

一人称視点映像を用いたランキング学習による 相対的地位の推定

樋口 未来^{1,a)} 米谷 竜¹ 木谷 クリス 真実² 佐藤 洋一¹

概要: 本稿では、ソーシャルインタラクションのさらなる理解を目的に、話者間の相対的地位を推定する手法を提案する。地位とは、例えば教授と学生等の社会的立場であり、相対的地位は、相手が目上か目下かといった社会的関係である。我々は、人との接し方が相手に応じて変化することに着目し、会話中の非言語コミュニケーションを解析することにより相対的地位を推定する。ここで、相手がどの程度目上かといった、相対的地位の大きさを定量的に定めることが難しいことが問題となる。これに対し、本研究では、相対的地位の大小関係（例えば、学生から見た教授の相対的地位は、学生から見た助教の相対的地位よりも大きい）を決めることは難しくないということに注目し、大小関係の拘束条件を用いたランキング学習により相対的地位を求める。また、人との接し方の差異を抽出するために、会話相手を近距離かつ正面から撮像可能な一人称視点映像を用い、深層畳み込みニューラルネットワーク（CNN）により特徴量を抽出する。さらに、一人称視点映像のカメラ装着者の頭部の動きの影響を除去するために、映像中の顔の座標を基に画像を補正し、プーリング処理を行う Face-aligned spatio-temporal pooling を提案する。実験では、28組の対面会話映像を用いて提案手法の有効性を確認した。

Estimating Relative Social Status using Learning-to-Rank in First-person Perspectives

MIRAI HIGUCHI^{1,a)} RYO YONETANI¹ KRIS M. KITANI² YOICHI SATO¹

Abstract: We address the novel problem of *relative social status estimation* which aims to estimate the relative difference in the social status of one person compared to another using passive wearable first-person cameras mounted on heads. While relative social status is highly subjective and hard to measure quantitatively, it has a measurable effect on human behavior (*e.g.*, rigid posture, nods) during a social interaction. In other words, people adapt their behavior adequately according to the social context of interaction. It is, therefore, possible for us to compare relative differences between two social interaction scenes, namely, in which scene people have a wider gap in their social status. Based on these observations, we develop a data-driven ranking algorithm that learns this comparison between paired interaction sequences to regress the relative social status for a new interaction sequence. To fully exploit subtle behavior induced by various body parts, we introduce a face-aligned spatio-temporal feature pooling scheme using convolutional neural networks (CNN). We present a dataset of 28 different face-to-face conversations in realistic environments. Our experiments show that our framework is able to robustly categorize different degrees of relative social status from first-person videos, and reveal which human behaviors are important for estimating relative social status.

¹ 東京大学生産技術研究所
Institute of Industrial Science, the University of Tokyo,
Meguro-ku, Tokyo 153-8904, Japan

² カーネギーメロン大学
Carnegie Mellon University, Pittsburgh, PA 15213

^{a)} mhiguchi@iis.u-tokyo.ac.jp

1. はじめに

本研究では、一人称視点映像中の人と人とのインタラクションを解析することで、2人の人物間の社会的地位の差（相対的地位）を推定することを目的とする。社会的地

位を推定することは、コミュニケーション支援、ライフログ [1], インタラクティブロボット [2], さらには心理学分野のための人と人のインタラクションの自動解析など、さまざまな応用が考えられる。本研究では、所属する組織内の地位、研究経験、社会経験に起因する、人物間の社会的地位の差（相対的地位）を推定する。

本研究が推定対象とする相対的地位は、人と人のコミュニケーションに様々な影響を及ぼしている。例えば、社会的地位が同等の2人の会話であれば、人はより楽な姿勢を取り、より多くのハンドジェスチャを用いる（図1(a), (b)）。逆に、教授と学生のように目上と目下の関係にある場合は、目下の人物はより正しい姿勢を取り、大きなハンドジェスチャの使用は控える傾向にある（図1(c), (d)）。具体的には、図1(c)のように、目下の人物の視点の映像中では会話相手が画像の中央付近に映っている。これは、カメラの装着者である目下の人物が正しい姿勢を維持しているためである。また、図1(d)のように、目上の人物の視点の映像中に映っている目下の人物は、図1(b)の同輩と話しているときと比べて、ハンドジェスチャーの使用頻度が低い。このように社会的地位が人の非言語コミュニケーションに影響を及ぼすことは、心理学の研究でも報告されている [3]。本研究では、このような会話相手に応じて生じる非言語コミュニケーションの差異を手掛かりに、“ある人物が会話相手をどの程度目上、目下と見なしているか”という相対的地位を推定する。

非言語コミュニケーションの解析に基づいて、相対的地位を推定するためには、以下の2つの問題を解決する必要がある。

- (1) 社会的地位の真値: 相対的地位のような相手に対してどう感じているかという概念は主観的であるため、定量的なラベル、数値を与えることは必ずしも容易ではない。コンピュータで社会的地位を推定するためには、このようなラベル付け、数値化の難しい概念を扱うことのできる手法が必要となる。
- (2) 特徴量の抽出: 人は人とコミュニケーションを取る際に、会話相手に応じて柔軟に接し方を調整する。会話相手に応じて生じる非言語コミュニケーションの差異は、頷きの大きさや速度、体の姿勢といった微小なものである。相対的地位を推定するためには、そのような微小な動きや姿勢の変化を観測する必要がある。

1つ目の問題を解決するために、本稿では教師有り機械学習の一つであるランキング学習を導入する。相対的地位は、定量的に数値化できないが、2つの相対的地位を比較するとその大小関係は人にとって判断が容易である。例えば、学生からみた教授の相対的地位は、学生間の相対的地位よりも明らかに大きい。2つの相対的地位の間の大小関係を拘束条件として用い、ランキング学習により話者間の社会的地位がどの程度離れているかを推定する。

2つ目の問題に対しては、固定カメラによるインタラクションの解析 [5], [6] の代わりに、2台のウェアラブルカメラで撮影した一対の一人称視点映像を用いることで問題の解決をはかる。会話相手によって生じる非言語コミュニケーションの微小な差異を観測するためには、人の真正面にカメラを設置する必要があるが [7], [8], 固定カメラを日常生活の様々な場所に導入することは難しい。一方、頭部装着型のカメラで撮影できる一人称視点映像は、インタラクションの対象を真正面かつ近距離で、自然な状況下で撮影できるという利点がある。また、2人のインタラクションにおいて、お互いにカメラを装着することで得られる一対の一人称視点映像を用いることが、微小な動作、反応を認識するために有効であることが報告されている [9]。そこで、本研究では、2人の頭部に装着した2台のカメラから得られる一対の一人称視点映像を用いて、2種類の特徴量を抽出する。1つ目は、カメラ装着者の頭部の動きに関する特徴であり、カメラの自己運動を特徴量として用いる。2つ目は、映像に映っている上半身の動きとアピランスの特徴を、深層畳み込みニューラルネットワーク (CNN) により抽出する。しかしながら、カメラ装着者の頭部の動きが頻繁に生じる一人称視点映像から、映像に映っている上半身の特徴量を安定して抽出することが難しい。そこで我々は、Face-aligned spatio-temporal pooling を提案する。この提案手法により、通常の動作認識で用いられる画像全体の特徴量プーリングと比較して、カメラ装着者の頭部の動きの影響を除去することで、頑健に特徴量をプーリングすることができる。

本研究のアイデアを下記にまとめる。

- 相対的地位の推定: 人物間の社会的地位の差（相対的地位）がインタラクション中に推定可能であることに着目する。会話相手に応じて非言語コミュニケーションに差異が生じることを手掛かりに、相対的地位を推定可能であることを示す。
- ランキング学習の導入: 相対的地位は、定量的に数値化できない概念であるが、2つの相対的地位を比較するとその大小関係は人にとって判断が容易であることに着目し、教師有り機械学習の一つであるランキング学習を導入する。人と人のインタラクションの解析に、ランキング学習が適用できることを示す。
- 一人称視点映像による非言語コミュニケーションの解析: 固定カメラの代わりに、一人称視点映像を用いることで、屋外等の日常的なシーンで利用できる枠組みを提案する。また、話者それぞれにウェアラブルカメラを装着し、お互いを真正面かつ近距離から撮影することで、頭部の動き、上半身の動きの両方を精度良く解析する。
- CNNを用いた Face-aligned spatio-temporal pooling による上半身の特徴量抽出: CNNを用いることで、手、



図 1 対面会話中の一人称視点映像の例。

肘、肩等の身体の部位を検出することなく、上半身の動き、姿勢に関する特徴量を抽出する。さらに提案手法の Face-aligned spatio-temporal pooling により、カメラ装着者の頭部の動きが頻繁に生じる一人称視点映像を用いた場合であっても、安定した特徴量の抽出、およびプーリングが可能となる。

提案手法の有効性を確認するために、28 組の 1 対 1 の対面会話映像（合計 56 個の一人称視点映像）をウェアラブルカメラにより収集した。実験では、提案手法により相対的地位を精度良く推定できることを確認するとともに、相対的地位の推定にどの特徴が有用であったかを明らかにした。

2. 関連研究

ソーシャルインタラクション認識: 非言語コミュニケーションから、性格や会話の主導権といった人間の社会的な属性を推定する研究が、近年盛んにおこなわれている [10]。従来研究に共通するアプローチとして、教師有りの識別器を用いて問題を定式化する手法があげられる。教師有りの 2 クラス識別を用いた従来研究としては、非言語コミュニケーションの特徴から、性格 [5], [6]、会話の主導権 [7] を識別する研究があげられる。一方、本稿で認識対象とする相対的地位は定量的な真値を与えることが難しく、一般的な学習による識別器や回帰推定を適用することができない。

Jayagopi らは、それぞれの被験者にプロジェクトマネージャー等の社会的役割を割り当て、グループ会議中に会話を主導している人物を推定する手法を提案した。一方、我々は、社会的な役割を割り当てることなく、所属する組織内の地位、研究経験、社会経験に起因する社会的地位の差を求める [7]。さらに最近では、熊野らがグループでの会話中における話者の感情（同意、非同意など）を、話者の顔の向き、顔の表情、会話に対する反応時間から、ページアンネットワークにより推定する手法を報告している [8]。これらの従来研究も、人と人のインタラクションに関する問題を扱っているが、社会的地位という概念は扱って

ない。

心理学の研究では、社会的地位がインタラクション中の人の振る舞いにどのように影響するかを、手作業で解析した結果を報告している [3]。心理学の従来研究と異なり、本研究の目的は、人と人の社会的な関係を自動的に推定する枠組みを開発することである。

一人称視点映像を用いた行動認識: 近年、一人称視点映像が、行動認識、インタラクションの解析などの様々な応用に有用であることが報告されている [11]。Fathi らは、カメラ装着者の自己動作と物体の関連性（近接性、動きの方向）を解析することで、食事の準備の行動を認識する手法を提案した [12], [13]。また、大垣らは、カメラ装着者の視線の動きと頭部の動きから、書く、本を読むといった室内の行動を認識する手法を [14]、Kitani らは、頭部の動きの方向や頻度から屋外の行動を認識する手法を報告している [15]。さらに、カメラのエゴモーションの特徴から長時間にわたるカメラ装着者の行動を推定する試みや [22]、エゴモーションの特徴を時系列方向にプーリングすることでカメラ装着者の多様な行動を分類する試みがなされている [16]。これらの従来研究はカメラ装着者の動作、行動の推定には有用であるが、人と人とのインタラクションにおける人物間の相対的地位の推定に直接適用することはできない。

近年、一人称視点映像を用いた人と人のインタラクションの解析に関する研究も行われている。Fathi らは、一人称視点の向き、周囲の人物の顔の向きから、グループインタラクションが 1 対 1 の会話か、1 対多の会話か、複数人での議論かを推定する手法を提案している [17]。さらに大きなグループインタラクションにおいて、複数の人物が同時に注視している領域を推定する研究が報告されている [18], [19]。また、Ryoo らは、一人称視点映像のグローバルモーションと、映像中のローカルモーションから握手、ハグといった人と人のインタラクションを推定する手法を [20]、米谷らは、お互いにカメラを装着して得られる一対の一人称視点映像を用いて、微小な動作、反応を認識する手法を提案した [9]。一方、本研究では、一対の一人称視点映像を用いて、会話相手に応じて生じる顔つきや姿勢の変化といった非言語コミュニケーションの僅かな変化を解析し、話者間の相対的地位を推定する。

3. ランキング学習による相対的地位の推定

相対的地位は定量的に数値化できない概念であり、一般的な学習による回帰推定を適用できないことが問題となる。そこで我々は、2 つの相対的地位を比較した場合、その大小関係は人にとって判断が難しくないことに着目する。例えば、学生からみた教授の相対的地位は、学生間の相対的地位より明らかに大きい。本稿では、この相対的地位間の大小関係を拘束条件として用いることができる、ランキン

グ学習を導入する。

また、2人の話者の頭部に装着した2台のカメラから得られる一対の一人称視点映像を用いて、頭部の動きの特徴、上半身の特徴（姿勢、手の動き、体の動きなど）を抽出する。

3.1 ランキング学習

本研究の目的は、会話中の2人の話者間の社会的地位を、一対の一人称視点映像から推定することである。シンプルな手法として、機械学習の枠組みを用いて、2つの一人称視点映像から得られる特徴量を相対的地位に直接マッピングする手法が考えられる。具体的には、映像の組を (V_i^A, V_i^B) 、映像の組から抽出した特徴量ベクトルを $\Phi(V_i^A, V_i^B)$ とすると、特徴量ベクトル $\Phi(V_i^A, V_i^B)$ から相対的地位のスコア R_i ($R_i \in \mathbb{R}$) を計算するための関数を求める手法が考えられる。ただし、 V_i^A と V_i^B は、学習用データ中の i 番目の会話映像における人物 A と人物 B のウェアラブルカメラにより取得した映像である。例えば、線形回帰であればパラメータベクトル w^T を用いて、 $R_i = w^T \Phi(V_i^A, V_i^B)$ と書ける。

しかしながら、相対的地位は曖昧な概念であり定量的に定義することができないため、前述の手法では相対的地位のスコア R_i を推定することが難しい。そこで、我々は2組の会話映像の間の順序関係 $(V_i^A, V_i^B) \succ (V_j^A, V_j^B)$ if $R_i > R_j$ を導入する。そして、ランキング学習手法の一つである RankSVM [4] を用いて、パラメータベクトル w を次式の拘束条件に基づいて求める。

$$(V_i^A, V_i^B) \succ (V_j^A, V_j^B) \Leftrightarrow w^T \Phi(V_i^A, V_i^B) > w^T \Phi(V_j^A, V_j^B). \quad (1)$$

テスト時は、学習したパラメータベクトル w を用いて、未知の一対の一人称視点映像から人物 A の人物 B に対する相対的地位をランキングスコアとして計算する ($R_{\text{new}} = w^T \Phi(V_{\text{new}}^A, V_{\text{new}}^B)$)。この提案手法では、2つの相対的地位の大小関係を拘束条件として用いることで、相対的地位の大きさ（どの程度会話相手が目上か、目下か）を推定できる。

3.2 特徴量抽出

心理学分野の研究で、社会的地位がインタラクション時の非言語コミュニケーションに影響を及ぼすことが報告されている [3]。また、人間は、自分の真剣さや感情といった内部の心理状態を非言語な手段（顔、姿勢、ハンドジェスチャー等）で相手に伝達していると考えられる。そこで、我々は、相対的地位の推定には、話者の頭部の動き、姿勢、手のジェスチャーなど非言語コミュニケーションが重要であると考え、それらの特徴量として抽出する。

具体的には、2人の人物が頭部に装着した2台のカメラ

から得られる一対の一人称視点映像を用いて、各人物の頭部の動き、および上半身の動きとアビアランスに関する特徴量を抽出する。頭部の動き特徴には、一人称映像のグローバルな動きを用いることができる。つまり、人物 A が装着したカメラから得られる一人称視点映像を用いて、人物 A の頭部の動きの特徴を求める。また、人物 B が装着しているカメラから得られる映像を用いて、人物 A の上半身の特徴を抽出する。

頭部の動き特徴の抽出と Temporal Pooling: 頭部の動きの特徴として、カメラの自己運動から求められる Cumulative displacement (CD) curves を用いる [22]。CD curves は、事前に定義された小領域毎のローカルモーションを長時間の一人称視点映像から求め、累積することで求められる。図 2 に示すように CD curves には、カメラ装着者の頭部の非言語コミュニケーション（顔や姿勢の変化）の傾向が含まれる。この CD curves から、CD Curves の傾き、動きの大きさなどの 1 フレーム毎の特徴量を生成する [22]。

しかしながら、1 フレーム毎の特徴量は、歩行や自動車の運転といった長時間の比較的単調な行動認識のために設計された特徴量であるため、会話のように相手の反応や会話の流れに柔軟かつ突発的に調整される人間の振る舞いの描写には適していない。そこで、我々は、Pooling 処理により特徴量を変換する。本稿では、顔や姿勢の変化の頻度といった情報を抽出するために Temporal max pooling を用いる。カメラ装着者が前後方向に頭部を動かした場合、各小領域中のローカルモーションは画像上で放射状に現れる。この放射状に現れる特徴を保持するために、頭部の動き特徴では Spatial pooling は行わない。

上半身の特徴の抽出: 人物 A の上半身は、会話相手の人物 B のカメラで撮像された映像から観測できる。しかし、頭部に装着するウェアラブルカメラの映像は、カメラ装着者（人物 B）の頭部の動きが頻繁に発生するため、上半身の各部位（手、肘、肩）を安定かつ高精度に検出することが難しい。一方、そのようなカメラ装着者の頭部の動きの影響を大きく受ける一人称視点映像であっても、顔追跡技術 [23] を用いることで顔を安定して検出し、追跡することができる。

顔は安定して検出できることから、本研究では全フレームの顔の位置を基に一人称視点映像を補正し（Face-aligned 画像）、補正後の Face-aligned 画像系列を用いて Two-stream CNN [21] により上半身の特徴量を抽出する。まず、Lucas-Kanade 法 [24] により求めたホモグラフィ行列を用いて、各フレームを会話開始前の正しい姿勢で撮影した初期フレームの座標系に変換する。その後、上半身周辺の画像を、初期フレームの顔の座標を基に補正後の画像から切り出す（図 3）。この処理により、カメラ装着者の頭部が頻繁に動く場合でも、上半身の動きとアビアランスに関する特徴を

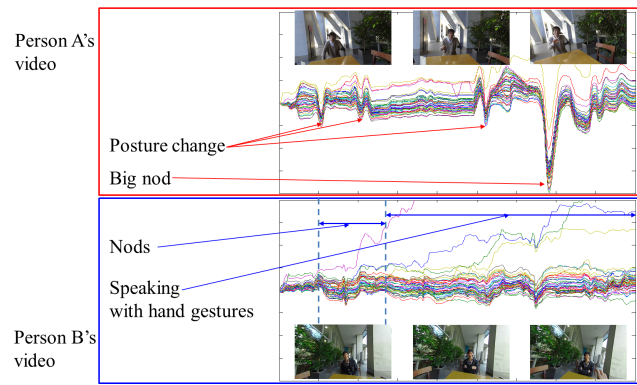


図 2 Cumulative displacement curves (y 成分) の例。人物 A は姿勢を何度か変え、その後大きくうなずいている。人物 B は何度か頷いた後、ハンドジェスチャーを使いながら発話している。

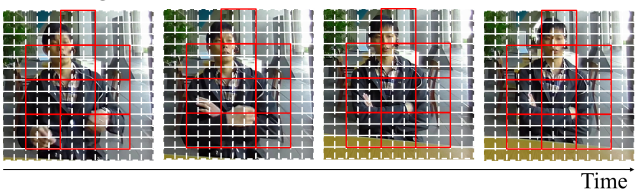


図 3 Face-aligned 画像の例

正しく抽出することが可能となる。

本稿では、動作認識用のデータセット (UCF101 [25] および HMDB51 [26]) を用いて学習済みの Two-stream CNN を特徴抽出器として採用し、Two-stream CNN の中間層の出力を上半身の特徴量とする。また、Two-stream CNN の入力データは、face-aligned 画像系列から求めたオプティカルフローから、face-aligned 画像のグローバルな動きを除去した局所的な動きベクトルの系列とする。

Face-aligned spatio-temporal pooling: 人と人のインタラクションは、相手の反応や会話の流れなどに応じて適応的に調整されるため、いつ、どの身体部位に相対的地位に関連する非言語コミュニケーションが現れるか定義できない。そこで、上半身の特徴に関しては、時系列方向と空間方向に特徴量のプーリング処理を行う。図 3 に示す通り相対的地位に関連する非言語コミュニケーション (ハンドジェスチャーや頭部の姿勢など) は、Face-aligned 画像上であっても様々な位置で観測され得る。体の姿勢推定の技術を用いても、手、肘、肩などの体の部位の位置を安定して求めることは難しいため、体の部位の位置同定は行わず図 3 の赤い矩形 (グリッド) に示す固定の領域を用いる。

まず、Two-stream CNN を用いて各セル毎 (白い矩形) にローカルモーションとローカルアピランスの特徴量系列を抽出する。その後、すべてのグリッド内で Max pooling 関数により時空間方向のプーリングを行う。

最終的には、一対の一人視点映像から人物 A の頭部の動き特徴と上半身の特徴、および人物 B の頭部の動き特徴と上半身の特徴を求め、顔の動き特徴と上半身の特徴をそれ

ぞれ PCA により次元圧縮を施し、連結することで特徴量ベクトル $\Phi(V_i^A, V_i^B)$ を生成する。

4. 実験

提案手法の有効性を確認するために、新たな一人称視点映像データセットを構築し、次節で述べる通り手作業で相対的地位の大小関係のアノテーションを行った。本章で述べる実験の目的は、1) ランキング学習を用いた提案手法により相対的地位が推定であることを確認すること、2) どの特徴量が重要かを明らかにすることである。

比較手法として、提案手法の Face-aligned spatio-temporal pooling の代わりに、画像全体で Spatio-temporal pooling を行う手法を実装した。また、頭部の動き特徴のみを用いた場合と、上半身の特徴のみを用いた場合を比較した。

データ収集: 評価実験のために、様々な組み合わせの 2 人の被験者が頭部にカメラ *1 を装着し、データセット 1 とデータセット 2 の合計で 28 組の会話映像 (56 個の一人称視点映像) を収集した。2 人の被験者は、椅子に着座した状態で机越しに会話をした。すべての会話シーンの映像の長さは、5400 フレームから 7200 フレームとした。図 4 に撮影した会話シーンの対となる一人称視点映像の例を示す。2 人の被験者は、ハンドジェスチャーや頷きといったボディランゲージを用いると同時に、体の姿勢や手の位置を変えながらお互いに会話していることが見て取れる。

被験者: すべての被験者は、大学院生か博士研究員のアジア人とした。データセット 1 の被験者は、男性 5 人と女性 2 人であった。また、データセット 2 の被験者は男性 3 人であり、うち 2 人はデータセット 1 の被験者に含まれない。被験者は、2 人が一組となって研究について自由に会話をした。

データセット 1: データセット 1 は、22 個の会話シーンから成る。全会話シーンに対して、被験者の研究室内の実際のポジションに基づいて相対的地位の大小関係のアノテーションを行った (表 1)。例えば、修士 1 年生からみた博士研究生の相対的地位は、修士 1 年生からみた修士 2 年生の相対的地位より大きい、といった具合にアノテーションを生成した。推定できる相対的地位の分解能を検証するために、アノテーション結果をそのまま用いる Full ranks, 分解能を落とした 7 ranks と 3 ranks の 3 種類のアノテーションを生成した。

データセット 2: データセット 2 では、被験者は実際の相対的地位ではなく、与えられた相対的地位 (目上, 目下) を演じた。このデータセット 2 は、6 個の会話シーンから成る。具体的には、人物 A が人物 B に対して目上、人物 B が人物 A に対して目上、人物 A が人物 C に対して目上

*1 Panasonic HX-A500 で解像度 1920×1080 の映像を 60fps で撮影した。

といったように、3人の被験者の全組み合わせで目上と目下を入れ替えて、会話シーンの一人称視点映像の組を撮影した。

評価方法: 推定性能を評価するために、データセット1の22個の会話シーンのうち20個のシーンを用いてRankSVMの学習を行い、残り2個の会話シーンで性能を評価することを、学習データとテストデータを入れ替えて11回行った(11-分割交差検証)。同一の被験者の組が2シーンずつ存在するため(例えば、シーン1では人物Aと人物Bが会話、シーン2では人物Bと人物Aが会話)、同一の被験者の組が学習データとテストデータに含まれないように11-分割交差検証とした。

また、データセット1の22シーンから115組のビデオクリップ、データセット2から33組のビデオクリップを生成した。ビデオクリップは、1800フレームの長さで、900フレームの重複を許して各会話シーンから切り出した。ランキングスコア R_i は各ビデオクリップ毎に求め、各ビデオクリップ毎に評価した。本稿では、上半身の特徴と頭部の動き特徴をそれぞれPCAで64次元に圧縮した。したがって、2つの一人称視点映像から得られるすべての特徴量は256次元とした。また、提案手法で用いるRankSVMには、すべての実験で線形カーネルを用いた。

評価指標: 相対的地位の推定結果を評価するために、本稿では、Kendall rank correlation coefficient (KRCC)を用いた。ここで、 $(Q_1, R_1), (Q_2, R_2), \dots, (Q_N, R_N)$ をアノテーションから得られる真値 Q と推定結果 R のペア、 N をテストデータの総数とすると、KRCCは式2と定義できる。ただし、 N_c は順序関係が一致したペアの総数(if $Q_i > Q_j$ and $R_i > R_j$ or if $Q_i < Q_j$ and $R_i < R_j$), N_d は順序関係が一致しなかったペアの総数(if $Q_i > Q_j$ and $R_i < R_j$ or if $Q_i < Q_j$ and $R_i > R_j$), そして N_p は相対的地位が等しいペア($Q_i = Q_j$)を除いた総数($Q_i > Q_j$ and $Q_i < Q_j$ excepting $Q_i = Q_j$)である。

$$\tau = (N_c - N_d)/N_p \quad (2)$$

4.1 実験結果

表2に示す通り、Face-aligned spatio-temporal poolingを用いた提案手法により、Full ranksのアノテーションでKRCCを0.699に、7 ranksのアノテーションでKRCCを0.664に、3 ranksのアノテーションでKRCCを0.914に改善することができた。また、表3の通り、上半身の特徴が頭部の動き特徴よりも相対的地位の推定に寄与していることがわかった。ただし、両方の特徴を用いたときが最も性能が高く、頭部の動き特徴も相対的地位の推定に必要であるといえる。

データセット2を用いた評価も行った。学習には、データセット1のすべてのデータ、および3 ranksのアノテ

	提案手法	比較手法
Full ranks	0.699	0.315
7 ranks	0.664	0.281
3 ranks	0.914	0.456

表2 提案手法と比較手法の推定精度の比較

ションデータを用いた。評価の結果、KRCCは0.427であった。学習データと異なる条件下のデータにおいても、提案手法により相対的地位を推定できることを確認した。また、人間が成長する過程で得た社会経験に基づいて相対的地位を演じた場合であっても、相対的地位を推定できることを確認した。

4.2 相対的地位に有用な特徴量の解析結果

上半身のどの部位の特徴が相対的地位の推定においてより重要かを明らかにするために、上半身の特徴の重要度を各グリッド毎に計算し、可視化した(図5)。特徴量ベクトル $\Phi(V_i^a, V_i^b)$ に各グリッド毎の上半身の特徴のみを用いてKRCCを計算し、そのKRCCの値を重要度とした。本実験には、データセット1の会話映像、および3 ranksのアノテーションデータを用いた。この結果から、頭部、肩、肘が、その他の部位より重要であることが分かる。これは、体、手の姿勢、姿勢の変化が、相対的地位の影響を強く受けているためと考えられる。

また、アノテーションデータのランク(目上、目下)毎に、各グリッドのランキングスコアの平均値を求め、可視化した(図6)。本評価実験では、2人の被験者の社会的地位が等しいデータは除外した。アノテーションデータのランクが目下の場合は、ランキングスコアが負の値を取るため正負を反転して可視化した。この実験結果から、一人称視点映像中に映っている会話相手がカメラ装着者よりも目上である場合、体の真正面の手の領域が相対的地位の推定に重要であることが分かった。

4.3 考察

提案手法により、インタラクション中の非言語コミュニケーションを手掛かりに、話者間の相対的地位を推定し得ることを確認した。しかし、男性と女性とでは、同じ相対的地位でも用いる非言語コミュニケーションが異なり得るが、提案手法では性別による差異は考慮していない。例えば、図4の下段のように女性は男性に比べてハンドジェスチャの使用頻度が低い傾向にあり、腕組みや姿勢を崩す動作も少ない。また、本稿では、組織内の地位、研究経験、社会経験に由来する社会的地位に着目したが、それ以外に権力、支配、富、特権階級などさまざまな要因が社会的地位に関連し得る。今後、推定性能を改善するためには、性別など様々な要因を考慮する必要がある。

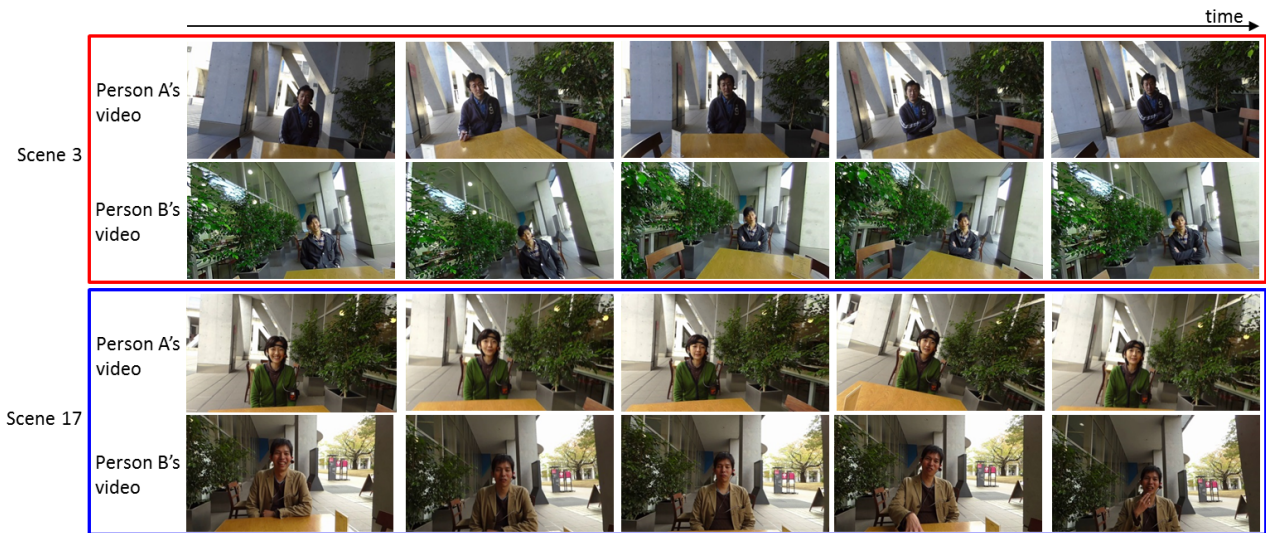


図 4 一人称視点映像の例

アノテーション	
full ranks	$2 < 5 = 10 < 13 = 15 < 20 < 3 = 8 = 18 = 21 < 11 = 12 < 22 < 4 = 7 = 17 < 19 < 14 = 16 < 6 = 9 < 1$
7 ranks	$2 = 5 = 10 < 13 = 15 < 20 = 3 = 8 = 18 < 21 = 11 = 12 = 22 < 4 = 7 = 17 = 19 < 14 = 16 < 6 = 9 = 1$
3 ranks	$2 = 5 = 10 = 13 = 15 < 20 = 3 = 8 = 18 = 21 = 11 = 12 = 22 = 4 = 7 = 17 = 19 < 14 = 16 = 6 = 9 = 1$

表 1 組織内の実際のポジションに基づいたアノテーション結果

	頭部の動き特徴	上半身の特徴
Full ranks	0.032	0.506
7 ranks	0.028	0.355
3 ranks	0.286	0.777

表 3 頭部の動き特徴と上半身の特徴の推定性能の比較



図 5 上半身の特徴の重要度

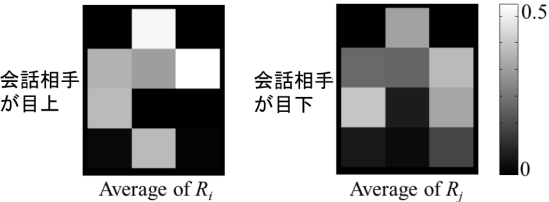


図 6 上半身の特徴の重要度 (相対的地位毎)

5. おわりに

本稿では、ソーシャルインタラクションのさらなる理解を目的に、一対の一人称視点映像を用いて話者間の社会的

地位の差 (相対的地位) を推定する手法を提案した。本研究では、話者の姿勢の変化やハンドジェスチャー等の非言語コミュニケーションが会話相手に応じて適応的に変化することに着目し、非言語コミュニケーションの特徴量を用いてランキング学習により相対的地位の大きさを推定した。また、カメラ装着者の頭部の動きが頻繁に生じる一人称視点映像から、映像中の上半身の特徴を安定に抽出するために CNN を用いた Face-aligned spatio-temporal pooling を提案した。実験では、28 組の対面会話映像を用いて提案手法の有効性を確認した。さらに、頭部の動き特徴と上半身の特徴の両方を用いることが相対的地位の推定に有効であるが、特に上半身の特徴が重要であることを明らかにした。

参考文献

- [1] M. Blum, A. S. Pentland, and G. Troster, "Insense: Interest-based life logging," IEEE MultiMedia, 13, 4, pp. 40–48, 2006.
- [2] O. Celiktutan, and H. Gunes, "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience," In Proc. of IEEE RO-MAN, 2015.
- [3] A. Leffler, D. L. Gillespie, and J. C. Conaty, "The effects of status differentiation on nonverbal behavior," In Social Psychology Quarterly, pp. 153–161, 1982.
- [4] T. Joachims, "Optimizing search engines using click-through data," In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 133–142, 2002.
- [5] B. Lepri, S. Ramanathan, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion - a systematic study," In T. Affective Com-

- puting, pp. 443–455, 2012.
- [6] S. Ramanathan, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, “On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions,” In Proc. ICMI, pp. 3–10, 2013.
- [7] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, “Modeling dominance in group conversations using nonverbal activity cues,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 17, No. 3, pp. 501–513, 2009.
- [8] S. Kumano, K. Otsuka, M. Matsuda, and J. Yamato, “Analyzing perceived empathy based on reaction time in behavioral mimicry,” In *IEICE Transactions*, pp. 2008–2020, 2014.
- [9] R. Yonetani, K. Kitani, Y. Sato, “Recognizing Micro-Actions and Reactions from Paired Egocentric Videos,” In Proc. of IEEE conf. on CVPR, pp. f3414–3421, 2016.
- [10] D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups: A review,” In *Image Vision Comput.*, pp. 1775–1787, 2009.
- [11] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, “The evolution of first person vision methods: A survey,” *CoPR abs/1409.1484*, 2014.
- [12] A. Fathi, A. Farhadi, and J. M. Rehg, “Understanding egocentric activities,” In Proc of ICCV, 2011.
- [13] A. Fathi, X. Ren, and J. M. Rehg, “Learning to recognize objects in egocentric activities,” In Proc. of IEEE Conf. on CVPR, pp. 3281–3288, 2011. pp. 1226–1233, 2012.
- [14] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, “Coupling eye-motion and ego-motion features for first-person activity recognition,” In *CVPR Workshops on Ego-Centric Vision*, 2012.
- [15] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, “Fast unsupervised ego-action learning for first-person sports videos,” *Proc. of IEEE Conf. on CVPR*, pp. 3241–3248, 2011.
- [16] M. S. Ryoo, B. Rothrock, and L. Matthies, “Pooled motion features for first-person videos,” *CoRR abs/1406.2199*, 2014.
- [17] A. Fathi, J. K. Hodgins, and J. M. Rehg, “Social interactions: A first-person perspective,” In Proc. of IEEE Conf. on CVPR, pp. 1226–1233, 2012.
- [18] H. S. Park, and J. Shi, “3D social saliency from head-mounted cameras,” In *NIPS*, pp. 431–439, 2012.
- [19] H. S. Park, and J. Shi, “Social saliency prediction,” *Proc. of IEEE Conf. on CVPR*, 2015.
- [20] M. S. Ryoo, and L. Matthies, “First-person activity recognition: What are they doing to me?,” *Proc. of IEEE Conf. on CVPR*, pp. 2730–2737, 2013.
- [21] K. Simonyan, and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *CoRR abs/1406.2199*, 2014.
- [22] Y. Poley, C. Arora, and S. Peleg, “Temporal segmentation of egocentric videos,” In Proc. IEEE Conf. on CVPR, pp. 2537–2544, 2014.
- [23] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” In Proc. IEEE Conf. on CVPR, pp. 532–539, 2013.
- [24] S. Baker, R. Gross, and I. Matthews, “Lucas-kanade 20 years on: A unifying framework: Part 3,” *International Journal of Computer Vision* 56, pp. 221–255, 2002.
- [25] K. Soomro, A. R. Zamir, and M. Shah, “MCF101: A dataset of 101 human actions classes from videos in the wild,” *CoPR abs/1212.0402*, 2012.
- [26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: a large video database for human motion recognition,” in *Proc of ICCV*, 2011.