The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

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

Evaluation of bidirectional method between motion and context for human motion recognition

小椋 忠志 *1　稲邑 哲也 *1*2
Tadashi Ogura　Tetsunari Inamura

*1総合研究大学院大学
SOKENDAI(The Graduate University for Advanced Studies)

*2国立情報学研究所
National Institute of Informatics

It is generally difficult to recognize and distinguish human motion which are different kinds of motions but whose motion patterns are similar to each other. Conventionally, we have proposed a method of alternately performing two processes, motion recognition using context information and re-estimation of context based on recognition results. However, the performance evaluation of reasonable repetition times in these two recognition processes was not discussed. In addition, only an unrealistic ideal distribution is used as the motion appearance probability, and utility in the probability distribution including noise has not been discussed. In this paper, we aim to investigate these problems and clarify the conditions for performance improvement. Through experiments, a high motion recognition ratio was obtained when the number of iterations was 5 and 10. Furthermore, we confirmed that the proposed method maintains a high motion recognition ratio even if using noisy motion appearance probability. From these results, we have concluded that the proposed method has utility for practical motion patterns.

1. はじめに

「バイバイする」という動作と「窓を拭く」という動作は，身体動作のみに着目すると，どちらも手を振るという動作でとてもよく似ている．動作のみに着目した認識手法ではこのようなよく似た動作を区別することが難しく誤認識の要因となる．実際に人が人を観測している際には，それまでの観測情報から文脈を考慮出来るため，これらの動作を誤認識することは考えにくい．例えば,「窓を拭く」という動作の観測の前に,「机を拭く」「床を掃く」という動作を観測できたのであれば，観測者は「掃除」をしているという観測対象者の動作のカテゴリを推測することが出来る．本稿では，この動作のカテゴリを文脈と呼び，文脈を活用した動作認識手法について述べる．

生活支援ロボットのような人とのコミュニケーションが求められる場面では，人とロボットで共有可能な概念となる文脈が求められる．Recurrent Neural Network(RNN) に基づく動作認識手法 [Du 15] は，コンテクスト層を所持しているものの，その内部状態の可読性が低くコミュニケーションのための文脈情報の抽出は難しい．Attamimi らは，動作を含む複数のモダリティを用いたカテゴリ分類手法を提案している [Attamimi 14]．提案されているカテゴリ分類は人とのコミュニケーションに適したものであるが，カテゴリや文脈そのものの時間的関係性は考慮されていない．そのため，動作認識に用いるための文脈は，その時々に応じて適切に観測情報から推定することが求められる．

本研究は，似ている動作への誤認識を低減することを目的として次のような手法を提案する．冒頭の「掃除」のような可読性のある文脈情報と，パターン認識手法を組み合わせて動作を認識する．また，動作の認識結果に基づいて，現在の文脈情報を更新する．そして，この 2 つのプロセスが双方向に繰り返し実行される．この動作と文脈の双方向な認識手法によって，時系列な観測データに対してその時々に応じた文脈情報を活用し，似た動作への誤認識低減を実現する．

連絡先: 小椋忠志，総合研究大学院大学，東京都千代田区一ツ橋 2-1-2，t-ogura@nii.ac.jp

本研究におけるこれまでの研究成果 [Ogura 18] では，双方向な認識における繰り返し処理において，最適な繰り返し回数など詳細なアルゴリズムの性能評価について議論がされていなかった．また，文脈情報として動作の観測を予測する動作出現確率 (Motion Appearance Probability Distribution; 以下 MAPD) について，ノイズの少ない理想的な分布のみを用いており，ノイズを含む確率分布における有用性についても議論されていなかった．そこで本稿では，これらの点を調査し，性能向上のための条件を明らかにすることを目的とする．適切な繰り返し回数は何回であるか，よりノイズの多い MAPD を用いた場合でも認識性能を保つことができるか，という 2 点について実験を通して評価する．

2. 文脈と動作の双方向推定手法の概要

提案する手法は大きく分けて 2 つの処理によって構成される．一つ目は文脈を用いた動作認識で，二つ目は動作認識結果に基づく文脈の推定である．手法のより詳しい説明は文献 [Ogura 18] に示されているため，ここでは手法の概要について簡単に述べる．

2.1 文脈を用いた動作認識

提案手法の処理手順の概要を図 1に示す．文脈情報は日常生活における動作のカテゴリを対象に，図 1右上のように確率分布として扱われる．文脈の確率分布は図 1中央のように離散的に取り扱われる．時刻 τ の時の動作パターン $\boldsymbol{o}_\tau$ が観測されたとき，HMM によって認識尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ が求められる．ここで，HMM である λ_i は動作 m_i に対応する．離散の粒 k が所属する文脈 j を q_k とすると，離散の粒 k における認識処理の際に用いられる文脈情報としての MAPD は $P(m_i|j = q_k)$ である．HMM による認識尤度 $P(\boldsymbol{o}_\tau|\lambda_i)$ と現在の文脈に基づく MAPD$P(m_i|j = q_k)$ を組み合わせる処理 [Ogura 18] によって，文脈を用いた動作認識を実現する．

2.2 動作認識結果に基づく文脈の推定

動作の認識結果は，2.1節の離散的処理によって，図 1の下部のように離散分布として得られる．離散の粒 k の認識結果

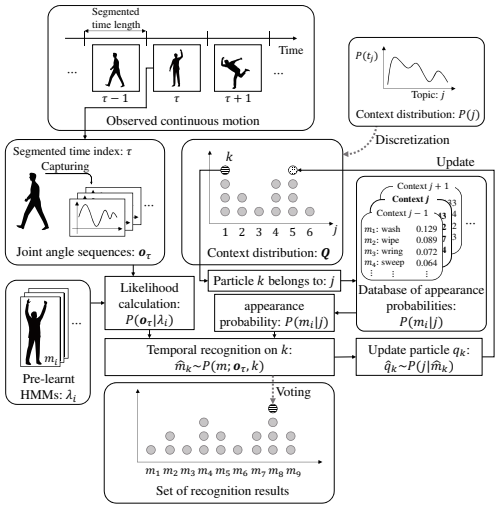


図 1: 文脈情報を用いた動作認識と文脈分布の更新手順

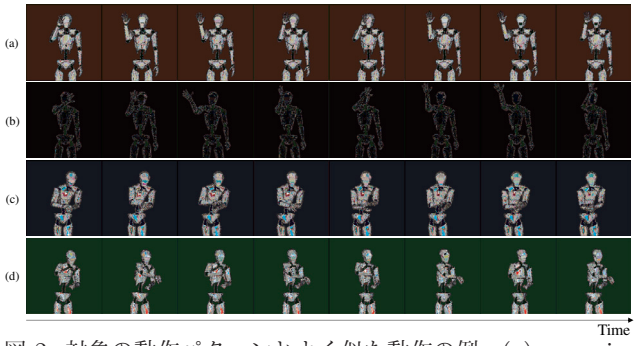


図 2: 対象の動作パターンとよく似た動作の例．(a) m_6: wiping window．(b) m_{33}: waving hand．(c) m_8: washing．(d) m_{13}: frying with pan．(a) と (b) の動作，(c) と (d) の動作はそれぞれよく似ており，誤認識を引き起こしやすい．

を $\hat{m}_k$ とする．k が次に所属する文脈 j を $\hat{q}_k$ とすると，$\hat{q}_k$ を得るための確率分布 $P(j|\hat{m}_k)$ を，ベイズの定理に基づいて次のように求める．

$$P(j|\hat{m}_k) = \frac{P(\hat{m}_k|j)P(j)}{P(\hat{m}_k)} \tag{1}$$

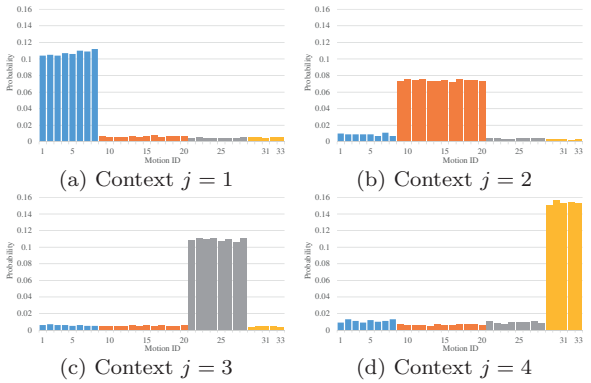
ここで，$P(\hat{m}_k|j)$ は MAPD$P(m = \hat{m}_k|j)$ によって得られ，同様に $P(\hat{m}_k)$ は MAPD を参照し j についての周辺化によって得られる．また本稿では，$P(j)$ には図 1の左上部に示されるような，一時刻前の時刻 $\tau - 1$ の時の文脈の確率分布を用いる．

2.1節における文脈を用いた動作認識と本節の動作認識結果に基づく文脈の推定は，1 つの時刻において複数回繰り返し実行される．

3. 評価実験

3.1 対象のデータ

ここでは 2 つの実験に共通する対象のデータについて述べる．4 種類の文脈に所属する合計 33 種類の日常動作を用意した [Ogura 18]．図 2に実際にキャプチャした動作パターンを示す．対象の動作の中では図 2(a) と図 2(b) の様によく似た身体動作を含んでいる．観測の身体動作が対象とする関節角は，



(a) Context $j = 1$　(b) Context $j = 2$

(c) Context $j = 3$　(d) Context $j = 4$

図 3: 実験に用いる文脈 j ごとの動作出現確率 $P(m_i|j)$

頭・両肩・両手首の各 3 次元と，肘・全指の各 1 次元の，合計 27 次元である．各動作は 60[Hz] でキャプチャされ，約 4 秒間の長さとする．Left-to-right モデルの HMM を用い，状態数は 16，GMM の混合数は 5 で学習を行った．

図 3に本実験に用いる MAPD を示す．今回の実験では，$P(m_i|j)$ の値は手動で与えた．図 3を見て分かるように，ある文脈はそのカテゴリに属する動作を出力する確率が高く，異なるカテゴリに属する動作はほとんど出力しない．この実験では，文脈の数 J は 4，離散の粒の数 K は 400 で，文脈の確率分布は一様分布を初期状態とした．

3.2 比較手法

この実験では提案手法を含む 4 つの手法を用いて実施した．

(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．

$$i_{hmm} = \arg\max_i P(\boldsymbol{o}_\tau|\lambda_i) \tag{2}$$

(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認識に用いる．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．一方で，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，HMM による尤度と次の動作の確率値を統合するスコア $S_{ngram}(\boldsymbol{o}_\tau, i)$ を次式のように定義する．

$$\begin{aligned} S_{ngram}(\boldsymbol{o}_\tau, i) = \exp\{&\alpha\left(\log P(\boldsymbol{o}_\tau|\lambda_i) - C\right) \\ &+ \log P(m_i^\tau|\hat{m_i}^{\tau-1}, \cdots, \hat{m_i}^{\tau-N+1})\} \end{aligned} \tag{3}$$

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 の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．

$$i_{ngram} = \arg\max_i S_{ngram}(\boldsymbol{o}_\tau, i) \tag{4}$$

表 1: シーケンスデータ A に対する各手法による認識率

	Recognition Ratio				
(i)	0.66116				
			2-gram	3-gram	4-gram
(ii)			0.70110	0.63499	0.55647
	Recognition Ratio on Repeat Cycles				
	once	2 cycles	3 cycles	5 cycles	10 cycles
(iii)	0.66942	0.66942	0.66529	0.67493	0.66116
(iv)	0.61846	0.69697	0.70661	**0.72727**	0.71488

(iii) Not sequential
この手法では，提案する手法のうち，文脈の確率分布を毎時刻一様分布にリセットする．とある時刻の入力情報に対して，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．

(iv) Our method
提案する手法である．

3.3 2 種類のシーケンスデータに対する実験

3.3.1 2 種類のシーケンスデータについて

この実験では，2 種類のシーケンスデータを認識の対象とする．ここで，シーケンスデータとは複数の動作を繋ぎ合わせた連続動作のことを示す．1 つのシーケンスデータの中では動作だけでなく文脈も次々に変化している．具体的には次の 2 種類である．

シーケンスデータ A
シーケンスデータ A における 1 系列のデータは，動作ラベル m_1 から m_{33} まで順番に動作を繋げたものである．このシーケンスデータを 22 種類用意し，認識の対象に用いる．シーケンスデータ A は，22 種類のデータにおいて動作の順番が固定されている条件となっている．シーケンスデータ A の目的は，22 種類のデータに対する認識結果を比較し，文脈の分布の変化の追従性の評価を行う点にある．またこのシーケンスデータは，1 系列が 33 動作なのに対して 4 種類の文脈が次々に変化していくため，文脈の切り替わりが早いという特徴がある．

シーケンスデータ B
このシーケンスデータは動作の順番が固定されていないデータであり，1 つの系列に約 200 動作が含まれている．1 つの系列の中では 9 回文脈が切り替わり，文脈に従った動作が順番に実行される．この実験では合計 10 種類のシーケンスデータが用意されている．シーケンスデータ B は，動作の順番が固定でなく，そして文脈の切り替わりが比較的緩やかであり，先のシーケンスデータ A と比べてより実践的なテストデータであることを示している．

この 2 種類のシーケンスデータを対象として，手法ごとに動作の認識を行った．

3.3.2 シーケンスデータ A に対する実験結果

すべての手法による認識率を表 1に示す．(iii) と (iv) の手法は，上下層において双方向の認識処理を持つ認識手法であるため，双方向の認識処理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，1，2，3，5，10 回と変更して実施した．また，N-gram に基づく手法 (ii) では，2-gram，3-gram，4-gram の 3 種類を対象とした．表 1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 回での手法 (iv) である．22 種類のシーケンスデータの平均の認識率の比較では，提案手法が他の手法よりも優れていることが示された．

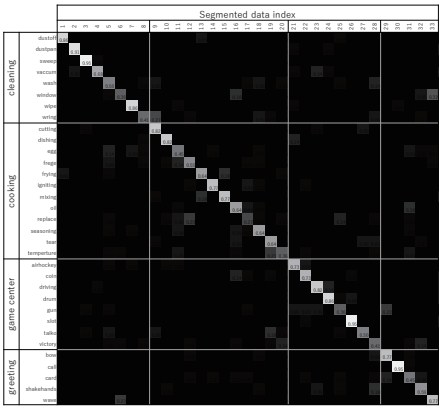


図 4: 手法 (i) による認識結果の混同行列

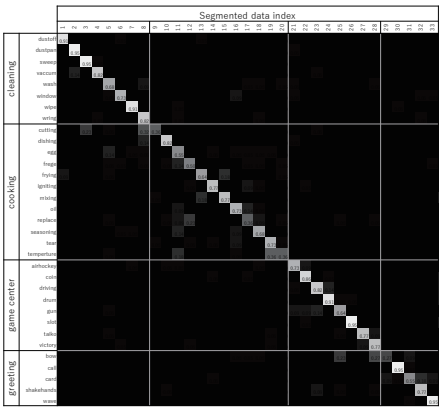


図 5: 手法 (iv) による認識結果の混同行列

図 4に手法 (i) によ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のインデックスを示す．この混同行列は 22 種類のシーケンスデータを対象にした認識結果を，割合と濃度で示したものである．用意されたシーケンスデータは動作 m_1 から m_{33} まで順番に実施されているため，正しい認識結果を示した場合，図の左上から右下まで対角線上に白くなる．図 4の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 2の (a) と (b) を見てもわかるようにこれらの動作はよく似ている．

図 5には，ループ構造の繰り返し回数 5 回の手法 (iv) による認識の混同行列を示す．手法 (i) の誤認識が確認された動作 m_6 と動作 m_{33} に注目すると誤認識は減少している．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が向上したからである．

22 種類のシーケンスデータに対する文脈の分布の平均と標準偏差を図 6に示す．横軸は分割された動作のインデックスで，縦軸は離散の粒の数である．また，網掛けの色は対象の動作が所属する文脈を示し，線の色と網掛けの色は対応している．平均の値をみると，おおよそ正しい文脈の推定が行われていることがわかる．一方で，動作のインデックス 9 番を見ると，文脈 $j = 2$ の動作であるのに対し文脈 $j = 1$ の離散の粒の数のほうが多くなっている．この原因は，これまでの観測情報から文脈 $j = 1$ の方が確率が高いためである．そのため，図 5の文脈が切り替わる直後に認識率が低下している．

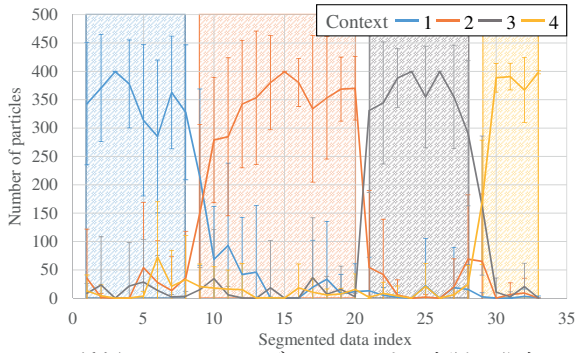
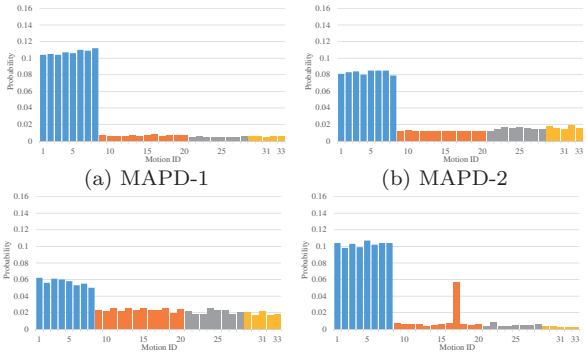


図 6: 22 種類のシーケンスデータにおける文脈の分布の平均と標準偏差の推移



(a) MAPD-1 (b) MAPD-2

(c) MAPD-3 (d) MAPD-4

図 7: 文脈 $j = 1$ の時における動作出現確率 MAPD-1，MAPD-2，MAPD-3，MAPD-4

3.3.3 シーケンスデータ B に対する実験結果

次に，シーケンスデータ B に対する認識結果を表 2に示す．手法 (i) によ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．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．

2 種類のシーケンスデータを対象にした実験では，A，B それぞれ繰り返し回数 5，10 回の時に最も高い認識率を得た．繰り返し回数 5 回と 10 回での認識率の違いは大きくはないが，この繰り返し回数の差は，文脈が切り替わる頻度に関係性があると考察する．

3.4 動作出現確率の検討のための実験

図 3の MAPD は，対象の文脈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，MAPD にノイズが多い場合の提案手法のパフォーマンスを評価する．

3.3節での 2 種類のシーケンスデータに対する実験に用いた MAPD を MAPD-1 と呼び，新たに MAPD-2，MAPD-3，MAPD-4 を追加した．4 つの MAPD のうち文脈 $j = 1$ の MAPD を比較したものを，図 7に示す．MAPD-2 と MAPD-3 を見てわかるようにこれらは MAPD-1 の出現確率と比較して，文脈に属する動作の確率が低く，異なる文脈に属する動作の確率が高くなっている．MAPD-4 による文脈 $j = 1$ の MAPD を図 7(d) に示す．MAPD-4 は，MAPD-1 に対して他の文脈の動作の確率が一部高くなってしまっているような，分布である．

表 3に，手法 (iv) による，各 MAPD に対する認識率を示す．いずれの結果においても，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，MAPD の変化に伴い，やや低くなっている．一方で，実践的なシーケンスデータに対する認識率は，すべての MAPD において，高い認識率を保っている．

表 2: シーケンスデータ B に対する各手法による認識率

	Recognition Ratio				
(i)	0.67350				
		2-gram	3-gram	4-gram	
(ii)		0.75300	0.76100	0.65150	
	Recognition Ratio on Repeat Cycles				
	once	2 cycles	3 cycles	5 cycles	10 cycles
(iv)	0.63450	0.76550	0.75850	0.75800	**0.76600**

表 3: 手法 (iv) による MAPD を変化させた際の認識率

	MAPD-1	MAPD-2	MAPD-3	MAPD-4
Sequence A	0.72727	0.67906	0.68182	0.67080
Sequence B	0.75800	0.75250	0.72350	0.77100

4. おわりに

本稿では，提案する文脈と動作の双方向な認識手法を評価するため，2 つの実験を行った．2 種類のシーケンスデータを対象にした実験では，急激な文脈を変化を含むデータを対象にした際には繰り返し回数 5 回，文脈の変化が緩やかなデータの際には繰り返し回数 10 回の時に，最も高い認識率を得た．また，同条件で実施した他の手法と比較して，提案する手法が優れた手法であることを示した．MAPD の検討のための実験では，よりノイズの多い MAPD をを用いた場合においても，高い認識率を保つことを示した．これらの結果から，提案手法は理想的な条件ではない，実践的な動作パターンを対象とした場合でも，有用性が確保されるという結論を得た．

本稿では，文脈と動作の双方向な認識手法の有効性を示した対象は動作パターンのみだったが，この考え方は他の認識対象においても十分適用可能であると考える．例えば，動作の文脈を用いることで，観測対象者が持つ道具の認識性能を高めることが可能である．また，所持する道具情報を用いることで，動作の認識を見直すことが可能となる．今後は，このような動作パターン以外への拡張に取り組む．

謝辞

本研究は，JST，CREST（グラント番号 JPMJCR15E3）の支援を受けたものである．

参考文献

[Du 15] Du, Y., Wang, W., and Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.1110–1118, (2015).

[Attamimi 14] Attamimi, M., Fadlil, M., Abe, K., Nakamura, T., Funakoshi, K., and Nagai, T.: Integration of various concepts and grounding of word meanings using multi-layered multimodal LDA for sentence generation, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2194–2201, (2014).

[Ogura 18] 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).