4R-06

Diffusion Model による体特徴点データ拡張を用いた少数データからの人物行動認識

濵部翔太† 山田啓一‡
名城大学† 名城大学‡

1. はじめに

近年，ビデオ画像からの人物行動認識を用いた製品やサービスが，数多く開発・実用化されている．人物行動認識の精度向上は，更なる製品やサービスの発展に不可欠である．

一般に，ビデオ画像から人物行動を精度よく認識するには，大量の訓練データが必要である．訓練データの数を増やす方法として，データ拡張がある．近年，敵対的生成ネットワーク（GAN[1]）や拡散モデル（Diffusion Model[2]）といった生成モデルを用いたデータ拡張が研究されている[3]が，特にビデオ画像の場合には，生成モデル自体の訓練にも大量のデータが必要となる課題がある．

ビデオ画像からの人物行動認識の手法として，[4]-[6]などが挙げられる．これらのうち[6]では，ビデオ画像を体特徴点の時間推移を表す 1 枚の画像（Encoded Human Pose Image, EHPI）に変換し，これを畳み込みニューラルネットワーク（CNN）により識別することで行動認識を行っている．

本論文では，ビデオ画像からの人物行動認識を目的に，訓練用のビデオ画像を EHPI に変換して生成モデルを訓練し，この生成モデルを用いてデータ拡張することによる手法を提案する．提案手法は，生成された EHPI を識別器で選別し前記 CNN の訓練に用いることを特徴とする．生成モデルに Diffusion Model を用い，実験によりその有効性を示す．

2. 提案手法

図 1 に提案手法の概要を示す．ビデオ画像から人物の体特徴点を抽出し EHPI に変換する．EHPI は，縦方向がビデオフレームに対応した時間，横方向が体特徴点，画素値が体特徴点の xy 座標を表す画像である．この EHPI を用いて生成モデルを訓練し，訓練後の生成モデルを用いてデータ拡張のための EHPI を生成する．特徴点座標の時間推移が集約された単一フレーム画像を用いることで，比較的少数の訓練データでも人物行動についての生成モデルの訓練が可能となると期待される．生成された EHPI は必ずしも人の取り得る姿勢を表現しているとは限らないため，これを人の取り得る姿勢の知識を用いて選別する．このようにして生成し選別された EHPI を元の EHPI に加え，EHPI から人物行動を識別する CNN の訓練に用いる．

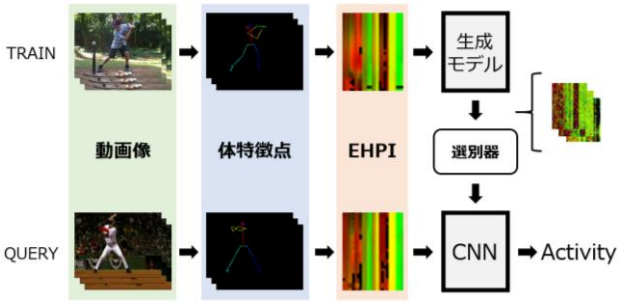


図 1 提案手法

3. 実験

3.1 実験データセット

実験には，21 クラスの人物行動のビデオ画像が含まれる JHMDB[7]を用いた．このデータセットには各クラス 39～55 シーケンスのビデオ画像が含まれるが，本実験では，これらのうちの各 11 シーケンスを用いた．

3.2 実験方法

各シーケンスから，開始フレームを 1 フレームごとにシフトしながら，区間長 32 フレームのサブシーケンスを 1～9 個抽出した．抽出された区間の各フレームにおける体特徴点を，OpenPose[8]を用いて検出した．これら 32 フレームの体特徴点の xy 座標から EHPI を作成した．EHPI に変換する際，1 シーケンスにおける首の座標の平均値が(0,0)となるよう平行移動し，xy 座標を正規化した．

生成モデルには Diffusion Model を用いた．比較のため，生成モデルに GAN を用いた場合についても実験した．Diffusion Model の訓練の際は，エポック数を 80 とした．GAN の訓練の際は，潜在変数を 100，バッチサイズを 32，エポック数を 50 とした．

生成された EHPI が人の取り得る姿勢であるか識別するために，本実験では簡易的な方法として，人物の首の長さ（鼻から首の特徴点の長さ）の変化に注目した．生成された EHPI の首の長さの変化量が閾値以上のとき人の取り得る姿勢ではないと判定し，それ以外の EHPI を CNN の訓練のための拡張データとした．

EHPI からの行動識別には，フィルタサイズ 2×5 の 4 層の畳み込み層と 1 層の全結合層からなる CNN を用いた．CNN の訓練には，元の EHPI に，その A 倍の生成 EHPI を加えたものを用いた．生成された EHPI の割合を変化させ，識別精度との関係を調べた．CNN の訓練の際，バッチサイズを 32，エポック数を 150 とした．初期学習率を 0.01 とし，100 エポック以降 20 エポックごとに学習率を 0.1 倍とした．正則化パラメータを 5×10^{-5} として，L2 正則化を行った．

訓練データの分散を増やすために，以下のようにデータを拡張した．

Human Activity Recognition from Limited Data using Diffusion Model Based Augmentation of Human Body Feature Points
† SHOTA HAMABE, Meijo University
‡ KEIICHI YAMADA, Meijo University

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

1) 50%の確率で左右反転
2) 反転されたデータの 25%の確率で左右の関節のインデックスを反転
3) 25%の確率で両足のインデックスを削除
4) 両足のインデックスが削除されたデータの 6.25%の確率で両膝のインデックスを削除

上で示した 2 では，撮影した方向と逆向きで撮影したことを再現する．3 と 4 の処理を行った理由は，下半身が写っていない行動クラスや下半身が隠れているシーケンスが存在するからである．

3.3 実験結果

図 2 に実験結果を示す．同図より，本論文の提案手法である Diffusion Model と生成 EHPI 選別器を用いた手法（A=0.15）の正解率が，57.1%となり，従来手法（A=0）の正解率 53.1%よりも 4 ポイント向上した．また，生成 EHPI を選別することなく用いた場合の正解率は 56.5%，生成モデルに GAN を用いた場合の正解率は 56.3%となった．これら何れと比べても，提案手法の正解率が高いことが分かる．

図 3 に訓練データに加える生成 EHPI の割合と正解率の関係のグラフを示す．提案手法では A=0.2 まで約 57%の正解率となったが，生成 EHPI を選別しなかった場合は A=0.2 での正解率が 54.9%と低くなった．加える割合 A がさらに大きいと，正解率が下がった．

3.4 考察

図 3 において，生成モデルで生成された EHPI を加える割合 A を大きくしたとき，正解率が一旦向上したあと

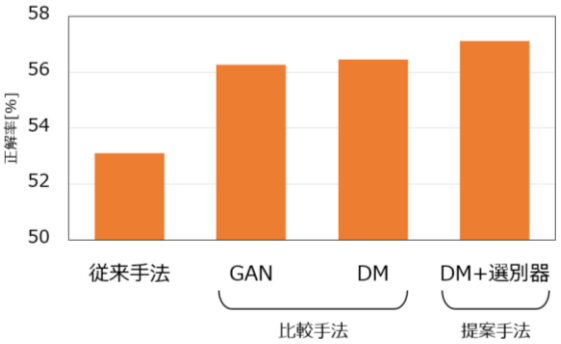


図 2 実験結果

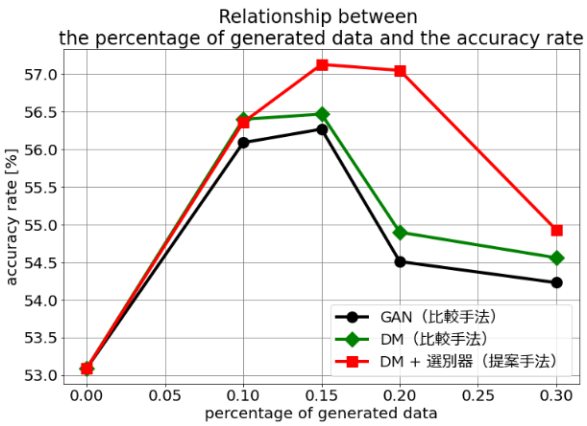
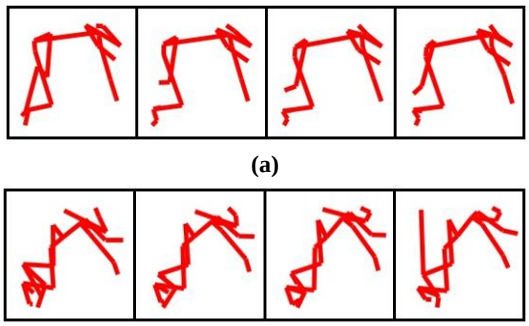


図 3 訓練データに加える
生成 EHPI の割合と正解率の関係



(a)

(b)

図 4 生成・選別された EHPI の体特徴点の例

低下した．また，生成 EHPI の選別器を用いたときの正解率が，選別を用いない場合の正解率よりも向上した．このような結果になったのは，生成されたデータが与える，正の効果と負の効果が正解率に影響するからだと考えられる．正の効果と負の効果とは，生成されたデータの質が良い場合と悪い場合の効果である．生成 EHPI の選別器を導入することで，拡張データの質が良くなったと考えられるが，生成 EHPI を加える割合をさらに増やすと，正解率の向上が低下した．

図 4(a)と(b)に，Diffusion Model で生成され選別された EHPI を体特徴点に逆変換した例を示す．動作は”push”（物を押す動作）であり，連続する 4 つのフレームを並べた画像である．この 2 つを比較すると，(b)は(a)よりも人の形を保っていないことが分かる．

これらの結果から，生成 EHPI の選別方法を改良することで正解率のさらなる向上が可能ではないかと考えられる．

4.まとめ

本論文では，ビデオ画像からの人物行動認識を対象に，生成モデルを用いて EHPI をデータ拡張することによる手法を提案した．実験の結果，21 クラスの行動識別問題において，生成された EHPI を訓練データに加えることにより，正解率が 53%から 57%に向上し，提案手法の有効性が確認された．

今後の課題として，データ拡張のための EHPI 生成選別法の更なる検討を行うことが挙げられる．

参考文献

[1] I. J. Goodfellow, et al., “Generative Adversarial Nets,” NIPS (2014)
[2] Jonathan Ho, et al., “Denoising Diffusion Probabilistic Models,” NeurIPS (2020)
[3] Xinyue Zhu, et al., “Emotion Classification with Data Augmentation Using Generative Adversarial Networks”, PAKDD (2018)
[4] K. Simonyan, et al., "Two-stream convolutional networks for action recognition in videos," NIPS (2014)
[5] M. Zolfaghari, et al., "ECO: Efficient Convolutional Network for Online Video Understanding," ECCV (2018)
[6] Ludl, et al., "Simple yet efficient real-time pose-based action recognition," IEEE ITSC (2019)
[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. "Towards understanding action recognition," ICCV (2013)
[8] Z. Cao, et al.," OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," IEEE Trans. PAMI (2019)

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.