身体動作における文脈情報の構成法と
文脈を活用した誤認識低減に関する研究

小椋 忠志

博士（情報学）

総合研究大学院大学

SOKENDAI (The Graduate University for Advanced Studies)

2019 年 9 月

目次

第 1 章　序論　1

1.1　研究背景　1

1.1.1　導入　1

1.1.2　動作認識における課題　2

1.1.3　研究目的　3

1.2　本論文の研究貢献　4

1.3　本論文の構成　5

第 2 章　身体動作における文脈情報の構成法と文脈を活用した誤認識低減　7

2.1　情報学における文脈に関する関連研究　7

2.1.1　自然言語処理における文脈　8

2.1.2　音声認識における文脈　9

2.1.3　人の動作における文脈　10

2.1.3.1　隠れマルコフモデルに基づく動作認識における文脈　10

2.1.3.2 動画上の時空間特徴に基づく動作認識における文脈 10
2.1.3.3 身体部位ごとの重みづけに基づく動作認識における文脈 11
2.1.3.4 ディープラーニングに基づく動作認識における文脈 12
2.1.3.5 N-grams に基づく動作認識における文脈 . . . 14
2.1.3.6 トピックモデルに基づく動作認識における文脈 14
2.2 身体動作における文脈情報の構成法 15
2.2.1 着目すべき部位の選択制御を行う文脈 18
2.2.2 日常動作における動作のカテゴリーとしての文脈 18
2.2.3 関連研究と本論文の手法との比較 20

第 3 章 文脈に基づく身体部位の着目制御と動作認識 23
3.1 身体部位の着目制御と動作認識 24
3.1.1 部位選択ベクトルの導入 25
3.1.2 パーティクルを用いた動作トピックに基づく動作と部位選択の仮説相互推定 26
3.2 同一身体上で実施される複数動作を対象にした検証実験 29
3.2.1 対象のデータ 29
3.2.2 実験結果 31
3.3 検討と課題 33
3.4 まとめ 33

第 4 章 動作のカテゴリーを扱う文脈と動作の双方向推定 35
4.1 文脈を用いた身体動作認識 35

4.1.1 現在の文脈を活用した認識手法 36
4.1.2 日常動作を例とした認識実験 38
4.1.3 複数の文脈に基づく動作を対象とした検証実験 42
4.2 時系列の変化を考慮した文脈と動作の双方向推定 47
4.2.1 文脈と動作の双方向推定手法の構成と実装 47
4.2.1.1 トピックに基づく動作認識処理 47
4.2.1.2 動作認識結果に基づくトピック分布の再推定処理 51
4.2.2 検証実験 52
4.2.2.1 対象のデータ 52
4.2.2.2 比較手法 57
4.2.2.3 シーケンスデータ A における検証実験 60
4.2.2.4 シーケンスデータ B における検証実験 66
4.2.2.5 コーパスの検討のための実験 68
4.2.2.6 有意性テスト 75
4.2.2.7 パラメータの検討のための実験 76
4.3 課題と検討 80
4.3.1 ループ処理における繰り返し回数について 80
4.3.2 文脈切り替わり時の認識率低下について 81
4.4 まとめ 82

第 5 章 結論 83
5.1 結論 83
5.2 提案手法の有効範囲と限界 84
5.3 今後の課題 86

5.3.1 2 つの取り組みの統合 86
5.3.2 動作以外の情報による文脈の利用 87
5.3.3 動作データの収集について 88

参考文献 91

謝辞 99

著者文献及び発表目録 101

図目次

2.1 文脈と動作出現確率の関係性 19
2.2 文脈と動作の時系列における関係性 19

3.1 パーティクルによる単位動作の認識と動作トピック推定の処理の流れ 27
3.2 提案手法によるパーティクルごとの認識結果 28
3.3 部位選択ベクトル $\boldsymbol{v}$ を用いない手法によるパーティクルごとの認識結果 29
3.4 提案手法によるパーティクル集合 $\boldsymbol{Q}$ 29
3.5 部位選択ベクトル $\boldsymbol{v}$ を用いない手法によるパーティクル集合 $\boldsymbol{Q}$ 29

4.1 文脈を用いた認識手法の概要図 36
4.2 文脈を考慮しない場合の認識結果 41
4.3 トピックモデルを考慮した認識結果 41
4.4 正解の動作ラベルの系列 45

4.5 トピックの情報を用いない手法による認識結果の系列 46
4.6 トピックの情報を用いる手法による認識結果の系列 46
4.7 パーティクルによるトピック情報を用いた動作認識とトピック分布の更新手順 48
4.8 HMM における尤度計算の際における確率密度関数 50
4.9 対象の動作の様子とよく似た動作の例．(a)m_6: wiping window．(b)m_{33}: bye-bye．(c)m_5: washing．(d)m_{13}: frying with pan．(a) と (b) の動作，(c) と (d) の動作はそれぞれよく似ており，誤認識を引き起こしやすい． 54
4.10 コーパス 1 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$ 55
4.11 学習とテストための動作サンプルの分け方とシーケンスデータの作り方 60
4.12 手法 (i) における認識の混同行列 62
4.13 手法 (v) における認識の混同行列 63
4.14 22 のシーケンスデータにおけるトピックの分布の平均と標準偏差の推移 64
4.15 手法 (iii) におけるトピックの認識結果とその正答率 65
4.16 手法 (v) におけるトピックの認識結果とその正答率 65
4.17 コーパス 2 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$ 69
4.18 コーパス 3 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$ 70
4.19 コーパス 4 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$ 71

4.20 β を 0.01 から 100 まで変化させた際の認識率の推移 77
4.21 β を 0.1 から 2.5 まで変化させた際の認識率の推移 78
4.22 β を 0.5 から 2.0 まで変化させた際の認識率の推移 79

表目次

2.1 文脈に関する関連研究との比較 20

3.1 本章で用いる記号の定義 24
3.2 実験に用いる各部位における単一動作 30
3.3 LDA による m_i，$\boldsymbol{v}_n$ の出現確率 $P(m_i, \boldsymbol{v}_n|t_j)$ 31

4.1 HMM で学習した動作パターンの一覧 38
4.2 実験条件と HMM に係るパラメータ 39
4.3 LDA によるトピック分類とトピックごとの動作 m_i の出現確率 40
4.4 実験の対象となる動作の一覧 43
4.5 LDA によって得られるトピックごとの動作出現確率一覧 44
4.6 α の値ごとの認識率 44
4.7 実験で取り扱う対象の動作のラベルとその動作が所属するトピックの一覧 56
4.8 シーケンスデータ A に対する各手法における認識率 61

4.9 シーケンスデータ B に対する各手法における認識率 67
4.10 4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの Kullback-Leibler(KL) 情報量とその合計値 . . 68
4.11 4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの JS 情報量とその合計値 72
4.12 4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの Cosign 類似度とその合計値 72
4.13 4 つのコーパスによる動作出現確率に対するあいまいさ $\psi(t_j)$ の計算結果 74
4.14 動作出現確率のためのコーパスを変化させた際における認識率 . 74
4.15 シーケンスデータ A に対する認識率における ANOVA の結果と p 値 75
4.16 シーケンスデータ B に対する認識率における ANOVA の結果と p 値 76
4.17 β を 0.01 から 100 まで変化させた際の認識率一覧 77
4.18 β を 0.1 から 2.5 まで変化させた際の認識率一覧 78
4.19 HMM 単体の認識が間違わない場合のリピート回数ごとの認識率 80

1

序論

1.1　研究背景

1.1.1　導入

近年，ロボットに関する技術の発展により，お掃除ロボットルンバやサービスロボットペッパーのように，我々の生活において身近な存在となってきた．人の生活環境内で活躍するパートナーとしてのロボットは，インタラクションを通じてより生活が便利になるような手助けを行うことが求められる．それらの生活支援ロボットは，人の自然な振る舞いを通して人とのコミュニケーションを行う必要があるといった点が，従来のような産業用ロボットとは大きく異なる．

RoboCup や WorldRobotSummit といった知能ロボットの競技会では，人との共存を目指した実世界の課題にチャレンジする試みであるが，生活支援ロボットのための知能は，いまだ実用レベルではないことを競技会の結果が物語っている．生活支援ロボットには人工知能を基にした様々な要素技術を組み合わせた総合的な能力が求められる．例えば要素技術のひとつとして，音声認識の世界では，Siri や Google Assistant といった認識システムを代表として，近年非常に高い精度を得ることを可能にしている．一方で，ロボットが観測できる情報から自発的に行動を起こせるようになるためには，人が今何をしているのかを認識する必要がある．その要素技術のひとつが「動作認識」である．動作の情報は音声の発話と異なり，決まった文法や単語，または音素のようなあらかじめ定義されているものがないことから，数多くの課題が残されている．

1.1.2　動作認識における課題

人の動作を対象にした身体動作認識の歴史は古く，数多くのアプローチが研究されてきている．これらの認識手法の多くは，とある入力データがどの動作に分類されるか，ということを単純に認識する手法である．学習済みのパターンと全照合するような従来のパターン認識手法による動作認識では，人にとってありえないと感じるような誤認識をたびたび起こす恐れがある．人にとってありえないと感じるのは，過去の観測情報や場面といった文脈を考慮しているからであり，文脈に逸脱していると考える点にある．

文脈を扱う手法の多くは，特徴量抽出の表現方法の 1 つを文脈と呼んでいる．これらの手法に共通する点は，時系列の情報を含んだ時空間表現を文脈と呼ぶ点にある．これらの手法の情報処理は，この時空間表現を分類器にかけて動作の認識を行うような手順であり，言い換えれば「文脈から動作」という一方向な認識である．本論文では，文脈は直接観測できない隠れた状態の情報であるとする．

文脈の情報は，人と共有可能な表現であるべきである．理由は，生活支援ロボットのような家庭内で活躍するロボットにはコミュニケーションを通した情報の共有が求められる点にある．すなわち，生活支援ロボットのための文脈情報を用いた動作認識は人と共有可能な文脈を取り扱う必要がある．コミュニケーションを通じて文脈情報を人と共有するためには，その文脈は発話可能な表現であることが求められる．その際における，文脈を推定するための有益な情報の 1 つが，動作がどういうものであったかという動作認識結果である．その時々における適切な文脈は刻一刻と変化していくため，動作の観測情報から適切な文脈を逐次推定することが求められる．動作の観測情報から適切な文脈を逐次推定するような「動作から文脈」という方向の情報処理も日常生活を対象にした動作認識には必要とされる．すなわち「文脈から動作」および「動作から文脈」という双方向な情報処理が，生活環境を対象とした動作認識には求められている．しかしながら，この双方向な関係性を実装し動作の認識に活用した手法がこれまで提案されていない．

1.1.3 研究目的

本論文の目的は，身体動作を認識する上での文脈情報の構成法はどうあるべきかを確立すること，そして，その文脈情報を活用することで身体動作の誤認識を防ぐということである本論文が扱う動作認識における誤認識に対する課題の一つ目が，どの身体部位に着目するべきかという問題点である．例えば，左手で電話をしながら右手メモを取るという複数動作が同一身体上に同時に観測されたときに，全身の情報を認識に用いると，本来認識したい動作が行われている部位とは異なる部位の動作によって誤認識が起こりうる．この問題に対して本論文では，認識するためにはどの身体部位に着目すればよいかという文脈を扱うことで誤認識の低減を狙う．もう一つの課題は，似ている動作への誤認識という問題点であ

る．例えば，右手を左右に横に振っている動作が観測されるとき，挨拶という文脈でその動作を見ると「バイバイ」と認識できる．一方で掃除という文脈でその動作を見ると「窓を拭く」という動作であり，「バイバイ」という動作が誤認識であったことが分かる．この問題に対して本論文では，これまでの観測からどのようなカテゴリの動作を行ってきたかという文脈を扱うことで誤認識の低減を狙う．

1.2　本論文の研究貢献

本論文の研究貢献は次の 2 つである．(1) 文脈と動作が互いに影響する仕組みを構成し，具体的な文脈情報の構成法を示すこと．(2)2 つの例題を通して互いに影響するループ構造が動作の認識を改善することを確認すること．

これまでに提案されてきた動作と文脈の関係性を用いた手法は次の 2 つのようないずれかのような特徴がある．(A) あらかじめ与えられた文脈情報を用いて動作の認識をする．(B) 観測から文脈を分類するのみで，文脈情報は認識の改善に用いられていない．この 2 つに共通していることは，文脈から動作，または，動作から文脈と言ったように，情報処理が一方向である点である．限られたシーンやタスクに特化させたような場面では一方向の認識によって十分な効果が得られる可能性がある．しかしながら，日常生活の人の動作を対象とした場合，一方向の認識では不十分である．その理由は，日常生活の人の動作においては，次々と動作のシーンやタスクが切り替わっていくためである．すなわち，身体動作を認識するための文脈情報は刻一刻と変化している．文脈情報が刻一刻と変化する場面において，これまでの研究の特徴を当てはめてみる．前者の (A) における問題点は，適切な文脈情報は誰が与えるのか？刻一刻と変化する文脈の変化を自分で推測し追従しなくてはならない点である．後者の (B) における問題点は，

場面ごとの適切な文脈情報は認識の改善に有益なはずなのに，活用されていない点である．これらの問題点をまとめると，動作と文脈が互いに影響しあうような双方向な情報処理が求められる．日常生活のような刻一刻と変化する場面に対応するためには，文脈と動作が互いに影響する双方向な情報処理が求められるにもかかわらず，これまで研究がされていない．

提案する手法を用いることで得られる利点は，場面に応じた文脈情報の活用によって不適切な誤認識が減るという点である．例えば，似ている動作への認識の改善である．全く異なる目的や場面で実行される動作であっても，複数の動作が非常によく似ている場合に誤認識が起こりうる．場面に応じた文脈情報は，場面にそぐわない誤認識を防ぎ，認識の改善に役立つ．もう一つの例は，多量の観測を対象にした認識の改善である．数多くのセンサ情報のうちすべての情報を認識に用いた場合，本当に認識したい動作とは異なる情報によって，認識のノイズとなり誤認識が起こりうる．場面に応じた文脈情報は，場面にそぐわない情報の領域を排除し，本当に認識したい対象への認識の改善に役立つ．この 2 つの例題を通して，文脈と動作が互いに影響するループ構造を用いることで認識の改善に役立つことを示す．

以上のことから本論文は，家庭用ロボットを始めとする日常生活を対象にした認識の要素技術として，文脈と動作の双方向な関係性を構成する第一歩目の研究を示したものである．

1.3 本論文の構成

2 章では，動作認識手法における近年の動向について触れ，文脈を用いた手法を中心に周辺研究について整理した上で，本研究の立ち位置について確認する．動作認識における課題点を 2 つ設定し，文脈を扱うアプローチによってそれぞ

れの課題の解決を目指す．1 つ目は，観測の全ての情報を扱うことによって複雑な動作をうまく認識できなくなる問題点である．この問題点に対して提案する手法では，認識のための着目すべき観測の領域を選択制御する文脈を扱う手法によって誤認識を防ぐ．3 章では，身体部位を例にとって，同一身体上で複数行われる複雑な動作を対象に，文脈による選択制御を伴う手法によって，動作の認識が改善すること目指す．もう 1 つの課題は，全く異なるカテゴリであるにも関わらず，動作が似ていることによって起こる誤認識である．この問題点に対して提案する手法では，動作のカテゴリとしての文脈を扱う手法によって，その誤認識を防ぐことを目指す．4 章では，日常動作を例にとりながら，似たような動作への誤認識の低減について取り組む．また，時系列データも対象にしながら，実験結果を通して提案する手法の有効性を示す．5 章では，2 つの取り組みを総括し，本研究における結論と今後の課題について述べる．

2

身体動作における文脈情報の構成法と文脈を活用した誤認識低減

2.1 情報学における文脈に関する関連研究

文脈の意味について，大辞林では次のように示されている．

1. 文における個々の語または個々の文の間の論理的な関係・続き具合。文の脈絡。コンテクスト。 「前後の－から意味を判断する」
2. 一般に、すじみち・脈絡。また、ある事柄の背景や周辺の状況。

しかしながら，これらの定義は情報学における実装上の定義としては曖昧で，文脈をモデル化するためには不十分な情報である．

情報学の分野においては，文脈は様々な方法によって表現され，研究されてきている．しかしながら，各々の研究で活用されている文脈に対して，情報学における文脈を明確に定義することは難しい．そこで本研究では，文脈を次のように定義し，周辺研究について文脈がどのように活用されているのか着目しながら整理する．

直接観測できない隠れた状態の情報

2.1 節では，この「文脈」に着目しながら，本論文が目指す「動作認識における文脈」はどうあるべきかについて関連する研究と比較しながら調査を行う．

2.1.1　自然言語処理における文脈

文脈という概念の多くは自然言語処理の分野において発展してきた．N-grams [1] は前後 N 個の単語の関係を学習することで，主に文法を対象とした文章の解読や生成の技術として広く使われている．単語の前後関係を見て推測する点においてこの手法は文脈を扱う手法である．またトピックモデル [2] は文書におけるトピック（話題）を推定し，分類等に用いられる手法である．このトピックは文書内の単語に共通するジャンルに相当するような分類を行うことから，トピックは単語に関わる事前情報として重要な要因となり，文脈であると言える．これらの自然言語処理における文脈情報の活用法は，古くから提案されており，自然言語処理の界隈では頻繁に用いられてきている．とりわけ，自然言語処理における文脈に関する研究の発達は，膨大にある文書や言語によって決められた文法ルール，そして辞書に登録されるような単語等が資産として潤沢にあることで実現されてきている．自然言語処理に対して，動作認識では文脈の情報を獲得で

きる程のデータ量が無く，辞書のように決まっている定義や，文法のようなルールが存在しない．そのため，これらの技術を活用するための工夫が必要となる．

2.1.2 音声認識における文脈

音声認識は，時系列の信号処理に基づく認識手法として捉えると，動作認識と非常によく似た特徴を持つ．音声認識では音声信号に対して，音素モデル，単語モデル，そして言語モデルというそれぞれのモデルの関係性により，それぞれの認識を改めることを行う [3]．ここで言う言語モデルは，文法に相当し，単語を推定する手助けとなる．音声認識を正しく実行するため，音声信号の波形に対してどのような単語が尤もらしいのかという認識は，N-grams による手法に基づいた情報を活用することで，逐次改善されるような仕組みが実装されている [4]．しかしながら，音声認識における音素も，自然言語処理における単語の辞書や文法と同じように，明確な定義がなされていることから，その情報を用いることで音声認識手法は発展していったと言える．一方で，動作認識においては，それらの音素に対応する単語や文法が存在しない．Taniguchi らは，これらの音素モデル，単語モデル，言語モデルを入力データから教師無しで同時に学習する手法を提案している [5]．この教師無しで獲得するような手法を用いないことには，音声認識と動作認識の手法を同等に扱うことはできない．そのため，動作認識のための文脈を実現するためには，身体動作の持つ音声とは異なる性質に適応する必要がある．

2.1.3　人の動作における文脈

2.1.3.1　隠れマルコフモデルに基づく動作認識における文脈

Janus らは，隠れマルコフモデル (Hidden Markov Model; 以下 HMM) に基づいた統計的相関に基づく教師無しの動作分節化手法 [6] を提案しており，Kulic らは，そのリアルタイム化を実装している [7]．この手法は，リハビリの動作解析 [8] や，ロボットの身体における動作生成へと拡張 [9] されている．Nakamura らはガウス過程 (Gaussian Process; 以下 GP) と隠れセミマルコフモデル (Hidden Semi-Markov Model; 以下 HSMM) を組み合わせて，連続する身体動作の分節化を実現している [10]．Taniguchi らは，HMM に基づく信号の記号化と言語モデルに基づく単語分節化を組み合わせることで，2 つの階層を持つ構造のモデルを提案 [11] し，動作の分節化を実現している．また，Taniguchi らの手法は 2 つの階層を同時に学習するようモデルが拡張されている [5]．いずれの手法も，どこで動作を区切ればいいかということを，全体の信号系列を考慮しながら実施されるため，文脈のようなものを考慮していると言える．

このように近年においても，様々な確率統計的なアプローチで動作の分節化が研究されている．これらの分節化技術は動作認識の前処理として有効な手法ではあるが，本研究では，動作の分節化を研究の対象にはせず，より先の認識プロセスについて取り組む．本論文の実験の一部では，あらかじめ適切な分節化が行われるものとして扱う．

2.1.3.2　動画上の時空間特徴に基づく動作認識における文脈

Zhang らは，画像特徴抽出手法である Shape Context を動画に拡張することで，動作の認識へと応用している [12]．また，画像上に抽出される特徴点を時系列に拡張し，それを文脈と呼んで認識に役立てる手法 [13, 14] においても，動

画上の行動認識に利用されている．これらの手法は，特徴を時空間空間に変換するこれらの方法は，処理における途中変換式によって得られるとある特徴を文脈として扱う．つまり，その文脈の機能は，特徴の変換としてのみ扱われ，その動作は直感的にどんな分類に所属するのか，例えば「掃除をしている」「どの部分に着目するべきか」というような，人に解釈が容易な前提知識として扱われるものではない．ほとんどの文脈を用いる手法は，その文脈を人々が解釈することが困難なパラメータのみで扱われている．産業用ロボットや定点カメラのような認識のみが求められる場合は，人々の解釈を考慮する利点は少ない．一方，人の生活空間内でコミュニケーションをとるロボットを想定した場合，文脈の表現を人と共有できるかそうでないかの大きな違いが生まれる．本論文で定義する文脈は，人が簡単に理解しやすい動作に関わる前提知識のようなものであり，この文脈と動作の関係について焦点を当てる．

これらの手法は認識において文脈を利用していると言えるものの，文脈そのものがどういうものであるかということを認識するに至っていない．

2.1.3.3 身体部位ごとの重みづけに基づく動作認識における文脈

身体のうち局所部分において実施されている身体動作に対して，身体の全身を対象に認識を行ってしまうと，関係の無い身体部分の動作によって誤認識を引き起こす恐れがあるという問題点がある．この問題に対して，身体における各関節情報に重みづけを行うことで，単一動作の認識率を向上させる研究が行われている [15–17]．一方で，上記のような二つの動作が同時に起こるような複合動作に対して認識を行う場合，いずれかの動作しか認識できなくなるか，まったく違った認識結果が表れてしまうことが懸念される．Wei らは身体の局所部分の動作認識手法と 2 つの動作の前後関係を使って，同一身体上で実施される複数動作の認識を実現している [18]．この手法では動作ラベル間の時間的前後関係性を用

いているが，上記の例のように多くの複合動作は 2 つの動作の目的や意味が異なっており，関連性が弱い場合がある．同一身体上で関連の弱い複数の動作が実行される場合は，目的の動作ごとに部位を取捨選択し，各々認識する必要があると考える．すなわち，「今どこに着目するべきか」という情報を複数推測しながら認識を行うことで，この問題への対応を行う．

2.1.3.4　ディープラーニングに基づく動作認識における文脈

身体動作の検出手法は，ディープラーニングの台頭により，近年急速に発達してきた．その傾向は，画像処理・動画解析を主としており，身体の形状を推定している．OpenPose [19] は動画から 2 次元の人の姿勢を推定する手法のパッケージであり，オープンソース化されている[*1]ことから，様々な応用が展開されている．Ohashi らは OpenPose を基に複数のカメラを用いることで人の 3 次元姿勢を推定する VideoCapture という手法を提案している [20]．これらの身体の姿勢推定手法は，画像のうち人の頭は肩より上にあり下の方には脚があり，と言ったような学習データに基づく前提知識が備わっている．逆に，学習データにはないようなアクロバットな動きや逆立ちになるような動きに対しては正しく推定することが難しくなる．人体の形の事前知識はある種文脈と言え，その文脈はアクロバットな動きに対応出来ていない．

本来，家庭環境やオフィスで活躍するサービスロボットを想定した場合，ロボット自身の持つセンサーによって人の動作を検出・推定することが想像に容易い．ここまで紹介した姿勢を推定する手法の発展を考慮すると，精度高く細部の人の姿勢を推定できるようになると推測できる．本論文では，人の姿勢は精度よく細部まで推定できることを前提とし，「では，その動作はどう言った意味を持

[*1] GitHub - CMU-Perceptual-Computing-Lab_openpose_ OpenPose_ Real-time multi-person keypoint detection library for body, face, hands, and foot estimation, https://github.com/CMU-Perceptual-Computing-Lab/openpose

つ動作なのか」ということを認識する，動作認識の領域について取り組む．

ディープラーニングによって動作を認識する手法の多く [21–24] は動画像をそのまま，もしくは下処理をして，畳み込みニューラルネットワーク (Convolutional Neural Network; 以下 CNN) に基づく画像特徴変換によって動作の識別を行う．CNN を動画上における行動認識に拡張する際には，時系列の変化を表現可能な変換を実施している．この時系列の前後の変化を考慮することから，文脈を扱う手法であるとも言える．ディープラーニングを用いた動作認識手法において，身体のスケルトン情報を用いた場合の多くは再帰型ニューラルネットワーク (Recurrent Neural Network; 以下 RNN) に基づいた時系列信号解析によるものが多い [25, 26]．またこの 2 つの技術を組み合わせる手法 [27] も提案されている．また，Du らは身体部位ごとに RNN による学習を行い，さらにそれを統合するような階層的な RNN による身体動作認識手法を提案している [28]．RNN は構造上コンテクスト層を持ち状態を管理する役割を持つことから，文脈を利用した手法であると言える．

ただし，これらの文脈情報は人との共有が難しく，計算の経過としてのパラメータに過ぎない．

Tani らは RNN にパラメトリックバイアス (PB) を加えた RNNPB と呼ばれる手法を提案し，行動の解析に役立てている [29]．RNNPB は，学習を行ったシーケンス間の類似構造を PB 空間で確認できるという特性があり，人が見て分かりやすい文脈の表現方法を提供可能であると言える．一方で本論文では，コミュニケーションを通じてロボットと人がどんな文脈ということを共有するために，文脈情報は発話可能な単語であることに重点を置き，文脈を扱う．

2.1.3.5　N-grams に基づく動作認識における文脈

自然言語処理の領域において，文脈を考慮する手法として N-grams に基づく手法が良く利用されている．そのため，この N-grams を用いることで，動作の認識を改善するという取り組みも，文脈を考慮する手法であると言える．N-gram を用いた動作認識への応用した手法の一つとして，観測された動作の情報からその動作がどんな動作であるかを説明する文章を生成する手法が提案されている [30, 31]．また，N-grams の特徴を生かし，観測された動作から次の動作を予測する手法が提案されている [32]．Taniguchi らは，HMM に基づく動作の記号化とその上位として N-grams に基づいた文脈を相互に学習する手法を提案している [11]．Taniguchi らの手法は，連続する動作の分節化を目的にしたものであり，動作やその文脈そのものを認識する手法とは異なる．

これらの文脈に基づく手法をもとにまとめると，本研究は，(1) 文脈というものがどういうものを扱っているのか人にわかりやすいものであること，(2) 動作だけではなく文脈そのものも認識すること，この 2 点を満たす手法を実現する．

2.1.3.6　トピックモデルに基づく動作認識における文脈

本研究では，文脈を扱うためにトピックモデル [2] と呼ばれる自然言語処理の手法に着目する．トピックモデルを用いた動作認識手法はいくつか存在するものの，そのほとんどは観測情報のトピック分類のみといった一方向の認識のみ実装されている．動画上の特徴量に基づいたトピックモデルを用いたアクションの分類手法 [33–35] では，その特徴量を単語として扱うようなトピック分類を行い，その分類されたトピックそのものを動作認識の結果としている．ユビキタスコンピューティングの分野では，ウェアラブルデバイスによって得られるセンサ情報に基づいて行動を分類する手法の提案がされている [36]．また，車の運転における行動の分析においては，トピックモデルに基づいた行動の分類手法が提案

されている [36–38]．Bargi らの手法 [39]，Malgireddy らの手法 [40] は，HMM に基づいて行動を分析したのち，トピックモデルによる分類を行っている．しかしながらこれらの手法 [36–40] は人の身体動作を対象にはしていない．また，トピックモデルを使用しているものの，それは観測データを分類するのみにとどまり，その認識の方向は 1 方向のみである．Attamimi らの手法 [41] は，身体動作の情報を含む複数の概念を統合するような統合概念の形成を行い，統合された概念とモダリティの一つである身体動作との双方向な関係性を用いて学習している．しかしながら，この手法 [41] は認識プロセスにおいて観測情報を統合概念として分類するのみで，動作の認識が間違っていてもそれを修正するような双方向な関係性を用いた認識処理は実装されていない．Demiris らの手法 [42] は，階層構造を持ち認識処理が上下層にループする仕組みを持つ手法の一つである．この手法は，物体との関係性に焦点を当てた行動を制御する手法であって，身体動作を認識する手法とは異なる．

これらのトピックモデルに基づく手法をもとにまとめると，本研究は，上位層の文脈と下位層の動作の認識が一方向ではなく双方向に行われる手法の実現を行う．

2.2 身体動作における文脈情報の構成法

本論文の目的は，身体動作を認識する上での文脈情報の構成法はどうあるべきかを確立すること，そして，その文脈情報を活用することで身体動作の誤認識を防ぐということである．ここで文脈情報の構成法とは，文脈を計算機モデルとして表現する適切な方法である．そして本研究では，身体動作における文脈の情報は解決したい問題点ごとにどう扱われるべきか異なるとして，問題点に対してどのように文脈情報を構成するべきか，2 つの問題点を題材に議論する．ここで

は，文脈情報を問題点ごとに対応させ，その問題を解決させることが，身体動作における文脈情報の構成法の確立であると定義する．

本研究では，これまで示してきた動作認識に求められる文脈情報を扱いその文脈を考慮した動作の認識を実装する．動作を下位の層であるとすると，文脈はその動作の認識を助ける上位の層の概念となる．この上位層の文脈と下位層の動作は一方向に認識されるのではなく，双方向な関係性を用いた認識であるべきである．この理由は，文脈が分かれば動作が正しく認識できる，また，動作を見ていればその人の文脈が分かる．この文脈と動作の関係性を用いる必要があるためである．

関連研究を基に動作の認識における問題点は，次のように整理される．ひとつは，複数の動作が身体上において混ざり合って実行される場合，本来認識したい身体部位で実行された動作とは異なる身体部位で実行されている他の動作を要因として，正しく認識できなくなるという点である．この問題点を解決するためには，どのようなことに着目して認識していけばいいかという事前知識としての文脈を用いることが求められる．具体的には，「歩く」と「走る」が混ざった動作は，その動きの速度に着目することでこれらを分離することが出来る．また，「電話する」と「メモを取る」が混ざった動作は，左手を見れば電話をしていて右手を見ればメモを取っているということを分離することが出来る．

ふたつめの問題点は，異なる場面や条件下の動作であっても，動作そのものがとても類似している場合，似ている他の動作に誤認識してしまうという点である．この問題点を解決するためには，それらの動作がどのようなカテゴリーに属するかという事前知識としての文脈を用いることが求められる．具体的には，「掃除」をしているのであるから，「バイバイ」という動作にも見えるがそれは「窓を拭く」という動作であるというように，似ている動作を区別することが出来る．このように，カテゴリとしての文脈は類似している動作への誤認識を防ぐ

ことが可能であると言える．

本論文では，これら二つの問題点に対して，それぞれ文脈を扱うというアプローチを用いて，その有用性について検証を行う．本論文では，これらの文脈の取り扱いや認識の対象に動作のみを利用する．しかしながら，この文脈を扱うにあたって動作以外の情報も大変重要なものである．例えば，人が手に何を持っているかというような道具の情報は，動作の認識における重要な事前知識となる．具体的には，動かしている手に雑巾を持っていれば，その動作を「バイバイ」という動作として認識されることは考えにくく，「窓を拭く」という正しい認識をするにあたって重要な情報となる．また，その雑巾を手に持っているということを認識できなければ，その人の動作に着目することで，手に持っている道具を推測可能であると言える．具体的には，「机を拭く」「窓を拭く」という動作を観測できれば，手に持っている道具が雑巾であることを推定できるのである．このように本論文で扱う文脈は，道具と動作の関係性においても適用可能な枠組みであり，今後取り組むべき課題の一つである．

これらの問題に対して，本論文では具体的に次のように取り組む．文脈とは動作の認識に対してどういう役割を持つか，という点において本論文では 2 つの観点から取り組む．ひとつ目は「どのようなことに着目すればいいか」ということを扱う文脈，もうひとつは「その動作はどのようなカテゴリーに属するか」ということを扱う文脈である．

共通のアプローチとして，本研究では次のような手段によって問題解決を目指す．まず，今どんな文脈であるかということを確率分布で表現する．その文脈の確率分布に基づいて動作出現確率を参照し，動作に対するパターン認識に作用させる．ここで動作出現確率とは，動作の出現を予測する確率分布であり，文脈を動作の関係性を記述する重要な要素となる．動作出現確率とパターン認識手法に基づく認識尤度を組み合わせることで，現在の文脈に基づいた動作の認識を実

現する．また，その動作の認識結果を基に，現在の文脈の分布の再推定を行う．この文脈から動作へ，動作から文脈へという認識を実現する．

2.2.1　着目すべき部位の選択制御を行う文脈

この取り組みでは，「どのようなことに着目すればいいか」を扱う文脈において，その典型例である「全身のうちどの部位を着目すべきか」という選択制御を文脈に基づいて行う．文脈は，どこに着目するべきかという情報を「部位選択ベクトル」として所持している．また，HMM による動作の学習は，身体部位ごとに行い，部位選択ベクトルに基づいて身体部位ごとの HMM を選択したのち認識を行う．その認識結果に基づいて，文脈の分布を再推定する．このループ処理を繰り返し実行することで，現在どの部位に着目するべきなのか，という文脈と，身体部位ごとの動作の認識を実現する．さらに，同一身体上で実行されている複数動作に対しても，複数の文脈が候補として挙がることで，複数の動作を同時認識できることを示唆し，全身を対象とした認識手法と比較して認識率の向上を実現する．

2.2.2　日常動作における動作のカテゴリーとしての文脈

この取り組みにおける文脈の定義は「認識のための事前分布としての動作出現確率を扱う上位の概念」とする．動作出現確率とは，観測される動作の候補を予測する確率分布である．図 2.1 に文脈と動作出現確率の関係性の例を示す．この取り組みでは，人が行う日常生活での動作を例として，動作を観測されるシーンに基づいたカテゴリーに分類し，そのカテゴリーを文脈として取り扱う．図 2.1 に示した例は，この動作のカテゴリーとしての文脈と動作出現確率の関係性を示

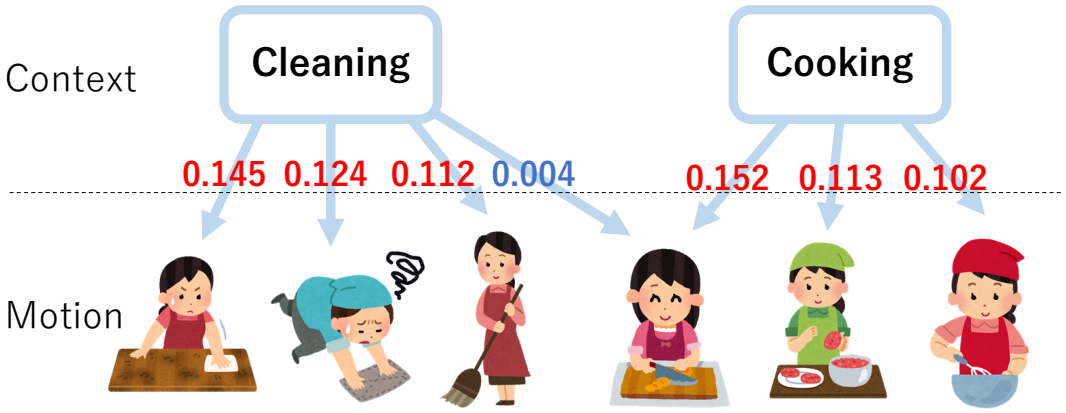


図 2.1　文脈と動作出現確率の関係性

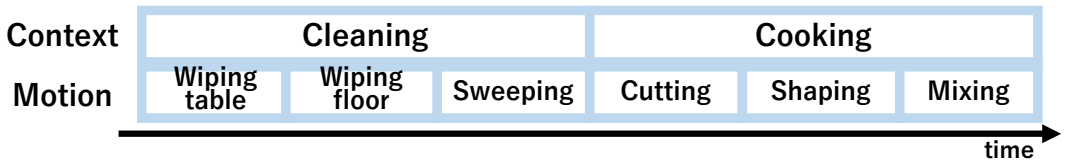


図 2.2　文脈と動作の時系列における関係性

したものであり，この文脈における動作出現確率とはどういう意味を持つか，具体的な例を用いて説明する．「掃除」の文脈では「窓を拭く」「机を拭く」「床を掃く」という動作が観測される確率が高く，その文脈の中では「野菜を切る」といったような動作が観測される確率が低い，ということを表現している．日常生活における観測を例にすると，その観測における文脈が急激に変化することは考えにくい．つまり，「掃除」をしている人は，しばらくの間「掃除」に関わる動作を行う可能性が高い．すなわち，文脈は時系列の観測に対して効果的な役割を持ち，動作の認識において重要な情報を持つことになる．この取り組みでは，実際に時系列の観測データを対象に実験を行い，他の文脈に基づく手法と比較しながら，提案手法の優位性について検証する．

表 2.1　文脈に関する関連研究との比較

比較項目 / 代表的な手法	[18]	[15]	[33]	[41]	Proposed
身体動作認識	✓	✓	✓	partially	✓
同時発生動作の認識	✓	partially			partially
類似性動作の識別	partially			partially	partially
人と共有可能な文脈				✓	✓
時系列な認識	✓		✓		✓
双方向な手法					✓

2.2.3　関連研究と本論文の手法との比較

表 2.1 に文脈に関する関連研究と提案手法の比較の表を示す．本論文で論じる解決するべき課題点との比較を行う．Wei らの手法 [18] は身体動作上の同時発生する複数動作の認識を目的としている．動作の時間的前後関係性を学習パラメータとして保持していることから，文脈情報のような情報を扱っていると言える．しかしながら，Wei らの手法 [18] が扱う文脈情報に相当するパラメータは，人と共有することは非常に困難なものである．Goutsu らの手法 [15] は身体部位ごとの重みづけを行い，さらに動作の説明として適切な文章を生成する手法である．動作の認識に身体部位ごとの情報を用いており，言語処理手法と組み合わせている点において，本論文の提案手法と類似するが，時系列の文脈情報（着目するべき身体部位）の変化の追跡や，人と共有可能な文脈情報を扱ってはいない．Tavenard らの手法 [33] は，身体動作認識においてトピックモデルを用い時空間な観測情報の分類を行っている．トピックモデルを用いているという点で提案手法と共通する部分があるが，トピックモデルは動作ラベルの分類器としてのみ利用されており，文脈と動作の関係性については言及されていない．Attamimi らの手法 [41] は，物体のカテゴリ分類手法であるが，物体に関わるマルチモーダル情報の一つに身体動作情報を用いている．Latent Dirichlet Allocation(LDA) を応用し，動作情報を含むマルチモーダル情報の統合概念の形成と，その概念を活用した未学習物体の分類を可能にしている．この手法 [41] は，人とのコミュニケーションを基に概念を形成しており，人と共有可能な文脈を扱っているとい

える．しかしながら，物体認識のための手法であるため，時系列に刻一刻と変化するような動作や文脈の認識への応用は工夫が必要となる．

Baker らは人の行動の理解をモデリングするために，ベイズ推定に基づく逆方向の行動推定手法を提案している [43]．Baker らの研究では，人の行動を分析するため，単純化された 2 次元のナビゲーションタスクを例に，行動からゴールを予測するような人の考え方のモデリングを行っている．このアプローチは本論文における「動作から文脈を推定する」仕組みととてもよく似ている．一方で Baker らの手法では，観測から行動を予想する確率分布は，被験者によるアンケートによって生成されており，動作の観測に対する認識問題に適用するには認識のためのアルゴリズムを組み合わせるような工夫が必要になる．

これらの手法と比較し，本論文で提案する動作認識手法は，人と共有な文脈を扱い，時系列な文脈と動作の関係性を活用している．また，文脈情報の構成法として，着目すべき領域の選択やカテゴリとして文脈情報を扱うことにより，同一身体上で発生する複数動作の誤認識や，似ている動作への誤認識をそれぞれ低減する．このように提案する手法は，周辺研究と比較しても新しい着想によるアプローチであり，本論文で取り扱う課題点を解決する手法である．

3

文脈に基づく身体部位の着目制御と動作認識

数多く観測される情報のうち認識に適した情報を選択できない場合，認識したい対象とは異なる領域の情報によって誤認識が発生してしまうことが十分に考えられる．例えば，テーブルを拭くという身体動作を認識したい時に，腕がどう動いているか，雑巾を手に持っているかという情報はおそらく重要であり，その時に脚がどう動いたか，便器の蓋が閉まっているかどうかというような情報はおそらくあまり重要ではないといえる．この問題点をミニマルに扱うと，全身のうちどの身体部位に着目するべきかという課題に置き換えることができる．例えば，電話をしながらメモを取るという同一身体上で実施される複数動作が観測された場

合，認識に全身の情報を用いると，認識したいメモを取るという動作にたいして電話をするという動作がノイズとなり，誤認識のもととなる．

本章では，身体のうち局所的に実施されている動作に対して，全身を対象にした認識を実行してしまうと誤認識が生じてしまう問題に対して，「今どこに着目するべきか」という点を扱う文脈を用いて解決を狙う．

3.1　身体部位の着目制御と動作認識

表 3.1　本章で用いる記号の定義

$\boldsymbol{p}$	動作パターン
w	トピックモデルが対象とする単語
m_i	i 番目の動作ラベル
M	動作ラベルの数
t_j	j 番目の動作トピック
J	動作トピックの数
c_l	l 番目の分割された身体部位
L	部位の分割数
$\boldsymbol{o}$	観測された身体動作信号
$\boldsymbol{Q}$	パーティクルの集合
k	k 番目のパーティクル
$\boldsymbol{v}_n$	n 番目の部位選択ベクトル
b^l	部位 l に対する重み
$\hat{m}$	サンプリングされた動作ラベル
$\hat{\beta}$	サンプリングされた部位選択ベクトル
$\hat{k}$	サンプリングされたパーティクル
$P(m, \boldsymbol{v}\|t_j)$	トピック t_j に対応した m，$\boldsymbol{v}$ の出現確率
$S(\boldsymbol{o}, i)$	身体部位選択処理後の動作信号 $\boldsymbol{o}$ の動作ラベル m_i に対する統合尤度

今どこに着目するべきかという情報を文脈によって取り扱うために，「部位選択ベクトル」というものを導入する．この部位選択ベクトルは，どの部位に着目するべきかという情報を基に，HMM に基づいた部位ごとの認識尤度を選択する役割を持つ．また，動作の認識結果と文脈の双方向な繰り返し処理を実装するため，「パーティクル」という概念を用いて，離散的に処理を実行する．

3.1.1 部位選択ベクトルの導入

本章では，身体部位の選択を行うベクトル $\boldsymbol{v}$ を以下のように導入する．

$$\boldsymbol{v} = [b^1, b^2, \cdots, b^l, \cdots, b^L]^{\mathrm{T}} \quad (b^l = \{0, 1\}) \tag{3.1}$$

ここで，b^l の l は分割された身体部位 c_l に対応しており，L はその分割数である．部位選択ベクトル $\boldsymbol{v}$ は，全身を複数の身体部位 c_l に区切った後，どの身体部位 c_l を選択するかという情報を持つベクトルである．b^l は身体部位 c_l を選択するか選択しないかという情報を持ち，b^l の値が 1 であればその部位 c_l を動作の認識尤度計算に用いる．

この部位選択ベクトルは，動作コーパスと呼ばれる観測された動作の情報を記録したコーパスから獲得する．動作コーパスは，アノテーターに依頼しアノテーターが観測対象者の連続した動作を観察し，動作パターン $\boldsymbol{p}$ に対し動作ラベル m とその動作がどの身体部位によって実施されたかという情報 (= 部位選択ベクトル)$\boldsymbol{v}$ をアノテートして貰うことで生成される．

ここでは，その動作コーパスから動作パターン $\boldsymbol{p}$，動作ラベル m を取り出し，隠れマルコフモデル (Hidden Markov Model; 以下 HMM) の学習データとして用いる．また，同様に動作ラベル m と部位選択ベクトル $\boldsymbol{v}$ を取り出し，トピックモデルの対象となるデータベースを作成する．そのデータベースの塊を文書 ($document$) とすると，単語 $w(word)$ が複数連なる文 ($sentence$) があったとして，その文を複数含むものといったように表すことができる．その文書における単語 w は以下の情報を持つ．

$$w \in \{m, \boldsymbol{v}\} \tag{3.2}$$

この文書は実験者によってトップダウンに作成したものを用いる．

この文書を対象として，トピックモデルの一種である LDA を適用する．LDA におけるトピック t_j での単語 w の出現確率 $P(w|t_j)$ に倣い，動作ラベル m_i と部位選択ベクトル $\boldsymbol{v}_n$ の組み合わせを LDA における単語 w として扱う．すなわち，LDA によって出力される出現確率 $P(m_i, \boldsymbol{v}_n|t_j)$ を用いる．　ここからは，トピックモデルによる実装された「どこに着目するべきか」という文脈を，便宜上「トピック」と呼ぶことにする．

3.1.2　パーティクルを用いた動作トピックに基づく動作と部位選択の仮説相互推定

図 3.1 に提案する認識手順の概要を示す．本手法では現在の動作トピックが t_j である確率分布 $P(t_j)$ をパーティクルの集合 $\boldsymbol{Q} \in \{1, 2, \cdots, k, \cdots\}$ によって表現する．パーティクル k はいずれかの動作トピック t_j に属し，パーティクル k が所属するトピックを q_k とする．パーティクル集合 $\boldsymbol{Q}$ のうち，動作トピック t_j に属するパーティクル k の数が動作トピック t_j の確率 $P(j)$ として表現する．また，部位選択ベクトルの選択，単位動作の認識及び次の属するパーティクルの選択はパーティクル k ごとに行う．パーティクル k は次の手順によって更新される．まず，パーティクル k が示すトピック t_j から，動作ラベル $\hat{m}$ と部位選択ベクトル $\hat{\boldsymbol{v}}$ を以下のようにサンプリングする．

$$\hat{m}, \hat{\boldsymbol{v}} \sim P(m_i, \boldsymbol{v}_n|t_j) \tag{3.3}$$

ここで $P(m_i, \boldsymbol{v}_n|t_j)$ は LDA によって生成される，トピック t_j に対応した m, $\boldsymbol{v}$ の出現確率である．単位動作の認識処理は，サンプリングされた部位選択ベクトル $\hat{\boldsymbol{v}}$ を用いる．観測された動作信号は，あらかじめ定義された身体部位ごとに L 個に分割される．また，同様にして動作ラベル m_i に対応する HMM のパ

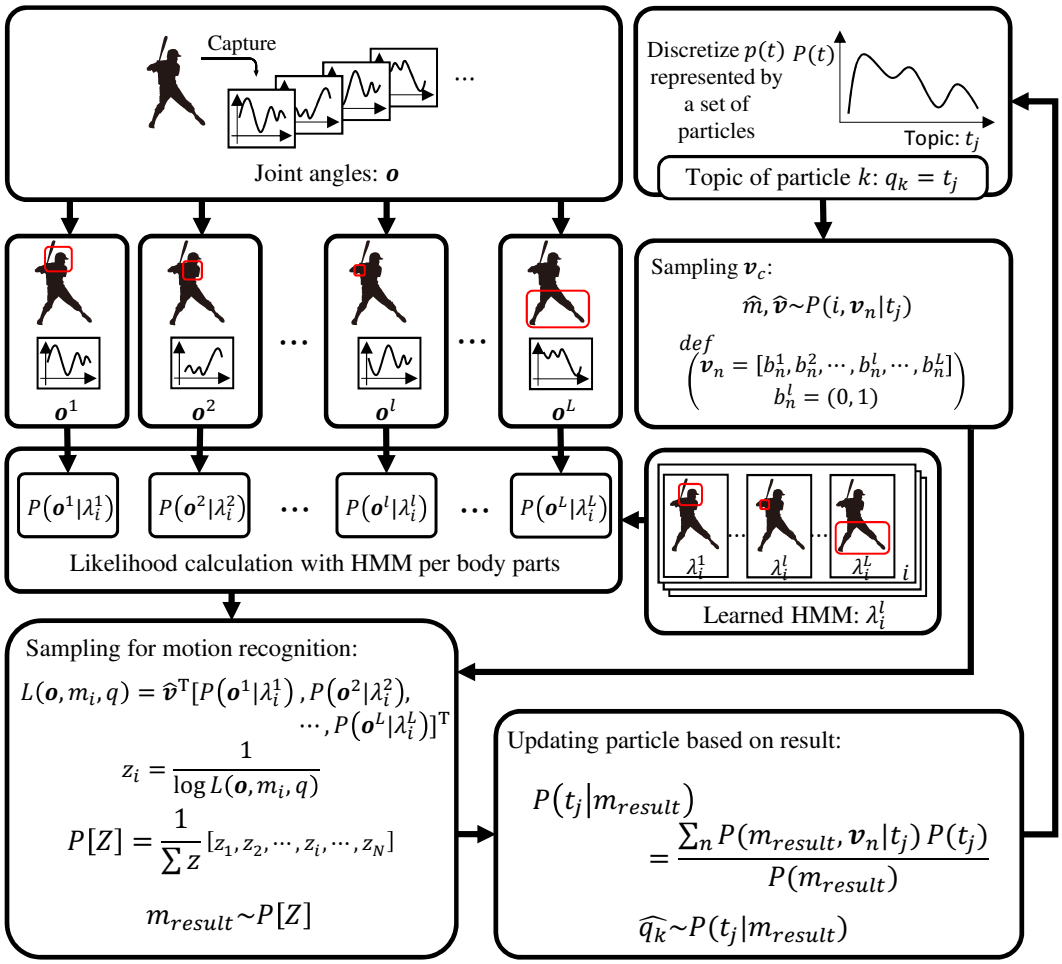


図 3.1　パーティクルによる単位動作の認識と動作トピック推定の処理の流れ

ラメータ λ_i^l を，部位 l ごとに学習している．その λ_i^l を用いて観測された動作信号に対する尤度 $P(\boldsymbol{o}^l|\lambda_i^l)$ を部位 l ごとに求める．観測された動作信号に対する尤度 $P(\boldsymbol{o}^l|\lambda_i^l)$ と部位選択ベクトル $\hat{\boldsymbol{v}}$ を用いて，パーティクル k による認識結果 m_{result} のサンプリングを以下のように行う．

$$m_{result} \sim P\,[Z] \tag{3.4}$$

ただし，$P\,[Z]$ は以下の式によって計算される．

$$P\,[Z] = \frac{1}{\sum z}\,[z_1, z_2, \cdots, z_i, \cdots, z_M] \tag{3.5}$$

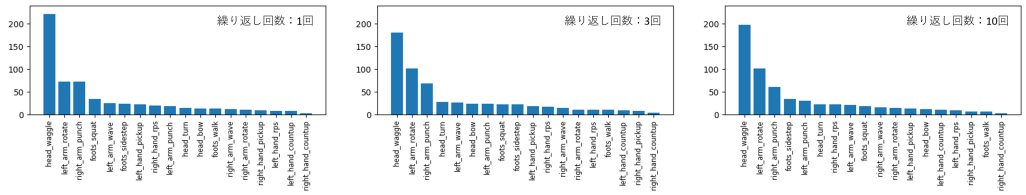


図 3.2　提案手法によるパーティクルごとの認識結果

ここで z_i は次の式 (3.6) で計算される．

$$z_i = \frac{1}{\log S(\boldsymbol{o}, i)} \tag{3.6}$$

$S(\boldsymbol{o}, i)$ は，部位選択を行った後の統合認識尤度を意味し，次のように計算する．

$$\begin{aligned} S(\boldsymbol{o}, i) = \boldsymbol{v}_n^{\mathrm{T}} \big[& P(\boldsymbol{o}^1 | \lambda_i^1), P(\boldsymbol{o}^2 | \lambda_i^2), \cdots, \\ & (\boldsymbol{o}^l | \lambda_i^l), \cdots, P(\boldsymbol{o}^L | \lambda_i^L) \big]^{\mathrm{T}} \end{aligned} \tag{3.7}$$

最後に，認識結果 m_{result} が所属しているであろうトピック t_j に基づきパーティクル k の更新を以下のように行う．

$$q_{\hat{k}} \sim P(t_j | m_{result}) \tag{3.8}$$

ただし，$P(t_j | m_{result})$ は次式のように求める．

$$P(t_j | m_{result}) = \frac{\sum_n P(m_{result}, \boldsymbol{v}_n | t_j) P(t_j)}{P(m_{result})} \tag{3.9}$$

更新後のパーティクル $\hat{k}$ はいずれかのトピック t_j に属する．これらの処理をパーティクル集合 $\boldsymbol{Q} \in \{1, 2, \cdots, k, \cdots\}$ の全てのパーティクル $\{1, 2, \cdots, k, \cdots\}$ に対して実行し，認識結果 m_{result} のヒストグラムと更新後のパーティクルの分布を得る．また，この一連の流れを複数回繰り返し実行することで，動作トピックに基づいて動作認識と部位選択の仮設を相互に推定する．

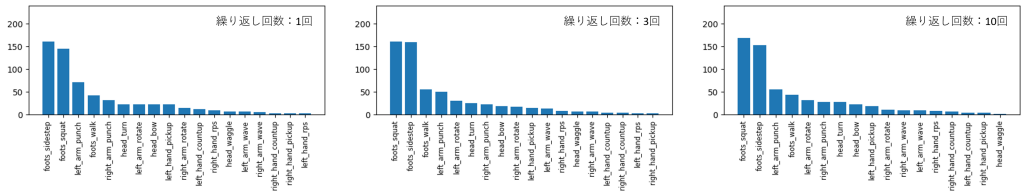


図 3.3　部位選択ベクトル $\boldsymbol{v}$ を用いない手法によるパーティクルごとの認識結果

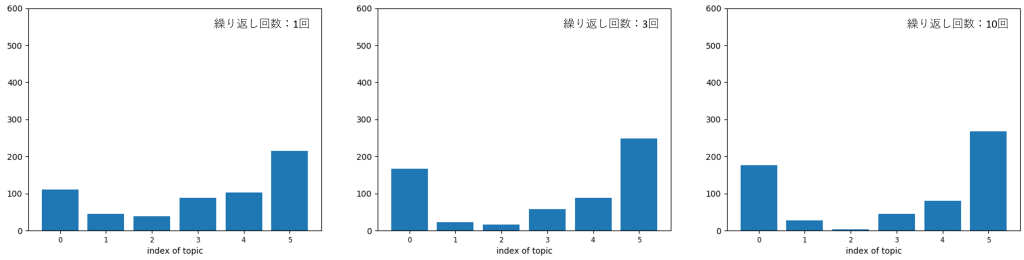


図 3.4　提案手法によるパーティクル集合 $\boldsymbol{Q}$

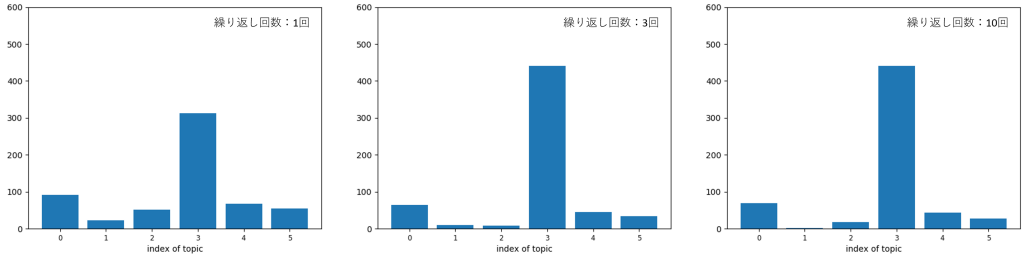


図 3.5　部位選択ベクトル $\boldsymbol{v}$ を用いない手法によるパーティクル集合 $\boldsymbol{Q}$

3.2　同一身体上で実施される複数動作を対象にした検証実験

3.2.1　対象のデータ

表 3.2 に実験に用いる単位動作の一覧を示す．本実験では，身体の部分で実施される動作に対する認識を目的としているため，分かりやすく身体部位ごとの動作を対象とする．具体的に図 3.1 に示す分割数 L は 6 であり，信号は $\{c_1$：頭

表 3.2　実験に用いる各部位における単一動作

c_l: Body part	m_i: Motion label
c_1: head	m_1: head bow
	m_2: head turn
	m_3: head waggle
c_2: left arm	m_4: left arm punch
	m_5: left arm rotate
	m_6: left arm wave
c_3: left hand	m_7: left hand countup
	m_8: left hand pickup
	m_9: left hand rps
c_4: right arm	m_{10}: right arm punch
	m_{11}: right arm rotate
	m_{12}: right arm wave
c_5: right hand	m_{13}: right hand countup
	m_{14}: right hand pickup
	m_{15}: right hand rps
c_6: foots	m_{16}: foots squat
	m_{17}: foots sidestep
	m_{18}: foots walk

部，c_2：左腕，c_3：左手，c_4：右腕，c_5：右手，c_6：脚部 } に分割される．表 3.2 に示す動作ラベル m_i は全部で 18 種で，分割された部位それぞれに 3 種の単位動作が用意されている．これらの動作ラベル m_i に対応する動作パターンに対して，身体部位 c_l ごとに HMM で学習した．LDA の対象となる動作コーパスは，*word* が 1〜50 個ならぶ *sentence* を 10000 文持つ *document* を用意した．本来であれば，連続した人の日常動作を観察して得ることが好ましいが，本実験の対象動作は日常動作とは異なる．そのため動作コーパスは，トピックや部位選択ベクトルがある程度偏るよう恣意的なものを用意した．トピック数 J は 6 として LDA を実行し，出現確率 $P(m_i, \boldsymbol{v}_n|t_j)$ を得る．また動作トピックの初期の分布は一様分布を仮定し，k のパーティクル数は 600 で，一つの動作トピック t_j に対してパーティクル数が 100 になるように用意する．

表 3.3 LDA による m_i，$\boldsymbol{v}_n$ の出現確率 $P(m_i, \boldsymbol{v}_n|t_j)$

topic t_1	0.134 * “left_arm_rotate-[0,1,0,0,0,0]” 0.125 * “left_arm_wave-[0,1,0,0,0,0]” 0.123 * “left_arm_punch-[0,1,0,0,0,0]” ⋮
topic t_2	0.139 * “right_hand_pickup-[0,0,0,0,1,0]” 0.137 * “right_hand_rps-[0,0,0,0,1,0]” 0.131 * “right_hand_countup-[0,0,0,0,1,0]” ⋮
topic t_3	0.147 * “left_hand_pickup-[0,0,1,0,0,0]” 0.131 * “left_hand_rps-[0,0,1,0,0,0]” 0.122 * “left_hand_countup-[0,0,1,0,0,0]” ⋮
topic t_4	0.124 * “foots_squat-[0,0,0,0,0,1]” 0.122 * “foots_walk-[0,0,0,0,0,1]” 0.120 * “foots_sidestep-[0,0,0,0,0,1]” ⋯
topic t_5	0.142 * “right_arm_rotate-[0,0,0,1,0,0]” 0.130 * “right_arm_punch-[0,0,0,1,0,0]” 0.128 * “right_arm_wave-[0,0,0,1,0,0]” ⋮
topic t_6	0.119 * “head_waggle-[1,0,0,0,0,0]” 0.111 * “head_turn-[1,0,0,0,0,0]” 0.109 * “head_bow-[1,0,0,0,0,0]” ⋮

3.2.2 実験結果

認識対象となる観測データ $\boldsymbol{o}$ は，表 3.2 の動作ラベル m_i のうち 3 つの動作が同時に同一身体上で実施されるものとする．ただし，3 つの動作を同一身体上で実演することは困難であることから，単一に収録された動作ラベル m_i に対応する動作パターンを複数切り貼りすることで作成する．具体的には，m_3: head waggle を基準とし，m_3: head waggle の身体部位 c_2，c_4 をそれぞれ，m_5: left arm rotate の c_2，m_{10}: right arm punch の c_4 に差し替えたものを用意した．

また，提案手法の比較手法として，本手法のうち部位選択ベクトル $\boldsymbol{v}$ を用いない手法 (以下比較手法) を設定した．

図 3.2 と図 3.3 にそれぞれ提案手法と比較手法によるパーティクルごとの認識結果を示す．縦軸は結果を示したパーティクルの数，横軸は認識結果のラベルで，数の多い順に並ぶ．グラフは左から，繰り返し回数 1 回，3 回，10 回の結果である．提案手法の結果では，m_3: head waggle が一つ大きく表れ，次に m_5: left arm rotate，m_{10}: right arm punch と並ぶ．これらの認識結果は用意した認識対象となる観測データ $\boldsymbol{o}$ のラベルと同値であり，同一身体上での複数動作を同時に認識可能になることを示唆する．繰り返し回数を重ねることでの大きな結果の変化はないが，繰り返し回数 1 回の時では m_3: head waggle が次の m_5: left arm rotate と比較して大きく離れているのに対し，繰り返し回数 10 回の時では両者の差は縮まっている．これは，繰り返し推定するうちに m_3: head waggle のみだけでなく他の単位動作にも候補として浮上してくるプロセスが実装できていると言える．一方で，図 3.3 の比較手法による結果では，m_{16}: foots squat，m_{17}: foots sidestep の 2 つが高く表れており，繰り返し回数いずれの結果においても，正しい認識を行えていない．

図 3.4 と図 3.5 に提案手法及び比較手法におけるパーティクル集合の分布を示す．縦軸はパーティクルの数で，横軸は動作トピックのインデックスを示す．LDA によるトピック分類の結果を表 3.3 に示す．動作トピック $\{t_1, t_2, t_3, t_4, t_5, t_6\}$ はそれぞれ $\{c_2$: 左腕，c_5 : 右手，c_3 : 左手，c_6 : 脚部，c_4 : 右腕，c_1 : 頭部 $\}$ の部位に関わる動作が出現する確率が高い傾向にある．これに対し図 3.4 では，t_6，t_1 が高く表れ，続いてに t_5，t_4 と表れる．一方で図 3.5 では，単位動作認識結果の影響もあり，t_4 に多く集まってしまっている．複数動作を同時認識する際に，複数のトピックを考慮することが難しい結果になっていると言える．これらの結果から，本提案手法は，身体動作の部位を選択し，

単位動作の認識とトピックの相互推定を行うことで，同一身体上で実施される複数の動作を同時に認識可能な手法であることを明らかにした．

3.3 検討と課題

本章で実験に用いた身体動作は部位ごとに用意された意味を持たない動作であった．局所的であっても，とある意味を持つ動作が 2 つ以上の複数部位に着目するべきである条件を考慮する必要がある．今回の実験で示すことはできなかったが，その場合も 2 つ以上のトピックの分布が励起され，その複数の部位に着目するべきであるという分布を生成可能であると考えられる．また，2 か所以上の複数部位に着目すべきということを管理するトピックを設けることで，身体部位単体のトピックとは異なる扱いを得ることが可能になると推測する．動作によっては，複数の部位に着目したいものの，その部位ごとの重要度が異なることも考えられる．その問題に対応するためには，式 (3.1) における b^l の値を連続値にすることで，重みづけが行えると考える．

この実験では，時系列な動作の変化や，文脈そのものの変化を考慮した条件下では実施しなかった．繰り返し処理した後，形の変化したトピックの確率分布を用いることで，これまでの観測に基づいた文脈を用いることが可能であり，すなわち時系列の文脈を考慮する認識が実現可能になる．この本研究の仕組みにおける時系列データに対する実験は，文脈の定義は異なるものの，4 章において実施され，その有効性が検証される．

3.4 まとめ

本章では，本研究の文脈を活用した誤認識低減の取り組みの一つとして，「今どこに着目するべきか」という点を扱う文脈について着目した．部位選択ベクト

ルと呼ばれる，どの部位を選択するかという情報を持つベクトル導入し，文脈と動作の双方向な関係性を持つ推定手法を確立した．実験では，同一身体上で実施される複数動作を対象に，全身を対象にした手法と比較して，よりよい認識を実現した．また，双方向な関係性を持つ推定手法により，今どの部位に着目するべきか，という情報を持つ文脈の推定を可能にした．

4

動作のカテゴリーを扱う文脈と動作の双方向推定

本章では，似ているが全く異なったカテゴリーの動作への誤認識の低減を目的として，動作のカテゴリーを扱う文脈を利用して問題の解決を狙う．具体的な例として，日常生活における動作を対象とし，「掃除」「料理」「ゲームセンター」「挨拶」という 4 つの文脈に焦点を当てる．

4.1　文脈を用いた身体動作認識

とてもよく似ているが観測される場面が全く異なる動作の誤認識を低減させるため，今の文脈の情報を用いる認識手法を実現する．現在の適切な文脈が与え

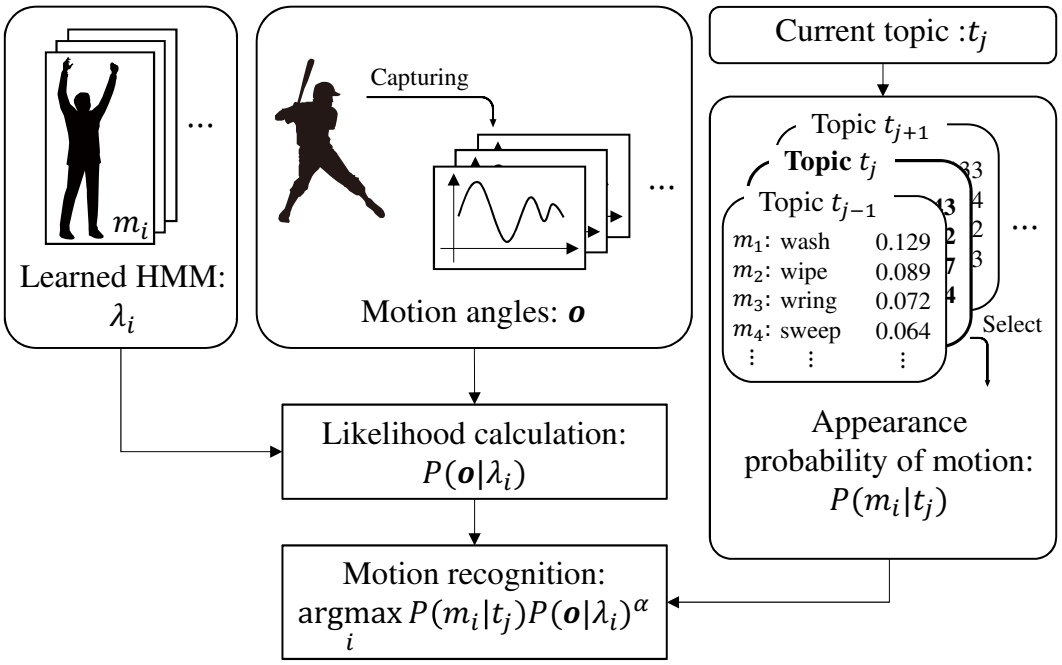


図 4.1 文脈を用いた認識手法の概要図

られる場合と文脈を用いない場合とを比較して，文脈を用いることの重要性を明らかにする．

4.1.1 現在の文脈を活用した認識手法

図 4.1 に提案する身体動作の認識に文脈が作用する手法の概要図を示す．通常のトピックモデルは，学習結果としてトピックごとに図 4.1 の右側に示されるのような単語 w_i ごとの出現確率 $P(w_i|t_j)$ を生成する．この関係性を動作の分野に応用すると，とある一連動作の文脈 t 上で出現しうる動作 m_i の出現確率 $P(m_i|t_j)$ を意味することとなる．人の動作から得られる身体動作信号 $\boldsymbol{o}$ は，隠れマルコフモデル (Hidden Markov Model; 以下 HMM) などの識別器によってあらかじめ学習された身体動作 m_i との尤度が算出される．一連動作の持つ現在のトピック t_j の動作出現頻度 $P(m_i|t_j)$ を，識別器によって算出された尤度に作

用させることで，本研究の目指す文脈に基づく動作の認識精度向上を実現する．ここからは，文脈はトピックモデルにおけるトピックによって実装されているため，実装レベルの説明ではトピックという言葉を用いる．

動作 m_i に対応する HMM を λ_i，観測された身体動作信号を $\boldsymbol{o}$ とすると，身体動作信号 $\boldsymbol{o}$ が観測されたときそれが動作 m_i である確率 $P(\lambda_i|\boldsymbol{o})$ はベイズの定理から以下の式に示すのような関係をもつ．

$$P(\lambda_i|\boldsymbol{o}) = \frac{P(\boldsymbol{o}|\lambda_i)P(\lambda_i)}{P(\boldsymbol{o})} \tag{4.1}$$

$P(\boldsymbol{o}|\lambda_i)$ は HMM に対する動作の認識尤度である．また，$P(\boldsymbol{o})$ はすべての i において同じ値を示すため，i についての比較を行う際には以下の比例関係から省略可能である．

$$P(\lambda_i|\boldsymbol{o}) \propto P(\boldsymbol{o}|\lambda_i)P(\lambda_i) \tag{4.2}$$

ここで，事前確率 $P(\lambda_i)$ は，λ_i と i について対応する m_i を参照し，現在のトピックにおける動作 m_i の出現頻度 $P(m_i|t_j)$ によって与えられるものと仮定すると，次式のように表すことが出来る．

$$S(i, \boldsymbol{o}, t_j) = P(\boldsymbol{o}|\lambda_i)P(m_i|t_j) \tag{4.3}$$

ここで $S(i, \boldsymbol{o}, t_j)$ は i，$\boldsymbol{o}$，t_j に関わるスコアを意味し，計算の途中経過の表現である．トピックに基づいた認識結果 m_k の k は以下の式のように求まる．

$$k = \arg\max_i P(\boldsymbol{o}|\lambda_i)P(m_i|t_j) \tag{4.4}$$

トピックモデルとして Latent Dirichlet Allocation (LDA) [44] を，身体動作信号の識別器として Gaussian Mixture Model-HMM (GMM-HMM) を用いた．GMM-HMM における尤度 $P(\boldsymbol{o}|\lambda_i)$ は確率密度関数であるため，それを微小区

表 4.1　HMM で学習した動作パターンの一覧

m_i	動作パターン
m_1:	ちりとりに入れる
m_2:	床を掃く
m_3:	雑巾を洗う
m_4:	窓を拭く
m_5:	机を拭く
m_6:	雑巾を絞る
m_7:	手を振る
m_8:	エアホッケーする

間 α で積分し確率値に変換するという処理を近似的に行う．すなわち，

$$\arg\max_i \{ \log P(m_i|t_j) + \alpha \log P(\boldsymbol{o}|\lambda_i) \} \tag{4.5}$$

を計算する．ここでは，α は経験則的に定数を入れることとする．

4.1.2　日常動作を例とした認識実験

文脈を用いた手法の有効性を検証するため，「掃除」という文脈上での一連動作を例とした身体動作認識実験を行った．表 4.1 に学習した動作パターンの一覧を示す．ここでは，m_1 から m_6 までの 6 種類の動作が「掃除」に関する動作であり，m_7 と m_8 は「掃除」とは関係の無い動作である．実際の身体動作は，m_4「窓を拭く」と m_7「手を振る」が，m_5「机を拭く」と m_8「エアホッケーする」がそれぞれよく似た動作となっている．掃除の文脈上で行われる動作 6 種類とその動作とよく似た 2 種類を GMM-HMM で学習した．本実験における条件とパラメータ値を表 4.2 に示す．実験に用いる身体動作はモーションキャプチャデバイスによって取得する．取得した各関節の回転量のうち，首・両肩・両手首の 3 軸と両肘・各手の 5 本の指の 1 軸，計 27 つの関節回転量を用いた．Left-to-Right モデルの HMM を用い，状態数は 15，GMM の混合数は 10 として学習した．

表 4.2 実験条件と HMM に係るパラメータ

パラメータ名	値
フレーム周波数	60 Hz
身体動作信号次元数	27
身体動作信号系列長	200
各動作パターンのサンプル数	40
HMM 状態数	15
GMM 混合数	10

次に掃除の文脈に基づいた 90 秒間の一連動作を収録した．収録した一連動作から学習時と同じ 200 フレームを 1 フレームずつずらしながら抽出し，GMM-HMM によって最尤の認識結果を得る．時間推移に対する実際の動作と GMM-HMM による認識結果を図 4.2 に示す．掃除の一連動作に対して，床を掃く，ちりとりに入れる，雑巾を洗う，雑巾を絞る，の 4 動作は正しく認識されている．一方で，机を拭く，窓を拭く，という動作は，それらによく似た掃除以外の文脈で出現する，エアホッケーする，手を振る，という動作へと一部で誤認識されている．そこで動作の認識の際に，現在の文脈の考慮を取り入れた式 (4.5) を適用する．分類対象の系列群として，掃除のほかに，ゲームセンター，あいさつ，料理の 4 種の文脈に従った動作 m_i が 10 個前後連なる系列を文脈ごとに 4 つ，計 16 系列を用意した．それらを LDA によって 4 つのトピックに分類した結果を表 4.3 に示す．自動分類されたトピックごとの単語動作 m_i の出現確率 $P(m_i|t_j)$ が出力されている．トピック t_1，t_2，t_3，t_4 に対応する文脈はそれぞれ，あいさつ，料理，掃除，ゲームセンターとなるように出現する動作単位シンボル m_i の分類が行えていることが確認できる．本実験では現在の文脈は掃除であるとして，t_3 の動作出現確率 $P(m_i|t_3)$ を用いて，トピックモデルに基づく認識を行う．また式 (4.5) における α は 0.003 とし，同様の掃除一連動作に対し認識を行った．トピックモデルに基づいた認識結果を図 4.3 に示す． 図 4.2 では

表 4.3　LDA によるトピック分類とトピックごとの動作 m_i の出現確率

トピック t_1	0.470*お辞儀する	0.211*握手する
	0.194*手を振る	0.051*名刺を交換する
	0.030*呼び込む	0.002*卵を割る
	0.001*運転する	0.001*混ぜる
	0.001*ドラムを叩く	0.001*アイスホッケーをする
	0.001*ガッツポーズ	0.001*食材をちぎる
トピック t_2	0.162*調味料を入れる	0.123*食材を移す
	0.123*炒める	0.101*盛り付ける
	0.081*着火する	0.081*油を入れる
	0.080*混ぜる	0.061*食材を洗う
	0.049*食材を切る	0.042*温度を確認する
	0.042*冷蔵庫を開ける	0.040*卵を割る
トピック t_3	0.245*雑巾を絞る	0.245*雑巾を洗う
	0.162*窓を拭く	0.107*床を掃く
	0.107*ちりとりで集める	0.084*机を拭く
	0.016*掃除機をかける	0.016*ほこりをはたく
	0.001*卵を割る	0.001*混ぜる
	0.001*ドラムを叩く	0.001*食材をちぎる
トピック t_4	0.476*お金を入れる	0.118*太鼓を叩く
	0.115*運転する	0.102*銃を構える
	0.100*ドラムを叩く	0.035*アイスホッケーをする
	0.019*ガッツポーズ	0.011*スロットをする
	0.001*呼び込む	0.001*卵を割る
	0.001*手を振る	0.001*混ぜる

誤認識となっていた，机を拭く，窓を拭く，という動作が，別の文脈のよく似た動作に誤認識することなく，認識精度が向上している．この実験結果から，掃除という文脈に従った認識結果により，似た動作であっても認識誤りを低減させたことを確認した．

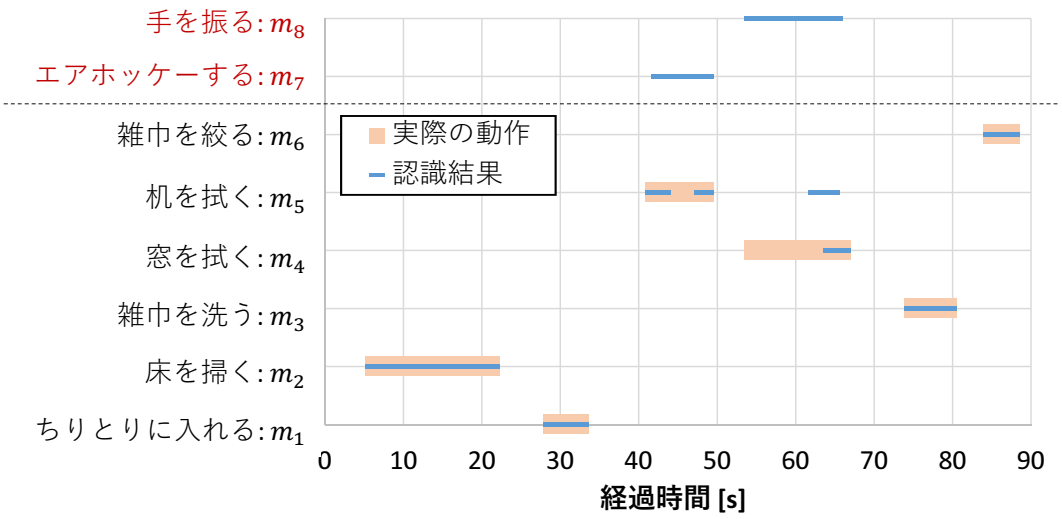


図 4.2　文脈を考慮しない場合の認識結果

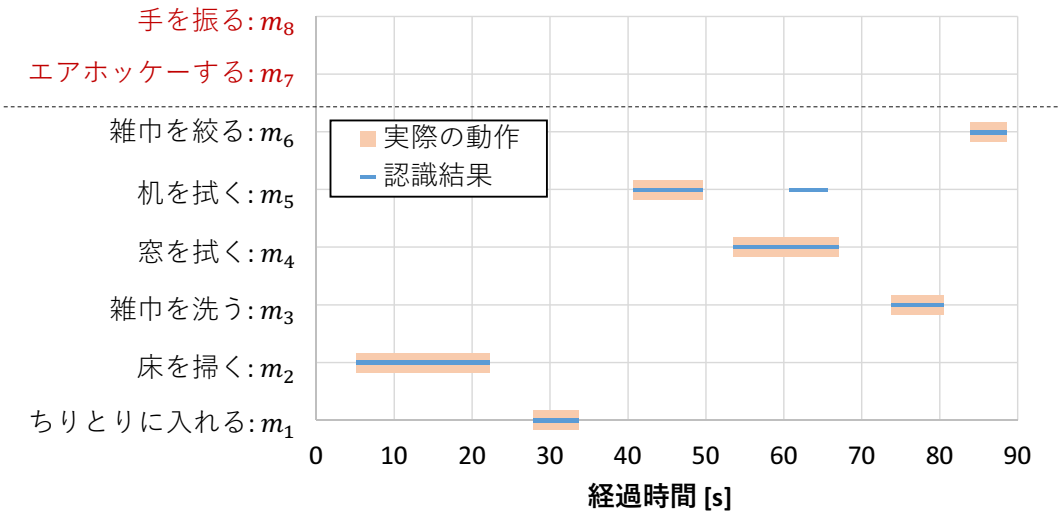


図 4.3　トピックモデルを考慮した認識結果

4.1.3　複数の文脈に基づく動作を対象とした検証実験

先ほどの実験では，掃除のみのトピックを対象とした実験を行い，その対象動作も 8 種のみであった．複数の文脈に基づく動作に対して，提案手法の有効性を検証するため，複数の文脈上での一連動作を例とした身体動作認識実験を行った．表 4.4 に学習した動作パターンの一覧を示す．掃除，料理，ゲームセンター，あいさつの 4 つの文脈上で行われる動作 34 種類を GMM-HMM で学習した．実験に用いる身体動作はモーションキャプチャデバイスによって取得する．取得した各関節の回転量のうち，首・両肩・両手首の 3 軸と両肘・各手の 5 本の指の 1 軸，計 27 つの関節回転量を用いた．Left-to-Right モデルの HMM を用い，状態数 32，GMM の混合数は 10 として学習した．

次に動作パターンを連続して実施する一連動作を収録した．連続動作は，図 4.4 に示した動作パターンを m_0 から m_33 まで順番に連続して約 600 秒間実施したものである．その中でそれぞれの動作パターン m_i はおよそ 15 秒間実施した．図 4.4 にその計測された動作パターンの正解ラベルの系列を示す．動作が番号順に実施されているため，階段状になる．収録した一連動作から 200 フレームを 5 フレームずつずらしながら抽出し，GMM-HMM によって最尤の認識結果を得る．時間推移に対する GMM-HMM による認識結果を図 4.5 に示す． 動作パターンが分類されるトピック毎に色分けし示している．掃除動作は 0 秒から 125 秒程で行われているが，その際に他のトピック上の動作へと誤認識されていることが確認できる．そこで動作の認識の際に，現在の文脈の考慮を取り入れた式 (4.5) を適用する．分類対象の系列群として，表 4.4 と同様の，掃除・料理・ゲームセンター・あいさつの 4 種の文脈に従った動作 m_i が 10 個前後に連なるテキストの系列を文脈ごとに 4 つ，計 16 系列を用意した．それらを LDA によって 4 つのトピックに分類した結果を表 4.5 に示す．自動分類されたトピッ

表 4.4 実験の対象となる動作の一覧

m_i 動作パターン	トピック
m_0: 埃をはたく m_1: ちりとりに入れる m_2: 床を掃く m_3: 掃除機をかける m_4: 雑巾を洗う m_5: 窓を拭く m_6: 机を拭く m_7: 雑巾を絞る	掃除
m_8: 食材を切る m_9: 盛り付ける m_{10}: 卵を割る m_{11}: 冷蔵庫を開ける m_{12}: 炒める m_{13}: 点火する m_{14}: 混ぜる m_{15}: 油を注ぐ m_{16}: 食材をちぎる m_{17}: 食材を移す m_{18}: 調味料を入れる m_{19}: 温度を確認する m_{20}: 食材を洗う	料理
m_{21}: エアホッケーをする m_{22}: お金を入れる m_{23}: 運転する m_{24}: ドラムをたたく m_{25}: 銃を構える m_{26}: スロットをする m_{27}: 太鼓を叩く m_{28}: ガッツポーズ	ゲームセンター
m_{29}: お辞儀する m_{30}: 呼び込む m_{31}: 名刺を交換する m_{32}: 握手する m_{33}: 手を振る	あいさつ

クごとの単語動作 m_i の出現確率 $P(m_i|t_j)$ が出力されている．トピック t_1，t_2，t_3，t_4 に対応する文脈はそれぞれ，あいさつ，料理，掃除，ゲームセンターとなるように出現する動作単位シンボル m_i の分類が行えていることが確認できる．本実験では現在のトピックを与え，t_j の動作出現確率 $P(m_i|t_j)$ を用いて，式 (4.5) によるトピックモデルに基づく認識を行う．その際の式 (4.5) における

表 4.5　LDA によって得られるトピックごとの動作出現確率一覧

t_1	0.470*お辞儀する	0.211*握手する
	0.194*手を振る	0.051*名刺を交換する
	0.030*呼び込む	0.002*卵を割る
t_2	0.162*調味料を入れる	0.123*食材を移す
	0.123*炒める	0.101*盛り付ける
	0.081*着火する	0.081*油を入れる
	0.080*混ぜる	0.061*食材を洗う
	0.049*食材を切る	0.042*温度を確認する
	0.042*冷蔵庫を開ける	0.040*卵を割る
t_3	0.245*雑巾を絞る	0.245*雑巾を洗う
	0.162*窓を拭く	0.107*床を掃く
	0.107*ちりとりで集める	0.084*机を拭く
	0.016*掃除機をかける	0.016*ほこりをはたく
t_4	0.476*お金を入れる	0.118*太鼓を叩く
	0.115*運転する	0.102*銃を構える
	0.100*ドラムを叩く	0.035*エアホッケーをする
	0.019*ガッツポーズ	0.011*スロットをする

表 4.6　α の値ごとの認識率

Condition	Recognition rate
Without topic model	0.20091
$\alpha = 0.0005$	0.32750
$\alpha = 0.0010$	0.29727
$\alpha = 0.0020$	0.26523
$\alpha = 0.0030$	0.24750

α は 0.0005 で行った．トピックモデルに基づいた認識結果を図 4.6 に示す．トピックモデルを用いない場合と比較し，別のトピックの動作への誤認識の頻度が減少している．表 4.6 にトピックモデルを用いない場合と提案手法の α の値を変化させた際の認識率を示す．この認識率は実際の動作に対する認識結果の正答率を示している．トピックモデルを用いない場合と比較して，提案手法では認識率が向上していることが確認できる．この実験結果から，トピックモデルを用

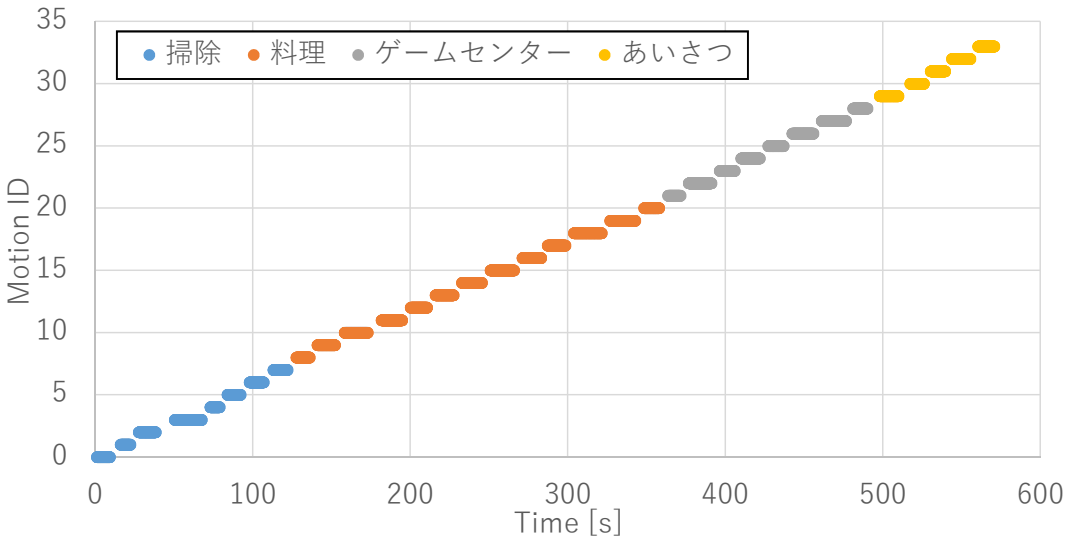


図 4.4　正解の動作ラベルの系列

いた認識結果により，似た動作でもトピック内での出現確率の低い動作の確率が低下し，認識精度が向上したことを確認した．

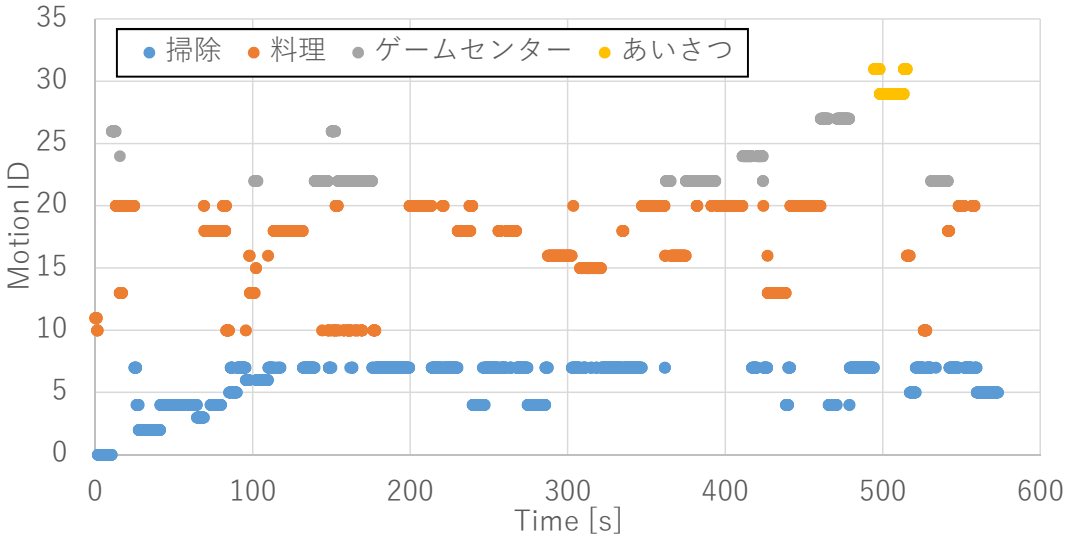


図 4.5　トピックの情報を用いない手法による認識結果の系列

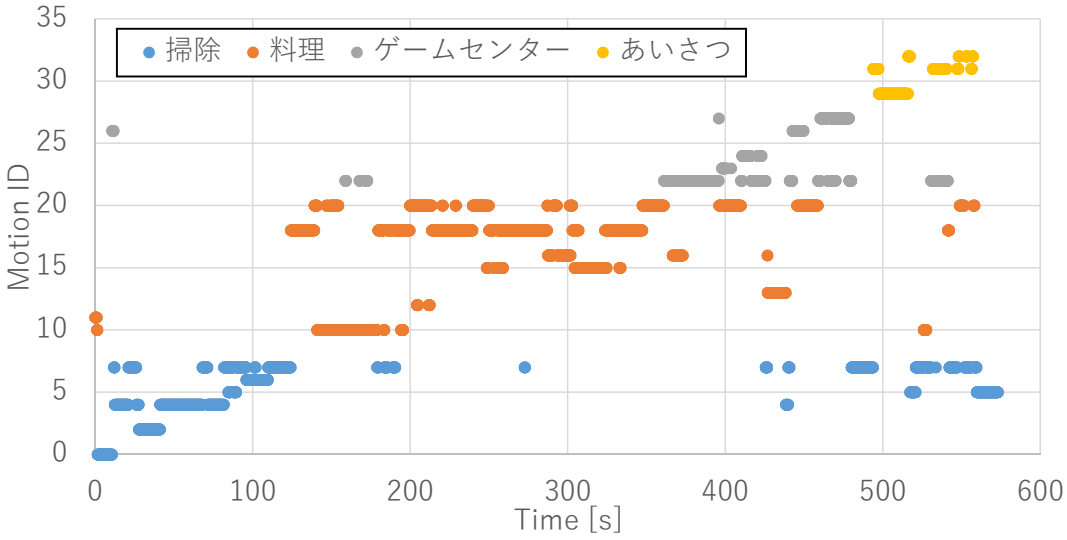


図 4.6　トピックの情報を用いる手法による認識結果の系列

4.2 時系列の変化を考慮した文脈と動作の双方向推定

4.2.1 文脈と動作の双方向推定手法の構成と実装

本手法は，上位層の文脈と下位層の動作認識が互いに影響しあう仕組みを構築する．提案する手法の処理の手順と概要を図 4.7 に示す．この図を基に，1) トピックに基づく動作認識，2) 動作認識結果に基づくトピック分布の再推定，の 2 つの手順に分けて説明を行う．

4.2.1.1 トピックに基づく動作認識処理

トピックの確率 $P(t_j)$ は離散的な粒の集合として表現し，その粒 1 つをここではパーティクルと呼ぶ．図中央部にあるグラフは縦軸が粒の個数，横軸はトピックの ID を示す．このように，とあるパーティクルはいずれかのトピックに所属する．k 番目のパーティクル k が所属するトピックを q_k とし，q_k は次のようなトピックのインデックスを持つ．

$$j \in \{1, 2, \cdots, J\}. \tag{4.6}$$

ここで J はトピックの数を，j 番目のトピックは t_j を表す．また，q_k の集合であるトピックの離散分布 $\boldsymbol{Q}$ を次のように定義する．

$$\boldsymbol{Q} = \{q_1, q_2, \cdots, q_K\}. \tag{4.7}$$

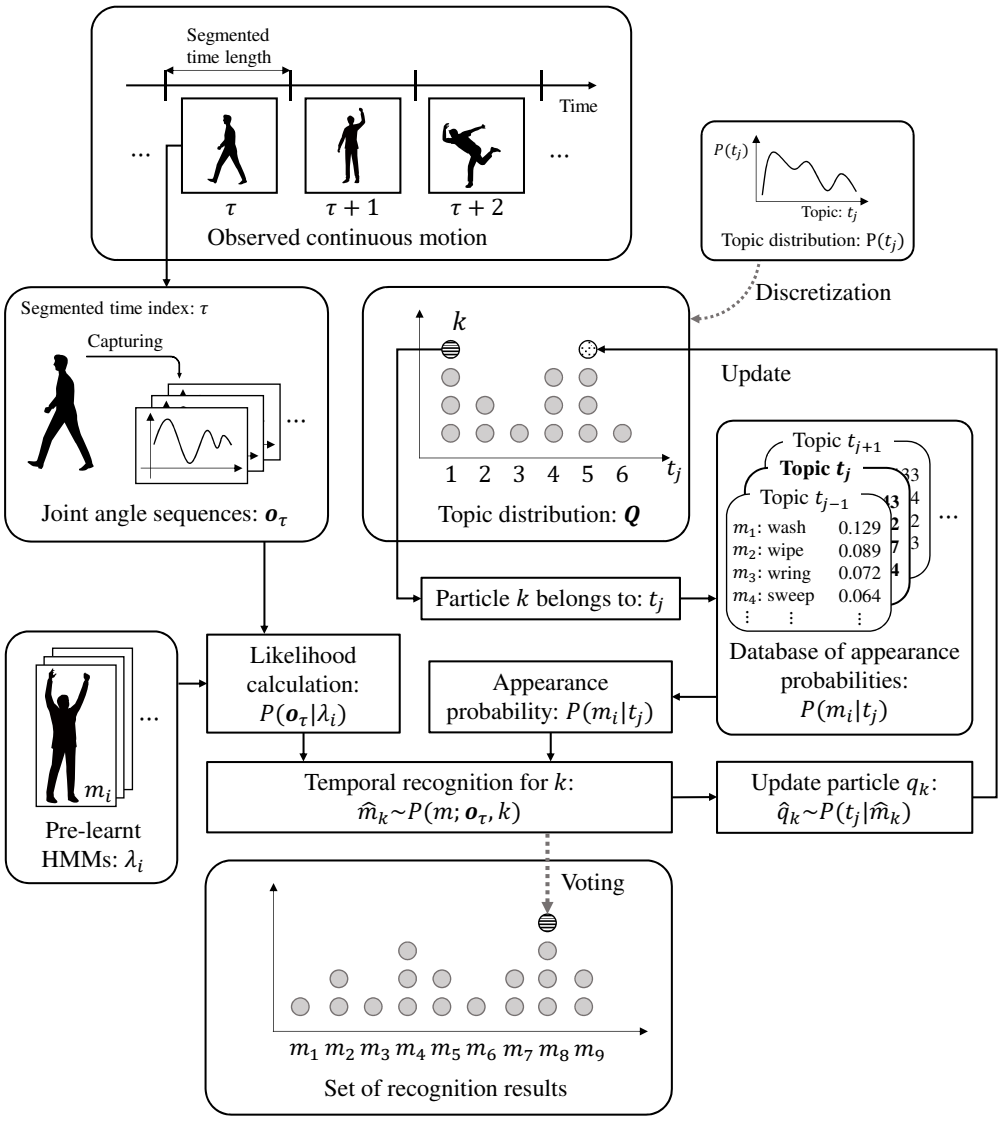


図 4.7 パーティクルによるトピック情報を用いた動作認識とトピック分布の更新手順

ここで K はパーティクルの数である．この離散分布からトピックの確率分布 $P(t_j)$ を次にように連続化して得ることが出来る．

$$P(t_j) = \frac{n(\boldsymbol{R})}{K}. \tag{4.8}$$

$$\boldsymbol{R} \in \{q_k; q_k = j\}. \tag{4.9}$$

ここで，$n(X)$ は X の数を数える関数で，$\boldsymbol{R}$ はパーティクル q_k が t_j になる部分集合である．この後の処理はパーティクルごとに実行される．

この手法は，時系列に観測される長時間の動作を対象とする．図上部に示すように，観測される連続動作は，あらかじめ動作プリミティブごとにセグメントされる．とある時間 τ に観測されるセグメントされたとある時間長の動作プリミティブにおける関節角の時系列データを $\boldsymbol{o}_\tau$ とする．その関節角 $\boldsymbol{o}_\tau$ に対する尤度 $P(\boldsymbol{o}_\tau|\lambda)$ を，学習済みの HMMλ_i を用いて計算する．「床を掃く」「机を拭く」といったような動作プリミティブの動作ラベルを m_i とする．HMMλ_i は動作ラベル m_i に対応する．とあるパーティクル k が所属するトピック t_j における動作 m_i の出現確率 $P(m_i|t_j)$ を参照する．HMM による尤度 $P(\boldsymbol{o}_\tau|\lambda)$ および動作出現確率 $P(m_i|t_j)$ を統合するような計算を行いたい．例えばその統合したスコアを $S(\boldsymbol{o}_\tau, i, t_j)$ とすると，次のような計算を想定する．

$$S(\boldsymbol{o}_\tau, i, t_j) = P(\boldsymbol{o}_\tau|\lambda_i)P(m_i|t_j). \tag{4.10}$$

この計算式は，次のようなベイズの定理からヒントを得たものである．

$$P(\lambda_i|\boldsymbol{o}_\tau) = \frac{P(\boldsymbol{o}_\tau|\lambda_i)P(m_i)}{P(\boldsymbol{o}_\tau)}. \tag{4.11}$$

ここで $P(\lambda_i|\boldsymbol{o}_\tau)$ は $\boldsymbol{o}_\tau$ が観測されたときにおける λ_i の事後確率であり，$P(m_i)$ は m_i の事前確率となる．$P(\boldsymbol{o}_\tau)$ は $\boldsymbol{o}_\tau$ の事前確率であるが，計算上すべての i に対して同じ値が入るため，省略してスコアを求める．ここで $P(m_i)$ に，とあるパーティクルが所属するトピック t_j によって得られる出現確率 $P(m_i|t_j)$ を代入すると目標の計算式となる．

しかしながら，HMM によって求められる尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ は実際には計算の都合上積分処理が省かれている．4.8 に HMM における尤度計算の際における確率密度関数を示す．本手法における HMM は連続型の GMM を用いたものを採

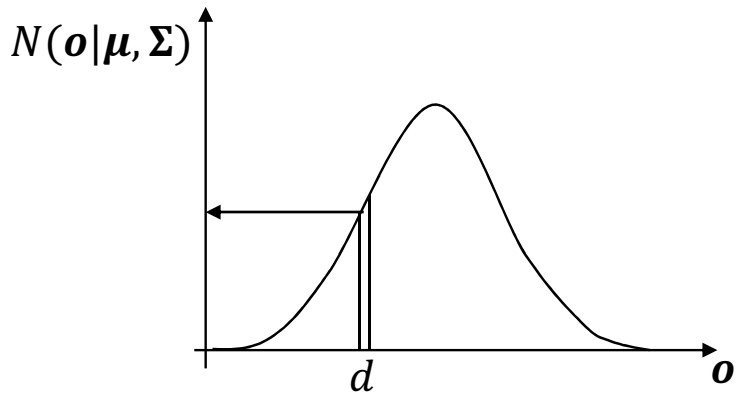


図 4.8　HMM における尤度計算の際における確率密度関数

用しているため，確率密度関数は図のように $N(\boldsymbol{o}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ で表現される．また，$\boldsymbol{o}$ は実際には多次元であるがここでは省略する．確率密度関数から確率を得るためには，図のように微小区間 d で積分処理を行う必要がある．しかしながら，実装の段階ではその積分処理がしばしば省かれる．そのため，その積分処理を事後処理的に担う調整パラメーター α を次のように導入する．

$$S(\boldsymbol{o}_\tau, i, t_j) = P(\boldsymbol{o}_\tau|\lambda_i)^\alpha P(m_i|t_j). \tag{4.12}$$

また計算の都合上，扱う数字のオーダーがとても低くなるため，実際には対数での実装を次のように行う．

$$S(\boldsymbol{o}_\tau, i, t_j) = \exp\{\alpha\left(\log P(\boldsymbol{o}_\tau|\lambda_i) - C\right) + \log P(m_i|t_j)\}, \tag{4.13}$$

ここで C は尤度の値が 1 を超えないようにする定数である．上述のような積分処理が省かれることによって，尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ の値がしばしば 1 を超えてしまうため，定数 C を次のように定義する．

$$C = 2\left|\max \log P(\boldsymbol{o}_\tau|\lambda_i)\right|. \tag{4.14}$$

α は HMM による尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ と出現確率 $P(m_i|t_j)$ を調整するような役割を

持つ．この実験では試験的に α を次のように設定した．

$$\alpha = \frac{\max \log P(m_i|t_j)}{\max \left(\log P(\boldsymbol{o}_\tau|\lambda_i) - C\right)}. \tag{4.15}$$

この式では，α は HMM による尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ と出現確率 $P(m_i|t_j)$ に応じてダイナミックにパラメータが決定される．この意図は，HMM による尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ を出現確率 $P(m_i|t_j)$ と同等に扱うための調整を意味する．

その後，認識結果をサンプリングするための確率分布 $P(m; \boldsymbol{o}_\tau, t_j)$ を先ほどの統合したスコア $S(\boldsymbol{o}_\tau, i, t_j)$ を用いて次のように求める．

$$P(m; \boldsymbol{o}_\tau, t_j) = \frac{S(\boldsymbol{o}_\tau, i, t_j)}{\sum_i S(\boldsymbol{o}_\tau, i, t_j)}, \tag{4.16}$$

ここで，$P(m; \boldsymbol{o}_\tau, t_j)$ は認識結果が m_i になる確率を表す．最後に，認識結果 $\hat{m}$ を確率分布 $P(m; \boldsymbol{o}_\tau, t_j)$ を用いて次のようにサンプリングする．

$$\hat{m} \sim P(m; \boldsymbol{o}_\tau, t_j). \tag{4.17}$$

これらの手順によって上位層の文脈を用いた下位層の動作認識を実装する．

4.2.1.2 動作認識結果に基づくトピック分布の再推定処理

続いて，下位層の動作認識の結果に基づいた上位層の文脈を更新する手順について説明する．この手順も同様にパーティクルごとの処理によって実装される．具体的には，パーティクル k が次に所属する文脈を $\hat{q}_k$ とすると，$\hat{q}_k$ を次のようにサンプリングする．

$$\hat{q}_k \sim P(t_j|\hat{m}), \tag{4.18}$$

ここで $P(t_j|\hat{m})$ は，認識結果 $\hat{m}$ において次の文脈に t_j が選択される確率を表す確率分布である．次のトピックを選択する確率分布 $P(t_j|\hat{m})$ は，認識結果 $\hat{m}$

に基づいて次のように計算する．

$$P(t_j|\hat{m}) = \frac{P(\hat{m}|t_j)P(t_j)}{P(\hat{m})}, \tag{4.19}$$

ここで，$P(t_j)$ は t_j の周辺確率で，この値をどのように設定するかという点をのちの実験において議論する．$P(\hat{m})$ は $\hat{m}$ の周辺確率で次のように計算することが出来る．

$$P(\hat{m}) = \sum_{t_1}^{J} P(\hat{m}|t_j). \tag{4.20}$$

また，$P(\hat{m}|t_j)$ は，すでに得られている出現確率 $P(m_i|t_j)$ において，$m_i = \hat{m}$ となる条件下の確率分布である．

これらの処理の流れを Algorithm 1 に示す．3 行目から 11 行目までの手順はパーティクル 1 つに対して行う処理である．この処理をすべてのパーティクルにおいて実行すると，認識結果はパーティクルの数だけ得られる．すなわち認識結果は分布として得ることが出来る．またこの処理の流れのようにパーティクル分布が更新される．この更新処理を複数繰り返すことによって，上下層のループ処理が実現され双方向な情報処理が実現される．

4.2.2　検証実験

4.2.2.1　対象のデータ

表 4.7 に本実験で対象となる動作のラベルとその動作が所属するトピックを示す．また，その動作の例を図 4.9 に示す．この図は，全身モーションキャプチャデバイスによって得られた動作をビジュアライズしたものを時系列に並べたものである．横軸は時間の流れを表し，右に行くごとに時間が経過する．(a) の「wiping window」の動作と (b) の「bye-bye」という動作を見ると，とてもよく

Algorithm 1 文脈に基づく動作の認識と認識結果に基づくトピック分布の更新手順

for $\boldsymbol{\tau} = 1$ to End of data **do**
  for $k = 1$ to K **do**
    $t_j = q_k$
    $C = 2\,|\max \log P(\boldsymbol{o}_\tau|\lambda_i)|$ ▷ Eq. (4.14)
    Calculate α using Eq. (4.15) ▷ Eq. (4.15)
    Calculate $S(\boldsymbol{o}_\tau, i, t_j)$ using Eq. (4.13) ▷ Eq. (4.13)
    $P(m; \boldsymbol{o}_\tau, t_j) = S(\boldsymbol{o}_\tau, i, t_j) / \sum_i S(\boldsymbol{o}_\tau, i, t_j)$ ▷ Eq. (4.16)
    $\hat{m} \sim P(m; \boldsymbol{o}_\tau, t_j)$ ▷ Eq. (4.17)
    $P(\hat{m}) = \sum_{t_1}^{J} P(\hat{m}|t_j)$ ▷ Eq. (4.20)
    $P(t_j|\hat{m}) = P(\hat{m}|t_j)P(t_j)/P(\hat{m})$ ▷ Eq. (4.19)
    $\hat{q}_k \sim P(t_j|\hat{m})$ ▷ Eq. (4.18)
  end for
  Update $\boldsymbol{Q}$ using $\{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_K\}$
  $i_{result} = \arg\max_i n((\hat{m} = m_i); \{\hat{m}_1, \hat{m}_2, \cdots, \hat{m}\})$
end for

似ていることが確認できる．同様に，(c) の「washing」と (d) の「frying with pan」を見てもその二つの動作を区別することは容易ではない．そして，それぞれ 2 つの動作は全く異なる文脈上での動作である．しかしながら，文脈情報を用いずにこれらの関節角の情報のみで認識をしようと試みると，人間であってもこれらを区別することが難しいことが分かる．所属するトピックが異なるものの動作はよく似ているという動作を含んだ動作を対象としている．すなわち，この実験ではこの間違いやすい動作を対象に，提案する手法がその誤認識を防ぎ，認識率が向上することを明らかにすることを目的としている．

時系列の関節角データは，「Perception Neuron」という名前の全身モーションキャプチャデバイスを用いる．このデバイスによって得られる関節角情報のうち，本実験では，27 軸の関節角を対象に記録する．この 27 軸のうち，15 軸は頭・両肩・両手首において各 3 軸，および，12 軸は両肘・両手の指において各 1 軸として構成される．各動作ラベルにおいて 22 サンプルの時系列関節角データ

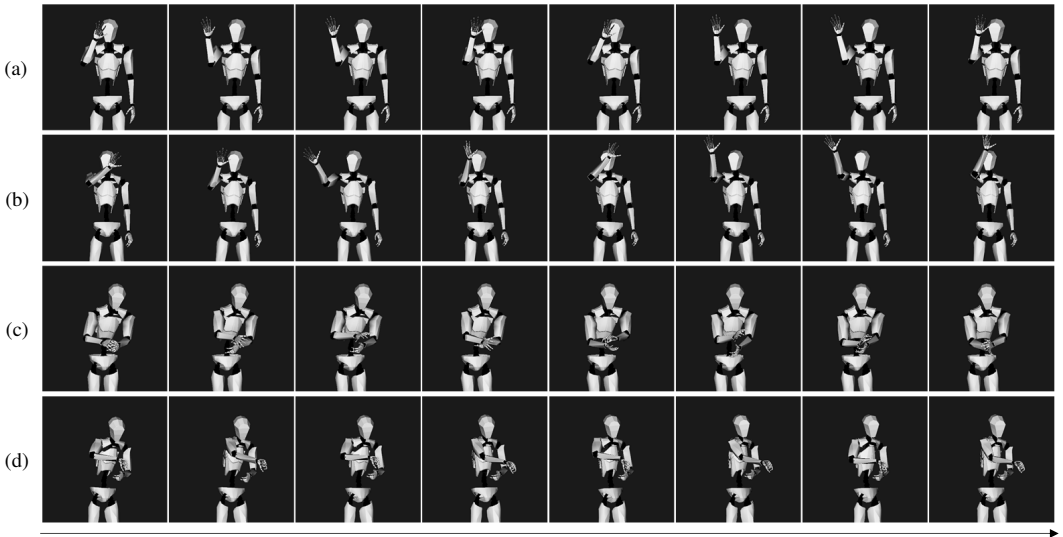
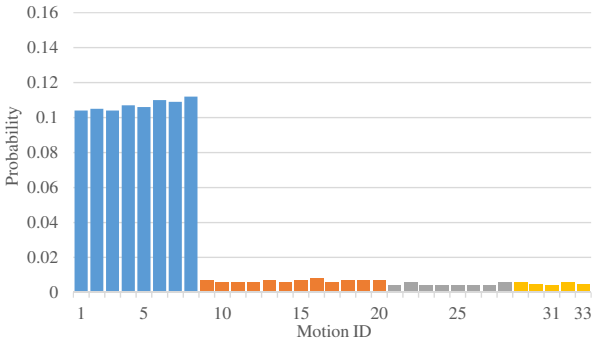


図 4.9　対象の動作の様子とよく似た動作の例．(a)m_6: wiping window. (b)m_{33}: bye-bye. (c)m_5: washing. (d)m_{13}: frying with pan. (a) と (b) の動作，(c) と (d) の動作はそれぞれよく似ており，誤認識を引き起こしやすい．

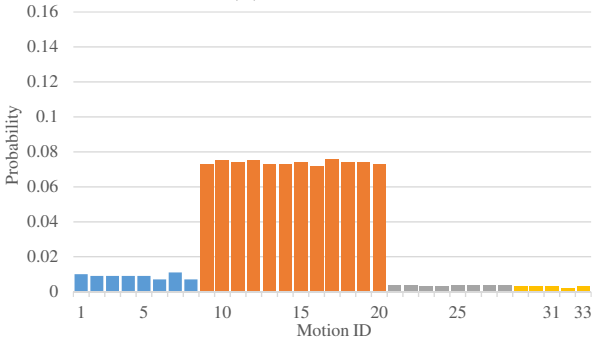
を収録．そのサンプルは全身モーションキャプチャデバイスによって 60[Hz] で収録されていて，1 つのサンプルは平均して 4 秒程度の長さである．

HMM は連続型の GMM-HMM で，left-to-right モデルを採用する．その際の隠れ状態数は 16 で，GMM の混合数は 5 とした．

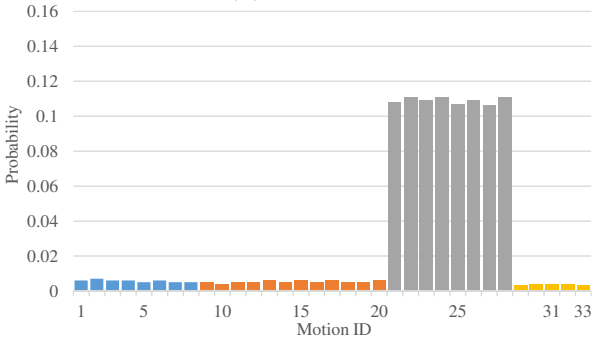
図 4.10 に本実験に用いるトピックごとの動作出現確率 $P(m_i|t_j)$ を示す．その出現確率は，とあるコーパスを対象にした LDA [2] による学習の結果から得る．ここで用いるコーパスを便宜上「コーパス 1 」と呼ぶことにする．このコーパスは 10,000 個のセンテンスから構成される．そしてその 1 つのセンテンスは，1〜50 個の動作ラベルを持つ．このセンテンスの持つ意味は，とある条件・シーン下で観測が予想される人の動作を順番に書き下したものである．例えば，「wiping a window」「wiping a desk」「sweeping the floor」といったように時系列に観測される動作が順番に書き下されている．これは，完全にランダムに発生するのではなく，とある条件・シーンで観測したもの，すなわち「掃除」をす



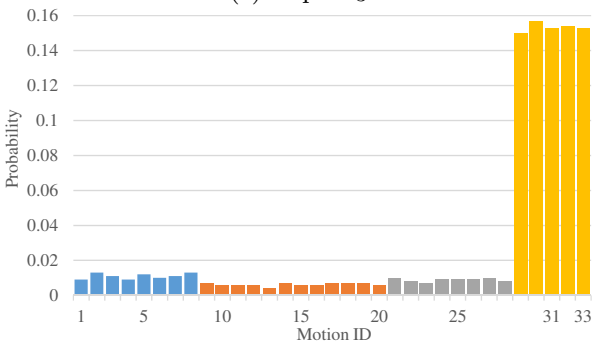
(a) Topic t_1



(b) Topic t_2



(c) Topic t_3



(d) Topic t_4

図 4.10　コーパス 1 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$

表 4.7　実験で取り扱う対象の動作のラベルとその動作が所属するトピックの一覧

m_i: Label name	t_j: Topic
m_1: dusting off	t_1: cleaning
m_2: dustpan	
m_3: sweeping floor	
m_4: vacuuming	
m_5: washing	
m_6: wiping window	
m_7: wiping table	
m_8: wringing cloth	
m_9: cutting	t_2: cooking
m_{10}: dishing foods	
m_{11}: breaking egg	
m_{12}: opening fridge	
m_{13}: frying with pan	
m_{14}: igniting	
m_{15}: mixing	
m_{16}: pouring oil	
m_{17}: replacing	
m_{18}: seasoning	
m_{19}: tearing	
m_{20}: check temperature	
m_{21}: air hockey	t_3: game center
m_{22}: inserting coins	
m_{23}: driving	
m_{24}: playing drum	
m_{25}: gun shooting	
m_{26}: playing slot	
m_{27}: playing taiko	
m_{28}: victory pose	
m_{29}: bowing	t_4: greeting
m_{30}: beckoning	
m_{31}: exchanging card	
m_{32}: shaking hands	
m_{33}: bye-bye	

る人の観測，といったように一つのセンテンスにおいてその文脈が決められており，それに従った観測を記述している．ただし，「掃除」という観測で，完全に「掃除」に関する動作しか観測されないのではなく，ランダムにある程度その他の動作も観測されるものとする．コーパス 1 は，図 4.10 に示した動作出現確率を生成するために調整された恣意的な文脈の偏りを持つ確率値に従ったプログラ

ムによって自動生成されたものである

この実験ではトピックの数 J は，表 4.7 に示したものと同様に 4 である．また，図 4.10 に見て分かるように，各動作が所属するトピックにおいて，高い出現確率を持つことが分かる．この実験では，パーティクルの数 K は 400 で，パーティクルの分布は一様分布を初期状態として用いる．

4.2.2.2 比較手法

提案する手法をほかの手法と比較するために，この実験では 6 つの手法を用いて実施される．その手法は次の (i) から (vi) である．

(i) Only HMM

この手法は HMM のみによる認識手法であり，すなわち文脈情報を用いない手法である．この手法による認識結果の動作インデックスを i_{hmm} とすると，その i_{hmm} は次のように計算される．

$$i_{hmm} = \arg\max_{i} P(\boldsymbol{o}_\tau|\lambda_i). \tag{4.21}$$

ここで $P(\boldsymbol{o}_\tau|\lambda_i)$ は HMM による尤度であり，その尤度が最大となるインデックスを取得する．

(ii) With N-grams

この手法は，N-grams から得られる次のラベルを推測する確率値を本研究で用いる動作出現確率として扱い，認識に用いる手法である．N-grams の学習に用いるコーパスは，LDA に対して用いたものと全く同様のものを用いる．N-grams は，単語の前後関係を確率で表現したモデルで，文章の文脈を考慮する手法であるといえる．しかしながら，HMM による尤度と組み合わせるためには工夫が必要となる．そのため，本研究での取り組みと同様に，式 (4.13) に倣って，統合スコア $S_{ngram}(\boldsymbol{o}_\tau, i)$ を次式のよう

に定義する．

$$S_{ngram}(\boldsymbol{o}_\tau, i) = \exp\{\alpha\,(\log P(\boldsymbol{o}_\tau|\lambda_i) - C) + \log P(m_i^\tau|\hat{m}_i{}^{\tau-1}, \cdots, \hat{m}_i{}^{\tau-N+1})\}, \quad (4.22)$$

ここで，$\hat{m}_i{}^{\tau-1}, \cdots, \hat{m}_i{}^{\tau-N+1}$ は 1 時刻前から $N-1$ 時刻前までの動作の認識結果であり，N は N-grams における学習の対象となる単語の連結数である．また，$P(m_i^\tau|\hat{m}_i{}^{\tau-1}, \cdots, \hat{m}_i{}^{\tau-N+1})$ は $\hat{m}_i{}^{\tau-1}, \cdots, \hat{m}_i{}^{\tau-N+1}$ の情報から次の動作が m_i^τ となる確率を示したもので，これは N-grams によって得ることが出来る．この手法による認識結果の動作インデックス i_{ngram} は次式のように計算を行う．

$$i_{ngram} = \arg\max_i S_{ngram}(\boldsymbol{o}_\tau, i). \quad (4.23)$$

比較手法にこの手法を設定する意図は，次の理由がある．文脈を考慮するための手法はいくつかある中で，N-grams は一般的で強力な手法である．しかしながら，本研究の扱うトピックとして上位と下位の層を構成するような構造を生成できない．すなわち，文脈そのものがどういう分類であるかということを認識できない手法に対して，提案する手法の有効性を検証する狙いがある．

(iii) Not sequential

この手法では，提案する手法におけるトピックの分布を，毎時刻一様分布にリセットする手法である．とある時刻における入力情報に対して，上位と下位のループ処理を実行するものの，認識そのものはその時刻で独立しており，次の時刻へ継承しないものとなる．この手法を用いる理由は，時系列を考慮しない手法と比較する目的のためである．この手法と比較することで，時系列によるトピックの推定が動作の認識に及ぼす影響を検証

する．

(iv) Uniform distribution

ここから 3 つの手法は提案手法で取り扱う確率分布 $P(t_j)$ をどのように設定するかという検証を行うためのものである．この $P(t_j)$ は式 (4.19) に用いる確率分布である．この手法 (iv) では，$P(t_j)$ に一様分布を用いる．

(v) Previous segment

この手法では，$P(t_j)$ に一時刻前のトピックの分布を用いる．一時刻前とは，図 4.7 を例にすると時刻 $\tau - 1$ の時を表す．

(vi) Previous cycles

この手法では，$P(t_j)$ に 1 つ前のパーティクル更新処理におけるトピックの分布を用いる．

この後 2 種類のシーケンスデータに対する 2 つの実験を行う．1 つは，動作 ID1 33 までという順番が固定されたシーケンスデータを対象に行い，そのシーケンスデータをここではシーケンスデータ A と名付ける．このデータを用いる理由は，同じ動作の変化の系列を複数回試行することで，一度きりではなく複数試行による傾向を見るためである．シーケンスデータ A の特徴は，動作 ID の 1 番から順番に実行していることから，文脈の切り替わるタイミングが早い条件であるということである．もう 1 つは，順番が固定されていない文脈の切り替わるタイミングが穏やかなシーケンスデータを対象に行い，このシーケンスデータをシーケンスデータ B と名付ける．シーケンスデータ A との対比として，文脈の切り替わるタイミングが異なる二つのシーケンスデータを用意するためにシーケンスデータ B を用意している．ここで，シーケンスデータとは，複数の動作が時系列に次々とつながっているデータである．本実験では，このシーケンスデータは，あらかじめ別々に収録された動作の個々を人工的につなぎ合わせることで

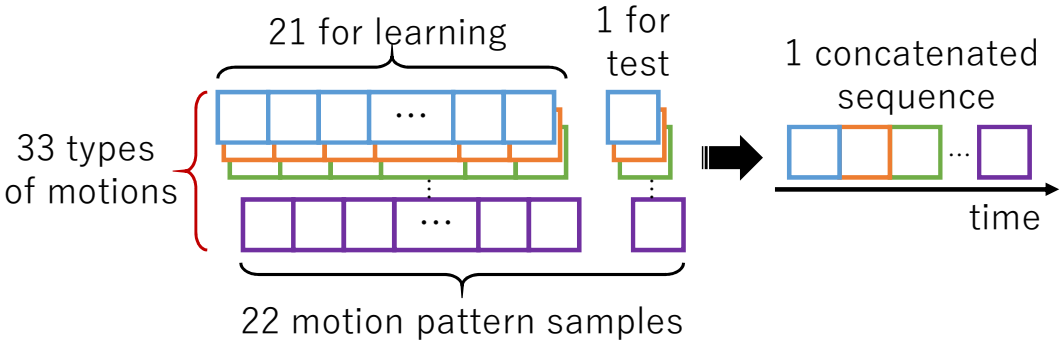


図 4.11　学習とテストための動作サンプルの分け方とシーケンスデータの作り方

作成され，認識の時には動作の個々ごとに自動で区切られるものとする．実験では複数のシーケンスデータを対象に行い，すべてのシーケンスデータに対するすべての認識結果の認識率を得る．

4.2.2.3　シーケンスデータ A における検証実験

図 4.11 に本実験における学習とテストための動作サンプルの分け方とシーケンスデータの作り方の説明図を示す．この実験では，動作の 22 サンプルのうち 21 サンプルを HMM の学習に用い，残りの 1 サンプルを対象のシーケンスデータに用いる．残りの 1 サンプルを動作 m_1 から動作 m_{33} まで順番に接合したものをシーケンスデータとする．この 22 サンプルの分け方は 22 通り存在するため，22 種のシーケンスデータを用意することが出来る．

すべての手法による認識率の一覧を表 4.8 に示す．これらは 22 のシーケンスデータに対する合計の認識率である．(iii) から (vi) の手法は，上下層において相互のループ構造を持つ認識手法であるため，その繰り返し回数による認識率の違いについても検討を行う．ループ構造の繰り返し回数を，この実験では 1，2，3，5，10 回と変更して実施した．また同様に，N-gram に基づく手法 (ii) においても，2-gram，3-gram，4-gram の 3 種類に対して実施した．この表を見ると，最も高い認識率を出した手法は，繰り返し回数 5 回における手法 (v) である．合

表 4.8 シーケンスデータ A に対する各手法における認識率

Method	Recognition Ratio				
(i) Only HMM	0.661157				
			2-gram	3-gram	4-gram
(ii) With N-grams			0.701101	0.634986	0.556473
	Recognition Ratio on Repeat Cycles				
	once	2 cycles	3 cycles	5 cycles	10 cycles
(iii) Not sequential	0.669421	0.669421	0.665289	0.674931	0.661157
(iv) Uniform distribution	0.707989	0.713499	0.705234	0.681818	0.669421
(v) Previous segment	0.618460	0.696970	0.706611	**0.727273**	0.714876
(vi) Previous cycles	0.614325	0.651515	0.652893	0.654270	0.666667

計の認識率において提案手法が他の手法と比較して高い認識率を持つことが分かった．ここからは，より詳しく実験結果について分析する．

図 4.12 に HMM のみの手法 (i) における認識結果の混同行列を示す．縦軸は動作ラベルで，横軸はシーケンスデータを動作ごとに区切ったのち，区切られた動作の時系列順序におけるインデックスを示す．この混同行列は 22 のシーケンスデータのうち，それぞれがどの認識結果を示したかを割合と濃度で示したものである．22 のシーケンスデータうち，そのインデックスの示す動作のすべてが同じ認識結果を示した場合，該当箇所は白くなり，その度合いを濃度で示している．また，グラフ内の薄い線は，トピックが切り替わる瞬間を表す．用意されたシーケンスデータは動作 ID において 1 から 33 まで順番に実施されているため，すべての認識において正しい認識結果を示した場合，グラフの左上から右下まで対角線上に白くなることになる．このグラフにおいて認識率の低い箇所を見ていくと，「wiping the window」の動作 (m_6) の多くが「bye-bye」の動作 (m_{33}) として誤認識されていることがわかる．図 4.9 の (a) と (b) を見てもわかるように，この誤認識はこれらの動作がとてもよく似ていることから引き起こされたものであると推測できる．

図 4.13 には，ループ構造の繰り返し回数 5 回の手法 (v) における認識の混同

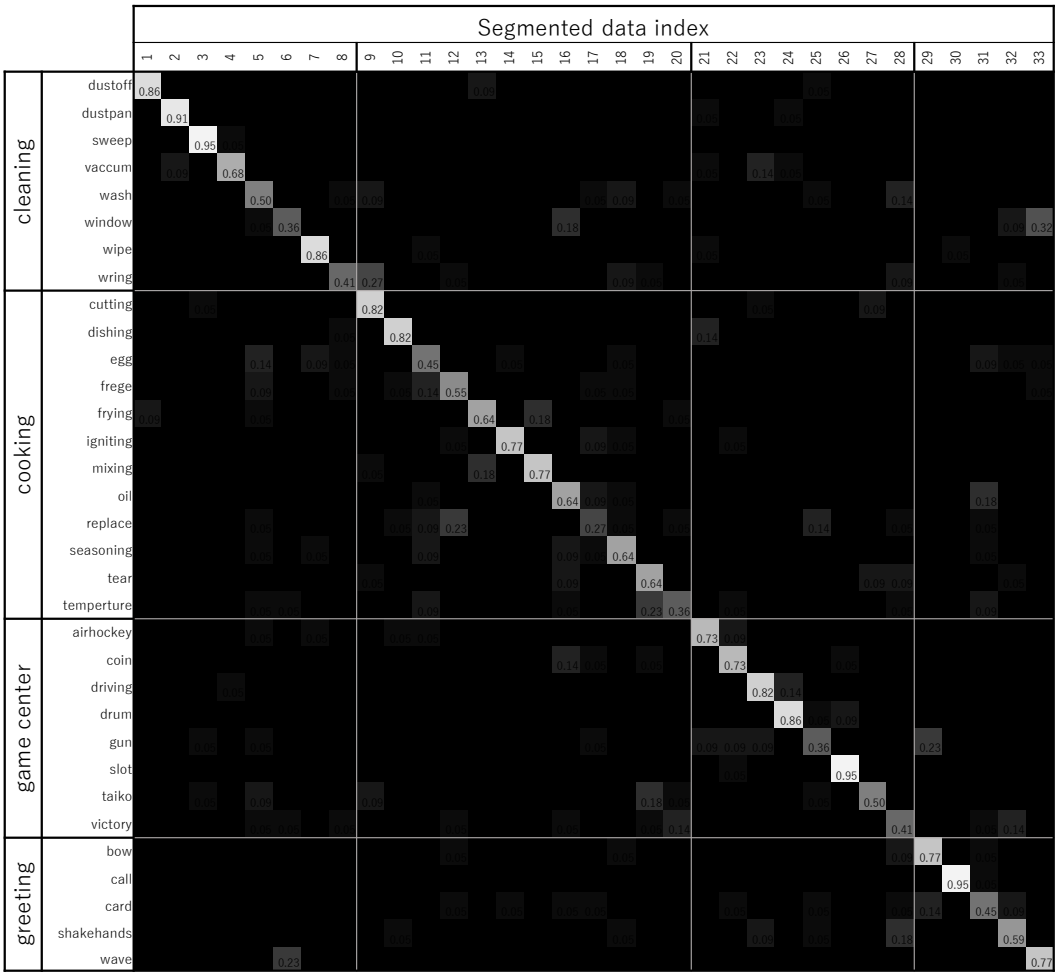


図 4.12　手法 (i) における認識の混同行列

行列を示す．先ほどの誤認識が確認された動作 m_6 と動作 m_{33} に注目するとその誤認識は HMM のみの手法 (i) と比較して激減していることがわかる．この誤認識が減少した理由は，これまでの観測情報から現在の文脈を正しく推測し，その文脈情報を用いて認識率を向上させたことによるものであると推測できる．この結果により，全体的に他の文脈の動作への誤認識が減少し認識率を向上させた．一方で，提案手法 (v) の認識結果における，文脈が切り替わる直後の認識率に注目すると，文脈の切り替わった直後の認識結果の多くにおいて HMM のみ

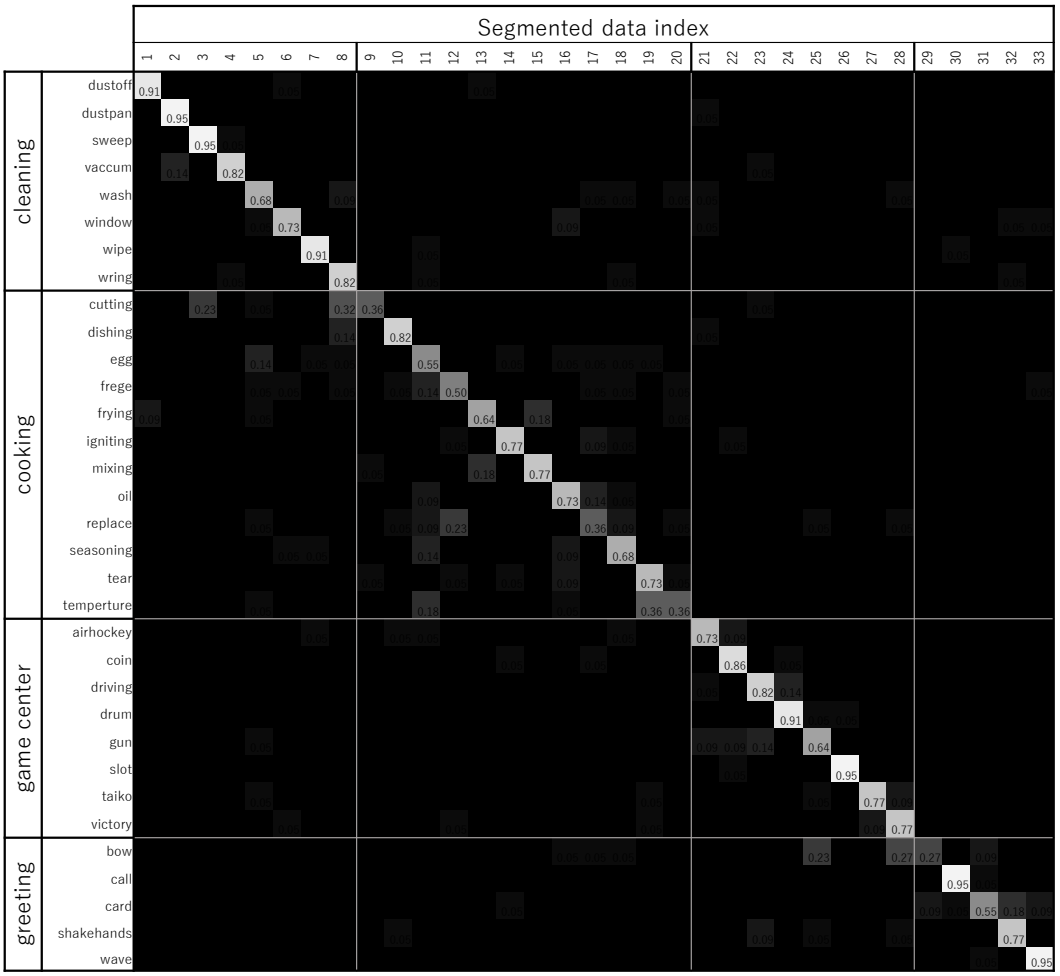


図 4.13　手法 (v) における認識の混同行列

の手法 (i) と比較して認識率が減少している．この認識率が低下している原因は，突然文脈が切り替わったことによって，提案手法がこれまでの観測情報をもとに推定している現在のトピックと認識対象となる動作が所属するトピックが異なるためである．しかしながら，この条件下においても提案する手法は，従来手法と比較して認識率を高く保っている．この結果から提案手法が従来手法と比較して優れていることが示される．

ではその際に，トピックそのものの認識結果はどうであったかを分析する．22

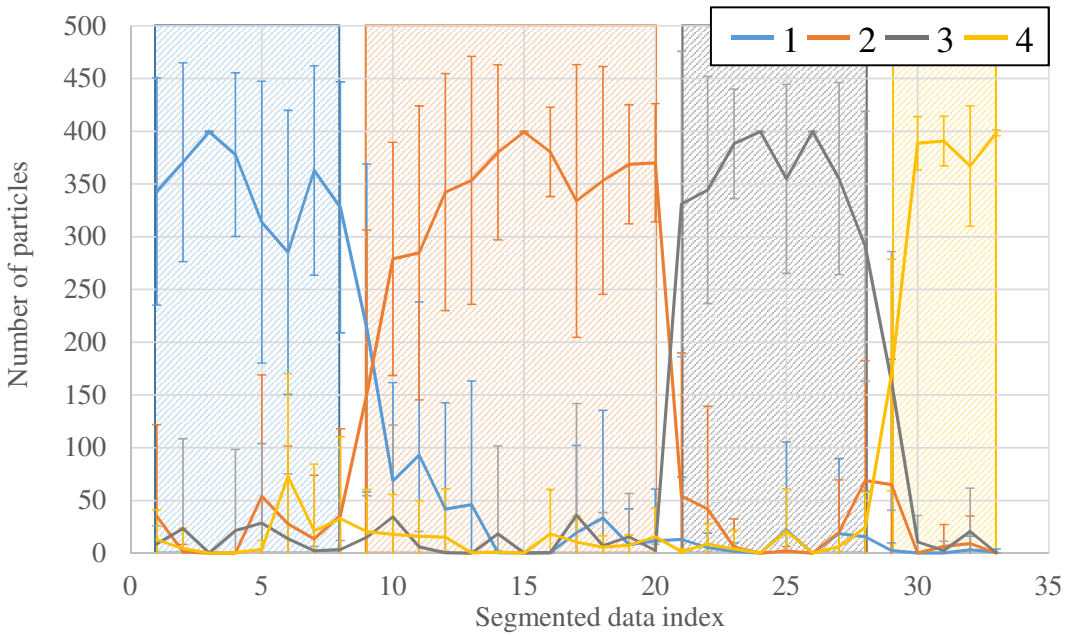


図 4.14　22 のシーケンスデータにおけるトピックの分布の平均と標準偏差の推移

のシーケンスデータに対するパーティクルによるトピックの分布の平均と標準偏差を図 4.14 に示す．横軸は同様に分割された動作のインデックスで，縦軸はパーティクルの数であり，その最大値は 400 である．また，図内の網掛け部分は対象の動作が所属する正しいトピックの範囲を示す．グラフの線の色と網掛け部分の色は対応しており，正しい認識がされている場合は，一番高い線の色が網掛け部分の色と一致することになる．平均の線を追っていくと，おおよそ正しいトピックの推定が行われていることがわかる．すなわち，図 4.13 において動作 m_6 と m_{33} の誤認識を減少させた理由として，正しくトピックが認識できており，そのトピックの情報を正しく利用できたことを示す．一方で，分割された動作のインデックス 9 番を見てみると，トピック 2 の動作を対象にしている条件下であるのにトピック 1 のパーティクルの数のほうが多くなっている．これはこれまでの観測情報から提案手法がトピック 1 の方が高い確率を持つように出力しているためである．そのため，図 4.13 のトピックが切り替わる直後に認

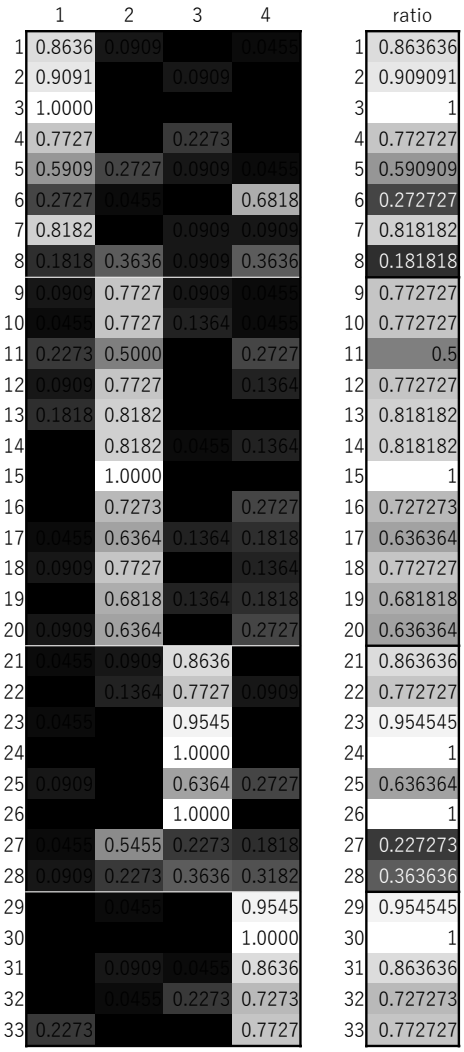


図 4.15 手法 (iii) におけるトピックの認識結果とその正答率

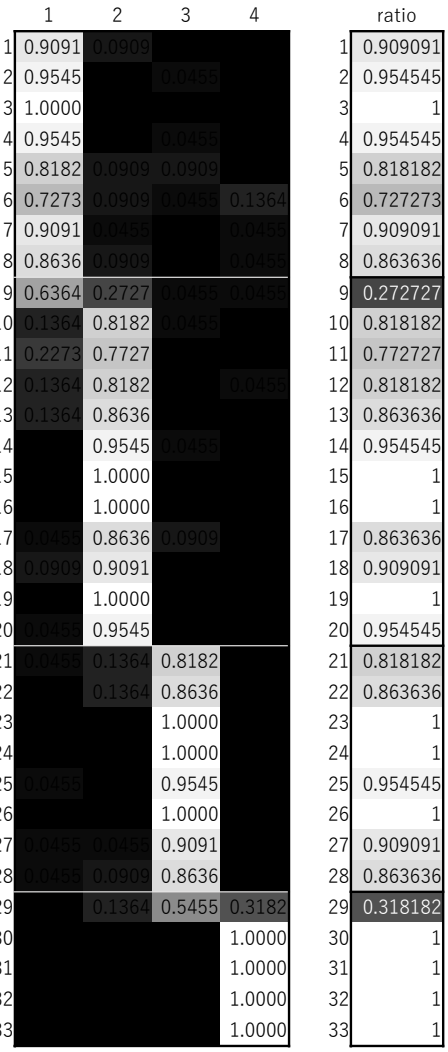


図 4.16 手法 (v) におけるトピックの認識結果とその正答率

識率が低下する要因となっていることがわかる．

手法 (iii) と手法 (v) におけるトピックの認識結果の混同行列と，分轄された動作のインデックスごとにおけるトピックの正答率を，それぞれ図 4.15 と図 4.16 に示す．各図の左側がトピックの認識結果の混同行列であり，縦軸が分割された動作のインデックスで横軸がトピックのインデックスである．数値は 22 のシーケンスデータのうち該当箇所の割合を示し，白が最大黒が最小となる濃度で表

す．また各図の右側はその分割された動作のインデックスにおけるトピックの正答率を示し，濃度で表している．各図に入っている横線は文脈が切り替わる瞬間である．比較手法として手法 (iii) を用いている理由は，時系列を考慮することによる優位性を検証するためである．手法 (i) と (ii) はトピックそのものの認識を行わないため比較ができない．図 4.15 を見ると，図 4.12 において認識間違いが起こりやすかった箇所において，トピックの認識率が低くなっている．例えば動作 m_6 は動作 m_{33} と間違いやすいため．トピック 1 ではなくトピック 4 の認識が多くなり，認識間違いが生じている．そして手法 (iii) における合計のトピック認識率は 0.741047 であった．それに対して，手法 (v) では合計のトピック認識率は 0.881543 であった．トピックの分布を次の時刻へ継承することで大きくトピックの認識率が向上していることがわかる．細部を見ても，認識すべき正しいトピックから大きく崩れることなく認識が行えていることがわかる．ただし，前述のとおり，トピックの切り替わり時においては認識間違いが起こりやすい．しかしながら，トピック認識率が大幅に向上していることから，提案する手法がより良い認識性能を持つことが示される．

これらの実験結果から，シーケンスデータ A を対象にして，提案手法が他の手法と比較して優れていることが明らかになった．

4.2.2.4　シーケンスデータ B における検証実験

先のシーケンスデータ A に対する実験では，観測される動作のトピックが，動作の種類の関係から，高速に切り替わる条件下となっている．しかしながら，実際の環境を想定すると，完全な順番で実施されることはなく，急に様々なトピックに切り替わる条件下はまれである．そのため，この実験では順番のないよりシーケンスデータ B を作成して実施する．

この実験では，10 のシーケンスデータを作成した．1 つのシーケンスデータ

表 4.9 シーケンスデータ B に対する各手法における認識率

Method	Recognition Ratio				
(i) Only HMM	0.6735				
			2-gram	3-gram	4-gram
(ii) With N-grams			0.753000	0.761000	0.651500
	Recognition Ratio on Repeat Cycles				
	once	2 cycles	3 cycles	5 cycles	10 cycles
(iii) Not sequential	0.636000	0.679000	0.692000	0.688500	0.688500
(iv) Uniform distribution	0.626500	0.678500	0.695500	0.691500	0.686500
(v) Previous segment	0.634500	0.765500	0.758500	0.758000	**0.766000**
(vi) Previous cycle	0.635000	0.679500	0.688500	0.673500	0.664000

の中では 10 のトピックについての観測を想定している．またその 1 つのトピック上では 10 から 30 個の動作が並ぶ．その 10 個のシーケンスデータの動作を累計すると 2000 個となった．すなわち，平均すると 1 つのシーケンスデータは約 200 個の動作を持つ．この実験に用いるシーケンスデータと先の実験との大きな違いは，トピックが切り替わる頻度である．先の実験では，5 から 12 動作でトピックが切り替わってしまっていたのに対して，この実験では 10 から 30 動作で切り替わる．

10 のシーケンスデータに対する各手法における認識結果を表 4.9 に示す．ここでは，提案手法における $P(t_j)$ はこれまでの実験から手法 (iii) から手法 (vi) において，手法 (v) が最も優れているとして扱い，HMM のみの手法 (i) と N-grams に基づく手法 (ii) との比較を行う．HMM のみの手法 (i) における認識率は先の実験結果による認識率と大きくは変わらない．N-gram に基づく手法 (ii) は 3-gram において大きく認識率の向上が見られた．これは，先の実験と比較してトピックの切り替わりが緩やかになったことから，これまでの観測情報を長く用いることによる優位性が増したと考えられる．提案手法では，繰り返し回数 10 回においてすべての手法の中で最も高い認識率を得た．この認識率は，先の実験の認識率よりも高いものとなり，提案手法にとってより良い条件であるこ

表 4.10　4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの Kullback-Leibler(KL) 情報量とその合計値

	Topic t_j				
	1	2	3	4	sum
Corpus 1	0.053022	0.074169	0.046605	0.054246	0.677410
Corpus 2	0.163383	0.159178	0.156190	0.162554	1.706054
Corpus 3	0.789590	0.471901	0.745162	1.088385	3.095037
Corpus 4	0.197376	0.132441	0.295696	0.366113	0.991626

とがわかる．

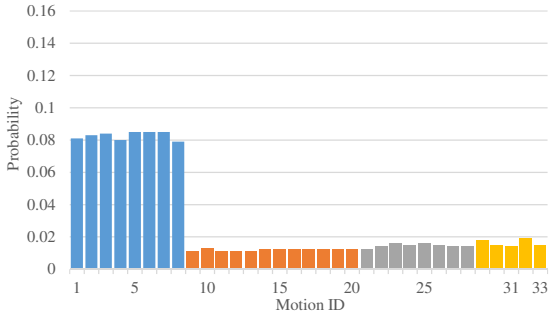
この実験結果から，提案手法はよりシーケンスデータ B に対してより良い性能を発揮し，他の手法と比較して優れていることが示された．

4.2.2.5　コーパスの検討のための実験

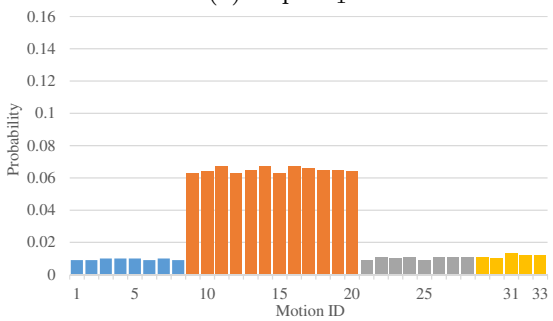
図 4.10 のトピックごとの動作出現確率は，対象のトピックに所属する動作とそれ以外の動作における確率がある程度はっきりとわかるくらいに差のある，いわゆる理想に近いデータであると言える．そこで，図 4.10 における動作出現確率の分布がよりあいまいに変化した際における提案手法のパフォーマンスの変化について検証を行う．

コーパス 1 に加えて，コーパス 2，コーパス 3，コーパス 4 の 3 つのコーパスを新たに追加した．コーパス 2 とコーパス 3 を対象に，LDA によって得られた動作出現確率をそれぞれ，図 4.17 と図 4.18 に示す．図 4.17 と図 4.18 を見てわかるようにこれらは図 4.10 の出現確率と比較してよりあいまいになっていることがわかる．

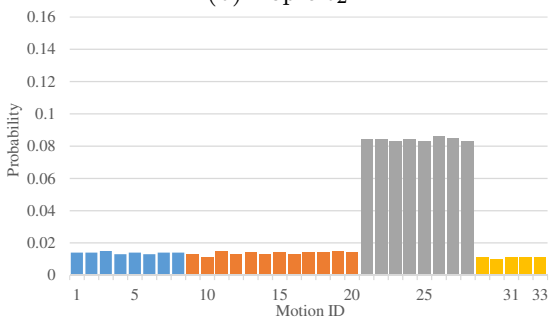
コーパス 4 による出現確率を図 4.19 に示す．この出現確率は，コーパス 1 に対して他のトピックの動作の確率が一部高くなってしまっているような，イレギュラーを含む分布である．これらの動作出現確率を対象として実験を行う．



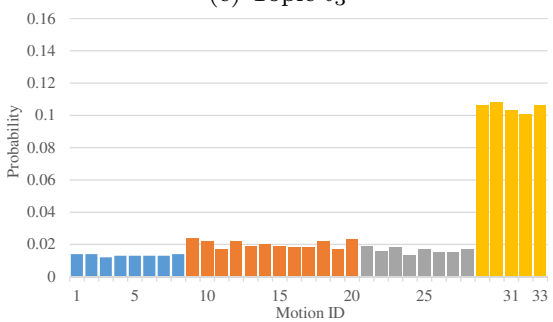
(a) Topic t_1



(b) Topic t_2

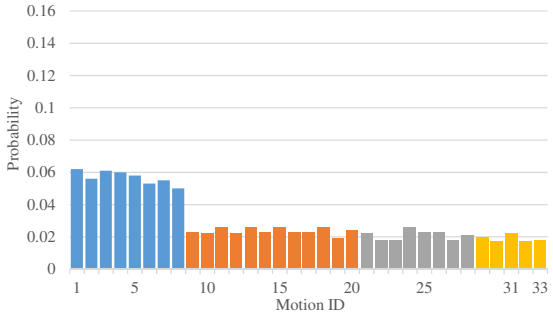


(c) Topic t_3

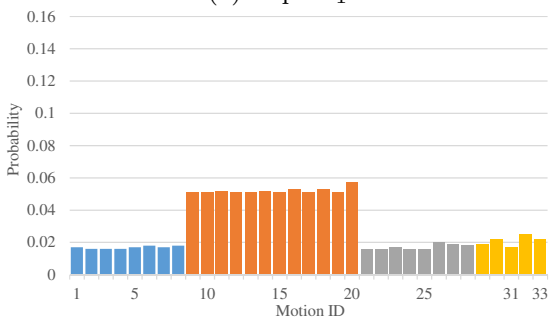


(d) Topic t_4

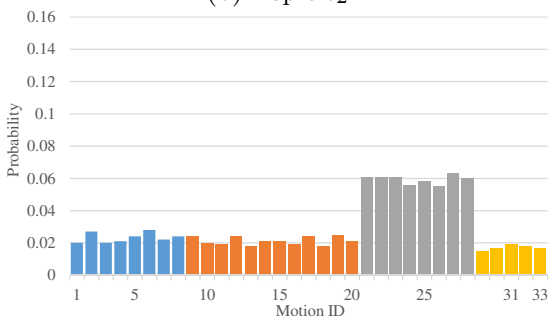
図 4.17 コーパス 2 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$



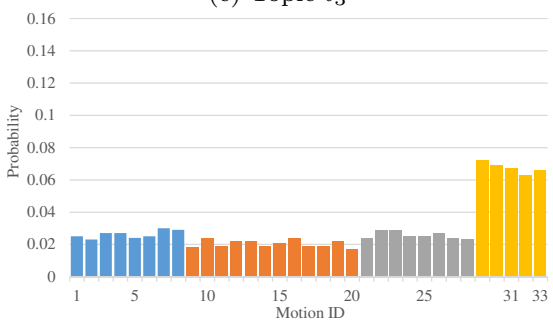
(a) Topic t_1



(b) Topic t_2

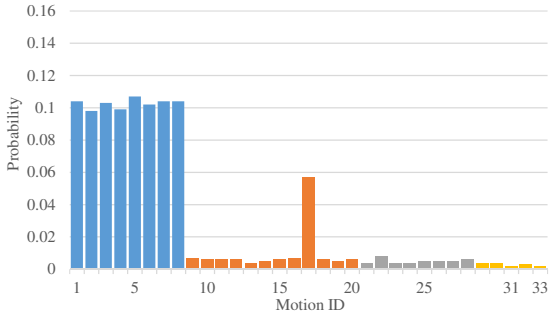


(c) Topic t_3

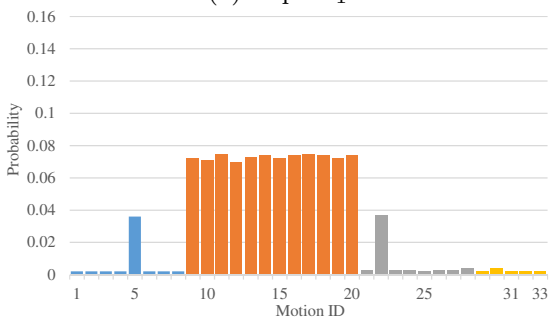


(d) Topic t_4

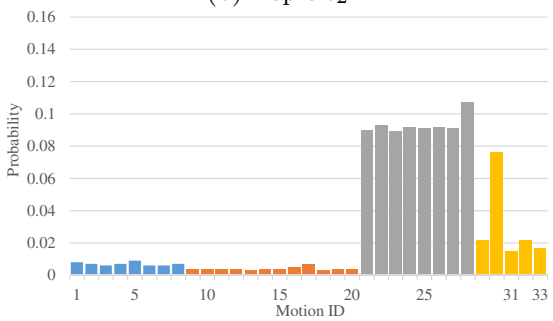
図 4.18　コーパス 3 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$



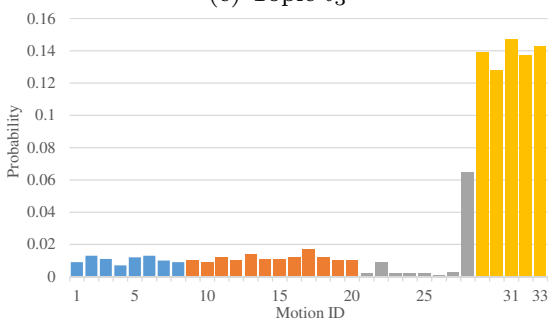
(a) Topic t_1



(b) Topic t_2



(c) Topic t_3



(d) Topic t_4

図 4.19 コーパス 4 を対象にした LDA によるトピック t_j ごとにおける動作出現確率 $P(m_i|t_j)$

表 4.11　4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの JS 情報量とその合計値

	Topic t_j				
	1	2	3	4	sum
Corpus 1	0.051933	0.041511	0.046126	0.088348	0.227919
Corpus 2	0.134400	0.082045	0.129758	0.203020	0.549225
Corpus 3	0.241771	0.152958	0.230081	0.315356	0.940167
Corpus 4	0.065719	0.043603	0.098987	0.120729	0.329039

表 4.12　4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの Cosign 類似度とその合計値

	Topic t_j				
	1	2	3	4	sum
Corpus 1	0.995109	0.993862	0.996365	0.991161	3.976497
Corpus 2	0.959460	0.978047	0.963470	0.930708	3.831687
Corpus 3	0.823710	0.907435	0.844129	0.765550	3.340826
Corpus 4	0.977293	0.978670	0.946232	0.965111	3.867308

所属するトピックの動作しか出現しないような理想的な動作出現確率分布を $P_{ideal}(m_i|t_j)$ とすると，図 4.10，図 4.17，図 4.18，4.19 に対する分布の比較を行うことができる．表 4.10 に 4 種類の出現確率における理想分布 $P_{ideal}(m_i|t_j)$ に対するトピック t_j ごとの Kullback-Leibler(KL) 情報量とその合計値を示す．その合計値 KL_{sum} は次のように計算する．

$$KL_{sum} = \Sigma_j D_{KL}(P(m_i|t_j)||P_{ideal}(m_i|t_j)) \tag{4.24}$$

ここで KL 情報量 $D_{KL}(P(i)||Q(i))$ は次式の通りである．

$$D_{KL}(P(i)||Q(i)) = \Sigma_i P(i) \log \frac{P(i)}{Q(i)} \tag{4.25}$$

表 4.10 を見てわかるように，KL 情報量はコーパス 1 の動作出現確率と比べてコーパス 2，コーパス 3 における動作出現確率は大きく離れていることがわかる．また，JS 情報量 (Jensen-Shannon divergence) についても調査を行った．JS 情報量 D_{JS} は次のように定義される．

$$D_{JS}(P(m_i|t_j)||P_{ideal}(m_i|t_j)) = \frac{1}{2}D_{KL}(P(m_i|t_j)||M) + \frac{1}{2}D_{KL}(P_{ideal}(m_i|t_j)||M) \quad (4.26)$$

ここで M は次のように定義される．

$$M = \frac{P(m_i|t_j) + P_{ideal}(m_i|t_j)}{2}. \quad (4.27)$$

この JS 情報量の計算結果を表 4.11 に示す．また，$P_{ideal}(m_i|t_j)$ と比較したコサイン類似度においても調査した．$P_{ideal}(m_i|t_j)$ とのコサイン類似度の結果を図 4.12 に示す．各要素は 1 に近いほど似ているとされており，合計値は 4 に近いほうが類似度が高い．JS 情報量およびコサイン類似度の結果の傾向は KL 情報量による結果と大きく変わらず，今後の議論は KL 情報量を中心に行う．

また，具体的に，所属する動作とそうでない動作とのあいまいさを検証するために，あいまいさ $\psi(t_j)$ を次のように定義する．

$$\psi(t_j) = \frac{\mathrm{ave}(P(m_i^{incorrect}|t_j))}{\mathrm{ave}(P(m_i^{correct}|t_j))} \quad (4.28)$$

ここで $P(m_i^{correct}|t_j))$ と $P(m_i^{incorrect}|t_j))$ はそれぞれ，トピック t_j に所属する動作とそうでない動作の出現確率を示す．また，$\mathrm{ave}(X)$ は X の平均値を求める関数である．この式は，トピック t_j に所属する動作が観測される頻度に対して，トピック t_j と関係のない動作がどの程度の割合で観測されるかということを計算している．それぞれのコーパスによる動作出現確率に対して，あいまいさ $\psi(t_j)$ の計算結果を表 4.13 に示す．コーパス 1，コーパス 2，コーパス 3 にお

表 4.13　4 つのコーパスによる動作出現確率に対するあいまいさ $\psi(t_j)$ の計算結果

	Topic t_j			
	1	2	3	4
Corpus 1	0.053022	0.074169	0.046605	0.054246
Corpus 2	0.163383	0.159178	0.156190	0.162554
Corpus 3	0.384000	0.346153	0.354357	0.350784
Corpus 4	0.068989	0.078277	0.110818	0.079250

けるあいまいさ $\psi(t_j)$ は，それぞれおおよそ 5[%]，15[%]，35[%] であった．この条件は過酷なものであり，例えば「掃除」という文脈に関する動作を観測しているはずであるにもかかわらず高い頻度で「掃除」ではない動作が観測されるような条件下であることになる．このような特徴を持つ動作出現確率 4 種類に対して実験を行う．

表 4.14　動作出現確率のためのコーパスを変化させた際における認識率

	Corpus 1	Corpus 2	Corpus 3	Corpus 4
Sequence data A	0.727273	0.679063	0.681818	0.670799
Sequence data B	0.758000	0.752500	0.723500	0.771000

表 4.14 に，繰り返し回数 5 回の提案する手法 (v) における 2 種類のシーケンスデータに対する，各コーパスによる動作出現確率を用いた際の認識率の一覧を示す．いずれの結果においても，HMM のみの手法 (i) と比較して高い認識率を保っていることがわかる．単純なシーケンスに対する認識率は，動作出現確率の変化に伴い，やや低くなっている．一方で，シーケンスデータ B に対する認識率は，すべてのコーパスによる出現確率において，非常に高い認識率を保っている．

この結果から，シーケンスデータ A のようなトピックが頻繁に切り替わる厳しい条件下においては，提案手法は表 4.10 のように KL 情報量がおおよそ 0.7

表 4.15 シーケンスデータ A に対する認識率における ANOVA の結果と p 値

	(i)	(ii)	(iii)	(iv)	(v)
(ii)	(0.1526)	–	–	–	–
(iii)	(1.0000)	. (0.0776)	–	–	–
(iv)	* (0.0301)	(0.8286)	** (0.0058)	–	–
(v)	*** (0.0002)	(0.1526)	*** (0.0001)	** (0.0085)	–
(vi)	(1.0000)	* (0.0125)	(1.0000)	(0.6093)	*** (0.0001)

Significance codes: 0 ≤ '***' ≤ 0.001 < '**' ≤ 0.01 < '*' ≤ 0.05 < '.' ≤ 0.1

以下の時により効果的な能力を持つことが示される．また，よりシーケンスデータ B に対しては，その KL 情報量が 3.0 を超えた場合においても高いパフォーマンスを発揮する．

4.2.2.6 有意性テスト

表 4.8 および表 4.9 では，複数回試行によるテスト全体の平均値での認識率の比較となっていた．しかしながら，分散を考慮しない平均値だけでの比較では，有意な差があるとは言えない．そこで，提案手法と比較手法に有意な差があるかどうかを検証するため，分散分析（Analysis of Variance; 以下 ANOVA）を行った．検定はホルム (Holm) の方法で多重比較を行った．

表 4.15 にシーケンスデータ A に対する認識率における ANOVA の結果を示す．また，表内括弧で示された数値は ANOVA における p 値を示す．通常 p 値が 0.05 以下であれば有意な差があるとされている．手法 (v) を基に他の手法と比較すると，手法 (i)(iii)(iv)(vi) に対して有意な差があることが明らかになった．一方で，手法 (ii) に対する有意な差は示されなかった．

表 4.16 にシーケンスデータ B に対する認識率における ANOVA の結果を示す．同様に，手法 (v) に対する結果では，手法 (i)(iii)(iv)(v) に対して有意な差があることが示された．一方で，いずれの実験結果においても，手法 (ii) に対して有意な差は見られなかった．提案手法は N-gram による手法では表現できな

表 4.16　シーケンスデータ B に対する認識率における ANOVA の結果と p 値

	(i)	(ii)	(iii)	(iv)	(v)
(ii)	** (0.0019)	–	–	–	–
(iii)	(0.2492)	** (0.0059)	–	–	–
(iv)	(0.1242)	** (0.0091)	(0.6451)	–	–
(v)	** (0.0023)	(0.3119)	** (0.0023)	** (0.0067)	–
(vi)	(0.5905)	** (0.0059)	(0.2124)	. (0.0999)	** (0.0049)

Significance codes: $0 \leq$ '***' $\leq 0.001 <$ '**' $\leq 0.01 <$ '*' $\leq 0.05 <$ '.' ≤ 0.1

い，今どんな文脈であるかという文脈の認識が行えるという利点を持ちながら，文脈を扱う N-gram による手法と同等の認識性能を持つことが示された．

4.2.2.7　パラメータの検討のための実験

式 (4.15) における α は，HMM による認識尤度と動作出現確率の重みを調整する役割を持つ．そのため，このパラメータを変更することで提案手法のパフォーマンスが向上することが考えられる．式 (4.15) では，双方の最大値を基準に調整を行ったが，その重みが適切か，実際にはどのような値が良いのか，を検証する．例えば，α を調整するような新たなパラメータ β を，式 (4.15) に導入し，α を次のように定義する．

$$\alpha = \frac{\max \log P(m_i|t_j)}{\max \left(\log P(\boldsymbol{o}_\tau|\lambda_i) - C\right)}\beta \tag{4.29}$$

ここで β は，HMM による認識尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ と動作出現確率 $P(m_i|t_j)$ の重みを調節する役割を持つ．この β の値を次のように変化させ，シーケンスデータ B に対する実験を再度行う．

$$\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\} \tag{4.30}$$

これらの β の値を用いて実験を行った結果を表 4.17 に示す．この実験では，$\beta = 1$ の時に最も高い認識率を得ることが出来ていることから，式 (4.15) での

表 4.17　β を 0.01 から 100 まで変化させた際の認識率一覧

β	0.01	0.02	0.05	0.1	0.2	0.5
ratio	0.111000	0.156000	0.274500	0.428000	0.631500	0.741500
β	**1**	2	5	10	20	50
ratio	**0.762000**	0.760000	0.725000	0.709000	0.693000	0.685000
β	100					
ratio	0.681000					

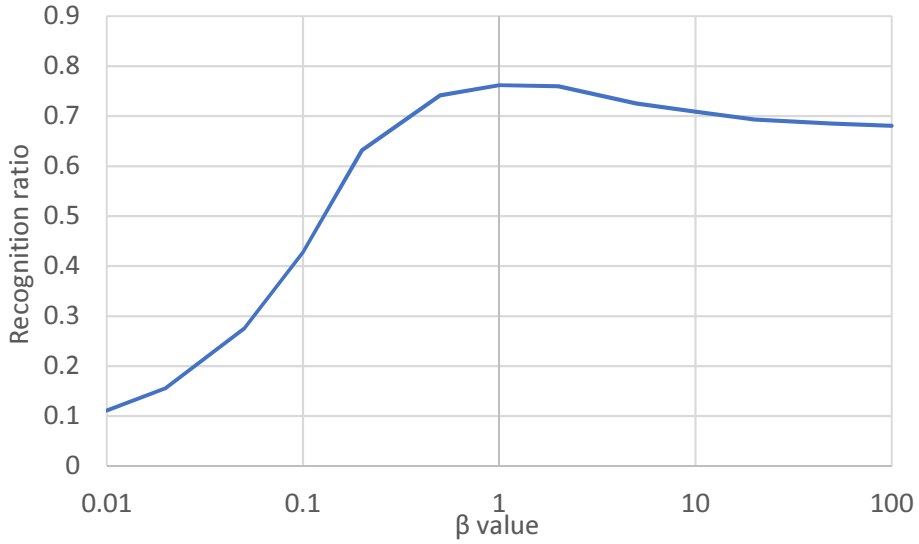


図 4.20　β を 0.01 から 100 まで変化させた際の認識率の推移

パラメータ設定について，まずまずの妥当性があることが示された．認識率の傾向性を確認するため，この結果をグラフ化したものを図 4.20 に示す．縦軸は認識率で横軸は β の変化を片対数グラフで示したものである．この傾向を見ると 0.5 から 2.0 前後に極大値が存在するように見える．そこで，0.5 から 2.0 前後に極大値が存在すると仮定して，0.1 から 2.5 に拡張し次の β の値を用いて実験を行った．

$$\begin{aligned}\beta \in \{&0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0,\\ &1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0,\\ &2.1, 2.2, 2.3, 2.4, 2.5\}\end{aligned} \tag{4.31}$$

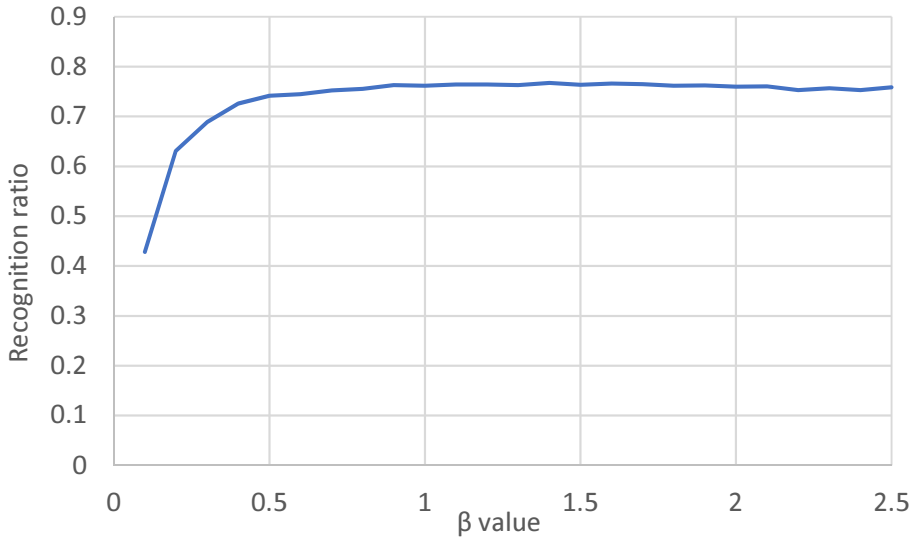


図 4.21　β を 0.1 から 2.5 まで変化させた際の認識率の推移

この β の値に対する認識率を表 4.18 に示す．この結果では $\beta = 1.4$ の時に最も

表 4.18　β を 0.1 から 2.5 まで変化させた際の認識率一覧

β	0.1	0.2	0.3	0.4	0.5
ratio	0.428000	0.631500	0.689000	0.726000	0.741500
β	0.6	0.7	0.8	0.9	1.0
ratio	0.745000	0.752500	0.755500	0.763000	0.762000
β	1.1	1.2	1.3	**1.4**	1.5
ratio	0.764000	0.764500	0.763000	**0.767500**	0.763500
β	1.6	1.7	1.8	1.9	2.0
ratio	0.766000	0.765000	0.762000	0.762500	0.760000
β	2.1	2.2	2.3	2.4	2.5
ratio	0.760500	0.753000	0.757000	0.753000	0.758500

高い認識率を得た．また同様に図 4.21 にこの結果をグラフ化したものを示す．$\beta = 0.8$ 付近以降からほぼ横ばいになっていることが分かる．さらに詳しく見るため，β の値が 0.5 から 2.0 の範囲に限定してより細かいグラフを見る．図 4.22 に β を 0.5 から 2.0 まで変化させた際の認識率の推移のグラフを示す．これまでの β の推移から，単調に極大値が存在すると推測していたものの，実際の認識率

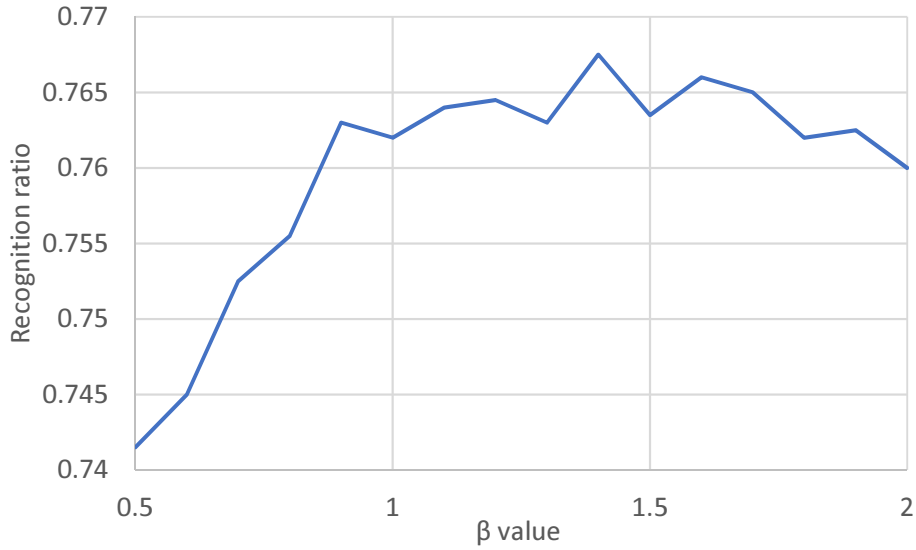


図 4.22　β を 0.5 から 2.0 まで変化させた際の認識率の推移

の推移はやや波打つ結果となっていた．この理由としては次のことが考えられる．HMM の認識や提案手法のパーティクルごとの処理におけるサンプリングは，確率に従って実行されるため，同じ条件で実行しても結果に多少の違いが現れる．そのわずかな違いは，ほとんどの場合において大差は生まれないものの，いくつかの要素（値）が非常に競っている場合，結果に違いが生まれる．その結果図 4.22 のように単調な極大値ではなく，波打つような結果になったと思われる．また，この結果は複数回実行することで，また違った値が現れるのではないかと推測できる．

この β の値を変化させる実験では，実験の結果から β の値を調整することで，提案手法における能力拡張性を示した．

4.3　課題と検討

4.3.1　ループ処理における繰り返し回数について

提案手法は，パーティクルを用いたループ構造による繰り返し処理が実装されている．本論文での実験では，繰り返し回数を $1, 2, 3, 5, 10$ 回と実行している．シーケンスデータ B に対する実験の結果 (図 4.9) では繰り返し回数 10 回の時において最も優れた結果を得たものの，シーケンスデータ A に対する実験の結果 (図 4.8) では繰り返し回数 5 回において最も優れた結果を得た．通常であれば，繰り返し回数を増やすごとにその認識精度が向上すること考えられる．一方で，繰り返し回数を増やすことによって，その瞬間における認識に対する重みが大きくなり，これまでの観測に基づく文脈の影響力が少なくなると考えられる．すなわち，繰り返し回数を増やすごとに，文脈の影響が無くなり HMM の認識性能に大きく依存することになるのではないかという仮説を立てて実験を行った．この実験では，HMM の認識率が 100[%] になるような入力データのみで構成されたシーケンスデータを認識の対象とする．このシーケンスデータに対して提案手法で繰り返し回数を変化させながら認識を行った際の認識率を表 4.19 に示す．この結果では，繰り返し回数を増やすごとに認識率がやや低下している傾向にあ

表 4.19　HMM 単体の認識が間違わない場合のリピート回数ごとの認識率

repeat	1	2	3	5	10	20
ratio	0.953168	0.951791	0.944904	0.946281	0.940771	0.944904

ることが分かる．具体的には，文脈が切り替わる瞬間の認識率が低下していると推測される．その理由は，繰り返すごとに HMM の認識の影響は大きくなるものの，これまでの観測における文脈の分布 $P(t_j)$ の影響も残っており，繰り返すことでパーティクルの分布の形が $P(t_j)$ の影響を少しずつ受けていると推測で

きる．この結果から，HMM の認識率が 100[%] になるような条件下では，提案手法は HMM 単体での認識以上の性能を出すことは難しいことが分かる．一方でそのような条件下は非常に限定的であり，HMM の認識率が 100[%] になる条件下で実施しているのであるから HMM の認識率を超えることが出来ないのは当然であり，実際には先の実験で示したように HMM 単体での認識では，文脈が異なるのにもかかわらず似た動作に誤認識してしまうという弱点を持つ．

この局所的なとある時刻の影響力を強く受ける構図を防ぎながらもこれまでの観測情報を生かす試みとしては，長期記憶の文脈情報を保持する試みが挙げられる．現在の仕組みでは，その文脈情報は 1 時刻前のもののみを用いるようになっている．文脈の分布そのものはそれまでの影響を受けて変化したものであるものの，ループ回数が増加するとともに，その文脈の分布はその瞬間の時刻の影響を強く受けてしまう．これを防ぐために，1 つ前の時刻よりさらに前の時刻の分布を引用したり何らかの変数で保持することで，これらの局所的な影響力だけでなく，これまでの観測情報を生かせるようになると考えている．

4.3.2 文脈切り替わり時の認識率低下について

提案する手法は，これまでの観測に基づいた文脈を用いるため，文脈が切り替わる瞬間の認識率が大幅に低下する．しかしながら，実際の観測では，瞬時にその場面が切り替わることは考えにくい．例えば，観測している人が瞬時に別に人に変わった場合，その瞬間のその人の動作を認識することは難しく，いくつかの観測を経たのちに，思い返すことでその認識を改めることが出来る．この，「思い返してみれば」という処理を付加させることで，提案手法における文脈切り替わり時の認識率低下を防ぐことが出来ると考えている．具体的には，通常の認識では 1 時刻前の文脈を用いるが，これを現在時刻から過去に逆順にたどらせることで，これまでの観測による文脈から過去の認識を改めることが可能になると考

えている．

4.4　まとめ

本章では，似ているが全く異なったカテゴリーの動作同士の誤認識の低減を目的として，動作のカテゴリーを扱う文脈を利用した手法の検討を行った．まず，現在はどんな動作のカテゴリーであるかという情報が与えられるものとして，HMM による動作の認識尤度と動作出現確率の統合処理による認識を実現した．実験では実際の日常動作を対象に，文脈を用いない手法と比較して文脈を用いた手法では，誤認識の低減が出来ることを確認した．さらに，時系列の変化を考慮した文脈と動作の双方向な関係を持つ推定手法を，パーティクルを用いた離散的ループ処理に基づいて，その構成を確立し実装した．比較手法に，HMM のみの手法，文脈を考慮する手法の 1 つである N-grams や，提案する枠組みであっても時系列を考慮しない手法を用意した．時系列なデータに対して，観測から現在の文脈を推定し，さらに認識においてその文脈を用いることで，他の手法と比較して認識率が優れていることを示した．その時系列なデータは，単純な物だけではなく，様々な文脈上で様々な動作が行われる合計 2000 動作を対象にしたものを用意し，認識の対象として実験を行った．また，動作出現確率を得るためのコーパスはどのような物が提案手法にとって有効となるのかを検証する実験も行い，文脈が頻繁に変化しない条件下においては非常に高い性能を持つことを示した．提案手法に仮に用いていたパラメータの数値についても検討を行い，提案手法の能力拡張性についても示唆した．これらの結果から，提案する手法は似ているが全く異なったカテゴリーの動作への誤認識の低減を実現したことを明らかにした．

5

結論

5.1 結論

本論文の目的は，文脈と動作の関係性に着目し，文脈と動作の双方向な情報処理を構成して，身体動作認識における誤認識を低減することであった．

2 章では，文脈を用いて動作を理解し，動作から文脈を推定する，という双方向な情報処理によるループ構造を持つ手法を提案し，その手法の必要性について述べた．

3 章では，文脈と動作の双方向な情報処理が誤認識の低減に有効な手法であることを示す例題の一つ目として，どの領域に着目するべきかという問題を対象とした．具体的な問題として，同一身体上で実行される複数の動作に対する誤認識

を取り扱った．実験結果では，全身を対象とした認識手法と比較して，誤認識を低減し正しく認識が行えることを示した．また，着目すべき身体の領域を正しく推定できることを示した．

4 章では，もう一つの例題として似ている動作への誤認識という問題点に対して，提案手法の有効性を論じた．これまでの観測からどのようなカテゴリの動作を行ってきたかという文脈を扱うことで誤認識の低減を実現した．実験では，これまでの動作の観測から文脈を推定し，現在の認識を改善する結果を示した．さらに，提案手法が文脈の切り替わりが緩やかである場合において，提案手法がより良い結果を示すこと，追加の実験によって明らかにした．

本論文は，文脈と動作が互いに影響するループ構造を持つ手法が，身体動作の認識における誤認識の低減に有効な手法であることを明らかにした．

5.2　提案手法の有効範囲と限界

提案手法では，扱いたい文脈情報に適した動作出現確率を用意する必要がある．文脈ごとの動作出現確率は，所属する文脈ではない文脈に対する動作出現確率がある程度あいまいであっても，誤認識を低減する能力は有していた．しかしながら，この動作出現確率が適切でない場合，HMM 単体の能力以上の性能を発揮することは難しい．また，本論文では文脈の数を 4 または 6 と設定したが，この数が 10，100 と増えていった場合，現在の手法の仕組みでは認識の性能を高めることが難しくなると考えられる．本論文での HMM 単体による認識性能は 6〜7 割程度であった．提案手法は，HMM 単体の認識性能に強く依存する面があり，HMM 単体での認識率が 2〜3 割程度であれば，正しく文脈を推定することが難しくうまく性能を発揮できない．また，HMM 単体での認識率が 10 割かそれに近い値であれば，HMM 単体での認識率よりも提案手法を用いるほうが認識

率が低くなる．HMM 単体による動作の認識は，古典的で基礎的な手法であり，DNN に基づく手法を代表とする近年の強力な分類手法と比較して性能で劣る面を持つ．しかしながら，本論文では単純な認識の性能を競うことを目的としていない．あくまでも，文脈と動作の関係性を構築したことが大事であり，HMM を他の強力な手法に置き換えて議論しても大筋に変化は無い．

3 章では動作認識における身体部位を例として，観測のうちどの情報に着目するべきかということを扱う文脈について，同一身体上で実施される複数動作の認識を目的に，提案手法の実装及び実験を行った．3 章の実験では，各身体部位単体で実行される動作を認識の対象とした．しかしながら動作の種類によっては，両手で行うようなものや，複数の身体部位が連動する動作もある．加えて，重みのように身体部位ごとの重要度というパラメータが必要な場面も考えられる．すなわち，単一の部位のみを扱う部位選択ベクトルではなく，複数部位や重みを含む部位重みベクトルのようなパラメータを使うような拡張が求められる．ただし，部位ごとに重みが異なるような部位選択ベクトルを扱う文脈が人と共有可能なものであるかどうかの議論を行う必要があると考える．

4 章では，4 つの日常動作のカテゴリを例にとり，似た動作を含むようなデータセットに対して，誤認識を低減する手法を扱った．加えて，時系列データを対象に，動作や文脈の時系列の変化に追従する仕組みを実装した．この実装では，動作の順番や文脈の順番を考慮するようなアルゴリズムを導入していない．N-grams を代表とする順番を扱う手法は，料理のレシピのような順番がある程度決まっている動作や，掃除の後には食器を洗うだろうというような文脈の変化の予測に寄与することが考えられる．一方で，掃除の各動作のように順序はさほど重要ではないものの，その文脈内での観測される動作が様々存在するような場面においては，順番を重んじる手法と比較してある程度有効であると考えられる．また，文脈の切り替わりが異なる 2 つのデータを対象とした実験では，文脈

の切り替わりが緩やかである場合において高いパフォーマンスを発揮し，動作出現確率の種類の許容範囲が広くなることを示した．この結果から，提案する手法は，文脈の切り替わりが激しい場合には従来手法と比較して有意な差が無いものの，文脈の切り替わりが緩やかである場合には効果的な認識が行えるといえる．本論文では，「掃除」を行っている人はしばらくの間「掃除」に関わる動作を実行するであろうという仮定を持つため，この仮定の中では良いパフォーマンスを発揮することが確認できた．

本論文では，「観測の全ての情報を用いることによる誤認識」と「よく似た動作に間違えてしまう誤認識」という 2 つの問題点を対象に，それぞれ文脈の持つ役割を変えることで，同じ認識のプロセスでの解決を行った．しかしながら，それぞれ文脈のもつ役割が異なることから，これらの問題を同時に解決するためにはさらなる工夫が必要であり，本論文の今後の課題である．

5.3　今後の課題

5.3.1　2 つの取り組みの統合

3 章の取り組みと 4 章の取り組みでは，それぞれが扱う文脈の定義が異なる．具体的には，「今どこの身体部位に着目するべきか」と「動作のカテゴリー」ということを扱う 2 つの文脈が存在していた．本研究では同じアプローチを用いて別々の課題の解決を実現したものの，これらの問題は同時に観測しうるものであり，同時に解決できることが好ましい．しかしながら，それぞれの取り組みの文脈の定義が異なることから直接応用することは難しい．これを解決する方法としては，2 つの文脈を同時に保持するという手段が挙げられる．3 章と 4 章のどちらの手法も認識結果に基づいて文脈の分布の再計算を行うため，認識結果が

得られる場合は独立して分布を再生成可能である．この段階の構成では，それぞれの文脈は完全に独立していて相互の関係性が無いものとなっている．これらの文脈の分布やその出現確率が互いに影響を及ぼすような仕組みが必要かどうかを検証しながら実装する．

5.3.2 動作以外の情報による文脈の利用

本論文では，動作認識として，人の動作を対象に人の行動を分析している．一方で，人の動作は，環境とのインタラクションや人のいる場所によってその行動がある程度予測が可能である場合がある．例えば，図 2.1 の下部に示すイラストは，身体を動かしていないにもかかわらず，その人が何をしているのかを推測することが容易である．今後の取り組みとしては，例えばその人がどんなものを手に持っているかという道具の情報を用いることが挙げられる．道具がどういうものであるかという情報は，動作のカテゴリーとしての文脈の推測の助けになり，道具がどういう動きをしているかという情報は，その人の動作を認識する助けになる．また，所持している道具がどんなものかわからない場合においても，身体動作の部分を着目したり道具の種類はわからなくても動作に着目したりすることで，あいまいな道具の種類を認識できるようになると考えている．また，その場所がどういう場所かという情報を基に，文脈の分布や動作出現確率を変化させることも有効な手段であると考えている．具体的な例としては，キッチンに立っている人は，おそらく料理をするだろうし，急にスポーツで遊ぶとは考えにくいからである．これらの動作以外の様々な概念は動作の認識において有益な情報であり，動作認識のための文脈を構成する上で考慮すべき情報であると言える．しかしながら，それらの情報の関係性や，観測の獲得は非常に複雑で，すべての関係性を記述することは困難であるため，本論文では最も重要な観測の情報である動作に着眼点を置いていた．一方で，複数の概念を組み合わせる取り組みとして，

Nakamura らの提案する Serket [10] が挙げられる．Serket はベイジアンネットワークによる学習モデルのモジュール化に寄与しており，提案手法における出現確率や文脈の分布は確率モデルの枠組みにおいて親和性が高い．提案手法は，人が理解しやすい文脈や動作を扱うことから，このように他の概念との親和性や拡張可能性が高いことが分かる．具体的には，場所や道具のような場面の因子が提案手法の文脈の分布に影響を与えることで容易な実装ができると考えている．

5.3.3　動作データの収集について

提案手法における最大の課題は，どのようにして動作の出現確率を得るかである．この出現確率は十分な量の動作観測データを対象に LDA を主としたトピックモデルに基づいて生成されることを前提としている．では，その大規模な観測データをどのようにして収集するかということが論点となる．この問題点は，動作や行動認識のフィールドにおいて非常に難しい問題である．これらをデータを得る手段としては，実際の人の動作を観測し，その動作のアノテーションを行うことで収集することが可能となる．その行動データを動作の信号と同時に獲得することは難しい．一方で提案する手法は，その行動データに基づく解析と動作認識の機構が分かれていることに利点がある．すなわち，人の行動の観測データがテキストレベルでも存在すると利用することが可能である．例えば，Virtual Home [45] のような取り組みを利用することで，行動に対するアノテーションのみを効率よく収集することが可能なものであると考える．しかしながら，今後の展開を想定すると，人の動きだけではなく物や環境の観測についてもデータとして保持することが好ましい．その場合はシミュレータの 1 つである SIGVerse [46] の取り組みを利用することで，仮想環境における人の動きとその環境情報の獲得を容易に実現可能である．仮想環境を用いるメリットとしては，様々な実験環境を現実と比較して低コストで実現可能である点がある．また，本

論文の付録に示すように，近年の VR に関するデバイスは急速に発展しており，より効率よく身体動作を収集可能になっていくと考えている．

参考文献

[1] Peter F Brown, Peter V DeSouza, Robert L Mercer, Vincent J Della Pietra JCL. Class-Based n-gram Models of Natural Language. Computational Linguistics. 1992;18(1950):467–479.

[2] Blei DM, Edu BB, Ng AY, et al. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993–1022.

[3] Lee K. Context-independent phonetic hidden markov models for speaker-independent continuous speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1990 April;38(4):599–609.

[4] Matsuda S, Hu X, Shiga Y, et al. Multilingual speech-to-speech translation system: Voicetra. In: 2013 IEEE 14th International Conference on Mobile Data Management; Vol. 2; June; 2013. p. 229–233.

[5] Taniguchi T, Nagasaka S, Nakashima R. Nonparametric Bayesian Double Articulation Analyzer for Direct Language Acquisition From Continuous Speech Signals. 2016;8(3):171–185.

[6] Janus B, Nakamura Y. Unsupervised probabilistic segmentation of mo-

tion data for mimesis modeling. In: ICAR '05. Proceedings., 12th International Conference on Advanced Robotics, 2005.; Vol. 2005. IEEE; 2005. p. 411–417.

[7] Kulić D, Takano W, Nakamura Y. Online segmentation and clustering from continuous observation of whole body motions. IEEE Transactions on Robotics. 2009;25(5):1158–1166.

[8] Lin JFS, Kulic D. Online segmentation of human motion for automated rehabilitation exercise analysis. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2014;22(1):168–180.

[9] Takano W, Nakamura Y. Real-time Unsupervised Segmentation of human whole-body motion and its application to humanoid robot acquisition of motion symbols. Robotics and Autonomous Systems. 2016; 75:260–272.

[10] Nakamura T, Nagai T, Taniguchi T. SERKET: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model. 2017 dec;.

[11] Taniguchi T, Nagasaka S. Double articulation analyzer for unsegmented human motion using Pitman-Yor language model and infinite hidden Markov model. 2011 IEEE/SICE International Symposium on System Integration, SII 2011. 2011;:250–255.

[12] Zhang Z, Hu Y, Chan S, et al. Motion Context : A New Representation for Human Action Recognition. Spinger-Verlag Berlin Heidelberg. 2008; (Mc):817–829.

[13] Ju Sun, Xiao Wu, Shuicheng Yan, et al. Hierarchical spatio-temporal context modeling for action recognition. In: 2009 IEEE Conference on

Computer Vision and Pattern Recognition; jun. IEEE; 2009. p. 2004–2011.

[14] Wu X, Xu D, Duan L, et al. Action recognition using context and appearance distribution features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2011; :489–496.

[15] Goutsu Y, Takano W, Nakamura Y. Classification of Multi-class Daily Human Motion using Discriminative Body Parts and Sentence Descriptions. International Journal of Computer Vision. 2018;126(5):495–514.

[16] Evangelidis G, Singh G, Horaud R. Skeletal quads: Human action recognition using joint quadruples. In: Proceedings - International Conference on Pattern Recognition; 2014. p. 4513–4518.

[17] Ofli F, Chaudhry R, Kurillo G, et al. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. Journal of Visual Communication and Image Representation. 2014 jan;25(1):24–38.

[18] Wei P, Zheng N, Zhao Y, et al. Concurrent action detection with structural prediction. Proceedings of the IEEE International Conference on Computer Vision. 2013;(1):3136–3143.

[19] Cao Z, Simon T, Wei SE, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); jul. IEEE; 2017. p. 1302–1310.

[20] Ohashi T, Ikegami Y, Yamamoto K, et al. Video Motion Capture from the Part Confidence Maps of Multi-Camera Images by Spatiotemporal Filtering Using the Human Skeletal Model; 2018. p. 4226–4231.

[21] Lea C, Flynn MD, Vidal R, et al. Temporal convolutional networks for action segmentation and detection. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017;2017-January:1003–1012.

[22] Varol G, Laptev I, Schmid C. Long-term Temporal Convolutions for Action Recognition To cite this version : Long-term Temporal Convolutions for Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;40(6):1510–1517.

[23] Li Z, Gavrilyuk K, Gavves E, et al. VideoLSTM convolves, attends and flows for action recognition. Computer Vision and Image Understanding. 2018;166(October 2017):41–50.

[24] Hou Y, Li Z, Wang P, et al. Skeleton optical spectra-based action recognition using convolutional neural networks. IeeexploreIeeeOrg. 2018; 28(3):807–811.

[25] Fragkiadaki K, Levine S, Felsen P, et al. Recurrent Network Models for Human Dynamics. 2015 IEEE International Conference on Computer Vision (ICCV). 2015;:4346–4354.

[26] Ordóñez FJ, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors (Switzerland). 2016;16(1).

[27] Wang L, Xiong Y, Wang Z, et al. Temporal Segment Networks for Action Recognition in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;(c):1.

[28] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. Proceedings of the IEEE Computer Society

Conference on Computer Vision and Pattern Recognition. 2015;:1110–1118.

[29] Tani J, Ito M, Sugita Y. Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. Neural Networks. 2004;17(8-9):1273–1289.

[30] Takano W, Imagawa H, Nakamura Y. Prediction of human behaviors in the future through symbolic inference. Proceedings - IEEE International Conference on Robotics and Automation. 2011;:1970–1975.

[31] Takano W, Nakamura Y. Bigram-based natural language model and statistical motion symbol model for scalable language of humanoid robots. Proceedings - IEEE International Conference on Robotics and Automation. 2012;:1232–1237.

[32] Takano W, Imagawa H, Kulić D, et al. What do you expect from a robot that tells your future? The crystal ball. IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings. 2010;:1780–1785.

[33] Tavenard R, Emonet R, Odobez JM. Time-sensitive topic models for action recognition in videos. 2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings. 2013;:2988–2992.

[34] Wang YWY, Mori G. Human Action Recognition by Semilatent Topic Models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009;31(10):1762–1774.

[35] Wang T, Liu C. Human Action Recognition Using Supervised pLSA. Int J Signal Process Image Process Pattern Recognit. 2013;6(4):403–414.

[36] Huynh T, Fritz M, Schiele B. Discovery of activity patterns using topic

models. Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp '08). 2008;:10–19.

[37] Bando T, Takenaka K, Nagasaka S, et al. Generating contextual description from driving behavioral data. IEEE Intelligent Vehicles Symposium, Proceedings. 2014;(Iv):183–189.

[38] Bando T, Takenaka K, Nagasaka S, et al. Unsupervised drive topic finding from driving behavioral data. IEEE Intelligent Vehicles Symposium, Proceedings. 2013;(Iv):177–182.

[39] Bargi A, Da Xu RY, Piccardi M. An online HDP-HMM for joint action segmentation and classification in motion capture data. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012;:1–7.

[40] Malgireddy MR, Nwogu I, Govindaraju V. Language-Motivated Approaches to Action Recognition. In: Journal of machine learning research. Vol. 14; 2017. p. 155–181.

[41] Attamimi M, Fadlil M, Abe K, et al. Integration of various concepts and grounding of word meanings using multi-layered multimodal LDA for sentence generation. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems; sep. IEEE; 2014. p. 2194–2201.

[42] Demiris Y, Khadhouri B. Hierarchical attentive multiple models for execution and recognition of actions. Robotics and Autonomous Systems. 2006;54(5):361–369.

[43] Baker CL, Saxe R, Tenenbaum JB. Action understanding as inverse planning. Cognition. 2009;113(3):329–349. Available from: http://dx.doi.org/10.1016/j.cognition.2009.07.005.

[44] Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning - ICML '06; New York, New York, USA. ACM Press; 2006. p. 113–120.

[45] Puig X, Ra K, Boben M, et al. VirtualHome : Simulating Household Activities via Programs. Cvpr 2018. 2018;.

[46] Mizuchi Y, Inamura T. Cloud-based multimodal human-robot interaction simulator utilizing ROS and unity frameworks. In: 2017 IEEE/SICE International Symposium on System Integration (SII); dec. IEEE; 2017. p. 948–955.

謝辞

数多くの方々の支援を受け，本論文を書き上げることが出来ました．指導教員の稲邑哲也先生には，忙しい中や夜遅くにおいても快く議論に応じてくださり，何度も何度も助けられました．また，私生活においても，体調を崩しやすく気落ちしやすい私ですが，励ましの言葉をかけてくださり非常に心が救われたことが数多く思い出され，感謝の思いでいっぱいです．公私ともに支えになりました稲邑哲也先生に，重ねて御礼申し上げます．予備公聴会および本公聴会で外部審査員として引き受けてくださいました早稲田大学の尾形哲也先生に感謝申し上げます．尾形先生は JST の ACT-I への提出の際にも親身になってアドバイスをくださり，書類作成や面接の際の励みになりました．重ねて御礼申し上げます．また，予備公聴会と本公聴会だけでなく，2 度の中間発表においても審査委員を快く引き受けてくださいました，山田誠二先生，佐藤健先生，市瀬龍太郎先生，山岸順一先生に感謝申し上げます．中間発表でのディスカッションはその後の研究の方針や新しい実験を決定する上で非常に参考になりました．また評価シートでの激励の言葉は研究を進める上での励みになりました．重ねて御礼申し上げます．

私が RA として所属させていただいた CREST のプロジェクトでは，数多くの会議に参加させていただき，研究者としてよい経験や知識の蓄積を行うことが出来ました．特に大阪大学の長井隆行先生を中心としたプロジェクトチームでは，合宿形式のハッカソンなどの研究室間の垣根を超えた濃密な時間を過ごすことが出来，他の研究室のノウハウに触れることが出来る非常に良い機会を頂きました．

研究室の特任研究員である，坂戸達陽さん，水地良明さん，郷津優介さんには，研究ミーティングをはじめとし，研究室内でのディスカッションも含め，多くの支援を受けました．日々の生活においても友好的に接してくださり，御礼申し上げます．また特任技術専門員の山田裕基さんには実装における相談や機器の取り扱い等，多くのことでお世話になりました．御礼申し上げます．

著者文献及び発表目録

論文誌への掲載

Tadashi Ogura, Tetsunari Inamura. Bidirectional estimation between context and motion in motion sequence in which context changes. Journal of Advanced Robotics, Vol. 33, Iss. 11, pp. 550-565, 2019.

国際会議（ポスター・査読あり）

Tadashi Ogura, Tatsuya Sakato, Tetsunari Inamura, Human Motion Recognition Based on Topic Model, The 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016), WeCI2.23, 2016.

Tatsuya Sakato, **Tadashi Ogura**, Tetsunari Inamura, Human Motion Recognition Based on Dynamic Topic Model, The 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016),

WeCI2.24, 2016.

国内会議（口頭発表・査読なし）

小椋 忠志，坂戸 達陽，稲邑 哲也，トピックモデルを考慮した身体動作認識，第 34 回日本ロボット学会学術講演会，2Z2-02，2016.

坂戸 達陽，小椋 忠志，稲邑 哲也，ダイナミックトピックモデルを利用した身体動作系列認識，第 34 回日本ロボット学会学術講演会，2Z2-03，2016.

横田 栞，水地 良明，小椋 忠志，崔 龍雲，稲邑 哲也，人の身体動作観察とレシピ情報に基づくロボットの調理動作の生成への取り組み，第 34 回日本ロボット学会学術講演会，3G2-03，2016.

小椋 忠志，稲邑 哲也，トピックモデルを用いた認識対象の選択制御とその動作認識への応用，2017 年度 人工知能学会全国大会（第 31 回），4D1-OS-37c-3，2017.

小椋 忠志，稲邑 哲也，同一身体上の複数動作認識に向けたトピックモデルに基づく動作と部位着目の仮説相互推定，第 35 回日本ロボット学会学術講演会，3I3-02，2017.

小椋 忠志，稲邑 哲也，長時間動作の文脈と身体動作の相互認識，2018 年度 人工知能学会全国大会（第 32 回），1O1-03，2018.

小椋 忠志，稲邑 哲也，動作と文脈の双方向な認識手法における身体動作認識性能評価，2019 年度 人工知能学会全国大会（第 33 回），2M5-J-10-03，2019.

国内会議（ポスター・査読なし）

小椋 忠志，坂戸 達陽，稲邑 哲也，トピックモデルを考慮した身体動作認識，計測自動制御学会 システム・情報部門 学術講演会 2016，GS13-9，2016.

坂戸 達陽，小椋 忠志，稲邑 哲也，Dynamic Topic Model を利用した身体動作系列のトピック推定，計測自動制御学会 システム・情報部門 学術講演会 2016，GS13-10，2016.

横田 栞，水地 良明，小椋 忠志，崔 龍雲，稲邑 哲也，ロボットの調理動作生成のためのレシピ情報に基づく人の身体動作パターンの収集，計測自動制御学会 システム・情報部門 学術講演会 2016，GS13-14，2016.