NAIST-IS-MT1551001

修士論文

頭部動作認識のための
自然会話映像データセットの構築と評価

秋山 解

2017 年 3 月 16 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

秋山 解

審査委員：

向川 康博 教授	（主指導教員）
中村 哲 教授	（副指導教員）
金出 武雄 客員教授	（副指導教員）
舩冨 卓哉 准教授	（副指導教員）
伍 洋 特任助教	（副指導教員）

頭部動作認識のための
自然会話映像データセットの構築と評価*

秋山 解

内容梗概

本研究は，会話中に表れる頭部動作を映像から自動的に認識し，人間のコミュニケーションを理解する手がかりとして活用できるようにすることを目指す．そのために，自然会話映像を撮影し，従来の頭部動作認識で扱われてきたものより多種類である 10 クラスの頭部動作をアノテーションしたデータセットを構築した．複数名によるアノテーションを分析し，人による頭部動作の認識の曖昧さを確認した．会話における頭部動作の統計を分析した結果，クラスごとの頭部動作の発生頻度に大きな偏りが確認され，稀にしか観測されない頭部動作も確認された．また，構築されたデータセットに基づき，映像から推定した頭部位置姿勢から頭部動作の検出および識別を試み，複数のアルゴリズムで比較した．その結果から，自然会話における頭部動作から多種類の頭部動作を認識する問題は困難であると結論付けられた．頭部動作の定義の曖昧さ，個人間分散，サンプル数の不足，頭部位置姿勢推定の精度に関する課題を指摘した．

キーワード

頭部動作認識，データセット構築，自然会話，非言語情報

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1551001, 2017 年 3 月 16 日.

Dataset construction and benchmarking for spontaneous head gesture recognition*

Kai Akiyama

Abstract

This study targets at automatic recognition of head gestures in spontaneous human conversations, for the purpose of understanding human interactions and assisting people in their communication with others. Most existing datasets are non-spontaneous and small-scale, with only quite few gesture types, which makes a gap from real applications of head gesture recognition. Moreover, no dataset is publicly available for fair comparison of recognition models.

Therefore, a new video dataset is built directly focusing on spontaneous human-human conversations with a larger scale and a better coverage of gesture types. The videos are labeled by multiple annotators for a more reliable ground-truth and a better understanding of how humans do this recognition task. A detailed statistical analysis of the dataset reveals underlying challenges on biased gesture types and ambiguity.

As a baseline for recognition experiments using the dataset, classification and detection are attempted with a deep learning model as well as representative models from existing solutions. Experimental results reveal some important open issues such as the ambiguity of gesture definitions, differences among people, imbalanced gesture types, influences of human pose estimation.

Keywords:

Head gesture recognition, Dataset construction, Spontaneous conversations, Non-verbal information

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1551001, March 16, 2017.

目次

図目次

表目次

第 1 章　はじめに

有史以前より，人類はコミュニケーションを取ることで互いの感情を理解し，社会を築いてきた．そのコミュニケーションをコンピュータにも理解させようという試みは，現在様々な方面から取り組まれている．

コンピュータが人間のコミュニケーションをよりよく理解するようになると，人間とロボットやバーチャルアバターとのより人間らしいインタラクションが期待されるだけでなく，障碍者のコミュニケーション支援や，人間同士のコミュニケーションの分析にも応用が可能である．

対面で行う会話に焦点を当てると，会話の理解に必要な情報の代表として，声を通して伝わる言語情報が挙げられる．この言語情報を音声から抽出しようとする技術が音声認識であり，これは近年スマートフォンへの実装によって急速に日常生活に浸透した．しかし，人間同士の会話では言語情報に限らず様々な情報が交換されている．声の調子，表情，手動作，頭部動作などはその例であり，非言語情報と総称される．

非言語情報はコミュニケーションにおいて大きな役割を果たしている [1]．非言語情報は必ずしも意識的に発せられるものではないが，情報の受け手が捉える印象に大きく影響を及ぼしているとされる．Mehrabian [2] の実験によれば，音声の意味的な情報，音声の聴覚的な情報，顔の視覚的な情報に矛盾があったとき，概ねそれぞれ 7%，38%，55% の割合で受け手による感情推定に影響を及ぼすという．

コンピュータが人間と同様にコミュニケーションを理解するためには，こうした非言語情報の認識が欠かせない．先に例として挙げた，声の調子，表情，手動作，頭部動作については，音声感情認識，表情認識，手動作認識，頭部動作認識としてそれぞれ研究が行われている．こうした分野の中で，本研究は会話中で特に頻繁に出現するとされる [3] 頭部動作認識に着目している．

1.1　頭部動作の自動認識

パターン認識は一般に，検出 (detection) と識別 (classification) の問題に大別される．検出は，与えられたデータの中から所望のクラスが出現する箇所を特定する

問題である．一方識別は，予め何らかのクラスに属することが判明しているデータに対して，そのクラスを同定する問題である．映像からの頭部動作認識についてこれらを当てはめると，検出は映像中から頭部動作が発生している時刻とその頭部動作クラスを特定することである．録画された映像から頭部動作に関する統計を得たいときや，逐次入力される映像からリアルタイムに頭部動作の発生を知りたいときなど，多くの応用では検出が求められる．これに対して，識別は任意の頭部動作のみを何らかの方法で抽出した映像サンプルを予め用意し，このサンプルに対して頭部動作クラスを特定することである．事前にサンプルの抽出が求められるため応用先が限られる一方，識別の問題は検出と比べシンプルであり，検出に至るまでの必要条件としてしばしば研究されている．

1.2　本研究における貢献

従来の頭部動作認識に関する研究では，4 種類以下と少数の頭部動作のみが研究対象とされており，また，自然会話とは乖離した条件において撮影された映像が用いられることが多かった．本研究では，より多くの種類，具体的には 10 種類の頭部動作を研究対象とし，新たに自然会話映像データセットを構築し，頭部動作のアノテーションに対する分析を行った．構築されたデータセットを用いた頭部動作認識のベースラインとして，頭部動作の検出および識別アルゴリズムの適用を試み，その性能を評価した．さらに，データセットの構築，解析，および頭部動作認識の実験に基づいて，より優れた頭部動作認識手法に向けて解決すべき課題を明らかにした．

第 2 章　自然会話映像データセットの設計

2.1　関連研究

頭部動作を撮影した映像データセットについて，ここでは自然会話の文脈における頭部動作であるか否かで大別し，表 2.1 にまとめる．

多数を占めているのは自然会話以外の頭部動作のデータセットであり，これはさらに 2 種類に分けられる．一方は，ただ単純にカメラに向かって演じられた頭部動作の映像であり，表 2.1 の最初の 3 つが相当する．もう一方は，コンピュータの画面を前にして撮影されたもの [7, 8, 11, 15] ，人型の仮想エージェントが表示された画面を前にして撮影されたもの [12, 14]，およびロボットを前に撮影されたもの [9, 10, 11] である．いずれも質問に対する Yes/No の答えに共起するうなずきや首振りが撮影対象とされている．しかし，収集された頭部動作が数百個程度と小規模で，頭部動作のクラス数も 1 種類から 3 種類程度と少ない．うなずきが収集されて

表 2.1　頭部動作のデータセットの概要

名称	年	自然会話	人数	映像長さ	サンプル数	クラス数	アノテータ数
N/A [4]	1999	No	26	N/A	108 frames	3	1
N/A [5]	2000	No	3	∼35s	450 frames	3	1
N/A [6]	2013	No	N/A	N/A	150	2	1
N/A [7]	2001	No	10	N/A	110	2	1
N/A [8]	2003	No	11	N/A	190	2	1
N/A [9]	2004	No	N/A	3h24m	2212	3	1
Self&iGlass [10]	2005	No	16	∼1h	288	2	1
MelHead [11]	2007	No	16	∼1h	274	1	1
WidgetsHead [11]	2007	No	12	79m	269	1	1
RAPPORT [12, 13]	2010	No	47	∼1h34m	587	1	1
RAPPORT [14, 13]	2013	No	45	∼1h30m	666	1	1
N/A [15, 16]	2014	No	10	N/A	600	6	自動
KTH-Idiap [17][18]	2014	Yes	9	1h	136	1	1
Ubimpressed [18]	2015	Yes	12	1h	407	1	1
NOMCO [19]	2011	Yes	12	∼1h	2293	11	4
ALICO [20]	2014	Yes	50	5h31m	2440	12	2
FIPCO（本研究）	2015	Yes	14	∼5h30m	4148	10	3

いる点については全データセットにおいて共通であり，2 種類の頭部動作が扱われる場合は首振りが，3 種類の場合は更にかしげが追加される．

人間同士の会話に表れる自然発生的な頭部動作と比較して，こうした意図的な頭部動作はより大きな動きとして表れる傾向にあるため，自動認識が容易であると考えられる．

自然発生的な頭部動作のデータセットは，人間の話し相手がいる環境での会話を撮影したものである．KTH-Idiap コーパス [17, 18] は円卓を囲む 4 人で行われる集団面接が撮影されたものであり，Ubimpresssed データセット [18] は一対一の面接が撮影された．いずれも自然会話であるものの規模が小さく，うなずきのみが数百個程度アノテーションされている．これら 2 つのデータセットは自動認識のために構築されたものである一方，NOMCO [19] および ALICO [20] は言語学や感情化学の分野において構築された．これらにおいては多様な頭部動作が扱われており，10 種類以上の頭部動作が数千個アノテーションされている．また，アノテーションが主観に基づくことを考慮し，アノテーションが複数のアノテータによって行われている．なお，複数のアノテータを用いたのはこれらのデータセットのみであった．しかし，これらのデータセットに高度な自動認識アルゴリズムを適用した研究はない．

既存のデータセットに関しては，一部を除いてほとんどが一般の研究者にとって利用不可能な状況にある．そのため，各研究グループが独自にデータセットを撮影して実験を行っており，認識結果は比較できない．顔の表情認識や手動作認識と比較して頭部動作認識に関する研究があまり進められていないことには，こうした事情が関係している可能性がある．

2.2 データセットに対する要件

- 人同士の自然会話を撮影すること．
- 会話の中で意味を持って現れる多種類の頭部動作が実際に出現し，アノテーションされること．
- すべての映像を統一的に扱えるよう，映像間でカメラの視点や撮影方法を変更しないこと．

表 2.2　頭部動作の定義

主な移動・回転軸	ラベル名（英）	ラベル名（和）	説明
ピッチ	Nod	うなずき	顔が下がって上がる
	Jerk	上方うなずき	顔が上がって下がる
	Up	仰ぎ	顔をしばらく上げ続ける
	Down	俯き	顔をしばらく下げ続ける
	Ticks	連続うなずき	連続したうなずき
ロール	Tilt	かしげ	片方に頭を傾ける
ヨー	Shake	首振り	水平に頭を降る
	Turn	横を向く	水平に頭を回転させる
Z (スケール)	Forward	乗り出し	前に乗り出す
	Backward	退き	後ろに退く

- 同様の目的で，撮影される被験者はなるべく動き回らないこと．
- 固定カメラだけでなく，ウェアラブルカメラによるユースケースも考慮し，ウェアラブルカメラによる撮影も行うこと．

2.3　研究対象とする頭部動作

本研究において取り扱う頭部動作を表 2.2 に示す．以降，これらの頭部動作のラベルについては，表 2.2 中の ‘ラベル名（英）’ に従い表記する．これらの頭部動作は，NOMCO [19] および ALICO [20] における頭部動作のラベルを参考に決定した．うなずきについては，単独で発生する場合だけでなく，連続的に発生する場合が多く観測された．連続的に発生するうなずきをすべて個別にアノテーションすることは困難であるため，本データセットにおいては，単独で発生するものを Nod，間を置かず連続的に発生するものを Ticks と区別して定義した．そのため，一つの Ticks ラベルには連続する複数のうなずきが内包されている．

2.4　会話の設定

会話は一対一の対面で行うものとし，話し手・聞き手を定めない双方向の会話とする．撮影中に話題を見失い会話が停滞することのないよう，話題は事前に決定する．話題の選択にあたっては，話題の一覧（付録 A を参照）を提供し，撮影前に被験者同士で相談の上決定する．

第 3 章　自然会話映像データセットの構築

2 章での議論に基づき，自然会話映像データセットを構築した．

3.1　撮影対象

撮影対象となる被験者は，20 代の日本人男性 15 人である．その内 10 人は互いに面識のあるグループから選ばれており，全員について面識の有無を事前に確認した．会話を行う被験者の組み合わせは，互いに面識のある知り合い同士の組み合わせと初対面の組み合わせの両方を設定した．撮影では，毎回異なる組み合わせの被験者が会話に参加した．

3.2　撮影環境

会話映像の撮影は屋内および屋外で行った．被験者および機器の配置を図 3.1 に示す．被験者らは互いに向かい合って椅子に着席する．一方の被験者が撮影対象となり，他方の被験者がウェアラブルカメラを装着，隣に固定カメラを設置した．ウェアラブルカメラは装着者の額の位置から一人称視点映像を撮影し，同時にアイカメラで装着者の目の赤外線映像を撮影する．

固定カメラおよびウェアラブルカメラの間で撮影開始時刻の同期を取るために，撮影開始時は両カメラに時計が映るようにした．ただし時計が被験者らの視界に入ると会話を阻害する可能性があるため，撮影開始後は速やかに時計を取り除いた．

固定カメラは Logicool C920t，ウェアラブルカメラは Pupil Pro が用いられた．固定カメラは被験者の正面に設置されるため，被験者の注意を引かないよう，黒い紙で覆い，レンズ部分だけを切り抜いた．ウェアラブルカメラの装着者は装着時に Pupil 標準のキャリブレーションを行い，アイカメラの調整を行った．

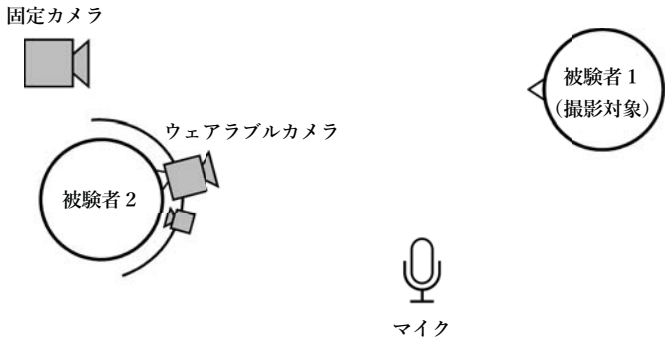

図 3.1　撮影における被験者と機器の配置

図 3.2　会話映像を撮影している様子

3.3　撮影された映像データセット

会話映像の撮影は全部で 44 回行われたが，撮影に不備のあった映像などを取り除き，最終的に被験者 14 人による 30 本分の会話映像のデータセットとした．映像は 1 本あたり約 11 分で，フレームレートは 24 フレーム毎秒である．固定カメラおよびウェアラブルカメラの映像の解像度は 1920 × 1080，アイカメラの映像の解像度は 640 × 480 である．30 本の各映像に関する撮影環境および被験者同士の関係は表 3.1 の通りである．

実際の映像の例は付録 B を参照されたい．

表 3.1　撮影環境と被験者同士の関係関係の内訳

		被験者同士の関係	
		知り合い	初対面
撮影環境	屋外	10	10
	屋内	5	5

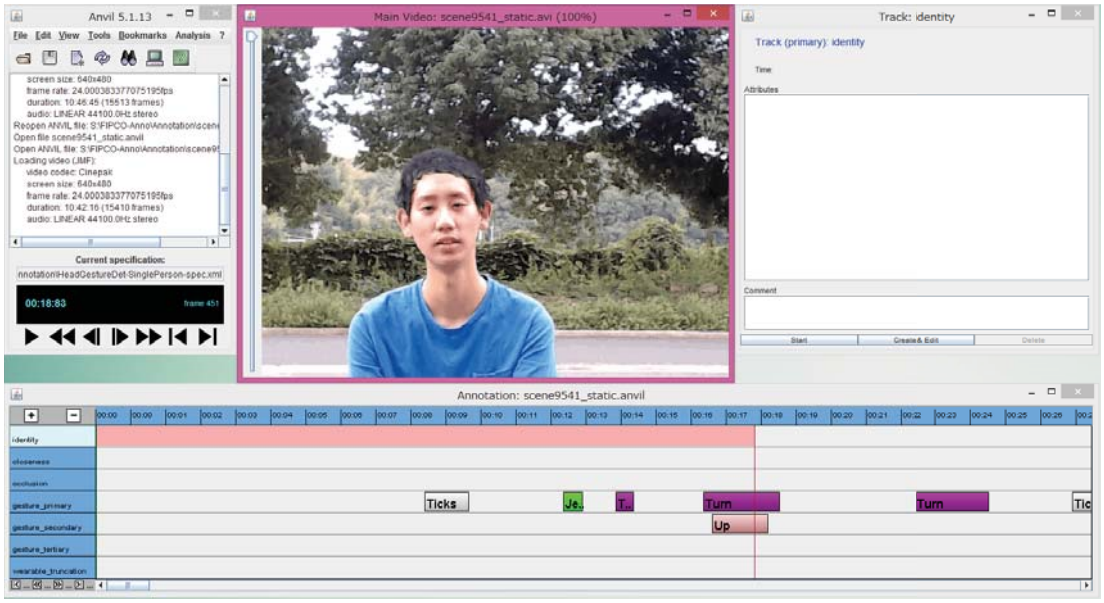

図 3.3　頭部動作のアノテーションを行っている画面

3.4　頭部動作のアノテーション

会話映像データセットに対して，頭部動作のアノテーションを行った．フリーウェア Anvil*を使用し，30 本すべての映像シーケンスが 3 人のアノテータによって独立にアノテーションされた．図3.3 はアノテーションを行っている画面である．

アノテーションの作業では，アノテータが映像を見ながら，頭部動作の発生して

*Anvil5，http://www.anvil-software.org/

表 3.2　3 人のアノテーションの一致度

比較するアノテータ	κ
A vs. B	0.45
B vs. C	0.46
C vs. A	0.44

表 3.3　統合前後のアノテーションの一致度

比較するアノテータ	κ
A vs. 統合後	0.58
B vs. 統合後	0.68
C vs. 統合後	0.63

いる箇所にラベルを配置し．そのクラスと 3 段階の強さを記録した．複数の種類の頭部動作が同時に発生した場合，あるいは頭部動作の分類が曖昧な場合に，最大で 3 つの頭部動作を重複してアノテーションすることを許容した．

こうして作成された 3 つのアノテーションから，各頭部動作クラスに対する全アノテータの見解を単一のアノテーションに統合し，これを真値として取り扱うこととした．アノテーションの統合は，各頭部動作クラスに関して，特定の場所におけるラベルの有無を多数決で決定し，合意に至ったラベルの開始点と終了点をそれぞれ平均した．具体的には，各頭部動作クラスに関して，3 つのアノテーション間で ‘重なり’ が閾値以上となるラベルのペアを列挙し，各ペアについて開始点と終了点をそれぞれ平均したものを，統合されたアノテーションとした．さらに，重複しているラベルについては，‘重なり’ の最も大きかったペアに由来するラベルのみを残すことで重複を解消した．なお，同時にアノテーションされている ‘強さ’ ついては，統合されるラベルの ‘強さ’ の平均をとった．

上述のラベルのペアに関する ‘重なり’ の尺度は，Intersection over Union (IoU) を採用した．また，合意しているとみなす ‘重なり’ の閾値は 0.5 とした．

3.5　アノテーションの一致度

複数のアノテータによって独立に行われたアノテーションについて，アノテータ間の一致度を計る尺度として Cohen’s Kappa（κ 係数）がある．Cohen’s Kappa は，2 者のアノテーションが一致している割合 p_0 および 2 者が正負の比率を保ったままランダムにアノテーションしたとき偶然一致する割合 p_e を用いて次の式で与えられる．

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{3.5.1}$$

$\kappa = 1$ のとき 2 者のアノテーションが完全に一致しており，$\kappa = 0$ のとき偶然による一致と同程度であると言える．

ここでは，特定のクラスのアノテーションの有無をフレーム毎に比較することで，Cohen's Kappa を計算する．

アノテータ A，B，C 間で各頭部動作クラスの Cohen's Kappa を計算したものを表 3.2 に示す．いずれの組み合わせにおいても一致度が十分に高いとされる目安の 0.6 より低く，アノテータ間で一定のアノテーションの基準が共有されていないことがわかる．一方，統合されたアノテーションと統合前の各アノテーションを比較した場合（表 3.3），いずれも統合前のアノテータ間の一致度と比較して向上しており，適切に統合が行われたことが確認できる．

3.6 頭部動作の統計

データセット全体のアノテーションより，頭部動作の内訳は図 3.4 のようになった．最も大きな割合を占める Nod は 1840 回観測された一方，Up，Shake，Forward，Backward に関しては 100 回以下であった．こうしたサンプル数の少ない頭部動作クラスは機械学習が困難であると予想される．

なお，Shake，すなわち首振りは多くの頭部動作認識に関する先行研究でうなずきと並んで取り扱われているが，本研究における自然会話データセットでは最も稀な頭部動作となった．ALICO において Shake が全ての頭部動作の 3.65% を占めているのと比べると，約 3 分の 1 の頻度である．これは ALICO が構築されたドイツと日本との文化の差によるものである可能性がある．

頭部動作の内訳を映像シーケンスごとに示したものが図 3.5，これを撮影されている被験者ごとに示したものが図 3.6 である．頭部動作の種類と数には大きな個人間分散があることがわかる．

頭部動作の強さの割合を撮影されている被験者ごとに示すと図 3.7 のようになる．全体的に弱い頭部動作が多く，1 と 1.5 を合わせると全体の 61% を占めている．個人間分散はここにも表れており，強さ 1 の頭部動作が約 20% 以下の被験者が 5 人

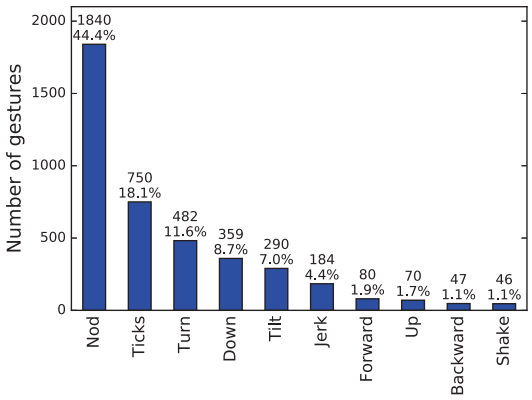

図 3.4　データセット全体における頭部動作の内訳．数字は頭部動作サンプルの数と割合を示している

いる一方で，半数が’weak’ である被験者が 3 人いる．

図 3.8 で会話を行っている被験者同士の面識の有無で頭部動作の内訳を比較すると，頭部動作の発生頻度は全体としてはほぼ等しい一方，内訳に差があることがわかる．初対面の場合，知り合いである場合と比べて Nod の割合が増加し，Ticks の割合が減少している．ここから導かれる仮説として，初対面の人と会話する場合は礼儀が重んじられるため，曖昧な反応である連続的なうなずきの代わりに，明確な単発のうなずきが好まれると考えられる．

図 3.9 は各クラスにおける頭部動作の継続時間の分布を示したバイオリンプロットである．Nod，Jerk，Shake はすべてのサンプルが約 2 秒以内に収まっている一方，それ以外のクラスでは分散が大きく，5 秒以上継続する長いサンプルも存在する．ここで使用しているアノテーションは 3 つのアノテーションの統合後の結果であるため，1 人のアノテータによって誤ったラベルが与えられたのではなく，少なくとも 2 人のアノテータの同意を得ているものである．この原因は，動作を表すラベルと状態を表すラベルが混在していることにある．分散の大きい Up，Down，Tilt，Turn，Forward，Backward のクラスについて，長時間に亘ってその状態が維持されることがあり，そうした事例にもラベルが与えられている．

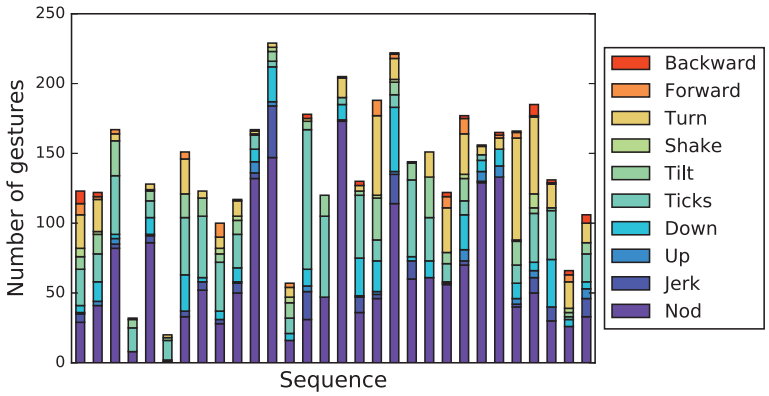

図 3.5　各シーケンスにおける頭部動作の内訳

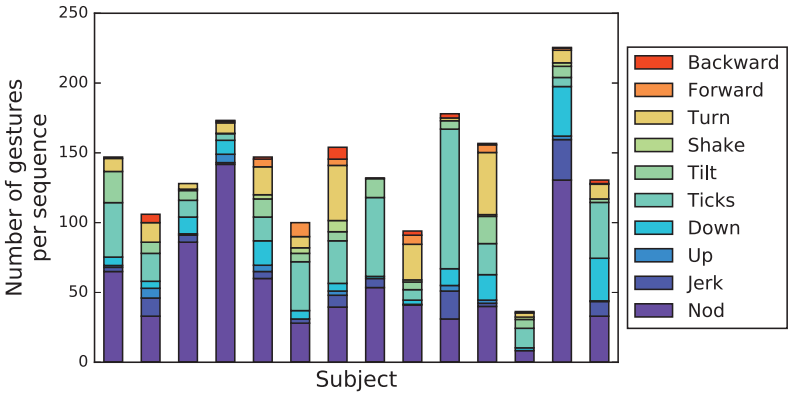

図 3.6　各被験者におけるシーケンスあたりの頭部動作の内訳

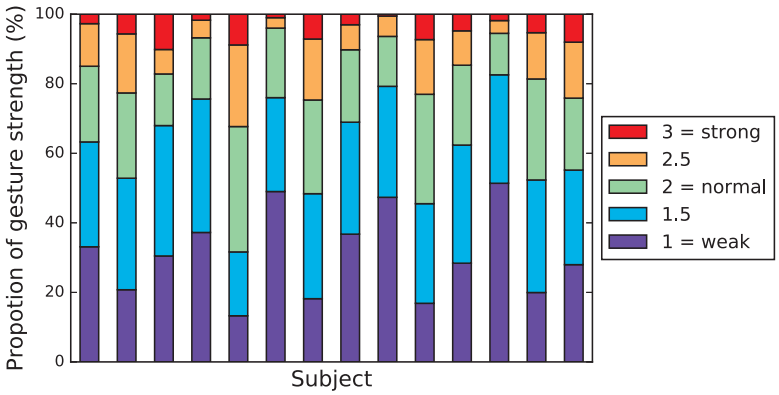

図 3.7　各被験者における頭部動作の強さの割合．被験者の並び順は図 3.6 と共通である

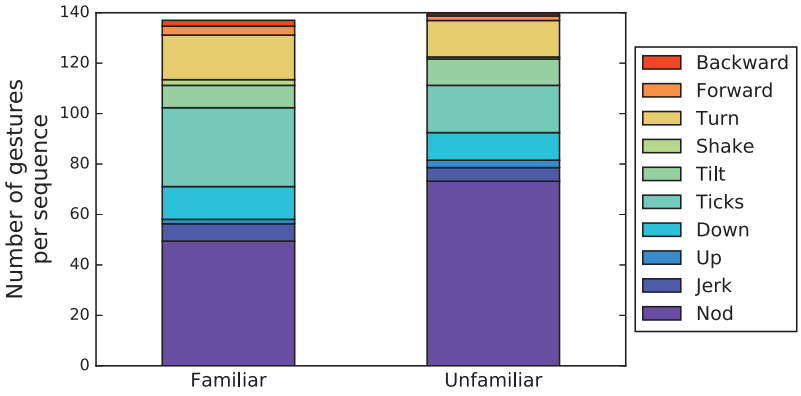

図 3.8　被験者同士の面識の有無による頭部動作の分布の差

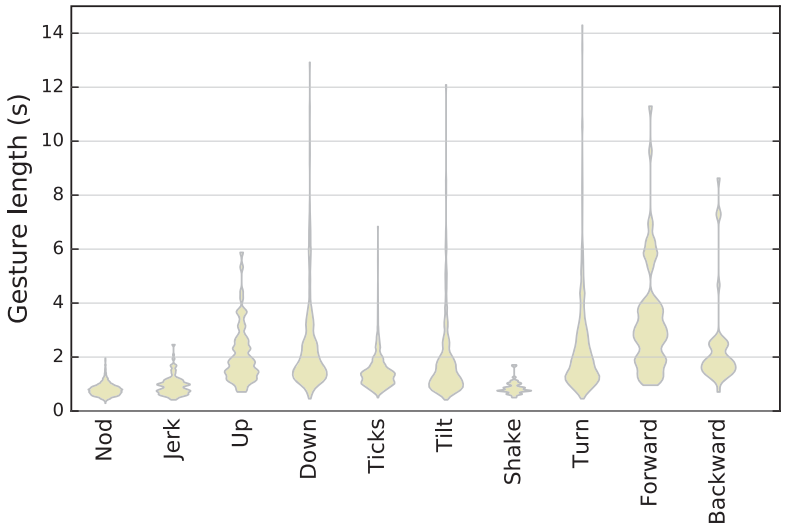

図 3.9　頭部動作の継続時間の分布

第 4 章　頭部動作認識アルゴリズムの適用

本研究で構築されたデータセットに対して，頭部動作を検出または識別するアルゴリズムを適用した．

4.1　関連研究

従来の頭部動作認識のアルゴリズムをここでは 3 種類に分類し，表 4.1 にまとめる．

まず機械学習に頼らない手法である．Kawato ら [5] のリアルタイムうなずき・

表 4.1　頭部動作認識アルゴリズムの概要

アルゴリズム	著者	年	比較対象
ルールベース	Kawato ら [5]	2000	
ルールベース	Anam ら [21]	2014	
Buffered State Machine	Galanakis ら [22]	2014	
単純ベイズ	Paggio ら [19]	2011	ZeroR
SVM	Morency ら [10]	2005	HMM
SVM	Chen ら [18]	2015	
TFSM	Davis ら [23]	2001	
TFSM	Chu ら [24]	2012	
HMM	Choi ら [4]	1999	直接観測
HMM	Kapoor ら [7]	2001	
HMM	Tan ら [8]	2003	
HMM	Fujie ら [9]	2004	
HMM	Gunes ら [25]	2010	
HMM	Wei ら [6]	2013	
HMM	Terven ら [15]	2014	
HMM	Saleh ら [26]	2015	
HMM	Terven ら [16]	2016	
LDCRF	Morency ら [11]	2007	HMM, SVM, CRF, HCRF
LDCRF	Ozkan ら [12, 13]	2010	CRF
LDCRF	Ramirez ら [27]	2011	SVM, 決定木, CRF
LDCRF	Ozkan ら [14, 13]	2013	CRF
	（本研究）	2017	SVM, HCRF, LDCRF, LSTM

首振り検出は，眉間のトラッキングによるルールベースの手法である．ルールベースのモデルは近年の研究にも見られ，視覚障碍者支援に向けた研究 [21] の中で用いられている．その他，Buffered State Machine [22] のように基本的な 6 種類の動き（上，下，左，右，左回転，右回転）を推定された頭部位置姿勢から認識する手法がある．こうした機械学習を用いない手法は，データセットのアノテーションを必要としない一方，ルールで記述可能な検出対象と整ったデータを用意する必要があり，柔軟性があるとは言えない．

次に，頭部動作をサンプルごとに機械学習によって識別する手法である．Paggio ら [19] は頭部動作の役割の識別に単純ベイズを用いた．Morency ら [10] および Chen ら [18] は，画像からの一般物体検出 [28] の場合と同様に，映像から SVM によるうなずきの検出が可能なことを示した．

最後に，最も多く用いられている時間的変動を扱った手法である．表 4.1 に示すように，頭部動作認識に Timed Finite State Machine (TFSM) を用いる研究 [23, 24] として見つかるのは 2 件である一方，隠れマルコフモデル (HMM) を用いる研究は少なくとも 9 件存在する．HMM は人の手による状態定義に依存しているが，自然な頭部動作がこうした定義に常に適合するとは考え難い．したがって，状態定義に頼らない手法として条件付確率場 (CRF) や Hidden-state CRF (HCRF) の方が適していると考えられる．特に Morency ら [11] が提案した Latent-Dynamic CRF (LDCRF) は，柔軟に隠れ状態をモデル化でき，従来のグラフィカルモデル (HMM，CRF,HCRF) や SVM，決定木よりも効果的とされる [13, 12, 14, 27]．ただし LDCRF は自然会話における頭部動作には適用されていないため，本研究でその性能を検証する．

4.2 適用する手法の概要

頭部動作を検出するシステムの概要を図 4.1 に示す．ここでは映像はすべて固定カメラによって撮影されたものを用いる．学習用映像は頭部位置姿勢推定および特徴量抽出によって時系列データに変換され，アノテーションから得られた真値のラベルと共に検出器の学習が行われる．多クラス検出へのアプローチとして，特定の頭部動作クラスのみを検出できる独立した 2 クラス検出器を頭部動作の種類だけ用

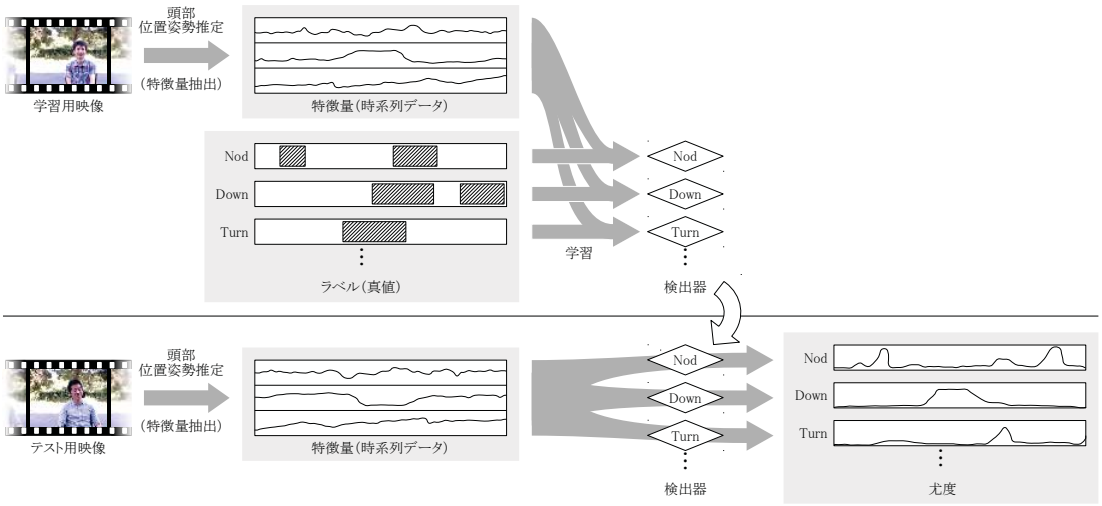

図 4.1　頭部動作検出システムの概要

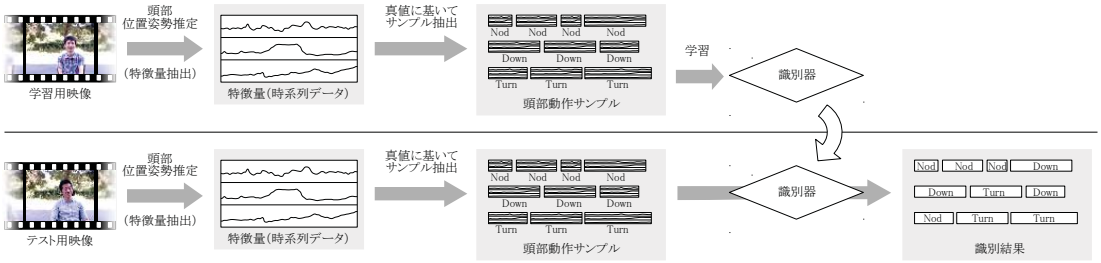

図 4.2　頭部動作識別システムの概要

意する手法を採用する．これはアノテーションにあたって異なるクラスのラベルが重複することを許しているためである．テストの際には，各検出器へ同一のデータを入力し，各頭部動作ごとの検出性能を評価する．これにより複数のクラスが重複して検出されることを許して評価を行うことができる．

一方頭部動作を識別するシステム（図 4.2）は，検出システムとは異なり一つの頭部動作サンプルに対して唯一の出力が期待されるため，多クラス識別器を使用する手法を採用する．

4.3 頭部位置姿勢の推定

まず会話映像に対して頭部の位置姿勢推定を行う．ここでは三次元顔モデルフィッティング手法である Zface [29] を使用し，画像上における顔の X 座標，Y 座標，スケール，ピッチ角，ロール角，ヨー角を推定した．以降，映像シーケンスは 6 変数から成る時系列データとして取り扱う．

なお，ノイズを軽減するため，位置姿勢の時系列データは 3 フレーム幅で平滑化処理を行った．

4.4 特徴量抽出

独自の特徴量として Histogram of Velocity and Acceleration (HoVA) を設計し，頭部位置姿勢の時系列データに適用した．実験では，HoVA 特徴量を用いない生の位置姿勢データを用いた場合と比較する．

HoVA 特徴量は，時系列データの速度・加速度（一階微分・二階微分）の変動を記述することができる．頭部動作は，頭部全体の位置姿勢の速度・加速度によって記述できると考えらえられ，また，速度・加速度を求める微分によって視点位置に依存しない検出が可能になる．HoVA 特徴量の利点として，セルに分割することでラベルのずれがある程度許容されるようになる点，ヒストグラム化およびヒストグラムの正規化によってサンプル毎の波形の差異に対してロバストになる点が挙げられる．

図 4.3 に従って HoVA 特徴量の詳細を述べる．まず，時系列データ $\mathbf{x}$ を一階微分した速度 $\mathbf{v} = \dot{\mathbf{x}}$ を求める．ここではマスク $[-1, 0, 1]$ の畳み込みにより一階微分を計算した．次に，時系列データを幅 w_c のセルに分割し，セルごとに正のビン V_i^P および負のビン V_i^N を用意する．セルごとの $\mathbf{v}$ の値が，対応する符号のビンに投票される．このとき，$\mathbf{v}$ の注目セル c_i の中心から半幅 w_c 以内にある正・負の要素数を，注目セル c_i の中心からの距離による線形重みをかけて，それぞれ V_i^P，V_i^N に加える．最後に，こうして得られたヒストグラムを隣接するセルによって正規化する．ここでは，左側セルとの正規化と右側セルとの正規化の平均をとった．すなわち最終的な正規化された時系列データ $\bar{\mathbf{V}}^P$ および $\bar{\mathbf{V}}^N$ を次の式で与える．

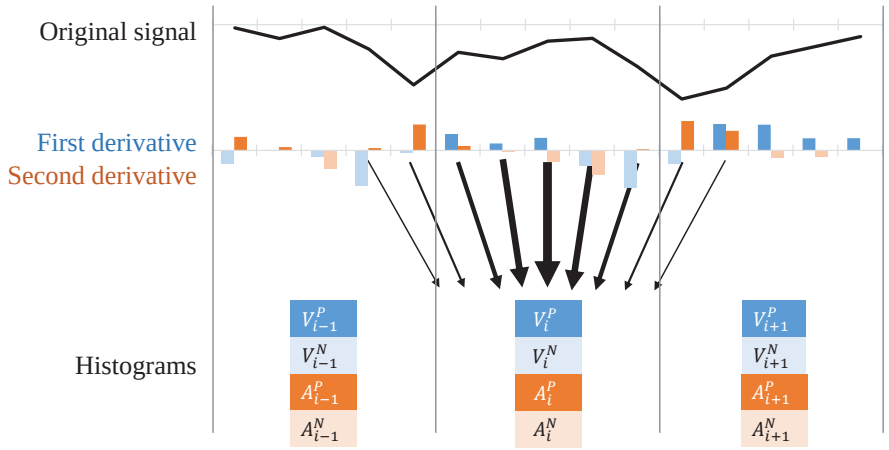

図 4.3　HoVA 特徴量抽出におけるヒストグラム化

$$\bar{V}_i^{P,N} = \frac{1}{2} \sum_{j \in \{0,1\}} \frac{V_i^{P,N}}{||V_i||_j} \tag{4.4.1}$$

$$||V_i||_j = \sqrt{(V_{i+j-1}^P)^2 + (V_{i+j-1}^N)^2 + (V_{i+j}^P)^2 + (V_{i+j}^N)^2} \tag{4.4.2}$$

以上を二階微分した加速度 $\mathbf{a} = \ddot{\mathbf{x}}$ に対しても行い，$\bar{\mathbf{A}}^P$ および $\bar{\mathbf{A}}^N$ を得る．二階微分のマスクは $[-1, 0, 2, 0, -1]$ を使用した．このようにして，時系列データ $\mathbf{x}$ から，長さ w_c のセルごとの 4 変数特徴量 $\bar{\mathbf{V}}^P, \bar{\mathbf{V}}^N, \bar{\mathbf{A}}^P, \bar{\mathbf{A}}^N$ を得る．頭部動作に適用すると，位置姿勢の 6 変数それぞれに対して HoVA 特徴量が計算され，最終的に 24 変数となる．

4.5　機械学習モデル

HoVA 特徴量または生の時系列データに基づいて識別や検出を行う機械学習モデルとして，SVM，LDCRF，HCRF，LSTM を適用した．

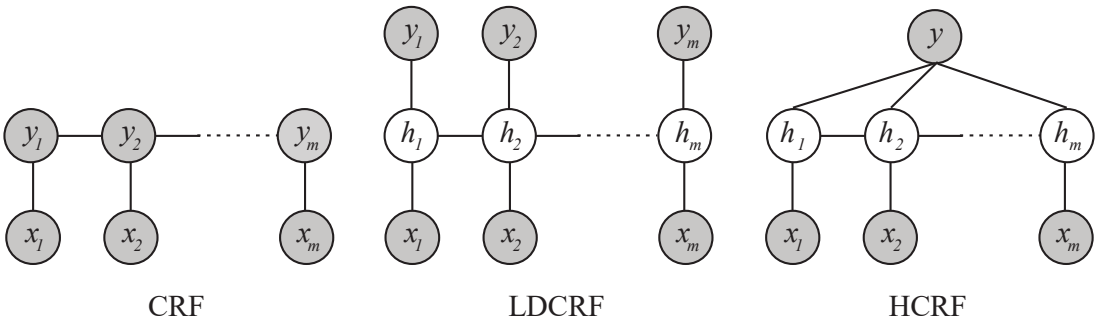

図 4.4　CRF およびその派生モデル

4.5.1　SVM

SVM は目的関数とデータの間のマージンを最大化する識別器として広く用いられており，モデルへの適合性と汎化性のバランスをパラメータによって調整可能である．比較的小規模でシンプルなモデルだが，識別・検出両方の問題でその性能が示されている．本研究では Linear SVM を，非グラフィカルモデルの代表として識別・検出の両方で用いる．

4.5.2　LDCRF，HCRF

Conditional Random Field (CRF，条件付き確率場) は，説明変数に基づく目的変数の条件付き確率を最大化するよう推論されるグラフィカル モデルだが，これを元にいくつかの派生モデル（図 4.4）が提案されている．

Latent-Dynamic CRF (LDCRF) は CRF から派生したモデルで，頭部動作検出に適用された実績がある．CRF では教師データとして明示的に与えられるラベル間の状態遷移がモデル化される一方，LDCRF はそれに加えて隠れ状態層によりラベルの暗示的な状態遷移をもモデル化することができる．

本研究では LDCRF を検出器として用いる．Morency [11] らの手法に従い，映像シーケンスから正事例の周辺と不事例のサンプルを抽出して学習を行う．HoVA 特徴量を用いる場合のウィンドウ幅は 3，ラベルあたりに割り当てる隠れ状態数は 2 とした．

Hidden CRF (HCRF) も LDCRF と同様に CRF から派生し，動作認識のため

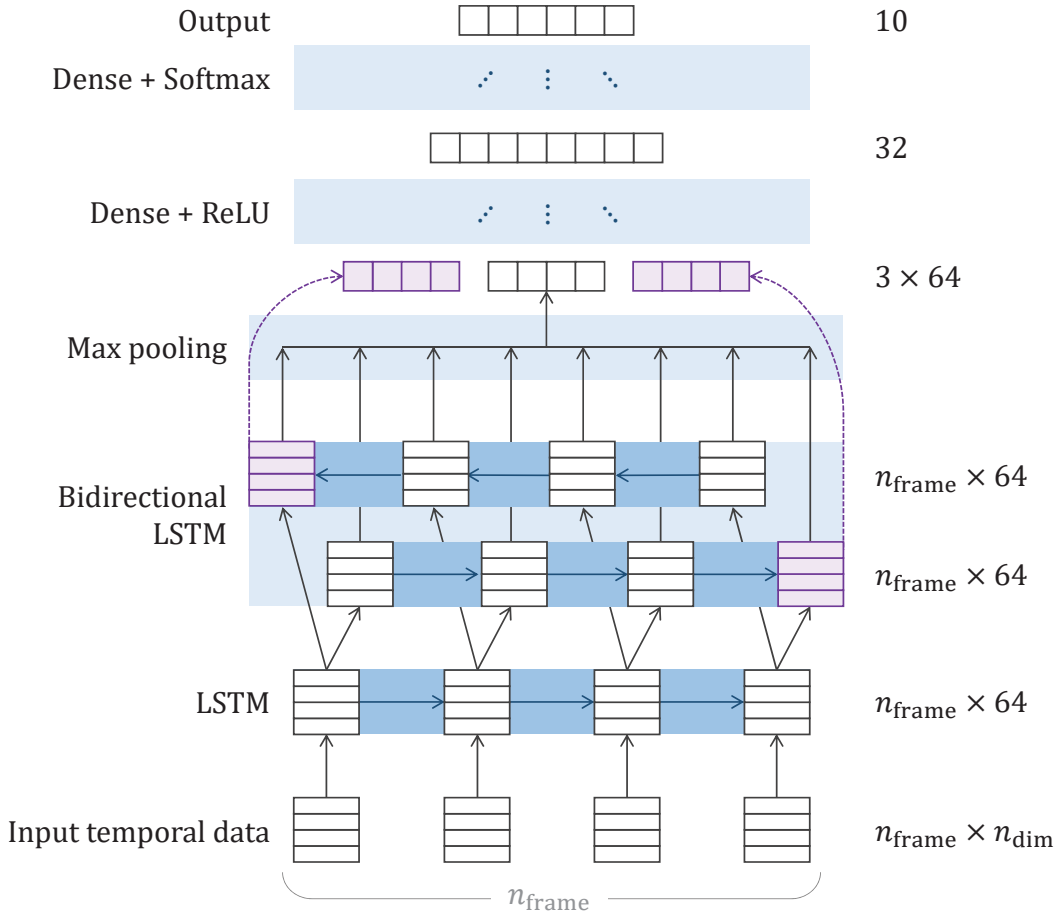

図 4.5　LSTM を用いた識別のためのモデル

のモデルとして提案された．ただし，ラベル内の隠れ状態の遷移に注目しており，ラベル間の遷移を扱えない点が LDCRF と異なる．本研究ではこれを識別器として用いる．

4.5.3　LSTM

Long-short Term Memory (LSTM) は再帰的ニューラルネットワークの一種であり，映像や信号などの時系列データに用いられるものである．

ここでは Shen ら [30] が提案する Hybrid Bidirectional LSTM なるモデルを採用する．可変長の頭部動作サンプルを入力とする識別問題へ適用するにあたって，

LSTM への入力が可変長行列となることによって出力も可変長行列となる一方，単一の識別結果を得るために LSTM から一定の大きさの出力行列を得る必要がある．一般的に，LSTM の最後尾の出力のみを用いる方法や，出力全体の時間方向の最大値または平均値を用いるプーリングによって解決される．これらに対して，Hybrid Bidirectional LSTM は Bidirectional LSTM の最後尾の出力とプーリングによる出力を併用し，性能の向上を図ったものである．

本研究で頭部動作の識別のために使用されたモデルが図 4.5 である．可変長の時系列データを入力とし，出力ベクトルは各要素が頭部動作のクラスを表す．Hybrid Bidirectional LSTM のモデルの前に通常の LSTM を追加した他，Shen ら [30] と同様に全結合層によって出力ベクトルを得た．

4.6 評価方法

4.6.1 交差検証

頭部動作認識モデルの性能は交差検証によって確認する．交差検証では，データセットを 2 つのサブセット，すなわち学習用セットおよびテスト用セットに分割する．学習用セットに含まれるサンプルのみでモデルを学習させ，テスト用セットに含まれるサンプルを使用してモデルをテストする．この一回の学習・テストに用いられるデータセットの分割をスプリットと呼ぶ．データセットの分割方法を変えることで複数のスプリットを生成し，各スプリットについて学習・テストを行い，すべての結果を総合してモデルの性能を測る．

さらに，本研究では学習用セット・テスト用セットをそれぞれ 2 つのサブセットに分割する方法を用いる．学習用セット (‘training set’) およびテスト用セット (‘test set’) から，それぞれ一部のサンプルが ‘training-validation set’ および ‘validation set’ として抽出される．training set で学習したモデルはまず training set 自身でテストされ，モデルの学習データへの適合性が評価される．次にこのモデルは training-validation set でテストされ，学習に用いたものと類似したデータにおける汎化性能が評価される．さらに validation set でテストされ，最終的なテストに用いるものと類似したデータにおける汎化性能が評価される．モデルのパラメータはこの validation set でのテスト結果を踏まえて調整される．最後に test

set でテストされ，これによって性能の評価値が算出される．

スプリットの生成方法は leave-one-person-out とする．すなわち特定の被験者が撮影対象となっている映像をテスト用として抽出し，テスト用の被験者を変えながらスプリットを生成する．training-validation set および validation set として，それぞれ training set および test set から頭部動作が最も頻繁に含まれる映像を抽出した．このとき，test set に含まれる映像が 1 本のみであった場合は validation set を抽出できない．これを避けるため，複数の映像に収録されている被験者のみをテスト用として使用する．データセットには合計 14 人の被験者が撮影されているが，この内 4 人は 1 本の映像にしか収録されていないため，結果としてスプリットの数は 10 個になった．

4.6.2 評価尺度

性能を評価するための尺度として，検出の性能評価には PR 曲線 (precision-recall curve) および AP (average precison) を，識別の評価尺度として混合行列および accuracy，F-measure を用いることとする．F-measure は，accuracy とは異なり，データ数の分布に偏りがある場合にも公平な評価を与える．

PR 曲線は，検出・不検出を区別する閾値を変化させたときの precision および recall の変動をグラフに示したものである．理想的な検出器では PR 曲線は precision $= 1$ の直線となる．AP は PR 曲線の積分値である．

4.7 実験結果

4.7.1 検出の結果

従来手法として代表的な SVM および LDCRF による検出性能を PR 曲線および AP で図 4.6 に示す．最もサンプル数が多かった Nod の検出性能が最も高く，反対に，最もサンプル数の少ない Shake 等は非常に低い結果となった．

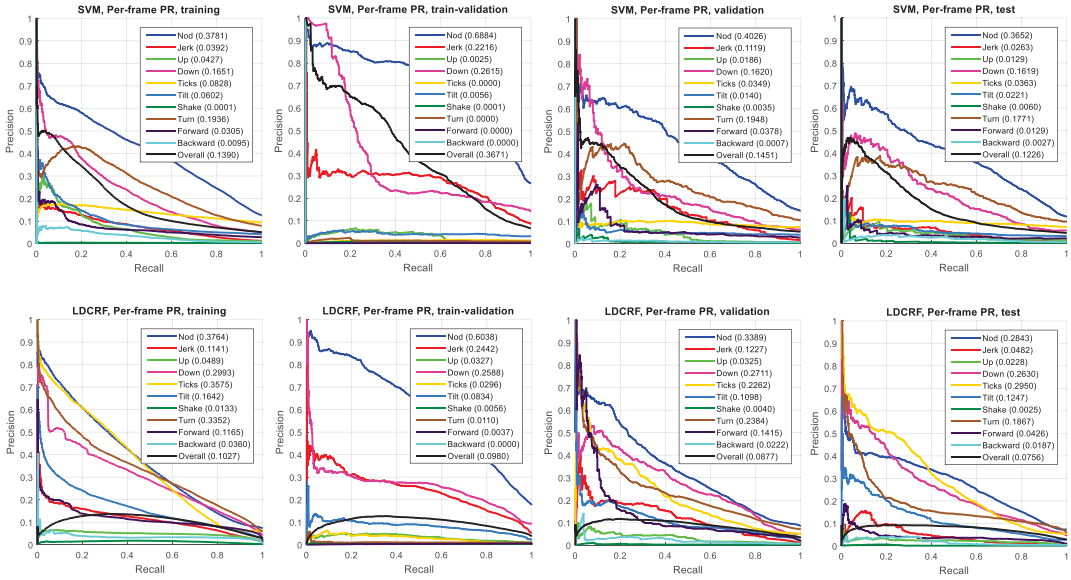

図 4.6　SVM および LDCRF による検出性能を示す PR 曲線．各グラフの凡例中の数字は average precision (AP) を示している

4.7.2　識別の結果

頭部動作認識に関する既存の手法として SVM，HCRF を適用する他，深層学習による手法として LSTM を適用し，識別性能を F-measure で比較すると図 4.7 のようになった．クラス間でサンプル数の偏りが大きいため，SVM では学習時に重みを付けた場合も試したが，僅かな性能向上にとどまった．HCRF，LSTM を用いた場合も，training set によるクローズドテスト以外では僅かな向上にとどまった．ただし，HCRF と LSTM では学習データに対するモデルの適合性が SVM より高かった．

HCRF および LSTM による識別について，test set でテストされた場合の混合行列を図 4.8 に示す．Shake については全く推定結果に現れないなど，サンプル数の少ないクラスに関する性能は低かった．しかし，最も頻繁に出現した Nod については 0.8 を超える高い精度を得ることができた．HCRF と比較して，LSTM では Nod および Up において性能が低下した一方，その他のクラスでは性能が向上した．すなわち，多種類の頭部動作を同列に扱う場合は，LSTM の方が優れているといえる．

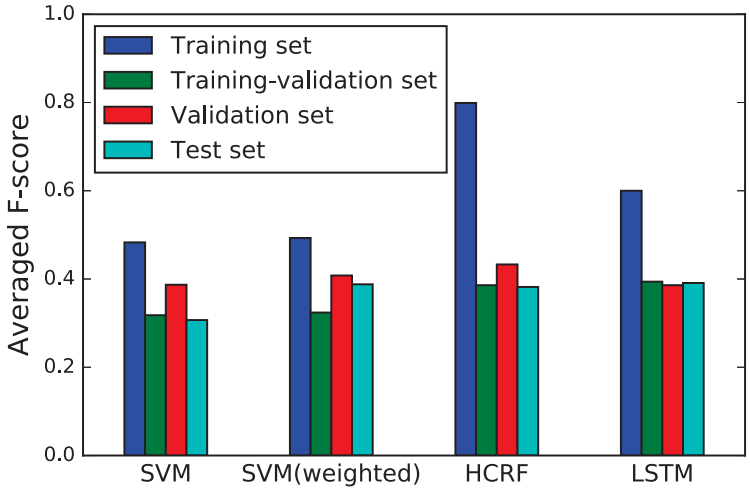

図 4.7　各識別器による識別性能

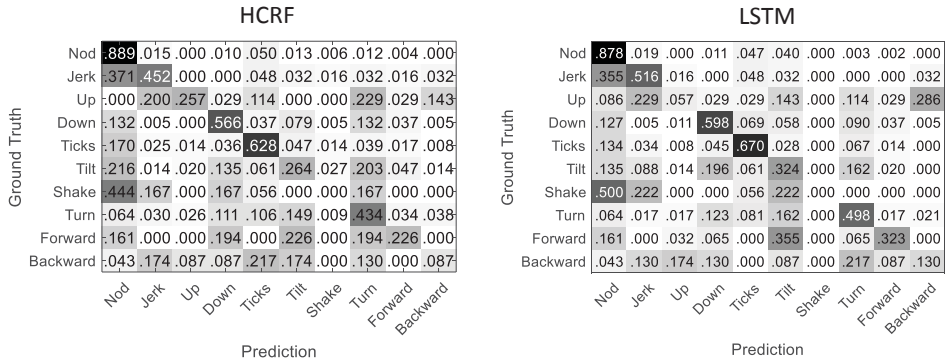

図 4.8　HCRF および LSTM による識別結果の混合行列

4.7.3　HoVA 特徴量の有効性

HoVA 特徴量を使用した場合と，HoVA 特徴量を抽出せず頭部位置姿勢の生データを使用した場合で識別性能を図 4.9 に比較した．いずれの識別機においても HoVA 特徴量を用いた場合の方が F-score が高くなっており，頭部動作認識において HoVA 特徴量が生データより適していることが示された．

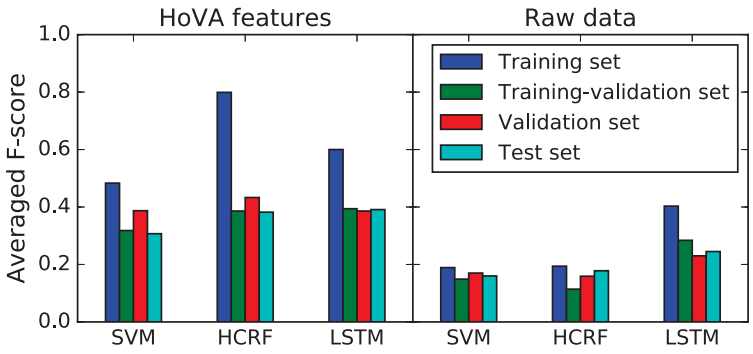

図 4.9　HoVA 特徴量と生データでの各識別器による識別性能

4.7.4　LSTM について

図 4.9 における生データの結果を見ると，LSTM の識別性能が最も優れている．したがって，特徴量の設計によっては LSTM が最も適したモデルである可能性がある．

LSTM の学習における各ステップでの accuracy の推移をグラフにしたものが図 4.10，および学習率をこれより高く設定したものが図 4.11 である．前者はモデルが高速に収束しているが，後者は収束までに時間がかかっている．また，後者のモデルは training set 以外のデータへの汎化性が僅かに低いが，学習データに対する適合性が非常に高く，十分なサンプル数があれば，より高い性能を得られる可能性がある．

図 4.10 について，test set を各スプリットに分けて示したのが図 4.12 ってある．スプリット，すなわち撮影される被験者によって，性能が大きく左右されることが分かる．頭部動作の個人間分散が大きいことは，この結果にも表れているといえる．なお，他の LSTM 以外のモデルについても同様な現象が確認された．

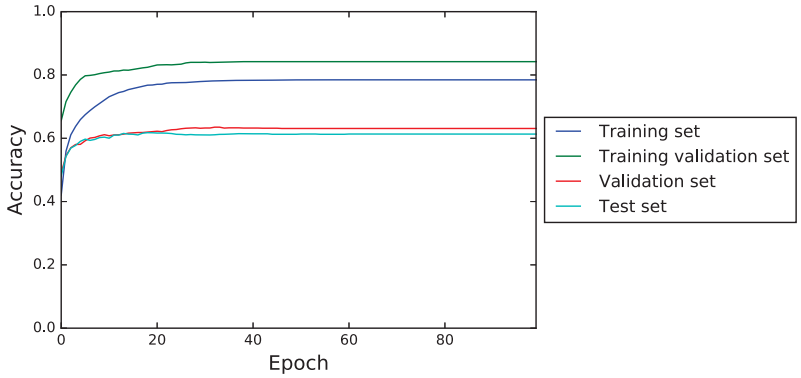

図 4.10　LSTM による識別精度の推移．Accuracy はすべてのスプリットの平均を取っている

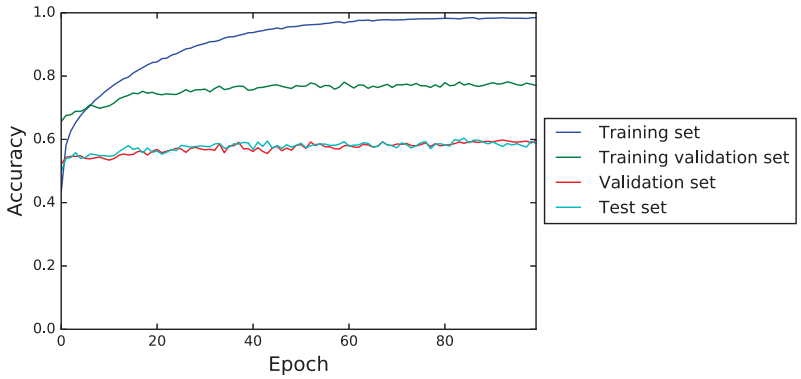

図 4.11　学習率を高く設定した場合の LSTM による識別制度の推移

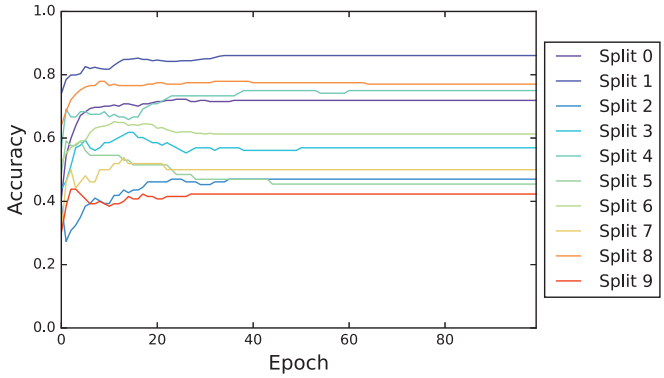

図 4.12　LSTM による test set の識別精度の推移

第 5 章　おわりに

5.1　明らかになった課題

本研究におけるデータセットの収集，分析および頭部動作認識の実験を通して明らかになった課題を挙げる．

頭部動作の定義が曖昧であること．人の頭部動作にいくつかの種類があることは明らかだが，それらの分類や，出現の有無を判断する明確な基準が存在しない．アノテーションで用いられる基準はアノテータの主観に任されており，その主観による基準も常に一定に保つことは困難である．アノテーションの基準の曖昧さは，3.5 節で示した通り Cohen's Kappa の低さに表れている．したがって，自然会話における頭部動作を Nod や Shake などといった高レベルの抽象表現として認識する問題は非常に困難であり，根本的に [22] のような基本的な動きの統計で扱う方が現実的である可能性がある．

頭部位置姿勢データの S/N 比が低いこと．ZFace による頭部位置姿勢推定には誤差があり，頭部位置姿勢データにノイズとなって現れる．このノイズと細かい頭部動作との判別が困難である一方，映像から人が判別することは可能である．したがって，より高精度な頭部位置姿勢手法を用いるか，あるいは頭部位置姿勢のような画像情報の高レベルな抽象表現ではなく低レベルな抽象表現を活用することで改善が期待される．後者の例として，映像のフレームとオプティカルフローに同時に畳みこみニューラルネットワークを適用する手法が挙げられる．

頭部動作に個人差があること．前 2 項に関連して，頭部動作がどの程度明確に表れるかは個人による差が大きく，頭部動作の有無を判別する万人に共通した閾値を設けることは困難である．個人に合わせて適切な閾値を決定する，あるいは複数のモデルの中から個人に適切なモデルを選択するなどの対応が考えられる．

稀にしか出現しない頭部動作があること．本研究で定義した 10 種類の分類では，すべての頭部動作の中で Nod および Ticks が計 6 割を占めた一方，Shake など 1% にしか及ばないものもあった．こうした稀にしか出現しない頭部動作は，その頭部動作クラスを網羅的に表現可能な可能な，十分なサンプル数を自然会話の中から収集することが難しい．したがって，稀な頭部動作は何らかの方法でより多くのサンプルを収集する必要があり，自然会話以外の文脈から収集されたデータで学習

したモデルが効果的な可能性もある．同時に，こうした稀な頭部動作を認識することの必要性を再検討する必要がある．

5.2　ウェアラブルカメラによる撮影に関する問題点

会話映像を撮影するにあたって，固定カメラによる撮影だけでなくウェアラブルカメラによる一人称視点映像の撮影も行った．当初この一人称視点映像をホモグラフィ変換によって補正し，固定カメラと同様に扱うことを検討していた．しかし，ウェアラブルカメラによって撮影された被験者の映像は，装着者の頭部動作によって大きな移動が発生し，撮影されている被験者が画面外に見切れる場面が散見された．見切れが発生していない箇所に限定しても，映像の補正は困難を極め，固定カメラと同様に扱える程度の補正結果は得られなかった．

頭部装着型のウェアラブルカメラを用いて装着者自身の頭部動作を認識できる可能性はある．しかし，話し相手の振る舞いをウェアラブルカメラで捉えるためには，胸部など安定した部位にカメラを装着する方が適切である．

ウェアラブルカメラに搭載された近赤外線のアイカメラを用いて視線方向の推定を試みた．ここで用いた視線推定手法は，瞳孔の位置を視線方向にマッピングするものであるため，黒い瞳孔を正確に捉える必要がある，しかし，屋外で撮影されたものは近赤外線の環境光が強く，眼球表面で反射して視線推定の妨げとなった．また，被験者によっては瞼が瞳孔の位置を推定する上で大きな障害となった．これらの理由により，安定した視線方向の推定に成功したアイカメラの映像は多くなかった．

5.3　まとめ

本研究では，自然会話の中に表れる頭部動作を自動的に認識できるシステムを目指し，共有可能な自然会話映像データセットを構築した．従来の頭部動作認識では，主に自然会話以外の状況で撮影された映像を用いて，1 種類から 3 種類程度の頭部動作クラスのみが扱われてきた．一方，本研究では言語学等の分野における会話映像コーパスを参考に 10 種類の頭部動作クラスを定義し，3 名にのアノテータによっ

てアノテーションを行った．また，アノテーションを統合して頭部動作認識の問題に利用できる形にした．

統合前のアノテーションの一致度を分析したところ，いずれのアノテータも互いに一致度が低く，人による頭部動作認識の曖昧さが明らかになった．また，頭部動作の内訳を分析することで，クラスごとの頭部動作の発生頻度に大きな偏りが確認され，稀にしか観測されない頭部動作も確認された．

さらに，構築されたデータセットを元に，複数のモデルによる頭部動作の検出および識別を試みた．ここで試したいずれのモデルについても，性能が十分に高いとは言えず，実用化には更なる研究が必要である．サンプル数が少ない Shake 等のクラスに関しては特に特に性能が低いため，より多くのサンプルを収集する必要がある．本研究で提案した HoVA 特徴量は，これを使用しない場合と比べて優れた結果が得られ，頭部動作認識に適していることが示された．従来用いられてきたモデルと比較して，大量の学習データが得られた場合は，LSTM が最良の選択肢となり得ることが期待される結果が得られた．

自然会話における頭部動作の認識は，人間にとっても困難な問題であり，データやアノテーションの収集方法に工夫が必要であると考えられる．

謝辞

本研究に取り組むにあたって，親身になってご指導いただきました，ロボットビジョン研究室の伍洋 特任助教，そして光メディアインタフェース研究室の向川康博 教授，舩冨卓哉 准教授，久保尋之 助教に，心より感謝申し上げます．また，研究に対する精神や，論文の書き方など，多方面からのご指導をいただきました，金出武雄 客員教授にも，感謝の念が尽きません．海を越え，カーネギーメロン大学から，研究に関する具体的な助言をいただきました，Kris M. Kitani 先生，Laszlo A. Jeni 先生にも，この場を借りてお礼申し上げます．そして，私の生活や研究をいつも支えてくれた，光メディアインタフェース研究室の皆様のことを，私は忘れることはないでしょう．博士前期課程における 2 年間にわたり，家族，親戚，友人，先生，事務員，様々な方のご支援といただきましたこと，改めて感謝の念をここに記します．

参考文献

[1] Evelyn Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, Vol. 32, No. 7, pp. 855–878, 2000.

[2] Albert Mehrabian, et al. *Silent messages*, Vol. 8. Wadsworth Belmont, CA, 1971.

[3] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 1983.

[4] H.-I. Choi and P.-K. Rhee. Head gesture recognition using HMMs. *Expert Systems with Applications*, Vol. 17, No. 3, pp. 213–221, 1999.

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

[6] 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 *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pp. 1–4. IEEE, 2013.

[7] Ashish Kapoor and Rosalind W. Picard. A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pp. 1–5. ACM, 2001.

[8] 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, October 2003.

[9] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pp. 159–164. IEEE, 2004.

[10] Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pp. 18–24. ACM, 2005.

[11] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8. IEEE, 2007.

[12] Derya Ozkan, Kenji Sagae, and Louis-Philippe Morency. Latent mixture of discriminative experts for multimodal prediction modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 860–868. Association for Computational Linguistics, 2010.

[13] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, pp. 125–138. Springer, 2007.

[14] Derya Ozkan and Louis-Philippe Morency. Latent mixture of discriminative experts. *IEEE Transactions on Multimedia*, Vol. 15, No. 2, pp. 326–338, 2013.

[15] Juan R. Terven, Joaquin Salas, and Bogdan Raducanu. Robust head gestures recognition for assistive technology. In *Mexican Conference on Pattern Recognition*, pp. 152–161. Springer, 2014.

[16] 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, January 2016.

[17] Catharine Oertel, Kenneth A. Funes Mora, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, UM3I '14, pp. 27–32, New York, NY, USA,

2014. ACM.

[18] Y. Chen, Y. Yu, and J. M. Odobez. Head Nod Detection from a Full 3d Model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 528–536, December 2015.

[19] Patrizia Paggio and Costanza Navarretta. Learning to classify the feedback function of head movements in a Danish corpus of first encounters. In *Proceedings of ICMI2011 Workshop Multimodal Corpora for Machine Learning*, 2011.

[20] Hendrik Buschmeier, Zofia Malisz, Joanna Skubisz, Marcin Wlodarczak, Ipke Wachsmuth, Stefan Kopp, and Petra Wagner. ALICO: A multimodal corpus for the study of active listening. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation, 26-31 May, Reykjavik, Iceland*, pp. 3638–3643, 2014.

[21] ASM Iftekhar Anam, Shahinur Alam, and Mohammed Yeasin. Expression: a google glass based assistive solution for social signal processing. pp. 295–296. ACM Press, 2014.

[22] George Galanakis, Pavlos Katsifarakis, Xenophon Zabulis, and Ilia Adami. Recognition of simple head gestures based on head pose estimation analysis. *AMBIENT*, Vol. 2014, pp. 88–96, 2014.

[23] James W. Davis and Serge Vaks. A perceptual user interface for recognizing head gesture acknowledgements. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pp. 1–7. ACM, 2001.

[24] Shaowei Chu and Jiro Tanaka. Head nod and shake gesture interface for a self-portrait camera. In *The Fifth International Conference on Advances in Computer-Human Interactions (ACHI 2012)*, pp. 112–117, 2012.

[25] Hatice Gunes and Maja Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *International conference on intelligent virtual agents*, pp. 371–377. Springer, 2010.

[26] Salah Saleh and Karsten Berns. Nonverbal communication with a hu-

manoid robot via head gestures. pp. 1–8. ACM Press, 2015.

[27] Geovany A. Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction*, pp. 396–406. Springer, 2011.

[28] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 32, No. 9, pp. 1627–1645, 2010.

[29] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1, pp. 1–8. IEEE, 2015.

[30] Lei Shen and Junlin Zhang. Empirical evaluation of rnn architectures on sentence classification task. *arXiv preprint arXiv:1609.09171*, 2016.

発表リスト

[1] Wu Yang, Kai Akiyama, Kris Kitani, László Jeni, Yasuhiro Mukaigawa, “Head Gesture Recognition in Spontaneous Human Conversations: A Benchmark”, CVPR 4th Workshop on Egocentric (First-Person) Vision, July 2016.
[2] 秋山解, 伍洋, Kris Kitani, László Jeni, “自然会話における頭部動作の検出”, 情処研報 CVIM 203, September 2016.

付録 A　被験者に提供された話題候補の一覧

会話の話題

話題は撮影の前に話し相手の方と相談して決めていただきます．以下の話題リストから自由に 4 つ程度選び，覚えてください．会話中はリストを見ることができません．どれから始めても構いません．ひとつの話題について語りきったと感じたら別の話題に進んでください．選んだすべての話題について話す必要はありませんが，各グループから最低 1 つは話してください．（終了 5 分前や 2 分前のアナウンスを参考にして話題を切り替えても結構です．）質問の趣旨に従って，あまり脱線しすぎないようにしてください．

グループ 1

どちらが良いですか？それぞれの良いところ，悪いところは何ですか？

- きのこ，たけのこ
- カピバラ，アルパカ
- 紙の書籍，電子書籍
- 任天堂 (Wii)，SCE (PlayStation)
- Windows，Mac
- LaTeX，Word
- 就職，進学

グループ 2

- あなた方は明日，計 1 千万円を手に入れます．
  2 人で 1 週間以内に使い切る計画を立ててください．
- あなた方はこれから，2 人で喫茶店を立ち上げます．
  他に類を見ないどんな特徴で客を集めますか？
- あなたの生き写しのクローンが現れました．
  どう対処していきますか？
- 神様に頼めば何でも 2 つ手に入るとすれば，何を頼みますか？
- あなた方は江戸時代初期にタイムスリップしました．
  今あなた方が伝授できるどんな知識や技術を教えれば，最も感謝されるでしょうか？

付録 B　撮影された映像のサンプル

図 B.1　固定カメラの映像の例

図 B.2　ウェアラブルカメラの映像の例

図 B.3　アイカメラの映像の例