G-1

音声ラベルを用いた 3DCNN による人物動作分類に関する検討

A Study on Human Motion Classification Using 3DCNN with Speech Labels

○王雨桐[1]，関弘翔[2]，細野裕行[2]
*Yutong Wang[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: Recent research shows that replace standard categorical labels with high dimensional, high entropy labels can lead to more robust and data-efficient feature learning in image classification task. In this research, we build the human motion classification task by using speech label as the supervised signal and evaluate the availability.

1. まえがき

動画像の分類は，画像の分類ほど確立されていない課題である．人物動作認識で使用されるディープニューラルネットワーク(DNN)には，画像認識で使われる二次元の畳み込みニューラルネットワーク(2D-CNN)を時間軸方向に拡張した三次元の畳み込みニューラルネットワーク(3D-CNN)を使用するモデルなどがある[1].

近年，教師あり学習における教師ラベルのモダリティ表現を検討することで，学習モデルの品質に大きな影響を与えることが報告されている[2]．音声ラベルの使用は，学習データ量が少ない場合，より識別性の高い特徴の学習を促すことが期待できる．音声を教師信号として用いることで，画像分類課題に対する新しいパラダイムが提案されている[2]．本研究では，動画は画像に対して時間方向の次元が追加されたデータであり，画像認識で有効な手法が動画認識においても適用できる可能性があると考えた．

本研究の目的は，通常のカテゴリラベルでなく，高次元，高エントロピーの音声ラベルを用いて，頑健性と効率性の高い動画分類器を構築し，有用性を評価することである．以上を基にして，音声ラベルを用いた人物動作分類器の構築とその評価を報告する．

2. ラベルの処理

データセットのテキストラベルからテキスト音声合成システムで英語の音声を自動に発話して，ログメルスペクトログラムを生成する．

3. モデルの構築

本研究では，UCF101[3]などの人物動作認識用のデータセットを対象として，動画識別用 3D-CNN モデルを構築する． 提案するモデルのアーキテクチャを Fig. 1 に示す．提案モデルは，動画エンコーダと音声ラベルデコーダの２つの部分から構成されて，3DResNet モデルを動画エンコーダとして使用し，一つの全結合層といくつかの転置畳み込み層を音声ラベルデコーダとして使用する．音声ラベルデコーダ内部で，バッチ正規化と活性化関数 ReLU を備えている．

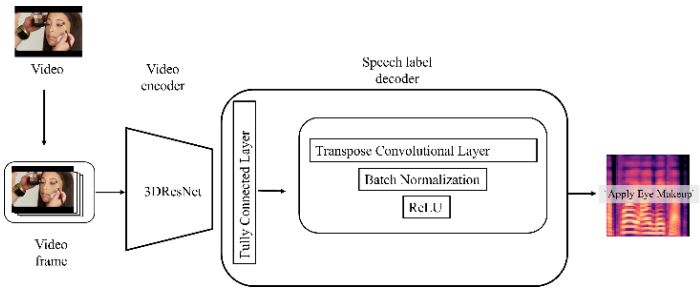

Figure 1. An overview of human motion classification models with speech labels

4. 有用性の評価

有用性を効率性と頑健性から評価する．効率性の検証のため，学習データを減少し，カテゴリラベルと音声ラベルを用いたモデルの認識精度を比較する．

また，テストデータに敵対的摂動を加えたとき，摂動の大きさを増加しながら，カテゴリラベルと音声ラベルを用いたモデルの認識精度を比較することで頑健性を評価する．

5. まとめ

本研究では音声ラベルを用いた人物動作分類器を提案し，モデル構築を行った．このモデルに対して，カテゴリラベルと音声ラベルを用いて，少量な学習データの場合の認識精度と敵対的攻撃に対する頑健性を検討していく．

参考文献

[1] Hara Kensho, et al. “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?” Proceedings of the IEEE conference on CVPR. pp. 6546-6555. 2018.

[2] Boyuan Chen, et al.: “Beyond Categorical Label Representations for Image Classification,” ICLR 2021

[3] Soomro Khurram, et al.: "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402. 2012.

１：日大理工・院（前）・情報　２：日大理工・教員・情報