修士論文

単眼カメラを配置した腕輪型デバイスを用いてクリックジェスチャ認識を実現する手法の提案

五十嵐大和

主指導教員 宮田 一乗 教授

北陸先端科学技術大学院大学先端科学技術研究科（先端科学技術専攻）

令和4年3月

概要

単眼カメラを手の甲が映るように配置した腕輪型デバイスから得た映像を処理することで物理的なマウスと同様のクリック認識を実現する手法を提案する．本研究では，スマートウォッチのような腕輪型デバイスにそのようなカメラが搭載されることを想定している．実際に昨今のスマートウォッチには写真撮影目的でカメラが埋め込まれたモデルも発売されている．しかしこの場合，カメラの映像では指の動きは観測できず，マウス操作では手全体の動作もわずかである．そのため，マウスのクリックジェスチャといった小さな変化を手の甲の映像のみから予測可能であるかどうかを調査する．本稿では手の輪郭や手の甲の僅かな変化を詳細に観測するため「ClickNet」を設計した．ClickNet はクリックジェスチャを予測する CNNLSTM ネットワークである．ClickNet を学習させるため被験者を募りデータ収集を行い，ラベル付けを行うことでデータセットを作成した．このデータセットを用いてモデルを学習させ評価をした結果，ジェスチャ予測の F1 スコアが 0.88390 となった．

本研究では次のような新規性がある．まず手首に装着したデバイスを用いることでウェアラブルでない動作推定手法と比べ，携帯性に優れるとともに，ユーザが努力せずとも手を継続的に視野に収めることが可能である．次に小型の単眼カメラを用いるため特別なセンサが不必要であるとともに，既存のスマートウォッチの内蔵カメラを活用することも考えられるためコストを低く済ませることができる．アクセサリとして腕時計を日常的に身に付ける場合が多く，スマートウォッチという腕時計の形状をしたデバイスも普及しているため，抵抗感なく身に付けやすい形のデバイスになっている．さらに，既存手法で問題となっていたセマンティックセグメンテーションやオプティカルフローの計算コストの問題を蒸留を使うことで解決した．

目 次

図 目 次

表 目 次

第1章　はじめに

近年，ヘッドマウントディスプレイ (HMD) を用いる VR/AR デバイスや小型タッチパネルを用いる時計型デバイスが普及している．HMD 型デバイスではボディトラッキングやコントローラを，時計型デバイスでは身につけた腕時計サイズのタッチパネルをそれぞれ入力インターフェースとして用いる．これらの入力インターフェースを用いることで，ハンズフリーの利便性や没入感を実現している．一方，上記の入力インターフェースでは，マウスによるカーソル操作やスマートフォンなど大きな操作面でのタッチ操作の場合と比べ，テキスト入力や精密なポインティングといった細かい動作を行うことが困難である．ポインティングや決定といった動作は GUI とのインタラクションでは基本的な動作であり，GUI 操作のしやすさなどのようなユーザーエクスペリエンス (以下 UX) に大きく影響すると考えている．これをモバイルな環境やバーチャル空間内など物理マウスが手元にない，持ち歩いていない場合でもある場合と同様にマウスに似た入力方法を利用可能にするシステムを提案する．

課題の解決策としてハンドトラッキングにより手の動きを追跡し入力として利用することが挙げられる. 既存のハンドトラッキングデバイスとして, データグローブなど手袋型のデバイスや LeapMotion[1] といった設置型のトラッキング用カメラがある．さらに，ハンドトラッキング/ハンドジェスチャ認識の研究は盛んであり，設置型，ウェアラブル型の双方で研究が進められている．既存手法をモバイル環境で利用することを考えた場合に次のような問題点が挙げられる．

- **携帯性** - Kinect[2]など設置型のカメラが必要な場合は持ち運びに難がある．
- **トラッキング範囲** - カメラの範囲外に出るとトラッキングが不可能である．
- **デバイスのサイズ** - グローブ型やセンサが大きい場合場合手の動きを阻害する場合がある．
- **特別なセンサ** - 赤外線カメラ [43] や EMG[43]，FSR[17] などの特殊なセンサが必要となる．

これらの課題点やモバイル環境，仮想空間内での利用を想定し，スマートウォッチの利用を考えた．腕に装着するデバイスはトラッキング対象の手を常時観測可

[1] https://developer.leapmotion.com/
[2] https://developer.microsoft.com/ja-jp/windows/kinect/

能であるとともに小型であり，携帯性，トラッキング範囲，デバイスサイズの問題を解決できる．加えて，スマートウォッチの側面に埋め込み可能な単眼カメラを用いた手法を構築することで，特別なセンサをデバイスに埋め込む必要がなくなるためコスト面でも有用であると考えている．

腕輪型デバイスに装着した単眼カメラを用いてハンドトラッキングを行う手法として BackHandPose[88] がある．BackHandPose は RGB 映像に加えて Opisthenar[90] で用いられた Modified MHI(Motion History Image[8]) を利用することで皮膚の変形を観察し，ハンドトラッキングを可能にしている．

既存手法から腕輪型デバイスを用いたハンズフリーマウスを実現することは可能であると考えた．ハードウェアとして小型魚眼カメラを設置した腕輪型デバイスを設計した．このデバイスの映像から右/左クリックジェスチャを認識するため，BackHandPose で提案された DorsalNet を参考にネットワーク「ClickNet」を設計した．システムはまず，映像から手のみを抜き出すため FastSCNN[61] を用いた Hand Segmentation を行い，これを処理することで OpticalFlow を計算する．OpticalFlow は ClickNet によって処理され，それぞれのジェスチャである確率が出力として返される．

手法の有効性を確認するため，データを集めデータセットを作成するとともにそのデータセットを用いてネットワークを学習させることで精度を検証した．指標として F1 スコアを用いて検証した結果，全体のデータを用いた検証では F1 スコア 0.88385，個人データを用いた交差検証では最高で F1 スコア 0.91957，平均で 0.86788 を得た．

本稿は次のような新規性がある．

1. マウス動作の際の僅かな動きを検知し予測可能なネットワーク：マウス動作によって起こる手の甲の変化を観察し，ラベル付けを行うことにより，手の甲の動きのみでクリック操作の学習が可能になった．さらにオプティカルフローを利用することにより手の甲の細かな変化を学習しやすくしている．

2. モバイルデバイスに適した小型のネットワーク：MobileNetV3 のアーキテクチャを活用することで関連研究より小型かつ CPU 上での計算に適したネットワークになっている．

3. MARS[14] を活用した処理の高速化：MARS の蒸留方法を活用し，RGB 画像の入力のみで，OpticalFlow を入力としたモデルと同様に動き情報をエンコードすることが可能になっている．これにより OpticalFlow の計算（平均処理時間：0.09373 秒）とセマンティックセグメンテーションの処理 (平均処理時間：0.00491 秒) が不要になり，平均 0.00521 秒で処理可能になった．

第2章　関連研究

本稿に関連する研究領域として，以下が挙げられる．本章ではこれらの既存研究について述べ，最後にこれらの研究と比べた本研究の利点を述べる．

- ハンドジェスチャ認識/トラッキング
- ニューラルネットワークを用いた画像処理
- バーチャルマウス/ポインティングデバイス

2.1　ハンドジェスチャ認識/トラッキング

2.1.1　ウェアラブルでない手法

ウェアラブルでない方法として，まずマーカーを用いたモーションキャプチャ手法 [27] が挙げられる．マーカーベースの手法は高精度にトラッキングが可能であるがトラッキング対象へのマーカー装着に加え，外部にも器具 (カメラなど) の設置が必要である．そのため昨今では，マーカーレスの手法が研究されている．カメラベースのトラッキング手法としては，まず，映像にモデルフィッティングを適用し，手の形状を再構築する手法が考案された．この中には複数のカメラを使用する手法 [55] や深度カメラを用いる手法 [34, 62]，単眼カメラ一つで推定を行う研究 [48] がある．昨今ではニューラルネットワークを用いた手法が広く使われるようになり，画像から直接，推定を行う手法 [25, 35, 69, 75, 81, 94] が主流になっている．ニューラルネットワークにより，カメラのみで高精度なトラッキングが可能になったが，これらの手法は外部 (手が観測可能な位置) に設置したカメラが必要であり，携帯性に難があるため日常の様々な場面で同様に使いたいと考えた場合にふさわしくない．

カメラ以外を用いた手法としては LED[50]，Wi-Fi[42, 79, 78]，音響センサ [60] を利用した研究が存在する．これらは精度，測定可能な距離，そして指の動きが推定できないなど自由度が少ない．

ウェアラブルでない手法の中では J.Song らの論文 [74] がスマートフォンのカメラを活用しており，モバイル環境での活用も期待できる．しかし，片手でカメラを持ち，もう片方の手でジェスチャをするため，両手がふさがってしまう問題がある．

2.1.2 ウェアラブルデバイスを用いた手法 (腕以外に装着するデバイス)

体にデバイスを設置することで外部のセンサが必要なくなり，モバイルな環境で利用することが可能になる．

手，腕以外に装着する手法として，肩に載せる手法 [73]，胸部に設置する手法 [33, 65]，靴に設置する手法 [5]，帽子に装着する手法 [89] がある．しかし，手がカメラの視野に入るように手を保つ必要があり，疲労が発生する．このようにタッチパネルやトラッキングを用いたインタフェースを利用していて起こる疲労はゴリラ腕 [37] と呼ばれており問題になっている．さらに，最近普及し始めている HMD を用いた手法 [38] もあるが，OculusQuest[1]，Microsoft Hololens[2]，HTC Vive[3] など，HMD 型の AR/VR に最適化された手法であり，モバイル環境での利用に適していないとともに，上で挙げたゴリラ腕問題が発生し得る．

手に装着する手法としてセンサ付きグローブを装着する手法 [12, 13, 26] が一般的である．グローブ型ハンドトラッキング分野のサーベイ論文として [19, 76] がある．グローブ型デバイスは高精度なトラッキングが可能なため，映画や産業で利用されるがデバイスのサイズや価格の面，そして装着にかかる時間など問題があり，日常生活での利用には適していないと考えられる．他に昨今に製品化された技術として OculusTouch[4]などのコントローラ型のハンドトラッキングがある．近接センサや静電容量センサにより手の部分的なトラッキングを可能にしており，Arimatsu らの論文 [2] や AirPicher[68] でも同様手に持つタイプのデバイスが提案されている．その他に手の甲にセンサを付ける手法 [52, 40, 47]，指輪型デバイスを装着する手法 [9] なども提案されているが，限定的なジェスチャしか認識できない，特別なセンサが必要となるなど日常的な利用には適さないと考えている．

2.1.3 ウェアラブルデバイスを用いた手法 (腕に装着するデバイス)

人間は日常的に腕時計を身に付けていることから，他の形状をしたデバイスと比べ腕輪型デバイスはほかの形状のデバイスに比べ親しみやすい傾向にあり，特にスマートウォッチは腕時計の代替として人気になっている．そのため腕輪型デバイスを用いたハンドジェスチャ認識，ハンドトラッキングに関する研究が盛んに行われている．

センサを用いた手法として，EMG[58, 67, 53]，FSR[17],EMG+FSR[58]，オプティカルセンサ [23, 56]，静電容量センサ [63, 82, 51] を用いた手法は手の表面の変化を観測することでジェスチャ認識をしている．特に EMG を用いた手法に関し

[1]https://www.oculus.com/quest/
[2]https://www.microsoft.com/ja-jp/hololens
[3]https://www.vive.com/jp/
[4]https://www.oculus.com/rift-s/

てはMyo[5]として製品化されており，5つの異なるハンドジェスチャを認識可能になっている．他に，EIT[92, 93]，生体音響信号 [49, 1, 18, 91]，超音波 [57, 71] を利用した手法があり，これらは手や腕の内部で起こる変化を観測することでジェスチャ認識をしている．NIR[56, 44] や距離センサ [23] を用いた形状認識による手法もある．これらの手法の欠点は特別なセンサが必要であり，ウェアラブルデバイスに組み込まれる可能性が低いことである．

一つの例外として挙げられるのがIMUを用いたSerendipity[86]である．近年のウェアラブルデバイス (ex. スマートフォン，スマートウォッチ，HMD) にIMUが組み込まれていることが多いため，モバイル環境での手法を考える際にIMUを利用することは合理的だと考えている．しかし，識別可能なジェスチャは5つのみであり，さらに50Hzで取得したデータを他のサーバに送信しリアルタイム計算を可能にしているためモバイル環境での利用は難しい．

カメラを用いた手法で一般的なものが手の内側を撮影する手法である．このタイプの手法としてIRレーザ+IRカメラを用いるDigits[43]，加速度センサとLEDのフラッシュを用いるDigiTap[84]，単眼カメラの映像から推定を行うWristCam[84]がある．これらの手法は指先が映像に映っているかどうかに大きく依存する問題がある．次に手の外側 (甲) の映像から推定する手法がある．このタイプのデバイスは指先が見えにくくなるため，ジェスチャ推定が難しくなる．Chenら [10] は高い位置にカメラを配置することで指先を観測可能にし，ASL(American Sign Language)5種類を推定した．Opisthenar[90] では，IRカメラを利用し手の甲の変形を観測することで，11種類のASLと指のタッピングジェスチャの推定を行った．BackhandPose[88] ではOpisthenarの手法を単眼カメラで可能するとともに，ネットワークに変更を加えることでハンドトラッキングも可能にした．手の内側と外側の両方にカメラを設置する方法として，FingerTrak[32] がある．FingerTrakでは4つのサーマルカメラを腕の周囲に等間隔で配置することで手のシルエットを取得し，そのシルエットを学習させることでハンドトラッキングを行っている．

2.2 ニューラルネットワークを用いた画像処理

昨今ではジェスチャ認識，ボディトラッキングを含む画像処理分野においてニューラルネットワークを用いたアプローチが主流となっている．CNN(畳み込みニューラルネットワーク) により画像の特徴を学習することが可能になり，それにより画像から手の形状を認識することが可能になった．2020年には自然言語処理の分野でSOTAとなっているTransformerを画像処理に適用したViT(Vision Tranformer)[21] が提案された．さらに2021年にはMLPを画像処理に適用したMLP-Mixer[80] が提案された．しかしこれらのモデルは画像一枚毎に計算するため，動画を処理する場合には適していない．ここでCNNを拡張し動画認識を行え

[5]https://www.bynorth.com/

るようにする研究が行われている．この分野で提案されているネットワーク構造として，CNN と RNN を組み合わせた CNN-LSTM や LRCN[20]，元の映像とオプティカルフローを入力として別々に計算を行い最後に結果を結合する two-stream network[72, 24]，3 次元で畳み込みを行う 3D CNN[39] などがある．3D CNN は動画認識分野で良く用いられている手法だが推論時間，消費メモリの観点でコストが大きい．Transformer をベースとした動画認識モデル [7, 3, 59] も提案されている．

2.3 ポインティングデバイス/インタフェース

ポインティングデバイスとはコンピュータとのインタラクションを可能にするデバイスであり，カーソルを移動しアイコンを選択することができるようになる．デスクトップパソコンにおけるポインティングデバイスの中で事実上の標準となっているのが 1960 年代にエンゲルバートによって発明されたマウスである．ユーザはマウスを平らな場所で移動させることで移動距離に比例してカーソルを移動することが可能になる．似ているポインティングデバイスとして，ボールを回転させることでカーソルを移動するトラックボールがある．マウスは何十年もの間利用されてきたが，ラップトップ，スマートフォンなど外出先で利用されるようなデバイスの普及が進んだことにより，平らな場所がないと利用できないマウスの代替としてタッチパネルといったポインティングデバイスが利用されるようになった．昨今ではスマートウォッチ，ヘッドマウントディスプレイなど新たなデバイスの普及が進んでおり，そのようなデバイス上で効率的に選択操作を行えるインタフェースが必要になっている．

モバイル環境や手元にマウスがない場合 (HMD 型 VR など) でのポインティング操作を容易にするアプローチとしてインタフェースを空間に拡張することや動きを阻害しないような動作を入力として用いることが挙げられる．例として，ハンドトラッキングを用いる手法 [4]，アイトラッキングを用いる手法 [54, 36, 11]，ヘッドトラッキングを用いる手法 [85, 64, 83]，脳波を用いる手法 [45]，靴を用いる手法 [30]，指先のトラッキングを用いた手法 [29, 16], シルエットを用いた手法 [70], 腕のトラッキングを用いた手法 [41, 87] などがある．

2.4 関連研究と比べた提案手法の利点

関連研究と比較し本研究は以下のような利点がある．

- 手首に装着したデバイスを用いるためウェアラブルでない動作推定手法に比べ, 携帯性に優れるとともに，ユーザが努力せずとも手を継続的に視野に収めることが可能である．

- 小型の単眼カメラを用いるため特別なセンサが不必要であり，既存のスマートウォッチの内蔵カメラを活用することも考えられるため実装コストが低い.
- 人々はアクセサリとして腕時計を普段から身につけているとともに，スマートウォッチも普及してきているため，手首にデバイスを装着することに対する抵抗感が少なく，身につけやすい.
- 既存手法で問題となっていたセマンティックセグメンテーションやオプティカルフローの計算コストの問題を蒸留を使うことで解決した.

第3章　提案手法

この研究は腕輪型デバイスに設置したカメラから得られる映像からマウス操作を可能にすることを目的としている．目的を達成するため手首にマウントし手の甲を撮影可能なカメラを 3D プリンタと小型 RGB カメラボードを用いて製作し，製作したデバイスから得られる映像からポインタの動き，クリック操作を予測するシステムを設計した．

3.1　ハードウェア

手首にマウントしたカメラから手の甲全体を撮影するため，小型の広角 RGB カメラを搭載したボード (FHD01M-L170-JP) を使用した (FOV170 度)．このカメラから得られる 640x480 の映像を 224x224 にクロップして使用している．カメラマウントを図 3.1 のように 3D プリンタを用いて製作し，ネジでカメラボードを固定した．このときカメラマウントは手の表面から 8mm 離れた位置にレンズを配置するとともに，12 度傾けることで手の根本まで撮影可能にしている．手首への固定はカメラマウントに縫い付けたベルクロテープを用いて行う．このデバイスを手首に装着した場合に得られる画像を図 3.2 に示す．

3.2　データセットの作成

3.2.1　データ収集

マウス動作推定ネットワークの学習のためのクリックジェスチャの分類ラベル，マウス動作の回帰ラベルを得るため 6 人の被験者からデータを収集した．データ収集では被験者の属性データ (性別，利き手) に加え，デバイスから得られる映像とジェスチャラベル，ポインタの座標を一つにまとめたデータを 1 フレームとして時系列順に 20fps で保存する．一人に付き約 10000 フレームのデータを保存し，被験者 6 名 (男性 5 名，女性 1 名) で合計約 60000 フレームのデータを得た．

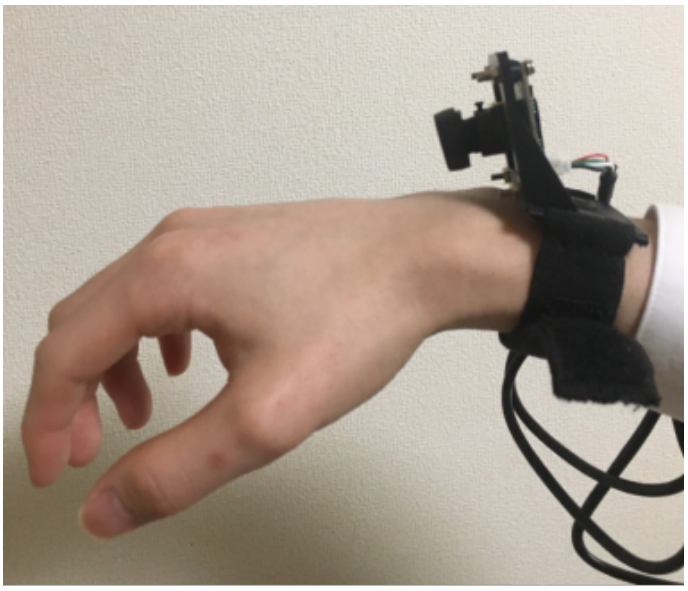

図 3.1: 実験に用いるデバイス

図 3.2: デバイスから得られる映像

3.2.2　データ形式

クリックジェスチャのラベル

データセットのクリック分類ラベルについて，観察よりクリックボタンを離してから約 2, 3, 4 フレーム後で手の甲に変化が起こることが見て取れた．そのため学習時にはクリックボタンを離したフレーム (図 3.4 では t0，図 3.5 では t3) から 2,3,4 フレーム目をクリックラベルとしている．右クリックラベルとオプティカルフローに関して対応の例を図 3.5 に，左クリックラベルとオプティカルフローに関して対応の例を図 3.4 に示す．ここで画像は 20fps のサンプリング間隔で撮影して

いる．OpticalFlow の視覚化には Baker らが提案した手法 [6] を使用している．視覚化に使用されている色相環を図 3.3 に示す．ラベルの番号はそれぞれ以下を意味している．

0. ジェスチャなし
1. 左クリック
2. 右クリック

画像から右クリックの場合は中指の腱に沿った顕著な変化が，左クリックの場合は中指の腱に沿った右クリックより淡い変化に加えて広範囲の淡い変化が起こっている．

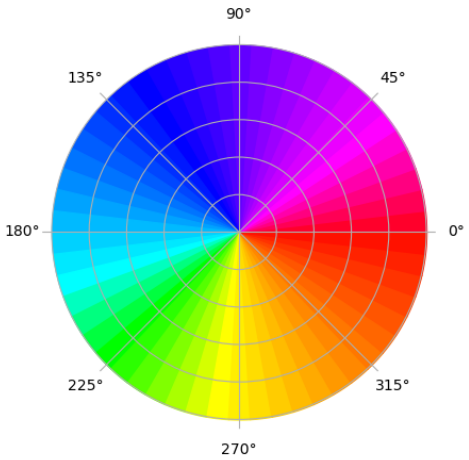

図 3.3: OpticalFlow の視覚化に使用した色相環

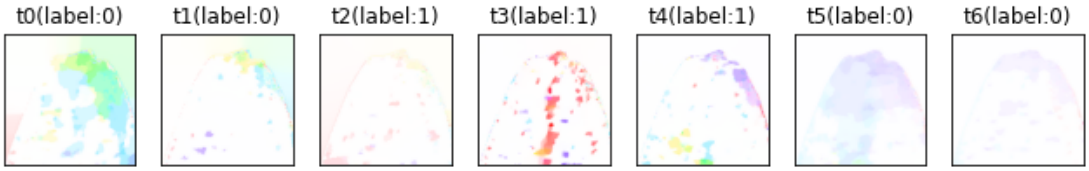

図 3.4: データセットのラベルとオプティカルフローの対応（左クリック）

ポインタ操作のラベル

画面上のポインタ座標を [x, y] の形で保存している．

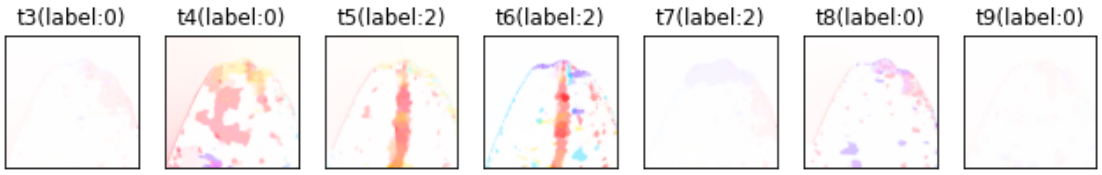

図 3.5: データセットのラベルとオプティカルフローの対応（右クリック）

3.2.3 収集方法

データ収集は図 3.6 の GUI を用いて行った．被験者はデバイスを装着した状態で画面上をランダムに動くボタンを順にクリックしていく．システム上ではデバイスが撮影した画像とラベルがフレームとしてまとめられ，時系列順に保存されるようになっている．被験者は左クリックでボタンをクリックしていく 1 分間のセクションを 5 回，同様に右クリックで行うセクションを 5 回行い，1 人につき合計で 10 分間のデータ収集を行った．ここで被験者は 1 セクションにつき 80 ～ 100 回程度のクリックを行った．そのため一人につき左クリック，右クリックそれぞれに関して 400 ～ 500 回程度のデータを収集できた．

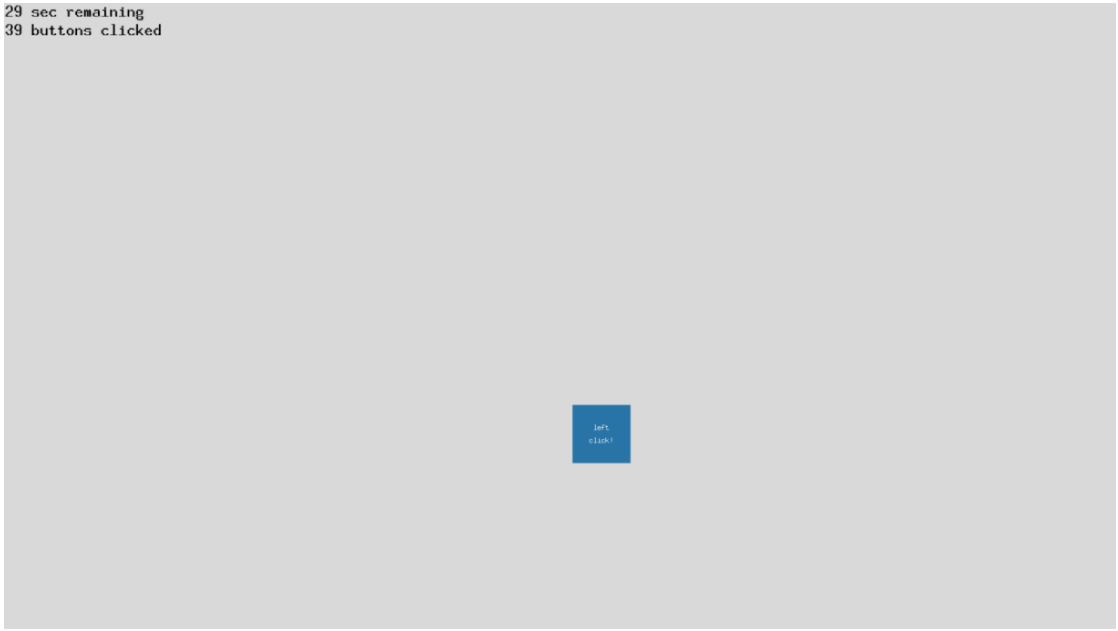

図 3.6: データ収集する際の GUI

3.3 ソフトウェアアーキテクチャ

提案するシステムはデバイスから得られるRGB映像を以下のように処理し出力を得る．モデルの概形を図3.7に示す．

1. データの前処理
    - ハンドセグメンテーション - セマンティックセグメンテーションによりRGB画像から手の領域のみを抽出する．
    - モーションのエンコード - ネットワークに入力するためのOptical Flowをハンドセグメンテーションをして得られた画像から生成する．
2. マウス動作推定ネットワーク - 学習時は生成したOpticalFlowを時系列順に2枚ずつ入力とし，クリックジェスチャの推定を行う．推論時はデバイスから得られるRGB映像を用いる．

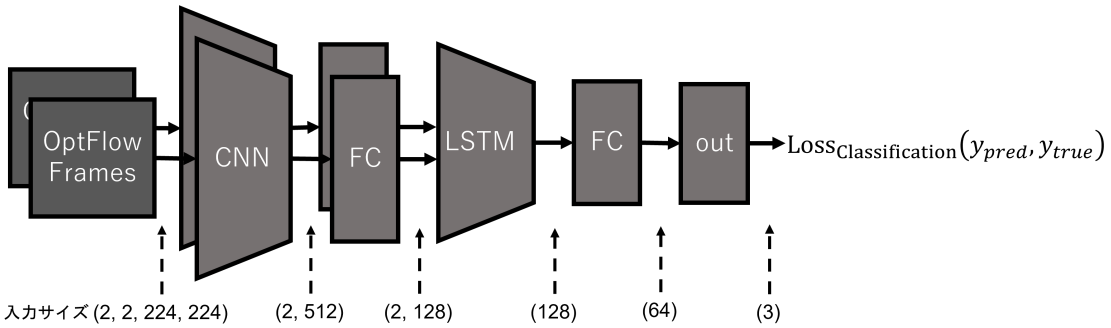

図 3.7: ネットワークのアーキテクチャ

3.3.1 データの前処理

ハンドセグメンテーション

画像から動作推定を行う際に背景によるノイズが入らないようハンドセグメンテーションを行う．セマンティックセグメンテーションを行うアルゴリズムの中でFastSCNN[61]を利用した．3.2節の方法でデータ収集した画像データから手動で設定した色の閾値により二値化するとともに輪郭収縮 (Contour Erosion) 処理を施すことでラベルを生成した．データ拡張として明るさや色相を変化させることで，最終的に100000の学習用データを得た．損失関数としてDiceLoss，学習率0.0001で設定し学習した結果，テスト用データセットでIoU:0.9912の結果を得た．デバイスから得た画像とその画像にセグメンテーションを適用した例を図3.8に示す．

図 3.8: セグメンテーションを適用する前と後の画像

モーションのエンコード

マウス動作推定ネットワークとして動画認識で用いられる 2-stream network[72] を参考に隣接フレームの動作がエンコードされた画像（eg. オプティカルフロー）を入力として使用した．動作をエンコードするアルゴリズムとして TV-L1 OpticalFlow[66](図 3.9) や Opisthenar[90] で用いられた modified MHI(Motion History Image:図 3.10) といった手法がある．OpticalFlow の視覚化には Baker ら [6] によって提案された色相環を利用している．ハンドセグメンテーション処理をした画像に OpticalFlow や MHI のアルゴリズムを適用して生成した画像を入力として CNNLSTM（タイムステップ 2）を学習した結果をホールドアウト検証で比較し，精度の良い TVl1 OpticalFlow を Motion Flow の入力として利用することにした．ここで画像のエンコーダとして Resnet18 を利用している．比較した結果を表 3.1 に示す．このとき，指標として F1 スコア（4 章で後述する式 4.2）を用いた．

表 3.1: モーションエンコーディングによる精度の違い

入力	F1 スコア
TVL1 OpticalFlow	0.8195
Motion History Image	0.7167

3.3.2 マウス動作推定ネットワーク:ClickNet

マウス動作推定ネットワークとして連続した画像からクリックジェスチャの推定を行うエンコーダデコーダネットワークを設計した．入力として 3.3.1 で述べた

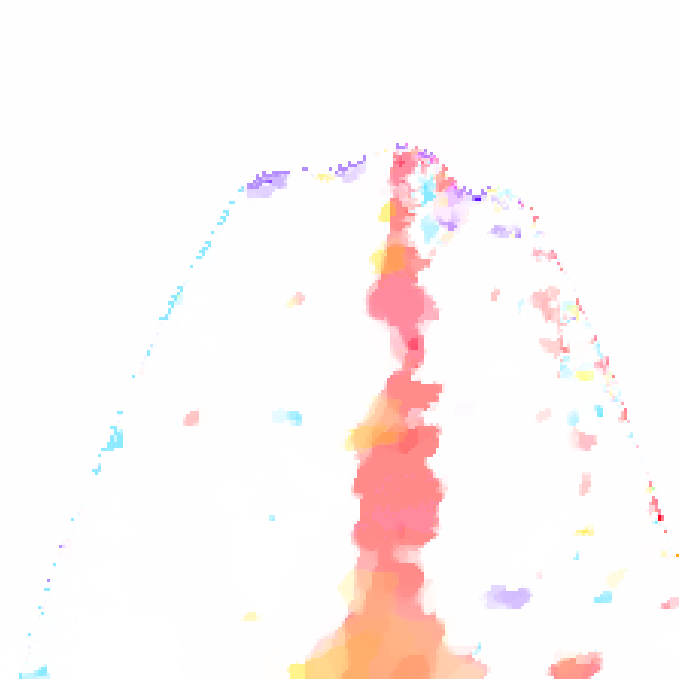

図 3.9: TVL1 オプティカルフロー [66]

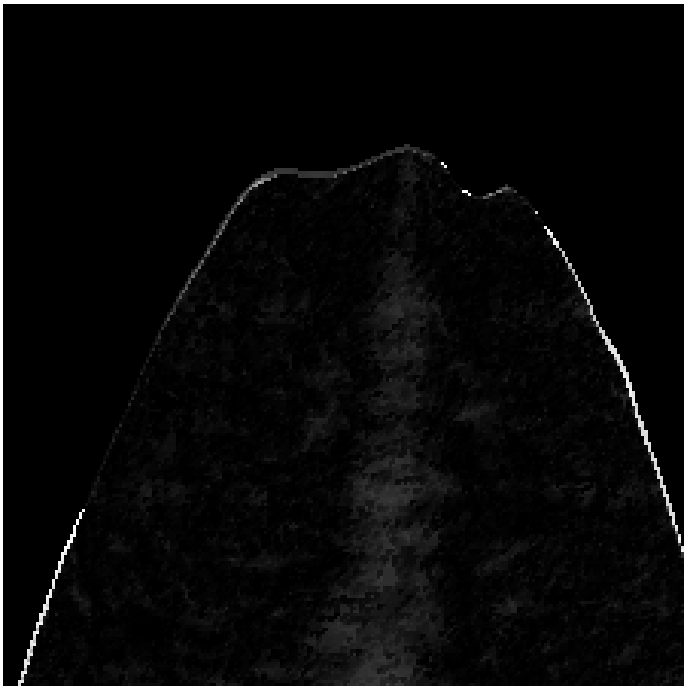

図 3.10: モーションヒストリーイメージ [90]

連続したモーションをエンコードした画像 2 枚を用いる．本研究では入力からクリックジェスチャ（左右のボタンを押す動作）のラベルを予測するネットワーク

を学習させた．このアーキテクチャの概形を 3.7 に示す．

まず，入力を OpticalFlow のみ，セグメンテーションを適用した RGB 画像のみ，両者の 2 ストリームにした場合の違いでホールドアウト検証を行い，入力を OpticalFlow のみ (MotionFlow のみ) にすることを選択した．検証した結果を表 3.2 に示す．入力が複数の場合は図 3.7 のモデルを並列に並べ，最後に FC 層で結果を結合したモデルを使用して学習している．結果から入力を 2 種類にした場合と比べて入力を OpticalFlow のみとした場合の精度が高くなっている．次に CNNLSTM の画像エンコーダとして，ResNet18[28] とモバイルデバイス向けに設計されている MobileNetV3[31] でホールドアウト検証を行い，MobileNetV3-Small に LSTM を組み合わせたアーキテクチャを選択した．比較した結果を表 3.3 に示す．ResNet18 は MobileNetV3 と比較してパラメータ数が多いため，このタスクでは過学習を起こしやすいと考えている．ここで指標として F1 スコア（4 章で後述する式 4.2）を用いた．3DCNN は計算コストの面でモバイルデバイス上でリアルタイムで推論を行うことが難しいため比較の対象にしていない．それと比較して CNNLSTM はモデルサイズが小さくすむため，モバイルデバイスに適していると考えた．

表 3.2: 入力を変えた場合の精度の違い

入力	F1 スコア
OpticalFlow のみ	0.8195
RGB 画像のみ	0.6779
OpticalFlow + RGB 画像	0.8143

表 3.3: エンコーダによる精度の違い

エンコーダ	F1 スコア
Resnet18	0.8195
MobileNetV3-Small	0.8542

3.3.3 MARS[14] による推論の高速化

ClickNet を用いてリアルタイムで推論する場合の課題として TVL1 OpticalFlow の推論コストの高さが挙げられる．MARS[14] は OpticalFlow で学習したネットワークが持つ情報を元の RGB 画像を入力とするネットワークに蒸留することで TwoStream ネットワークにおける MotionFlow の推論コストを下げる手法である．具体的な学習方法を図 3.11 に示す．この手法を用いることにより，セマンティックセグメンテーション (GPU で平均 0.00491 秒) や OpticalFlow(GPU で平均 0.08361 秒) の計算 が不要になり，一回の推論にかかる時間を平均 0.09373 秒から平均 0.00521 秒まで減らすことができ，約 18 倍の高速化ができた．

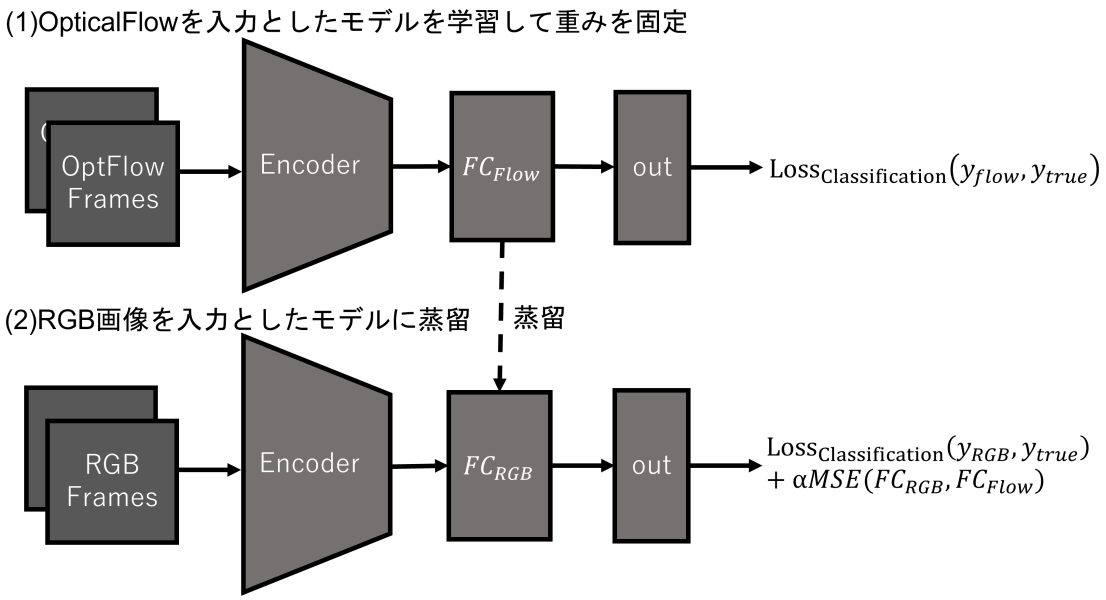

図 3.11: MARS[14] を用いた RGB ストリームへの蒸留方法

第4章　実験・評価

本章ではClickNetについて行った評価実験の方法と結果，そしてMARSを適用した結果について述べる．

4.1　ClickNetの評価

4.1.1　実行環境

モデルの訓練，評価はNVIDIA GTX1060，Intel Core i7 7700を搭載したコンピュータ上で行った．

4.1.2　ClickNetの実装

前章で提案したClickNetをPytorch[1]を用いて実装した．クラスの不均衡(図4.1)に対処するためクリック分類の損失関数としてClass Balanced Focal Loss[15]を用いた（β=0.999, γ=2.0）．Class Balanced Focal Lossを式4.1に示す．最適化アルゴリズムとしてはAdam[46]を用いている．学習率は初期値0.00001からスケジューラとしてPytorch組み込みのReduceLROnPlateau(patience = 2)のアルゴリズムを利用して学習率減衰させている．最後にデータ拡張としてデバイスの傾きを考慮するため，画像に回転（-10度, -5度，5度，10度）を加えたものをデータセットに加えている．データ拡張による精度の違いをホールドアウト検証した結果を表4.1に示す．

表 4.1: データ拡張による精度の違い

	F1スコア
データ拡張なし	0.8542
データ拡張あり	0.8839

$$CB_{focal}(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_i^t)^{\gamma} \log\left(p_i^t\right) \tag{4.1}$$

[1]https://pytorch.org/

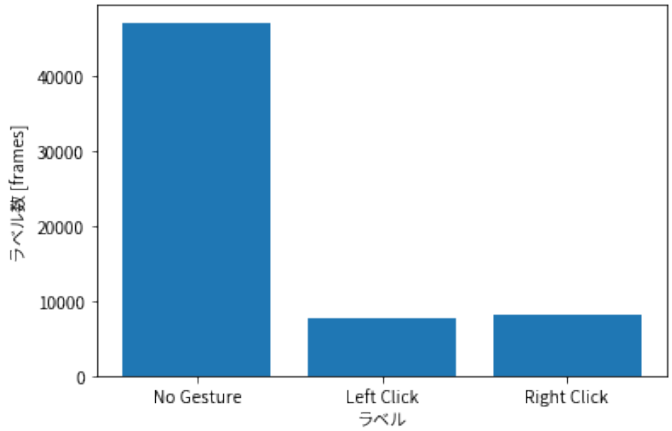

図 4.1: ラベルの分布

4.1.3 評価方法

以下の 2 種類の検証を行い，モデルの評価を行う．クリックジェスチャ分類の評価指標として F1 スコアを用いた．計算式を式 4.2 に示す．

- 全体のデータセットでのホールドアウト検証：全体のデータをを訓練用データとテスト用データに分けて学習させ，学習結果を確認することで全体のデータにおける精度を確認する．
- 個人データでのホールドアウト検証：一人のデータを訓練用データとテスト用データに分け，個人のデータにおける精度を確認するとともに，どのような要素が精度に影響を及ぼしているかを考察する．
- Leave-One-Out 検証：一人のデータを Validation 用のデータとして残し，他のデータを訓練データとして用いることで，未知のデータでも予測可能なモデルになっているかどうかを確認する．

$$
\begin{aligned}
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F1Score &= \frac{2 * Precision * Recall}{Precision + Recall}
\end{aligned} \tag{4.2}
$$

4.1.4 評価結果

それぞれの評価方法を用いて学習させた場合の平均精度を表 4.2 に示す．

表 4.2: それぞれの評価方法での平均精度

評価方法	F1 スコア
全データ	0.88390
個人データ	0.86788
Leave-One-Out	0.56054

全体データでの学習結果

全体のデータでホールドアウト検証を行った結果，F1 スコアは 0.88389 になった．予測結果の混合行列を図 4.2 に示す．

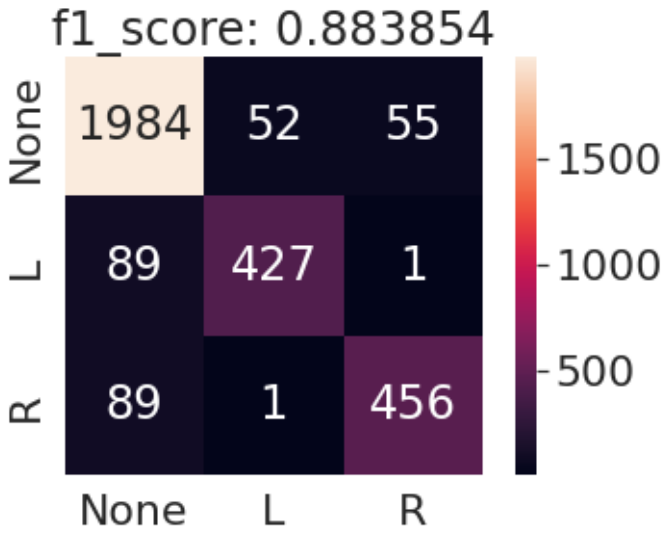

図 4.2: 全体のデータで学習を行った場合の混合行列

個人データでの学習結果

個人データでホールドアウト検証を行った結果，平均 F1 スコアは 0.86788 となった．それぞれのモデルの推論精度を表として表 4.3 に，予測結果を混合行列として図 4.3 に示す．

Leave-One-Out 検証の結果

Leave-One-Out 検証を行った結果，平均 F1 スコアは 0.56054 となった．それぞれのモデルの推論精度を表として表 4.4 に，予測結果を混合行列として図 4.4 に示す．この結果は，全体データの場合と個人データの場合の結果と比べて低い精度となっており，過学習が原因として考えられる．

表 4.3: 被験者による精度の違い

被験者	F1 スコア
female1	0.91957
male1	0.86524
male2	0.87712
male3	0.87663
male4	0.82272
male5	0.84602

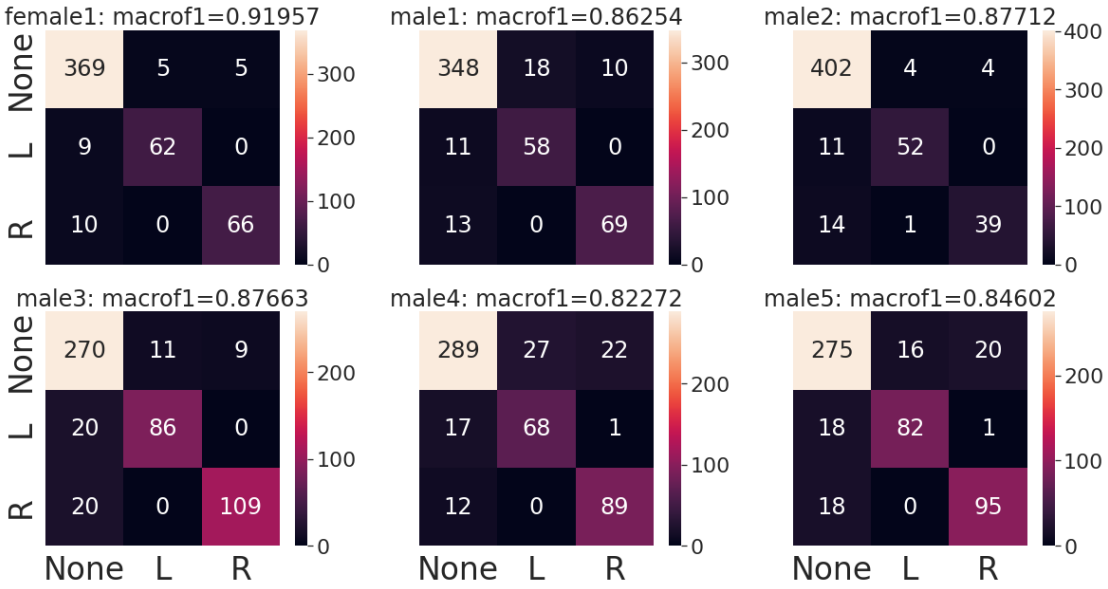

図 4.3: 被験者それぞれのデータで学習を行った場合の混合行列

4.1.5 考察

ClickNet に対して 3 種類の評価を行った結果，以下のような学びを得た．

データ拡張

データ拡張を行ったことにより F1 スコアで 0.03 もの差が出た．ウェアラブルデバイスを用いたシステムにおいて，使用中のデバイスのずれや傾きの変化，ユーザによるデバイスの付け方などによる画像の変化が大きく，データ拡張によりそのような変化に対応しやすくなったと考えられる．

表 4.4: Leave-One-Out 検証の結果

被験者 (テストデータ)	F1 スコア
female1	0.71990
male1	0.71620
male2	0.46220
male3	0.49120
male4	0.52110
male5	0.62920

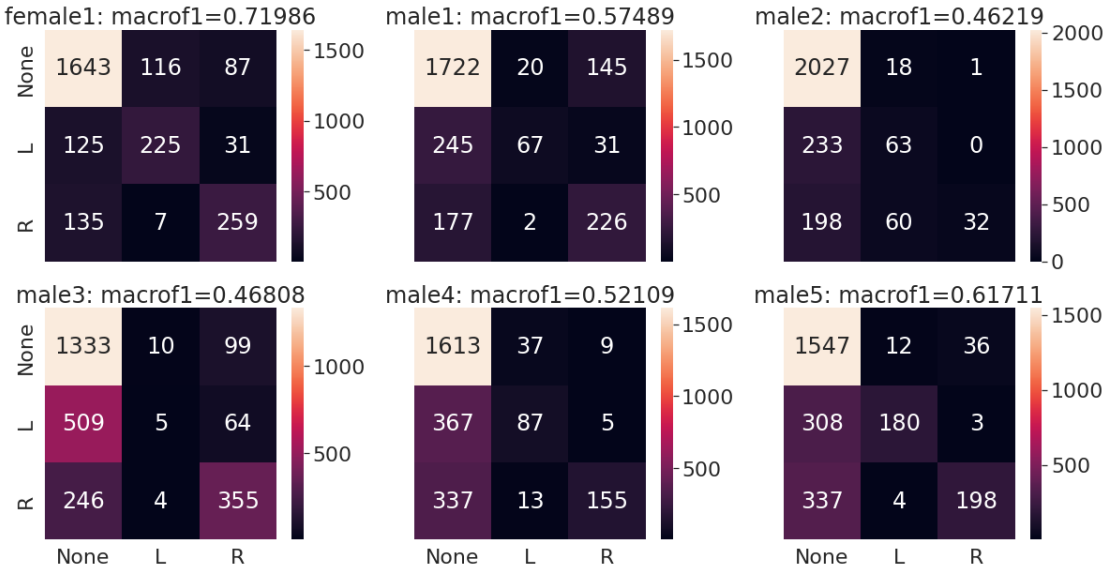

図 4.4: Leave-One-Out で学習させた場合の混合行列

Noisy Label での学習

精度は全体のデータを用いた場合で 0.88 となった．混合行列より左クリックラベルと右クリックラベル間の識別は高精度にできているが，ジェスチャなしラベルとクリックラベル間の識別が上手くできていないことが見て取れる．この理由としてラベルが Noisy であることが挙げられる．クリック後 2,3,4 フレーム後で決め打ちしてクリックラベルを付けているため，図 3.4，図 3.5 のようにジェスチャなしラベルでクリックと同じ変化，ジェスチャありラベルで変化が起きていない場合があり，それを理由として誤った予測が出ている可能性がある．

精度が変わる要素

精度が良くなる要素として，ユーザの手がやせ型かどうかがあると考えられる．最も精度の高い female1(図 4.5) のようにやせ型の場合，腱の動きが顕著に観測可

能になるため，OpticalFlow で観測しやすくなる．精度が低くなる場合として焦点距離よりも手の位置が近くぼやけてしまう場合が挙げられる．例を図 4.6 に示す．この例では，手の位置がカメラに近くぼやけてしまっているため OpticalFlow にも全体的な淡い変化が起きてしまっている．

図 4.5: 精度が高い例 (female1))

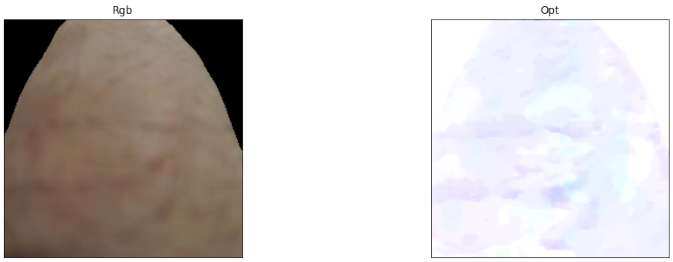

図 4.6: 精度が低い例 (male4))

未知のデータの予測

Leave-One-Out 検証では全体データと個人データで検証した場合と比べ平均 F1 スコア 0.56 と大幅に低い値となった．この理由としてデータセットの網羅性の低さが挙げられる．6 人から取ったデータセットで 5 人を訓練データとして 1 人をテストデータとして学習しているため，手の形やクセなどの知識が十分に得られていなかったと考えられる．

4.2 MARS によって軽量化したモデルの評価

4.2.1 MARS の実装

3.3.3 項で説明した MARS を pytorch を用いて実装，学習させた．ここで MARS を学習させる際の損失関数は MARS の論文を参考に式 4.3($\alpha = 0.01$) を用いている．MSE を式 4.4 に示す．式 4.3 中の α は FC 層の出力から計算する MSE と分類

問題の損失関数のオーダーを近づけるためのハイパーパラメータであり，分類の損失関数として用いたClassBalancedLossのオーダーがMSEより小さいため，0.0001とおいている．

$$\begin{aligned} Loss_{MARS} &= Loss_{classification} + \alpha Loss_{fc} \\ &= CB_{focal}(\mathbf{p}, y) + \alpha MSE(fc_{student}, fc_{teacher}) \end{aligned} \tag{4.3}$$

$$MSE(\hat{y}, y) = \frac{1}{n}\sum_{i+1}^{n}(\hat{y}_i - y_i)^2 \tag{4.4}$$

4.2.2 評価方法

全てのデータを訓練データとテストデータに分けてホールドアウト検証を行った．評価指標としてF1スコアを用いた．(式4.2).

4.2.3 評価結果

システムの精度，推論速度について，ClickNet，ClickNet(MARSを適用)，元のRGB画像を入力としたモデルの3つで比較を行った結果を表4.5に示す．MARSで学習したモデルでジェスチャ分類を行った結果の混合行列を図4.7に示す．MARSの評価より3DCNNだけでなく，CNNLSTMにおいてもMARSは有効な蒸留方法であり，さらにマスク処理を施した画像から生成したOpticalFlowを用いていても精度をあまり犠牲にすることなく蒸留可能であることが分かる．さらにMARSの論文ではグリッドサーチによりαの値を決めていたため，αの値の調整をすればもう少し精度が上がる可能性がある．

表 4.5: ClickNetにMARSを適用した場合としていない場合の比較

モデル	入力画像	推論速度 (秒)	F1スコア
元画像 (RGB) を入力としたモデル	RGB	0.00555	0.82980
ClickNet(OpticalFLow)	OpticalFlow	0.09373	0.88390
ClickNet(MARS)	RGB	0.00555	0.87307

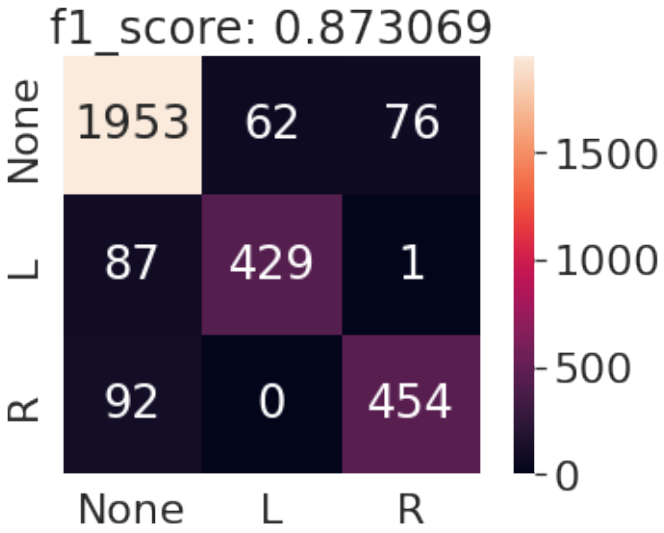

図 4.7: ClickNet に MARS を適用した場合の混合行列

第5章　おわりに

5.1　今後の展望

ラベル付けの精度

PC 側でクリックが検知されたフレームから 2,3,4 フレーム目を決め打ちしてクリックラベルにしているため，手の甲に変化が起こるタイミングが少しずれていたりすると，変化が起こっていないにもかかわらずクリックラベルになっているフレーム，変化が起こっているのにもかかわらずクリックラベルになっていないフレームが存在し得る．それが理由として，誤検知が増加してしまっている可能背がある．このような Noisy Label は機械学習では大きな課題となっており，広く研究が進められているため Noisy Label に対応できるような学習方法を導入するのは一つの解決策だと考えている．他にも，手間がかかるが手作業でラベル付けを行うことでも解決できる可能性がある．

個人差 (手の形，クセなど)

個人データでの検証により，人により認識精度にムラがあることが分かった．手の形，クセなどいろいろなファクターが考えられるが，一番大きな要因になっていそうなのは，手の甲にある腱が観測可能であるかどうかである．やせ型の場合，腱がカメラから顕著に観測可能であり，精度が高くなる傾向にある．FlowNet[22] や PWC-Net[77] などより高精度に動作をエンコード可能なシステムの利用などが方法としてあると考えている．

未知のデータの予測

LeaveOneOut 検証では，過学習が起きたためテストデータで大幅に低い精度になった．学習データにテストデータに適用し得る情報が含まれていなかったといえる．6 人のみのデータを用いた検証であったため，データセットにおいて手の形やクセなどの網羅性が低く，テストデータに対応しきれなかったと考えられる．混合行列をみると全く学習していないわけではなく，ある程度学習しようとしている傾向もあるためデータセットを増やしていけば対応できそうであると考えている．

マウスのポインタ操作

本稿ではクリックジェスチャのみに焦点を当て，ポインタ操作について検証は行わなかった．しかし，マウス操作においてクリックジェスチャに加えポインタ操作は重要であり，マウスを実現するためにはこれを可能にする必要がある．方法としてはスマートウォッチに組み込まれることの多い IMU(慣性センサ) といったセンサを利用したマルチモダルネットワークなどが考えられる．

5.2 まとめ

本稿では手首に身に付けたデバイスから得られる画像を用いてクリックジェスチャを予測する手法を提案した．

データ収集とラベル付けを行い，設計したモデルを学習させた．モデルに対して 3 種類の評価を行い，全体のデータでは F1 スコア 0.88390，個人データでは F1 スコア 0.86788 を達成した．しかし，データセットの網羅性や Noisy ラベルにより，精度が出ない場合もあることが分かった．

さらに，OpticalFlow の計算コストという問題もあり，この点に関して MARS のテクニックを用いた蒸留を行った．学習させ評価を行った結果，精度の減少があったが推論速度を 0.08818 秒削減でき，1 フレーム毎の推論速度を 0.00555 秒まで減らすことができた．

謝辞

本研究を進めるにあたり，暖かく見守ってくださった宮田一乗教授と謝浩然講師，アドバイスをくださった研究室の皆様方，データ収集に協力してくださった方々に深く感謝を申し上げます.

参考文献

[1] Brian Amento, Will Hill, and Loren Terveen. “The Sound of One Hand: A Wrist-Mounted Bio-Acoustic Fingertip Gesture Interface”. In: *CHI ’02 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’02. New York, NY, USA: Association for Computing Machinery, 2002, pp. 724–725. ISBN: 1581134541. DOI: 10.1145/506443.506566. URL: https://doi.org/10.1145/506443.506566.

[2] Kazuyuki Arimatsu and Hideki Mori. “Evaluation of Machine Learning Techniques for Hand Pose Estimation on Handheld Device with Proximity Sensor”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376712.

[3] Anurag Arnab et al. “ViViT: A Video Vision Transformer”. In: (Mar. 2021). URL: https://arxiv.org/abs/2103.15691v2.

[4] Daniel Bachmann, Frank Weichert, and Gerhard Rinkenauer. “Evaluation of the Leap Motion Controller as a New Contact-Free Pointing Device”. In: *Sensors* 15.1 (2015), pp. 214–233. ISSN: 1424-8220. DOI: 10.3390/s150100214. URL: https://www.mdpi.com/1424-8220/15/1/214.

[5] Gilles Bailly et al. “ShoeSense: A New Perspective on Gestural Interaction and Wearable Applications”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 1239–1248. ISBN: 9781450310154. DOI: 10.1145/2207676.2208576. URL: https://doi.org/10.1145/2207676.2208576.

[6] Simon Baker et al. “A database and evaluation methodology for optical flow”. In: *Proceedings of the IEEE International Conference on Computer Vision* (2007). DOI: 10.1109/ICCV.2007.4408903.

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. “Is Space-Time Attention All You Need for Video Understanding?” In: (Feb. 2021). URL: https://arxiv.org/abs/2102.05095v4.

[8] Aaron F Bobick and James W Davis. “The Recognition of Human Movement Using Temporal Templates”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 23.3 (2001), pp. 257–267. ISSN: 0162-8828. DOI: 10.1109/34.910878. URL: https://doi.org/10.1109/34.910878.

[9] Liwei Chan et al. “CyclopsRing: Enabling Whole-Hand and Context-Aware Interactions Through a Fisheye Ring”. In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. UIST ’15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 549–556. ISBN: 9781450337793. DOI: 10.1145/2807442.2807450. URL: https://doi.org/10.1145/2807442.2807450.

[10] Feiyu Chen et al. “Finger Angle-Based Hand Gesture Recognition for Smart Infrastructure Using Wearable Wrist-Worn Camera”. In: *Applied Sciences* 8.3 (2018). ISSN: 2076-3417. DOI: 10.3390/app8030369. URL: https://www.mdpi.com/2076-3417/8/3/369.

[11] Xiuli Chen et al. “An adaptive model of gaze-based selection”. In: *Conference on Human Factors in Computing Systems - Proceedings* (May 2021). DOI: 10.1145/3411764.3445177.

[12] Simone Ciotti et al. “A Synergy-Based Optimally Designed Sensing Glove for Functional Grasp Recognition”. In: *Sensors* 16 (2016), p. 811. DOI: 10.3390/s16060811.

[13] J Connolly et al. “IMU Sensor-Based Electronic Goniometric Glove for Clinical Finger Movement Analysis”. In: *IEEE Sensors Journal* 18.3 (2018), pp. 1273–1281. ISSN: 1558-1748. DOI: 10.1109/JSEN.2017.2776262.

[14] Nieves Crasto et al. “MARS: Motion-Augmented RGB Stream for Action Recognition”. In: (2019), pp. 7874–7883. DOI: 10.1109/CVPR.2019.00807{\"{i}}. URL: http://www.europe.naverlabs.com/Research/.

[15] Yin Cui et al. “Class-Balanced Loss Based on Effective Number of Samples”. In: ().

[16] Shome S Das. “Real Time Direction Estimation for Pointing Interactions Using a Depth Sensor and a Nine Axis Inertial Motion Unit”. In: *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’20. New York, NY, USA: Association for Computing Machinery, 2020. ISBN: 9781450377737. DOI: 10.1145/3389189.3392618. URL: https://doi.org/10.1145/3389189.3392618.

[17] Artem Dementyev and Joseph A Paradiso. "WristFlex: Low-Power Gesture Input with Wrist-Worn Pressure Sensors". In: *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. UIST '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 161–166. ISBN: 9781450330695. DOI: 10.1145/2642918.2647396. URL: https://doi.org/10.1145/2642918.2647396.

[18] Travis Deyle et al. "Hambone: A Bio-Acoustic Gesture Interface". In: Nov. 2007, pp. 3–10. ISBN: 978-1-4244-1452-9. DOI: 10.1109/ISWC.2007.4373768.

[19] L Dipietro, A M Sabatini, and P Dario. "A Survey of Glove-Based Systems and Their Applications". In: *Trans. Sys. Man Cyber Part C* 38.4 (2008), pp. 461–482. ISSN: 1094-6977. DOI: 10.1109/TSMCC.2008.923862. URL: https://doi.org/10.1109/TSMCC.2008.923862.

[20] Jeff Donahue et al. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (2017), pp. 677–691. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2599174. URL: https://doi.org/10.1109/TPAMI.2016.2599174.

[21] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (Oct. 2020). URL: https://arxiv.org/abs/2010.11929v2.

[22] Philipp Fischer et al. "FlowNet: Learning Optical Flow with Convolutional Networks". In: (Apr. 2015). URL: http://arxiv.org/abs/1504.06852.

[23] Rui Fukui et al. "Hand Shape Classification with a Wrist Contour Sensor: Development of a Prototype Device". In: *Proceedings of the 13th International Conference on Ubiquitous Computing*. UbiComp '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 311–314. ISBN: 9781450306300. DOI: 10.1145/2030112.2030154. URL: https://doi.org/10.1145/2030112.2030154.

[24] Harshala Gammulle et al. *Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition*. 2017.

[25] Liuhao Ge et al. "3D Hand Shape and Pose Estimation from a Single {RGB} Image". In: *CoRR* abs/1903.0 (2019). URL: http://arxiv.org/abs/1903.00812.

[26] Oliver Glauser et al. "Interactive Hand Pose Estimation Using a Stretch-Sensing Soft Glove". In: *ACM Trans. Graph.* 38.4 (2019). ISSN: 0730-0301. DOI: 10.1145/3306346.3322957. URL: https://doi.org/10.1145/3306346.3322957.

[27] Shangchen Han et al. “Online Optical Marker-Based Hand Tracking with Deep Labels”. In: *ACM Trans. Graph.* 37.4 (2018). ISSN: 0730-0301. DOI: 10.1145/3197517.3201399. URL: https://doi.org/10.1145/3197517.3201399.

[28] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.0 (2015). URL: http://arxiv.org/abs/1512.03385.

[29] K. Horiuchi and T. Matsumaru. “Short Range Fingertip Pointing Operation Interface by Depth Camera”. In: *2018 IEEE International Conference on Robotics and Biomimetics, ROBIO 2018*. Institute of Electrical and Electronics Engineers Inc., July 2018, pp. 132–137. ISBN: 9781728103761. DOI: 10.1109/ROBIO.2018.8665254.

[30] Daniel Horodniczy and Jeremy R. Cooperstock. “Free the hands! Enhanced target selection via a variable-friction shoe”. In: *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2017-May. Association for Computing Machinery, May 2017, pp. 255–259. ISBN: 9781450346559. DOI: 10.1145/3025453.3025625.

[31] Andrew Howard et al. “Searching for MobileNetV3”. In: *Proceedings of the IEEE International Conference on Computer Vision* 2019-October (May 2019), pp. 1314–1324. ISSN: 15505499. DOI: 10.1109/ICCV.2019.00140. URL: https://arxiv.org/abs/1905.02244v5.

[32] Fang Hu et al. “FingerTrak: Continuous 3D Hand Pose Tracking by Deep Learning Hand Silhouettes Captured by Miniature Thermal Cameras on-Wrist”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. Vol. 4. 2. Association for Computing Machinery, June 2020, pp. 1–24. DOI: 10.1145/3397306. URL: https://dl.acm.org/doi/abs/10.1145/3397306.

[33] Dong-Hyun Hwang et al. “MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera”. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 98–111. ISBN: 9781450375146. DOI: 10.1145/3379337.3415856. URL: https://doi.org/10.1145/3379337.3415856.

[34] Antonis Argyros Iason Oikonomidis Nikolaos Kyriazis. “Efficient model-based 3D tracking of hand articulations using Kinect”. In:

[35] Umar Iqbal et al. “Hand Pose Estimation via Latent 2.5D Heatmap Regression”. In: *CoRR* abs/1804.0 (2018). URL: http://arxiv.org/abs/1804.09534.

[36] Robert J.K. Jacob. "What you look at is what you get: Eye movement-based interaction techniques". In: *Conference on Human Factors in Computing Systems - Proceedings* (Mar. 1990), pp. 11–18. DOI: 10.1145/97243.97246.

[37] Sujin Jang et al. "Modeling Cumulative Arm Fatigue in Mid-Air Interaction Based on Perceived Exertion and Kinetics of Arm Motion". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 3328–3339. ISBN: 9781450346559. URL: https://doi.org/10.1145/3025453.3025523.

[38] Y Jang et al. "3D Finger CAPE: Clicking Action and Position Estimation under Self-Occlusions in Egocentric Viewpoint". In: *IEEE Transactions on Visualization and Computer Graphics* 21.4 (2015), pp. 501–510. DOI: 10.1109/TVCG.2015.2391860.

[39] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.1 (2013), pp. 221–231. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.59. URL: https://doi.org/10.1109/TPAMI.2012.59.

[40] S Jiang et al. "Stretchable e-Skin Patch for Gesture Recognition on the Back of the Hand". In: *IEEE Transactions on Industrial Electronics* 67.1 (2020), pp. 647–657. DOI: 10.1109/TIE.2019.2914621.

[41] Keiko Katsuragawa et al. "Watchpoint: Freehand pointing with a smartwatch in a ubiquitous display environment". In: *Proceedings of the Workshop on Advanced Visual Interfaces AVI*. Vol. 07-10-June. Association for Computing Machinery, June 2016, pp. 128–135. ISBN: 9781450341318. DOI: 10.1145/2909132.2909263.

[42] Bryce Kellogg, Vamsi Talla, and Shyamnath Gollakota. "Bringing Gesture Recognition to All Devices". In: *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*. NSDI'14. USA: USENIX Association, 2014, pp. 303–316. ISBN: 9781931971096.

[43] David Kim et al. "Digits: Freehand 3D Interactions Anywhere Using a Wrist-Worn Gloveless Sensor". In: *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*. UIST '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 167–176. ISBN: 9781450315807. DOI: 10.1145/2380116.2380139. URL: https://doi.org/10.1145/2380116.2380139.

[44] Jungsoo Kim et al. “The Gesture Watch: A Wireless Contact-Free Gesture Based Wrist Interface”. In: *Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers*. ISWC ’07. USA: IEEE Computer Society, 2007, pp. 1–8. ISBN: 9781424414529. DOI: 10.1109/ISWC.2007.4373770. URL: https://doi.org/10.1109/ISWC.2007.4373770.

[45] S Kim et al. “Point-and-Click Cursor Control With an Intracortical Neural Interface System by Humans With Tetraplegia”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19.2 (2011), pp. 193–203. DOI: 10.1109/TNSRE.2011.2107750.

[46] Diederik P. Kingma and Jimmy Lei Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Dec. 2014). URL: https://arxiv.org/abs/1412.6980v9.

[47] Wakaba Kuno, Maki Sugimoto, and Yuta Sugiura. “Finger posture estimation by measuring skin deformation on back of hand”. In: *Kyokai Joho Imeji Zasshi/Journal of the Institute of Image Information and Television Engineers* 73.3 (2019), pp. 595–601. ISSN: 1342-6907. DOI: 10.3169/itej.73.595.

[48] Martin de La Gorce, David J Fleet, and Nikos Paragios. “Model-Based 3D Hand Pose Estimation from Monocular Video”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.9 (2011), pp. 1793–1805. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.33. URL: https://doi.org/10.1109/TPAMI.2011.33.

[49] Gierad Laput, Robert Xiao, and Chris Harrison. “ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers”. In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 321–333. ISBN: 9781450341899. DOI: 10.1145/2984511.2984582. URL: https://doi.org/10.1145/2984511.2984582.

[50] Tianxing Li et al. “Reconstructing Hand Poses Using Visible Light”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.3 (2017). DOI: 10.1145/3130937. URL: https://doi.org/10.1145/3130937.

[51] X Liang, H Heidari, and R Dahiya. “Wearable Capacitive-Based Wrist-Worn Gesture Sensing System”. In: *2017 New Generation of CAS (NGCAS)*. 2017, pp. 181–184. DOI: 10.1109/NGCAS.2017.80.

[52] Jhe-Wei Lin et al. "BackHand: Sensing Hand Gestures via Back of the Hand". In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. UIST '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 557–564. ISBN: 9781450337793. DOI: 10.1145/2807442.2807462. URL: https://doi.org/10.1145/2807442.2807462.

[53] Z Lu et al. "A Hand Gesture Recognition Framework and Wearable Gesture-Based Interaction Prototype for Mobile Devices". In: *IEEE Transactions on Human-Machine Systems* 44.2 (2014), pp. 293–299. DOI: 10.1109/THMS.2014.2302794.

[54] Christof Lutteroth, Moiz Penkar, and Gerald Weber. "Gaze vs. Mouse: A fast and accurate gaze-only click alternative". In: *UIST 2015 - Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Nov. 2015), pp. 385–394. DOI: 10.1145/2807442.2807461.

[55] Vadim Lyubanenko et al. "Multi-camera finger tracking and 3D trajectory reconstruction for HCI studies". In: *Advanced Concepts for Intelligent Vision Systems : 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings*. Ed. by Jacques Blanc-Talon et al. Lecture Notes in Computer Science 10617. International: Springer, 2017, pp. 63–74. ISBN: 978-3-319-70352-7. DOI: 10.1007/978-3-319-70353-4{_}6.

[56] Jess McIntosh, Asier Marzo, and Mike Fraser. "SensIR: Detecting Hand Gestures with a Wearable Bracelet Using Infrared Transmission and Reflection". In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. UIST '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 593–597. ISBN: 9781450349819. DOI: 10.1145/3126594.3126604. URL: https://doi.org/10.1145/3126594.3126604.

[57] Jess McIntosh et al. "EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1923–1934. ISBN: 9781450346559. URL: https://doi.org/10.1145/3025453.3025807.

[58] Jess McIntosh et al. "EMPress: Practical Hand Gesture Classification with Wrist-Mounted EMG and Pressure Sensing". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 2332–2342. ISBN: 9781450333627. URL: https://doi.org/10.1145/2858036.2858093.

[59] Daniel Neimark et al. “Video Transformer Network”. In: (Feb. 2021), pp. 3156–3165. DOI: 10.1109/iccvw54120.2021.00355. URL: https://arxiv.org/abs/2102.00719v3.

[60] Corey R Pittman and Joseph J LaViola. “Multiwave: Complex Hand Gesture Recognition Using the Doppler Effect”. In: *Proceedings of the 43rd Graphics Interface Conference*. GI ’17. Waterloo, CAN: Canadian Human-Computer Communications Society, 2017, pp. 97–106. ISBN: 9780994786821.

[61] Rudra P.K. Poudel, Stephan Liwicki, and Roberto Cipolla. “Fast-SCNN: Fast Semantic Segmentation Network”. In: *30th British Machine Vision Conference 2019, BMVC 2019* (Feb. 2019). URL: https://arxiv.org/abs/1902.04502v1.

[62] Chen Qian et al. “Realtime and Robust Hand Tracking from Depth”. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR ’14. USA: IEEE Computer Society, 2014, pp. 1106–1113. ISBN: 9781479951185. DOI: 10.1109/CVPR.2014.145. URL: https://doi.org/10.1109/CVPR.2014.145.

[63] Jun Rekimoto. “GestureWrist and GesturePad: Unobtrusive Wearable Interaction Devices”. In: *Proceedings of the 5th IEEE International Symposium on Wearable Computers*. ISWC ’01. USA: IEEE Computer Society, 2001, p. 21. ISBN: 0769513182.

[64] Andreia Sias Rodrigues et al. “Evaluation of a Head-Tracking Pointing Device for Users with Motor Disabilities”. In: *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 156–162. ISBN: 9781450352277. DOI: 10.1145/3056540.3056552. URL: https://doi.org/10.1145/3056540.3056552.

[65] Grégory Rogez et al. “3D Hand Pose Detection in Egocentric {RGB-D} Images”. In: *CoRR* abs/1412.0 (2014). URL: http://arxiv.org/abs/1412.0065.

[66] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. “TV-L1 Optical Flow Estimation”. In: *Image Processing On Line* 3 (July 2013), pp. 137–150. ISSN: 2105-1232. DOI: 10.5201/IPOL.2013.26.

[67] T Scott Saponas et al. “Enabling Always-Available Input with Muscle-Computer Interfaces”. In: *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*. UIST ’09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 167–176. ISBN: 9781605587455.

DOI: 10.1145/1622176.1622208. URL: https://doi.org/10.1145/1622176.1622208.

[68] Kyeongeun Seo and Hyeonjoong Cho. "AirPincher: A Handheld Device for Recognizing Delicate Mid-Air Hand Gestures". In: *Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology*. UIST'14 Adjunct. New York, NY, USA: Association for Computing Machinery, 2014, pp. 83–84. ISBN: 9781450330688. DOI: 10.1145/2658779.2658787. URL: https://doi.org/10.1145/2658779.2658787.

[69] Toby Sharp et al. "Accurate, Robust, and Flexible Real-Time Hand Tracking". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2015, pp. 3633–3642. ISBN: 9781450331456. URL: https://doi.org/10.1145/2702123.2702179.

[70] Jun Shingu et al. "Depth based shadow pointing interface for public displays". In: *UIST 2016 Adjunct - Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. Association for Computing Machinery, Inc, Oct. 2016, pp. 79–80. ISBN: 9781450345316. DOI: 10.1145/2984751.2985710.

[71] S Sikdar et al. "Novel Method for Predicting Dexterous Individual Finger Movements by Imaging Muscle Activity Using a Wearable Ultrasonic System". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.1 (2014), pp. 69–76. DOI: 10.1109/TNSRE.2013.2274657.

[72] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576.

[73] Mohamed Soliman et al. "FingerInput: Capturing Expressive Single-Hand Thumb-to-Finger Microgestures". In: *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*. ISS '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 177–187. ISBN: 9781450356947. DOI: 10.1145/3279778.3279799. URL: https://doi.org/10.1145/3279778.3279799.

[74] Jie Song et al. "In-Air Gestures around Unmodified Mobile Devices". In: *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. UIST '14. New York, NY, USA: Association for Computing Ma-

chinery, 2014, pp. 319–329. ISBN: 9781450330695. DOI: 10.1145/2642918.2647373. URL: https://doi.org/10.1145/2642918.2647373.

[75] Adrian Spurr et al. "Cross-modal Deep Variational Hand Pose Estimation". In: *CoRR* abs/1803.1 (2018). URL: http://arxiv.org/abs/1803.11404.

[76] David J Sturman and David Zeltzer. "A Survey of Glove-Based Input". In: *IEEE Comput. Graph. Appl.* 14.1 (1994), pp. 30–39. ISSN: 0272-1716. DOI: 10.1109/38.250916. URL: https://doi.org/10.1109/38.250916.

[77] Deqing Sun et al. "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume". In: *CoRR* abs/1709.0 (2017). URL: http://arxiv.org/abs/1709.02371.

[78] Li Sun et al. "WiDraw: Enabling Hands-Free Drawing in the Air on Commodity WiFi Devices". In: *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. MobiCom '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 77–89. ISBN: 9781450336192. DOI: 10.1145/2789168.2790129. URL: https://doi.org/10.1145/2789168.2790129.

[79] Zengshan Tian et al. "WiCatch: A Wi-Fi Based Hand Gesture Recognition System". In: *IEEE Access* PP (2018), p. 1. DOI: 10.1109/ACCESS.2018.2814575.

[80] Ilya Tolstikhin et al. "MLP-Mixer: An all-MLP Architecture for Vision". In: (May 2021). URL: https://arxiv.org/abs/2105.01601v4.

[81] Jonathan Tompson et al. "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks". In: *ACM Trans. Graph.* 33.5 (2014). ISSN: 0730-0301. DOI: 10.1145/2629500. URL: https://doi.org/10.1145/2629500.

[82] Hoang Truong et al. "CapBand: Battery-Free Successive Capacitance Sensing Wristband for Hand Gesture Recognition". In: *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. SenSys '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 54–67. ISBN: 9781450359528. DOI: 10.1145/3274783.3274854. URL: https://doi.org/10.1145/3274783.3274854.

[83] S Uchino et al. "Development of a Pointing Device that Directly Measures the Tilt Angles of a Head". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019, pp. 2851–2856. DOI: 10.1109/SMC.2019.8913882.

[84] Andrew Vardy, John Robinson, and Li-Te Cheng. "The WristCam as Input Device". In: *Proceedings of the 3rd IEEE International Symposium on Wearable Computers*. ISWC '99. USA: IEEE Computer Society, 1999, p. 199. ISBN: 0769504280.

[85] Edwin Walsh, Walter Daems, and Jan Steckel. "An optical head-pose tracking sensor for pointing devices using IR-LED based markers and a low-cost camera". In: *2015 IEEE SENSORS - Proceedings*. Institute of Electrical and Electronics Engineers Inc., Dec. 2015. ISBN: 9781479982028. DOI: 10.1109/ICSENS.2015.7370112.

[86] Hongyi Wen, Julian Ramos Rojas, and Anind K Dey. "Serendipity: Finger Gesture Recognition Using an Off-the-Shelf Smartwatch". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3847–3851. ISBN: 9781450333627. URL: https://doi.org/10.1145/2858036.2858466.

[87] R B Widodo et al. "The IMU and Bend Sensor as a Pointing Device and Click Method". In: *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. 2019, pp. 294–297. DOI: 10.1109/ISITIA.2019.8937086.

[88] Erwin Wu et al. "Back-hand-pose: 3D hand pose estimation for a wrist-worn camera via dorsum deformation network". In: *UIST 2020 - Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 2020, pp. 1147–1160. ISBN: 9781450375146. DOI: 10.1145/3379337.3415897.

[89] Weipeng Xu et al. "Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera". In: *CoRR* abs/1803.0 (2018). URL: http://arxiv.org/abs/1803.05959.

[90] Hui Shyong Yeo et al. "Opisthenar: Hand poses and finger tapping recognition by observing back of hand using embedded wrist camera". In: *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 2019. DOI: 10.1145/3332165.3347867.

[91] Cheng Zhang et al. "FingerPing: Recognizing Fine-Grained Hand Poses Using Active Acoustic On-Body Sensing". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–10. ISBN: 9781450356206. DOI: 10.1145/3173574.3174011. URL: https://doi.org/10.1145/3173574.3174011.

[92] Yang Zhang and Chris Harrison. "Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition". In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. UIST '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 167–173. ISBN: 9781450337793. DOI: 10.1145/2807442.2807480. URL: https://doi.org/10.1145/2807442.2807480.

[93] Yang Zhang, Robert Xiao, and Chris Harrison. "Advancing Hand Gesture Recognition with High Resolution Electrical Impedance Tomography". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 843–850. ISBN: 9781450341899. DOI: 10.1145/2984511.2984574. URL: https://doi.org/10.1145/2984511.2984574.

[94] Christian Zimmermann and Thomas Brox. "Learning to Estimate 3D Hand Pose from Single {RGB} Images". In: *CoRR* abs/1705.0 (2017). URL: http://arxiv.org/abs/1705.01389.