2W-06

動画認識技術を用いたCSI自動正解ラベリングによる行動認識

石坂拓海† 田中悠貴† 斎藤隆仁†† 池田大造†† 峰野博史†

†静岡大学情報学部 ††株式会社NTTドコモ

1. はじめに

昨今の技術進展により人物の行動認識に関する研究[1]が盛んに行われている．既存の行動認識手法は，設置したカメラや無線機器等を用いた手法などが提案されている．環境に設置したカメラを用いた手法では，時空間特徴を用いた手法[1]などが提案されており，無線機器等を用いた手法としてはWi-Fiチャネル状態情報 (Channel State Information: CSI) を用いた手法が提案されている．特にCSIを用いた研究ではDeep-Learningを用いた手法が増えているが，大量のデータが必要なだけでなく，正解ラベリングが手動で行われているため，CSIデータセット作成に手間を必要とし，誤りが生じる可能性がある．また，これまで認識できていなかった複雑な状況での行動認識もCSIを用いた手法では難しかった．

本研究では，動画認識技術を用いて検出された行動を正解ラベルとし，CSIへ自動でラベリングすることで，正解ラベリングや行動ごとにデータ収集する労力削減を図りつつ，これまでCSIで認識が困難だった行動も対象とする，動画とCSIを用いた行動認識手法を検討する．学習時には動画データ取得のためにカメラを用いるが，推論時はCSIのみで実施できるため，運用時のプライバシーにも配慮できると考える．本稿では特に，手動で収集した少量の動画学習データに対し，正解ラベル生成のための動画による行動認識精度向上に関して検討する．

2. 関連研究

動画認識技術を用いた行動認識の研究は，Deep-Learningを用いることで，歩くなどの単純な行動から物体を使った複雑な行動まで様々な行動を対象に行われている．例えば，長期の時間依存性を学習するTimeception[2]は，料理や皿洗いといった日常生活の動画データを多く含むデータセットCharades[3]を用いて高精度を実現した．また，3DCNN (3D Convolutional Neural Network) にTransformerを適用した手法[4]では，データセットAVA-kinetics[5]を用いて，既存手法より高い精度を示した．ここで学習に使用された動画データセットは，ファイルサイズが約50GBの大規模データセットであり，各行動に対する学習データも数百~数万と膨大である．ただし，数十程度の特異な行動に関しては，認識性能が低くなる傾向が見られる．

一方，CSIを用いた研究もDeep-Learningを用いた手法が増えており，LSTM (Long Short Term Memory) を用いたCSIベースの行動認識[6]も提案されている．歩くや立つといった6種類の日常行動の認識において，LSTMを用いた手法で約90%の性能を示している．ただし，Deep-Learningを用いた手法は大量のデータセットを必要とし，正解ラベリングを手動で行っているため，多くの時間を必要とし，ラベリングの誤りが生じる可能性も高い．

Behavior Recognition by CSI Automatic Correct Labeling using Video Recognition Technology

Takumi Ishisaka†, Yuki Tanaka†, Takato Saito††, Taizo Ikeda††, Hiroshi Mineno†

† Faculty of Informatics, Shizuoka University

†† NTT DOCOMO, Inc.

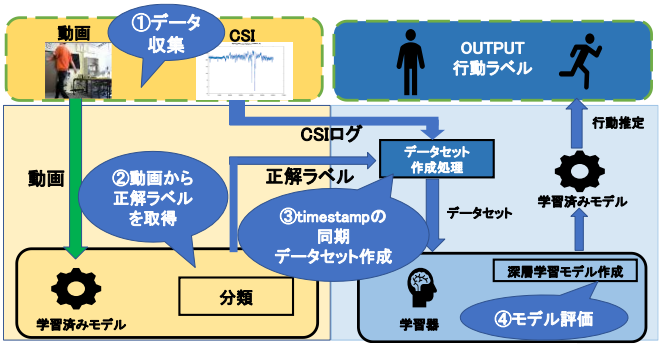

図1 提案手法の概要

3. 動画認識技術を用いたCSI自動正解ラベリングによる行動認識

本研究では，動画認識技術を用いて検出された行動を正解ラベルとし，CSIへ自動でラベリングすることで，正解ラベリングや行動ごとにデータ収集する労力削減を図る動画による正解ラベルを用いたCSIベース行動認識手法を提案する（図1）．これにより，これまで認識できていなかった複雑な状況での行動も学習が容易となる．

CSIの自動正解ラベリングに用いる動画・CSIには，各認識行動における動画時間の短い動画を用いるのではなく，各認識行動を包含する長時間の1つの動画を用いる．この入力動画から正解ラベルを出力する行動認識モデルの学習器として3D Resnet[7]を用いることとした．画像認識タスクにおいて高い予測性能を持つResnetに時系列情報を学習可能にした3D Resnetは，動画の画像情報のみを使用して学習可能であり，動画による行動認識において主流な手法となりつつある．また，学習用の動画データの前処理として，動画内の2フレーム間でグレースケール化による画素差分を用いて適切なフレームを抽出しトリミングを行う．動画のトリミングを行うことで動画内の行動特徴を絞り込むよう前処理を行う．

CSIの自動正解ラベリングで，動画解析結果の時刻とCSIの時刻を同期させる必要があるため，データ収集時にネットワークタイムプロトコルNTP (Network Time Protocol) を用いて時刻同期を行う．さらに，CSIのタイムスタンプは受信機に搭載されたNICに依存するため，適切な実時間に変換する．

CSIによる行動認識モデルの学習器にはLSTMを用いる．LSTMは再帰的ニューラルネットワークRNN (Recurrent Neural Network) において主流な手法で，その他主流なモデルである隠れマルコフモデルHMM (Hidden Marcov model) やランダムフォレストと比較して，高い性能を得られることが確認されている[6].

4. 評価実験

4.1. 実験方法

正解ラベル生成のための動画による行動認識精度向上に関して基礎実験を実施した．認識対象とする行動は

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

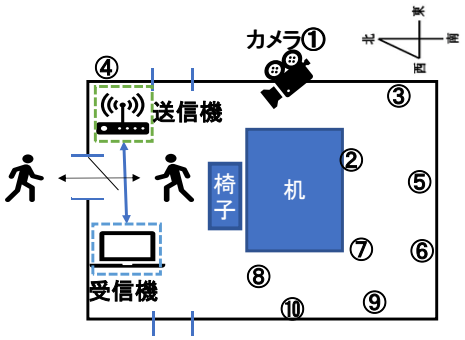

図 2 実験環境

「部屋への入室」「部屋からの退出」「椅子に座る」「椅子から立つ」「行動なし」の 5 種類とした．その理由は，類似した行動は動画を用いた場合は識別可能だが，CSI では認識が困難であると考えたため，CSI で行動認識できそうな一連の行動を対象とした．実験環境は生活環境を想定しており，机や椅子，棚，窓などが存在する（図 2）．壁は一部鉄筋コンクリート，一部鉄骨で，ドアは締め切った状態で実験を行った．動画撮影には Nvidia Jetson Nano と Web カメラ（logicool CN270）を使用し，CSI 収集には Intel5300 を搭載した受信機ノート PC と Wi-Fi ルーター（Buffalo WSR-2533DHPL）を送信機とした．

まず実験 1 では，動画による行動認識モデル学習用のデータ収集を行った．動画による行動認識モデルの汎用性を確認するため，図 2 中の①から⑩の 10 か所に Web カメラを設置し，異なる複数の画角の動画を撮影した．各箇所で 5 種類の行動×10 試行分のデータ収集を行い，合計 500 個の動画データを収集した．また，収集した動画データは，左右反転，グレースケール化，ぼかし，RGB シフトを組み合わせた 9 種類のデータ拡張方法で行い，合計 5000 個の動画データでモデル学習を実施した．

次に実験 2 では，CSI による行動認識モデル学習用のデータ収集を行った．図 2 のように部屋内のドア周辺に Wi-Fi 送受信機を設置し，カメラを③の位置に設置した．無線通信はフレネルゾーンという楕円形の無線通信路が存在し，送受信機間の直線を中心とした楕円体を通じて送受信される．そのため，フレネルゾーンに人が入出し，行動することで CSI に顕著な変化が表れると考え，送受信機を設置した．CSI と動画は時刻同期を行い，「行動なし→部屋に入る→椅子に座る→椅子から立つ→部屋から出る→行動なし」という行動順で 5 種類の行動を包含した 1 連の動画×20 試行分のデータ収集を行った．

4.2. 実験結果

図 3 に，実験 1 で構築したモデルを用いてテストデータに対して推論を行った正解値と推定値のヒートマップを示す．「椅子から立つ」と「部屋から出る」に関して若干誤りがあるが，平均すると約 92%の精度で識別できている．図 4 に，本モデルを用いて実験 2 で収集した一連の動画に対して推論を行った結果を示す．図 4 左上のグラフは，理想的な順序で行動認識した場合の結果を示し，20 試行分のデータでの結果はほぼ同じだったため代表的な 5 つのグラフのみ掲載する．一連の動画においてもおおむね正確な行動認識ができているが，「椅子から立つ」と「部屋から出る」に関して若干誤りがある．若干の改善の余地はあるが，時刻同期された CSI データへ適切に正解ラベリングを行うことができ，ラベリングやデータ

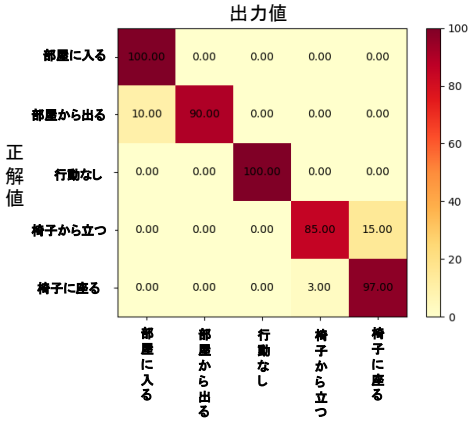

図 3 動画による行動認識モデル推論結果

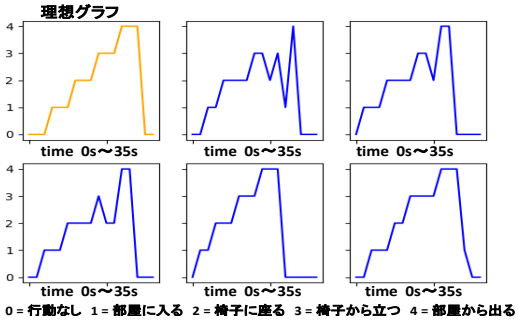

図 4 時間経過で変化する行動ラベル

収集の労力削減が図れる可能性を確認した．CSI で認識が困難な行動に対しても，本動画認識技術を用いてラベリングを行うことで認識可能になる可能性がある．

5. おわりに

本稿では，動画認識技術を用いて検出された行動を正解ラベルとし，CSI へ自動ラベリングすることで正解ラベリングや行動ごとにデータ収集する労力削減を図りつつ，これまで認識できていなかった複雑な状況での行動も対象とする，動画と CSI を用いた行動認識手法の実現可能性を確認した．今後，動画認識精度の向上を図りつつ，実生活における物体の移動，背景変化，体型等の違いによる動画認識精度の変化を分析しつつ，本提案手法の有効性実証を進める．

謝辞

本研究の一部は，東北大学電気通信研究所における共同プロジェクト研究の支援によって行われた．

参考文献

[1] L. Minh Dang, et al.: Sensor-based and vision-based human activity recognition, Pattern Recognition, Vol.108 (2020).
[2] H. Noureldien, et al.: Timeception for Complex Action Recognition, CVPR, pp.254-263 (2019).
[3] G. A. Sigurdsson, et al.: Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, CVPR (2016).
[4] G. Rohit, et al.: Video Action Transformer Network, CVPR, pp.244-253 (2019).
[5] A. Li, et al.: The AVA-Kinetics Localized Human Actions Video Dataset,CVPR (2019).
[6] Y. Siamak, et al.: A Survey on Behavior Recognition Using Wi-Fi Channel State Information, IEEE Commun. Mag., 55(10) (2017).
[7] Kensho Hara, et al.: Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, CVPR (2017).

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.