H-007

# 一人称視点動画における食材の変化を利用した詳細調理行動認識の検討
# Fine-Grained Cooking Action Recognition in Egocentric Videos with Capturing Ingredients' Transformation

岡本 淳志† 井上 勝文† 吉岡 理文†
Atsushi Okamoto Katsufumi Inoue Michifumi Yoshioka

## 1. はじめに

近年, ウェアラブルカメラで撮影した一人称視点動画の解析が盛んに行われている. 中でも調理行動においては, 撮影者の行動を認識することで, 自動レシピ生成などの調理支援サービスが可能となる. この分野における先行研究では, Convolutional Neural Networks(CNN)モデルが“切る”, “炒める”, といった大まかな調理行動について正確に認識できている[1, 2]. しかし, より正確な調理支援のためには切り方, 混ぜ方といった詳細な調理行動の認識が必要となる. そこで, 本研究では詳細な調理行動の認識を目的とする. その第一歩として, 基本的な調理行動の一つである切り方に着目し, “乱切り”, “千切り”などの調理行動の認識精度向上を目指す.

従来の行動認識手法では, 手の動きや形, 調理器具を捉えるモデルが提案されている[1, 2]. しかし, 切り方においてはそれらの情報が類似しているため認識が困難である. 一方で, 切り方は調理後の食材の形によって名付けられることが多く, その形が特徴的なものが多い. そこで, 本研究では食材の変化に着目し, これを捉えることで切り方を認識できると仮定する. この仮定のもと, 前後のフレームから中間フレームを復元するフレーム補間によって, 食材の“変化”という情報から行動認識するモデルを提案する. 食材変化を高精度に捉えるには, 高精度なフレーム補間が必要となるが, 一般的な元画像と生成画像との画素単位の比較で学習する手法を用いると, ボケた画像が生成され, 食材の形状の復元が難しいという問題がある. そこで本研究では, Generative Adversarial Network[3]のフレームワークにおける識別機能を導入し, 画像が生成されたものか否かを見抜くよう学習させる. これにより, 元画像に近い画像を生成する生成器を作成する. 本研究では, このように作成した生成器より抽出される特徴量は食材の変化情報が含まれていると考え, この情報より行動を認識し, その精度を検証する.

## 2. 関連研究

主に動画を用いた行動認識手法では, 外観情報と動き情報を捉える 3DCNN が広く用いられる. Carreira らの手法である I3D[4]では, RGB 画像と Optical Flow 画像を個別の 3DCNN に入力し, 学習させる. これにより, それぞれ外観情報と動き情報を捉える. Urabe らの手法[5]では,一人称視点の調理動画において, 外観情報を捉えるための 2DCNN と, 動き情報を捉えるために 3DCNN を組み合わせた手法を提案している. 2 つの CNN の予測結果を平均する, または掛け合わせることで高精度な調理行動認識を可能としている. また背景情報の削減のため, 画像から手の周辺領域を切り出し CNN に入力している. 本研究でもその知見を利用し,前処理として手領域からその先端部分の領域を切り出す. Michibata らは RGB 画像, Flow 画像を入力した 2DCNN に加え, 手領域を示す 2 値画像を入力する 2DCNN を導入したモデルを提案している[1]. 一人称視点の調理行動認識において高精度を達成しているが, 手の動き等が類似した詳細行動においては誤認識が多いことが報告されている.

図 1 前処理による領域抽出の例. (左)元画像に手領域を黄色でマスクした画像. (右)抽出された領域.

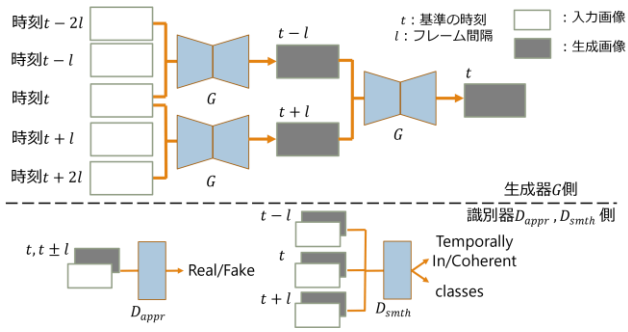


図 2 行動認識ネットワークの概要. (上部)画像生成の流れ. (下部)切り方識別及び画像の正誤識別の流れ.

このため, 外観情報と動き情報に着目した[1, 4, 5]では, 詳細行動の判別は難しい. 本研究では,その解決を目的とする.

## 3. 提案手法

提案手法は食材周辺領域を切り出す前処理部と, 中間フレームを生成, 及び切り方を判定する行動認識部に分かれる.

### 3.1 前処理部

調理動画にはその行動に関係ない背景情報, 物体が多く存在する. 行動に関与する食材に着目して行動認識するためには, 画像の中でも食材が写る箇所に注目する必要がある. そのために前処理部は, 手の先端付近に食材が映ると仮定し, その周辺領域を抽出する. 図 1 に前処理を通した食材周辺領域の抽出結果を示す. まず, EGTEA Gaze+データセット[6]で学習した RefineNet[7]を用いて手領域を推定し, 各連結領域の重心を計算する. そして, 手領域の先端付近の正方形領域を切り出す. このとき, 切り出す中心の x 座標を重心の中点に, y 座標を手領域の最上位ピクセルと等しくなるよう設定している. これにより, 調理行動に関係ない物体や, 背景の影響が軽減される.

### 3.2 行動認識部

動画内に写る物体の変化を捉えるために, フレーム補間モデルを行動認識に応用する. 特に本研究では, ある時刻の前後の画像から基準の時刻の画像が正しく生成できれば食材の変化を捉えられている, と仮定する. この仮定のもと, 行動認識部ではフレーム補間と同時に切り方を認識する. 行動認識部は図 2 に示すように, 生成器$G$と, 2 つの識別器$D_{appr}, D_{smth}$の 3 種の CNN で構成され, 時刻$t$を基準に間隔$l$で計 5 フレーム使用する. まず, 前後の時刻の画像から中間画像を復元する$G$を用いて, 時刻$t-2l, t, t+2l$の画像から時刻$t-l,\ t+l$の画像を生成し, これらの生成画像から同様に時刻$t$の画像を生成する. そして生成した画像, 及び元画像の双方を図 2 に示すように 2 種の識別器$D_{appr}, D_{smth}$に入力する. $D_{appr}$が元画像かどうかを推定し, $D_{smth}$が時系列が自然かどうか, 及び切り方を推定する. フレーム補間と切り方の分類を同時に学習することで, 形状の変化を捉えたクラス分類を実現する.

損失関数として, 生成画像と元画像の L1 ロス, $D_{appr}, D_{smth}$における敵対的ロス, 切り方の予測と教師データとの交差エントロピーロスの総和をとっている. また, 推論時には, 生成画像と元画像をそれぞれ$D_{smth}$に通して得られた切り方の予測確率を平均して, 最終の予測とする.

## 4. 実験

### 4.1 実験条件

本実験では, 調理行動の中でも器具が共通かつ食材の形が特徴的である, “薄切り”, “千切り”, “短冊切り”, “みじん切り”, “乱切り”の 5 種類の切り方について認識精度を評価した. また, 環境変化に対する頑健性を検証するため, 複数の環境を使用した. 井上ら[2]の作成した 1 環境のデータセットに, 異なる 3 つの環境で撮影したデータセットを追加した. [2]の撮影環境でのデータを Data A, 新たに撮影したデータを環境ごとにそれぞれ Data B, C, D とする. 学習, 検証データには Data A を使用し, テストデータには Data A 以外を使用した. また, 精度比較として, 既存の行動認識 CNN である I3D[4]でも同様に実験した. このとき, 前処理部において食材がどの切り方でも写るように, 切り出す正方形の一辺は元画像の 0.7 倍とした. 入力フレーム数は共通して 41 フレーム長であり, I3D では全フレームを入力し, 提案手法では 10 フレーム間隔で 5 フレーム抜き出した. 学習エポックは 30 エポック中 Data A の検証データで最高精度であるものを採用した.

### 4.2 実験結果

各手法における 3 回平均の認識精度を表 1 に示す. I3D(RGB), I3D(Flow)はそれぞれ RGB 画像, Flow 画像を用いて学習したモデルであり, I3D はそれらの平均をとったものである. 提案手法は, 学習時と異なる環境である Data B, C, D では従来手法に勝る結果となった. この結果から, 提案手法の方がより環境差異に強いことが確認できた. これは前処理を通して背景情報が削減され, より切り方に関係のある食材周辺の情報から判断するようになったことに起因したと考えられる. しかし, 食材の変化を安定して捉えるのは難しく, Data A における精度が他手法に比べ低い結果となった. 全体的な精度向上, 及び安定した認識のためには, 外観情報などを捉える既存のネットワークとの併用などが考えられる.

表 1　切り方認識の精度[%]

| Methods | Data A | Data B | Data C | Data D |
|---|---|---|---|---|
| I3D(RGB) | **96.3 ± 0.6** | 38.9 ± 3.7 | 8.6 ± 3.7 | 27.2 ± 9.9 |
| I3D(Flow) | 62.6 ± 7.8 | 28.5 ± 6.5 | 21.8 ± 9.7 | 24.2 ± 9.3 |
| I3D | 64.9 ± 3.7 | 29.3 ± 7.9 | 25.8 ± 13 | 24.8 ± 12 |
| **Ours** | 47.0 ± 8.4 | **43.1 ± 7.9** | **46.3 ± 7.5** | **33.3 ± 7.8** |

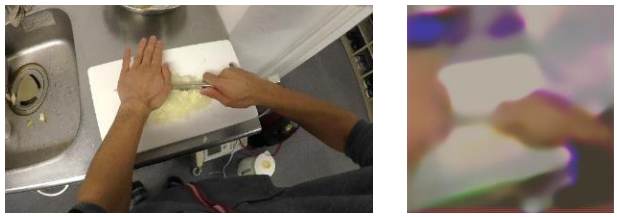

**図 3　$G$による生成画像の例. (左)元画像. (右)生成画像.**

$G$による生成画像の例を図 3 に示す. 手領域は復元されているが, 食材部分はぼやけて消えている. これより, 元画像の L1 ロスにより大域的に似ている画像は生成できているものの, 敵対的損失による画像生成の強化があまりできていないようにみられた. また, このことが切り方の認識精度における不安定さの一因となっていると考えられる.

## 5. おわりに

本研究では切り方の認識において, 食材の変化に着目したモデルを提案し, 認識精度を評価した. 食材の変化に着目することで, どの環境においても一定の認識精度が得られたが, 一方で認識精度が不安定であることも確認された. 今後の課題として, 既存のネットワークを組み込んだモデルの提案, 識別器の改良などが挙げられる.

また, 本研究では詳細な調理行動のみに限定して認識することを目的とした. 今後の展望としては, 的確な調理支援サービスの実現のため, 認識結果を用いた動的なレシピ文生成などを検討していく.

### 参考文献

[1] S. Michibata et al., “Cooking activity recognition in egocentric videos with a hand mask image branch in the multi-stream cnn”, In Proc. of CEA, (2020).
[2] 井上ら, “一人称視点動画を用いた詳細調理行動判別の検討～3d-cnn を用いた認識手法の場合～”, 第 20 回画像の認識・理解シンポジウム Extended Abstract 集, (2017).
[3] I. J. Goodfellow et al., “Generative adversarial nets”, In Proc. of NIPS, (2014).
[4] J. Carreira, A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset”, In Proc. of CVPR, (2017).
[5] S. Urabe et al., “Cooking activities recognition in egocentric videos using combining 2dcnn and 3dcnn”, In Proc. of CEA, (2018).
[6] Y. Li et al., “In the eye of the beholder: Gaze and actions in first person vision”, In Proc. of ECCV, (2018).
[7] G. Lin et al., “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation”, In Proc. of CVPR, (2017).

† 大阪公立大学 大学院情報学研究科　Graduate School of Informatics, Osaka Metropolitan University