Vol.2019-CVIM-217 No.14
2019/5/30

深層学習を用いた動画像からのソーシャルタッチ検出

赤塚 大地[1,a)]　中澤 篤志[1]　西田 豊明[1]

概要：ソーシャルタッチは人同士のコミュニケーションの中で重要な役割を担っている. 我々は，ソーシャルタッチを多く利用する介護シーンを想定し，介護中の介護者の左右の手がそれぞれ被介護者の身体に触れているか否か，さらには触れているならば身体のどの部位に触れているかを判定する手法を開発した．提案手法の詳細としては天井に設置されたカメラ画像に対して，介護者を認識する体領域画像と被介護者を認識する体領域画像，オプティカルフロー画像を生成し，接触判定用の CNN に入力することによって手と体の接触状況を判定する．ベースライン手法として CNN にカメラ画像のみを入力する手法も用意し比較したところ，ベースライン手法の精度は 52.8%だったのに対し，提案手法では 61.8%と精度の向上が確認できた．

キーワード：畳み込みニューラルネットワーク，行動認識

Social Touch Detection from Video with Deep Learning

Abstract: Social touch plays an important role in communication between people. We assume a care scene where caregivers use social touch a lot, and We have developed a method that detects touch between their left and right hands while working and a care receiver's body, and furthermore, evaluates whether they're touching, and if so, which part of care receiver's body they're touching. As details of the proposed method, semantic segmentation images for recognition caregivers, semantic segmentation images for recognition care receivers, and optical flow images are generated from frame images by using camera installed on a ceiling, and input them to CNN for detection touch. By doing this, this system detects contact between their hand and body.We prepare a method to input only camera frame images to CNN as baseline and compare it with proposal methods.While the accuracy of the baseline method was 52.8%, the accuracy was improved to 61.8% by the proposed method.

Keywords: convolutional neural network, action recognition

1. はじめに

身体的な触れ合いは人同士の非言語的なコミュニケーション方法の 1 つとして重要な役割がある．このような触れ合いをソーシャルタッチと呼び、我々はこれを使うことで会話情報からは伝わりにくい細かな意思や親密な感情を他者に伝達すること可能としている．ソーシャルタッチは様々な形を取り、例えば相手に対して握手をする、抱きしめる、なでるといった動作を行うことにより相手に好印象を与えたり自身の要望を受け入れやすくさせるといった効果が確認されている [12]．このようにソーシャルタッチは人のコミュニケーションにおいて極めて重要な役割を担っている．

しかしながらインタラクションデバイスが仲介するようなリモートなコミュニケーションが当たり前になった現代において、コミュニケーションを図る際に多くの場面でソーシャルタッチが欠如してしまう傾向にある．現在普及しているリモートコミュニケーションは言語情報・音声情報・映像情報を媒体として行われるが、触覚情報をサポートする通信システムはほとんど普及していない．この場合対話相手との間でしばしば意思の齟齬が生じる．このためリモートコミュニケーションを行う場合に触覚情報を伝達するデバイスを開発 [8] することも重要である．また今後ますます増えるであろう人とコンピュータとのインタラクションにおいてもソーシャルタッチのやり取りを行う手段を確保することが必要であり、これに関しても研究が進め

[1] 京都大学 kyoto University
[a)] akatsuka.daichi.54v@st.kyoto-u.ac.jp

られている [9]. また高齢者や認知症のケアなど介護の現場においてもソーシャルタッチは重要な役割を果たしている．認知症が進行した時に現れる症状の 1 つに暴言を吐いたり、暴力をふるったりすることがある [4]．これらは介護従事者に対する負担を増加させ、さらには介護サービスの質を低下させることにも繋がる．このような症状を緩和させる 1 つの方法としてユマニチュード [6] が挙げられる．ユマニチュードは「見つめる」「話しかける」「触れる」「立たせる」の 4 つのコミュニケーションを柱としている．ユマニチュードを上手く活用するためには介護従事者にそれに関わる技能が要求される．このため石川らの研究 [5] ではユマニチュードの技能を評価する方法が提案されている．石川らは技能評価の手法として gaze,speech,touch 等の個々の動作の集合からなる Intra-modality，各 Intra-modality 間の関係の集合からなる Inter-modality，アイコンタクトや言語対話などのような相手との関係構築ための行為の集合からなる Multimodal-interaction の 3 層に分けて，階層的な評価手法を提案し，その有効性を示している．またユマニチュードの 4 つの技能の中でも本研究に関連する「触れる」は患者に対し、安心感や信頼感を持たせることができ、ユマニチュードの技術として不可欠である．この「触れる」という行為がどのような形でどれぐらいの頻度を行われているかは介護従事者のユマニチュード技術を図る上で重要な指標となる．

以上のように人の触れるという行為には単にコミュニケーションにおいて感情の伝達という役割があるだけではなく，介護の現場においても認知症の人に対するコミュニケーション技術として重要な役割を果たす．そのため，介護の技術の訓練シーンにおいて，触れることを自動的に認識できるようなシステムを使えば，介護従事者のユマニチュード技術の評価を効率的に行うことができ，十分な技術を持った介護従事者の育成の促進することができる．
本研究では上記の接触（ソーシャルタッチ）を認識するためのシステムを構築することを目的とする．具体的には映像には相手の体に触る人と触られる人の 2 人が映るようにし，触る方は介護を提供する側，触れられる方は介護を受ける側を想定する．さらに触れられる人はベッドで横になっている状態であるとして，カメラを両者の上方に設置する．このカメラからの映像をもとに触れる人が実際に触れているか否か，さらには触れているならば相手の体のどの部位に触れているかを識別する．

本稿の具体的な構成は以下のようになる．2 章で関連研究を述べ，本研究の位置付けを明らかにする．3 章では本研究で提案するシステムを述べる．4 章で本研究で提案したシステムの評価，ベースライン手法との比較を行う．5 章では本研究で得られた結果をまとめ，今後の課題について述べる．

2. 関連研究

2.1 行動認識に関する研究

カメラ映像から行動認識を行う研究として Ma らの手法 [7] が挙げられる．Ma らの手法は頭部装着型の一人称カメラ映像から装着者がどのような行為 (Action) をしているか，それがどんな物体 (Object) に対して行われているか，そしてこれら 2 つを統合し，活動 (activity) として認識することを目的としている．一人称カメラ映像として使用しているデータセットとしては GTEA, GTEA gaze, GTEA gaze+の 3 つが用いられている．これらのデータセットは装着者の手と何かしらの物体とのインタラクションを含むような映像を含んでいる．Ma らの手法で提案された行動認識のネットワークは大きく分けて 2 つのストリームで構成されている．片方のネットワークはカメラの装着者の手が何らかの働きかけを行っている対象物を識別するネットワークであり，ここでは Object Net と呼んでいる．Object Net は事前処理として物体のローカライゼーションを行う．セグメンテーションネットワークで自身の腕をセグメントし，それと位置のヒートマップをもとに物体の位置を特定する．その位置情報をもとに画像をクロッピングして，Obeject Net に入力する．もう片方のネットワークは人の動きを識別するためのものであり，ここでは Action Net と呼んでいる．Action Net に対しては Optical flow 画像を入力し，「投げる」「かき混ぜる」といった Action クラスを識別する．これら 2 つのネットワークを 3 層の全結合層で接続し，最終的に装着者が何を使ってどんな動作をしているかというその人の活動を認識する．

この Ma らの手法とは別に同じ一人称視点の行動認識の研究として Singh らの手法 [11] が挙げられる．Singh らは目の動きや手の動き，さらに頭の動きがその人の行動認識の重要な手がかりとなると考え，これらの動きの連携を学習するための Ego ConvNet を導入する．Ego ConvNet には手の動きを捉えるためのハンドマスク画像，頭の動きを捉えるものとして画像を 2D ホモロジー変換したもの，目の動きを見るために顕著性マップ画像を入力し，学習させる．この Ego ConvNet は主に 2D 畳み込み層を用いて構成されたネットワークと時間方向に 3D 畳み込み層を用いて構成したネットワークを最終的に結合させる構造を取っている．以上のように作られた Ego ConvNet に Simonyan らによって提唱された 2 ストリームのアーキテクチャ [10] を追加する．Spatial stream は RGB ビデオフレームを入力し，外観認識する一方で temporal stream には Optical flow 画像を入力し，外観の動きを捉える．Simonyan らのアーキテクチャは UFC-101 などの三人称視点のカメラ映像に対する行動認識ができることが示されているが，Singh らはこれにフレーム画像やフロー画像からでは捉えること

のできないヘッドマウントカメラの装着者の動きを認識する Ego ConvNet を組み合わせた点が革新的といえる．

また一人称カメラ映像からの行動認識だけでなく，そこからその行動の評価も行うシステムを構築する研究もある．それが Bertasius らの手法でバスケットボール選手に一人称カメラを取り付け，その映像をもとに選手のパフォーマンス評価を行うものである．本研究では行動評価までは行っていないが，いずれユマニチュード技能の評価も行うシステムを考えた場合，Bertasius らの研究は応用できる部分があるのではないかと考えている．Bertasius らの評価システムにはビデオとそのビデオに対してプロのバスケットボール選手が付けた評価ラベルがデータセットとして与えられている．手法の詳細としてはまずビデオに対して評価する上で重要な箇所をクロッピングする．これは動きの激しいカメラ映像から効率的にパフォーマンス評価を行うためである．次にプレイ上のイベントを検出する．この手法では３つのイベント，「誰かがシュートを打つ」，「自分がボールを保持する」，「ショットを打つ」を先ほどのクロッピングした画像とプレイヤーのコート上の位置をもとに検出している．最後に混合ガウス分布を用いて評価特徴量を得て，最終的なパフォーマンス評価を行っている．

3. 提案手法

3.1 提案手法の概要

本研究の目的は介護者が被介護者に触れているかどうかや体のどの部位に触れているかを判定するシステムを構築することである．さらに追加条件として図１に示すように被介護者は画像内で頭が左側に位置するようにベッドで横になっており，介護者は被介護者から見て右手，つまり画像内でベッドの下側に起立している状況を想定している．この目的を達成するため本研究では２つの手法を提案する．基本的なアルゴリズムの概要を図２に示す．まず認識対象の動画を各フレームに分ける．そのフレームごとに２種類のセマンティックセグメンテーション生成ネットワークに入力し，結果を得る．ここで，２種類のセマンティックセグメンテーションのネットワークとは介護者をセグメントするためのネットワークと被介護者の体の各部位をセグメントするためのネットワークである．同時にセマンティックセグメンテーションとは別に各フレーム間からオプティカルフローも生成する．オプティカルフローは画像間の各画素の変化を捉えたものであり，画像内の物体の動きを捉えることができる．ここでは主に介護者の動きを見るために導入した．これら２種類のセマンティックセグメンテーションについては次節で，オプティカルフローについては次々節で詳しく説明する．この各フレームやフレーム間に対して生成した合計３種類の画像を接触判定のネットワークに入力する手法が提案手法１である．さらにこの３種類の画像に加えて，元のフレームの画像も接触判定のネットワークに入力する手法が提案手法２である．またこれらに加え，Baseline 手法として接触判定のネットワークにフレーム画像のみを入力する方法も実装した．以上の３つの手法の接触判定ネットワークの詳しい構造については本章の第４節で説明する．

図 1　入力 RGB 実画像の例

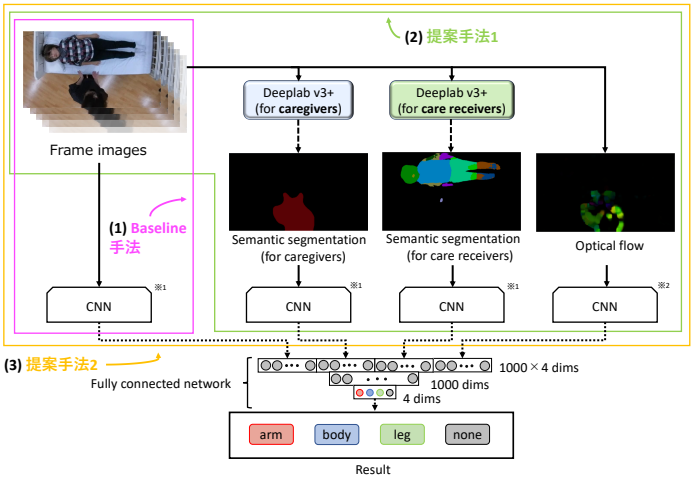

図 2　提案手法の概要図

なおこのシステム内の接触判定ネットワークを学習あるいはテストするために使用した動画は kodak カメラを使って撮影した．

3.2 体領域の認識

この節では介護者を認識するためのセマンティックセグメンテーションと被介護者の体の部位を認識するためのセマンティックセグメンテーションの２種類のセグメンテーションについて説明する．

3.2.1 介護者の認識

接触判定のネットワークに画像内のどの画素が介護者を表しているのかを示すためにセマンティックセグメンテーションの技術を用いてセグメンテーション画像を生成する．つまり与えられた画像に対して立っている状態の人とそれ以外の背景の部分の２つに塗り分けた画像を自動的に作れるようにすることが目的である．具体的には本研究では Chen らによる deeplab v3+[2] を用いた．学習には Kinect v2 を直立した人を真上から見下ろす位置に設置し，

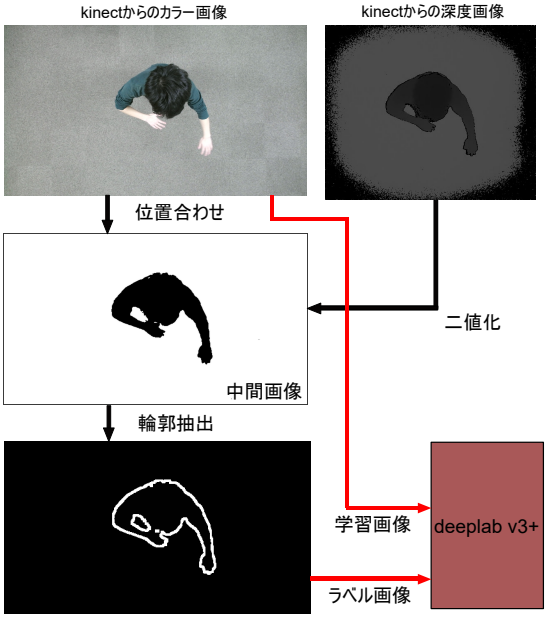

図 3　deeplab v3+学習データのラベル画像の生成

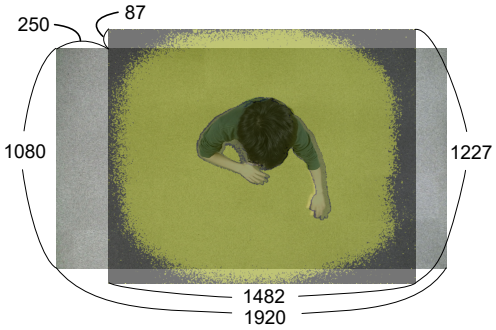

図 4　カラー画像と深度画像の位置合わせ

得られた距離画像から人領域の切り出しと対応するカラー画像を取得し，セグメンテーションの学習に用いた．具体的には，カラー画像とそれに対応する深度画像から図 3 の中央の中間画像を生成する．これは深度画像の人の部分を黒に，それ以外の背景を白にし，さらに図 4 のように大きさがカラー画像と同じになるように拡大と位置調整を行った．なお図 4 は見やすさのため深度画像を黄色に着色している．また Kinect v2 のカラーカメラと深度カメラは物理的な位置が異なるためそれぞれ撮影される画像もわずかなズレが生じている．こうして作成した中間画像に対してさらに人の輪郭を強調する操作を行う．これは deeplab v3+がラベル画像として受理できるような形式にするためである．ここで強調された輪郭は白色 (明度 255)，背景は黒色 (明度 0)，人の部分は明度 1 にする．このようにして深度画像をラベル画像に変換する．

実際に学習データを収集する際には服の色にセグメンテーションが影響されないように注意する必要がある．このため今回は赤・青・緑・黒・白の 5 色の服を着ながら撮影を行った．さらに本研究の目的は介護者の接触判定なので

図 5　介護シーン画像にセマンティックセグメンテーションを適用した結果

介護者の腕・手を見ることは非常に重要であると考えられる．このセグメンテーションを行う理由の 1 つに接触判定のネットワークに明示的に介護者の場所を示すことによって同時に介護者の腕の場所も示すことを期待している．このためできる限り介護者の腕が上手くセグメンテーションできないという事態を避けられることが望ましく，本研究では各色の服に対し長袖と半袖の 2 パターンを用意し両方の場合でも上手くセグメンテーションできるように配慮した．

上記の点に留意しながら 5fps で学習データを 13750 枚撮影した．このデータセットを 2 つに分割し，片方は 12723 枚の訓練データ，もう片方は 1027 枚のテストデータとした．このデータセットを用いて実際に学習させた結果，図 5 に示すようなセグメンテーション画像を得ることができた．セグメンテーション画像内の赤色の箇所が介護者を意味しており，それ以外の黒色の箇所は背景を表している．図 5 の場合では上手く介護者のみをセグメンテーションできており，期待通りの画像を得ることができたが，介護者が被介護者を覆い隠している場合など状況によっては被介護者の腕や足を余分にセグメントすることもあった．また立っている人が複数人いる状況下ではその人々を全てセグメントした画像が生成される．なお図 4 や図 3 では介護者が画像内の下側を向いているが，最終的には上下を反転させて画像を生成している．

3.2.2　被介護者の認識

接触判定のネットワークに画像内のどの画素が被介護者を表しているのか，さらには体のどの部位を表しているかを示すためにセマンティックセグメンテーションの技術を用いてセグメンテーション画像を生成する．先ほどの介護者をセグメンテーションする場合とは異なり，横になっている人をセグメンテーションし，また頭部・胴体・上腕・前腕・腿などより細かく見分けられなければならない．このため介護者をセグメンテーションする際に用いたオリジナルの学習データとは別のデータセットを用いる．

本研究で用いたデータセットは PASCAL VOC で配布されたアノテーション付きの画像データセットである PASCAL Part Dataset[3] である．これを学習データとして deeplab v3+を学習させ，実際にセグメンテーションした結果が図 6 である．図から分かるように介護者をセグメンテーションすることなく，上手く被介護者の各部位が塗

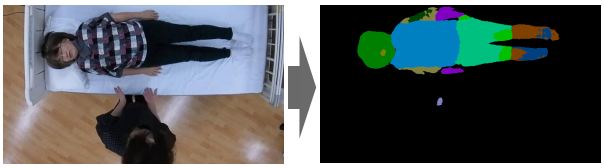

図 6　PASCAL Part Dataset で学習させたセマンティックセグメンテーション

り分けられていることが確認できる．なおこの PASCAL part Dataset には直立した人を横から撮影した画像を多く含んでいるため，単純に図 6 中の実画像をテストデータとしてそのまま deeplab v3+に入力しても上手くセグメンテーションしてくれない．そこで 1 度入力画像を 90°時計回りに回転させたものを入力して，セグメンテーション画像の生成後に元に戻している．しかしながら介護者のセグメンテーションの時と同様に介護者が被介護者の大部分を覆い隠している場合は上手くセグメンテーションできないことがあった．

3.3　Optical flow 画像

Optical flow はフロー推定の 1 種で，画像間の差を捉えることができる．具体的には図 7 に示すように画像内の各画素の動きのベクトルの向きを HSV 色空間の色相に，大きさを明度に置き換えて表現している．基本的に被介護者はベッドで横になっており，ほとんど動かないと仮定しているため，今回の場合動画内で動くものは介護者しかなく Optical flow 画像では介護者の動きのみを捉えていると考えることができる．また今回は Optical flow の手法の 1 つである Gunnar Farneback 法を利用した．

Optical flow は 2 枚の画像間で定義されるので n 枚の画像に対しては n-1 枚が生成る．そのためここでは最後に生成された n-1 番目の Optical flow 画像を複製し，n 番目の画像とすることで実画像と Optical flow 画像の数を合わせている．また各実画像に対して連続する 4 フレーム分の Optical flow 画像を対応させて，接触判定のネットワークに入力している．つまり k 番目の実画像には k-3,k-2,k-1,k 番目の Optical flow 画像を対応させていることになるが，ここで 1,2,3 番目の実画像は-2,-1,0 番目の Optical flow 画像が存在しないために割り当てることができない．このため 1 番目の Optical flow 画像を-2,-1,0 番目の画像として複製することによってこの問題を解決した．

3.4　ネットワーク構造

この節ではベースライン手法と 2 種類の提案手法で使われている接触判定のネットワークの具体的な構造について説明する．

3.4.1　Baseline 法: RGB 画像のみによる方法

まずベースライン手法では RGB 実画像のみを接触判定の

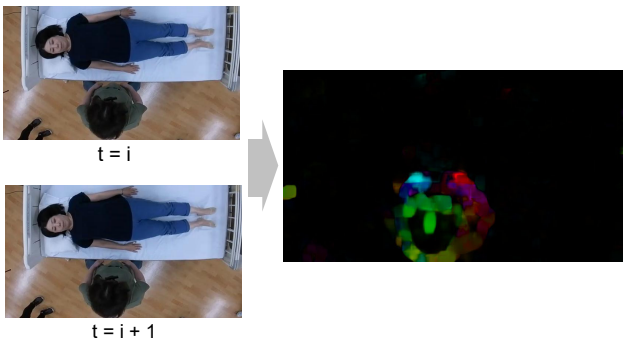

図 7　2 画像から生成されたオプティカルフロー

ネットワークに入力する．この手法は後に説明する 2 種類の提案手法が出した精度がどの程度良いものなのかを測るための比較手法である．そのため出来るだけシンプルなカメラからの映像をそのまま使う方法を採用した．RGB 実画像をネットワークに入力する際には 224*224[pixel] にまで縮小している．よってネットワークの Input は 224*224*3channel となっている．このネットワークは主に Chatfield らが提案した CNN-M[1] をもとに設計している．ネットワークは大きく分けると 5 つの畳み込み層と 3 つの全結合層からなっている．1 つ目の 2 次元畳み込み層はカーネルが 7 × 7，フィルター数は 96 で，この層の次に接続している LRN は局所的応答正規化 (Local Response Normalization) の略である．その後 MaxPooling を経て縮小させる．2 層目にはカーネル 5 × 5，フィルター数 256 の 2 次元畳み込み層，LRN，ゼロパディング，MaxPooling を経る．その後カーネル 3 × 3，フィルター数 512 の 2 次元畳み込み層，ゼロパディングを 3 回設けて最後に MaxPooling を経る．ここまでが前半の畳み込み層部分でこれを Flatten 層で平滑化してから 3 つの全結合層で 4096 次元→ 1000 次元→ 4 次元にして 4 クラス識別をしている．また全結合層間で 50 %のドロップアウトも行っている．なお 5 つの畳み込み層と前半の 2 つの全結合層では活性化関数を ReLU 関数に設定して，最後の全結合層のみ softmax 関数にしている．

3.4.2　提案手法 1: 体領域画像と optical flow 画像による方法

介護者認識用と被介護者認識用のセマンティックセグメンテーション画像と Optical flow 画像をネットワークに入力する手法を提案手法 1 とする．3 つの Input には Optical flow 画像，セマンティックセグメンテーション画像を入力する．図 7 に載せた Optical flow 画像は 3channel のカラー画像に見えるが Optical flow は各画素に対してベクトルを割り当てたものなので 2channel の画像で表現することができる．そのため 2channel で 224*224[pixel] の Optical flow 画像を 4 フレーム分入力するため Input は 224*224*8channel となる．このネットワークのマージ層

よりも入力側はベースライン手法のネットワークから最後の全結合層を除いたものと全く同一のものが 3 つ並列になった構造になっている．マージ層以降は 2 つの全結合層で 1000 次元→ 4 次元にして 4 クラス識別を行っている．またベースライン手法と同様に活性化関数は最後の全結合層のみ softmax 関数でそれ以外は ReLU 関数である．

3.4.3 提案手法 2: 体領域画像・optical flow 画像・RGB 実画像による方法

提案手法 1 をさらに RGB 実画像も利用するように拡張した手法を提案手法 2 とする．提案手法 2 では提案手法 1 で 3 並列構造だったものをそのまま 4 並列にし実画像が入力できるようにした．新たに追加したネットワーク部は Input は実画像が RGB 形式の画像なので 224*224*3channel となっている

4. 実験

実験は各手法を用いて 4 つの動画に対して交差検定を行った．表 1 に示すように実験 1 ではテストデータとして動画 1 を，訓練データはそれ以外の動画を使って実験を行い，実験 2 ・実験 3 ・実験 4 も 1 つの動画をテストデータにして，それ以外の動画を訓練データにした．また各実験には右手と左手をそれぞれ別々に学習させたので，最終的には各手法に対して 8 回学習をさせたことになる．なお表中の括弧内の数値は総フレーム数，つまり訓練/テストデータ数である．

また今回実験を行うのに際して NVIDIA 社の GeForce GTX 1080 Ti を 2 つ用いて学習させた．さらにすべての手法において学習時のバッチサイズは 32，学習率は 0.0002，実行 epoch 数は 30 に設定した．

下にベースライン手法・提案手法 1，提案手法 2 の実験時のパラメータ・学習時間についてそれぞれ説明する．なお各実験の右手と左手における 1epoch 当たりの学習時間は同じである．

- ベースライン手法
  - メインメモリ容量：64GB
  - 1epoch 当たりの学習時間：
    実験 1 -134s，実験 2 -145s，実験 3 -115s，実験 4 -126s
- 提案手法 1
  - メインメモリ容量：128GB
  - 1epoch 当たりの学習時間：
    実験 1 -296s，実験 2 -328s，実験 3 -262s，実験 4 -299s
- 提案手法 2
  - メインメモリ容量：128GB
  - 1epoch 当たりの学習時間：
    実験 1 -392s，実験 2 -415s，実験 3 -320s，実験 4 -383s

図 8 判定結果の一例

4.1 結果と考察

図 8 に判定結果を可視化したものの一例を示す．これは提案手法 2 において実験 2 の左手を学習させて，その学習モデルを用いて動画 2 を判定させた時の 1 フレームである．図中の Predict は判定結果，GT は Grand Truth の略，その下の 4 つの青いバーは各クラスにおける予測値を表している．この場合だと arm の予測値がもっとも大きくなっており，正しい判定結果が出ている．

さらに表 2 にベースライン手法・提案手法 1 ・提案手法 2 の 15epoch 目における実験結果を示す．表の各行は 3 つの各手法を用いて行った 4 つの実験それぞれの右手と左手の適合率の平均値と再現率の平均値を載せた．また正答率は 4 つの実験での正答率をそれぞれのテストデータ数をもとに加重平均を取った値である．また表中の合計とは各手法における右手と左手の正答率の平均値である．この合計が各手法における本実験の精度となる．なお表中の数値の単位は全て%である．適合率と再現率と正答率については以下のように定義する．

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

表 2 の合計の列を比較するとベースライン手法は 53.0 %，提案手法 1 は 61.8 %，提案手法 2 は 61.8 %となっている．つまり提案手法 1 と 2 の間にはほとんど精度の差はなく，一方で 2 つの提案手法とベースライン手法の間には約 8 %分から 9 %分の差が見られるため，2 つの提案手法が単純な RGB 実画像をネットワークに入力する方法よりも優れていることが分かる．しかしながら提案手法 1 と提案手法 2 の間に差がないという結果はフレーム画像が精度向上に貢献していない，また接触判定のネットワーク内でフレーム画像に対して極めて小さい重みしか学習していないと考えられる．

さらにより細かく見ると各手法において左手の正答率が右手の正答率を平均して約 5.57 %分上回っている．この

	実験 1	実験 2	実験 3	実験 4
訓練データ	2,3,4（22388）	1,3,4（24327）	1,2,4（18620）	1,2,3（22142）
テストデータ	1（6771）	2（4832）	3（10539）	4（7017）

表 1　各実験で用いた動画の番号とデータ数

手法名			none	arm	body	leg	正答率	合計
ベースライン手法	left	適合率	36.1	50.3	74.6	84.6	57.6	53.0
		再現率	35.9	39.1	85.7	75.5		
	right	適合率	27.6	58.6	34.6	57.0	48.3	
		再現率	21.5	18.7	25.9	88.7		
提案手法 1	left	適合率	50.6	58.9	77.3	85.3	63.1	**61.8**
		再現率	27.3	63.2	85.4	80.5		
	right	適合率	38.1	72.3	45.8	72.1	**60.5**	
		再現率	36.7	59.5	26.8	85.2		
提案手法 2	left	適合率	52.4	61.4	76.9	86.4	**64.2**	**61.8**
		再現率	28.8	63.7	86.3	81.3		
	right	適合率	33.5	72.5	41.7	70.9	59.4	
		再現率	34.2	54.8	23.6	86.8		

表 2　ベースライン手法・提案手法 1 ・提案手法 2 の実験結果

図 9　判定失敗の例 1

図 10　判定失敗の例 2

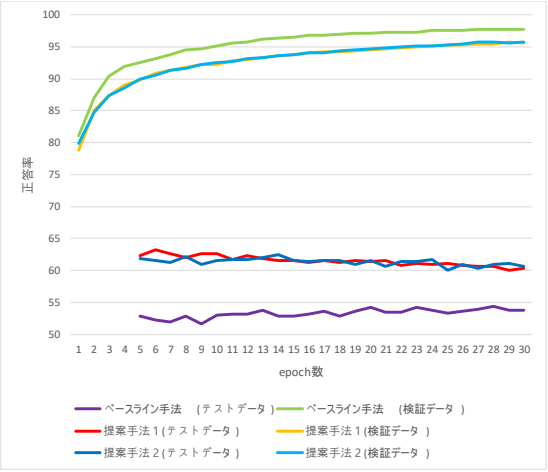

図 11　各手法における検証データとテストデータに対する正答率

ように左右の手で精度に差が生じる原因として被介護者の頭が左になるようにカメラを設置するという条件が考えられる．介護者の左側には常に被介護者の頭があり，右側には被介護者の足があることから映像が左右非対称となり，結果としてこのような右手・左手の精度の差が生じた．

またクラス別の適合率と再現率を見ると全ての手法で none の適合率と再現率が共にその他のクラスのものと比較して低くなっていることが確認できる．つまり本来相手の体に触れていないフレームでどこかに触れていると判定してしまったり，あるいはどこかに触れているのに触れていないと判定している事象が多発しているのであるが，これに関連する判定ミスの一例として図 9 のような場面がある．図 9 は提案手法 2 を用いて左手について実験 2 を行った結果を可視化したものだが，この画像内では実際に触れいないフレームで腕に触れていると判定している．このように介護者の腕が被介護者の上部に空中に浮いているような状況下では天井に設置されたカメラからでは腕の精密な高さまでは識別することはできず，結果として判定に失敗する．本実験ではこのような事象が散発していた．

さらに別の失敗例として図 10 に示すような場面で判定ミスが見られた．図 10 は図 9 と同じ手法・実験での左手の判定結果であるが，この場合介護者の腕が介護者自身の頭や胴体に隠れて画面内に映っていない．このような場合も非常に判定ミスが起きやすいことが確認された．

以上のような失敗例に対してはカラー画像のみが撮影できる天井カメラではなく，深度カメラの撮影できるカメラを設置することによって介護者の腕の高さを認識させるといった方法やまた介護者の腕が隠れている場合は「判定不能」という結果を出すようにして，システムを信頼性を向上させるといった方法が考えられる．また今回確認された実験時の問題点として学習データ数の少なさが挙げられる．図 11 は 3 つの手法それぞれのテストデータと検証データ

を各 epoch で学習されたモデルを用いて判定させた時の正答率である．検証データは学習データの 40 %を割り当てている．図から提案手法 1 ・ 2 ではテストデータの正答率と検証データの正答率との間に約 35 %分ほどの極めて大きな乖離があることが確認できる．これに対処するにはデータ数を増やしたり，データとして利用する動画の種類を増やす必要があると考えられる．

5. 結論・今後の課題

本稿ではソーシャルタッチの重要性から人の触れるという行為の検出の自動化が意義を持つという研究背景に基づいて，深層学習を用いた触れる動作の検出を行うシステムの構築を目的とした．本システムを用いた実験の結果，ベースラインとなる手法と比較した場合，我々が提案した手法を利用した方が精度の向上が確認でき，提案手法の有用性が示された．一方，4 章で述べたように腕の高さを認識できない，あるいは腕が天井カメラからの映像には映らないといったことから生じる誤判定も散見され，精度低下を招く原因となった．また学習データの多様性の不足に原因があると考えられる問題も確認された．

さらなる精度向上を目指すには上記の問題点を克服することが不可欠である．解決法として深度カメラの増設や判定不能クラスの追加，学習データの追加が挙げられる．一方，深度カメラの増設などは非常に有効な方法であると考えているが，システムの要求環境を安易に複雑化させることはシステムのユーザビリティの低下を招く可能性もあるので慎重に検討することが必要である．

また今回提案したシステムには被介護者が画面の左側を頭にして横になっているという使用条件が課せられている．現実の介護現場では寝たきりの高齢者ばかりではなく，被介護者がベッドの脇に座るなどの状況も十分考えられる．そのような状況下でも正しい接触判定をできるようにすることはシステムの汎用性を向上させることに繋がる．さらに本システムは介護者が相手の体のどこに触れているかを見る際，腕・胴体・足といった大まかな位置しか判定できない．しかし実際に腕 1 つ取っても，手に触れているか，あるいは上腕に触れているかで相手に与える心理的な影響や感情のニュアンスが異なる場合がある．システムを介護者のユマニチュード技能の評価に利用するならば，このような細かな体の位置も判定できることが望ましい．さらに本システムをより有用性の高いものにするならば，触れているかどうかやどこに触れているかだけでなく，相手の体にどのように触れているかが認識できるとなお良い．例えば相手にそっと優しく触れる場合と乱暴に触れる場合では与える印象は真逆になる．このように考えれば触れ方はユマニチュードの「触れる」という技能の重要な要素の 1 つであることが分かる．以上のような点についてシステムを改善することが今後の課題となると考えている．

謝辞 本研究は科研費 17H01779, 26249029, 15H02738, および, JST CREST, JPMJCR17A5 の支援を受けている．

参考文献

[1] Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets, *arXiv preprint arXiv:1405.3531* (2014).
[2] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation, *arXiv preprint arXiv:1802.02611* (2018).
[3] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision*, Vol. 88, No. 2, pp. 303–338 (online), DOI: 10.1007/s11263-009-0275-4 (2010).
[4] for Health, N. I. and Britain), C. E. G.: *Dementia: Supporting People with Dementia and Their Carers in Health and Social Care: Quick Reference Guide* (2006).
[5] Ishikawa, S., Ito, M., Honda, M. and Takebayashi, Y.: The skill representation of a multimodal communication care method for people with dementia, *JJAP Conference Proceedings*, Vol. 011616, p. 4 (online), DOI: 10.7567/JJAPCP.4.011616 (2016).
[6] Ito, M. and Honda, M.: An examination of the influence of Humanitude caregiving on the behavior of older adults with dementia in Japan, *Proceedings of the 8th International Association of Gerontology and Geriatrics European Region Congress* (2015).
[7] Ma, M., Fan, H. and Kitani, K. M.: Going deeper into first-person activity recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1894–1903 (2016).
[8] Nakanishi, H., Tanaka, K. and Wada, Y.: Remote handshaking: touch enhances video-mediated social telepresence, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 2143–2152 (2014).
[9] Silvera-Tawil, D., Rye, D. and Velonaki, M.: Interpretation of social touch on an artificial arm covered with an EIT-based sensitive skin, *International Journal of Social Robotics*, Vol. 6, No. 4, pp. 489–505 (2014).
[10] Simonyan, K. and Zisserman, A.: Two-stream convolutional networks for action recognition in videos, *Advances in neural information processing systems*, pp. 568–576 (2014).
[11] Singh, S., Arora, C. and Jawahar, C. V.: First Person Action Recognition Using Deep Learned Descriptors, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2620–2628 (2016).
[12] Toet, A.: Social Touch in HumanComputer Interaction, *Frontiers in Digital Humanities*, Vol. 2, pp. 1–13 (online), DOI: 10.3389/fdigh.2015.00002 (2015).