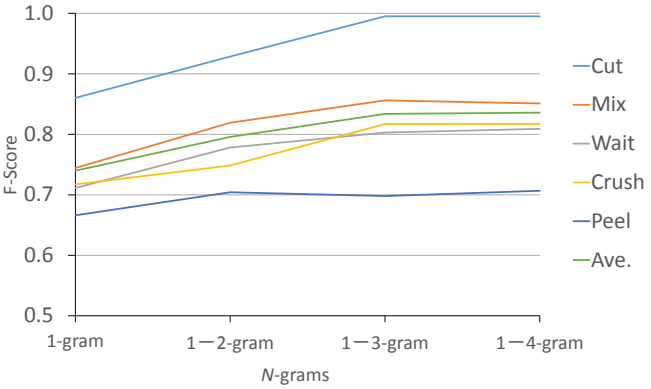

図 6　用いる N-gram 群と精度の関係

また，提案手法では N-gram の頻度ヒストグラムを結合したものを特徴ベクトルとして用いるが，適切な N の値を決定する必要がある．図 6 に用いた N-gram 群と F 値の関係を調理動作ごとに示す．概ね $N = 3$ までは用いた N-gram 群の N の値が大きいほど精度が向上したが，4-gram まで用いても精度の向上はほとんど見られなかった．さらに，4-gram まで用いる場合には 3-gram まで用いる場合に比べ莫大な計算時間がかかる．このことより以下の実験では $N = 3$ までの N-gram 群の頻度ヒストグラムを連結したものを特徴ベクトルとして識別に用いた．

3.2 比 較 手 法

ここでは動作識別において従来用いられていた，文献 [6] の手法における特徴抽出方法ついて述べる．「O」と「Z」を含まない 24 種類の記号を統計処理した N-gram の頻度ヒストグラムを

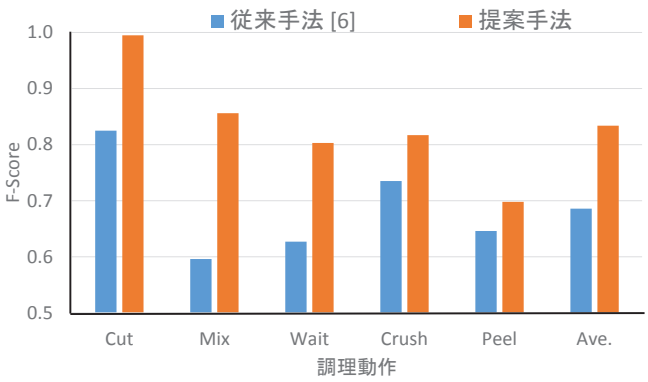

図 7　調理動作識別結果における従来手法との比較

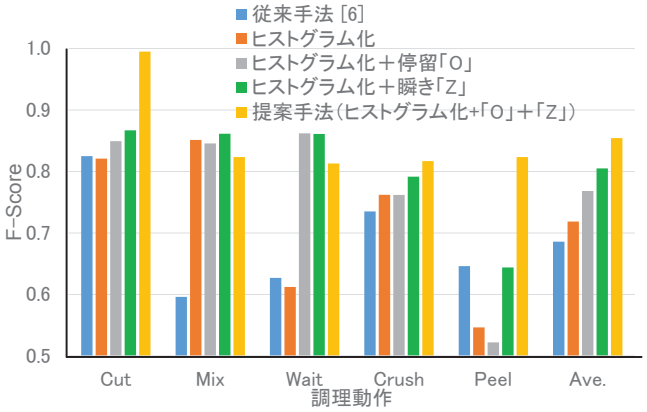

図 8　追加記号による調理動作識別精度の変化

作成し，その統計量を特徴とする．この際 1-gram から 4-gram までの計 4 つの Wordbook を集計し，それぞれから 5 つの特徴量を抽出する．具体的には，(1) 最大出現語の出現数，(2) 全出現語の平均出現数，(3) 出現語数，(4) 出現数の分散，(5) 最大出現数と最小出現数の差である．

3.3　識 別 結 果

上記の実験条件において，調理動作と精度の関係を図 7 に示す．図より全ての動作において比較手法 [6] より高精度の識別を行えたことが分かる．提案手法の平均 F 値は比較手法 [6] と比べて 0.168 の精度向上を果たし，0.854 となった．なお Wait の精度が低いのは，被験者によって待機中の動作に個人差があったためであると考えられる．

3.4　特徴量間の比較

提案手法では従来手法に比べ，停留を考慮するために原点「O」を，瞬きを考慮するために「Z」を記号に加えた．本実験では注視点が計測されなかった区間を瞬きと仮定した．そして瞬き区間においては，その前後の座標で線形補間した座標に事前処理を適用した．この記号追加の有効性を検証する．従来手法 [6],「O」も「Z」も含まない従来の記号で作成した N-gram から統計量を抽出せずに全体をヒストグラム化しただけのもの，それに加えて「O」単体 を考慮したもの,「Z」単体を考慮したもの，両方とも考慮したもの（提案手法）の精度を比較したものを図 8 に示す．平均精度に目を向けると，それぞれの記号が精度向上に貢献していることが分かる．また,「Mix」,「Wait」においては提案手法の精度が低下しているが，これは特徴次元が増えたことで雑音が混ざり，クラスタリング時に動作の典型的なパターンを表現するようなサンプルが間引かれすぎたことが原因であると考えられる.「Peel」においては，追加した記号が意図したような貢献を果たしていないが，これは調理者が皮むきを手元に引き付けて行ったため，視対象が視線計測装置のキャリブレーション平面から離れてしまい，視線を正確に計測できなかったことがその原因として考えられる．

3.5　多クラス分類による比較

多クラス分類での性能評価を行うために，上記で行った 2 クラス分類における各動作の尤度を利用し，その尤度が最も高かった動作を予測結果とした．その識別結果の Confusion Matrix

推定されたクラス

実際のクラス	Cut	Mix	Wait	Crush	Peel	Recall
Cut	0.67	0.31	0.02	0.00	0.00	0.67
Mix	0.01	0.98	0.01	0.00	0.00	0.98
Wait	0.04	0.45	0.51	0.00	0.00	0.51
Crush	0.05	0.53	0.00	0.41	0.00	0.41
Peel	0.17	0.53	0.06	0.01	0.23	0.23
Precision	0.88	0.47	0.90	0.99	1.00	

図 9　多クラス識別結果

（分類表）を図 9 に示す．分類表においては行単位でクラスが統一されており，列単位にその中でどのクラスにどの割合で識別が行われたかを示している．識別結果「Cut」,「Mix」,「Wait」はいずれも正しいクラスを推定できた標本が過半数を超えたが,「Crush」,「Peel」に関しては正しく分類できなかった標本が多い．また,「Mix」においては recall が非常に高く，それ以外のクラスでは precision が高くなっている．このことは「Mix」に含まれる視線運動パターンが他の動作においても共通で出現していることによるものと考えられる．

4.　む　す　び

本発表では視線運動を用いて 5 種類の基本的な調理動作の識別について報告した．実験では動作が混ざらない時区間を対象に 2 クラス識別で各動作の識別を行ったところ，高い精度が得られたものの，多クラス識別においては精度良く識別することができないクラスが存在した．

本研究では人間の内部状態（認知・判断）を反映する視線情報のみに着目したが，調理者の注視対象を考慮可能な画像特徴（動作）を併用することで更なる精度向上が可能であると考える．また，動作が切り替わる直前に，調理者の意識が次の動作に移る際に，調理者の注意が次の動作に関連する調理器具や食材に移ると考えられるため，そのような特徴的な視線パターンを抽出したい．

今後は，識別クラス数を増加させた上で，動作区間の切り替

わりを推定し，同一動作の区間を自動検出し，多クラス識別で前後区間との平滑化により安定した結果を得られるようにするなどの改良を検討する．このようにして調理動作が識別可能となれば，これを調理支援に応用するために，視線運動パターンをさらに詳細に解析する予定である．例えば，調理熟練者と初級者では現れる N-gram の出現順序と持続時間が異なる可能性がある．このような特徴に基づいてコツを抽出するなどの応用を目指していく．

5. 謝　　辞

本研究の一部は，科学研究費補助金および実世界データ循環学リーダー人材養成プログラムの支援による．

文　　献

[1] K. Doman, C. Kuai, T. Takahashi, I. Ide, and H. Murase, “Video CooKing: Towards the synthesis of multimedia cooking recipes,” Proc. 17th Int. Conf. on Multimedia Modeling (MMM2011), pp. 135–145, Jan. 2011.

[2] Y. Hayashi, K. Doman, I. Ide, D. Deguchi, and H. Murase, “Automatic authoring of domestic cooking video based on the description of cooking instructions,” Proc. 5th Int. Workshop on Multimedia for Cooking and Eating Activities, pp.21–26, Oct. 2013.

[3] 平山高嗣, “人間の内部状態を顕在化する視覚的インタラクション,” 情処研報, 2013-CVIM-188-27, Aug. 2013.

[4] Y. Li, A. Fathi, and J. Rehg, “Learning to predict gaze in egocentric video,” Proc. 2013 IEEE Int. Conf. on Computer Vision (ICCV2013), pp.3216–3223, Dec. 2013.

[5] 井上 裕哉, 平山 高嗣, 井手 一郎, 出口 大輔, 村瀬 洋 “視線情報の分析に基づく調理行動理解に向けて,” 情処技報, マルチメディア・仮想環境基礎研究会（MVE), Vol. 114, No. 487, pp.47–48, Mar. 2015.

[6] A. Bulling, J. Ward, H. Gellersen, and G. Troster, “Eye movement analysis for activity recognition using electrooculography,” Proc. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 33, No. 4, pp.741–751, Feb. 2011.

[7] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, “Fast unsupervised ego-action learning for first-person sports videos,” Proc. 24th IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR2011), pp.3241–3248, June 2011.

[8] K. Ogaki, K. Kitani, and Y. Sugano, “Coupling eye-motion and ego-motion features for first-person activity recognition,” Proc. IEEE Workshop on Egocentric Vision in Conjunction with CVPR2012, pp.1–7, June 2012.

[9] T. Nakano, Y. Yamamoto, K. Kitajo, T. Takahashi, and S. Kitazawa, “Synchronization of spontaneous eyeblinks while viewing video stories,” Proc. R. Soc.B, Vol.276, pp.3635–3644, July 2009.

[10] R. Collobert and S. Bengio, “Support vector machines for large-scale regression problems,” J. Machine Learning Research, Vol.1, pp.143–160, Feb. 2001.

[11] D. Vinod, “Integer programming and the theory of grouping,” J. American Statistical Assoc., Vol.64, No.326, pp.506–519, June 1969.

[12] `http://eyemark.jp/product/emr_9/index.html`

[13] D. Irwin, “Fixation location and fixation duration as indices of cognitive processing,” The Interface of Language, Vision, and Action: Eye Movements and the Visual World, pp.105–134, Psychology Press, New York, 2004.