学位申請論文

深度映像による時系列的位置情報を用いた
部位分割及び部位着目による
人物動作認識手法の提案

平成30年3月

東京理科大学大学院　理工学研究科　経営工学専攻
斎藤　裕佑

論文要旨

人物動作認識技術とは，人間が何らかの意図を持った行動を行う際の，その人物の関節位置や角度の変化，シルエットの様子などのアナログな情報をデジタルな信号へと変換し，そのデジタルな信号から自動的にその動作を言語化する技術のことである．人物動作認識技術は様々な場面への応用が期待されており，例えば健康管理を支援するシステム，ネットワークロボット等への応用が考えられる．具体的には，食べる, 話すといった身体動作を認識することによって，健康管理システムや生活支援システムへと応用できる．さらにその状況の中でも，咀嚼と健康には深い関わりがある [23] ことから，咀嚼状況を理解することで，より詳細に健康に関する支援を行うことができる．

しかしながら，屋内環境は，照明環境が変化しやすい環境であると言える．従来の動作認識の研究においては RGB カメラを用いたものが多くあり, ある程度の照明環境に対応する手法も考えられてはいる [24] が，通常の RGB カメラの場合極端な照明環境の変化に対応することはできない．また，家具や生活品などが多く，動作認識の際，それらはデバイスと認識対象者との間の障害物となりうる．しかし，本来人間の動作は，その動作が見えている部分のみで完結している場合であれば，その部分の情報のみから何の動作をしているかを認識することができる．

人物動作認識技術の既存研究は，主に二つの体系に大別することができる．一つ目に，RGB カメラセンサによって撮影された RGB カメラ映像から人物領域に関する特徴を取得する画像差分による動作認識の研究である．しかし，RGB カメラセンサによる手法は，屋内における照明環境の変化に対応が困難となる．また，二つ目は骨格モデルに基づいた動作認識である．特殊なマーカーと複数のカメラによるモーションキャプチャ(MoCap System) は発達し，動作認識に有

用だと考えられるが，認識時にも装備が必要であったり高価であるなど，一般的な利用はほぼ不可能であると考えられる．マーカーを必要としない，色彩映像によるモーションキャプチャは現在研究が進められているが，光学式のモーションキャプチャは日常の照明環境や閉塞に弱く，また磁気式のモーションキャプチャでは，普段使用する家電製品や金属製品に精度が大きく左右されてしまう．また，照明環境に依存しない深度映像による骨格認識を用いた研究としては，Xia ら [33] や Wang ら [31] の研究が挙げられる．しかし，赤外線を利用しているためにデバイスと認識対象との間の障害物（これを閉塞という）によって認識のズレが生じてしまうという問題点があり，既存研究 [33] では，屋内環境にあるような，身体の一部を隠してしまうほどの閉塞を考慮していなかった．

そこで本論文では，照明変化に頑健な深度情報を用いて，屋内実環境における閉塞に依存しない動作認識手法を提案することを目的とする．まず身体全体の関節動作認識に着目する．実生活環境における閉塞を考慮するため, 身体骨格における関節を 6 つの部位に分け, その部位ごとに独立した特徴を生成する手法を提案する．本提案手法では, 部位ごとに独立した特徴を生成するために, 各部位における基準関節の方向情報を，特徴を生成する際の空間の軸と定める．そして一回の動作において，各部位の基準関節の方向情報を軸とした空間内にて，基準関節以外の関節の時系列位置情報の推移をヒストグラムに変換する．さらに，関節ごとのヒストグラムを統合し，統合されたヒストグラムを部位ごとの特徴量とする．次に，閉塞がない環境での動作によって Random Forest[1] を用いて分類器を作成しておく．その上で，未知の動作データに対して，各部位の特徴量に変換し，各部位の分類器に入力する．そして各部位の動作に対する確率推算値を重みとして集約し，動作を認識する部位アンサンブル予測による動作認識手法を提案する．

そして，提案手法が屋内環境における閉塞に対して有効であるかを示すために，閉塞がある場合とそうで無い場合の両方について同様に，日常動作における骨格の三次元位置情報の推移を記録したデータセットを作成し，評価を行った．閉塞のない状態でのデータを学習データ，閉塞のある状態でのデータをテストデータとし，既存手法 [33] を本データに適用した場合と提案手法を比較し

た．その結果，既存手法が 83.2%であるのに対し，本提案手法は 84.7%となり，提案手法の有効性を示すことができた．部位アンサンブル予測により，可視可能な部位の特徴の重みが重要視されやすく，部位に着目した動作認識手法を実現したと言える．

次に身体骨格における一部分として，顔における口や顎関節などの動きによる動作認識に着目する．顔の特徴的な表出点位置情報を用いて顔動作を認識する手法を提案する．鼻と顎の距離, 口の縦方向の距離, 横方向の距離において, それぞれの基本統計量を特徴の基本とした．さらにノイズ除去を行った上で, 定常性に関するパラメータを抽出し，特徴として加えた．そして, 複数の距離情報から得られた特徴に対して, Random Forest[1] による動作の分類を行う手法を提案した．また, 顔動作における定常性の特徴の有効性を示すために，食べる, 話す, 何もしない時のデータセットを作成し, 基本特徴量のみの精度と比較した．その結果，基本特徴量のみの場合は 80.0%であるのに対し，本提案手法は 83.3%となり，定常性に関する特徴の有効性を示すことができた．つまり，顔部位において口と顎の特徴的な表出点位置情報から，提案手法によって日常的な動作を判別することが可能であることを示した．

さらに，身体動作と顔動作の認識手法の融合について考察した．身体と顔を別々に動作認識した手法と，身体 6 部位，顔を 1 部位とした時のアンサンブル手法について考察した．また身体と顔の複合的動作におけるデータの相互作用について言及し，先の二つの手法それぞれについて，その影響を考察した．

以上より，照明変化に頑健な深度情報を用いて，屋内実環境における閉塞に依存しない動作認識手法を実現したと言える．

目 次

図 目 次

表 目 次

第1章　序論

序論では，本研究の背景となる動作認識技術，及びその応用システムについて，現状と問題点について示し，本論の目的について述べる．

1.1　背景

近年, センサデバイスの発達に伴い, 人物動作認識技術が発展してきている．人物動作認識技術とは，RGB カメラセンサや人感センサを用いることで，人間の動作という事象をデジタル信号へと置き換え, 人物の行動や状態を把握し理解することを支援する技術のことである．人物動作認識技術は，防犯システム [1] や監視システムなどのセキュリティ分野に応用されており，病院における見守りシステムや在宅介護システムなどへの応用が期待されている．このような背景として，近年核家族化や単身赴任などによる単身者が増えたことや，高齢化社会となり孤独死などが増えたことが原因として挙げられる．例えば，単身世帯の数は年々増加しており，特に高齢者の単身世帯が目立っている [36]．また, 孤独死は 65 歳以上の高齢者だけでなく, 40 代から 50 代においても多くみられる現象である [37]．すなわち，単身者や，家族と同居してはいるが外出している等の理由で誰も周りに居ない状態の人物においては，誰かが身近にいるような状態の人物と比べて，何らかの異変が起こった時のリスクが増えると考えられる．したがって，まるで身近に誰かがいるように，対象人物の状況を理解し，かつその状況に応じたサポート機能や情報提供を行うようなシステムが求められている．さらに，対象人物が「何を行っているか」を詳細に理解することによって，より具体的なシステムへと応用が可能であると考えられる．例えば，食べる, 話すといった身体動作を認識することによって，健康管理システ

[1] Alicelab, http://alicelab.jp/

ムや生活支援システムへと応用できる．さらにその状況の中でも，咀嚼と健康には深い関わりがある [23] ことから，咀嚼状況を理解することで，より詳細に健康に関する支援を行うことができる．

しかしながら，屋内環境は，照明環境が変化しやすい環境であると言える．従来の動作認識の研究においては RGB カメラセンサを用いたものが多くあり，ある程度の照明環境に対応する手法も考えられてはいる [24] が，通常の RGB カメラセンサの場合極端な照明環境の変化に対応することはできない．また，家具や生活品などが多く，動作認識の際，それらはデバイスと認識対象者との間の障害物となりうる．しかし，本来人間の動作は，その動作が見えている部分のみで完結している場合であれば，その部分の情報のみから何の動作をしているかを認識することができる．したがって，本研究では照明環境に依存しにくい深度情報を用いて，屋内実環境における人物の可視部分のみから動作認識を可能にすることを目指すものである．

1.2　人物動作について

そもそも人間は多数の関節から形成され，それらが複雑な動きをすることで，身体活動を行っている．身体活動は目的や動機の有無，また意図の強さなどに応じて，各研究者達が言葉の定義を行ってきた [13][11][18]．

ここで，森 [19] は動作認識の際，人間の身体活動を以下の 4 段階に分類した．

行為（Act）

明確な目的，動機を持ち，思慮・選択・決心を経て意識的に行われるものであり，反射的行動や本能的行動，睡眠時などの無意識行動とは異なる．行為把握（Act comprehension）の対象となる．

行動（Behavior）

全体的で何らかの意味付けを与えられる身体の働きのことである．行動理解（Behavior understanding）の対象となる

動作（Action）

身体的で，識別･分節化が可能な物理的運動･動作認識（Action recognition）の対象となる．

動き（Motion）
: 物理的な体・部位の運動・変化，動き計測（Motion measurement）の対象となる．

その上で森 [19] は，時系列的に入力されるパントマイムのような人の身体的な運動そのものを動作とし，対象となるモノや環境の状態と合わせて理解した言葉を行動としている．例えば,「ご飯を箸で食べる」行動の一部として,「食べる」という動作が存在する．本研究では，このような「動作」を分類し，認識することができる手法を提案する．

1.3　動作認識に関する既存研究

人物動作認識技術の既存研究は，主に二つの体系に大別することができる．一つ目に，RGB カメラセンサによって撮影された色彩映像から人物領域に関する特徴を取得する画像差分的な動作認識の研究である．画像差分的な動作認識における画像特徴としては，広域特徴と局所特徴が存在する．広域特徴に関する研究では，Bobick ら [2] や Yu ら [34] などが挙げられる．広域特徴の共通点としては，まず背景差分などによって人物領域を抽出し，その後注目領域から特徴を符号化することで特徴量を生成していることである．したがって，広域特徴を用いる手法では，前処理となる人物領域抽出の段階における人物領域自体の精度や，また使用するデバイスの視点や照明環境などによるノイズ，デバイスと認識対象の間の閉塞によって認識精度が過敏に変化してしまう．次に局所特徴を用いた研究では，Laptev ら [15] や南里ら [22] が挙げられる．局所特徴は，まず例えば映像内における画像的あるいは時間的に急激な変化が起きた点などの，時空間における特徴点を算出する．そして局所的な特徴を周辺領域において算出し，その特徴の記述子を統合し特徴量とする．局所特徴は広域特徴に比べ，ノイズや部分的閉塞に頑健であるが，単位の異なる異種特徴が混合してい

ることが多く，また一般に高い次元を持っている．したがって，コードブック等の記述子を用いて，ヒストグラムなどの特徴量へと圧縮されることが多い．

二つ目に，骨格モデルに基づいた動作認識の研究がある．早くは Cambell ら [5] の研究が挙げられる．特殊なマーカーと複数のカメラセンサによるモーションキャプチャ(MoCap System) は発達し，動作認識に有用だと考えられるが，認識時にも装備が必要であったり高価であるなど，一般的な利用はほぼ不可能であると考えられる．マーカーを必要としない，色彩映像によるモーションキャプチャは現在研究が進められているが，光学式のモーションキャプチャは日常の照明環境や閉塞に弱く，また磁気式のモーションキャプチャでは，普段使用する家電製品や金属製品に精度が大きく左右されてしまう．

対して本研究では，骨格モデルの認識には深度カメラセンサである Kinect[2] デバイスと，Kinect for windows SDK[3] を用いる．Kinect および同 SDK には，Shotton ら [27] による骨格認識技術が内包されており，30[fps] にて人間の三次元的な骨格を取得できる．さらに，人物動作は空間的なものであるため, 本研究では空間的な人物動作認識について焦点を当てる．また，赤外線による深度情報は屋内の照明環境などによる影響も少ないため，通常画像よりも望ましいと言える．ここで，骨格認識を用いた研究としては，Xia ら [33] や Wang ら [31] の研究が挙げられる．しかし，赤外線を利用しているために，デバイスと認識対象との間の障害物（これを閉塞という）によって認識のズレが生じてしまうという問題点がある．

1.4　屋内動作認識に関する問題点

本研究では，屋内における動作認識に焦点を当てる．まず屋内においては，電気などの照明環境は変化しやすいが，人物は暗い場所でも動作を行う．したがって，本研究では赤外線による深度情報による動作認識に言及する．

そして，次に閉塞について議論する．閉塞が起こりうる状況は，まずデバイスと認識対象との間の障害物が存在する場合である．屋内環境において，実生

[2]Microsoft Kinect, http://www.xbox.com/ja-JP/kinect
[3]Kinect for windows, http://www.microsoft.com/en-us/kinectforwindows/

活環境においては閉塞を引き起こしやすい家具などが配置されていることが多い．例えば，机や椅子などが挙げられるが，環境によって机の高さなども変わってくる．しかし，そのような場合でも，手と顔の動きのみ認識できれば，食べるなどの動作を識別することは可能であると考えられる．

ここで，骨格モデルに基づいた動作認識では，まず姿勢を推定したのちに，骨格情報の時系列変化を用いて特徴量へと変換する．閉塞が起こりうる状態とは，先に述べたデバイスと障害物によるものだけではなく，認識対象者自身の部位が他の部位を隠す場合もある．閉塞が見られる場合，姿勢推定自体の精度に影響を及ぼす．本研究においては，これらの閉塞が起こった場合でも，可視可能な部分の特徴に着目できるような動作認識手法の提案を目指すものである．

1.5　本研究の目的

先に述べた通り，屋内実環境には，照明環境が変化しやすく，また家具などの配置によってデバイスが認識対象者全てを捉えることはできない可能性が高い環境となっている．したがって本研究の目的は，照明変化に頑健な深度情報を用いて，屋内実環境における人物の可視部分に着目して動作認識を可能にする手法を提案することである．

本目的を達成するために，深度映像を用いた骨格認識技術 [27] によって算出した身体骨格の三次元位置情報を用いる．その上で，デバイスと認識対象者との間に起こる閉塞を考慮し，骨格を 6 つの部位に分けた上で，それぞれの部位が独立した特徴を生成することができる手法を提案する．その際，関節位置の三次元位置情報の時系列推移を，動作の際支点となる関節を中心とした角度のヒストグラムへと変換し，特徴量として用いることで動作認識を実現する．また，身体の一部として，顔の特徴的な点の三次元位置情報の時系列推移を用いて顔の動作を認識する手法の提案を行う．顎関節と口の特徴的な位置の三次元位置情報の推移から，定常性を考慮した上で顔の動作認識を行うことができるかどうかを検証する．

1.6　認識対象とする動作

ここで本研究において，屋内においての動作を以下のように分類する．

1. 移動を伴う動作

何もしないままただ移動のみを目的として歩く動作を行う場合と，そうでない場合とが存在する．後者の場合の一例として，何かモノを移動させることを目的として，手を保持したまま移動する「運ぶ」を対象とする．つまり本研究では，歩く，運ぶ，を対象とする．

2. 腕のみで行うことができる動作

座ったままでも，立ったままでも，その状態に関係なくただ腕のみあるいはその周辺の関節だけを動作させることで目的を達成する行動が存在する．その行動における動作は数多く存在するが，特に屋内の場合，何かを拾ったり，ゴミをゴミ箱へ投げたりする動作が考えられる．つまり本研究では，投げる，拾うを対象とする．

3. 姿勢の状態自体が大きく変化する動作

特に屋内の場合，椅子やソファなどが置かれていることが多く，またそのモノに関わる動作としては立つ，座る，が主な動作として挙げることができる．したがって本研究では，立つ，座る，を対象とする．

4. 顔における動作

また，動作認識において，より細かく人物の動作を認識するために，身体の一部としての顔に着目する．ここで，屋内において，机などの閉塞がある状態でも，顔の部分は見えていることが多い．また顔を身体動作と独立して動くことができる．例えば，食べる際の咀嚼動作は，顎関節を動かすことによって行うことができる上に，健康と関連が深い．また，認識対象者が話す動作を認識できれば，より対象者の振る舞いに応じたサービスへと応用できると考えられる．したがって本研究では，顔の動作として，特に食べる，話す，を対象とする．

1.7　本論文の構成と概要

第一章「序論」では，人物動作認識について，その利用イメージや，そもそも動作とはどう定義されるかについて論じた．そして，既存の人物動作認識に関する研究体系を説明し，さらに屋内環境における人物動作についての問題点について論じた．そして，本研究で取り組む課題について言及し，目的を明らかにした．

第二章「関連研究」では，動作認識手法を画像差分による手法と骨格モデルを用いた手法を，それぞれについてさらに RGB カメラセンサと深度カメラセンサを用いた手法紹介する．その上で，深度映像を利用した骨格モデルによる手法について着目し，現状の問題点を指摘した．また，顔における動作認識手法においても，既存研究を紹介した上で，現状の問題点を指摘した．その上で，本研究の位置付けを行った．

第三章「部位分割的動作認識手法」では，まず身体動作認識において，閉塞となる障害物の多い実生活環境に対応するため，身体骨格における関節を 6 つの部位に分け, 部位ごとに独立した特徴を生成する手法を提案する．ここで，部位ごとに独立した特徴を生成するために，各部位において動作の際に支点となる関節を一つ設定し，その関節を基準とした特徴量を生成する．さらに，各部位において Random Forest[1] を用いて分類器を作成しておき，予測する入力データに対して，各分類器に入力した際に出力される動作ラベルごとの確率推算値を重みとして用いる部位アンサンブル予測による動作認識手法を提案する．提案手法が閉塞に対する有効性を示すために，閉塞状況がある場合とそうで無い場合の両方について同様に，日常動作における骨格の三次元位置情報の推移を記録したデータセットを作成し，評価を行う．評価方法として，本データセットに対して，Xia[33] の手法を適用した上で，閉塞のない環境でのデータを学習データ，閉塞のある環境でのデータをテストデータとし，本提案手法と認識精度を比較する．

第四章「顔動作認識手法」では，身体骨格における一部分としての顔の骨格及びパーツの動作に着目する．鼻と顎の距離, 口の縦方向の距離, 横方向の距離

において, それぞれの基本統計量を特徴の基本とした顔動作認識手法を提案する．さらにノイズ除去を行った上で, 定常性に関するパラメータを自己相関によって求め，その値を特徴として加える．そして, 複数の距離情報から得られた特徴量に対して, Random Forest[1] による動作の分類を行う．また, 提案手法において，定常性の効果を検証するために，食べる, 話す, 何もしない時のデータセットを作成し, 定常性を考慮しない場合の基本特徴量のみを特徴として用いた場合と，定常性によるパラメータを特徴として加えた場合とで分類性能を比較し評価する．

第五章「考察」では，第 3 章と第 4 章の結果を踏まえた上で，身体動作認識と顔動作認識の融合について考察する．本研究が想定する環境を明確にした上で，身体動作認識と顔動作認識を別々に行いラベルを結合する手法と，身体 6 部位，顔を 1 部位とした時の部位アンサンブル予測による手法について述べる．また，身体動作と顔動作の複合的動作の際に，身体及び顔のデータ間の相互作用について考察する．さらに，先述した二つの手法について，身体動作と顔動作の複合的動作における相互作用による影響について考察する．

第六章「結論」では，本論文の成果を述べる．本論文の成果とは，人物動作認識技術において，屋内実環境における閉塞に依存しない動作認識手法を提案したことである．また，本研究によりどのようなシステムが可能になるかを今後の展望として述べる．

第2章　関連研究

本研究では，まず動作認識技術を骨格モデルを用いない画像差分的な動作認識手法について説明する．次に，骨格モデルに基づく関節位置情報を用いた骨格的な動作認識手法について説明する．また，身体部分の一部として顔の動作認識に関する既存研究の紹介を行った上で，本研究の位置付けを行う．

2.1　画像差分を用いた動作認識手法

2.1.1　RGB 映像による画像差分による認識手法

まず動作認識技術において，RGB 映像のみで骨格認識を利用しない，画像差分を用いた認識手法が存在する．画像差分的認識手法の動作認識における画像特徴としては，広域特徴と局所特徴が存在する．広域特徴に関する研究では，Bobick ら [2] や Yu ら [34] が挙げられる．Bobick ら [2] は各フレームで作成したシルエットの変化領域を抽出し，一時的なテンプレートを作成しマッチングさせることで動作を認識している．また Yu ら [34] は, 画像から人物領域の輪郭を算出し, 重心から人物の頭, 手, 脚の画像領域を推定する variable star skeleton representation (VSS) を提案し，また動作中の画像領域がどの部位の画像領域に属するかのヒストグラムを作成し，動作認識に用いている．広域特徴の共通点としては，まず背景差分などによって人物領域を抽出し，その後注目領域から特徴を符号化することで特徴量を生成していることである．したがって，広域特徴を用いる手法では，前処理となる人物領域抽出の段階における人物領域自体の精度や，また使用するデバイスの視点や照明環境などによるノイズ，デバイスと認識対象の間の閉塞によって認識精度が過敏に変化してしまう．

次に局所特徴を用いた研究では，Laptev ら [15] や南里ら [22] が挙げられる．

局所特徴は，まず例えば映像内における画像的あるいは時間的に急激な変化が起きた点などの，時空間における特徴点を算出する．そして局所的な特徴を周辺領域において算出し，その特徴の記述子を統合し特徴量とする．Laptev ら [15] は Harris のコーナー検出手法を時系列方向に拡張し，STIP として活用した．しかし，継続して行われる動作に関しては STIP が生成されないため，長時間に区切られるような継続的な動作の認識には適していないことが知られている．また，南里ら [22] は，局所特徴を利用した異常動作の検出を試みている．局所特徴は広域特徴に比べ，ノイズや部分的閉塞に頑健であるが，単位の異なる異種特徴が混合していることが多く，また一般に高い次元を持っている．したがって，コードブック等の記述子を用いて，ヒストグラムなどの特徴量へと圧縮されることが多い．本研究においては，骨格位置の三次元位置情報の時系列推移，すなわち四次元情報を各関節ごとに扱うため，高次元になる特徴をヒストグラムに圧縮する．ここで，画像差分的認識手法では RGB 映像のみを用いるため，通用公開されているデータセットを使用される場合が多い．データセットの例として Hollywood human action dataset[4] などがある．これらのデータセットは映画や Web 上の映像で構成されており，あらかじめ正解ラベルが付与されている．動作認識において時間情報は動作を識別する上で非常に重要であり，動作認識を目的とした多くの画像特徴は時間軸を考慮した特徴生成を行っている．しかしながら，RGB 映像を用いているため，極端な照明環境の変化に対応しているとは言い難い．

2.1.2　深度映像における画像差分による認識手法

また，深度映像を用いるが, 骨格認識技術を用いない画像差分的認識手法を挙げる．Li ら [16] は，人間の骨格情報ではなく，人間と背景の分別のみ利用し，人間のシルエット全体の深度情報を利用した動作認識手法を提案した．また Jalal ら [9] は, 深度シルエットからラドン変換を用いて三次元シルエットを再構成し, 動作認識を行う手法を提案した．また掃除や料理などの家庭内の動作に対して動作認識精度の評価を行っている．さらに, Oreifej ら [25] は，深度画像におけ

[4]Hollywood human action dataset，http://www.di.ens.fr/ laptev/download.html

る人物領域の時系列な特徴を用いている．また，Chaaraoui ら [6] は, 人体骨格の一部と, 深度画像における人物の輪郭から特徴を抽出している．しかしながら，深度画像における人物領域を用いた手法は，骨格関節の複雑な構造を考慮していないため，人間の空間的で微細な動きを捉えきれない．郷津ら [8] も，一般的に深度画像を利用した画像差分的認識手法よりも，骨格モデルによる動作認識手法の方が分類精度が高い傾向にあるとしている．したがって本研究においては，深度映像における骨格的な動作認識手法を提案する．

2.2　骨格モデルに基づいた動作認識手法

2.2.1　接触型骨格認識技術を用いた手法

骨格モデルに基づいた動作認識手法の研究としては，早くは Cambell ら [5] の研究が挙げられる．また, 村尾ら [20] は手首, 足首, 腰に加速度センサを取り付け，歩く動作などに見られる繰り返しによって起こる，センサデータの定常性を考慮した自己相関を用いた動作認識手法を提案している．また近年郷津ら [8] は，接触型の骨格認識技術を用いて，動作認識に留まらず,「新聞を読む」といった微細な動作を分類し，行動を説明する短文を生成する手法を提案している．このように，特殊なマーカーと複数のカメラによる骨格認識技術は発達し，動作認識に有用だと考えられる．しかし，認識時にも装備が必要であったり高価であるなど，一般的な利用はほぼ不可能であると考えられる．マーカーを必要としない，RGB 映像による骨格認識技術は現在研究が進められているが，光学式の骨格認識技術は日常の照明環境や閉塞に弱く，また磁気式の骨格認識技術では，普段使用する家電製品や金属製品に精度が大きく左右されてしまう．

2.2.2　非接触型骨格認識技術

ここで，深度画像を用いた骨格認識技術が Shotton ら [27] によって提案された．深度画像における関節位置とその周囲の深度情報を予め Random Forest[1] によって学習させておき，入力された深度画像から，各画素がどの関節に所属するかを識別する．その後，密度推定を行って最瀕値を推定することで，関節

の三次元位置を推測する．また, Shen[26] は動作の際に支点となる関節を静的な関節とし，他の関節位置の予測精度を向上させている．さらに，Taylor ら [29] は閉塞部分の関節を推測する手法を提案している．これらの研究から，深度映像を用いて人物の骨格を 30[FPS] 以上の速さで推定でき，深度映像を用いた骨格的認識手法が発展してきた．本研究においても，深度画像を用いた骨格認識技術によって推定された骨格の関節位置情報を用いることで，骨格的認識手法の提案を行う．

2.2.3　RGB 映像による骨格モデルによる認識手法

近年, 深層学習の発展により，RGB 映像から人物骨格位置を算出することが可能となっている．深層学習とは，多層のニューラルネットワークによる機械学習手法の一種であり, 学習のための計算機能力の向上や，WWW(world wide web) の発達による学習データの収集の容易さ等により, 昨今急速に進展を遂げている．Toshev ら [30] は，深層学習によって，RGB 画像から人物の骨格をする手法を提案している．さらに，姿勢の事前知識に基づいて，通常画像から三次元的な姿勢を推定する研究も行われてきている [12]．すなわち，近年 RGB 映像からでも骨格認識技術を行うことができるようになった．そして，その RGB 映像の骨格認識技術を用いることで，RGB 映像での骨格的認識手法の研究が発展している．Wang ら [32] は，RGB 画像から人体の骨格関節を平面的に 14 に分け推定し，さらにその 14 の二次元的な骨格関節を 5 つのグループに分けて動作を分類している．しかしながら，RGB 映像を用いた手法は，照明環境の変化に対応することが困難である．

2.2.4　深度映像による骨格モデルによる認識手法

次に, 深度映像による骨格認識を利用した，骨格モデルによる動作認識手法を取り上げる．Xia ら [33] は得られた骨格の三次元位置情報の推移を，腰を中心とした半球体上の空間を角度によって 84 分割した上でヒストグラムを作成し，クラス分類を行うことで動作認識を実現させている．しかし，実生活環境での利用を考えた場合，家具や生活品などの閉塞になりうる障害物が多いが，Xia ら [33] はその閉塞を考慮に入れていない．

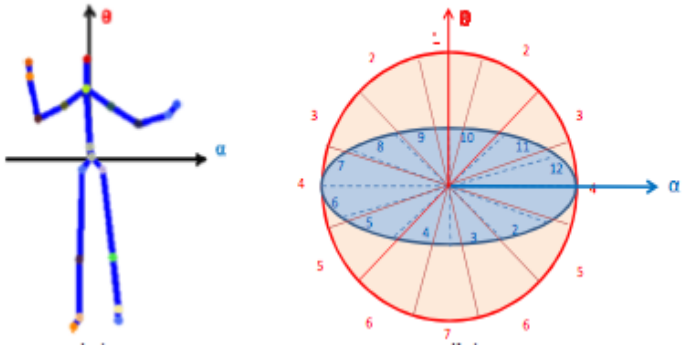


図 2.1: 骨格位置情報によるヒストグラム作成法 [33] における空間分割

また，Wang ら [31] は，三次元骨格位置同士の距離を特徴量とし，またある骨格位置付近の深度情報の空間的占有パターンを Local Occupancy Patterns(LOP) 特徴量として定義し，これらの特徴量を時間的フーリエ階層構造 (FTP, Fourier Temporal Pyramid) によって特徴付けたものをクラス分類している．さらに同研究では，FTP 特徴量を結合する構造をアクションレットとして定義した．その上でデータマイニング手法を使用することで，各動作における最も顕著なアクションレットを発見し，4 つの組の関節を信頼度とあいまい性を評価することによってアクションレットに含めた．しかしながら，このアクションレットの組は，右腕と左脚, または右腕と左腕の両方に見られていたり，同じ動作でも大きく異なる関節の組が選択されている．したがって，その一方の関節あるいは周辺を塞ぐように閉塞が起こっていた場合には適していない．

2.3　顔部分の動作認識手法に関する研究

田口ら [28] は，外耳に装着して，外耳道の咀嚼に伴う変形を検知して咀嚼動作を認識する咀嚼認識装置及び手法を開発している．また，Li ら [17] は, 1cm に満たない加速度センサを口内の歯に埋め込み, 咳をする, 咀嚼する, 飲む, 話すといった動作を識別する手法を提案している．しかし，身体動作認識における Cambell ら [5] の研究と同様に，ユーザーに対して常にこれらのデバイスを装着することを強いることになる．また口内に埋め込む場合，センサの故障の際のメンテナンスが困難である事や，センサを誤飲してしまう恐れがある．

さらに，Cadavid ら [4] は通常の RGB カメラを用いて, AAM(Active Appearance Model)[7] による顔の特徴部位を用いて, 開口咀嚼, 閉口咀嚼, 話す, 感情表現をする, 何もしない, といった動作を識別している．しかし, Cadavid ら [4] の手法では, 通常のカメラを使用しているため，環境光などの状態によっては識別が不可能となると考えられる．また人間の動作は特定の周波数で動いているわけではないため，フーリエ変換による周波数推定は適していない．

2.4　本研究の位置付け

ここで，本研究の位置付けを行う．人物動作認識技術は，画像情報から直接，骨格モデルを用いずに人物動作を認識する画像差分的な認識手法と，骨格モデルに基づいて，対象となる人物の骨格の関節位置情報を推定した上で動作認識を行う骨格的認識手法に大別することができる．まず本研究は，深度情報を用いた骨格認識技術 [27] による骨格的認識手法である．既存の深度情報による骨格的認識手法 [33][31] では考慮されていなかった屋内実環境での利用を考慮し，閉塞に頑健な動作認識手法を提案する．解剖学的検知 [3][14] を参考に動作の際支点となる関節 (基準関節) を設定し，その関節が必ず 1 つ入るように関節を 6 つの部位に分けることで，それぞれの部位単体での特徴を生成する．さらに静的な関節の方向情報を用いることで，部位ごとに独立した特徴量を生成する手法を提案する．

また, 身体の一部位のさらに細かな部分である顔における動作認識としても，

非接触型であり，また深度映像による顔の特徴的な点を用いた骨格的認識手法である．さらに, 本研究の対象とする動作において，食べる動作における顔の顎関節や, 話す動作における口の開閉の様子など，顔において反復性が見られる部分が多い点に着目する．ここで，定常性が見られる動作に対して，自己相関を用いた動作認識手法が有効であるとされている [20]．そこで，顔の動作においては，反復時における三次元位置データ推移の定常性を考慮する上で，自己相関を用いて特徴量を生成する．

第3章　部位分割的動作認識手法

本研究の目的は，照明変化に頑健な深度情報を用いて，屋内実環境における人物の可視部分に着目して動作認識を可能にする手法を提案することである．本目的を達成するために，深度映像を用いた骨格認識技術 [27] によって算出した身体骨格の三次元位置情報を用いる．その上で，デバイスと認識対象者との間に起こる閉塞を考慮し，解剖学的検知 [3][14] から，骨格関節位置群を上半身中心部位，下半身中心部位，右腕部位，左腕部位，右脚部位，左脚部位の 6 つの部位に分割する．なお，各部位に一つ，動作の際に支点となる関節（基準関節）を定めた．そして，各部位において独立した特徴を生成するために，基準関節における方向情報を用いて，部位ごとに空間の軸を定める．基準関節における方向情報を用いることで，それぞれの部位において独立した座標系を持つ空間となる．したがって，いくつかの部位が閉塞によって隠れてしまった場合の影響を最小限に抑えるような特徴を生成することができるようになる．

特徴量となるヒストグラムの生成においては，Xia ら [33] らの手法を応用し，各部位におけるヒストグラムを生成する．ここで，Xia ら [33] らは腰周辺の関節から空間の軸を決定し，身体全体に対し一つの座標系においてヒストグラムを生成していた．本研究においては各部位の基準関節の方向情報が求められているため，各部位ごとに Xia ら [33] らの手法を適用できる．具体的には，まず基準関節の座標軸（X 軸，Y 軸，Z 軸）による平面（YZ 平面，XZ 平面，XY 平面）のうち，二つの平面を選択する．次に，各部位における空間を選択した二つの平面に分け，さらに各平面を角度によって分割する．そして，各平面の領域の組み合わせをヒストグラムの階級とする．ヒストグラムへのカウントの方法としては，動作中の各フレームにおいて，三次元関節位置の時系列推移を各平面に射影することで該当する階級にカウントしていく．さらに，動作の全

フレームにおける各階級におけるカウントの累積を，全カウント数で除算することで，動作時間における正規化を施し，関節ごとのヒストグラムを生成する．また，各部位において，関節ごとのヒストグラムを結合し，部位ごとのヒストグラムとする．

次に，このヒストグラムを特徴量として機械学習を用いることにより動作を認識する．あらかじめ，閉塞されていない状態のデータを用いて，各部位において Random Forest[1] を用いて学習を行っておく．ここで，Ramdom Forest による分類器は，入力されたデータに対して，動作ラベルごとにどの程度信頼性を持って動作を判別しているかを表す確率推算値を出力できる．したがって，未知のデータが入力された場合，各部位における動作ごとの確率推算値が出力される．最後に，動作ラベルごとに全部位におけるそれぞれの確率推算値を重みとして総和をとり，最も確率推算値の総和が最も高い動作ラベルを，出力する動作として認識する．

3.1　設計方針

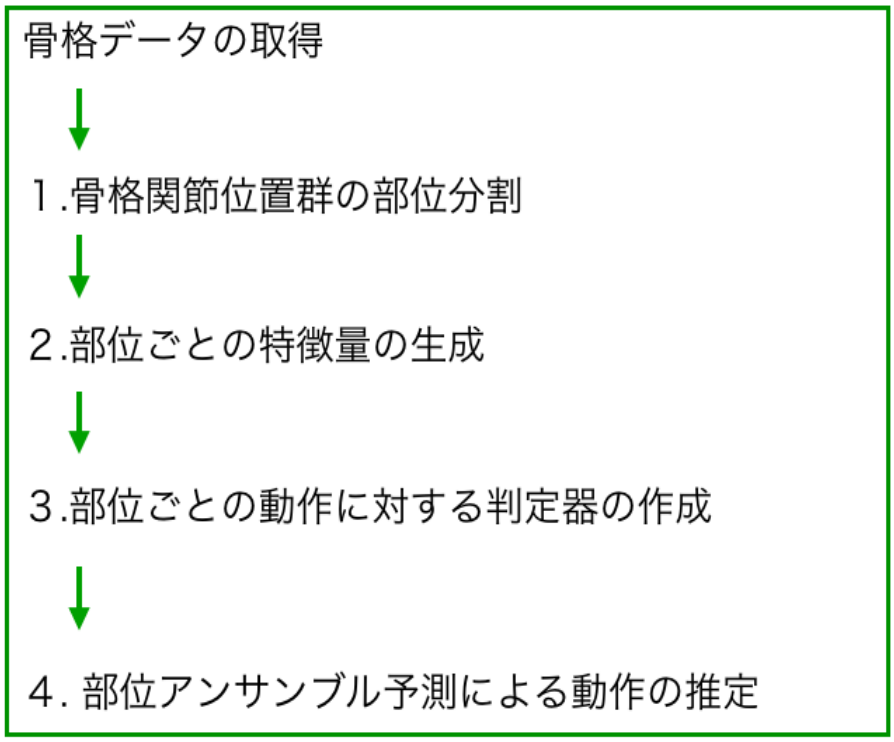


図 3.1: システムフロー

ある一つの動作を認識する際の，深度情報および骨格認識によって骨格の関節位置群を追跡してから，特徴量となるヒストグラムを生成するまでの全体の流れを，図 3.1 および以下に示す.

1. 深度情報を用いた骨格認識によって，骨格位置の三次元位置情報と関節の方向情報を得る．さらに，得られた骨格位置情報において，6 つの関節の部位（上半身中心部位，下半身中心部位，右腕部位，左腕部位，右脚部位，左脚部位）に分割する．さらに，それぞれの部位において，動作の際に支点となるような基準関節を設定する．骨格関節群を部位分けすることで，可視部分のみで特徴量を生成することができる．

2. 次に，部位ごとに独立した特徴の生成について説明する．

   (i)6 つの部位に骨格を分割したのち，部位ごとに独立したヒストグラムを生成するために，部位の関節の位置情報を，基準関節の方向情報を空間の軸とした座標系へと変換する．まず，基準関節の方向情報を用いて，基準関節での空間の座標系を決定する．基準関節の方向情報を用いることで，部位ごとに独立した特徴を生成することができる．

   (ii) 基準関節における座標系の空間を決定したのち，各部位ごとにおけるヒストグラムを生成する．まず基準関節の座標軸（X 軸，Y 軸，Z 軸）による平面（YZ 平面，XZ 平面，XY 平面）のうち，二つの平面を選択する．そして各部位における空間を選択した二つの平面に分け，さらに各平面を角度によって分割する．そして分割されてできた各平面の領域の組み合わせをヒストグラムの階級とする．

   (iii) 階級を決定したのち，各部位内での基準以外の関節の位置情報を，二つの平面へそれぞれ射影し，角度情報に変換する．そして，各平面での領域の組み合わせに属する階級にカウントしていく．動作した時間の長さに対し影響しないよう，追跡値および推測値のカウント回数から，各階級の値を総カウント数によって除算することで正規化する．また，ヒストグラムを各部位に属する基準関節以外の全ての関節において生成する．

   (iv) そして各部位において関節ごとのヒストグラムを結合し，部位ごとに一つのヒストグラムを特徴とする．

3. 部位ごとに特徴を生成したのち，部位アンサンブル予測による動作認識を行う．まずあらかじめ閉塞がない状態のデータを学習データとして用いて，部位ごとに Random Forest[1] による分類器を生成する．ここで，Ramdom Forest による分類器は，入力されたデータに対して，動作ラベルごとにどの程度信頼性を持って動作を判別しているかを表す確率推算値を出力できる．

4. 次に，認識する動作，すなわち未知の動作データが入力された場合について説明する．あらかじめ学習させておいた各部位における分類器によって，一部位に対して動作ごとの確率推算値が出力される．最後に，動作ラベルごとに全部位におけるそれぞれの確率推算値を重みとして総和をとり，最も確率推算値の総和が最も高い動作ラベルを，出力する動作として認識する．全身で一つの分類器を用いる既存手法 [33] とは異なり，本研究では部位ごとに学習した分類器による確率推算値を重みとして用いており，この動作認識手法を部位アンサンブル予測と定義する．

3.1.1　骨格関節位置群の部位分割

本研究の主なアイデアは，個々の関節グループの動きは独立したものが多い，という考えに基づいている．例えば，”水を飲む”という動作において，重要な動きは手の動きであり，その時足の動きや体の動きの影響は少ない．したがって，手の動きに注視することでラベル付けはできるはずである．ここで，人間の動作は，各関節の回転動作によって構成されることに着目する．解剖学の検知 [3][14] から，人体の関節的な動きに関わる部分は上肢と下肢にわけられる．さらに上肢と下肢のそれぞれにおいて，肩から先は可動性が高い自由上肢，股関節から先は可動性が高い自由下肢とされている．したがって，本研究においては，上肢と下肢における中心部として上半身中心部位，下半身中心部位を定める．さらに左右の自由上肢および自由下肢として，左腕部位，右腕部位，左脚部位，右脚部位に分割する．すなわち，骨格関節位置群を上半身中心部位，下半身中心部位，左腕部位，右腕部位，左脚部位，右脚部位の 6 つに分割する．また基準関節を，まず上半身中心部と下半身中心部において，それぞれで身体の軸となる関節である肩の中央，腰の中央とし，さらに左右の自由上肢と自由下肢の根元となる関節である左肩，右肩，左お尻，右お尻とする．このように，骨格関節位置群を 6 つの部位に分割することで，部位単体における特徴を生成することができ，さらに他部位が閉塞によって隠れてしまった場合でも，可視部分の部位のみで動作を認識することができる．加えて，それぞれの部位において基準関節を設定することで，部位ごとに独立した特徴を作ることができる．ここで，本研究においてはこのように骨格を 6 つの部位に分割したが，今後骨格認識技術が発展し，さらに手指などの細かい関節が認識できるようになった場合，さらに細かく部位を分割することで，より微細な動きを捉え，高度な動作認識を行うことができると考えられる．

骨格関節位置群の部位分割について，さらに具体的に述べる．まず深度情報から骨格認識により，骨格関節群の三次元位置情報を取得する．骨格認識によって得られる関節の数を N 個とした場合，検出された骨格の関節は，腰の中央を始点として（O_n : n = 0，1，…，N．O_0 は腰の中央）と表すことができる．そして，それぞれの関節位置の三次元位置情報は，($P_{O_n} = (P_{O_n}x, P_{O_n}y, P_{O_n}z)$, n = 0，1，…，20，$O_0$ は腰の中央）と表すことができる．次に，得られた骨格関節群を図 3.2 のように，上半身中心部，下半身中心部，左腕部，右腕部，左脚部右脚部，の 6 つの部位に分割する．また図 3.2 に示す通り，本研究では各部位ごとに 1 カ所ずつ，基準関節をそれぞれ肩の中央，腰の中央，左肩，右肩，左お尻，右お尻としている．ここで最終的にヒストグラムデータは，基準関節位置に対して，ヒストグラムを生成する関節位置の角度情報の蓄積によるものである．したがって，この時点において，関節位置情報群を個人差が発生しないように正規化を行う必要がないことに留意されたい．

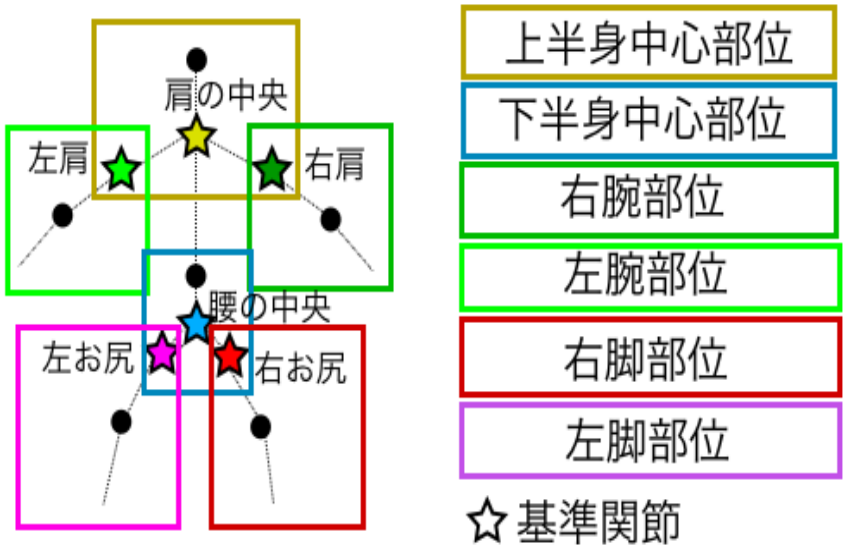


図 3.2: 6 つの部位の分割方法

3.1.2　各部位における特徴量の生成

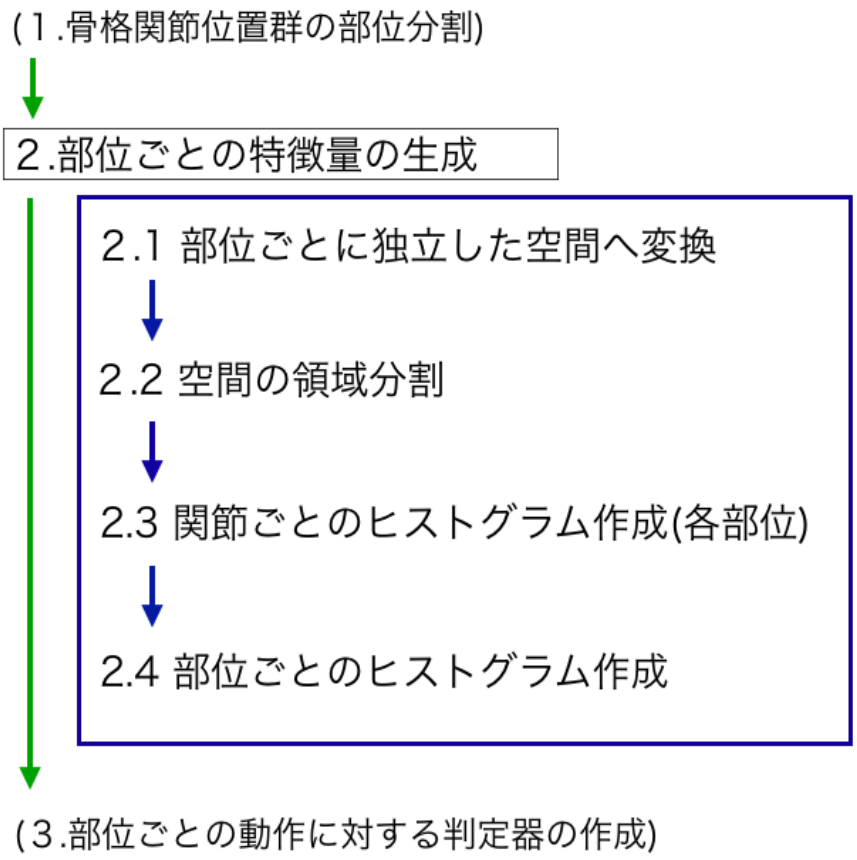


図 3.3: 各部位における特徴量の生成の流れ

6 つの部位に骨格を分割したのち，部位ごとに独立した特徴量を生成するためのシステムの流れを，図 3.3 に示す．まず，(i) 部位ごとに設定された基準関節の方向情報を軸とした空間へと変換する．その部位に属する深度カメラセンサからの絶対的な関節位置情報を，基準関節に対する相対的な関節位置情報へ変換する．(ii) 基準関節の位置情報を原点及び方向情報を軸とした空間において，空間を角度によって分割する．この時，一旦空間における二つの平面に特徴を作成する関節の位置情報を射影し，それぞれの平面において角度をとることで，三次元空間における角度情報を過不足なく表現することができる．そして，各平面に対し角度によって領域を決定しておく．この領域が，各平面にお

けるヒストグラムの階級となる．(iii) そして，各平面に射影された関節位置がどの角度の領域に属するかを，各平面におけるヒストグラムの頻度としてカウントしていく．そして，二つのヒストグラムを組み合わせ，各関節のヒストグラムを生成する．(iv) さらに，各部位において，基準以外の関節のヒストグラムを結合し，部位ごとの特徴量としてこれを用いる．

(i) 基準関節の方向情報を軸とした空間変換方法

ヒストグラムの作成は，部位ごとに定められた基準関節を原点とした三次元空間を利用する．まず，深度カメラセンサ自身における空間の直交座標系における座標軸を，実世界空間における空間の基準の座標系とする．ここで，関節の方向情報について述べておく．図 3.5 のように，腰の中心部を基準とした直交座標系を，隣接する関節位置への方向が Y 軸となるように回転していく．この回転は，回転行列として保存される．すなわち，関節の方向情報とは，その関節における直交座標系が，基準とした腰の中心部からどれだけ回転させたかを表す情報である．次に，ある一つの部位における基準関節の方向情報を，その部位における空間の座標系とする．深度カメラセンサ自身が実世界空間における空間の基準の座標系であるのに対し，部位における基準関節の方向情報が空間の基準の座標系となる．このように，基準関節の方向情報を，部位における空間の軸とすることで，部位ごとに独立した特徴を生成することができる．基準関節における方向情報を部位における基準の空間座標系とすることで，他の部位の状態に影響されない特徴量を生成することができる．例えば，投げるという動作における場合を考える．身体全体が傾くなどしたときの利き腕の部位における各関節のヒストグラムは，直立しているときと比べても，利き腕の動作が同じ動作であるならば，利き腕の関節のヒストグラムは身体の傾きに影響を受けない．

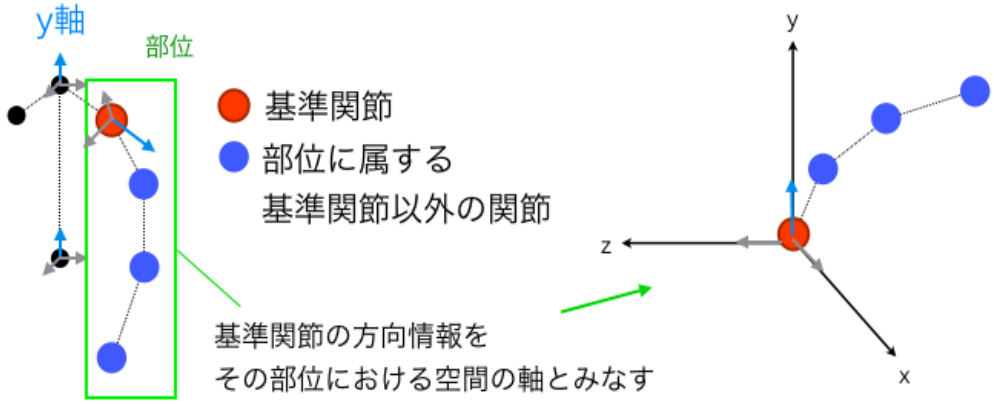


図 3.4: 部位内における空間軸の決定

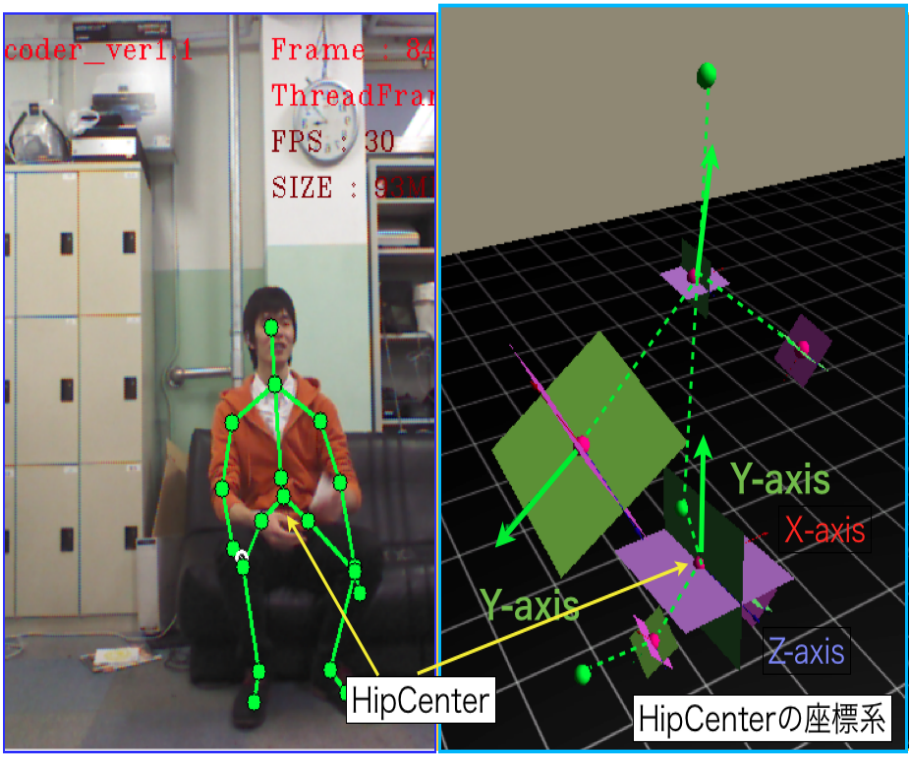


図 3.5: 関節の方向について

(ii) 関節ごとのヒストグラムの階級の決定

分割された領域およびその領域へのカウント方法について，図 3.7 に示す．ヒストグラムにおいて，分割された領域を階級という．基準関節における座標系の空間において，まず基準関節の座標軸（X 軸，Y 軸，Z 軸）による平面（YZ 平面，XZ 平面，XY 平面）のうち，二つの平面を選択する．そして各部位における空間を選択した二つの平面に分け，ここで二つの平面のうち，片方の平面は 360 度全平面をある角度 θ で分割し，もう一方の平面は全面ではなく半分の 180 度のみの平面を角度 θ で分割する．これは，のちの二つの平面でのヒストグラムを統合する際，360 度全平面を用いた平面によって，もう一方の使用しない半分の領域はすでにその範囲となっているためである．

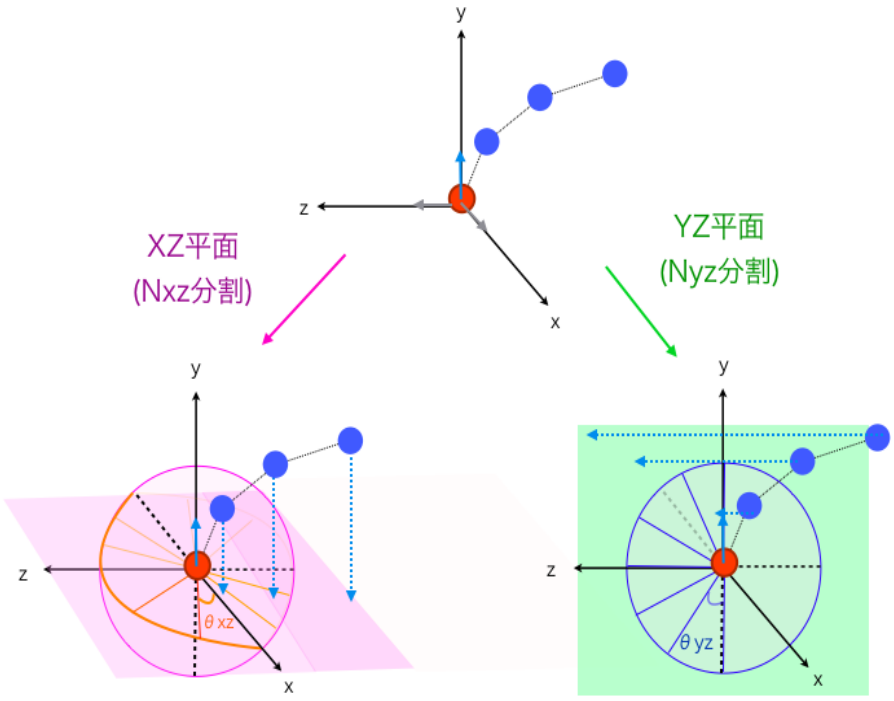


図 3.6: 階級決定のための空間分割とカウント方法

さらに各平面を角度によって分割することでヒストグラムの階級を決定する．ここで，本研究では図 3.7 のように，基準関節の座標系において，XZ 平面（0 度から 360 度）および YZ 平面（0 度から 180 度）のそれぞれにおいて空間を分割する．すなわち，XZ 平面における分割数を N_{xz}，YZ 平面における分割数を N_{yz}，XZ 平面における分割角度を θ_{xz}，YZ 平面における分割角度を θ_{yz} とすると，ヒストグラムの階級は，一つの階級を「(XZ 平面における領域) _ (XZ 平面における領域)」で表すことができる．階級を決定したのち，各部位内での基準以外の関節の位置情報を，二つの平面へそれぞれ射影し，その時の角度情報を該当する階級にカウントする．

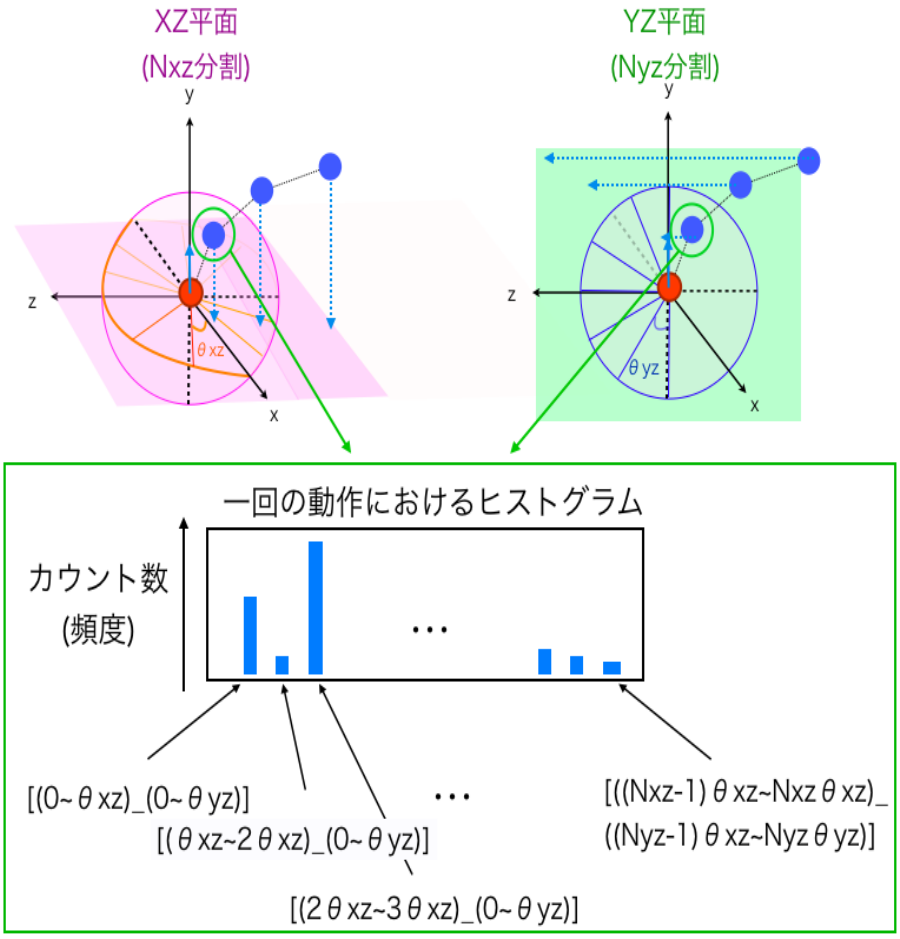


図 3.7: 一つの関節におけるヒストグラムの生成

(iii) 関節ごとのヒストグラムの生成

また，XZ 平面および YZ 平面に分割された領域に，ヒストグラムを生成する関節の三次元位置を，時系列的にカウントしていく方法を以下に示す．(i) まず，基準関節の空間座標系内における，ヒストグラムを生成する関節位置の三次元ベクトルを算出する．(ii) 次に，その三次元ベクトルを座標系における垂直面および水平面に射影し，それぞれの面における二次元ベクトルを算出する．(iii) そして，それぞれの面において，x 軸および y 軸との角度を算出し，該当する領域へカウントし，さらに総カウント数によって正規化する．

(iv) 部位ごとのヒストグラムの生成

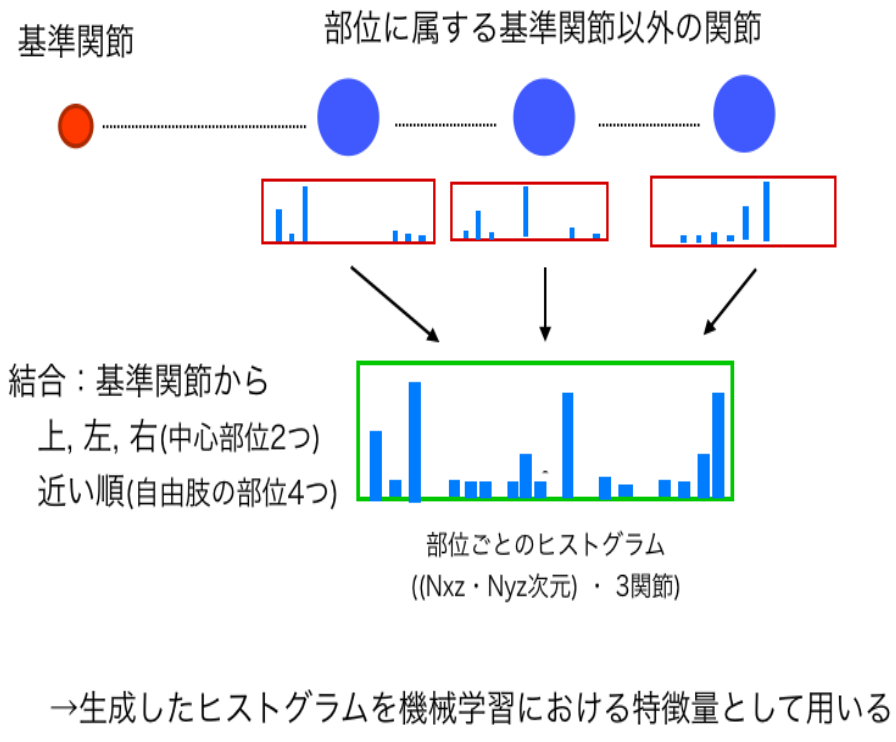


図 3.8: 関節ヒストグラムの結合による特徴量の生成について

最後に，基準関節以外の全関節におけるヒストグラムを機械学習の特徴量として用いることで，動作を認識することができる．この時，一つの関節あたり

のヒストグラムの次元数は，N_{xz} と N_{xz} の積で表すことができる．関節ごとのヒストグラムの結合について，図 3.8 に示す．本研究においては，6 つの部位にそれぞれ 3 つずつ基準関節以外の関節があるため，部位ごとに関節を結合した場合は $3N_{xz}N_{xz}$ となる．結合する順は，体幹にある部位（上半身中心部位，下半身中心部位）は，基準関節から上，左，右に属する関節の順にヒストグラムが結合される．また，可動性の高い自由肢にある部位（左腕部位，右腕部位，左脚部位，右脚部位）においては，基準関節から接続されている関節が近い順にヒストグラムが結合される．

3.1.3　各部位における分類器の生成

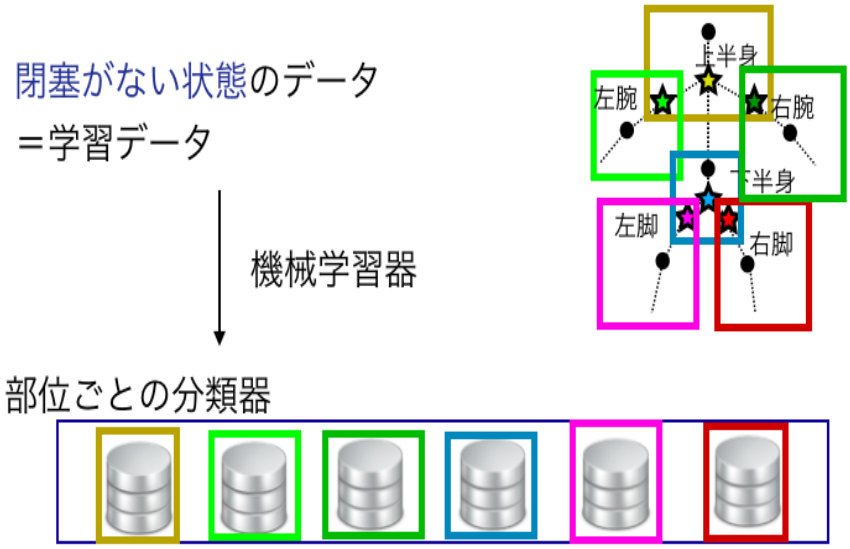


図 3.9: 各部位の分類器の生成について

各部位における特徴量の生成がされたことで，各部位それぞれにおいて分類器を生成することができるようになる．各部位の分類器の生成について，図 3.9 に示す．まずあらかじめ，動作予測を行う前に，事前に動作の特徴と動作ラベルの組み合わせを学習しておく必要がある．本研究においては，この学習を部位ごとに行い，各部位の分類器を生成する．ここで，学習データとしては，閉塞のない環境での動作データを取得しておき，Random Forest によって学習し，部位ごとの分類器を生成する．

Random Forest とは

動作データは複数のラベルのデータが存在するため，多クラス分類が自然に行うことができる Random Forest[1] を利用する．Random Forests はその名が示しているとおり，複数の木（tree）を用いて森（forest）を構成して識別などを行う機械学習アルゴリズムである．ここでいう木は決定木（decision tree）のことであり，個々の決定木は高い識別性能をもつわけではないが，それらを複数用い

てそれぞれの結果を補うことによって高い予測性能を得ることが一つの特徴である．これは機械学習の分野ではアンサンブル学習（ensemble learning）と呼ばれており，個々の決定木がアンサンブル学習における弱識別器（weak classifier）に相当している．Random Forest は多クラス識別に適しているだけでなく，予測の際にラベルごとの予測確率として確率推算値を求めることができる．確率推算値は，Random Forest による分類器において，まず各木の予測確率を各葉ノードで出力された予測ラベルのその木における割合として算出しておく．そして，各木で算出された平均予測確率を確率推算値として求める．本研究においては，まず Random Forest を用いて部位ごとに分類器を生成しておく．この時，部位ごとの特徴は独立している．そして，動作判定時，各部位において出力された動作ごとの確率推算値を重みとして利用することで，さらにアンサンブル的な予測を行う．

3.1.4　部位アンサンブル予測による動作認識

設計方針の最後として，認識する動作データ，つまり未知の入力データに対する動作ラベル判定の方法について説明する．部位アンサンブル予測による動作認識のイメージを表した図を，以下の図 3.10 に示す．まず，未知の動作データが入力データとして与えられた際，その動作データから部位ごとの特徴量を生成する．部位ごとに動作データから返還された特徴量を，対応する部位の分類器に入力する．次に，分類器から動作ごとの確率推算値が出力される．そして，各部位の動作ごとの確率推算値を，部位間で総和をとることで，全部位における動作ごとの確率推算値が求められる．その全部位における動作ごとの確率推算値の中から，最も高い値のラベルを動作ラベルとして出力する．

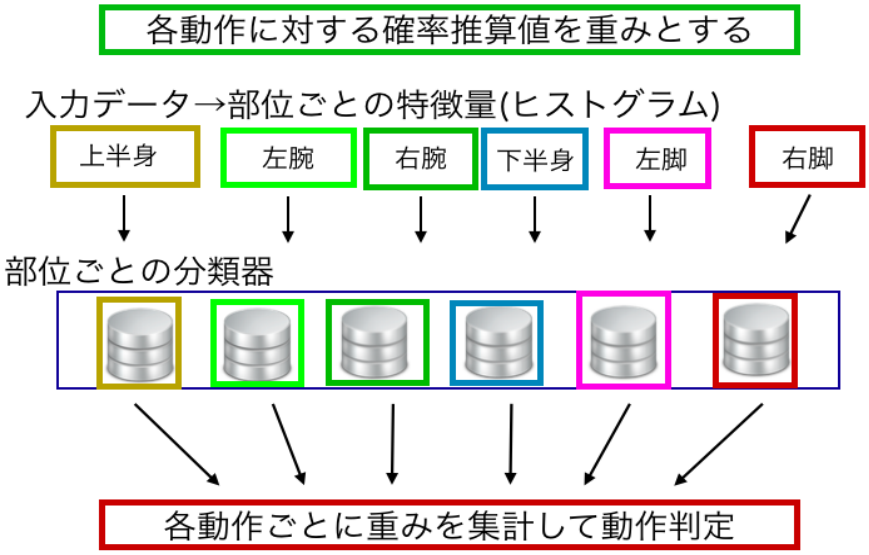


図 3.10: 部位アンサンブルによる予測について

3.2　実装

3.2.1　使用する機材について

まず本研究では，第一章でも述べた通り，深度カメラセンサとして Kinect v1 を用いる．Kinect for Windows SDK により，Kinect v1 から得られた深度情報から，骨格認識技術 [27] によって図 A.1 のような身体の 20 か所の関節の三次元位置を得ることができる．すなわち，骨格認識による 20 箇所の関節の三次元位置座標（$P_{O_n} = (P_{O_n}x, P_{O_n}y, P_{O_n}z)$, n = 0，1，…，20，$O_0$ は腰の中央）となる．さらに，同 SDK を用いることで，3.1 章で説明した方法によって，各関節の方向情報（$Q_{O_n} = (Q_{O_n}x, Q_{O_n}y, Q_{O_n}z, Q_{O_n}w)$, n = 0，1，…，20，$O_0$ は腰の中央）を取得することができる．また，全関節の三次元位置 P_{O_n} および方向情報 Q_{O_n} を 30[FPS](1 秒間に 30 回) の速さでを追跡することができる．しかし，遮蔽物が kinect と対象の間にあるピクセル領域においては閉塞が生じてしまい，深度情報を取得することができないといったデメリットもある．ここで，Taylor ら [29] によって，閉塞が生じている部分の骨格関節の位置情報を推測する手法が提案されており，同 SDK では骨格の関節位置の推測値として取得す

ることができる．

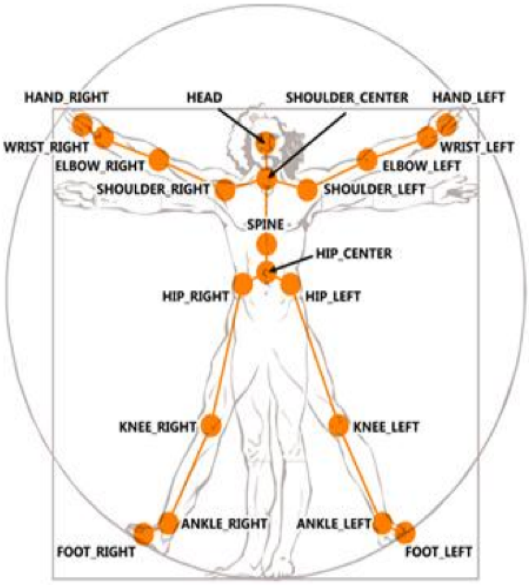


図 3.11: 割り振られている各部位の名称

3.2.2　得られる情報について

具体的に，取得するデータについて記述しておく．実験の際，動作データとして，(i) 動作を視覚的に記録しておくための動画データ（avi 形式，非圧縮），(ii) 日時，撮影開始からのフレーム数などの基本情報と，骨格認識による 20 箇所の関節の三次元位置座標を記録したデータ（csv 形式），(iii) 関節の方向情報と取得した時のフレーム数を記録したデータ（csv 形式）の，動作 1 回につき 3 種類のデータを取得しておく．動作データを作成する際，動画データから，目視によって任意に動作の開始と終了を判断する．そして動作の時間的範囲において，関節位置データと関節の傾きデータから動作にあたる箇所を抜粋し，これを 1 つの動作データとする．

動作を視覚的に記録しておくための動画データについて

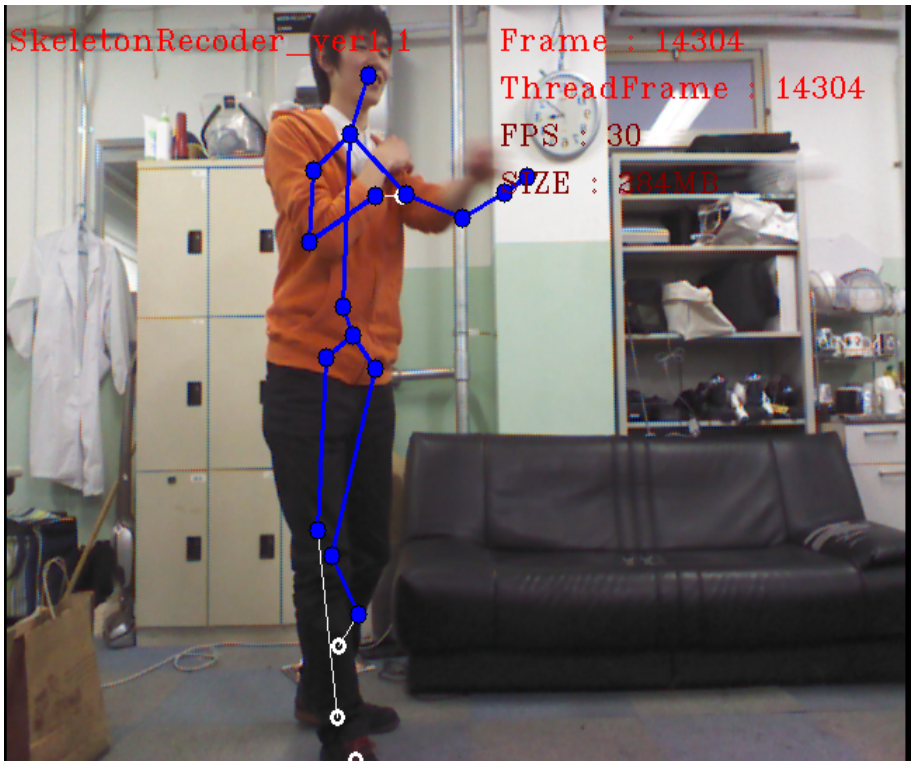


図 3.12: 動作を視覚的に記録しておくための動画データについて

動作を視覚的に記録しておくための動画データに関して，ある時間におけるキャプチャ画像を図 3.12 に示しておく．図 3.12 に示す通り，この動画データでは，各骨格の推定された三次元位置情報を画面上に射影した様子を含めて描画している．具体的には，関節を丸点，および丸点を繋ぐ線で構成された棒人間のような図が，撮影された RGB カメラ映像に加えて描画するようになっている．撮影を開始してからどれだけ何ミリ秒経過したかを示すフレーム数，1 秒間に何回骨格データを取得しているかを示す FPS（Frames Per Second）値，最後に動画データのデータサイズ (単位はメガバイト数) を記録している．動作開始と動作終了時のフレーム数を抜き出し，そのフレーム間における関節位置と関節の向きのデータを抜粋する．なお，図 3.12 における青い点と白い点はそれぞれ，認識できた関節位置と推測によって算出した関節位置を表す．

関節の三次元位置座標を記録したデータについて

関節の三次元位置座標を記録したデータは csv 形式で保存される．取得するデータの属性と内容および例を以下の表 3.1 に示す．

表 3.1: 関節位置情報を格納するデータの内容

データ名	説明	例
ファイル名	csv ファイルの ファイル名	20131217T150848
日付	取得日時	20131217
関節名	三次元位置情報を取得した 関節名	hipcenter
KinectAngle	Kinect デバイスの チルトモーターの傾き (垂直方向)	4
ミリ秒	取得した時の開始から 経過した時間 (ミリ秒)	5102
TrackingID	複数人数画面内に入ったときに 識別するため割り振られる ID	254
TrackingState	認識が実測値か推測値かどうか	tracked(実測値)
IsNearmode	Kinect が近距離測定モードかどうか	Default(通常時)
x	取得した三次元位置情報の x 座標	0.2045
y	取得した三次元位置情報の y 座標	-0.361
z	取得した三次元位置情報の z 座標	2.3439

先述した動画データから動作の開始されたフレーム数と終了したフレーム数を判断し，該当部分を抜粋することで動作データにおける骨格全体の関節データを作成していく．例えば,「投げる」というデータが 3000[frame] 目から 6000[frame] 目であった場合，その間のすべての属性の情報を抜粋し,「投げる」という動作データにおける関節位置の推移を示したデータが 1 つ作成されることになる．ここで，ファイル名や日時は便宜的に，それぞれ年月日時分秒および年月日の順に桁が並べられた数字データとしておく．また，Kinect v1 には垂直に振ることができるチルトモーターが搭載されており，チルトモーターの角度も取得しておく．

関節の向き情報を記録したデータについて

関節の向き情報と取得した時のフレーム数を記録したデータは，関節の三次元位置座標を記録したデータと同様に csv 形式で保存される．取得するデータの属性と内容および例を以下の表 3.2 に示す．

表 3.2: 関節の向き情報を格納するデータの内容

データ名	説明	例
ファイル名	csv ファイルの ファイル名	20131217T150848
日付	取得日時	20131217
ミリ秒	取得した時の開始から 経過した時間 (ミリ秒)	5102
TrackingID	複数人数画面内に入ったときに 識別するため割り振られる ID	254
KinectAngle	Kinect デバイスの チルトモーターの傾き (垂直方向)	4
親関節	向きを取得した 関節名	hipcenter
子関節	三次元位置情報を取得した 関節名	spine
Qx	取得した空間座標軸の四元数の x 成分 (クォータニオンの x 成分)	-0.04895
...	...	...
M11	取得した空間座標軸の回転行列 (4 行 4 列ある回転行列の 1 行 1 列目成分)	-0.99521
...	...	...

関節の位置情報と同様，先述した動画データから動作の開始されたフレーム数と終了したフレーム数を判断し，該当部分を抜粋することで動作データにおける関節の向き情報データを作成していく．親関節と子関節については，第三章において先述した通り，親関節位置から子関節位置に伸びるベクトルを y 軸とし，腰の中心部を基点としてその子関節まで回転していった時の向きが，取得した関節の向きとなる．ここで，取得した回転情報はすべて，センサーデバイスに対する回転情報であること (Kinect for windoes SDK では Absolute orientation と表現されるもの) であることに留意されたい．またこのとき，関節の向き情報は回転情報としてクォータニオン（四元数）あるいは回転行列として保存される．

骨格の関節間の階層構造と部位分割について

また，骨格における関節間の繋がりから，図 3.13 下部のような階層構造によって骨格の関節同士の関係性を表現することができる．腰の中心部（HipCenter）を基点として，左手（HandLeft），右手（HandRight），左足（FootLeft），右足（FootRight）そして頭部（Head）の 5 つの末端へと，親となる関節から子となる関節へ骨格の階層構造が設定されている．本研究では図 3.13 上部のように，骨格を 6 つの部位に分割した上で，動作の際に支点となる関節 6 つを基準関節として設定する．

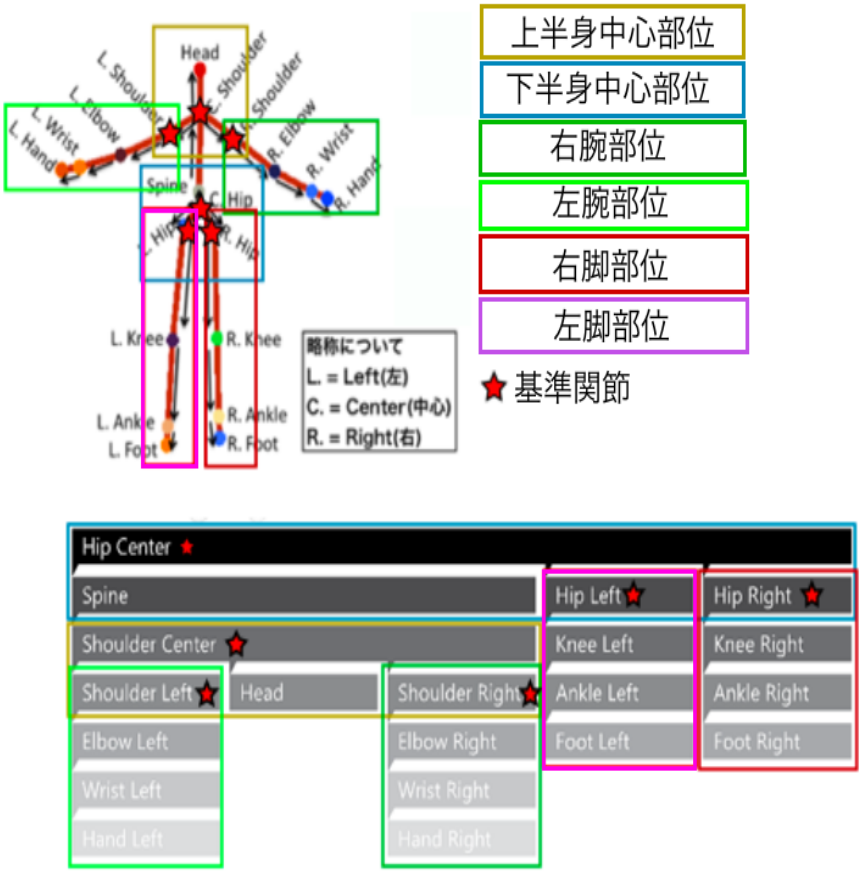


図 3.13: 関節間の階層構造について

3.2.3　ヒストグラム作成のための軸決定と射影

ここで，基準関節における空間の座標軸は，腰の中心部からの回転によって計算される．本研究では先述した通り，すでに Kinect for windows SDK によって算出された，深度カメラセンサの座標軸からの回転行列を取得している．クォータニオンおよび回転行列の両方を取得してはいるが，本研究では関節の向きに回転行列，ヒストグラムを作成するための射影などにおいてはクォータニオンを使用している．

基準関節における空間の座標軸の決定

実験によって取得した回転行列から，各部位における基準関節における向きを決定する．この時のみ，深度カメラセンサの座標系から三軸周りにそれぞれ一回ずつ回転させられればよいので，クォータニオンではなく 4 行 4 列の回転行列を利用した．

各関節位置の各平面への射影について

まず基準関節の座標系内における，ヒストグラムを生成する関節の位置ベクトル $\overrightarrow{P'_{O_n}}$ を求める．実世界の座標系（深度カメラセンサを基準とする座標系）における，ヒストグラムを生成する関節位置座標を P_{O_n}，基準関節の位置座標 S_{O_m} として次のように算出する．ただし O_m は，関節 O_n が属する部位においての基準関節とする．

$$\overrightarrow{P'_{O_n}} = \overrightarrow{P_{O_n} - S_{O_m}} \tag{3.1}$$

ここで，記録した回転行列とは逆の向きへの回転行列，すなわち基準関節における座標系から腰の中心部の座標系へ回転する回転行列 W’ を求めておく．W’ は，W を求める際に行った演算を逆に行うことで求めることができる．腰の中心部は基準となる座標系であり，その座標系の平面 $y = 0$ は地面に対し水平，Y 軸は地面に対し垂直となる．すなわち，基準関節における座標系内の関節ベクトル $\overrightarrow{P'_{O_n}}$ を，この座標系における位置ベクトル $\overrightarrow{V_{O_n}}$ へとさらに回転させる．

そして，その関節位置ベクトル $\overrightarrow{V_{O_n}}$ の xz 成分 $\overrightarrow{V_{O_n}xz}$，および yz 成分 $\overrightarrow{V_{O_n}yz}$ が，それぞれ基準関節における座標系の XZ 平面，YZ 平面に射影した座標と同値となる．

したがって，まず $\overrightarrow{V_{O_n}}$ を求める．基準関節における座標系内の関節ベクトル $\overrightarrow{V_{O_n}}$ の算出は，$\overrightarrow{P''_{O_n}} = (P'_{O_n}x, P'_{O_n}y, P'_{O_n}z, 1)$ と変換した上で，回転行列 W' による三軸それぞれにおける二次元平面における回転を施せばよい．この時，W によって基準関節の座標系を算出した時とは逆の順に，固定する軸を変更しなくてはならないことに留意されたい．$\overrightarrow{V_{O_n}}$ から，基準関節の座標系の XZ 平面に射影したベクトル $\overrightarrow{V_{O_n}xz}$，および YZ 平面に射影したベクトル $\overrightarrow{V_{O_n}yz}$ は，次のように求めることができる．ここで，T は行列の転置を表す．

$$\begin{cases} \overrightarrow{V_{O_n}xz} = \{V_{O_n}x, 0, V_{O_n}z\}^T \\ \overrightarrow{V_{O_n}yz} = \{0, V_{O_n}y, V_{O_n}z\}^T \end{cases} \tag{3.2}$$

射影したベクトルの角度の算出

最後に，基準関節における座標系の x 軸を表すベクトル $\vec{W}x$ と $\overrightarrow{V_{O_n}xz}$ の角度 θ_{xz}，y 軸を表すベクトル $\vec{W}y$ と $\overrightarrow{V_{O_n}yz}$ の角度 θ_{yz} をそれぞれ算出し，該当する階級にカウントしていけばよい．また，投げる（利き腕は右手）の動作において生成された右手および左手のヒストグラムを図 3.14 に示す．ここで，各階級の名前は簡易に表示するため，((XY 平面の角度領域の始点となる角度）_（YZ 平面の角度領域の始点となる角度)）としている．図 3.14 から，投げるという動作において，右手のヒストグラムは左手のヒストグラムに比べ，平らな形状となっている．この理由としては，動きの大きい右手は階級に対してカウントが分散され，動きの少ない左手は狭い範囲の階級にのみカウントされることが挙げられる．そして最終的に，ヒストグラムは動作時間中の全カウント数によって正規化される．

ここで，本研究における空間分割は，図 3.7 のように，基準関節の座標系において，XZ 平面（0 度から 360 度）および YZ 平面（0 度から 180 度）に空間を分割する．その時，本研究では分割する角度を 30 度に定める．すなわち，XZ

平面において N_{xz}=12，YZ 平面は N_{yz}=6 かつ $z > 0$ の範囲を満たす領域を用いる．ここで，YZ 平面に射影する際の z 軸方向の値が負だった場合，z 軸方向に-1 を乗算して $z > 0$ を満たすようにする．そして，XZ 平面における階級と YZ 平面における階級の組み合わせが，ある関節の動作一回に対するヒストグラムの階級となる．分割された領域の組み合わせは，各部位内の，基準関節を除く各関節において，生成されるヒストグラムの階級となる．

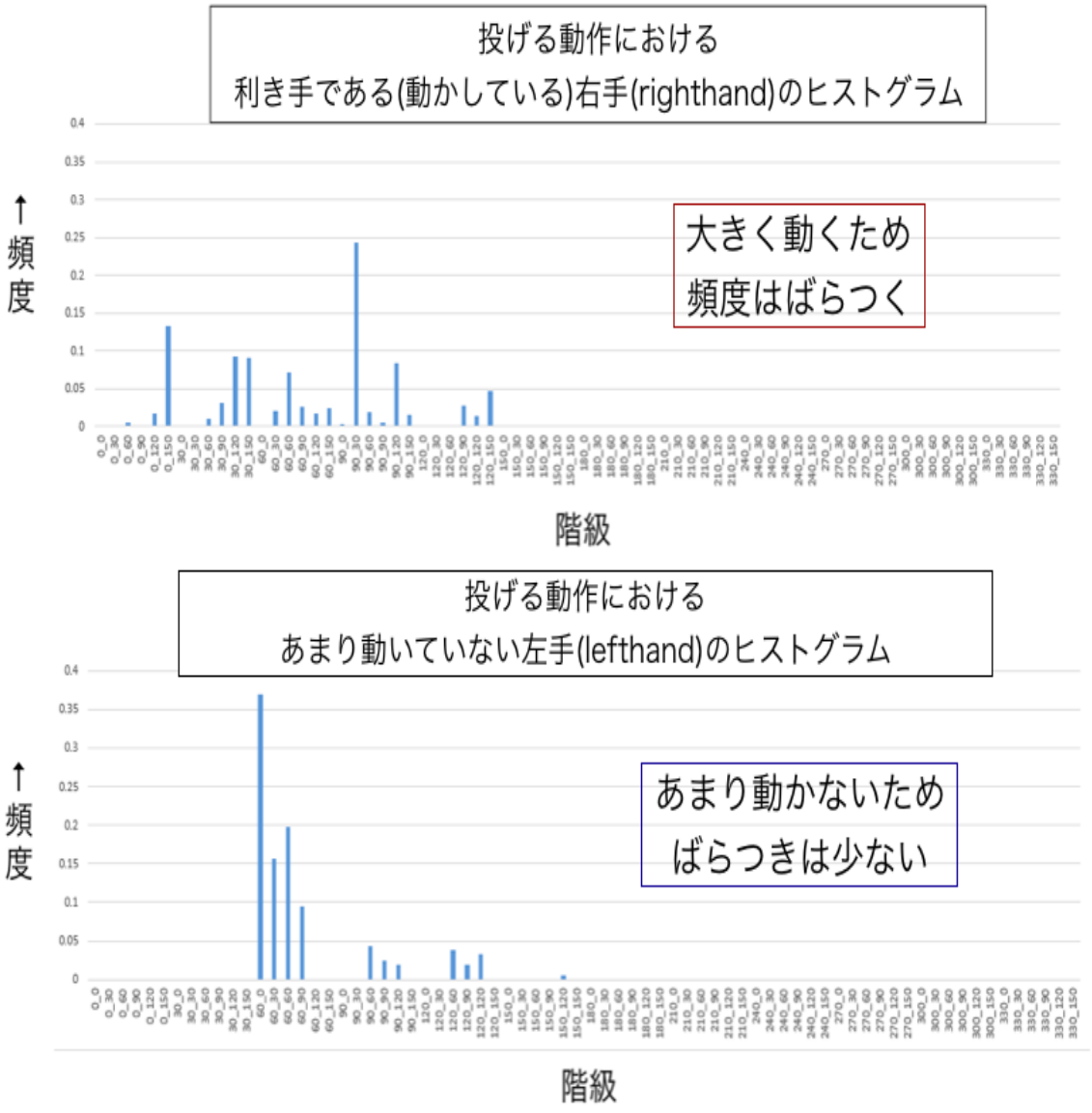


図 3.14: ”投げる”動作における，右手および左手のヒストグラム

3.2.4　仮想的な三次元空間への配置

ここで，各関節の位置や，基準関節の方向情報が視覚的に正しいかどうかを確認するため，本研究では仮想的な三次元空間に適した開発環境 Unity5.3.5[5] を用いて，各関節の位置や基準関節の方向情報，および XZ 平面（紫色）と YZ 平面（緑色）を表現し，さらにヒストグラムの生成も行った．深度カメラセンサにより取得した骨格の三次元位置情報を仮想的な三次元空間上に配置し，さらに基準関節の方向を表した結果を図 3.15 に示す．ここで，空間上の軸はそれぞれ X 軸が赤色，Y 軸が緑色，Z 軸が青色で示されている． 関節位置は緑色の球で示され，その中でも基準関節は紫色の球で示されている．さらに基準関節には，関節の方向を表すオブジェクトが付加されている．関節の向きを表すオブジェクトには，射影される平面である XZ 平面（紫色）と YZ 平面（緑色）が描写されている．

[5]Unity. https://unity3d.com/jp

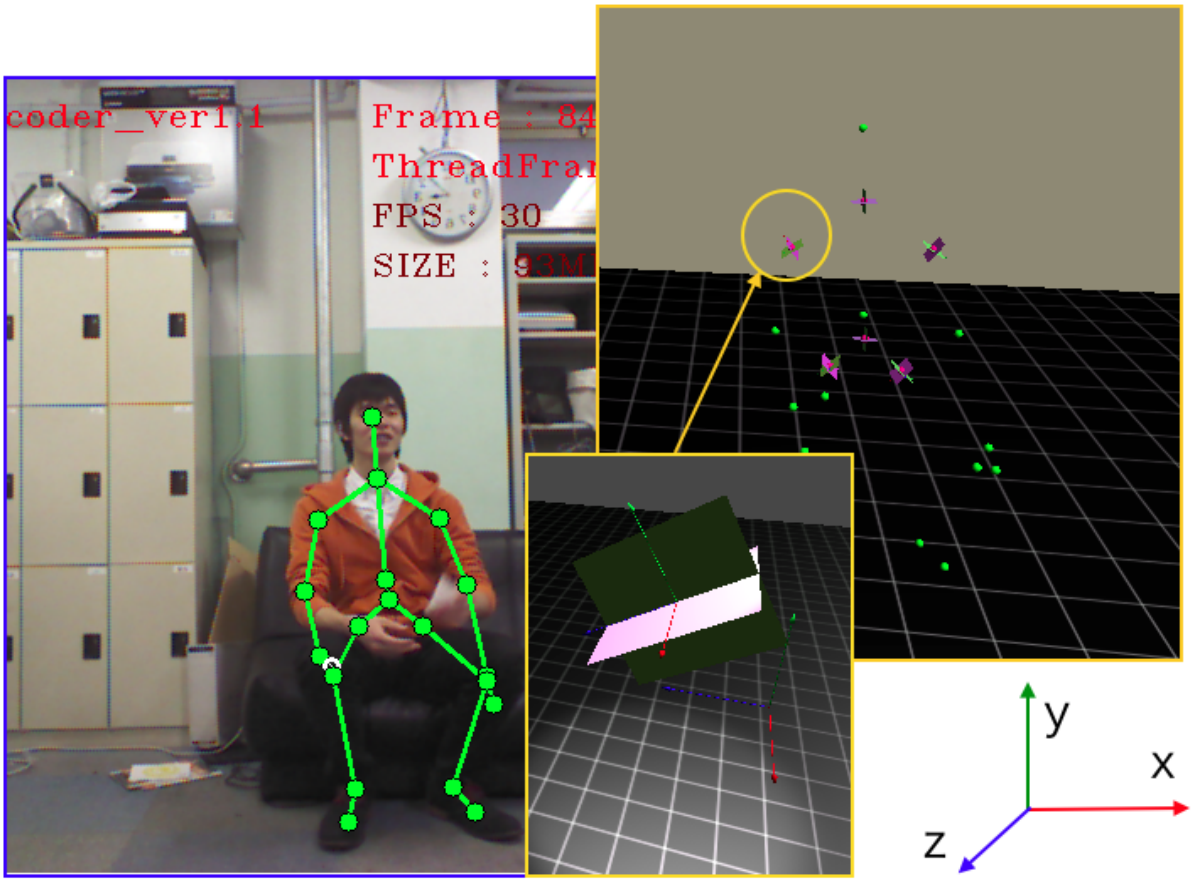


図 3.15: 関節位置と向きの空間上での表現

例えば左腕部位において，左肘関節のヒストグラムを生成するとき，XZ 平面（紫色）および YZ 平面（緑色）への射影が正確に計算されているかどうかを確認するために用いることができる．図 3.16 にその確認のために，基準関節の座標軸を拡大し，正確に射影が行われているかどうかを確認したときのキャプチャ画像を載せておく．図 3.16 において，左肘（赤い立方体）と，基準関節となる左肩のオブジェクトはそれぞれ拡大されている．また，左肘関節の XZ 平面（紫色）および YZ 平面（緑色）への射影はそれぞれ，平面上の立方体（XZ 平面へは紫色，YZ 平面へは黄色）として表している．

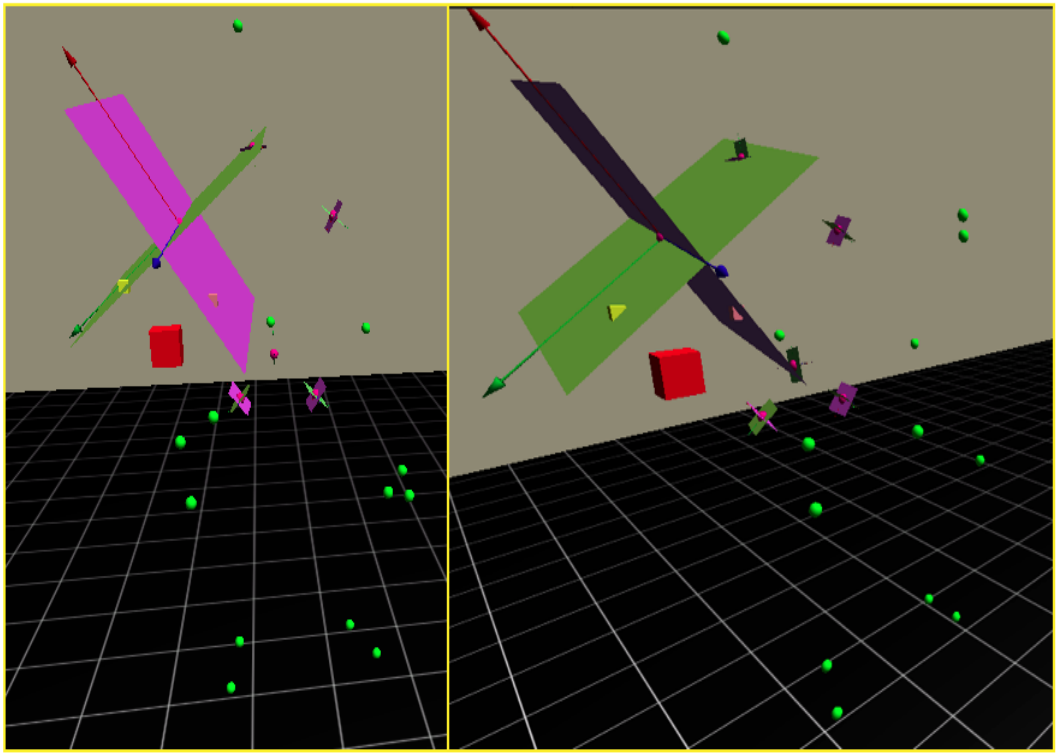


図 3.16: 基準関節と，左肘の XZ 平面（紫色）および YZ 平面（緑色）への射影について (左右とも同じ動作かつ同じ瞬間のデータを，それぞれ別方向から捉えた図.)

3.2.5　各部位における分類器の生成

設計方針で述べた通り，動作を識別する前に，あらかじめ部位ごとの分類器を Random Forest[1] によって行う．本研究では，python2.7[6]，機械学習ライブラリとして scikit-learn 0.16.1[7] を使用した．また，学習の際のパラメータとして，グリッドサーチによる最適なパラメータの探索を行った．以下の表に，パラメータ探索に使用した値の候補を示す．このパラメータは，本研究において作成した閉塞がないデータに対して比較対象として挙げている既存研究 [33] のヒストグラム生成手法を適用し，Random Forest を用いて 10 分割交差検証を行った際，最も良かった時の組み合わせを用いている．

表 3.3: グリッドサーチにおけるパラメータの候補 (身体動作認識)

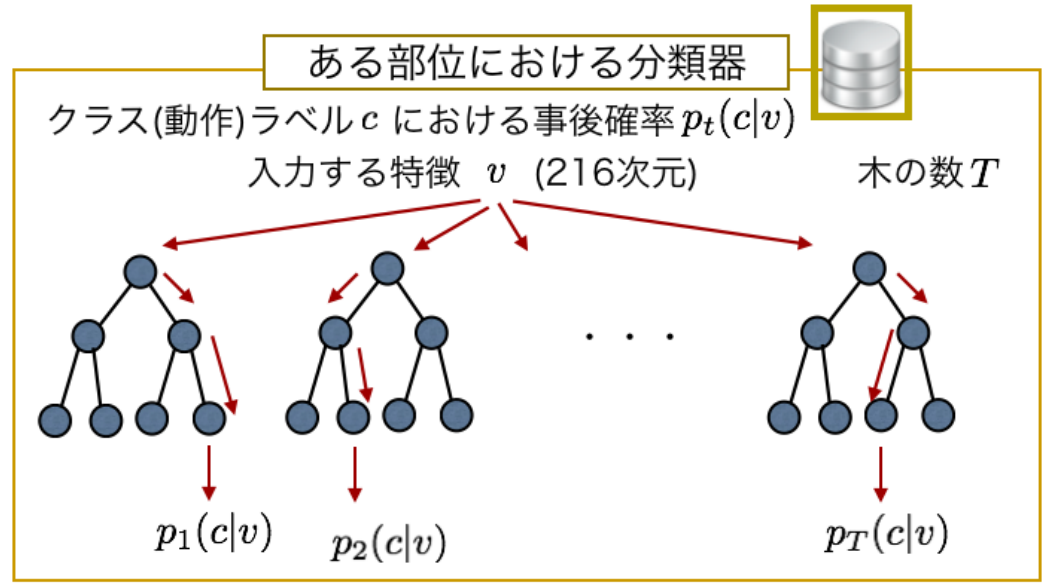
パラメータの種類	候補パラメータ
バギングに用いる決定木の個数	[5, 10, 100]
葉を構成するのに必要な最小限のサンプルの数	[3, 5]
最適な分割をするために考慮する特徴の数	[3, 5]
決定木の深さの最大値	[5, 10, 25, 100]

3.2.6　ある部位における動作ごとの確率の出力

ここで，入力データとして未知の動作データを判別する際の各部位における確率推算値について説明する．また，そのときの概念図を図 3.17 に示す．一つの部位に着目したとき，その部位に入力される特徴（216 次元）を v とする．分類器に v を入力した際の，決定木 t における動作ラベル c の事後確率は $p_t(c|v)$ と表すことができる．したがって，分類器における決定木の数を T とすると，クラスラベル c における確率推算値 $p(c|v)$ は以下のように各木における事後確率の相加平均として表すことができる．

$$p(c|v) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|v) \tag{3.3}$$

[6]Python. https://www.python.org/
[7]scikit-learn, http://scikit-learn.org/stable/



クラスラベル c における確率

$$p(c|v) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|v)$$

図 3.17: ある部位における動作ごとの確率の出力について

3.2.7　確率推算値の算出から動作ラベル推定

次に，各部位に動作データを入力し，各部位の各動作ラベルにおける確率推算値を重みとして，全部位において集約する方法について説明する．全部位を考慮すると，部位 r_n（n = 1, 2, 3, 4, 5, 6）における動作ラベルごとの確率推算値の集合 $P_C(r_n)$ は，動作ラベルを C ($C = \{c_1, c_2, ..., c_N\}$, 本研究では $N = 10$) として以下のように表すことができる．

$$P_C(r_n) = \{p_{r_n}(c_1|v_n), p_{r_n}(c_2|v_n), ..., p_{r_n}(c_N|v_n)\} \tag{3.4}$$

ここで，部位の集合を R とすると，本研究においては以下のように順序を設定した．また，下半身中央部位（lowercenter）, 上半身中央部位（uppercenter）, 左腕部位（leftarm）, 右腕部位（rightarm）, 左脚部位（leftleg）, 右脚部位（rightleg）である．

$$\begin{aligned} R &= \{r_1, r_2, r_3, r_4, r_5, r_6\} \\ &= \{lowercenter, uppercenter, leftarm, rightarm, leftleg, rightleg\} \end{aligned} \tag{3.5}$$

そして，動作ラベルごとの確率推算値を重みとして，全部位における動作ラベルごとの確率推算値の集合 $\sum P_C$ を導出する．各部位における動作ラベルごとの確率推算値を，動作ラベルごとに全部位での確率推算値の総和をとる．したがって，$\sum P_C$ は以下の式で表すことができる．

$$\sum P_C = \{\sum_{n=1}^{6} p_{r_n}(c_1|v_n), \sum_{n=1}^{6} p_{r_n}(c_2|v_n), ..., \sum_{n=1}^{6} p_{r_n}(c_N|v_n)\} \tag{3.6}$$

最後に，算出された全部位における動作ラベルごとの確率推算値の集合 $\sum P_C$ の中で最も値が高い動作ラベルを，最終的な予測結果として出力する．

3.3　実験

本研究の目的は，照明環境に依存しない深度情報を用いて，屋内実環境における人物の可視部分のみから動作認識を可能にする手法を提案することである．したがって，屋内実環境として，机という閉塞になりうる家具を配置した場合とそうで無い場合の両方の環境において，動作のデータセットをあらかじめ作成する．

実験の目的は，提案手法が閉塞がある環境において有効かどうかを示すことである．動作データセットにおいて，10 種類の分類（歩く (walk)，座る (sitdown)，立つ (standup)，拾う (pickup)，運ぶ (carry)，投げる (throw)，押す (push)，引く (pull)，波打つ (wave)，拍手する (chaphands)）の動作に対して精度検証を行う．この時，Xia ら [33] のヒストグラム生成手法を本データセットに適用し，部位アンサンブルによる手法と比較する．加えて，提案手法において部位アンサンブル予測を用いず，18 関節を単純に結合して動作 1 回における特徴とする手法についても比較する．また，学習データを閉塞がない環境でのデータ，テストデータを閉塞がある環境でのデータとして，3 つの手法の動作認識精度を比較する．

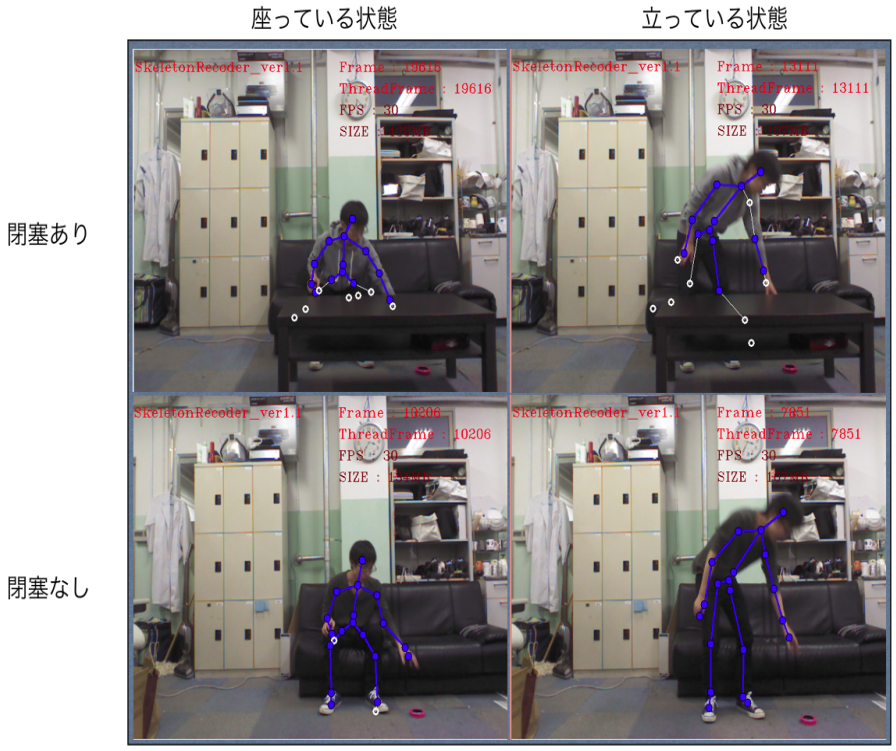


図 3.18: ある被験者における，閉塞の有無と姿勢の条件による”拾う”動作の様子

まず，収集した動作データについて説明する．被験者は 20 代の学生 8 人（うち一人は女性であり，また皆が右手を利き手としており，かつ身体的な障害を患っていない）を対象としている．動作の分類は，10 種類の分類（歩く (walk)，座る (sitdown)，立つ (standup)，拾う (pickup)，運ぶ (carry)，投げる (throw)，押す (push)，引く (pull)，波打つ (wave)，拍手する (chaphands)）の動作を取得する．

- (i) 移動を伴う動作（walk，carry）
- (ii) 主に腕を使用する動作（pickup，throw，wave，chaphands）
- (iii) 姿勢が変化する動作 (及びその他の動作)（sitdown，standup，(push，pull)）

そして，それぞれの動作において繰り返し回数を設定する．その内訳は，(i) 左右の移動方向それぞれに 3 回ずつ，(ii) 姿勢ごとに 5 回ずつ，(iii)5 回，と設

定した．さらに，机による閉塞があるかどうかを考慮した．すなわち，被験者 8 人，10 種類の動作とその分別における繰り返し回数（ (i)2 つの方向ごとに 3 回，(ii)2 つの姿勢ごとに 5 回，(iii)5 回），閉塞の有無の 2 種類の，合計 1,152 回の動作データを取得した．また動作 1 回分の時間的な長さの平均は 3,221[ms] となっている．収集したデータセットを用い，1,152 回の動作において，全 18 カ所の関節のヒストグラムを生成し，提案手法においては部位ごとの特徴量を生成する．また，Xia ら [33] の特徴となる関節は 9 関節, また 1 関節あたり 84 次元の特徴を生成する手法のため，動作 1 回あたり 756 次元の特徴となる．さらに，提案手法において，18 関節のヒストグラムを単純に結合した場合も含めて比較する．そして，3 つの手法について，実装で述べた項と同じパラメータ候補でグリッドサーチを行い，提案手法以外の手法においては Random Forest[33] による精度検証を行う．

3.4　実験結果

実験の結果を，以下の表 3.4 に示す．ここで，既存手法は Xia ら [33] の手法, 手法 1 は提案手法においてヒストグラム生成手法のみ採用し，部位アンサンブル予測を用いずに 18 関節のヒストグラムを結合した特徴を用いた手法となっている．そして手法 2 は提案手法において，提案通り部位アンサンブル予測を行う手法である．表 3.4 から，閉塞のある環境に対して，本提案手法が最も精度が高く（84.7%）なっている．したがって，本提案手法が閉塞がある環境において有効であると言える．

表 3.4: 閉塞ありのデータに対する手法ごとの認識精度

既存手法 [%]	手法 1 [%]	手法 2 (部位アンサンブル予測)[%]
83.2	81.8	84.7

3.5　考察

3.5.1　閉塞なしの環境における認識精度の比較

考察として，まず閉塞なしの環境における認識精度と閉塞ありの環境における認識精度の，既存手法 [33] と本提案手法の手法ごとの変化について着目する．まず，閉塞なしの環境のデータのみで 10 分割交差検証を行った結果，既存手法 [33] の手法においては 89.2%（実験結果 (閉塞あり) : 83.2%），及び本提案手法においては 85.0%（実験結果 (閉塞あり) : 84.7%）という結果となった．閉塞がない，すなわち全身が可視可能な状態である場合，既存手法 [33] の方が良い結果となった．これは，既存手法 [33] は腰を基準として全身の動きを動作特徴としていたのに対して，本提案手法は部位という局所的な部分に分割して特徴を捉えていることが原因として考えられる．つまり，局所的な特徴のみを考慮しているため，全体的な動作の特徴が捉えきれていないことが原因として考えられる．したがって，本提案手法において，局所的な特徴だけでなく，全身の動作特徴を考慮することで，精度が向上する可能性があると考えられる．しかしながら，本提案手法においては，閉塞ありの環境においても精度が低下しにくいという結果となり，さらに閉塞に対して有効であることを示している．ここで，動作における力学的な部位と部位の間の影響を考慮することが挙げられる．これは，動作を行う際，例えば投げるという動作は，手を挙げる前に，腰を回転させ腕を一旦後ろに下げる予備動作を行うことで，投げる動作を達成している．つまり，ある部位を動かした際の，付随して予備動作を行う部位が見えている前提ではあるが，予備動作を行う部位が力学的な影響を特徴として捉えることで，より部位間の動きの特徴を捉えることができるかもしれない．

3.5.2　各部位での予測における動作ごとの確率推算値の違い

また，部位アンサンブル予測の際，各部位での予測における動作ごとの確率推算値について着目する．ある一つの部位において，動作の特徴が現れるような場合は，その動作ラベルに対する確率推算値が高くなり他の動作ラベルに対しての確率推算値が低くなると考えられる．一方，動作の特徴が現れないよう

な場合，各動作ラベルにおける確率推算値は分散して平均化されると考えられる．表 3.5 に，右腕部位における，座る動作における確率推算値と投げる動作における確率推算値の例を示す．

表 3.5: 右腕部位における動作ごとの確率推算値の例 (枠内単位は%)

正解ラベル	予測ラベル (全部位)	歩く	座る	立つ	拾う	運ぶ	投げる
座る	拾う	5.3	14.8	18.5	20.3	15.8	4.5
投げる	投げる	4.5	1.8	1.8	9.0	2.5	62.3

正解ラベル	予測ラベル (全部位)	押す	引く	波打つ	拍手		
座る	拾う	6.1	4.5	5.8	4.3		
投げる	投げる	6.0	5.3	6.0	1.0		

例えば，投げる動作においての右腕部位は，入力されたデータに対して投げる動作の確率推算値が 62.3%と高くなっている．しかしながら，座る動作における右腕部位は，全動作ラベルに対する確率推算値が平均的になっている．したがって，このような違いを考慮した上で，閾値を用いるなどの処理によって，動作認識の精度を上げることができる可能性があると考えられる．

3.6　まとめ

第三章の目的は，身体動作において，照明環境に依存しない深度情報を用いて，屋内実環境における人物の可視部分のみから動作認識を可能にする手法を提案することである．本目的を達成するために，深度映像を用いた骨格認識技術 [27] によって算出した身体骨格の三次元位置情報を用いた．その上で，デバイスと認識対象者との間に起こる閉塞を考慮し，骨格関節位置群を上半身中心部位，下半身中心部位，右腕部位，左腕部位，右脚部位，左脚部位の 6 つの部位に分割した．ここで，各部位に一つ，動作の際に支点となる関節（基準関節）を定めておいた．そして，各部位において，基準関節における方向情報を用いて，部位ごとに空間の軸を定めることで，部位ごとに独立した特徴を生成した．さらに，部位アンサンブル予測による動作認識手法を提案した．部位アンサン

ブル予測とは，機械学習を用いて動作を認識する際，まず各部位において閉塞がない環境でのデータであらかじめ Random Forest[1] によって学習して分類器を生成しておき．動作認識時に各部位の分類器の入力データに対する確率推算値を重みとしてアンサンブル予測を行う手法のことである．また屋内実環境で考えられる身体の一部位が隠れてしまうほどの閉塞に対して，本提案手法による動作認識が有効であるかどうかを確認するための実験を行った．まず閉塞としての机がある状態とない状態の両方を考慮したデータセットを作成した．そして既存手法として Xia ら [33] の手法を本研究のデータセットに適用し，提案手法と比較した．その際，閉塞なしの環境における動作データを学習データ，閉塞ありの環境における動作データをテストデータとした．その結果，既存手法 [33] が 83.2%であったのに対し，本提案手法が 84.7%という結果となり，本提案手法の閉塞のある環境に対する有効性を示すことができた．今後の展望としては，部位という局所的な特徴だけでなく，腰や体幹を中心とした全身の動作を特徴として加えることが考えられる．また，部位と部位の間の，動作における力学的な影響を特徴として捉えることで，動作認識の精度が向上するかもしれない．加えて，より細かい動作を認識するために，身体の一部として顔の動作に着目することが考えられる．したがって，次章では，顔の動作に着目する．

第4章　顔動作認識手法

本章では，身体における一部分としての顔の動作に着目する．顔には可動性の高い関節として顎関節が，また目や口などの，関節ではないが動作する特徴的な部分が存在する．加えて，顔の動作は話すなどの日常的な動作や，食べる動作，その中でも咀嚼動作などの健康と繋がりのある動作が存在する．したがって，第四章では，顔において，さらに細かい可動的な部分として口の動きに着目し，動作認識を行う手法を提案する．身体における動作認識と同様に深度カメラセンサを用い，得られた深度情報から，図のような顔の顎関節や口の特徴的な位置（表出情報）を取得する．表出情報から鼻と顎の距離, 口の縦方向および横方向の距離情報を算出する．その上で，それぞれの距離の時系列情報を加工し，特徴量として用いる手法を提案する．そして，得られた特徴量から機械学習を用いて顔の動作を認識する．

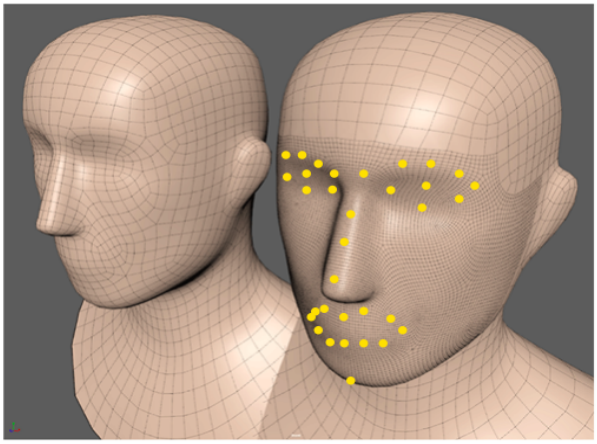


図 4.1: 顔の三次元的な表出情報

4.1　設計方針

まず顔における動作認識において，設計方針となるシステムの流れを以下に示す．

1. 初めに，深度情報による顔の表出情報取得手法によって顔の特徴点群の三次元的位置を取得する．ここで，例えば RealSense などのように，図の各点のように顔の表出情報における三次元位置を取得することができる技術が存在する．本研究においては，口について着目するため，顔の表出情報のうち口およびその周りにある鼻先の三次元位置情報の時系列推移を用いる．

2. 次に，鼻先と顎先の間，口の横幅，口の水平中央の縦幅，口の水平右側の縦幅，口の水平左側の縦幅の，5 つそれぞれの三次元距離情報を L_c，L_w，L_m，L_r，L_l として算出する．これは，顔における大きな動きとなる顎関節の動きを示す距離として L_c を選択した．また口の動きは顎関節と違い，縦横様々に動くため，横に一つ，縦に三つの距離を用いる．ここで，5 つそれぞれの距離情報の動作中の時系列推移は $L_c(\mathrm{t})$，$L_w(\mathrm{t})$，$L_m(\mathrm{t})$，$L_r(\mathrm{t})$，$L_l(\mathrm{t})$ として表すことができる．この 5 つの距離情報の時系列推移を用いることで，閉口咀嚼のような顎関節は動くが口は動かない場合や，話す時の左右の口の開きの差異などを特徴に含めることができる．

   また，そもそも顎関節や口の動きのような人間の動作を特徴とする場合，特定の周波数で動いているというわけではないため，約 4 秒間ほどの動作において，既存研究 [4] にあるようなフーリエ変換による周波数推定は適していない．したがって繰り返し動作を含むような場合，ノイズフィルタリングを行ったのちに，繰り返し性や時間方向に凸となっている部分を探す必要がある．そこで本研究では，5 つそれぞれの三次元距離情報の時系列推移に対し，心拍の認識 [10] にも用いられている矩形波窓を適用することによってノイズを除去する．ノイズ除去を行うことで，三次元距離情報の時系列推移における凸の部分をより詳細に取得することができる．

3. 矩形波窓 [10] を適用することによるノイズ除去を行ったのちに，自己相関関数を適用することによって，繰り返し動作などによって見られる信号の定常性を考慮し，その自己相関関数における初出ピークと初出インデックスを特徴量として用いる．さらに，動作中のデータ全てにおいて矩形波窓と自己相関関数窓を移動させた場合の，自己相関関数における初出ピークと初出インデックス，ピークが出現する回数，全てのピーク時における三次元距離の平均と分散を特徴に加える．これにより，信号の定常性に関する情報を特徴に加えることができる．また，動作中における三次元距離情報の時系列推移の基本的な統計量（平均値，分散値，最大値，最小値）を特徴として，機械学習にかけることで顔の動作を認識する．

4.1.1　顔の表出情報の取得

まず顔の表出情報を取得する．ここで，例えばRealSense[8] などのように，図の各点のように顔の表出情報における三次元位置を取得することができる技術が存在する．また他にも，深度画像によって顔の表出情報を推定する研究 [35] は行われてきている．そして顔の動作を識別するために，図 4.2 に，顔動作で用いる 5 つの距離について示す．第一に，顎関節の動きを捉えるため，鼻先と顎先の間 L_m を取得する．次に，口における動きを捉えるため，口の横幅 L_w，口の水平中央の縦幅 L_c，口の水平右側の縦幅 L_r，口の水平左側の縦幅を L_l として算出する．口の動きは顎関節と違い，縦横様々に動くため，横に一つ，縦に三つの距離を用いる．そして，各距離における時系列情報を取得する．

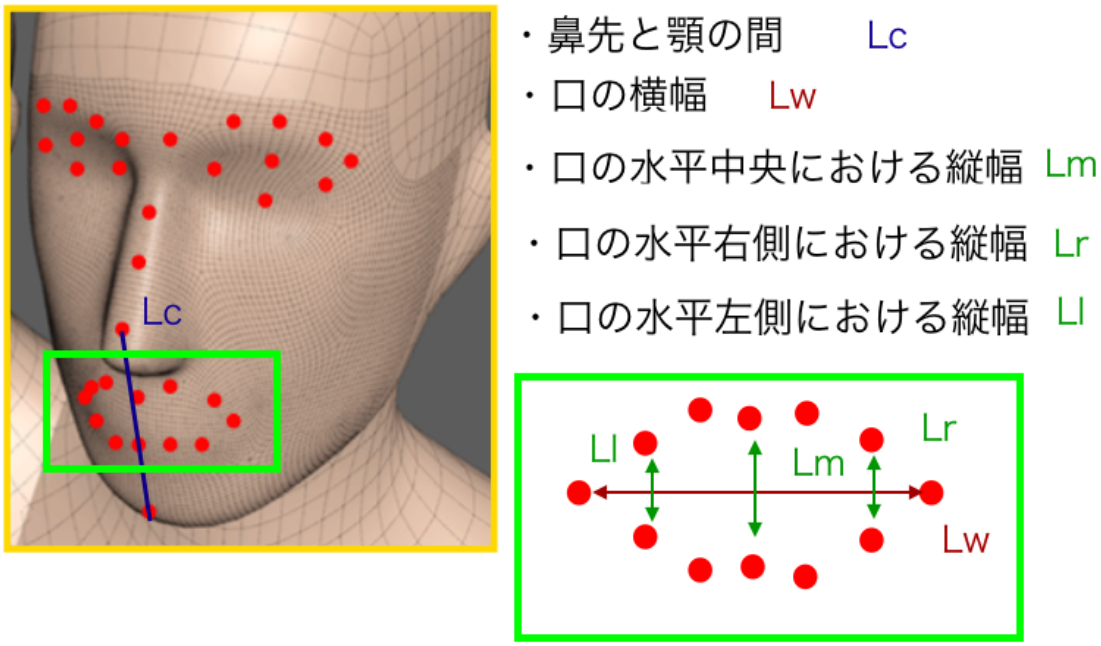


図 4.2: 顔の動作認識のための 5 つの距離

[8]Intel realsense device and sdk. http: //www.intel.com/content/www/us/en/architecture-and-technology/ realsense-overview.html

4.1.2　矩形波窓によるピーク検出のためのノイズ除去

動作中の顔の表出情報において推定された三次元位置情報にはノイズが含まれる．これは，データ取得自体の手法や，深度カメラセンサが情報を取得するタイミングなどによって発生してしまう．さらに，身体全体における関節の位置と比べ，顔の表出情報は距離情報の推定にノイズの影響を受けやすい．また，ピークを検出する際には，その凹凸部分を推定できれば良い．そこで本研究では，角ら [10] による矩形波窓を用いることで，ピーク検出のためのノイズ除去を行う．矩形波窓を利用することにより，値の上昇下降の変化のタイミングをより強く目立たせることができるようになり，ピークを検出しやすくなる．

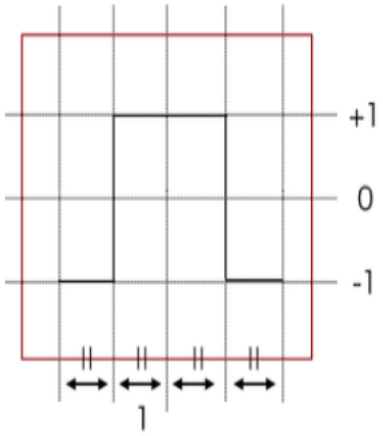


図 4.3: 矩形波窓 n について

4.1.3　自己相関関数による定常性判定

顔の動作には，しばしば繰り返しの動きが見られることがある．例えば，咀嚼動作における一回一回の噛む際の，鼻先と顎先の位置の距離 L_t がそれにあたる．距離 L_t のピークとなる点を認識する際，次に，フィルタリング後のデータに対して自己相関関数 (Auto-correlation Function, 以下 ACF) を適用し，その初出ピークの値から定常性，すなわち繰り返し動作が見られるかどうかを判別する．村尾ら [20] を参考に，入力の離散値 x_t に対し，ずれ τ における自己相関の値 R_{xx} は以下のように決定される．

$$R_{xx}(\tau) = \sum_{t=0}^{N-1} x(t)x(t-\tau) \tag{4.1}$$

さらに，[-1, +1] の範囲に自己相関の値を正規化する必要がある．R_xx は $\tau = 0$ の時に最大値をとるので，正規化した関数 R'_xx は以下の式で与えられる．

$$R'_{xx}(\tau) = \frac{R_{xx}(\tau)}{R_{xx}(0)} \tag{4.2}$$

そして，この関数 $R'_xx(\tau)$ における初出ピークの値と初出ピークが出現した時の時間を示すインデックス値，ピークが出現する回数，そして全てのピーク時における三次元距離の平均と分散を定常性を表す特徴量として用いる．

4.2　顔動作認識手法の実装

4.2.1　深度カメラセンサによる顔の特徴点群の取得

まず，鼻と顎先の三次元位置の距離情報を取得する必要がある．本研究では，Intel 社が提供している RealSense デバイス及び SDK[8] を利用する．RealSense デバイス及び SDK によって，当デバイスからおよそ 0.20[m] 1.2[m] の範囲に置いて顔の表出情報を 30[FPS] で取得することができる．図 4.4 に，表出情報の一つである顔のランドマーク情報の例を示す．本研究では，Visual studio(C++)[9] と OpenCV2.4.10 ライブラリ [10] を用いて，顔の表出情報の時系列情報を記録するシステムを作成した．また，本研究では 1 回の動作に用いるデータを 128 回分とする．1 回の動作にかかる時間は，1 秒間に 30 回データを取得することができるため，約 4 秒間の時間となる．

図 4.4: RealSense デバイス (左) と RealSendse SDK による顔のランドマーク情報の例

[8]Intel realsense device and sdk. http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html
[9]Visual Studio, Microsoft Corporation. https://www.microsoft.com/
[10]OpenCV. http://opencv.org/

4.2.2　矩形波窓によるフィルタリング

次に，5 つ距離の時系列情報に対し，矩形波窓によるフィルタリングをかける．この矩形波窓の大きさは，被験者によって最適な窓の大きさが変わる可能性がある．また，一回のフィルタリングではノイズを除去しきれないため，複数回繰り返してかけるものとする．本研究では顔動作認識を行う前に，矩形波窓と反復回数を事前実験にて最適なパラメータを決定する．

データの解析には，Python2.7 を利用した．図 4.5 に，咀嚼時の L_c の時系列情報に対して，窓の大きさ 20，反復回数 3 回のパラメータ設定にて矩形波窓を適用した時の例を示す．図 4.5 より，矩形波窓によって，ノイズを軽減した上で大きく凹凸となっている箇所を算出することができる．フィルタリング前のデータは，凹凸部分にノイズが走っており検出は困難となっているが，フィルタリング後のデータは値の上昇下降のタイミングを検出しやすくなっている．しかしながら，実際の凹凸のタイミングと時間的なズレがみられる．また，矩形波窓の大きさによってどれだけ細かい凹凸を検出するかが決定されるため，小さな咀嚼をノイズとみなされることがある．これは，顎関節の大きさなど，個人差によって変化するかもしれない．したがって，本研究では，顔の動作に対して，矩形波窓の大きさや回数はどの程度が適切かを事前実験にて求めた上で，そのパラメータを用いて本実験を行う．

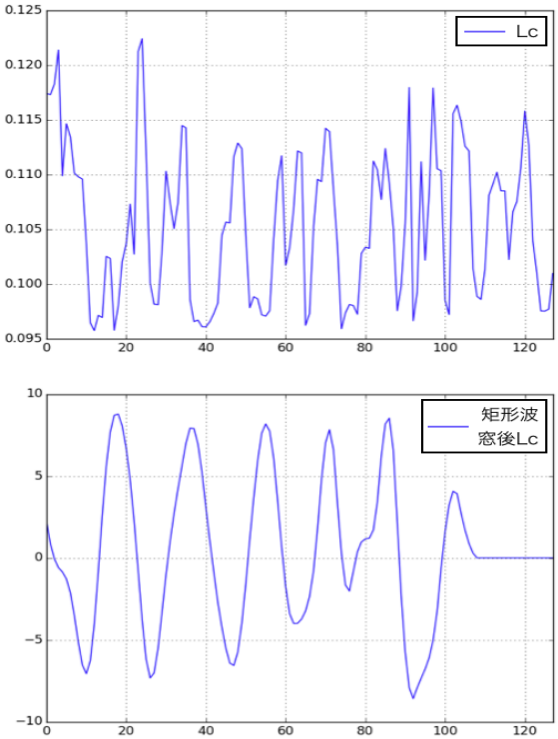


図 4.5: 矩形波窓による距離 L_t（上部）にフィルタリングを三回かけた結果の波形（下部）.

4.2.3　自己相関関数

顔の動作中の 5 つの三次元距離情報の時系列推移における定常性を計測するため，自己相関関数 (ACF) を用いる．図 4.6 に，食べる動作の間の自己相関関数を適用したデータの例を示す．ここで，本研究では動作一回のデータが 128[frame] であるのに対し，自己相関関数にかける範囲を 64[frame] としている．これは，自己相関関数を用いる場合，ずれ τ が最大の時において，式 4.1 において仮に自己相関関数にかける範囲が動作データの半数以上になってしまった場合は計算することができなくなる．したがって，本研究では自己相関関数にかける範囲を動作データの半分としている．

ここで，定常性が見られる場合，初期ピークが高く現れる．村尾ら [20] は，閾値を用いて定常性があるかどうかを判別していた．それに対し本研究では，初期ピークそのものの値を機械学習における特徴量として用いる．加えて，初期ピーク時のずれ τ の値，全ピークの回数，全ピーク間のずれ τ の間隔の平均および分散を定常性の特徴として用いる．

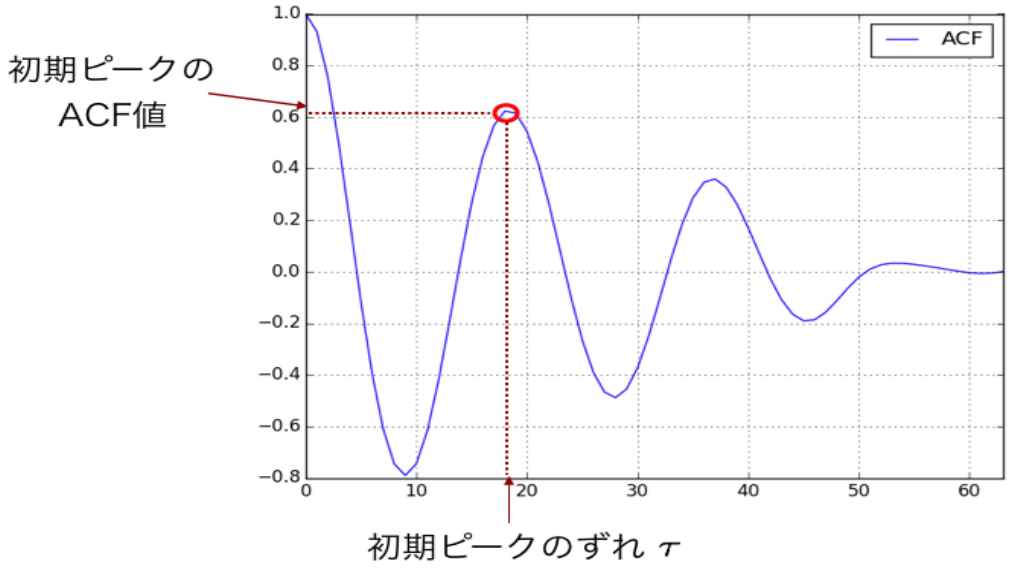


図 4.6: 食べる動作において自己相関関数を適用した例

4.2.4　全体の特徴量生成の流れと機械学習手法について

実装の最後に，全体の特徴量生成の流れを図 4.7 に示す．顔の表出情報の取得から，一回の動作における 5 つの距離の時系列的な変化のデータを算出する．ここで，それぞれの波形における基本統計量（平均，分散，最大，最小）を取得しておく．しかしながら，基本特徴量だけでは定常性を考慮できないため，矩形波フィルタをかけたのち，自己相関関数を適用することで，定常性に関する特徴を取得する．基本統計量のみを用いる場合，特徴量の次元は 5 種類の距離と 4 つの基本統計量の組み合わせのため，20 次元の特徴量となる．また，定常性を考慮した場合は 1 つの距離あたり 9 次元の特徴となるため，45 次元の特徴量となる．また，この特徴量を用いて，機械学習手法によって顔の動作を判別する．本研究においては，多クラス識別に適しているため，Random Forest[1] を用いる．

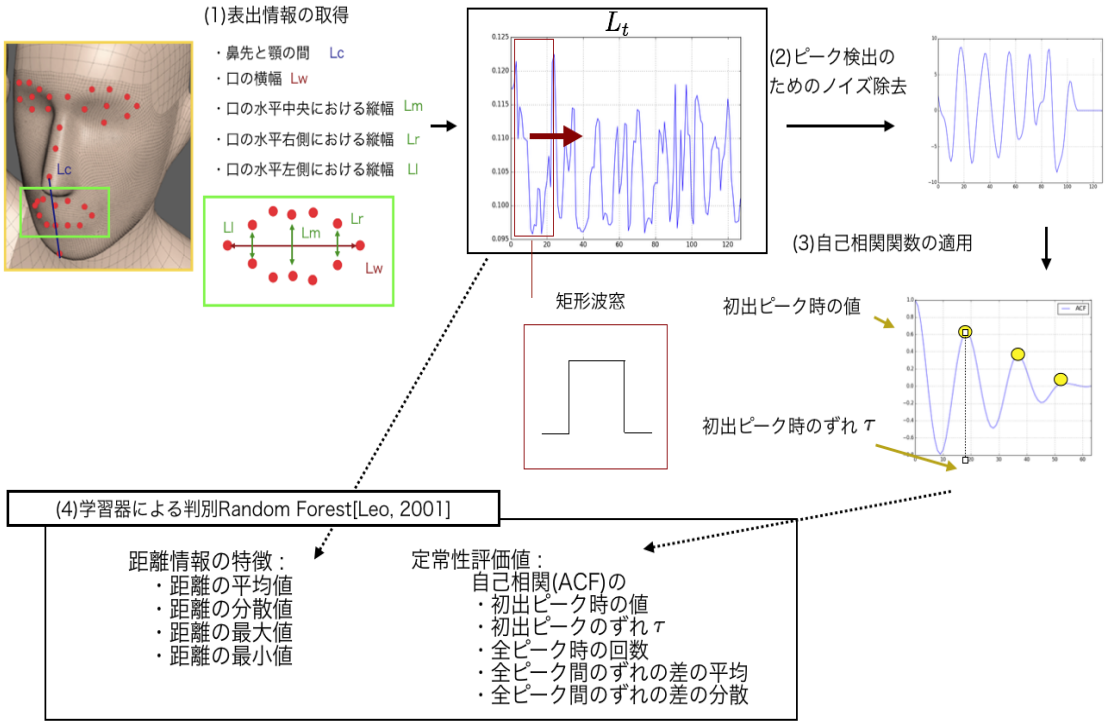


図 4.7: 顔動作認識における特徴量生成の流れ

4.3　予備実験（咀嚼動作認識実験）

4.3.1　使用するデータセットの作成

顔の動作認識における精度検証を行う前に，適切な矩形波窓の大きさと，かける反復回数を決定するための実験を行う．実験環境としては，被験者は RealSense デバイスの正面に座り，デバイスから 1.5[m] ほど離れて食事をする．また被験者には，図 4.8 に示すように白米を 10[g]，口に含んでから飲み込むまで咀嚼してもらう．被験者は 5 人で，20 代の被験者が 4 人，40 代の被験者が 1 人となっている．5 人の被験者，5 回の反復回数の全 25 データで構成されたデータセットを作成した．

図 4.8: 5 回分の白米 (10[g])

4.3.2　咀嚼回数の推定

評価方法としては，鼻先と顎の距離 $L_c(t)$ に矩形波フィルタをかけたのち，波形から咀嚼回数を推定する．具体的には，波形の下に凸となっている箇所の総数を予測咀嚼回数としてみなす．また，正解となる咀嚼回数は，波形データと時間をリンクさせた動画データから目視によって回数を判断した．表 4.1 に，試行した矩形波窓の大きさと反復回数の組み合わせにおける，咀嚼回数の被験者全員の平均推定精度を示す．ここで，評価する精度は，予測咀嚼回数を正解の咀嚼回数で割ったものとする．この結果から，本実験で用いるパラメータは，矩形波窓の大きさを 20，反復回数を 5 回とする．

表 4.1: 矩形波窓ごとの推定した咀嚼回数

		矩形波のサイズ [frame]				
		12	16	20	24	28
反復回数	2 回	44.16	58.96	72.64	84.03	85.86
	3 回	50.89	71.47	86.18	84.85	78.57
	4 回	55.42	76.67	87.87	80.79	72.61
	5 回	59.75	78.68	88.15	76.52	68.10
	6 回	59.37	79.08	88.11	76.52	67.74

4.4　実験（顔動作認識実験）

4.4.1　使用するデータセットの作成

顔の動作認識において，定常性を考慮することの有効性を示すために，食べる，話す，何もしないの 3 つの顔動作に関するデータセットを作成した．被験者は RealSense デバイスの正面に座り，デバイスから 1.5[m] ほど離れて各動作を行う．ここで，食べる動作においては，咀嚼実験と同様，被験者が食べる種類は米 10[g] である．被験者は 6 人で，全て 20 代の被験者であり，男性が 4 人，女性が 1 人となっている．6 人の被験者，3 種類の動作，5 回の反復回数の全 90 データで構成されたデータセットを作成し，これを評価に用いる．また，それぞれの試行において，1 回の動作データとして，128[frame]（1 秒間に 30[frame]

のデータの取得, 約 4 秒間）を 1 つのデータとした．これは，身体動作が約 4 秒間であったため，顔の動作認識においてもその時間が妥当であると考えたためである．また，矩形波相間フィルタの窓のサイズを事前実験の結果から最良である 20 とし，反復回数を 5 回とした．精度評価の機械学習手法には Random Forest[1] を用いる．ここで，定常性を考慮しない場合は 5 種類の距離に対して 1 つの距離あたり 4 次元であるため 20 次元となる．一方，定常性を考慮した場合の特徴量の次元数は，5 種類の距離に対して 1 つの距離あたり 9 次元であるため 45 次元となる．また，精度評価には 10 分割交差検証を用いるが，精度評価の前にそれぞれの手法において以下のパラメータ候補についてグリッドサーチを行っている．この時，データを 3 分割し，それぞれの手法で最も良かった時のパラメータを用いて 10 分割交差検証を行う．

表 4.2: グリッドサーチにおけるパラメータの候補 (顔動作認識)

パラメータの種類	候補パラメータ
バギングに用いる決定木の個数	[5, 10, 100]
葉を構成するのに必要な最小限のサンプルの数	[3, 5]
最適な分割をするために考慮する特徴の数	[3, 5]
決定木の深さの最大値	[5, 10, 25, 100]

4.4.2　実験結果

定常性を考慮しない場合と，定常性を考慮した場合についてそれぞれ 10 分割交差検証による精度評価を行ったところ，定常性を考慮しない場合は 80.0%，定常性を考慮した場合は 83.3%となった．したがって，定常性を考慮することは有効であると考えられる．

4.5　考察

顔動作の認識精度を向上させる要因として，予備実験時において，矩形波窓の大きさやかける反復回数による，被験者ごとの最適なパラメータの違いにつ

いて考察する．まず，窓サイズを変更した時の被験者ごとの咀嚼回数の精度を，表 4.3 に示す．表 4.3 から，被験者 D のみ，窓サイズが 16 の時が最も精度が良く，他の 4 人については，窓サイズが 20 の時に最も精度が良い．また，反復回数を変更した時の被験者ごとの咀嚼回数の精度を，表 4.4 に示す．表 4.4 については，A は 5 回が最も良く，B は 6 回，その他の 3 人は 4 回が最も良い精度となっている．以上の結果から，矩形波窓の大きさやかける反復回数は，被験者によって異なっている．これは，咀嚼一回一回の動作の大きさや，間隔が被験者によって異なるからであると考えることができる．したがって，顔の動作実験において，被験者ごとのパラメータをあらかじめ最適化したり，また使用するごとに徐々に個人にパラメータを特化していくことによって，より精度よく顔の動作認識を行うことができると考えられる．

表 4.3: 矩形波窓の窓サイズの変化による被験者ごとの咀嚼回数推定精度（反復回数=5）

被験者	矩形波のサイズ [frame]		
	16	20	24
A	85.3	94.7	88.4
B	74.1	91.2	84.1
C	79.4	92.1	74.1
D	88.5	80.4	55.2
E	66.1	82.3	81.2

表 4.4: 矩形波窓の反復回数の変化による被験者ごとの咀嚼回数推定精度（窓サイズ=20）

被験者	反復回数		
	4 回	5 回	6 回
A	93.4	94.7	93.9
B	87.1	91.2	94.5
C	93.3	92.1	92.2
D	83.0	80.4	78.4
E	82.5	82.3	81.6

第5章　考察

本章では，まず深度映像を用いた身体全体における動作認識手法及び顔動作認識する手法についての結果を踏まえた上で，それぞれの手法について考察する．その上で，身体動作認識手法と顔動作認識手法を組み合わせた場合についての考察を行う．二つの動作認識手法の組み合わせについては，想定する環境を設定した上で，身体動作と顔動作が同時に行われることを想定する．そして，身体動作と顔動作が同時に行われた場合に起こりうる，身体と顔の動きの相互作用について述べる．さらに動作認識手法を組み合わせる方法論について論じた上で，身体と顔の動きの相互作用を考慮する場合について考察する．

5.1　部位分割的動作認識手法に対する考察

第三章では，身体全体の動作を認識する手法として，骨格関節を 6 つの部位に分割し，それぞれの部位において動作の特徴を用いて分類器を生成した上で，部位アンサンブルによる予測を行うことで動作を認識する手法を提案した．また屋内実環境を想定し，センサデバイスと認識対象者の間に存在する障害物を閉塞として，閉塞の有無の両方の状態における動作のデータセットを作成した．そして実験として，閉塞のない状態で学習した上で，閉塞のある状態での動作に対する予測精度を Xia ら [33] の手法と比較した．結果として，閉塞のある状態においては Xia ら [33] の手法が 83.2%であるのに対し本提案手法の予測精度が 84.7%となり，閉塞のある屋内環境における本提案手法の有効性を示すことができた．

しかしながら，閉塞のないデータに対して 10 分割交差検証を行ったところ，Xia ら [33] の手法が 89.7%であるのに対し本提案手法の予測精度が 85.0%となった．これは，本提案手法は部位ごとの動きに対する特徴を局所的に捉えること

ができるが，身体の全体的な特徴を捉えきれてないからだと考えられる．そのため，例えば腰を基準とした各部位の角度情報を特徴に加えた上でアンサンブル予測を行うことで，局所的だけでなく身体の全体的な特徴を考慮して動作認識を行うことができるようになると考えられる．そして，動作認識精度が向上だけではなく，閉塞がある環境に対しても Xia ら [33] の精度を達成することができる可能性がある．

5.2　顔動作認識手法に対する考察

第四章では，身体部位における一部として，顔動作を認識する手法を提案した．顔の表出情報の 10 点から 5 つの距離を算出し，各距離の動作中の時系列変化における基本特徴量に加えて定常性を考慮した特徴を生成し，機械学習手法を用いて顔動作を認識する手法を提案した．定常性の考慮には，矩形波窓と自己相関関数を用いた．まず，予備実験として，矩形波窓の大きさとかける反復回数の最適な値を探るために，咀嚼回数の精度を検証したところ，適切な窓の大きさは 20, 反復回数は 5 回となった．さらに，この値を用いて，顔動作認識において定常性を考慮することの有効性を確認するために，食べる，話す，何もしないの動作認識の精度を，基本特徴のみの場合と定常性を含む特徴を用いた場合とで比較することで検証した．結果として，基本特徴のみの場合は 80.0%であるのに対し，定常性を含む特徴を用いた場合は 83.3%と精度の向上が見られた．また，被験者によって最適な矩形波窓の大きさが変わることから，認識対象者による違いを考慮することで，より正確に動作を認識できると考えられる．

5.3　身体動作認識と顔動作認識を融合する際の考察

次に，身体動作認識と顔動作認識を融合する場合について考察する．想定する環境について述べた上で，身体動作と顔動作を同時に行った時に，身体の動きと顔の動きが互いに影響することが考えらえる．その互いの相互作用について考慮した上で，考えられる手法を挙げていく．

5.3.1　想定する環境

初めに，身体動作認識と顔動作認識を融合する際の想定する環境について述べる．図 5.1 に，想定する環境の一例を示す．本研究で想定する環境は，屋内において，机などの，センサデバイスと認識対象者の間に閉塞となりうるような障害物が存在する環境を想定している．現実的な屋内の生活環境においては，机だけでなく，ソファや物置など，生活に必要でありながら，センサデバイスが対象人物全体を取得することを阻害するような物が多い．この解決方法としては，部屋の全範囲をカバーできるように，四方八方，また高い場所や低い場所に複数のセンサデバイスを配置する方法が考えられる．しかしながら，一般家庭などでそのような配置をすることは難しく，また病院などの施設に多いても，設置コストやデバイスの維持などのランニングコストがかかってしまう．そこで本研究では，テレビや机の前，あるいは病院におけるベッドなど，認識対象者が主に移動するあるいは停留する箇所が存在すると仮定し，その範囲内において動作を認識することを考える．

また本研究では，図 5.1 のように，身体骨格と顔の両方の動作を認識できる環境を想定する．現在の技術では，単一のデバイスで身体骨格と顔の表出情報を同時に，かつ処理速度を維持したまま，さらに同じ範囲で認識することは困難である．しかしながら，今後，技術の進展により身体骨格と顔の表出情報を同時に取得できると考えられる．したがって，想定する環境として，身体骨格と顔の表出情報を同時に取得できる環境での動作認識について考えていく．

図 5.1: 想定する環境について

5.3.2　複合的動作における認識手法について

身体と顔の複合的動作を考慮した上で，身体動作認識と顔動作認識を融合させることを考える．ここで，身体と顔の特徴を単純に結合して一つの特徴を生成し機械学習によって動作認識を行う手法は，第三章において，閉塞のある環境においては適さないことがわかっている．したがって，(i) 身体動作と顔動作を別々に認識しておき，動作ラベルを文書的に結合する手法と，(ii) 身体における 6 つの各部位の分類器と顔における口を一部位として分類器を生成し，7 つの部位において部位アンサンブル予測を行う手法の二つの手法について考察していく．

(i) 身体動作と顔動作を別々に動作認識しておき動作ラベルを結合する手法

一つ目の手法として，身体の動作認識及び顔の動作認識を別々に行った場合について考察する．この場合，身体動作及び顔動作の認識は，それぞれ第三章及び第四章で説明した手法で行う．すなわち，身体の動作認識においては，身体骨格の関節の三次元位置情報及び方向情報を用いて，部位分割によって部位ごとの特徴及び分類器を生成しておき，未知の入力データに対しては部位アンサンブル予測によって身体動作を認識する．一方，顔の動作認識においては，顔の表出情報の口周辺の 10 点から 5 種類の距離の特徴を算出する．こちらも，学習データによって分類器を生成しておき，未知のデータに対して顔の動作を認識する．そして二つの動作認識によって得られたそれぞれのラベルを，文書的に結合する．例えば，身体が歩く，また顔が食べる，動作を行っていた場合，歩きながら食べる，といったラベルが最終的に出力される．

(ii) 身体部位と顔部位において部位アンサンブル予測を行う手法

次に二つ目の手法として，身体部位における 6 部位と，顔における口周辺を 1 つの部位とみなし，7 部位の分類器を生成しておき，その 7 部位間で部位アンサンブル予測を行う場合について考察する．この手法においては，まず身体骨格の関節の三次元位置情報及び方向情報を用いて，部位分割によって部位ごとの特徴及び分類器を生成しておく．次に並行して，顔の表出情報の口周辺の 10 点から 5 種類の距離の特徴を算出し，学習データによって分類器を生成しておく．すなわち，身体で 6 つの部位ごとの分類器，顔で 1 つの部位ごとの分類器，合計で 7 つの部位の分類器が生成される．

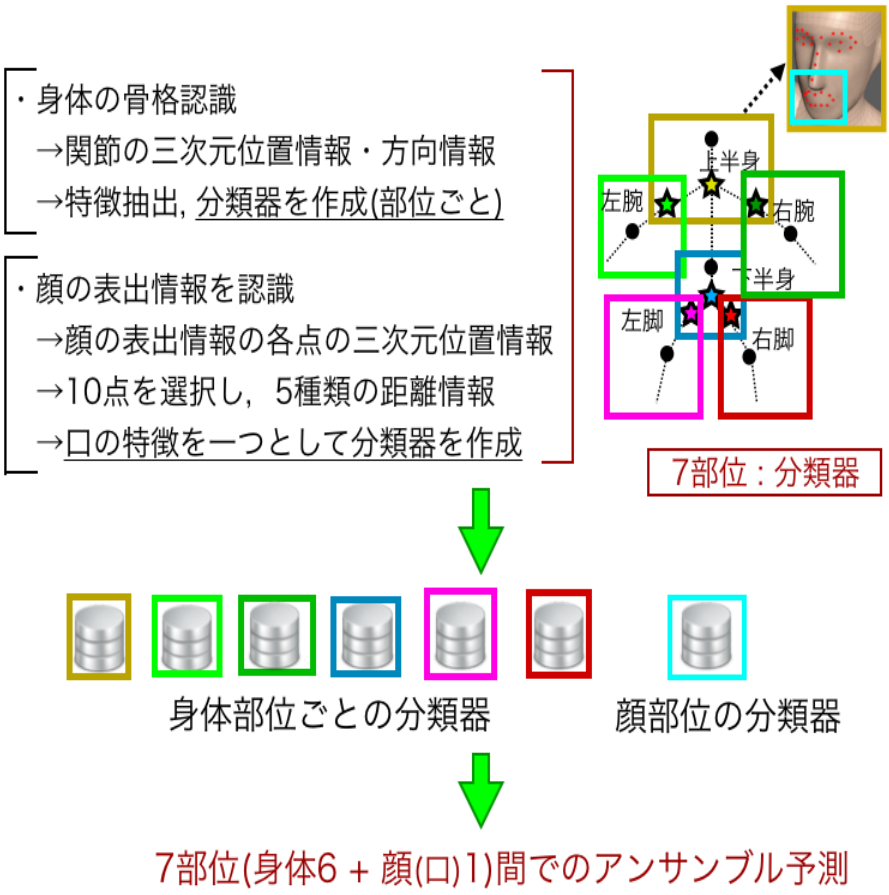


図 5.2: 身体部位と顔部位において部位アンサンブル予測を行う手法の流れ

5.3.3　身体と顔の複合的動作における相互作用について

身体動作と顔動作を同時に行ったときの全身動作を，身体と顔の複合的動作と定義する．身体と顔の複合的動作が起こった時，身体及び顔の振る舞いは，それぞれ別に動作した時に取得した振る舞いと異なると考えられる．例えば，立って歩きながら食べている時の顔の表出情報は，身体動作の影響を受けており，椅子に座りながら食べている時の顔の表出情報と異なると容易に予想できる．また，身体において立つ動作においても，携帯端末などで話しながら立つ場合と，顔で何も複合的動作を行わずに立つ場合とでは，立ち方が異なってくるため，身体骨格の関節位置や方向情報の時系列推移は一致しないと考えられる．さらに，村尾ら [21] は，両手首，腰，両足首に加速度センサを装着して動作データを取得した際，同じ身体の動作において，歩行中の手の動作識別が困難であるとしている．これは，歩く時の身体全体の振動が手に伝わり，手の動作が上手く識別できないためと考察している．屋内環境においては，歩くだけでなく，姿勢が大きく変化するような立つや座るなどの身体動作中においても，身体からの動きが顔の動作に影響しやすいと考えられる．例えば屋内では身体で立つ動作を行いながら顔で食べる動作と，座る動作を行いながら顔で食べる動作とでは，姿勢の変化による身体動作の影響が顔の表出情報に伝わる可能性は十分に考えられる．また，顔で食べている最中に，身体でゴミを投げた場合，その投げた振動が顔に伝わると考えられる．

したがって，身体と顔の複合的動作における相互作用を考慮しつつ，身体動作認識と顔動作認識を融合させる必要がある．ここで，先ほど紹介した二つの複合的動作における手法について，それぞれの手法で複合的動作における相互作用における影響を考察する

(i) 複合的動作における影響 (身体と顔を別々に動作認識し動作ラベルを結合する手法)

一つ目の手法である，身体と顔を別々に動作認識し動作ラベルを結合する手法において，複合的動作の場合について考察する．ここで，身体の分類器及び顔の分類器がそれぞれ複合的動作が行われていない場合のデータで学習してい

た場合，未知のデータが複合的動作であるならば，当然分類器は先述した身体と顔の複合的動作における相互作用について考慮されておらず，うまく認識できないと考えられる．したがって，学習データにおいても，身体及び顔の複合的動作が行われている状態を含めて取得すべきであると考えられる．

次に，学習データにおいても，身体及び顔の複合的動作が考慮された場合について考察する．その場合，単純に身体 10 動作と顔 3 動作の組み合わせ，すなわち 30 の動作における分類が可能となる．しかしながら，身体動作と顔動作を別々に動作認識しておき動作ラベルを結合する手法では，身体及び顔それぞれ別に独立した分類器であるとみなすことができるため，身体と顔の複合的動作における相互作用についての考慮が身体と顔で途切れてしまうことが考えられる．

(ii) 複合的動作における影響 (身体と顔で部位アンサンブル予測を用いる手法)

二つ目の手法である，顔を 1 部位とみなし，身体と顔で部位アンサンブル予測を用いる手法において，複合的動作の場合について考察する．考察する，まず身体の分類器及び顔の分類器がそれぞれ複合的動作が行われていない場合のデータで学習していた場合について考察する．この場合においては，(i) 身体動作と顔動作を別々に動作認識しておき動作ラベルを結合する手法と同様，未知のデータが複合的動作であっても，学習時にその影響は考慮されていない．したがって，やはり学習時に複合的動作における影響を含めて学習すべきであると考えられる．

そして，学習データにおいても，身体及び顔の複合的動作が考慮された場合について考察する．その場合，単純に身体 10 動作と顔 3 動作の組み合わせ，すなわち 30 の動作における分類が可能となる．また (i) 身体動作と顔動作を別々に動作認識しておき動作ラベルを結合する手法と比べ，身体と顔の間で部位アンサンブル予測を行うため，身体と顔の複合的動作における相互作用の影響を吸収できると考えられる．したがって，より複合的動作を考慮して動作を認識することができると考えられる．

5.4　身体動作認識と顔動作認識に関する今後の展望

最後に，身体動作認識と顔動作認識に関する今後の展望について述べる．まず，5.1 節で述べた通り，身体動作認識の手法としては，身体の全体的な特徴を加えることで，動作において局所的な部位だけでなく身体の全体的な動きを考慮して動作認識精度を向上させることができると考えられる．また，動作ごとの確率推算値の分布が平均的になるような部位に対して閾値を用いることで，より特徴の強い部位を考慮し，屋内環境に適した動作認識が可能となるかもしれない．さらに，顔の動作認識に関して，5.2 節で述べた通り，認識対象者による違いを考慮することで，より正確に動作を認識できると考えられる．また，身体動作認識と顔動作認識を融合させる際，身体部位と顔部位において部位アンサンブル予測を行うことで身体と顔の複合的動作における相互作用を考慮できると考えられる．さらに，身体において手指を一つの部位，あるいは顔における目をそれぞれ一つの部位とみなし，部位アンサンブル予測を用いることで，より多くの動作を認識することができる可能性がある．

そして，身体と顔の複合的動作を考慮して動作認識を行うことができるようなれば，さらに健康管理サービスや生活支援サービスの機能の充実へと繋がると考えられる．例えば，日常的に座る場合と，苦しみながら座る場合とを区別することができるようになり，健康管理サービスにおける体調把握機能などへと応用することができると考えられる．さらに，食べる動作における口の咀嚼の感覚の速さや，食べ物を口に持って行く腕の動作を認識することで，食べ方のアドバイスなどの機能へと発展させることができると考えられる．また，電話をかけながら話す動作を認識することで，環境音の音量を一時的に下げるなどの生活支援サービスの機能への応用が考えられる．

第6章　結論

本論文では，照明変化に頑健な深度映像を用いて，屋内実環境における人物の可視部分に着目して動作認識を可能にする手法を提案することを目的とした．本提案手法では，深度映像を用いた骨格認識技術によって算出された身体骨格の三次元位置情報の時系列推移および顔の特徴的な点の三次元位置情報の時系列推移を用いることで, 時空間的な動作認識を実現した．さらに，屋内実環境における閉塞を考慮し，実生活環境に適した人物動作認識手法を提案した．

第一章「序論」では，人物動作認識について，利用イメージや，動作とはどう定義されるかについて論じた．そして，既存の人物動作認識に関する研究体系を説明し，さらに屋内環境における人物動作についての問題点について論じた．そして，本研究で取り組む課題について言及し，目的を明らかにした．

第二章「関連研究」では，動作認識手法を画像差分による手法と骨格モデルを用いた手法を，それぞれについてさらに RGB カメラセンサと深度カメラセンサを用いた手法紹介した．特に深度映像を利用した骨格モデルによる手法について取り上げ，現状の問題点を指摘した．また顔における動作認識手法の紹介を行い，問題点を指摘した．その上で，本研究の位置付けを行った．

第三章「部位分割的動作認識手法」にて，身体動作認識において, 実生活環境における閉塞を考慮するため, 身体骨格における関節を 6 つの部位に分け, 部位ごとに独立した特徴を生成した．部位ごとに独立した特徴を生成するために, 各部位における基準関節を設定し，その方向情報を用いた．さらに，各部位において分類器を作成しておき，予測する入力データに対して，各分類器に入力した際に出力される動作ラベルごとの確率推算値を重みとして用いる，部位アンサンブル予測による動作認識手法を提案した．そして，提案手法が閉塞に対する有効性を示すために，閉塞状況がある場合とそうで無い場合の両方につい

て同様に，日常動作における関節情報の推移を記録したデータセットを作成し，評価を行った．この時，本データセットに対して，Xia ら [33] の手法を適用した上で本提案手法と比較したところ，閉塞のある環境に対して既存手法 [33] の精度が 83.2%であったのに対し，本提案手法の精度は 84.7%となった．この結果から，閉塞のある環境に対する本提案手法の有効性が示された．

次に第四章「顔動作認識手法」では，身体骨格における一部分として，顔の特徴的な表出点位置情報を用いて顔動作を認識する手法を提案した．鼻と顎の距離, 口の縦方向の距離, 横方向の距離において, それぞれの基本統計量を特徴の基本とした．さらにノイズ除去を行った上で, 定常性に関する自己相関の初出ピーク時の相関値とインデックスを特徴として加えた．そして, 複数の距離情報から得られた特徴に対して, Random Forest[1] による動作の分類を行う手法を提案した．また, 提案手法において，定常性の効果を検証するために，食べる, 話す, 何もしない時のデータセットを作成し, 分類性能を評価した．その結果，定常性を考慮しない場合は 80.0%，定常性を考慮した場合は 83.3%という認識精度となった．したがって，顔部位において口と顎の特徴的な表出点位置情報から，基本統計量に加えて定常性を考慮することで顔動作の認識精度が向上することが確認された．

そして第五章「考察」では，身体動作認識と顔動作認識のそれぞれの結果と考察を踏まえた上で，身体動作認識と顔動作認識を融合させる場合の手法について考察した．その際，身体動作と顔動作を同時に行った際の，身体骨格の動きと顔の表出情報の動きの相互作用について論じた上で，身体動作認識と顔動作認識を融合させる手法を二つに分け考察した．また，身体動作認識と顔動作認識の融合についての今後の展望について考察した．

最後に，今後の展望としては，本研究で提案した手法により屋内における日常的な動作を識別することで， 例えば単身者の日常生活を支援するようなサービスが可能になると考えられる．単身者がリビングでの食事する際, 咀嚼状況を把握することで, 身体的健康管理を支援するサービスへ繋がると考えられる．また, 単身者の手と顔の動作を識別することで，携帯端末で電話を行う際に TV などの環境音を自動で小さくするなどのサービスが考えられる．

謝辞

私の研究指導者である西山裕之教授は，私が学部生の頃からの恩師であり，博士課程単位取得満期退学以降も続けて，本論文作成のために御指導頂きました．特に，人物動作認識技術の必要性，身体と顔における動作認識の融合などの本研究の方針や方向性を始め，論文のまとめ方や発表方法，さらに研究者としてのあり方について，きめ細やかに御指導して下さいました．また，国際会議等で発表する機会を与えて頂いたことにより，海外の研究者との意見交換を介して本研究の成果をより向上させることができました．ここに，心より感謝致します．

本論文の研究指導を受け持って頂きました大和田勇人教授，原田拓准教授，滝本宗宏教授，竹村裕准教授の方々には，ご多忙なところ長い時間に渡って論文の内容を吟味して頂き，構成や記述形式などの詳細な御指導を頂くことで本論文をまとめることができました．誠に感謝しております．

最後に，長きに渡って私を育て見守って頂き，本論文の作成に陰ながら支援して下さいました私の両親に最大の感謝の言葉を贈ります．ありがとうございました．

参考文献

[1] L. Breiman.: ”Random forests,” Machine Learning, Vol. 45, No. 1, pp.5-32, 2001.

[2] A. F. Bobick, and J. W. Davis.: “The recognition of human movement using temporal templates,” IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI), Vol. 23, No. 3, pp.257-267, 2001.

[3] B. I. Bogart, V. H. Ort.: 『インテグレーテッドシリーズ 3 解剖学・発生学』 佐藤宏，大谷修，小澤一史，村上徹訳，東京化学同人，第 1 版第 1 刷．ISBN 978-4-8079-1645-0，2011．

[4] S. Cadavid, M. Abdel-Mottaleb and A. Helal.: ”Exploiting visual quasi-periodicity for real-time chewing event detection using active appearance models and support vector machines,” Personal and Ubiquitous Computing, Vol. 16, pp.729-739, 2012.

[5] L. Campbell and A. Bobick.: ”Recognition of human body motion using phase space constraints,” Proceedings of IEEE International Conference on Computer Vision, Fifth International Conference on, pp.624-630, 1995.

[6] A. A. Chaaraoui, J. R. Padilla-Lopez and F. Florez-Revuelta.: ”Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices,” IEEE international conference on computer vision workshops (ICCVW), pp.91-97, 2013.

[7] T. F. Cootes, G. J. Edwards and C. J. Taylor.: ”Active appearance models,” Computer Vision - ECCV’98. Lecture Notes in Computer Science. 1407. pp. 484, 1998.

[8] Y. Goutsu, W. Takano and Y. Nakamura.: ”Classification of Multi-class Daily Human Motion using Discriminative Body Parts and Sentence Descriptions,” International Journal of Computer Vision, pp.1-20, 2017.

[9] A. Jalal, M. Z. Uddin, J.T. Kim and T.S. Kim.: ”Recognition of Human Home Activities via Depth Silhouettes and R Transformation for Smart Homes,” Indoor and Built Environment, Vol. 21, pp.184-190, 2011.

[10] H. Kado, T. Ueda. Japan Patent Kokai 4-141440, 1992.

[11] 加賀谷拓, 羽倉淳, 藤田ハミド.: “無意図的動作に着目した人間のしぐさからの情動推定手法,” 日本ソフトウェア科学会第 22 回大会論文集, 3C-2, 2005.

[12] A. Kanazawa, M. J. Black, D. W. Jacobs and J. Malik.: ”End-to-end Recovery of Human Shape and Pose,” Cornel University Library, arXiv:1712.06584, 2017.

[13] 久保田新.: “運動・動作・行動の流れの中の「意志」と意識,” 理学療法学, Vol. 33, No.4, pp.199-201, 2006.

[14] 河野邦雄，伊藤隆造，坂本裕和，前島徹，樋口桂.: 『解剖学第 2 版』 財団法人 東洋療法学校協会, 医歯薬出版，第 2 版第 1 刷．ISBN 4-263-24207-6，2006．

[15] I. Laptev and T. Lindeberg.: “Space-time interest points,” Proceedings of the International Conference on Computer Vision (ICCV ’03), vol. 1, pp.432-439, 2003.

[16] W. Li, Z. Zhang and Z. Liu.: ”Action recognition based on a bag of 3d points,” Human Communicative Behavior Analysis Workshop (in conjunction with CVPR), pp.9-14, 2010.

[17] C. Li, Y.Chen, W. Chen, P. Huang and H. Chu.: ”Sensor-embedded teeth for oral activity recognition,” Proceedings of the 2013 International Symposium on Wearable Computers, pp.41-44, 2013.

[18] 松村京子.:“乳児の情動研究:非接触法による生理学的アプローチ,” ベビーサイエンス (学会誌), Vol. 6, pp.2-14, 2006.

[19] 森武俊.: ”動作認識：その手法と応用,” 第 13 回ロボット学会学術講演会予稿集, pp.237-240, 1995.

[20] K. Murao and T. Terada.: ”A motion recognition method by constancy decision,” Journal of Information Processing Society of Japan, Vol. 52, No. 6, pp.1968-1979, 2011.

[21] K. Murao and T. Terada.: ”A Conmined-activity Recognition Method with Accelerometers,” Journal of Information Processing, Vol. 24, No. 3, pp. 512 - 521, 2016.

[22] T. Nanri and N. Otsu.: ”複数人動画像からの異常動作検出,” 情報処理学会論文誌コンピュータビジョンとイメージメディア (CVIM), Vol. 46, pp. 43-50, 2005.

[23] I. Nasu and Y. Saito.: ”Active life expectancy for elderly japanese by chewing ability,” Japanese Society of Public Health, pp. 411- 423, 2006.

[24] N. M. Oliver, B. Rosario and A. P. Pentland.: ”A Bayesian Computer Vision System for Modeling Human Interactions,” IEEE Trans. PAMI, Vol. 22, No. 8, pp. 831-843, 2000.

[25] O. Oreifej and Z. Liu.: ”HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” IEEE computer society conference on computer vision and pattern recognition(CVPR), pp.716-723, 2013.

[26] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo and Z. Tu.: ”Exemplar-based human action pose correction and tagging,” Computer Vision and Pattern Recognition (CVPR), pp.1784-1791, 2012.

[27] J. Shotton, A. Fitzgabbon, M. Cook, T. Sharp, M. Finnochio, R. Moore, A. Kipman and A. Blake.: “Real-time human pose recognition in parts from a single depth image,” IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[28] K. Taguchi., et al. Japan Patent Kokai 2015-93050, 2015.

[29] J. Taylor, J. Shotton, T. Sharp and A. Fitzgibbon.: ”The Vitruvian Manifold: Inferring Dense Correspondences for One-Shot Human Pose Estimation,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), pp.103-110, 2012.

[30] A. Toshev and C. Szegedy.: ”DeepPose: Human Pose Estimation via Deep Neural Networks,” Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp.1653-1660, 2014.

[31] J. Wang, Z. Liu, Y. Wu and J. Yuan.: ”Mining Actionlet Ensemble for Action Recognition with Depth Cameras,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), pp.1290-1297, 2012.

[32] C. Wang, Y. Wang and A. L. Yuille.: "An approach to pose-based action recognition," In IEEE computer society conference on computer vision and pattern recognition (CVPR), pp.915-922, 2013.

[33] L. Xia, C. Chen and J. K. Aggarwal.: "View Invariant Human Action Recognition Using Histograms of 3D Joints," The 2nd International Workshop on Human Activity Understanding from 3D Data (HAU3D) in conjunction with IEEE CVPR 2012, Providence, RI, pp.20-27, 2012.

[34] E. Yu and J. K. Aggarwal.: "Human Action Recognition with Extremities as Semantic Posture Representation," Computer Vision and Pattern Recognition Workshops in IEEE Computer Society Conference on, pp.1-8. 2009.

[35] J. W. Zhang, C. Luo and F. Chen.: "Joint head pose and facial landmark regression from depth images," Computational Visual Media, Vol. 3, pp. 229-241, 2017.

[36] 厚生労働白書, 2011.

[37] 東京都監察医務院東京都 23 区における孤独死統計, 2013.

付 録A　深度カメラセンサ

A.1　Kinect v1

Kinect v1 は microsoft 社から Xbox360 用に 2010 年 11 月に発売された人間の身体の部位や動作を深度計測によって取得することができるデバイスである．

赤外線によりデバイスから物体までの距離を測ることができる深度センサと，通常の RGB(Red Green Blue) カメラセンサの 2 つのセンサ，またマルチアレイマイクロフォンマイクを搭載している．元は Xbox360 用のゲームプラットフォームを目的として開発されていたが，2011 年 6 月，microsoft 社は Kinect for Windows SDK（Software Development Kit）Beta を無償提供し，商用利用以外の目的に限られるが，各センサーからの情報をもとに，機能は限定されてはいるものの開発を行うことができる．また 2011 年 11 月，Kinect for Windows SDK Beta2 にバージョンアップされ，さらに 2012 年 2 月，正式に windows OS を実行環境とした Kinect for windows デバイスと開発用の SDK である Kinect for windows SDK をリリースした．本製品の内部と仕様を記述しておく（Microsoft 公式ページ [11] より）．

[11] https://blogs.msdn.microsoft.com/kinectforwindows/2012/12/07/inside-the-newest-kinect-for-windows-sdkinfrared-control/

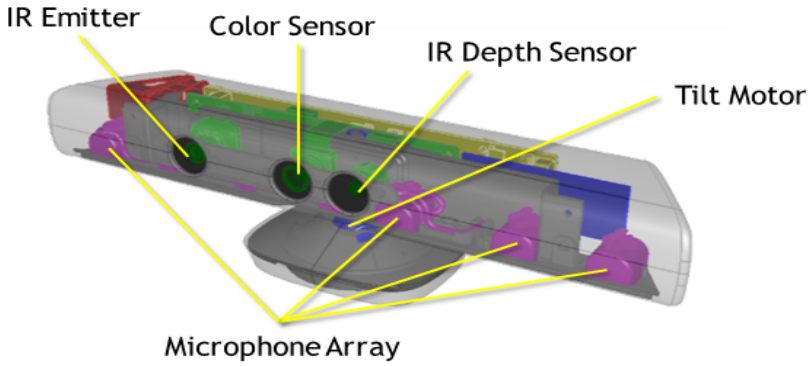


図 A.1: kinect センサの構造

表 A.1: kinect カメラと深度センサの仕様

仕様	内容
深度計測カメラ	640x480:16bit, 30fps
色計測カメラ	1280x960:32bit, 30fps
有効範囲	水平 57 度，垂直 43 度，± 27 度 (傾きの変更による)
センサの奥行き (通常)	0.8m ～ 4.0m
センサの奥行き (Near モード)	0.4m ～ 3.0m

A.2　Intel RealSense F200

Intel RealSense は，Intel 社から発売された深度センサモジュールのシリーズである．そのシリーズの一つとして，ノート PC のフロントカメラのような用途となる深度カメラセンサが Intel RealSense F200 が存在する．また，同 SDK には，手指の関節情報を検出する機能や音声認識機能，さらに本論文でも述べた通り，顔の表出情報を取得することができる機能を備えている．本製品の内部と仕様を記述しておく（Intel 公式ページより [12]）．

図 A.2: Intel RealSense F200 の構造

表 A.2: Intel RealSense F200 センサの仕様

仕様	内容
深度計測カメラ	640x480, 60fps(または，640x240, 120fps)
色計測カメラ	1920x1080, 30fps
有効範囲 (Depth)	対角線 90 度, 水平 73 度，垂直 59 度
手の認識範囲	0.20m ～ 0.55m
顔の認識範囲	0.35m ～ 0.70m

[12]https://software.intel.com/en-us/blogs/2015/01/26/can-your-webcam-do-this

付 録B　回転行列について

B.1　四元数

三次元空間でのx軸，y軸，z軸周りの回転を表す回転行列は，それぞれ次の通りである．

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & -sin\theta \\ 0 & sin\theta & cos\theta \end{bmatrix} \tag{B.1}$$

$$R_y(\theta) = \begin{bmatrix} cos\theta & 0 & sin\theta \\ 0 & 1 & 0 \\ -sin\theta & 0 & cos\theta \end{bmatrix} \tag{B.2}$$

$$R_z(\theta) = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{B.3}$$

Unityでは，他の3Dゲームエンジンや3Dレンダリングソフトなどで利用されるような，4x4行列を用いる．この行列は通常の三次元の回転行列に加え，最後の行または列が（0，0，0，1）で構成されている行列である．つまり，以下のような行列を指す．

$$R = \begin{bmatrix} x_1 & y_1 & z_1 & 0 \\ x_2 & y_2 & z_2 & 0 \\ x_3 & y_3 & z_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{B.4}$$

となっている．このようにオイラー角を利用した回転は，視覚的にわかりやすく，また演算も速いという利点がある．しかし，一つの回転軸が他の回転軸と並んでしまったとき，互いに回転がロックされるというジンバルロックという現象が起こる可能性がある．例えば，Z 軸を 90 度傾けると，X 軸と Y 軸が同軸となってしまい，姿勢制御が困難となってしまう．つまり，二回以上の回転計算をさせる時，オイラー角を用いた回転をさせるべきではないと言える．

B.2　クォータニオン

クォータニオン (四元数) は，William Rowan Hamilton によって記述された，複素数を拡張した数体型であり，特に 3D グラフィックスやコンピュータビジョンにおいて三次元での回転の計算によく用いられる．通常，回転に用いられるクォータニオンは $q = (\omega, V) = (\omega, x, y, z)$ で表される．

ここで，$V = (x, y, z)$ は虚部，ω は実部となる．回転量を θ，回転軸を (n_x, n_y, n_z) としたとき，この回転を表すクォータニオン q は

$$q = (cos\theta/2, n_x \cdot sin\theta/2, n_y \cdot sin\theta/2, n_z \cdot sin\theta/2) \tag{B.5}$$

と表すことができる．また，クォータニオンの回転の特徴として，乗算が容易であるという特徴がある．クォータニオンの乗算の定義は

$$\begin{aligned} q_0 q_1 &= (\omega_0, V_0)(\omega_1, V_1) = (\omega_0, x_0, y_0, z_0)(\omega_1, x_1, y_1, z_1) \\ &= (\omega_0\omega_1 - V_0 \cdot V_1, \omega_0 V_1 + \omega_1 V_0 + V_0 \times V_1) \end{aligned} \tag{B.6}$$

ここで，$V_0 \cdot V_1 = x_0x1 + y_0y1 + z_0z1$，$V_0 \times V_1 = (y_0z_1 - z_0y_1, z_0x_1 - x_0z_1, x_0y_1x_1)$ であるため，最終的に得られるクォータニオンの成分は以下のようになる．

$$\begin{pmatrix} \omega_0\omega_1 - x_0x1 - y_0y_1 - z_0z_1 \\ \omega_0x_1 + x_0\omega_1 + y_0z_1 - z_0y_1 \\ \omega_0y_1 + y_0\omega_1 + z_0x_1 - x_0z_1 \\ \omega_0z_1 + z_0\omega_1 + x_0y_1 - y_0x_1 \end{pmatrix} \tag{B.7}$$

つまり，回転を表すクォータニオンを合成する場合は，単純に乗算することで求めることができる．

しかし，新しいクォータニオンを回転によって作成する場合は異なる．回転を表す四元数は四つの実数の組であり，よってベクトルとしての長さが1であるという制約を課すことで，回転四元数の自由度を期待されるべき3に制限する．四元数は複素数の一般化として考えることができる．例えば，ケーリー＝ディクソンの構成法における，乗法と共役の一般化が挙げられる．このとき，四元数の乗法は実数の乗法と完全に同じではなく，可換にはならない．つまり，p，q が四元数ならば $pq = qp$ は真であるとは限らない．ちなみに，回転も同様に乗法を使って生成することができるが，行列や複素数の場合と異なり，二つの回転四元数を掛けて

$$x' = qxq^{-1} \tag{B.8}$$

となる．ここで，q は回転四元数，q^{-1} はその逆数で，x はベクトルとして扱われたクォータニオンである．

また，クォータニオンを利用することで，オイラー角による回転行列を使用する際に発生する可能性のある，ジンバルロック現象を回避することができる．