注視点を導入した Convolutional Recurrent Neural Network における調理行動認識

EP13064 小島隆義　　指導教授：山下隆義

1. はじめに

従来の Convolutional Recurrent Neural Network(CRNN)[1] による行動認識では, 画像全体を入力して行動の認識を行う. しかし, 行動によっては動作が画像の局所領域にて発生するため, どこの領域に着目するべきか学習できないことがある. そこで, 本研究では, 行動の認識に重要な領域を注視点として与えることで行動認識の高精度化を実現する.

2. CRNN

CRNN は, 畳み込み層とプーリング層, 全結合層から構成される Deep Convolutional Neural Network(DCNN) と, 系列データを扱う Recurrent Neural Network(RNN) を組み合わせることで, 動画像に対応した認識が可能である. 図 1 に CRNN のネットワーク構造を示す. DCNN で特徴抽出を行い, RNN で時系列情報を学習する.

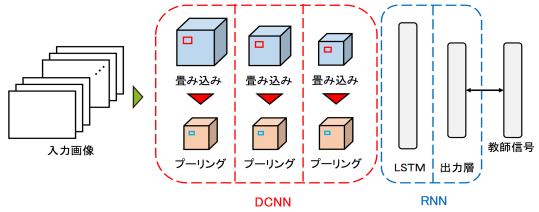


図 1：CRNN のネットワーク構造

3. 注視点を導入した CRNN による調理行動認識

本研究では, 2 種類の注視点情報の導入方法を検討する. 学習をはじめる前に, センサにより注視点を取得する. そして, 注視点を補助的に CRNN へ入力する方法と, 注視点を中心に切り出したフレームを CRNN に入力する 2 種類のネットワークを提案する.

3.1. 注視点の取得

The Eye Tribe Tracker を用いて 5 人の被験者の注視点を取得する. 同時に, 行動の切り替わるタイミングを入力してもらう. その入力が, あらかじめデータセットに用意されている各フレームの調理動作のラベルに近い被験者の注視点を利用する. 図 2 に注視点の取得方法を示す.

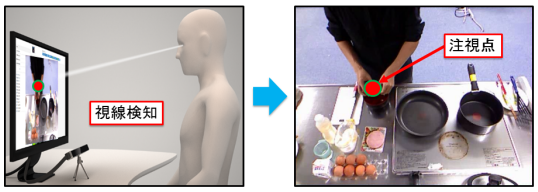


図 2：注視点の取得方法

3.2. 注視点入力 CRNN

注視点入力 CRNN のネットワーク構造を図 3 に示す. まず, 映像のフレームを入力し, 畳み込み層およびプーリング層で特徴ベクトルを生成する. そして特徴ベクトルを RNN 層に入力し, フレームに対する認識結果を出力する. RNN 層は内部に各フレームでの特徴情報が蓄積され, 系列を考慮した認識を可能とする. また, 注視点を入力として与えることで, 行動の認識に有効な領域に着目することができる.

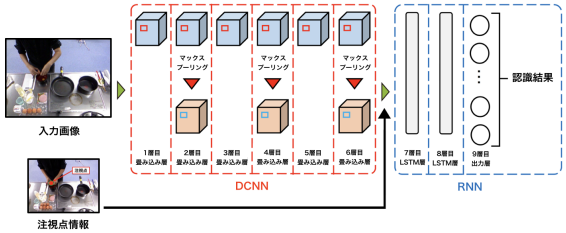


図 3：注視点入力 CRNN のネットワーク

3.3. 注視点切り出し CRNN

注視点切り出し CRNN のネットワーク構造を図 4 に示す. まず, 映像のフレームと注視点情報から, 注視点座標を中心に切り出す. 切り出したフレームを入力し, 畳み込み層およびプーリング層で特徴ベクトルを生成する. そして, 特徴ベクトルを RNN 層に入力し, フレームに対する認識結果を出力する. RNN 層は内部に各フレームでの特徴情報が蓄積され, 系列を考慮した認識を可能とする. また, 注視点を中心に切り出すことで, フレーム全体ではなく行動の認識に重要な領域のみを対象とする.

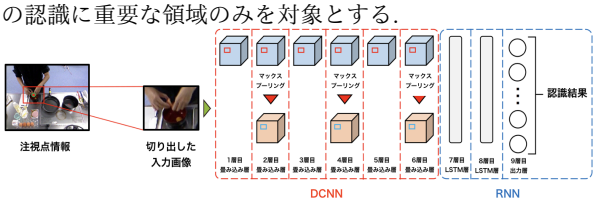


図 4：注視点切り出し CRNN のネットワーク

4. 評価実験

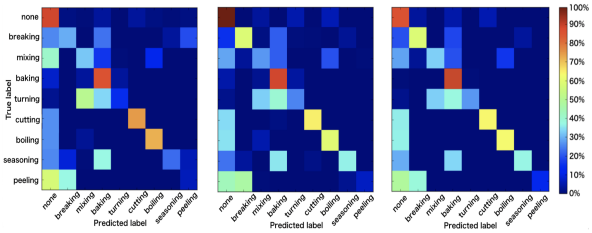
本実験では, Kitchen Scene Dataset を用いて調理行動認識を行う. Kitchen Scene Dataset は, 料理風景を撮影したデータセットであり, 7 人の被験者が 5 種類の卵料理を調理する様子を撮影している. Kitchen Scene Dataset は, 227,874 フレーム, 35 個の動画で構成されている. 調理行動は, 焼く, 混ぜる, 割る, 返す, 切る, ゆでる, 味付け, むく, 動作なしの 9 クラスである.

各手法の実験結果を表 1 に, Confusion Matrix による実験結果を図 5 に示す. 表 1 より, 全体の認識精度は注視点切り出し CRNN が最も良い. これにより, 注視点を中心とした領域のフレームを入力することは, 従来の CRNN のようにフレーム全体を入力するよりも有効である. クラス別では, 割る, まぜる, 焼く, 返す, 味付け, むくの 6 クラスの認識率が向上している. これは, 入力されたシーンごとに着目する領域が異なるため, 他クラスと誤認識しにくくなったからだと考えられる. 一方, 切る, ゆでるクラスの認識率は低下している. これは, 切る, ゆでるクラスは着目する領域が動作なしクラスの領域と重複することが多いため, 動作なしと誤認識しやすいからだと考えられる.

表 1：各手法の認識率 [%]

ネットワーク	割る	まぜる	焼く	返す	切る
CRNN	29.13	31.71	84.92	14.21	**74.57**
注視点入力 CRNN	58.47	27.11	88.23	25.33	64.63
注視点切り出し CRNN	**60.15**	**35.32**	**90.51**	**27.06**	65.72

ネットワーク	ゆでる	味付け	むく	動作なし	全体
CRNN	**71.11**	22.21	6.35	**87.39**	66.07
注視点入力 CRNN	59.44	**36.22**	7.11	84.91	66.28
注視点切り出し CRNN	62.63	34.94	**14.03**	81.47	**67.48**



(a) 従来の CRNN　(b) 注視点入力 CRNN　(c) 注視点切り出し CRNN

図 5：各手法の Confusion Matrix

5. おわりに

本研究では, 2 種類の注視点情報の導入方法として, 注視点を補助的に CRNN へ入力する方法と, 注視点を中心に切り出したフレームを CRNN に入力する方法を提案した. 評価実験により, 注視点を中心に切り出したフレームを CRNN に入力する方法が適していることを確認した. 今後は, 学習方法のさらなる効率化を検討する.

参考文献

[1] Z.Zuo, *et al.*, “Convolutional Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation”, CVPR , 2015.