# 複数人物の対面会話シーンを対象とした画像中の人物頭部追跡に基づく会話構造のモデル化と確率的推論

大塚　和弘[†,††]　　大和　淳司[†]　　村瀬　　洋[††]

† NTT コミュニケーション科学基礎研究所　〒 243–0198 神奈川県厚木市森の里若宮 3–1
†† 名古屋大学大学院 情報科学研究科　〒 464-8601 愛知県名古屋市千種区不老町
E-mail: †{otsuka,yamato}@eye.brl.ntt.co.jp, ††murase@is.nagoya-u.ac.jp

**あらまし**　本稿では，複数人物の対面会話シーンを対象とし，画像中の人物頭部追跡に基づいて会話構造のモデル化，及び，確率的推論を行う方法を提案する．本研究では，会話の構造として，会話参加者とその役割 (話し手，受け手など) の組合せに着目し，この会話の構造を参加者の行動から推定することを目標とする．ここでは参加者の行動として視線行動，つまり「人を見るという行動」に着目し，視線行動と会話の構造との関係を動的ベイジアンネットによりモデル化する．また，疎テンプレート Condensation 追跡法を導入し，視線行動の手掛かりとして各参加者の頭部方向を計測する．計測された頭部方向，及び，発話状態に基づき，マルコフ連鎖モンテカルロ法を用いて会話の構造を推定する方法を提案する．最後に，4 人会話を対象とした実験結果を示し，提案方法の有効性を確認する．
**キーワード**　対面会話，頭部追跡，視線，非言語行動，動的ベイジアンネット，マルコフ連鎖モンテカルロ法

# Modeling and Probabilistic Inference of Conversation Structures in Multiparty Face-to-face Setting Based on Visual Head Tracking

Kazuhiro OTSUKA[†,††], Junji YAMATO[†], and Hiroshi MURASE[††]

† NTT Communication Science Laboratories　3-1, Morinosato-Wakamiya, Atsugi, 243-0198
†† Graduate School of Information Science, Nagoya University　Furo-cho Chikusa-ku Nagoya 464-8601
E-mail: †{otsuka,yamato}@eye.brl.ntt.co.jp, ††murase@is.nagoya-u.ac.jp

**Abstract**　Novel model and method for probabilistic inference of conversation structures in multiparty face-to-face setting are proposed based on head tracking of the participants from video sequences. This study aims to identify the basic structure of conversations such who is talking to whom, by estimating the combination patterns of participants and their roles including speaker and addressees. To that end, our study focuses on participants' gaze behavior and models the relationship between gaze behavior and conversation structures using a dynamic Bayesian network. As a cue for determining gaze behavior, a visual head tracker based on sparse template Condensation is used to measure the head directions of participants. Using the head-direction data and utterance data, the conversation structures are identified using a Markov chain Monte Carlo method. Finally, experiments on four-person conversation confirm the effectiveness of the proposed model and method.
**Key words**　Face-to-face conversation, head tracking, eye gaze, non-verbal behavior, dynamic Bayesian network, Markov chain Monte Carlo method

## 1. はじめに

複数人物による対面会話は，情報の伝達・共有，他者の意図・感情の理解，グループの意思決定などにおいて，欠かすことができないコミュニケーションの一形態である．近年，空間や時間の壁を越えた他者との会話コミュニケーションを支援する技術として，遠隔会議システムや会議映像の自動編集／アーカイブ，会話エージェント／ロボットなどの発展に期待が寄せられている．現在，そのための基盤技術として，人間同士の対面会話シーンの自動的な認識・理解という研究テーマに注目が集まりつつあり，これまでにも，会議・会話中の各参加者の行動やグループ行動を認識する方法として，隠れマルコフモデルやその拡張である HMMs [1], Layered-HMM [2], Coupled-HMM [3], 動的ベイジアンネットワーク [4] などを用いた方法が提案され

ている．その一例として，McCowan らは，観測された画像情報（顔や手の位置・動き等）と音響情報（ピッチ，エネルギー等）に基づいて，会議中のグループ行動（発表，議論，ノート書き等）を HMMs により認識する方法を提案している．この方法を含め，これまで提案されている方法では，まず，観測された画像や音響信号から各人物の行動を認識し，その後，各人物の行動の直接的な組み合わせとしてインタラクションの認識を行うというアプローチに主眼が置かれている．しかし，このようなアプローチには，参加者の行動の背後にある心理的・社会的な側面を含めた会話の場の認識・理解という観点が欠落しており，その点より，現状では，会話シーンの自動認識・理解の研究は未だ萌芽期にあるといえる．

従来，複数人物による対面会話は，社会心理学等の分野において研究が行われており，参与枠組や参与役割と呼ばれる観点から会話の構造を捉える方法が知られている [5]．Goffman の参与枠組 [5] においては，会話参加者は，承認された参加者と承認されていない参加者（立ち聞きしている者）に分類され，さらに，承認された参加者は，話し手，受け手，及び，傍参与者に分類される．話し手，受け手などの役割は参与役割と呼ばれている．ここで，受け手は，話し手に話し掛けられている人物を指し，傍参与者は，承認された参加者のうち話し手でも受け手でもない人物を指す．話し手は，発話権（ターン）を保持している者として位置付けられる．また，受け手と傍参与者は聞き手とも呼ばれる．このような会話中の各参加者の役割は，会話の構造を表現するための不可欠な要素であると考えられる．従来，会話シーンの自動認識・理解の研究では，話し手の検出に主眼が置かれていたが，我々は，話し手以外の役割の推定も様々なアプリケーションにとって重要であると考える．例えば，文献 [6] では，会話映像の自動編集に関する実験を行い，話し手のみの映像や参加者全員を一画面に収めた映像よりも，話し手と受け手の映像が交互に切り替わる映像の方が，視聴者に対して「誰が誰に話し掛けているか」といった会話の構造をより正確かつ明瞭に伝達できることを示唆している．

著者らは，従来の直接的な行動認識のアプローチから一歩進め，会話参加者の行動が生成される背後にある会話現象の性質や構造に着目し，それらと人物の行動との関係についてモデルを構築し，インタラクションの認識を行うというアプローチを提案している [7], [8]．その第一歩として，本稿では，会話参加者とその参与役割との組み合わせに着目し，会話の各時点でそれらを推定することで，誰が誰に話しかけているか，誰が誰の話しを聞いているか，というような会話の基本的な構造を自動的に同定することを目標とする．また，そのための手掛かりとして，以下の観点より会話参加者の視線行動，つまり「人を見るという行動」に着目する．まず，これまでの社会心理学分野での研究によると，対面会話を遂行する上では，言語的な情報のみならず，非言語的な行動による情報の伝達・交換も重要な要素であることが指摘されている [9]．このような非言語行動には，視線や顔の表情，頷き，手振り・身振り，姿勢などがあるが，その中でも視線の役割は重要視されている [10], [11]．例えば，Kendon は，視線には，他者の行動をモニタリングする機能，自らの態度や意図を表出する機能，会話の流れを調整する機能が具わっていることを示唆している [10]．また，Goodwin は，会話の成立の上で，話し手と受け手の間の相互の視線インタラクションが不可欠であると主張している [11]．

以上のように，会話参加者の視線行動は，会話構造の推定の手掛かりとして有効であると考えられるが，その実現のためには，実際に会話中の視線行動を計測する必要がある．これまでにも人間の視線方向を計測する技術が提案されているが [12], [13]，自然な会話を妨げないように視線方向を正確，かつ，安定に計測することはいまだ困難なタスクとされる．そこで本研究では，より容易に計測が可能な頭部の姿勢・方向から視線の方向を推測するというアプローチを採用している．このアプローチは，人間には興味の対象をその視野の中央で捉えようとする性質があり，それにより，視線を向けた先の人物との位置関係に応じて，頭部や胴体の姿勢が変化する，という性質に立脚するものである．このアプローチの先駆的な例として，Stiefelhagen らは，4 人会話において，頭部方向に基づいて，誰が誰を見ているか判定可能であることを示している [14]．

また，本研究では，会話の各時点における参加者の行動は，その時点の会話の構造によって規定されるという仮説を立て，会話という現象を，会話の構造に相当する上位プロセスと参加者の行動という下位のプロセスとの間の相互作用によって時間発展する系と見立てて，確率的なモデルを構築し，このモデルに基づいて会話構造を推定するというアプローチを提案している [7], [8]．具体的な会話モデルとしては，マルコフ切替モデル (Markov-Switching Model) [15] と呼ばれる一種の動的ベイジアンネットワークを採用している．このモデルにおいては，視線方向は直接的な計測の対象ではなく，推定すべき変数の一つとみなされ，その推定の手掛かりとして人物頭部の方向が観測される．従来法 [7], [8] では，頭部方向の計測手段として磁気式のセンサーを採用し，それを各参加者の頭部に装着し計測を行っている．そのため，計測精度は高いものの，会話を行う環境や，参加者の行動が制約されるなど，適用可能な状況や応用が極めて限定されていた．そこで本稿では，このような問題に対処するために，各参加者を撮影した動画像から各人の頭部追跡を行うことで，頭部方向を非接触に計測する方法を新たに導入する．本稿の提案法は，頭部方向の計測に未校正カメラの画像が利用できるため，既存の会議システムなどで撮影されている参加者の映像を流用することができ，従来法と比較して格段に広い適用範囲を有することが特徴である．

本稿は以下のように構成される．まず，第 2 節において，会話モデル，及び，会話構造の推定法 [7], [8] を概説する．次に，第 3 節において頭部方向の推定法について述べる．続いて，第 4 節において実験結果を示す．第 5 節において提案法の課題等について議論し，第 6 節において本稿のまとめを述べる．

## 2. 会話モデル，及び，会話構造の推定法

### 2.1 対象とする会話構造

本研究では，$N$ 人物 ($N \geqq 3$) による対面会話を対象とし，参加者の行動と会話構造との関係をマルコフ切替モデル [15] と呼

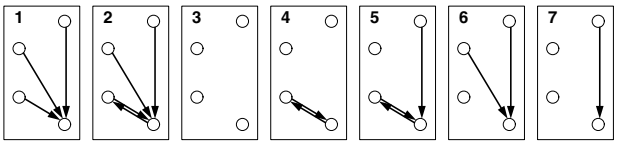


図 1　4 人会話における視線パターンの出現頻度（上位 7 位までを抜粋）．視線パターンを有向グラフにて表現（頂点＝人物，辺＝視線方向，出力辺がない頂点＝視線を逸らしている人物）．人物の入替えに不変な同型グラフのクラスについて，チャンスレベルに対する相対頻度の大きい順番に表示．

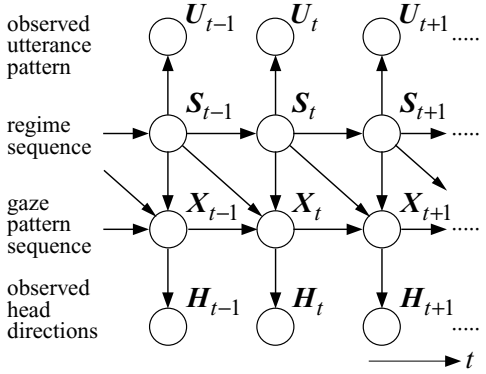


図 2　会話モデルの構造．

ばれる確率モデルを用いてモデル化を行う．ここでは，このモデルの上位プロセスを会話レジームと呼び，代表的な会話構造に相当すると想定する．また，会話レジームの状態は，マルコフ過程に従って時間変化するものと仮定し，このレジームの状態に依存して，参加者各人の行動や参加者間のインタラクションが確率的に生成されるものと考える．このモデルでは，発話権の交替に伴う参与役割の変化などの会話のダイナミクスは，会話レジームの状態遷移として表わされる．

会話レジームの状態の具体的な設定を行うために，本研究では，参加者の視線パターンの出現頻度に着目する．ここで視線パターンとは，ある時刻における全参加者の視線方向の集合を指す．参加者が $N$ 人の場合，各参加者の視線方向は，他の参加者のうち何れか一人の顔に向けられているか，あるいは，誰の顔からも視線を逸らしているかという $N$ 個の離散的かつ排他的な状態のいずれかにあるものと考える(注1)．従って，視線パターンは合計で $N^N$ 通り存在し得る．分析の一例として，図 1 には，ある 4 人会話（4.1 節のデータ G1-C1）を対象として，人手により検出された視線方向に基づく視線パターンの出現頻度を示す．このような分析により，一人物への視線の集中，二人物間の相互凝視のような特徴的な構造を持つ視線パターンが高い頻度で出現し，また，それらは比較的長時間継続する傾向があることが分かった．そこで本研究では，このような特徴的な視線パターンの構造と会話の構造との間には強い関連性があるものと予想し，以下に述べる，「一者集中」，「二者結合」，「分散」と呼ぶ 3 つの会話レジームのクラスを，推定すべき会話レジームとして用いる．

最初に，「一者集中」(convergence) と呼ぶレジームは，図 1 の#1,#6 のように，参加者の視線が一人の人物に集中する視線パターンをもたらす会話の構造に対応する．このレジームにおいては，最も多くの視線を受けている人物（中心人物と呼ぶ）が話し手となり，他の人物が受け手になり，いわゆる，モノローグと呼ばれる型の会話が進行していると想定される．このレジームを $R_i^C$ と記す．ただし，$i$ は中心人物を指す．$N$ 人会話の場合，$N$ 通りの一者集中レジームの状態 $\boldsymbol{R}^C = \{R_i^C\}_{i=1}^N$ が存在し得る．次に，「二者結合」(dyad-link) と呼ぶレジームは，図 1 の#4 のように，二人の人物が互いを見ている，つまり，相互凝視が生じるような会話の構造に対応する．ここでは，この相互凝視の関係にある二者が話し手または受け手となり，他の参加者は傍参与者となるものと想定され，いわゆる，ダイアローグと呼ばれる型の会話が進行していると考えられる．全体では，${}_NC_2$ 通りの二者結合レジームの状態 $\boldsymbol{R}^{DL} = \{R_{(i,j)}^{DL} | i = 1, \cdots, N; j = i+1, \cdots, N\}$ が存在し得る．最後に，「分散」(divergence) と呼ぶレジームは，上記二つのレジームに該当しない視線パターン（各参加者が，別々の人物を見る．視線を逸らしている等）が生じるような会話の構造に対応する．このレジームにおいては，グループとしての会話は生じていないと想定される．このレジームを $R^0$ と記す．

### 2.2　会話モデルの構造と会話レジームの状態推定

図 2 に会話モデルの構造を示す．このモデルには，隠れ状態変数として，会話レジームの状態の時系列 $\boldsymbol{S}_{1:T} = \{S_t\}_{t=1}^T$，及び，視線パターンの時系列 $\boldsymbol{X}_{1:T} = \{\boldsymbol{X}_t\}_{t=1}^T$ が含まれる．ここでは，時刻 $t = 1$ から $t = T$ までの離散時間区間をモデル化の対象とする．また，観測可能なデータとして，頭部方向の時系列 $\boldsymbol{H}_{1:T} = \{H_t\}_{t=1}^T$，及び，発話状態の時系列 $\boldsymbol{U}_{1:T} = \{U_t\}_{t=1}^T$ が含まれる．図 2 において，変数間の依存関係は有向辺として表現される．ある時刻 $t$ のレジームの状態 $S_t$ は，2.1 節で定義された $M(= N + {}_NC_2 + 1)$ 個のレジームの内，いずれか一つの状態 $S_t = R \in \boldsymbol{R} = \boldsymbol{R}^C \cup \boldsymbol{R}^{DL} \cup R^0$ をとる．また，視線パターン $\boldsymbol{X}_t$ は，各人物 $i \in \{1, \cdots, N\}$ の視線方向 $X_{i,t}$ の集合 $\boldsymbol{X}_t = \{X_{i,t}\}_{i=1}^N$ として定義される．ここで，人物 $i$ が人物 $j(\neq i)$ の顔を見る場合を $X_{i,t} = j$ と表し，人物 $i$ がいずれの人物の顔からも視線を逸らしている場合を $X_{i,t} = i$ と表す．頭部方向 $H_t$ は，各参加者の頭部方向の集合 $H_t = \{h_{i,t}\}_{i=1}^N$ として定義される．ただし，$h_{i,t}$ は，時刻 $t$ における人物 $i$ の頭部の水平方位角を表す．また，発話状態 $U_t$ は，各参加者の発話状態の集合 $U_t = \{u_{i,t}\}_{i=1}^N$ として定義される．ただし，$u_{i,t}$ は，人物 $i$ の発話状態を表わす（発話時=1，沈黙時=0）．

このモデルに基づいて，観測データである頭部方向の時系列 $\boldsymbol{H}_{1:T}$，及び，発話状態の時系列 $\boldsymbol{U}_{1:T}$ から，会話レジームの時系列 $\boldsymbol{S}_{1:T}$，視線パターンの時系列 $\boldsymbol{X}_{1:T}$，及び，モデルパラメータ $\varphi$ の同時事後確率分布の推定が行われる．本研究では，この問題の近似解法として，マルコフ連鎖モンテカルロ法 (MCMC) の一種であるギブスサンプラー [16] を用いる．なお，モデルパラメータ $\varphi$ には，各視線方向における頭部方向の尤度分布の平均値，分散や，各変数の遷移確率などが含まれる．

(注1)：本稿では，各参加者が注意を向ける先を限定するため，ノートや黒板などの道具は使用しないことを想定している．また，参加者は着席しており，会話中の参加者の増減や移動がないことを前提としている．

## 3. 頭部方向の推定方法と会話構造推定への適用

本節では，まず，本研究の会話シーン分析における頭部方向（頭部姿勢，顔向きとも呼ぶ）の追跡・推定に対する要求条件ついて述べた後，従来研究を外観し，要求条件を満たす方法として疎テンプレート Condensation 追跡法 (Sparse Template Condensation Tracker; 以後，STC 法と呼ぶ) を選択する (3.2 節)．続いて，STC 法の概要を説明し (3.3 節)，会話構造推定に適用する方法を述べる (3.4 節)．

### 3.1 会話シーン分析における要求条件

複数人物の対面会話シーンを対象として，頭部方向推定を行う場合，以下の条件・状況に対処することが求められる．

- 既存システム等の映像を流用するため，カメラパラメータなどの撮影環境に関する詳細情報が入手困難である．
- ステレオカメラ等，特殊な撮影手段は利用できず，一人物につき最大 1 台のカメラの利用を前提とする．
- 参加者の顔について事前にモデルを獲得することが困難．
- 不特定多数の人物に対して適用可能であること．
- 顔向きは正面顔から横顔まで多様に変化する (図 8(b))．
- 会話に伴う口の動きや表情の変化が大きい (図 8(c))．
- 顔を手で覆うなどオクルージョンが発生する (図 8(d))．

従来，この種の技術の主たるターゲットとして想定されていた HCI(Human-Computer Interaction) とは異なり，対面会話シーンの分析では，人間の自然な動作を対象とし，それを妨げないように撮影を行う必要があるため，より要求条件は厳しいものとなる．ただし，本稿ではオフライン処理を対象とする．

### 3.2 頭部方向推定の従来研究

従来，画像から人物の頭部方向を推定する代表的なアプローチとして，特徴点ベースの手法 [17] と，アピアランスベースの手法 [18] が知られている．特徴点ベースの手法では，目や鼻孔などの顔部品の特徴点の位置関係に基づいて，頭部方向の推定を行う．そのため，姿勢変動やオクルージョン (隠れ，遮蔽とも呼ぶ) などにより特徴点を安定に抽出できない状況には適用が困難である．また，アピアランスベースの手法は，事前に姿勢既知の条件で撮影された大量の顔画像データを用いた学習のプロセスを要し，また，遮蔽や顔領域の切り出しの精度に敏感であるという欠点を有する．また，Active Appearance Model (AAM) を用いるアプローチも近年，着目されている [19]．AAM は，メッシュ状の構造で顔形状を記述するモデルであり，顔面の細かな形状変化まで表現できる点が利点であるが，その構築には，メッシュの頂点を手動で与えるなど煩雑なプロセスが必要である．また，人物非依存を狙った Generic AAM は，一人物に特化した AAM と比較して，その性能が大きく劣ることが指摘されている [20]．そのため，これら従来法は，本研究には向かないと考えられる．

一方，近年，疎テンプレート Condesantion 追跡法と呼ばれる手法が提案されている [21]．この手法は，汎用的な物体追跡法であるが，人物の顔領域をテンプレートとして設定し，追跡を行うことで，人物頭部の方向を推定することが可能である．テンプレートは，一枚の画像のみから得られるため，事前の学習プロセスが不要である．また，ロバスト照合，Condensation による多重仮説の生成・検証，複数テンプレートの切替などにより，遮蔽や表情変化に対する高いロバスト性を実現している．これらの性質から STC 法は 3.1 節で述べた状況にも適応可能であると考え，本研究ではこれを採用する．

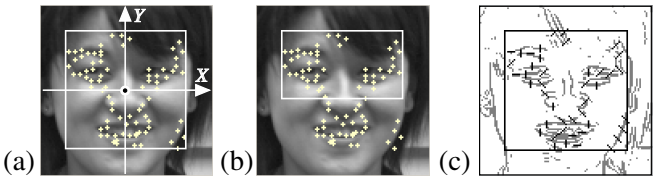


図 3 疎テンプレートの例 (グループ G1 の人物 3 (4.1 節参照))．領域，特徴点を示す．(a) 基本テンプレート $\mathcal{T}_0$，(b) 部分テンプレート $\mathcal{T}_1$，(c) ゼロ交差境界と境界ダイポール (線分で表す)．

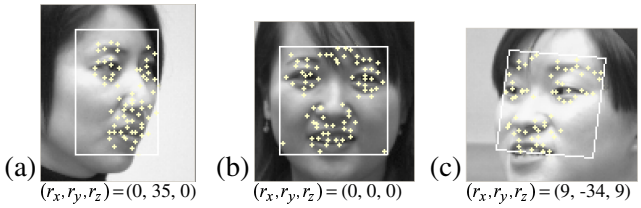


図 4 疎テンプレート設定の例 (グループ G1)，(a)P1, (b)P2, (c)P4. 設定時に与えた頭部姿勢を併記 (単位=度)．人物 i を Pi と記す．

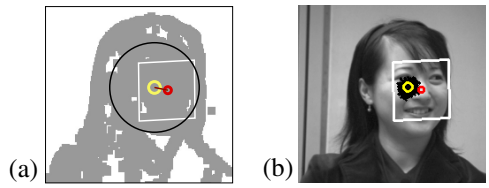


図 5 STC 法の動作の様子．(a) 抽出された頭部領域 (灰色で示す)，及び，頭部範囲を示す円 (黒い線)，(b) 追跡中のテンプレート (白枠)；黒点の集合：粒子の分布，大きい円：頭部中心，小さい円：テンプレート表面の中心．

### 3.3 疎テンプレート Condensation 追跡法 (STC 法)

疎テンプレートは，画像上の方形領域中の離散的な画素の集合から構成されるテンプレートであり，図 3(a) のように顔部品を含むように設定される．STC 法では，単一の画像から切り出された複数のテンプレートを切り替えて追跡を行うことで，部分的な遮蔽に対するロバスト性を高める機構が導入されている．ここではテンプレートの集合 $\boldsymbol{\mathcal{T}}$ は，顔面全体をカバーする基本テンプレート $\mathcal{T}_0$ と，$\mathcal{N}_T$ 個の部分テンプレート $\mathcal{T}_n$ $(n = 1, \cdots, \mathcal{N}_T; \mathcal{N}_T \geqq 0)$ の集合 $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_n\}_{n=0}^{\mathcal{N}_T}$ から構成される．部分テンプレートは，基本テンプレートの部分領域からなり，顔の一部が遮蔽された場合を想定し，残る可視領域に対応するものとして設定される．各テンプレート $\mathcal{T}_n \in \boldsymbol{\mathcal{T}}$ は，テンプレート画像 $J$ 上での各特徴点の位置 $\boldsymbol{p}_{n,m} = [X_{n,m}, Y_{n,m}]^T$ とその輝度値 $J(\boldsymbol{p}_{n,m})$ の組の集合 $\mathcal{T}_n = \{(\boldsymbol{p}_{n,m}, J(\boldsymbol{p}_{n,m}))\}_{m=1}^{\mathcal{N}_{Tn}}$ から構成される（$\mathcal{N}_{Tn}$ は特徴点数）．ただし，$\mathcal{T}_n \subset \mathcal{T}_0 (n > 0)$ である．特徴点は，画像の輝度分布の局所的な極大・極小点，及び，図 3(c) に示すようなゼロ交差境界に跨る境界ダイポールの両端点から選択される (詳細は [21])．特徴点数は，テンプレート領域内の全画素数の 1.5%程度と少なく，通常のテンプレートより格段に高速な照合が可能である．

STC 法では，各時刻においてテンプレートの状態を推定する問題を，Condensation の枠組みで捉え，観測情報に基づいてテンプレートの状態の確率分布を更新するステップと，その次時刻での分布を予測するという 2 つのステップを交互に実行する．テンプレートの状態の確率分布は，パーティクル (粒子) と呼ばれるサンプル集合 $\{\boldsymbol{s}_t^{(l)}, \omega_t^{(l)}\}_{l=1}^{\mathcal{N}_s}$ によって近似的に表現される．ここで $\mathcal{N}_s$ は粒子の個数を表し，$\boldsymbol{s}_t^{(l)}$，及び，$\omega_t^{(l)}$ は，それぞれ時刻 $t$ における $l$ 番目の粒子の状態，及び，重みを表す．各粒子 $\boldsymbol{s}_t^{(l)}$ の状態は，7 次元ベクトル $\boldsymbol{s}_t^{(l)} = [x_t^{(l)}, y_t^{(l)}, r_{x,t}^{(l)}, r_{y,t}^{(l)}, r_{z,t}^{(l)}, s_t^{(l)}, n_t^{(l)}]$ として定義される．ここで，$(x_t^{(l)}, y_t^{(l)})$ は，入力画像上でのテンプレートの位置を表す．また，$r_{x,t}^{(l)}, r_{y,t}^{(l)}, r_{z,t}^{(l)}$ は，それぞれ頷き，首振り，傾げに対応する頭部姿勢 (回転角) を表し，カメラに正対する顔向きを原点とする．さらに，$s_t^{(l)}$ はスケール，$n_t^{(l)}$ は，テンプレートの番号をそれぞれ表す．

更新のステップでは，各粒子の状態 $\boldsymbol{s}_t^{(l)}$ に対応するテンプレート $\mathcal{T}_{n'}(n' = n_t^{(l)}$ とおく) と入力画像 $I$ との照合が行われ，その照合誤差 $\varepsilon_t^{(l)}$ に基づいて，粒子の重み $\omega_t^{(l)}$ が更新される．照合誤差 $\varepsilon_t^{(l)}$ は，テンプレート $\mathcal{T}_{n'}$ 内の各特徴点 $\boldsymbol{p}_{n',m}(m = 1, \cdots, \mathcal{N}_{Tn'})$ の輝度値と，その特徴点が入力画像 $I$ 上に投影される点 $\boldsymbol{q}_{n',m}$ の輝度値との相対残差 $\kappa(J(\boldsymbol{p}_{n',m}), I(\boldsymbol{q}_{n',m}))$ に基づき計算される．ここで，$\boldsymbol{q}_{n',m}$ は，カメラモデルとして弱透視投影を仮定することで，$\boldsymbol{q}_{n',m} = s_t^{(l)} \cdot \boldsymbol{R}\, \boldsymbol{p}_{n',m} + [x_t^{(l)}, y_t^{(l)}]^T$ のように定義される．ただし，$\boldsymbol{R}$ は，$r_{x,t}^{(l)}, r_{y,t}^{(l)}, r_{z,t}^{(l)}$ によって定まる回転行列を表す．相対残差 $\kappa(J, I)$，及び，照合誤差 $\varepsilon_t^{(l)}$ は，

$$\kappa(J, I) = \frac{\alpha \cdot I - J}{J}, \varepsilon_t^{(l)} = \frac{1}{\mathcal{N}_{Tn'}} \sum_{m=1}^{\mathcal{N}_{Tn'}} \rho\left(\kappa(J(\boldsymbol{p}_{n',m}), I(\boldsymbol{q}_{n',m}))\right)$$

のように計算される．ここで $\rho(\cdot)$ は，$\rho(\kappa) = \kappa^2/(\kappa^2 + 1)$ と定義される一種のロバスト関数であり，これにより外れ値の影響を軽減した照合誤差が計算できる．また，$\alpha$ は，画像輝度の変化を吸収するための係数であり，照合誤差を最小とする値が選択される (詳細は [21])．各粒子の重みは照合誤差から $\omega_t^{(l)} \propto 1/\varepsilon_t^{(l)}, \sum_l \omega_t^{(l)} = 1$ のように計算される．その後，全粒子の中から，重みの値が大きい順番に上位 $\mathcal{N}_U(=10)$ 個が選択され，この選択された粒子の重み付き平均として，各時刻のテンプレート状態の推定値 $\bar{\boldsymbol{s}}_t^{(l)}$ が計算される．また，推定された頭部姿勢の内，首振り角に対応する $\bar{r}_{y,t}$ が水平方位角 $h_{i,t}$(2.2 節記載) として会話構造の推定のために出力される．また，選択された粒子の集合に基づいてリサンプリングが行われる．

次の予測のステップにおいては，現時刻 $t$ の各粒子の状態 $\boldsymbol{s}_t^{(l)}$ から，次時刻 $t+1$ における状態 $\widehat{\boldsymbol{s}}_{t+1}^{(l)}$ が予測される．状態のうち位置と姿勢については，各々独立なガウス分布 (平均=0) に従うシステムノイズが加えられる．また，スケールの値は，確率 $\xi_1(=0.955)$ にて持続し $(s_{t+1}^{(l)} = s_t^{(l)})$，残りの場合，均等な確率で増大または減少する $(s_{t+1}^{(l)} = s_t^{(l)} \times (1.02)^{+1 \text{ or } -1})$ という規則を適用する．テンプレートに関しては，確率 $\xi_2(=0.3)$ にて，基本テンプレートが選択され $(n_{t+1}^{(l)} = 0)$，それ以外の場合には，確率 $\xi_3(=0.4)$ にて持続し $(n_{t+1}^{(l)} = n_t^{(l)})$，さらに，それ以外の場合に他のテンプレートに均等の確率で遷移するような規則を適用する．この予測規則は，フィッティングの安定性の観点から基本テンプレートの適用を常に試行する一方，現時刻の遮蔽の状態が次時刻でも継続するという予想から，同じテンプレートの継続使用を試みるという考えに基づいている．

以上において，文献 [21] の STC 法と比較して，本研究の STC 法は，テンプレートの照合誤差の計算が簡素化され，また，テンプレートの予測規則が新たに導入された点が主に異なる．

### 3.4 STC 法の会話構造推定への適用

STC 法を会話構造推定へ適用するために試みた本稿独自の工夫について述べる．

**テンプレートの設定方法** 各人についてカメラに正対した平静時の顔が写っている画像フレームを手動で探し，その顔面領域をマウスで選択することで行う．なお，カメラに正対することのない人物については，図 4(a)(c) のように，傾いた状態のテンプレートを目視で設定する．また，図 8(d) のように手で口を覆うような動作が頻発している人物については，図 3(b) のように口を除く領域を部分テンプレートとして設定する．

**頭部回転角の符号不定性の解消** カメラモデルとして弱透視投影を採用している関係上，テンプレートの姿勢の回転角に符号の不定性の問題が存在する．これは回転角の符号を変えても画像上に投影される特徴点分布は同じになり，顔左右どちらを向いているか判断できないという問題である．ここでは，以下の方法により解決を図る．まず，テンプレートに奥行きのオフセットを導入し，テンプレート面を顔面に一致させ，位置 $(x, y)$ を頭部の中心（頭の回転中心を画像上に投影した位置）に変更する．この場合，特徴点の座標は $\boldsymbol{p} = [X, Y, Z]^T$ のように経験的に決めたオフセット成分 $Z$ を加えたものとする．次に，入力画像上で頭部領域の抽出を行い (図 5(a))，頭部中心 $(x, y)$ を中心とする円（頭の画像上での見かけの大きさに対応）が頭部領域内に包含されない場合，その度合いに応じて，対応する粒子の重みにペナルティーを課す．具体的には，この円の円周が頭部領域内に含まれる長さの割合を粒子の重みに乗じることで実施する．なお，頭部領域の抽出は，単純にフレーム間差分の累積により行う．以上の処理により，正しい頭部回転角の符号をもつ粒子のみが生き残ることで，不定性の問題が解消される．図 5(b) に，追跡中のテンプレート，及び，粒子の分布を図示する．

**追跡の初期化・再初期化** 初期時刻において，テンプレート状態の各要素について，その想定される範囲内の一様分布からの乱数生成により粒子集合の初期化を行う．また，追跡の失敗を，照合誤差の最小値が閾値以上になった場合として検出し，その場合には，更新ステップ中の上位 $\mathcal{N}_U$ 個の粒子の選択のプロセスを省く．これにより次時刻で再び対象を捉えるべく，より広い範囲に渡ってテンプレート照合を試みることができる．

**未校正カメラへの対応** 本研究では，各参加者の画像を別々の未校正カメラにより撮影する．そのため，計測される各人の頭部の方位角は，それぞれのカメラの座標系に依存したものとなる．文献 [7], [8] の提案法では，各人物に共通した世界座標系上での方位角を用いていたが，これは本質的な必要条件ではない．この方法では，各人物 $i$ が他者 $j(\neq i)$ に視線を向ける場合の頭部方向の尤度分布 $f(h_i|X_i = j) = N(\mu_{i,j}, \sigma_{i,j}^2)$ のパラメータ (平均 $\mu_{i,j}$，分散 $\sigma_{i,j}^2$) を推定しており，これは人物毎の別個

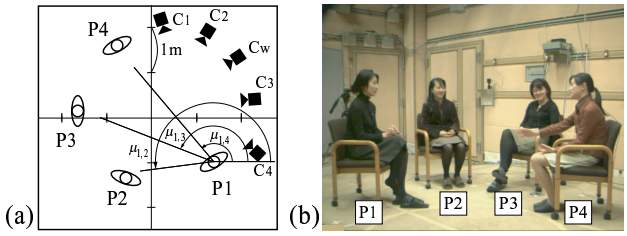


図 6　会話シーンの撮影環境．(a) 参加者，及び，カメラの配置，(b) カメラ $C_W$ から撮影された参加者の全体ショット (グループ G1)．

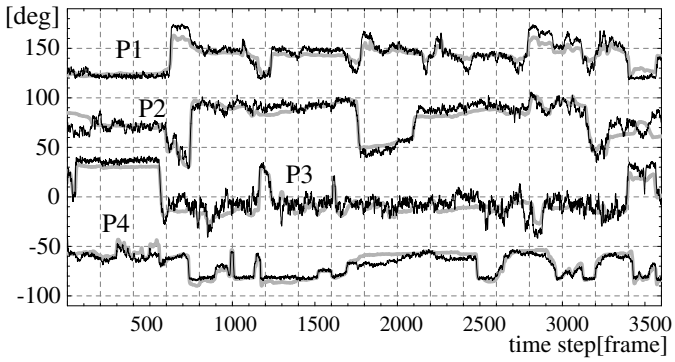


図 7　計測された頭部方向 (水平方位角) の時系列（G1-C1 の会話開始から 2 分間を抜粋）．実線：STC 追跡法による計測値，グレーの線：磁気式センサーの計測値．

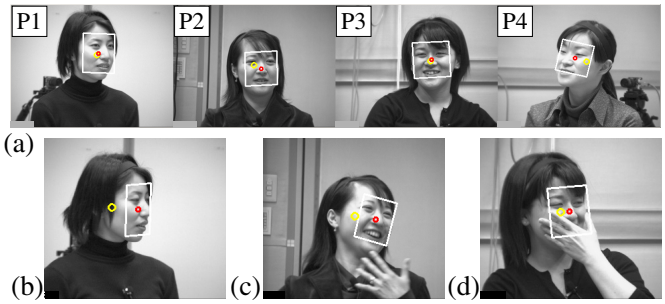


図 8　会話中の各参加者 (グループ G1) と追跡の様子．(a) 全参加者，(b) 横顔 (P1)，(c) 笑い (P2)，(d) 手で口を覆う動作 (P3)．

の座標系にもそのまま適用できる．ただし，この推定においては，これらパラメータの事前分布を設定する必要があるが，本稿では，STC 法の計測データを利用して，$\mu_{i,j}$ の事前分布 (これもまたガウス分布) の平均値 $\phi_{i,j}$ を設定するというアプローチを採る．具体的には，人物 1 を例にとると，この人物から見た他者の位置関係 (P3 は P2 の右にいる等) の事前知識を用いて，頭部方向尤度分布の平均値の間に $\mu_{1,4} < \mu_{1,3} < \mu_{1,2}$ の関係 (図 6(a) 参照) を仮定し，これに基づき，$\phi_{1,4}, \phi_{1,2}$ には，計測された頭部方位角の最小値，及び，最大値をそれぞれ与え，また，$\phi_{1,1}, \phi_{1,3}$ には，同平均値を割り当てることとした．なお，この設定は結果に対して敏感に影響しないことを実験により確認している．また，画像の水平軸が床面に平行であることは前提とする．

## 4. 実　　験

提案法の有効性を確認するため，4 人会話を対象とした実験を行った．本節では，まず，データセットについて述べた後，頭部方向の計測精度について検証する．次に，視線方向，及び，会話レジームの推定精度について定量評価の結果を示す．なお，著者の Web サイト(注2)にて実験結果の動画が閲覧可能である．

### 4.1　データセット

本稿では，文献 [7], [8] の手法と性能を比較するため，同文献と同一のデータセットを使用した．このデータは，4 名によるグループ会話を対象とし，20 代女性 8 名の参加者を 4 人ずつの 2 グループ G1,G2 に分け，それぞれ 2 つの議題について行った会話を収録したものである（以下，これら各会話データを G1-C1, G1-C2, G2-C1, G2-C2 と表記する）．参加者に対しては，提示された議題に対して議論を行い，5 分を目安にグループとして一つの結論を出すよう指示を与えた．一つの会話データの長さは 5.1~5.6 分であった．会話モデルの単位時間ステップは 1/30 秒である．会話モデルのパラメータは，文献 [7], [8] と同一とした．会話中の各参加者のバストショットを撮影し (図 8(a))，STC 法の入力とした (画像サイズ=320×240 画素, 30fps)．テンプレートは，G1 の人物 3 以外については，基本テンプレートのみを設定した．STC 法の粒子数は $\mathcal{N}_s$=2000 とした．

### 4.2　頭部方向の推定精度

図 7 には，会話データ G1-C1 について得られた各参加者の頭部方向（水平方位角）の時系列（最初の 3600 フレーム=120 秒を抜粋）を示す．また，図 7 には，各参加者の頭部にヘアバンドを用いて装着した磁気式センサー（POLHEMUS Fastrak™）の計測データも併せて示す．図 7 では，比較のため，STC 追跡法の計測値に対して，その平均値がセンサー出力の平均値に等しくなるようなバイアスが加えられている．図 7 中の STC 法による計測値には，絶え間ない微小振動がみられるものの，全体の傾向はセンサー出力と類似している．特に人物が視線をある人物から別の人物に移す場合に見られる頭部方向の急速な変化に対して，良好に追従していることが分かる．両者の計測値の平均偏差（差の絶対値の平均値）は，会話 G1-C1 の各人 $P1 \sim P4$ についてそれぞれ，5.1, 6.9, 10.0, 3.7[deg] であった．また，全データについての平均偏差は 5.9[deg] であった．

図 8 には，G1-C1 における頭部追跡の様子を示す．図 8(b) には，横顔による自己遮蔽が生じる場合，図 8(c) には，登録されたテンプレートとは異なる表情が生じる場合，図 8(d) には，部分的に顔領域が遮蔽された場合の各々において頑健に追跡が継続された例を示す．なお，各人物の追跡は，オフラインで実施され，各データの初期フレームから終端フレームまで自動的に追跡が行われた．また，データ中には，顔の全域を両手で覆うという完全オクルージョンが生じる区間が存在し，その区間における追跡精度は極端に劣化するものの，その状態から脱した時点で再初期化に成功し，追跡が続行された．処理速度は，動作周波数 3.2GHz の PC において，約 0.20[sec/frame] であった．以下の実験では，STC 法の計測値に対して平滑化などの前処理は施さないデータを使用した．なお，単純な時系列平滑化を導入した場合も，ほぼ同じ結果が得られることを予備実験により確認している．

(注2)：http://www.brl.ntt.co.jp/people/otsuka/miru2006.html

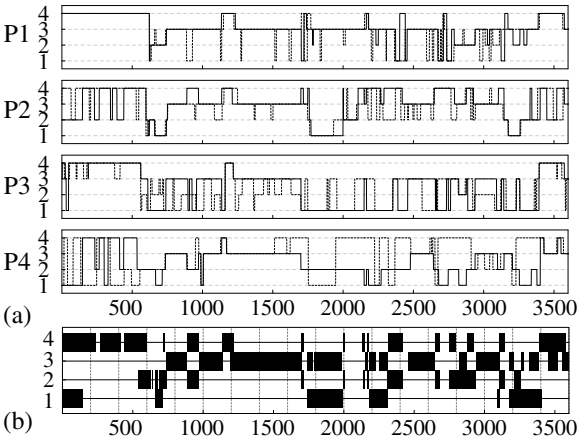


図 9　推定された系列（図 7 と同じデータ・時間区間）．(a) 視線方向 $\{X_{i,t}\}_{i=1}^{4}$（実線：推定値，破線：正解データ），(b) 会話レジームの状態；各時刻 $t$ において，一者集中レジーム $S_t = R_i^C$ の場合，$i$ の位置のみにバンド，二者結合レジーム $R_{(i,j)}^{DL}$ の場合，$i, j$ の位置にバンド，分散レジーム $R^0$ の場合は空白により示す．

表 1　推定された視線方向の推定精度 [%]．Proposed: 提案法の結果，Reference: 磁気式センサーで計測を行った結果．

| | G1-C1 | G1-C2 | G2-C1 | G2-C2 |
|---|---|---|---|---|
| Proposed | 64.8 | 57.3 | 67.5 | 69.3 |
| Reference | 71.1 | 59.3 | 72.4 | 75.9 |

表 2　会話レジームの平均正答率 [%]．Proposed: 提案法の結果，Reference: 磁気式センサーで計測を行った結果．

| | G1-C1 | G1-C2 | G2-C1 | G2-C2 |
|---|---|---|---|---|
| Proposed | 78.7 | 88.9 | 86.1 | 90.7 |
| Reference | 81.8 | 92.1 | 91.4 | 96.3 |

### 4.3　推定された視線方向の評価

図 9(a) には，提案法によって推定された視線方向と，対応する正解データを示す．視線方向の正解データは，人手により映像を観察することで作成された．表 1 には，推定方向と正解が一致したフレームの割合を推定精度として示す．比較のため，表 1 には，磁気式センサーによる結果も併記した．4.2 節で述べたように，STC 法の頭部方向データには，多くのノイズや誤差が含まれる．しかし，表 1 から分かるように，提案法の結果は，磁気センサーの結果とそれほど大差がないことから，頭部方向の計測手段として STC 法は有望な手段であることが示唆される．また，推定誤りの内訳をみると，視線を逸らした状態に関する誤りが多く (平均 70.4%) を占めている．これは，人間は頭部方向を変えずに視線方向を変化させることができるという性質に起因するものであり，センサーを用いた場合と同様に，頭部方向から視線方向を推定する際の本質的な限界といえる．

### 4.4　推定された視線パターン，及び，会話レジームの様子

図 10 には，会話 G1-C1 から推定された視線パターン，及び，会話レジームの様子を示す．最初 ($t = t_1$)，人物 4 が他の参加者に対して意見を述べている場面において，人物 4 への視線の集中，及び，人物 4 への一者集中レジーム $R_4^C$ が推定結果として得られた．次に ($t = t_2$)，人物 2 が人物 4 に対して応答を示す場面において，人物 2，4 の間の相互凝視，及び，この二者

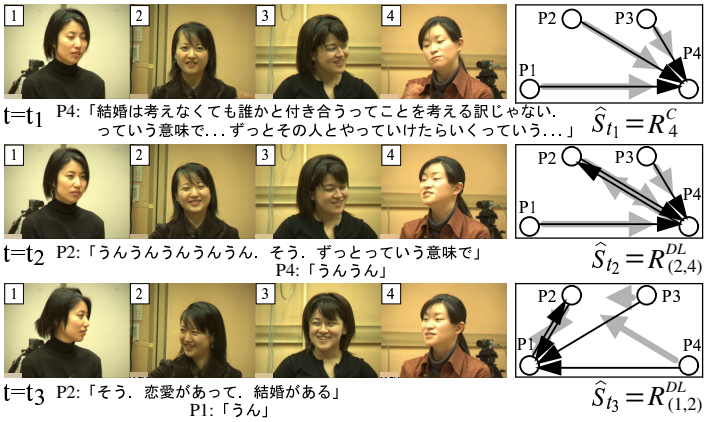


図 10　推定結果の様子 (G1-C1，$t_1 = 300, t_2 = 558, t_3 = 673$)．含む，各参加者の画像，視線方向（細い矢印：推定結果，グレーの矢印：正解データ），及び，推定されたレジーム $\hat{S}_t$．

間の二者結合レジーム $R_{(2,4)}^{DL}$ が推定された．その後 ($t = t_3$)，人物 4 が人物 2 に発話権を譲渡した後，人物 2 が人物 1 へ話し掛け，人物 1 がそれに応える場面においては，人物 1，2 間の二者結合レジーム $R_{(1,2)}^{DL}$ が推定された．センサーによる結果と比較して，提案法の結果は，推定された視線方向が若干，正確性に欠けるものの，会話レジームの遷移は，実際の会話の流れを反映したものであることが確認された．

### 4.5　会話レジームの評価

図 9(b) には，提案法によって推定された会話レジームの状態系列の一部を示す．また，表 2 には，各会話データについて，レジームの推定値の平均正答率を評価した結果を示す．レジームの正答率は，会話中の各フレームにおいて，推定されたレジームの状態と発話区間に付与されたアノテーションとの比較により計算される [7], [8]．ここでアノテーションは，各発話区間について，その種別（意見を述べる．質問をする．応答をする等），及び，発話の方向性（発話を向けた相手）を記述したものであり，人間の観察者が映像を観察することで作成した．表 2 から，提案法の正答率は，比較対象とした磁気センサーを用いた方法よりは劣るもの，それに十分に類するものであることが分かった．

以上の実験より提案法の有効性が確認され，また，STC 法による画像中の人物頭部追跡は，会話構造推定のための計測手段として有効であることが分かった．

## 5.　議　　論

**画像からの頭部方向の推定に関して**　今後の改良点としては，顔面の平面近似によるフィテッィングの不安定性を解消し，計測精度を高めるため，より顔面形状に近い立体的なテンプレートへの拡張が考えられる．また，テンプレート設定を自動化するため，顔検出技術 [22] の導入も検討課題としてあげられる．さらに，本稿では，一人一台のカメラを使用したが，現実の応用では，全方位カメラ等を利用し，より少数のカメラでより多くの参加者を捉えることが導入コストや利便性の観点から望ましい．そのため，画像中の各人の顔領域が小さく，また，複数の人物が混在する状況においても，安定に各人物の頭部を追跡

することが課題となる．また，現在，低解像度画像から直接，視線方向の推定を行う研究も進展しつつあり [23]，その結果との統合も興味深い課題である．

**確率的モデル化の利点** 会話構造推定のための他のアプローチの例として，頭部方向の単純な閾値処理により視線方向を推定し，その組み合わせにより直接，会話構造を求める方法が想定可能である．この方法は一見，簡易に見えるが，この場合，閾値を人手によって与える必要があり，各会話データ毎の試行錯誤が避けられない．それに対して提案法では，会話構造の推定と同時に，この閾値に相当する頭部方向尤度分布のパラメータの推定を行っており，その点，データ毎のパラメータの微調整が不要である．なお，この同時推定を実現する鍵は，会話のモデル (聞き手は話し手の方を見る傾向がある等) の利用にある．また，確率的モデル化のアプローチは，複数の異なる観測情報 (本稿では頭部方向，発話) を統合した会話モデルを構築する際の見通しの良い枠組みであり，会話という現象やその観測過程に内在する不確実性を取り扱うツールとしても有望である．

**会話構造推定の課題と方向性** 本稿では，参加者数固定かつ着席という限定的な状況を対象としたが，今後は，人物の移動や増減，資料の利用等，より一般的な状況への対処が課題としてあげられる．また，本稿では，3 クラスの会話レジームを仮説的に設定したが，現実には，参加者が複数のサブグループに分かれるなど多様な状況が生じうる．そのため，データに基づいて適応的に会話レジームを設定することも検討課題としてあげられる．また，会話に含まれる心理的・社会的な側面を定量化することも重要な応用課題である．その一例として，我々は現在，会話において誰が誰にどの程度の影響を与えているかといった影響力を定量化する方法を検討している [24]．さらに，本稿で計測した頭部運動の情報は，頷きなどの頭部ジェスチャーの検出にも利用でき，さらに，顔の表情や手振り・身振り等と併せた非言語行動の認識へと発展が期待できる．また，その他，提案法の応用としては，会議映像の自動編集・アーカイブ，会話エージェント・ロボットなど様々なものが考えられる．

## 6. むすび

本稿では，画像の認識・理解の新しい対象として人間同士の会話シーンを取り上げ，動画像から人物の動作を計測することで，会話の構造を推定するというアプローチを提案した．本研究では，その推定の手掛かりとして，会話参加者の「人を見るという行動」（視線行動）に着目し，会話の構造との関連性について確率的なモデルを構築した．また，視線行動の推測の手掛かりとして，画像上での人物頭部追跡による頭部方向の計測を新たに導入し，4 人会話を対象とした実験により，提案法の有効性を確認した．本稿の結果は，人と人とのコミュニケーションを支援するため技術として，画像の認識・理解技術が有用であることを示唆する例として位置づけることができる．

## 文　献


[1] I. McCowan, D. Perez, S. Bengio, G. Lathoud, M. Barnard and D. Zhang: “Automatic analysis of multimodal group actions in meetings,”, IEEE Trans. PAMI, **27**, 3 (2005).

[2] D. Zhang, D. G. Perez, S. Bengio, I. McCowan and G. Lathoud: “Modeling individual and group actions in meetings: A two-layers HMM framework”, Proc. 2nd. IEEE Workshop on Event Mining (2004).

[3] S. Basu, T. Choudhury and B. Clarkson: “Learning human interactions with the influence model”, MIT Media Lab. TR#539 (2001).

[4] A. Dielmann and S. Renals: “Dynamic Bayesian networks for meeting structuring”, Proc. IEEE ICASSP’04 (2004).

[5] E. Goffman: “Forms of Talks”, University of Pennsylvania Press, Philadelphia (1981).

[6] 竹前嘉修, 大塚和弘, 武川直樹：“対面の複数人対話を撮影対象とした対話参加者の視線に基づく映像切り替え方法とその効果”, 情報処理学会論文誌, **46**, 7, pp. 1752–1767 (2005).

[7] K. Otsuka, Y. Takemae, J. Yamato and H. Murase: “A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances”, Proc. ICMI’05, pp. 191–198 (2005).

[8] 大塚和弘, 竹前嘉修, 大和淳司, 村瀬洋：“複数人物の対面会話を対象としたマルコフ切替えモデルに基づく会話構造の確率的推論”, 情報処理学会論文誌, **47**, 7 (2006).

[9] M. Argyle: “Bodily Communication – 2nd ed.”, Routledge, London and New York (1988).

[10] A. Kendon: “Some functions of gaze-direction in social interaction”, Acta Psychologica, **26**, pp. 22–63 (1967).

[11] C. Goodwin: “Conversational Organization : Interaction between Speakers and Hearers”, Academic Press (1981).

[12] T. Ohno and N. Mukawa: “A free-head, simple calibration, gaze tracking system that enables gaze-based interaction”, Proc. Eye Tracking Research & Application Symposium (ETRA)2004, pp. 115–122 (2004).

[13] Y. Matsumoto and A. Zelinsky: “An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement”, Proc. Int. Conf. Automatic Face and Gesture Recognition ’04, pp. 499–504 (2000).

[14] R. Stiefelhagen, J. Yang and A. Waibel: “Modeling focus of attention for meeting index based on multiple cues”, IEEE Trans. Neural Networks, **13**, 4 (2002).

[15] C. J. Kim and C. R. Nelson: “State-Space Models with Regime Switching”, MIT Press (1999).

[16] W. R. Gilks, S. Richardson and D. J. Spiegelhalter: “Markov chain Monte Carlo in practice”, Chapman & Hall/CRC (1996).

[17] A. Nikolaidis and I. Pitas: “Facial feature extraction and pose determination”, Pattern Recognition, **33**, pp. 1783–1791 (2000).

[18] J. Sherrah, S. Gong and E. J. Ong: “Face distributions in similarity space under varying head pose”, Image and Vision Computing, **19**, pp. 807–819 (2001).

[19] J. Xiao, S. Baker, I. Matthews and T. Kanade: “Real-time combined 2D+3D active appearance models”, Proc. CVPR’04 (2004).

[20] R. Gross, I. Matthews and S. Baker: “Generic vs. person specific active appearance models”, Image and Vision Computing, **23**, pp. 1080–1093 (2005).

[21] 松原康晴, 尺長 健：“疎テンプレートマッチングとその実時間物体追跡への応用”, 情報処理学会論文誌：コンピュータビジョンとイメージメディア, **46**, SIG9(CVIM11), pp. 60–71 (2005).

[22] P. Viola and M. J. Jones: “Robust real-time face detection”, Int. J. Computer Vision, **57**, 2, pp. 137–154 (2004).

[23] 小野泰弘, 岡部孝弘, 佐藤洋一：“目領域の切り出しの不定性を考慮した低解像度画像からの視線方向推定”, MIRU’05, pp. 96–103 (2005).

[24] K. Otsuka, J. Yamato, Y. Takemae and H. Murase: “Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns”, Proc. ACM CHI’06 Extended Abstract, pp. 1175–1180 (2006).