修士論文

# 視線を考慮した
# 一人称視点映像からの
# 頷き検出

2017 年 2 月 3 日 提出

指導教員　佐藤 洋一 教授

情報理工学系研究科電子情報学専攻

48-156448　中野 雄介

# 概要

頷き動作には，相手の顔を見て大きく頷く動作や，ものを見ながら小さく頷く動作などバリエーションが存在する．本研究では，コンピュータビジョン技術によってこのような頷き動作を映像から検出する課題に取り組む．固定カメラを用いた従来手法では，撮影画像中に現れる顔の動きを計測し利用するため，小さな頷きの検出が難しい．また，対象人物の顔の一部が物体等によってカメラから隠蔽されている場合も，頷き検出は困難となる．これに対して本研究では，頭部動作をより直接的に観測できる手段として，人物の頭部に装着されたウェアラブルカメラを用いて記録可能な一人称視点映像を利用した手法を研究する．

一人称視点映像を用いた従来の頷き検出手法では，同動作の特徴として映像の大局的運動 (エゴモーション) が主に用いられてきた．これにより，前述の微細な動きや顔領域の隠蔽に関する問題は解決される．その一方で，動き情報のみを用いるアプローチでは，手元にある物体を注目する動作など，頷きに類似した動作を区別することが難しい．そこで本研究では，対象人物の一人称視点映像に加え，同人物の視線情報を利用するアプローチを提案する．本アプローチにより「会話中に頷きを行う際，人はしばしば相手の顔を見る」「手元にある物体に注目している場合は，その物体上に注視点が現れる」といった，頷きとその類似動作を区別するにあたって重要な特徴を利用することが可能となる．

提案手法の有効性を検証するため，本研究では 320 個の映像からなるデータセットを新たに構築し，頷き動作と類似動作の識別実験を行った．さらに，非統制環境下で長時間記録された映像からの頷き検出についても検討した．これらの実験の結果，頷きの認識において視線情報を利用することが有効であることが確認された．

# 目次

# 図目次

# 表目次

# 第1章

# 序論

## 1.1 背景と目的

頷きや首振り，首を傾げるといった頭部動作は，人と人の会話において重要な役割を果たしている [1]．例えば，会話において相手の話した内容に同意や理解を伝える際には，頷き動作がしばしば行われる．また，不同意の際には首振りが，疑問を表す際には首傾げが，それぞれ観測されることがある．このような頭部動作が会話においてどのように現れるかを解析することで，コミュニケーション支援に利用することができる．実際に従来研究では，会話中において誰が，誰に対して，どのように反応したかの認識 [2] や，面接におけるコミュニケーション能力の評価とフィードバック設計に頷き検出結果を利用する取り組み [3]，[4] が行われている．このような応用を実現するにあたっては，会話を何らかの手段で記録し，そこから自動的に頭部動作が認識できることが望ましい．また，コミュニケーション支援をリアルタイムで行う場合においても，頭部動作の自動認識は必須である．そこで本研究では，頭部動作のなかでも特に頷きに着目し，その自動認識手法を開発する．

いくつかの従来研究では，固定カメラを用いた頭部動作の認識が行われてきた [5]，[6]，[7]，[8]，[9]，[10]．これらの手法では，映像中の顔を追跡することによって認識を行っている．しかし，固定カメラによる手法では次のような課題がある．まず，会話中の頷き動作におけるバリエーションに対応ができない．例えば，

1. 相手の顔を見て大きく頷く動作
2. ものを見ながら小さく頷く動作

などの動作が観測されるが，頷き動作が小さい場合，映像中の顔の位置がほとんど変化しないため，検出が困難である．また，顔の前にあるものを見ている状態など，カメラに対して顔の一部が遮蔽されている場合も，頷きの検出は困難となる．さらに，屋内で向かい合って座っている場合など，固定カメラを利用できる環境は限定されている．

本研究では，このような課題に対処するため，頷き動作を行う人物が頭部に装着したウェアラブルカメラを用いて記録できる一人称視点映像の利用を検討する．一人称視点映像はカメラ装着者の頭部動作や手動作を認識するための有効手段としてコンピュータビジョン分野

で注目を集めており，近年多くの研究が発表されている [11]，[12]，[13]．また，同映像には会話相手が大きく映り込むという特性を利用し，グループ行動の解析に関する研究も行われている [14]，[15]，[16]．さらに，ウェアラブルカメラは固定カメラと異なり，利用できる環境が限定されないという利点もある．

このような特性を利用し，本研究では一人称視点映像を用いてカメラ装着人物の頷き動作を検出する手法を提案する．同人物が頷きを行うと，それにともなう頭部動作がカメラの運動として現れることが期待される．また，固定カメラを利用する場合と異なり頷きを行う人物を検出する必要がないため，手法の性能が人物検出や顔検出といった他の技術の性能に依存しないという利点もある．特に本研究では，一人称視点映像を用いた動作認識 [17] をさらに発展させ，頷きとその類似動作を区別することを試みる．

頷き認識の既存手法では，主に会話をするだけの状況が想定されてきた．たとえば，資料のあるミーティングや，道具のあるグループ作業などでは，自身の手元にある物体を注目するという動作も観測されうる．これらの動作は頭部運動という点では頷きと類似しているため，カメラ運動のみを利用する従来手法を用いて頷き認識を行うことは難しい．そこで，本研究では頭部運動に加え視線情報を用いて，このような頷きに類似した動作に影響を受けにくい認識を目指す．下にあるものを見る場合には，頷きの場合と比較すると，その物体に対して注視が発生しやすい．また，頷きの場合には，相手の顔を見ながら行っている場合が多いと考えられる．従って，視線情報を用いることで，注視が発生しているか，相手の顔を見ているかの情報を効果的に得られることが期待できる．

本研究の貢献は以下の通りである．

- 一人称視点映像における，映像記録者の視線を考慮した頷き動作認識手法を提案した．
- 提案手法の評価のために，視線情報が付与されている，頷き動作または類似動作のどちらかの行動が含まれた学習サンプルのデータセットを作成した．
- 作成したデータセットを用いて，頭部運動のみを用いた手法と視線情報を組み合わせた手法の比較実験を行った．その結果，視線情報を組み合わせた手法が識別に有効であることが確認できた．

## 1.2　論文の構成

本論文の構成は以下のようになっている．第 2 章では，本研究に関連した，映像からの頭部動作認識と，一人称視点映像解析における視線情報の利用についての先行研究を紹介する．第 3 章では，視線を考慮した頷き認識の手法を提案し，その詳細について説明する．第 4 章では，提案手法の評価のために行った実験と，その結果について説明する．第 5 章では，結論と展望について述べる．

# 第2章

# 関連研究

本研究では，一人称視点映像について，映像からの頭部動作認識と視線情報の利用を行っている．そこで，本章では，それぞれの要素における先行研究を紹介する．まず，2.1 節では映像からの頭部動作認識の研究を紹介する．次に，2.2 節で一人称視点映像解析における視線情報の利用に関する研究を紹介する．

## 2.1 映像からの頭部動作認識

本節では，映像を用いた頭部動作認識の研究と，頭部動作を用いた応用の研究について紹介する．

会話において，非言語コミュニケーションとして，頷き，首振り，首傾げなどの頭部動作が発生する．コミュニケーション解析において，これらの頭部動作を映像から自動で認識することは重要であり，さまざまなアプローチが行われてきた．

最初に，固定カメラの前で自発的に行った頭部動作を認識する研究が行われた．Kawato ら [5] は，人間の前に固定カメラを設置し，その顔映像から両目の間を追跡することで，頷きと首振りの検出を行っている．まず，顔映像から，離散フーリエ変換を用いて両目の間にある追跡可能な点を検出する．追跡した点のフレーム間での動きによって，“stable”，“transient”，“extreme”の 3 つの状態に分類する．この状態の遷移から頷きと首振りを検出する．Kapoor ら [6] は，瞳孔の追跡を行うことで，処理の簡略化を行っている．また，頷きと首振りの検出には，隠れマルコフモデル (HMM) [18] を用いている．Tan ら [7] は，カスケード分類器によって検出された目の座標を，Wei ら [8] は Kinect によって推定された頭の向きをそれぞれ用いて，HMM による頷きと首振りの検出を行った．

会話中の自然な頭部動作の認識も行われている．Nguyen ら [9] は，二人での会話における頷き動作の検出を行った．この研究では，それぞれの人の前に固定カメラを設置し，その固定カメラに映る顔映像中の目や口の周りといった特徴点を追跡することによって頷き動作の検出を行っている．追跡した点の座標の変化に対して短時間フーリエ変換を行い，SVM により学習することで，頷き動作か頷き動作でないかの識別を行った．また，Morency ら [19] は，ロボットやアバターとの会話における頷き検出を行っている．この研究では，条件付き確率

場 (CRF) に HMM における隠れ状態を付け加えた Latent-Dynamic Conditional Random Fields (LDCRF) が提案されている．この LDCRF を用いて，秋山ら [10] は，上方頷きや連続頷きなど 10 種類の頭部動作を検出する研究を行っている．この研究では，Histogram of Oriented Gradients (HoG) [20] を 1 次元時系列データに対して適用した，一階微分と二階微分の正負の勾配ヒストグラムをとる，Histogram of Velocity and Acceleration (HoVA) 特徴量が提案された．最近では，固定カメラによる手法以外にも，Terven らの研究 [21] のように，二人称視点映像を用いた頷きの検出も行われている．二人称視点映像とは，相手が装着したウェアラブルカメラによって記録された映像である．この二人称視点映像に映る顔の特徴点を追跡することによって，頷きの検出を行っている．

このように固定カメラを利用した手法では，映像中の顔から特徴点の検出を行い，それらを追跡することによって頭部動作の認識を行っている．そのため，以下のような課題がある．まず，会話中の頭部動作に対するバリエーションに対応ができない．例えば，頷き動作には，相手の顔を見て大きく頷く動作だけではなく，ものを見ながら小さく頷く動作も存在する．頷き動作が小さい場合，映像中の顔の位置がほとんど変化しないため，検出が困難である．また，顔の前にあるものを見ている状態など，カメラから顔が隠蔽される場合もある．このような場合，顔検出が困難となり，後段の認識手法の適用ができない．

これ以外にも，適用できる環境が限られているという課題もある．固定カメラを利用する場合は，屋内で座っている場合に範囲が限定され，人物が動く場合には適用できない．また，多人数のコミュニケーションにおける利用には適していない．多人数のコミュニケーションでは，話している人が変わると，その人の方を向く可能性が高い．そのため，顔の向きが正面ではなくなることが多く，顔の向きによっては固定カメラでの顔検出に失敗する可能性がある．顔検出が行えた場合でも，顔の向きごとに動作の学習を行う必要がある．

このような固定カメラによる手法の課題を解決するため，一人称視点映像を用いた頭部動作の認識手法が提案されている．ウェアラブルカメラを装着した人物が頷き動作を行った場合，その動作が小さくても頭部運動が映像の持つ大局的運動 (エゴモーション) として現れるという特徴がある．図 2.1 に，ウェアラブルカメラを装着した人物が頷き動作を行った場合の例を示す．この図では，Person B が頷き動作を行った場合の映像の変化を表している．左側の Person A による二人称視点映像では，顔の位置がほとんど変化していないのに対し，右側の Person B による一人称視点映像では，頷き動作が映像の動きとして現れていることがわかる．このようなエゴモーションを利用することにより，頭部動作の認識が可能になる．また，固定カメラを利用した手法と異なり，頭部動作を行っている人物の顔を検出する工程が必要ないため，顔が見ているものに対して隠れている場合や，顔の向きが変わる場合でも容易に適用可能である．

これまで，Yonetani ら [17] によって，エゴモーションを用いた頭部動作認識が行われている．この研究では，ウェアラブルカメラを装着した 2 人のコミュニケーションにおける，7 種類の動作の識別が行われている．7 種類の動作のうち，Positive(頷きなど)，Negative(首振り，首傾げなど) といった頭部動作によるものは，一人称視点映像の動きが重要であることが示されている．また，Poleg ら [12] によって，エゴモーションによる下を見る動作の検

出も行われている．

本研究では，一人称視点映像に加えて視線情報を用いた頷き動作の認識を行う．

映像から認識した頭部動作は，さまざまな形で応用が行われている．Hoque ら [3] と固定カメラを用いて頷きや首振りを検出し，表情や声などの情報と組み合わせて，被験者のコミュニケーション能力の評価とフィードバックを行うシステムを提案している．Naim ら [4] は，面接におけるコミュニケーションにおいて，これらの情報から面接官がつけたスコアの予測を行っている．また，非言語コミュニケーション以外の頭部動作も応用研究が行われている．Aran ら [22] は，固定カメラを用いて頭部運動を記録し，パーソナリティとの関係性評価を行っている．この研究では，頭部運動の標準偏差が外向性と相関があることが示されている．また，Hoshen ら [23] は，一人称視点映像の動きから，ウェアラブルカメラを装着している人を認識する研究を行った．Hoshen らは，歩く動作を行ったときの頭部運動の周波数スペクトルが人によって異なることを利用してウェアラブルカメラを装着している人の特定を行った．

## 2.2　一人称視点映像解析における視線情報の利用

近年，視線計測装置が小型化され，安価に手に入るようになった．そこで，視線情報を用いた研究が盛んに行われるようになってきている．本節では，一人称視点映像において視線情報を用いることで効果的に動作の認識や識別を行った研究を紹介する．

Fathi らの研究 [24] では，一人称視点映像における牛乳を注ぐなどの手を使う動作の認識を行っている．ここで，ウェアラブルカメラに装着された視線計測装置による視線情報を用いる．視線の座標付近の画像から特徴抽出を行うことで，動作認識の精度が向上することが示されている．Li らの研究 [25] では，一人称視点映像中の頭や手の動き，手の位置を用いて視線推定を行っている．推定した視線を用いて，物体のセグメンテーションや，行動認識を行っている．Xu ら [26] は，一人称視点映像の要約の研究を行っている．この研究では，入力映像をサブショット分割を行うが，映像中で注視が起こっている部分を検出し，サブショット分割に利用している．村上ら [27] は，3 人でのコミュニケーションにおいて，注目している相手の動作認識に対して視線情報を用いることの有効性を示している．一人称視点映像に，動作を行っているターゲットの人物と，別の人物の 2 人が映っている状況を考える．視線情報を用いることで，ターゲットの人物のみに絞って特徴抽出を行うことができる．Kera ら [28] は，グループ全員が，視線計測装置付きのウェアラブルカメラを装着した状態において，そのうち複数人が同じものを見ている状態の検出を行っている．一人称視点映像中に複数の物体が映っている状況が想定されている．視線情報を用いることで，その中のどの物体を見ているのかがわかる．これをグループ全員の一人称視点映像について行うことで，同時に複数人が同じ物体を見ている状態を検出する．

これらの研究に対し，本研究では，一人称視点映像における頷き動作と頷きと似た動作の識別について視線情報を用いる．このように視線情報を頷き認識に用いる試みはこれまでに報告されておらず，本研究における重要な新規性となっている。

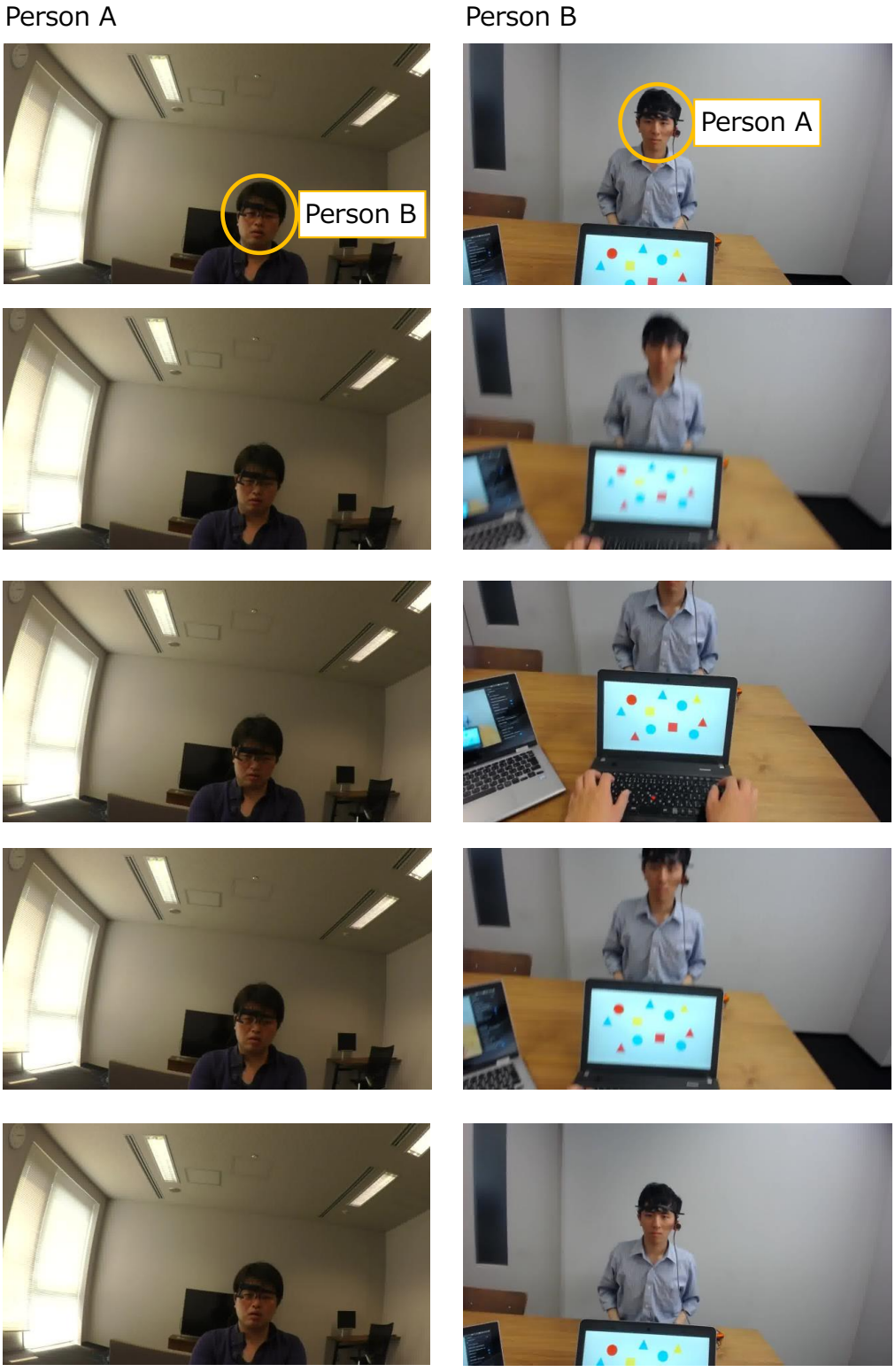


図 2.1: 頷きを行ったときの一人称視点映像の動き

# 第3章

# 提案手法

本章では，提案手法の詳細について説明する．図 3.1 に提案手法の概要を示す．本研究では，頷き動作または頷きと類似した動作 (以下，単純に類似動作と呼ぶ) のいずれかが収録されている一人称視点映像を識別することを目指す．これらの映像には，各フレームについて，ウェアラブルカメラを装着している人物の視線情報が，一人称視点映像中の注視点として与えられている．このような視線データを含む一人称視点映像を用いて頷き動作の認識を行う．

提案手法では，一人称視点映像中の頭部運動に加え，視線の動き，顔を見ているか，注視しているかの情報を用いる．具体的には，映像中に現れる会話相手の顔検出結果と視線情報を組み合わせることで，相手の顔を見ているかどうか，という情報を特徴として用いる．また，映像中における何らかの物体を注視しているかという情報も特徴として検討する．これらの特徴により相手の顔を見ながら頷く動作や，下にあるものに対して注視が発生している状態を効果的に認識できることが期待できる．

提案手法は以下の手続きにより構成される．まず，映像中の各フレームについて，頭部運動に加えて視線の動き，顔を見ているか，注視しているかの特徴を抽出し，それらを結合することでフレームの特徴量とする．次に，映像全体についてフレームごとの特徴量をまとめることで，全体の特徴量とする．具体的には，各フレームの特徴量をそのまま結合させた特徴量と，PoT (pooled time series) [29] による統計的な特徴量を生成して比較を行う．最終的に抽出された特徴をもとに，サポートベクターマシン (SVM) を用いて頷きと類似動作を識別する．

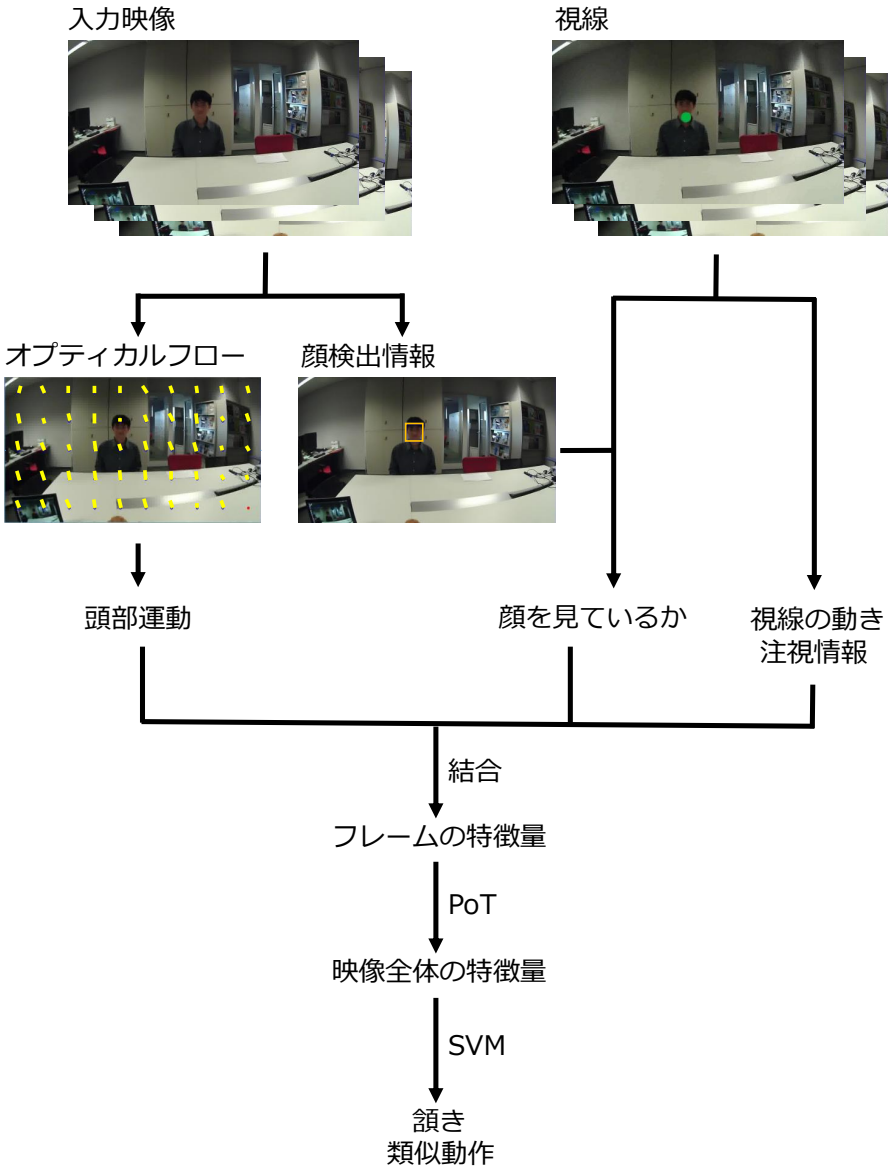
入力映像
視線
オプティカルフロー
顔検出情報
頭部運動
顔を見ているか
視線の動き
注視情報
結合
フレームの特徴量
PoT
映像全体の特徴量
SVM
頷き
類似動作

図 3.1: 提案手法の概要

図 3.2: 本研究における環境．右が頷きを行う人物，左が観測者である．

## 3.1　問題設定

本節では，本研究における状況設定について説明する．本研究では，複数人のコミュニケーションを想定している．ウェアラブルカメラを利用するため，多人数の場合や，移動を伴う場合でも適用可能である．提案手法では，会話を行うだけの状況に限らず，特に資料のあるミーティングや，道具のあるグループ作業など，手元に何らかの物体がある場合でも応用できることを目指している．図 3.2 のように，グループを構成する人物のうち，少なくとも頷き認識が適用される対象人物はウェアラブルカメラを装着しているという状況を想定する．また，同人物の視線データは，一人称視点映像中の 2 次元注視点の系列として与えられる．

本研究では，上のようにして与えられた一人称視点映像と視線データに対して，時間窓を用いて部分区間を抽出し，各区間について，頷きかそうでないかを認識する．特に本研究では視線を用いて頷きと類似動作を識別することを目指す．

## 3.2　1 フレームにおける特徴量の抽出

本節では，映像中の各フレームにおける特徴量抽出について述べる．本研究では，特徴量として，頭部運動，視線の動き，顔を見ているか，注視しているかの情報を用いる．まず，一人称視点映像から，オプティカルフローの抽出と顔検出を行う．これらに加えて，付与された視線情報を用いることで，それぞれの特徴量が抽出できる．それらを結合させることで，各フレームにおける特徴量を生成する．以下に，各特徴量の抽出の詳細について述べる．

### 3.2.1　頭部運動の抽出

一般に頷きは，一度頭が下に動き，上に上がって元に戻る動作，またはそれを複数回繰り返す動作であるため，頭部動作を正確に計測することが重要となる．本研究では一人称視点映像における動き情報を用いて，頭部動作を抽出する．例えば，頭部にウェアラブルカメラを装着している人物が頷き動作を行うと，カメラ自体も上下に移動する．その結果発生する一人称視点映像の大局的運動を，エゴモーションと呼ぶ．特に本研究では，Poleg らの研究 [12] で提案されているエゴモーションの抽出法を利用する．

具体的な手続きは以下の通りである．まず，勾配法 [30] の一種である，Lucas-Kanade 法 (LK 法) [31] を用いてオプティカルフローを計算する．Poleg らの手法では，映像中の各フレームを $W \times H$ のブロックに分割して，それぞれのブロックにおけるオプティカルフローを求めている．ここで，ブロック内の各画素のオプティカルフローは一定であると仮定している．本研究でも同様にして，各ブロックにおけるオプティカルフローを求める．このようにして求めた疎なオプティカルフローについて，全ブロックでの中央値をとることで 1 フレームにおけるオプティカルフローの代表値とする．背景など追跡できる点が存在しないブロックでは，オプティカルフローを正確に求めることができず，外れ値となることがある．中央値を利用することで，このようなノイズを軽減することができる．ここで，一人称視点映像の動きは，頭部運動と反対方向となるため，オプティカルフローの負の値をとることでエゴモーションを求める．本研究では，このようにして求めたエゴモーションを頭部運動を表す特徴量として用いる．本稿では，フレーム $t$ における頭部運動を $\boldsymbol{H}(t)$ と表し，頭が右上の方向に動く場合，すなわちオプティカルフローが左下方向の場合に正方向であるとする．値はフレームサイズにより正規化を行った．また，本研究の実験においては，Poleg らと同様に，$W = 10$，$H = 5$ を用いた．

図 3.3 に頷き動作と，下にあるものを見る動作におけるオプティカルフローの例を，図 3.4，3.5，3.6，3.7 に頭部運動の時系列データを示す．黄色の線は，各ブロックにおけるオプティカルフローの方向と大きさを表している．どちらの動作においても，開始点付近では上に移動し，終了点付近では下に移動しているため，頭部運動だけではこれらの動作の識別が困難であることがわかる．

### 3.2.2　視線の動きの抽出

3.2.1 節の図 3.3 からわかるように，頷き動作と，下にあるものを見る動作における頭の動きは類似している．しかしながら，これらの動作における視線の動きは異なると予想される．例えば，頷き動作では，置いてある物体の位置とは無関係に視線が上下に遷移するが，下を見る動作では，物体の位置で視線が一度停止すると考えられる．そこで，視線の動きの情報を用いることで，これらの動作を識別できることを目指す．各フレームにおいて，視線計測装置により，一人称視点映像中のどの座標を見ているかが記録されている．ここで，フレーム $t$ における視線の座標を $\boldsymbol{g}(t)$ とする．視線の動き $\boldsymbol{G}(t)$ は，対象のフレームとその前

のフレームにおける視線の座標の差分を用いることで，抽出できる．すなわち，

$$\boldsymbol{G}(t) = \boldsymbol{g}(t) - \boldsymbol{g}(t-1) \tag{3.1}$$

と表せる．視線が右上の方向に動く場合に正方向であるとする．値はフレームサイズにより正規化を行った．また，フレームにおける座標上での視線の動き $\boldsymbol{G}(t)$ の代わりに，頭部運動で補正を行った視線の動き $\boldsymbol{G}(t) + \boldsymbol{H}(t)$ を用いることが効果的であるかの検討も行う．$\boldsymbol{G}(t) + \boldsymbol{H}(t)$ を用いることで，頭の向きが変わっても実空間上で同じ場所を見ている場合は値が 0 になると考えられる．

図 3.8 に，3.2.1 節で示した 2 つの動作における視線の例を，図 3.9，3.10，3.11，3.12 にそれらの時系列データを示す．緑色の点は視線の座標を，赤色の線は視線の遷移を表している．頷き動作では，相手の顔付近を見続けているが，下にあるものを見る動作では，視線が顔の方向からラップトップの画面の方向に移動し，再び顔の方向に戻っていることがわかる．

### 3.2.3　顔を見ているかの情報の抽出

頷き動作を行う場合には，相手の顔を見ながら行っている場合が多いと考えられる．顔を見ているかの情報を用いることで，このような動作が効果的に認識できると期待できる．まず，各フレームにおいて，Viola-Jones 法 [32] により，顔検出を行う．これにより，顔が検出された場合は，顔のある座標と，顔領域の大きさが記録されている．顔を見ているかの情報は，顔の中心の座標と，視線の座標の距離を用いて抽出する．フレーム $t$ において，顔の中心の座標 $\boldsymbol{p}(t)$ と，視線の座標 $\boldsymbol{g}(t)$ の距離に対して，顔領域のバウンディングボックスの大きさ $s(t)$ で割ることで，正規化した距離 $d(t)$ を求めることができる．すなわち，

$$d(t) = \frac{||\boldsymbol{g}(t) - \boldsymbol{p}(t)||}{s(t)} \tag{3.2}$$

と表せる．これにより，顔領域の大きさに関係なく，相対的に顔からどれだけ離れた場所を見ているかがわかる．顔領域の $n$ 倍で正規化を行うことも可能であるが，本研究では $n = 1$ としている．特徴量 $L(t)$ は，この距離 $d(t)$ に対してガウス関数に似た逆数を利用する．すなわち，以下のように表せる．

$$L(t) = \exp(-\frac{d(t)^2}{2}) \tag{3.3}$$

これにより，顔の中心を見ている場合には値が 1 になり，顔から離れた場所を見ている場合には値が 0 に近くなる．なお，顔が検出されなかった場合には，$d(t) = \infty$，$L(t) = 0$ とする．

図 3.13，3.14 に，3.2.1 節で示した 2 つの動作における顔を見ているかの特徴量を時系列で表したグラフを示す．頷き動作では，動作は 8 フレーム目から 29 フレーム目で起こっている．下にあるものを見る動作では，動作は 13 フレーム目から 83 フレーム目で起こっている．頷き動作の場合では顔の近くを見て行っているため，値が高くなっている．逆に，下にあるものを見る動作では，顔から離れた場所を見ているため，値が 0 になっている．


図 3.3: オプティカルフローの例

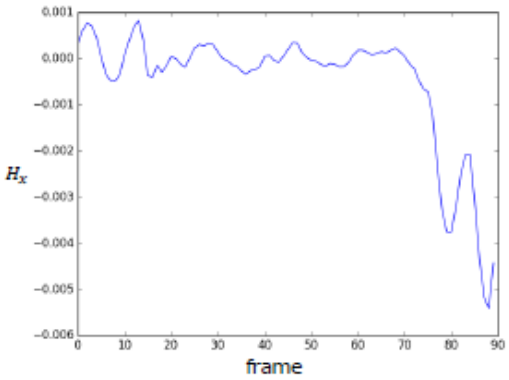


図 3.4: 頭部運動の x 成分 (頷き動作)

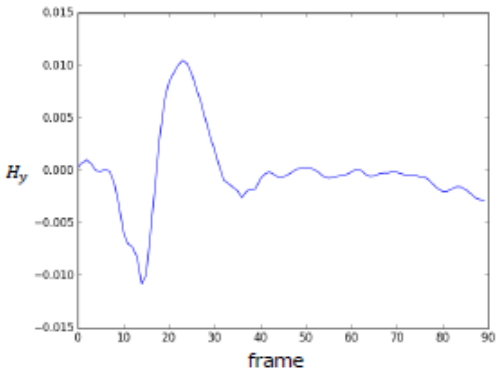


図 3.5: 頭部運動の y 成分 (頷き動作)

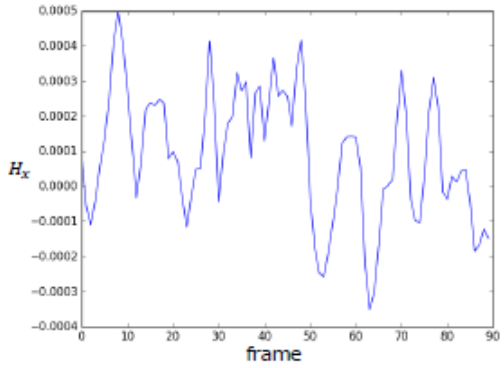


図 3.6: 頭部運動の x 成分 (下を見る動作)

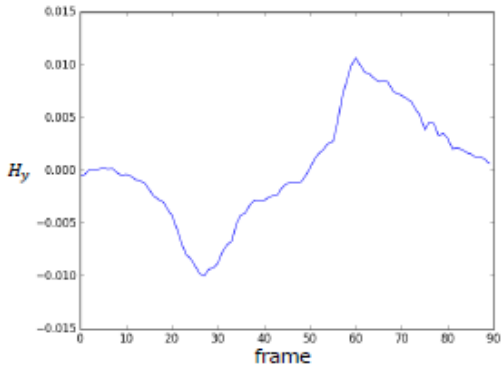


図 3.7: 頭部運動の y 成分 (下を見る動作)


図 3.8: 視線の例

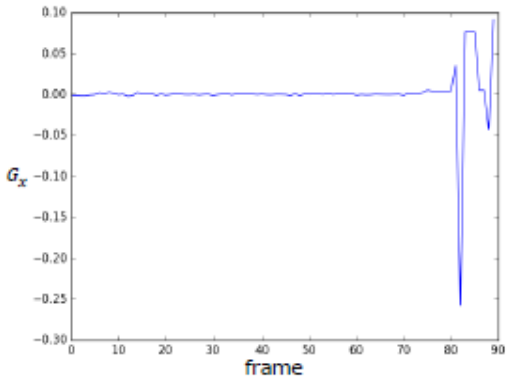


図 3.9: 視線の動きの x 成分 (頷き動作)

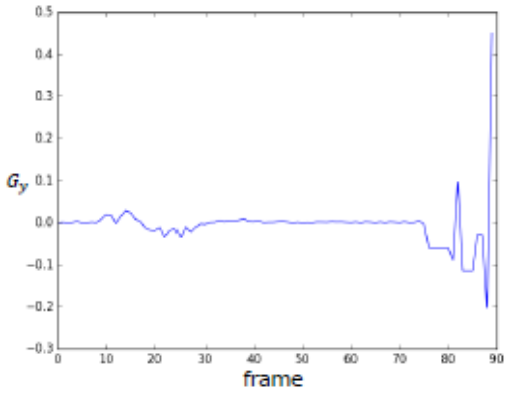


図 3.10: 視線の動きの y 成分 (頷き動作)

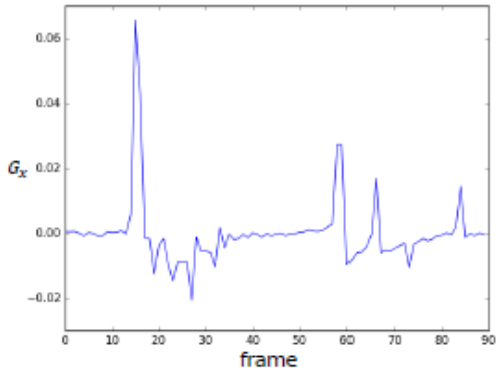


図 3.11: 視線の動きの x 成分 (下を見る動作)

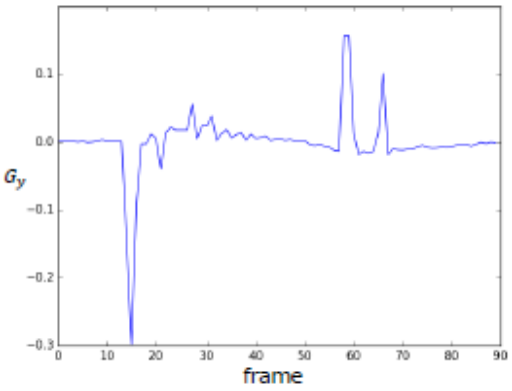


図 3.12: 視線の動きの y 成分 (下を見る動作)

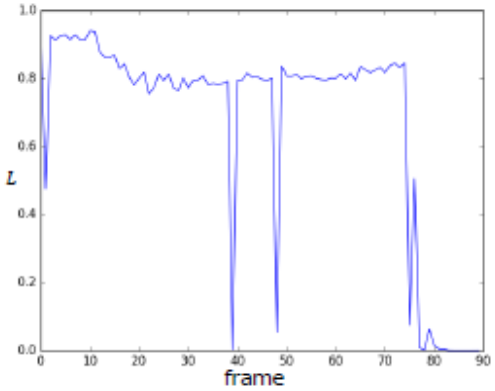


図 3.13: 顔を見ているか (頷き動作)

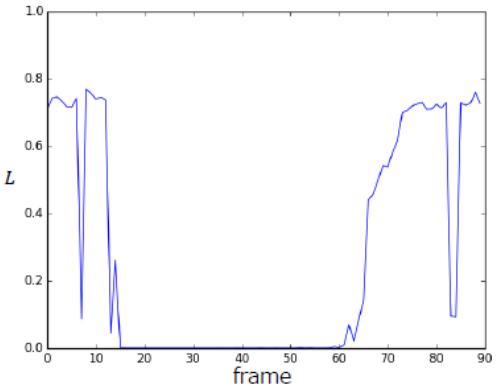


図 3.14: 顔を見ているか (下を見る動作)

### 3.2.4　注視しているかの情報の抽出

下にあるものを見る動作を行う場合，その物体に対して注視が発生するが，頷き動作を行う場合には発生しないと考えられる．そこで本研究では，対象物体を注視するような動作が発生しているかどうかという特徴を，視線の変化から抽出する．これにより，これら 2 つの動作の識別が効果的に行えると期待できる．注視しているかの判定は Salvucci and Goldberg[33] の I-DT (dispersion-threshold identication) に基づいて行う．I-DT では，一定時間 $\tau$ において，視線の変位 (dispersion) が一定以内であれば注視と定義する．すなわち，時刻 $t-\tau$ から時刻 $t$ までの視線の座標 $g$ について，変位 $D$ は，

$$D = \{\max(x) - \min(x)\} + \{\max(y) - \min(y)\} \tag{3.4}$$

と表せる．この変位 $D$ が閾値 $a$ 以下であれば，注視していることになる．ただし，閾値以下であるかで特徴量を生成した場合，$D$ が閾値付近の値をとった場合は，不安定になる．そこで本研究では，次の線形な関数を用いることで特徴量 $F(t)$ を生成する．

$$F(t) = \begin{cases} 1 - \frac{D(t)}{a} & (D(t) < a) \\ 0 & (otherwise) \end{cases}$$

これにより，同じ場所を見続けている場合には値が 1 になり，逆に，サッケードが発生している場合には値が 0 となる．

[33] の論文では，注視時間は通常 100ms 以上となるため，時間 $\tau$ は 100ms から 200ms の間で設定すればよいと記述されている．また，閾値 $a$ は，角度で表した場合に 0.5 度から 1 度の間で設定するのが適切であることも述べられている．本研究では，時間 $\tau$ は 100ms に設定した．また，閾値 $a$ は，0.02 とした．本研究で用いる Pupil Pro は，対角線の画角が 100 度程度であるので，1 度に相当する (x 成分と y 成分でそれぞれ 0.01) ．

入力映像は，30fps で収録を行っているため，時間 $\tau$ は 3 フレーム分となる．そこで，変位 $D$ は，

$$\begin{aligned} D(t) = &\{\max(0, G_x(t-1), G_x(t-1)+G_x(t)) - \min(0, G_x(t-1), G_x(t-1)+G_x(t))\} \\ &+ \{\max(0, G_y(t-1), G_y(t-1)+G_y(t)) - \min(0, G_y(t-1), G_y(t-1)+G_y(t))\} \end{aligned} \tag{3.5}$$

としても表せる．この $D(t)$ を用いて，

$$F(t) = \begin{cases} 1 - \frac{D(t)}{0.02} & (D(t) < 0.02) \\ 0 & (otherwise) \end{cases}$$

を求め，この $F(t)$ を，注視しているかの情報の特徴量とする．

図 3.15，3.16 に，3.2.1 節で示した 2 つの動作における顔を見ているかの特徴量を時系列で表したグラフを示す．頷き動作 (8 フレーム目から 29 フレーム目) の場合では顔の方向を見ているが，注視は行っていないため，低い値になっている．下にあるものを見る動作 (13

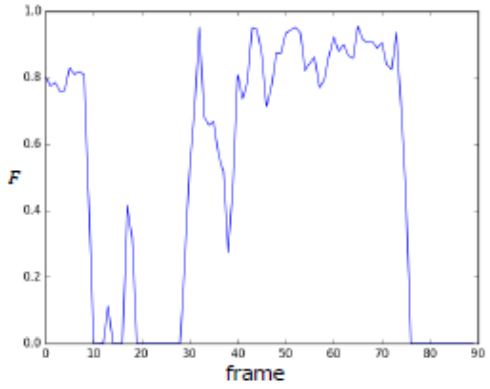


図 3.15: 注視しているか (頷き動作)

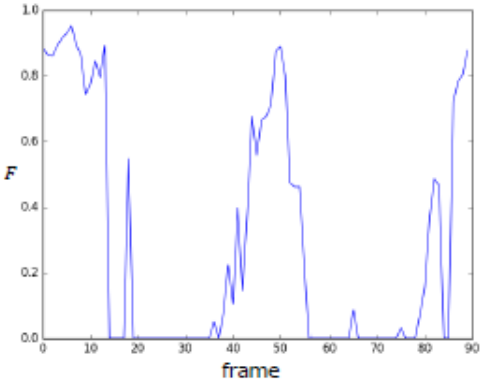


図 3.16: 注視しているか (下を見る動作)

フレーム目から 83 フレーム目) では，まず顔からラップトップの画面への視線の遷移が発生するため，値が 0 になる．次に，ラップトップの画面に対して注視が発生するため，値が 1 に近づく．その後，ラップトップの画面から顔に視線が戻るため，再び値が 0 になっている．

## 3.3　映像全体の特徴量の抽出

3.2 節で抽出した特徴量から映像全体の特徴量を生成する．これにより，識別器による動作の学習が可能になる．本研究では，各フレームの特徴量をそのまま結合させた特徴量のほかに，PoT (pooled time series) [29] を用いた特徴量生成も行う．

PoT を用いることで，映像中からどの時間の情報を用いるかという時間抽出と，統計的な特徴量の生成が可能になる．頷き動作は，発生している時間が短く，場合によっては連続して起こる．そこで，時間抽出を行うことで，映像中の一部の時間でしか発生していない頷きなどの動作を効果的に学習することができる．また，頷き動作は，人によって強度や持続時間が異なる．そこで，統計的な特徴量を生成することで，動作の違いに対してロバストになる．以下に，PoT の詳細について述べる．

PoT による特徴量の生成の概要を図 3.17 に示す．まず，フレーム $t$ における特徴ベクトルを $\boldsymbol{v}(t) \in \mathbb{R}^n$ とする．ここで，映像のフレーム数を $m$ とすると，特徴ベクトルの列 $\boldsymbol{v}(1), ..., \boldsymbol{v}(m)$ ができる．これらの特徴量に対して，$k$ 個の時間窓 $[t_1^s, t_1^e], ..., [t_k^s, t_k^e]$ を用いて時間抽出を行う．時間抽出された区間に対して後述する演算処理 $op$ を行うことで最終的な特徴量 $\boldsymbol{x}$ とする．

$$\boldsymbol{x} = (x_1^{op_1}[t_1^s, t_1^e], x_1^{op_2}[t_1^s, t_1^e], ..., x_n^{op_r}[t_k^s, t_k^e])^{\mathrm{T}} \tag{3.6}$$

演算処理 $op$ として，まず，最大値と和が挙げられている．

$$x_i^{max}[t^s, t^e] = \max\{v_i(t), t = t^s, \cdots, t^e\} \tag{3.7}$$

$$x_i^{\Sigma}[t^s, t^e] = \sum_{i=t^s}^{t^e} v_i(t)$$

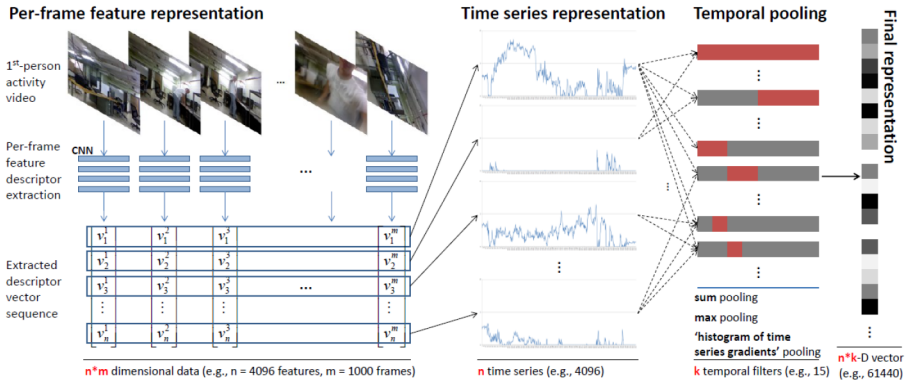


図 3.17: PoT による特徴量の生成 ([29] より引用)

また，勾配の正負の個数も挙げられている．

$$x_i^{\Delta^+}[t^s, t^e] = |\{t | v_i(t) - v_i(t-1) > 0 \wedge t^s \leq t \leq t^e\}| \tag{3.8}$$

$$x_i^{\Delta^-}[t^s, t^e] = |\{t | v_i(t) - v_i(t-1) < 0 \wedge t^s \leq t \leq t^e\}| \tag{3.9}$$

本研究では，これらの他に最小値

$$x_i^{min}[t^s, t^e] = \min\{v_i(t), t = t^s, \cdots, t^e\} \tag{3.10}$$

も用いる．

時間抽出に用いる時間窓は，全体 (1 種類)，3 分の 1 ずつ (3 種類)，9 分の 1 ずつ (9 種類) の計 13 種類とする．また，[29] の論文では，1 特徴量 1 区間に対して 1 つの演算処理を行っているが，本研究では，最大値，最小値，和，正の勾配の個数，負の勾配の個数という 5 種類の特徴すべてについて演算処理を行い特徴量を生成する．

## 3.4　実装

本節では，提案手法の実装の詳細やパラメータについて述べる．

本研究で用いる入力映像は，サイズを 1280 × 720px，フレームレートを 30fps として収録している．収録されたサンプル映像には，頷き動作あるいは類似動作が 1 つずつ含まれており，動作を行っている区間を含む 90 フレームまたは 45 フレームの学習用ラベルが付与されている．映像は 30fps で収録されているため，これらはそれぞれ 3 秒，1.5 秒に相当する．顔検出の精度を高めるため，サイズは変更せずに用いた．

提案手法では視線情報を用いた，頷き動作と類似動作の識別を行う．視線情報の収集には視線計測装置と一体になったウェアラブルカメラを用いる．これにより，各フレームにおけるウェアラブルカメラを装着している人物の視線の座標が得られる．ただし，視線推定に失敗し，視線情報が欠落したフレームも存在する．本研究では線形補間を行うことで欠損が発生しているフレームの視線情報を推定した．また，欠損が発生していなくてもノイズが含まれている場合もある．そこで，視線の座標を 3 フレームでスムージングした場合についても実験し，その効果を確かめる．

本研究では，特徴量として，頭部運動 (x 成分と y 成分)，視線の動き (x 成分と y 成分)，顔を見ているか，注視しているかの情報を用いる．すなわち，1 フレームあたりの特徴ベクトルは 6 次元となる．これらの特徴ベクトルを，結合させる，または PoT を用いることによって，映像全体の特徴量を生成して識別を行う．特徴ベクトルの次元は，そのまま結合させた特徴量が 540 次元 (90 フレーム) または 270 次元 (45 フレーム)，PoT を用いた特徴量が 390 次元となっている．識別は，線形 SVM を用いて行った．

# 第4章

# 実験

## 4.1 データセット

本研究では，一人称視点映像に加えて視線情報を用いることで頷き動作と類似動作の識別を行う．視線情報の付いた，頷き動作を行ったデータセットは存在しないため，今回新たに作成した．本データセットでは，被験者が視線計測装置の付いたウェアラブルカメラを装着し，相手に対して動作をする一人称視点映像を収録している．提案手法の評価のため，ものを持った状態や下に置いた状態での収録も行った．各映像サンプルについて，頷きまたは類似動作のどちらか 1 つの動作が含まれている．収録には，8 人の被験者が参加し，合計 320 のサンプル映像を作成した．

### 4.1.1 動作の種類

本研究で作成したデータセットでは，頷き動作と，類似動作の 2 種類の動作が含まれる．道具のある共同作業や，資料のあるミーティングでは，さまざまな頷き動作や類似動作が発生する．そこで，そのような動作に対応できるようにするため，表 4.1 に示す 4 種類ずつの動作について収録を行った．これらの動作の映像の例を図 4.2，4.3 に示す．

### 4.1.2 データセットの収録

本研究のデータセットでは，8 人の被験者の動作を，2 か所で収録した．1 人につき，頷き動作または類似動作の動作映像を 20 本ずつ，計 40 本の映像を作成した．収録は，あらかじめどちらの動作を行うか決めておき，1 本の映像の中で 1 つの動作を行うという形で行った．本データセットでは道具のある共同作業や資料のあるミーティングにも応用できることを目指している．道具のある共同作業では，被験者と相手が向き合って座る場合だけでなく，いずれかあるいは両方が立っている場合や，目の前に相手がいない場合も存在する．また，資料のあるミーティングでは，相手の顔を見ず，資料を見ながら頷く場合も考えられる．そこで，本データセットではそのような映像も収録を行った．視線付きの一人称視点映像の収録には Pupil Labs 社の Pupil Pro (図 4.1) を用いた．フレームレートは 30fps，フレームサイズは $1280 \times 720$px で収録している．データセット収録の様子を図 4.4 に示す．

表 4.1: 動作の種類

| | |
|---|---|
| 頷き動作 | 小さく 1 回頷く<br>小さく 2 回頷く<br>大きく 1 回頷く<br>大きく 2 回頷く |
| 類似動作 | ものを受け取る<br>ものを渡す<br>下にあるものを見る<br>下にあるものを顔を近づけて見る |

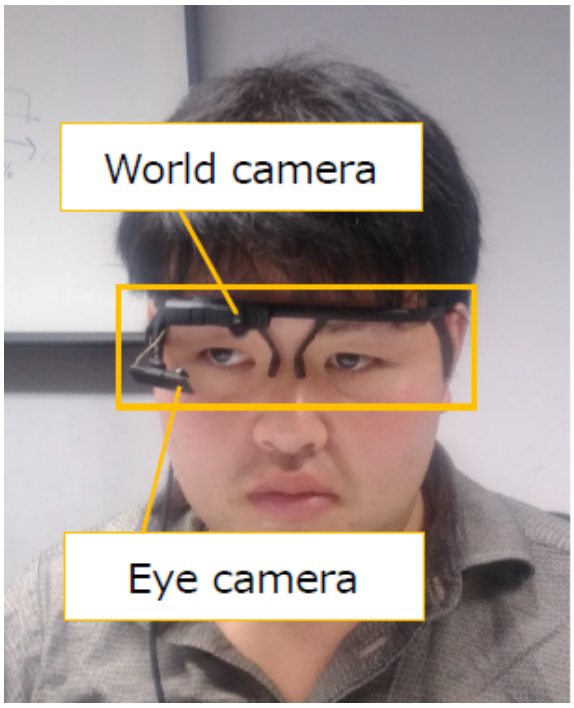


図 4.1: Pupil Pro．図に四角で示した部分がカメラになっている．

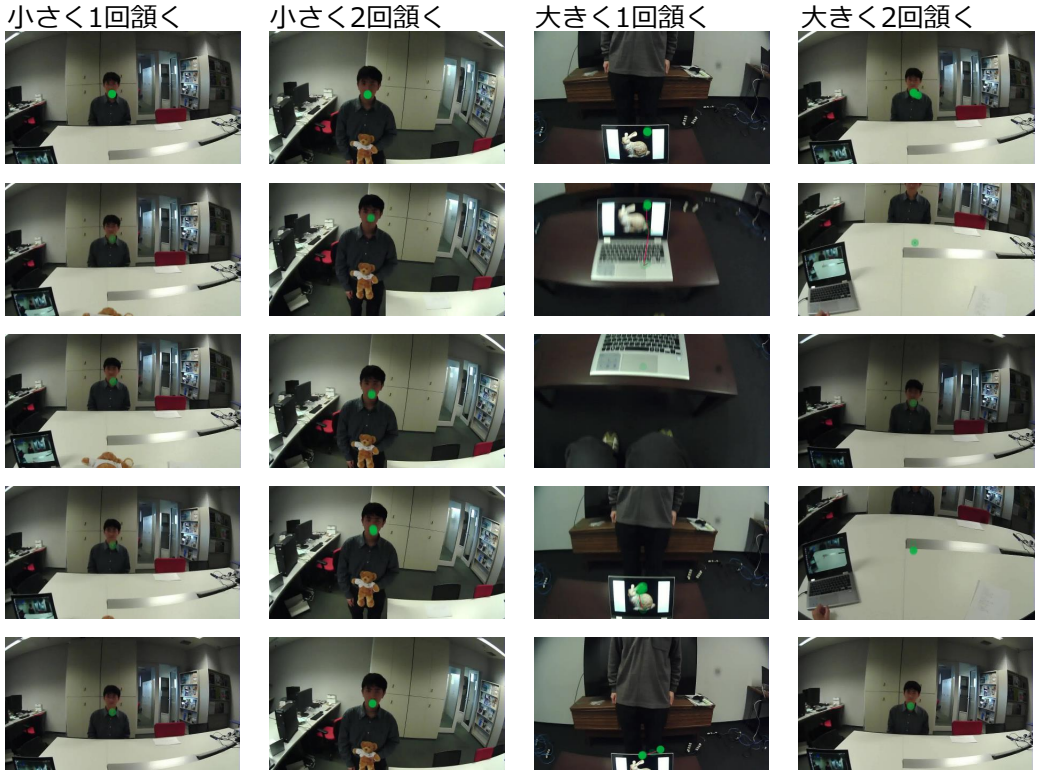


図 4.2: データセットの例 (頷き動作)


図 4.3: データセットの例 (類似動作)


図 4.4: データセット収録の様子

このようにして作成されたデータセットには，下にものがある場合や，被験者や相手がものを持っている場合も含まれている．これにより，頷きを行っているか，下にあるものを見ているのかの識別について，視線情報を用いることで認識精度が向上するのかの評価が可能になる．

### 4.1.3　学習データの作成

収録した各サンプル映像から，学習をさせるための情報を付与する．各映像には，頷き動作または類似動作が 1 つずつ含まれている．これらの映像に対して，動作が発生している時間が含まれるように学習に用いる開始点と終了点をつける．本データセットでは，ほぼすべての動作が 3 秒間で完結していたため，学習に用いるウィンドウサイズは 90 フレーム (3 秒間) とした．本データセットでは，動作の開始点を映像の開始点とすることはせず，動作が映像の中盤や後半に含まれているような学習データも作成した．これにより，動作が起こるタイミングによる過学習を防ぐことができ，動作が起こる時間の前後の状態遷移についても学習することができる．なお，動作の長さが 90 フレームを上回る場合には，動作の開始点の数フレーム前を映像の開始点に設定した．また，比較としてウィンドウサイズを 45 フレーム (1.5 秒間) にした学習データも作成した．

## 4.2　実験概要

本研究では，提案手法の評価のため，第 4.1 節で作成したデータセットを用いて，頷き動作か類似動作かの識別実験を行った．

提案手法では，一人称視点映像中の頭部運動に加え，視線の動き，顔を見ているか，注視しているかの情報を用いている．これらの視線に関する情報を用いることで，注視が発生しているか，相手の顔を見ているかの情報を効果的に得られることが期待できる．それぞれの特徴量追加に対する効果を評価するため，以下の 4 種類の手法で比較を行った．

- H: 頭部運動のみ (ベースライン)
- HG: 頭部運動+視線の動き
- HGL: 頭部運動+視線の動き+顔を見ているか
- HGLF: 頭部運動+視線の動き+顔を見ているか+注視しているか (提案手法)

8 人分のデータセットのうち 2 人分は，視線の欠損率が 40%以上だったため，提案手法の評価が困難であると判断し，本実験では除外した．残りの 6 人分のデータセットについて，1 人分をテストデータ，残りの 5 人分を学習データとして交差検証を行った．

## 4.3　実験結果

第 3 章で説明した学習条件のうち，ウインドウサイズ 90 フレーム，PoT あり，頭部運動補正なし，視線のスムージングありの場合が最も識別性能が高くなった．表 4.2 にその結果

を，図 4.5，4.6，4.7，4.8 にそれぞれの手法における ROC 曲線を示す．表の結果は 6 回のテストにおける平均値を示している．正解率はテスト映像の頷きか似た動作かの識別結果の正解率，AUC は ROC 曲線における下の部分の面積を表している．この条件では，提案手法がベースラインの識別性能を上回っている．また，注視しているかの情報を加えた場合が最も性能の向上が大きくなっていることがわかる．

この結果と，ウインドウサイズ 45 フレームの場合 (表 4.3)，PoT なしの場合 (表 4.4)，頭部運動補正ありの場合 (表 4.5)，視線のスムージングなしの場合 (表 4.6) をそれぞれ比較する．大まかな傾向として，ウインドウサイズ 45 フレームの場合と PoT なしの場合は，識別性能が下がっており，頭部運動補正ありの場合，視線のスムージングなしの場合は，識別性能の変化は小さいといえる．

表 4.2: 実験結果

| | 正解率 | AUC |
|---|---|---|
| H | 84.6% | 0.948 |
| HG | 86.3% | 0.958 |
| HGL | 87.1% | 0.954 |
| HGLF | 91.3% | 0.968 |

表 4.3: 45 フレームの場合の実験結果

| | 正解率 | AUC |
|---|---|---|
| H | 84.2% | 0.916 |
| HG | 87.9% | 0.933 |
| HGL | 87.1% | 0.924 |
| HGLF | 85.8% | 0.931 |

表 4.4: PoT なしの場合の実験結果

| | 正解率 | AUC |
|---|---|---|
| H | 75.4% | 0.815 |
| HG | 75.0% | 0.841 |
| HGL | 74.6% | 0.857 |
| HGLF | 81.3% | 0.917 |

表 4.5: 頭部運動補正ありの場合の実験結果

| | 正解率 | AUC |
|---|---|---|
| H | 84.6% | 0.948 |
| HG | 87.9% | 0.947 |
| HGL | 87.1% | 0.950 |
| HGLF | 87.5% | 0.952 |

表 4.6: 視線のスムージングなしの場合の実験結果

| | 正解率 | AUC |
|---|---|---|
| H | 84.6% | 0.948 |
| HG | 87.0% | 0.946 |
| HGL | 85.4% | 0.958 |
| HGLF | 90.4% | 0.966 |

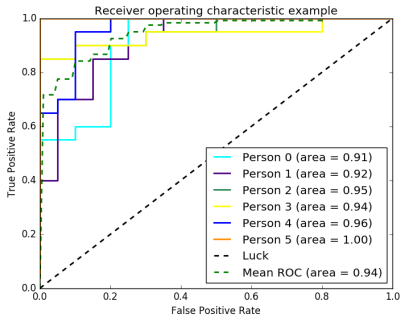


図 4.5: ROC 曲線 (H)

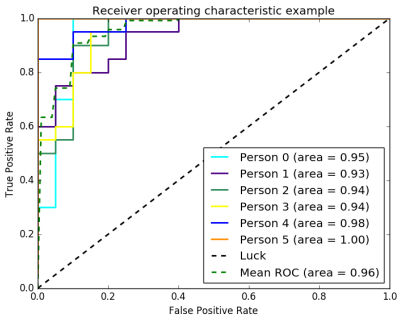


図 4.6: ROC 曲線 (HG)

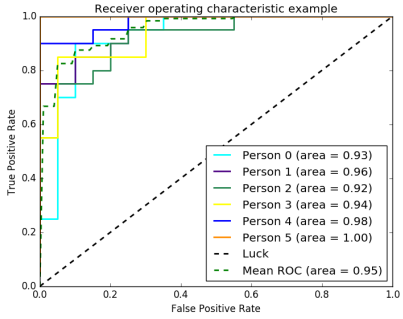


図 4.7: ROC 曲線 (HGL)

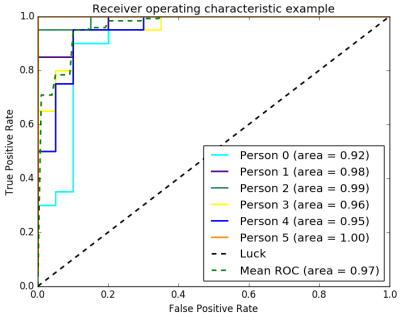


図 4.8: ROC 曲線 (HGLF)

## 4.4　考察

本実験では，H，HG，HGL，HGLF の 4 種類の手法を比較し，提案手法の評価と特徴量追加による効果の確認を行った．その結果，すべての条件において提案手法 (HGLF) がベースライン (H) の性能を上回っていることが確認できた．すなわち，視線情報を利用することは，頷きと類似動作の識別に対して有効であるといえる．

それぞれの特徴量追加を比較した場合，注視しているかの情報 (F) を追加した場合に最も識別性能が向上した．注視しているかの情報を利用することにより，頷き動作を行っていて下にあるものへの注視を行わない場合と，下にあるものに対して注視を行う場合の識別が効果的にできているといえる．逆に，顔を見ているかの情報 (L) を追加した効果は低かった．これは，相手が目の前にいない場合やラップトップや紙を見ながら頷くというように顔を見ていない状況も含まれており，頷き動作か類似動作かの 2 クラスの識別には効果的でないからだと考えられる．今回より細かいクラスの識別実験を行う場合には効果があると期待できる．例えば，ものを受け取る動作や渡す動作は必ず目の前に相手がいるため，相手の顔を見ているかの情報が重要になる．また，相手の顔を見ながら頷く動作を認識する場合もこの情報が必要であるといえる．

提案手法によって正しく認識できた例を図 4.9 に示す．左の図では，ラップトップを置いた状態で大きく 1 回頷く動作を行っている．視線の座標はラップトップの画面を指しているが，注視は行っていないため，ラップトップを見る動作ではないと認識されている．右の図では，ぬいぐるみを見て受け取る動作を行っている．ぬいぐるみに対して，注視が発生しているため，ぬいぐるみを見る動作であると認識できた．

提案手法でも認識ができなかった例を図 4.10 に示す．左の図では，ぬいぐるみを置いた状態で大きく 1 回頷く動作を行っている．この動作では，頭が下を向いているときに，服に対して注視が発生していると判定された．その結果，服を見る動作であると誤って認識された．右の図では，ラップトップを見る動作を行っている．この動作では，ラップトップに対して注視が発生していると判定されたにもかかわらず，頷きと認識されている．これは，顔とラップトップの間という大きな頭や視線の動きを伴った動作であることが理由として挙げられる．このように，注視が発生している時間に対して，頭や視線の動きが大き過ぎると，頷き動作と誤認識されると考えられる．

識別結果の詳細を図 4.11，4.12 に示す．頭部動作のみの場合と提案手法を比較すると，1 回頷く場合の精度が向上している．視線情報を用いることで，1 回頷く動作と類似動作との識別が効果的に行えた．また，ものの受け渡しを行う場合も精度の向上が見られた．これらの動作は，比較的長い時間注視が発生するため，注視情報の効果が高かったといえる．下にあるものを見る動作は，視線情報を用いることで，注視時間に対して頭部動作が比較的小さい場合は認識が行えたが，頭部動作が大きい場合は頷きと誤って認識された．その結果，総合的には精度の向上が見られなかった．2 回頷く動作は，類似動作に 2 回頭が上下する運動が存在しないため，また，下にあるものを顔を近づけて見る動作は，頭部動作が大きく，動作時間も長かったため，頭部動作のみでも認識できていた．

本実験では，学習条件の変化によって結果が変わるかについても検討を行った．ウインドウサイズ，PoT の有無，頭部運動補正の有無，視線のスムージングの有無についてそれぞれ考察する．

ウインドウサイズは，90 フレームの場合と 45 フレームの場合で比較を行った．本実験では，90 フレームで行った場合に高い性能が得られた．今回のデータセットでは，動作の時間が 45 フレームを超えるものが多かった．45 フレームのウインドウでは，動作の開始点から動作の途中までの情報しか得られなかったため，性能が得られなかった．逆に，90 フレームのウインドウでは，動作の前後まで含めて情報を得られたため，止まっている状態から動作をする状態への遷移，またはその逆についても学習が行えているといえる．

PoT なしの場合は，PoT ありの場合と比較して性能が低下した．この条件は，他の条件と比較すると，最も性能の低下が大きかった．頷き動作だけでも小さな頷きと大きな頷きが存在し，人によってもその強度には差がある．各フレームでの変位などをそのまま用いた場合，このようなバリエーションに対応するのが困難であるといえる．また，本データセットでは，動作の発生するタイミングを統一していない．フレームの特徴量をそのまま用いた場合は対応できなかったのに対し，PoT を用いた場合は時間抽出の効果によってどのタイミングで動作が行われても認識が可能であった．

頭部運動補正を行った場合では，視線の動きを加えた場合 (HG) は正解率が高くなった．頭部運動補正を行うことで見ている場所が空間上どれだけ変化したか表せていると考えられる．逆に，すべての特徴を用いた場合 (HGLF) では，性能が低下している．頭の向きが変化すると視線推定に誤差が発生する場合がある．例えば，図 4.13 の左の場合では，正面を見続けているにもかかわらず，下を向くと視線座標が下に移動している．また，右に示す，ガムテープの同じ点を注視し続ける場合も，下を向くと実際より下の座標で推定されている．このような誤差のため，正しく頭部動作による補正が行えていないといえる．また，注視の閾値を視線の座標の動きの場合と同じ小さい値に設定しているため，注視部分が注視ではないと判定されていると考えられる．

視線のスムージングの有無では，ほとんど違いがみられなかった．長時間にわたりノイズが含まれている区間では，3 フレームでスムージングを行ってもノイズを除去できない．また，正しくサッケードが取れている区間についてスムージングを行うと顔を見ているかや注視の判定で影響がでると考えられる．

本実験の結果を総合すると，頷き動作と類似動作の識別には視線情報を用いるのが効果的であるといえる．また，ウインドウサイズは動作の平均長より長い時間に設定し，特徴量は PoT を用いるとよいことがわかった．

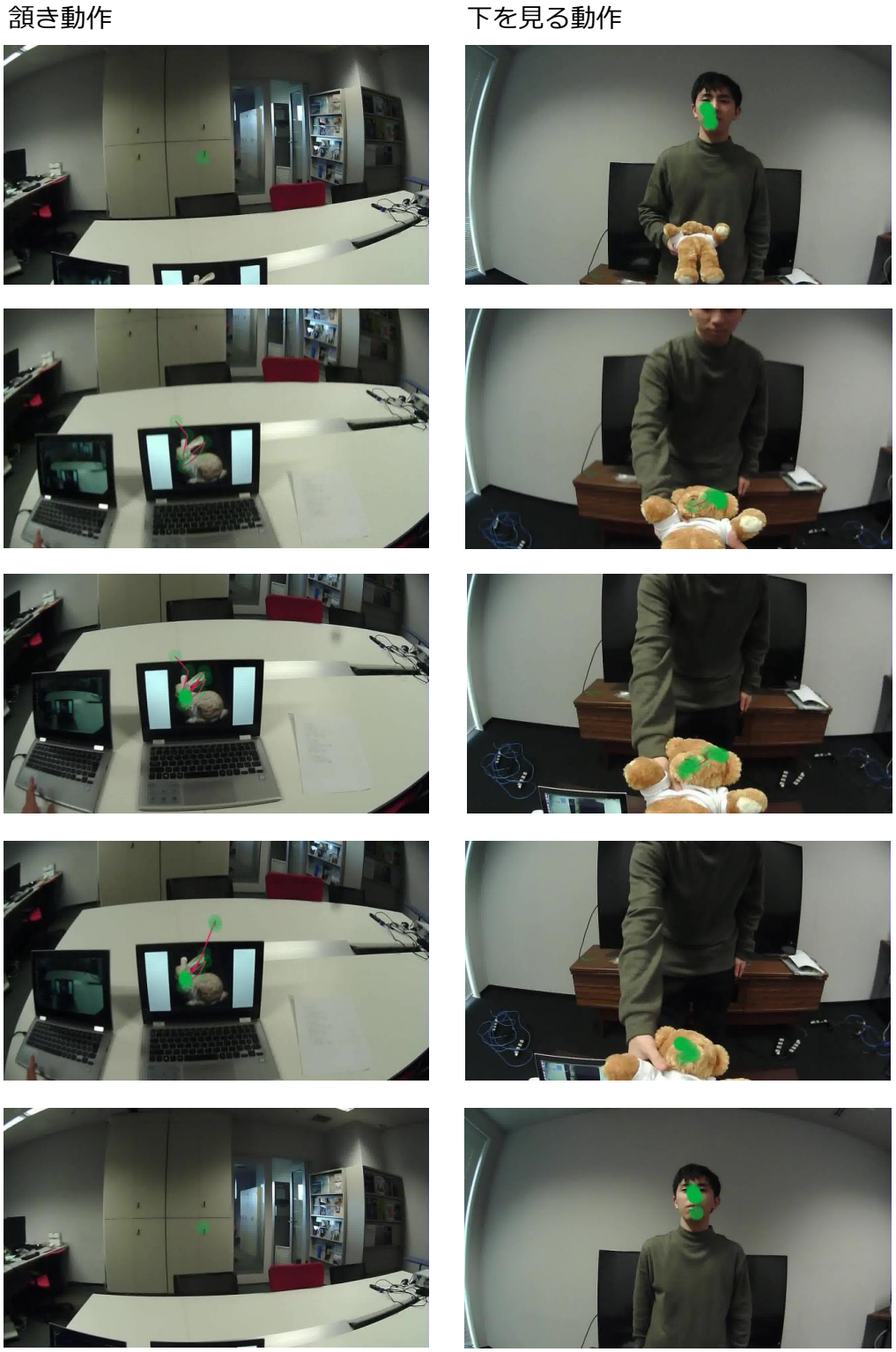


図 4.9: 正しく認識できた例

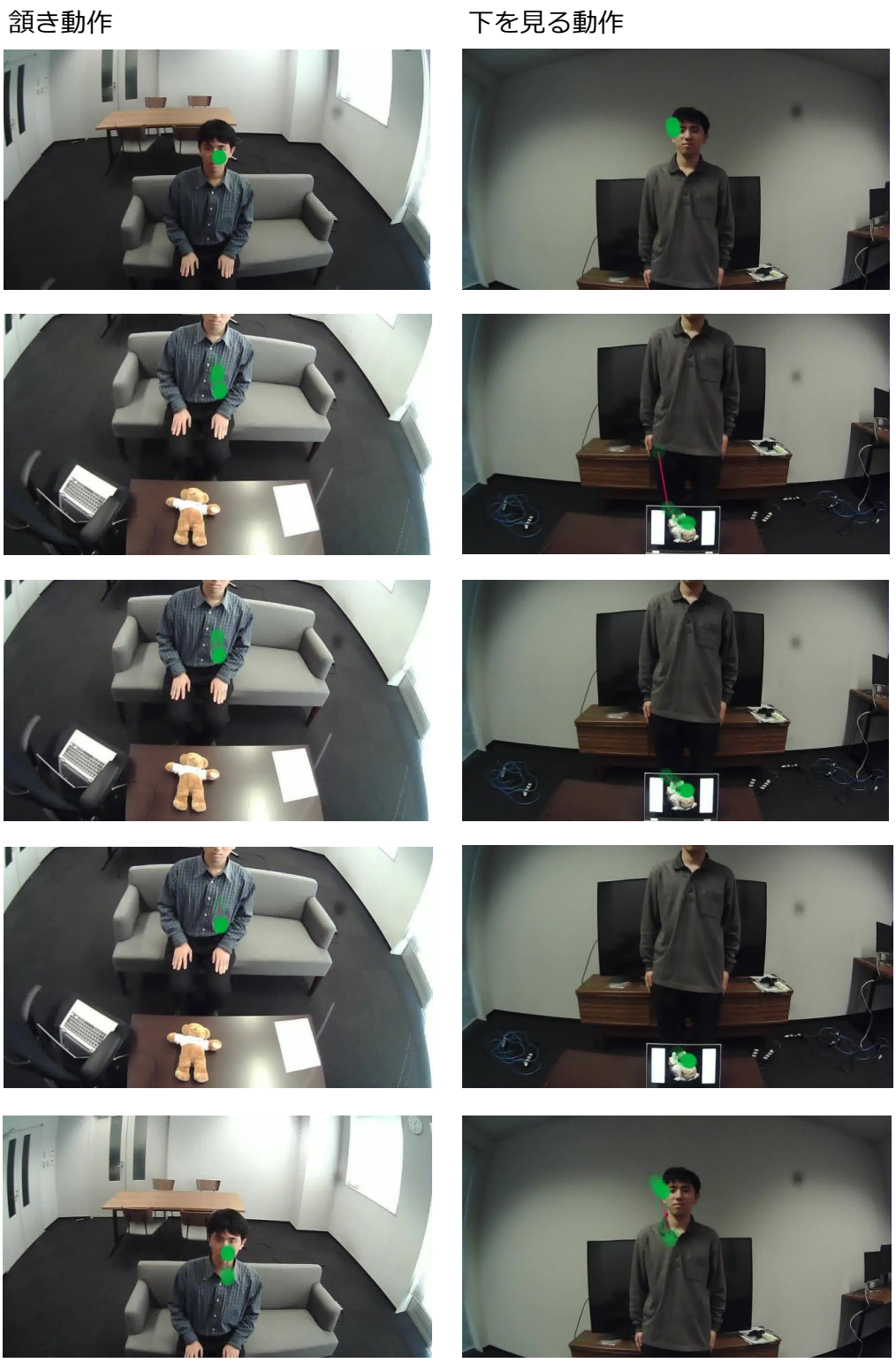


図 4.10: 誤って認識された例

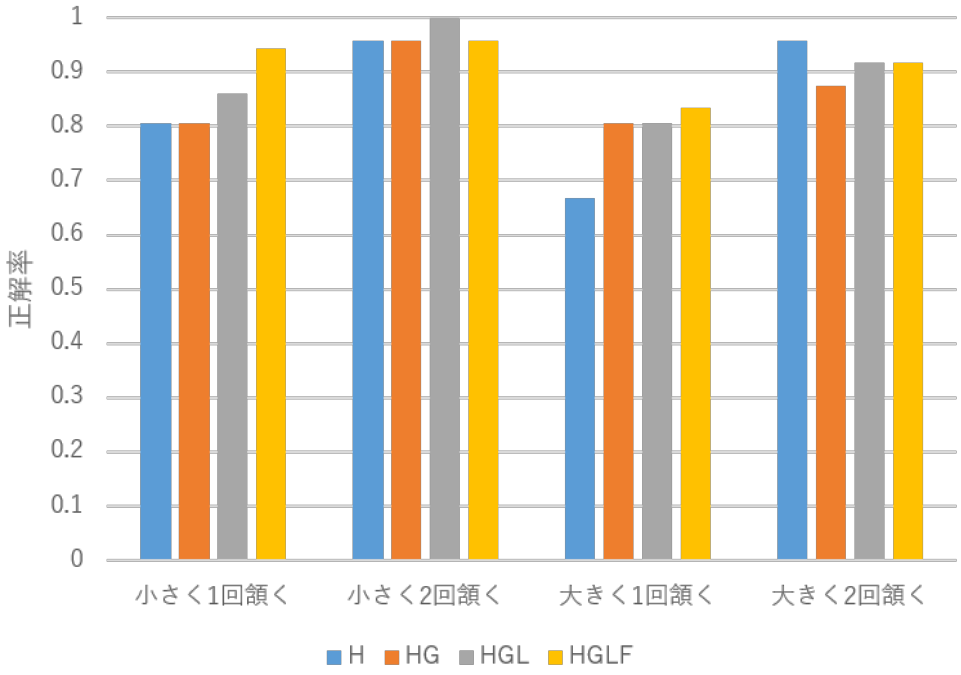


図 4.11: 識別結果の詳細 (頷き)

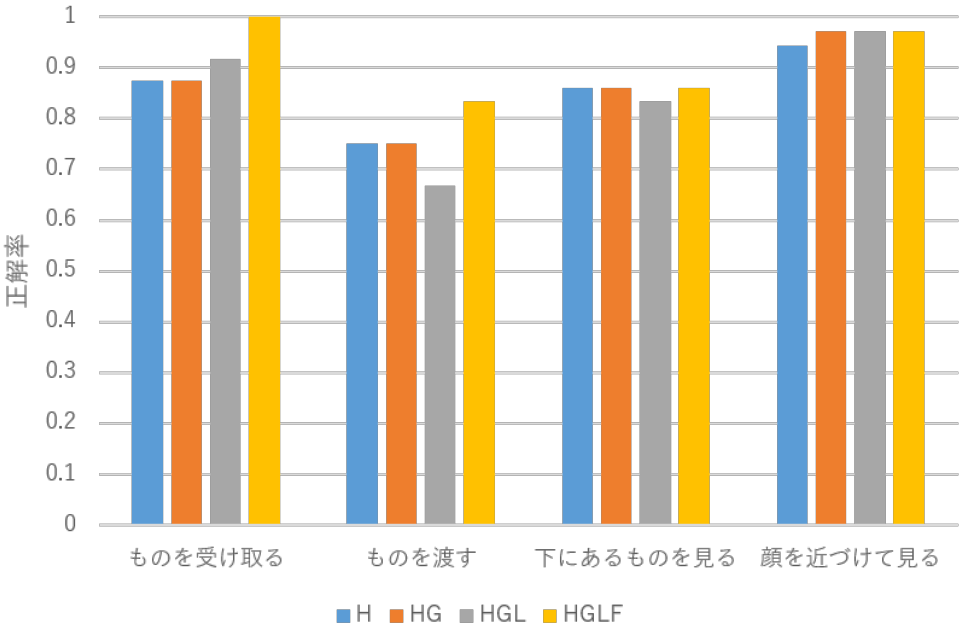


図 4.12: 識別結果の詳細 (類似動作)

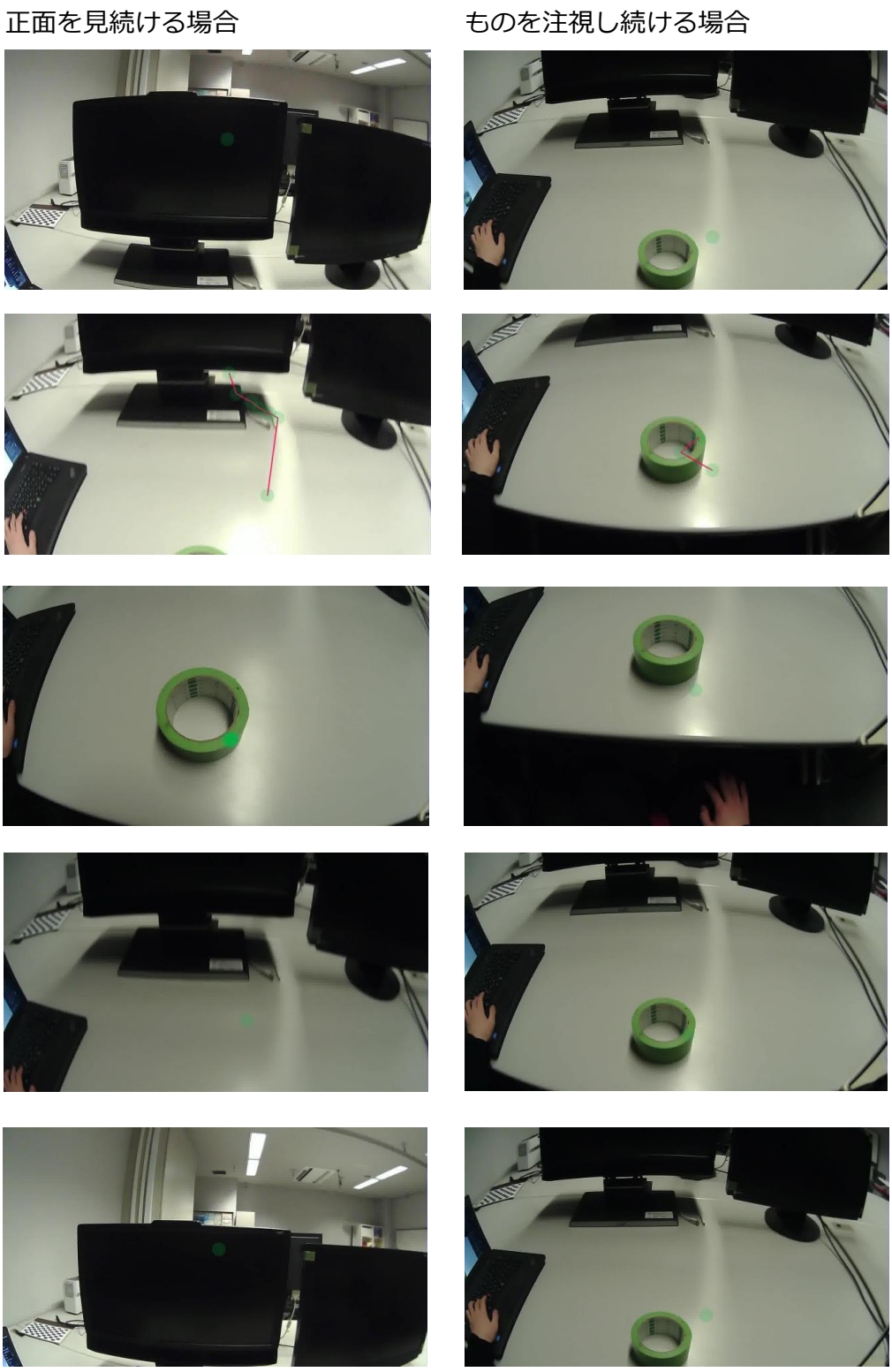


図 4.13: 頭の向きによる視線座標のずれ

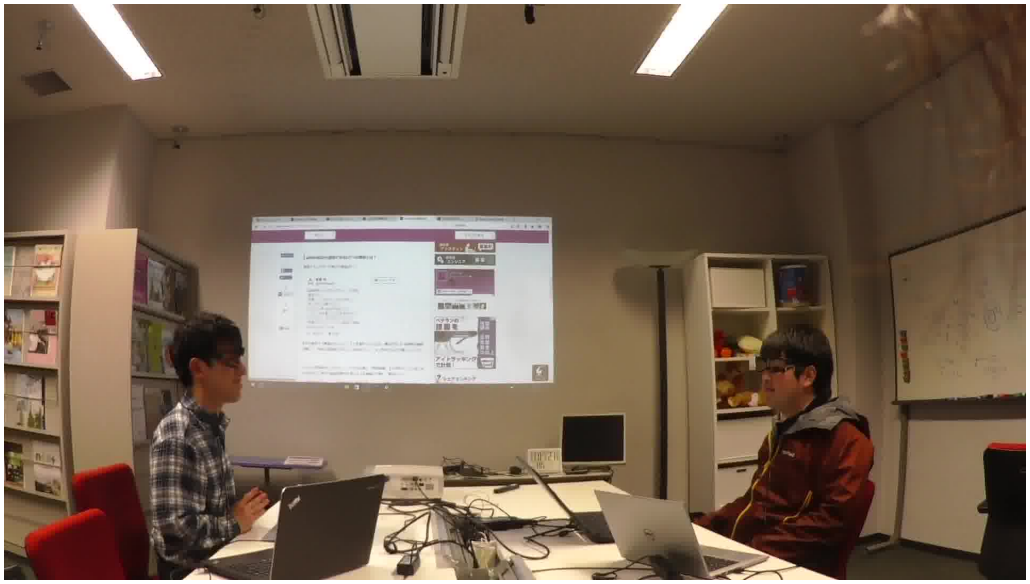

図 4.14: 会話映像の収録風景

## 4.5　検出実験

識別実験を応用し，非統制環境下で長時間記録された映像からの頷き検出についても検討を行った．検出に用いる映像として，2 人での会話の映像を記録した．下を見る動作も観測されるようにするため，図 4.14 のようにラップトップを置き，資料のあるミーティングを想定した．5 分程度のセッションを 4 セッション行い，合計 8 人分の映像を記録した．学習は，識別実験で最も性能が高かった，ウインドウサイズ 90 フレーム，PoT あり，頭部運動補正なし，視線のスムージングありの条件で行った．識別実験で用いたデータセットに，横を向く動作や静止した状態 120 サンプルを負事例として追加して，学習に用いるデータを作成した．検出実験はスライディングウインドウを用いる．ウインドウのステップ幅は 30 フレーム (1 秒間) として，それぞれの記録された映像のうち，4 分 30 秒ずつ，270 ウインドウについて頷きかそうでないかの識別を行う．正解ラベルは，90 フレームのうち，何フレーム以上頷き区間が含まれているかによって付与した．本実験ではその閾値として，18，27，36，45 フレームを用いた．

実験結果の ROC 曲線を図 4.15，4.16，4.17，4.18 に示す．どの閾値で行った場合でも，AUC が 0.6 程度にとどまる結果となった．原因として，まず，識別実験のデータセットを用いたことで過学習が起こっていることが考えられる．識別実験のデータセットでは，頷きか類似動作のどちらかを行うよう指示したため，識別が行いやすいデータセットになっている可能性がある．そのため，自然な会話中での動作には対応ができなかった．また，類似動作と横を向く動作や静止した状態をまとめて負事例としたことで学習が困難になったことも挙げられる．改善策として，頭部動作のみで，頷きと類似動作をまとめて，学習と検出を行い，視線情報を付与してこれらの動作の識別をすることで効果的に頷き検出が行えると考えられる．

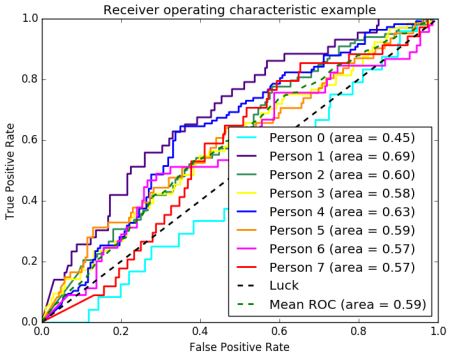


図 4.15: 検出結果 (閾値 18 フレーム)

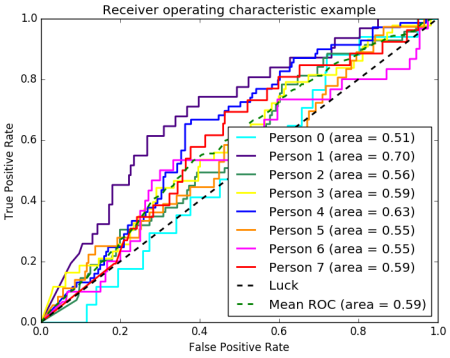


図 4.16: 検出結果 (閾値 27 フレーム)

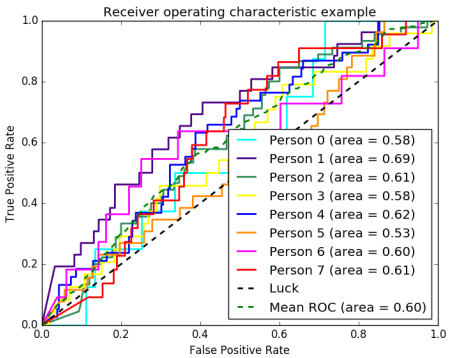


図 4.17: 検出結果 (閾値 36 フレーム)

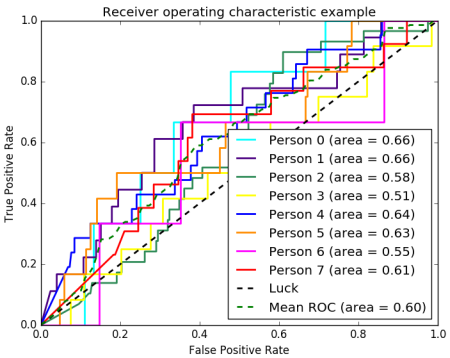


図 4.18: 検出結果 (閾値 45 フレーム)

# 第5章

# 結論

## 5.1 結論

本論文では，従来の固定カメラによる頷き検出手法の課題を説明し，一人称視点映像を用いることの意義について説明した．また，一人称視点映像から抽出できる頭部運動を用いた場合でも，頷き動作と下を見る動作の識別が困難であることを示した．その解決策として，頭部運動に加えて視線情報を用いた認識手法を提案した．視線情報として，視線の動き，顔を見ているか，注視しているかの情報を用いた．

また，頷き動作またはそれに類似した動作を行った，視線情報付きの一人称視点映像のデータセットを作成することで，提案手法の評価を可能にした．データセットは，資料のあるディスカッションや道具のある共同作業を想定し，下にものがある場合での収録も行った．これにより，頷きをしているか下にあるものを見ているか，従来の手法では識別が困難な状況を作成した．そして，作成したデータセットについて，識別実験を行うことで，視線情報を用いることの有効性を確認した．

## 5.2 課題と展望

本研究では，注視情報のパラメータとして，時間と変位の閾値を固定している．また，ウインドウサイズについても 2 種類のみで実験を行った．パラメータを変更することで結果が変わるのかや，適切なウインドウサイズについても検討する必要がある．

今回の実験では，注視が行われていても頷きと誤認識される動作が存在した．注視時間に対して頭や視線の動きが大きすぎる場合には頷きと認識されると考えられる．改善策として，注視している時間も含めて学習させることや，注視に重みをつけることが挙げられる．

また，本データセットでは，頷きか類似動作のどちらかを行うよう指示したため，識別が行いやすいデータセットになっている可能性がある．そのため，検出実験には対応ができなかった．今後の課題として，自然な会話中での動作について適用できるか検討する必要がある．また，横を向く動作が含まれる場合や，多人数で行う場合でも実験を行う必要がある．

さらに，実験では視線のスムージングの効果が見られなかった．全体を 3 フレームでスムージングした場合，長時間にわたりノイズが含まれている区間ではノイズを除去できない．

また，正しくサッケードが取れている区間についてスムージングを行うと顔を見ているかや注視の判定で影響がでる．そこで，長時間にわたりノイズが含まれている区間を検出し，その部分は長いフレームでスムージングを行うことが考えられる．また，サッケードが起こる区間について，視線推定の確信度が高い場合にはスムージングを行わないようにする必要がある．

本研究では，頷きか類似した動作かの 2 クラスの識別について実験を行った．今後の展望として，シーン情報を用いたより詳細な動作認識が考えられる．例えば，相手の顔を見て大きく頷いているか，あるいは，話とは関係のないものを見て小さく頷いているかという動作の認識が挙げられる．前者は，理解している状態であり，後者は，理解しているかわからない状態であると推定できる．また，資料のあるミーティングの場合には，資料を見ながら頷くことも考えられる．シーン情報を用いることで，ほかの人と同じ資料を見ているのか認識し，理解しているかの推定に応用することができる．また，このような推定に対して，話している人にフィードバックするシステムを設計することで，理解しているかわからない人に対して，自分の話していることを理解しているか確認するということも可能になる．

このような頷き認識の応用によって，コミュニケーションの解析や支援に役立てることができると見込まれる．

# 参考文献


[1] Senko K Maynard. Interactional functions of a nonverbal sign head movement in japanese dyadic casual conversation. *Journal of Pragmatics*, Vol. 11, No. 5, pp. 589–606, 1987.

[2] Kazuhiro Otsuka, Hiroshi Sawada, and Junji Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances. In *Proc. International Conference on Multimodal Interfaces*, pp. 255–262, 2007.

[3] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. Mach: My automated conversation coach. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 697–706, 2013.

[4] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG'15)*, Vol. 1, pp. 1–6, 2015.

[5] Shinjiro Kawato and Jun Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the" between-eyes". In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pp. 40–45, 2000.

[6] Ashish Kapoor and Rosalind W Picard. A real-time head nod and shake detector. In *Proc. Workshop on Perceptive User Interfaces*, pp. 1–5, 2001.

[7] Wenzhao Tan and Gang Rong. A real-time head nod and shake detector using hmms. *Expert Systems with Applications*, Vol. 25, No. 3, pp. 461–466, 2003.

[8] Haolin Wei, Patricia Scanlon, Yingbo Li, David S Monaghan, and Noel E O'Connor. Real-time head nod and shake detection for continuous human affect recognition. In *Proc. IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pp. 1–4, 2013.

[9] Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. Using self-context

for multimodal detection of head nods in face-to-face interactions. In *Proc. ACM International Conference on Multimodal Interaction*, pp. 289–292, 2012.

[10] 秋山解, 伍洋ほか. 自然会話における頭部動作検出. 研究報告コンピュータビジョンとイメージメディア (CVIM), Vol. 2016, No. 35, pp. 1–8, 2016.

[11] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1894–1903, 2016.

[12] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2537–2544, 2014.

[13] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact cnn for indexing egocentric videos. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[14] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proc. ACM Conference on Ubiquitous Computing*, pp. 699–704, 2012.

[15] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. Detecting bids for eye contact using a wearable camera. In *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG'15)*, Vol. 1, pp. 1–8, 2015.

[16] Shiro Kumano, Kazuhiro Otsuka, Ryo Ishii, and Junji Yamato. Collective first-person vision for automatic gaze analysis in multiparty conversations. *IEEE Transactions on Multimedia*, Vol. 19, No. 1, pp. 107–122, 2017.

[17] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2629–2638, 2016.

[18] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.

[19] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[20] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893, 2005.

[21] Juan R Terven, Bogdan Raducanu, María Elena Meza-de Luna, and Joaquín Salas. Head-gestures mirroring detection in dyadic social interactions with computer vision-based wearable devices. *Neurocomputing*, Vol. 175, pp. 866–876, 2016.

[22] Oya Aran and Daniel Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proc. ACM on International Conference on Multimodal Interaction*, pp. 11–18, 2013.

[23] Yedid Hoshen and Shmuel Peleg. An egocentric look at video photographer identity. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4284–4292, 2016.

[24] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pp. 314–327. Springer, 2012.

[25] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proc. IEEE International Conference on Computer Vision*, pp. 3216–3223, 2013.

[26] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2235–2244, 2015.

[27] 村上晋太郎, 米谷竜, 佐藤洋一ほか. 視線を利用した二人称視点動作認識. 研究報告コンピュータビジョンとイメージメディア (CVIM), Vol. 2016, No. 32, pp. 1–8, 2016.

[28] Hiroshi Kera, Ryo Yonetani, Keita Higuchi, and Yoichi Sato. Discovering objects of joint attention via first-person sensing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7–15, 2016.

[29] Michael S Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 896–904, 2015.

[30] Olivier Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT press, 1993.

[31] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, Vol. 81, pp. 674–679, 1981.

[32] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. I–I, 2001.

[33] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proc. Symposium on Eye Tracking Research & Applications*, pp. 71–78, 2000.

# 発表文献

中野雄介, 米谷竜, 樋口啓太, 佐藤洋一. 視線を考慮した一人称視点映像からの頷き検出. 電子情報通信学会総合大会, 2017 (発表予定) .

本研究に含まれない文献

T. Yamasaki, Y. Nakano, and K. Aizawa. A prediction model on 3D model compression and its printed quality based on subjective study. ACM SIGGRAPH 2015 Posters, 2015.
山崎俊彦, 中野雄介, 相澤清晴. 主観評価に基づく 3 次元モデルの圧縮品質とプリント品質の関係性予測. 映像情報メディア学会誌, 2015.

# 謝辞

佐藤洋一研究室での 2 年間は貴重なものでした．

東京大学生産技術研究所の佐藤洋一教授には，研究の方針について指導していただきました．また，研究で行き詰まった際には親切に相談に乗っていただきました．深く感謝いたします．

東京大学生産技術研究所の米谷竜助教には，コンピュータビジョンや機械学習について教えていただき，本研究での具体的な手法の設計でも相談に乗っていただきました．また，本論文の執筆に際し，丁寧にご指導いただきました．ありがとうございました．

東京大学生産技術研究所の樋口啓太特任助教には，研究の進捗について相談に乗っていただきました．また，研究生活が楽しくなるよう，明るく接していただきました．

秘書の鈴木咲恵さん，今川洋子さんには様々な事務手続きや研究生活の相談でお世話になりました．

佐藤研究室の皆様にはデータセットの収集にあたり，ご協力いただきました．本当にありがとうございました．

最後に，暖かく見守ってくださった家族と，修士論文でお世話になったすべての方々に感謝いたします．

2017 年 2 月 3 日