3E2-1

連続インタラクションデータからのインタラクションコンテキストの変化点の検出

The Detection of the Change Point of Interaction Context from Continuous Interaction Data

坂本佳愛[*1] 岡田将吾[*2] 西田豊明[*2]
Kae Sakamoto Shogo Okada Toyoaki Nishida

[*1]京都大学工学部情報学科 Kyoto University, Faculty of Engineering
[*2]京都大学大学院情報学研究科知能情報学専攻 Kyoto University, Graduate School of Informatics

In this paper, we propose a system that acquires the pairs of frequent occurrence patterns of the gesture and action from the time series data obtained by observing the interaction between multiple users and the robot in the filming task by the time series data mining and the template match. We also detect the multi-modal event, which is the change point of the interaction status, in order to understand the interaction context by the combination of multi-modal patterns.

The experiment shows that it is able to obtain the pair of the patterns of the gesture and action by the proposed system, also in the filming task in which complex and various non-verbal information is expressed, and to detect the change point of the interaction status by the multi-modal events.

As future tasks, the robot will generate the action controller with information including the past interaction, verbal information and present environment.

1. はじめに

複数人のユーザと自然なインタラクション可能なロボットを実現するためにはユーザの表出する非言語情報を認識し、それに応じて適切に行動を生成する機能が必要である [1]。本論文ではマルチモーダルを統合したパターンをインタラクションコンテキストとし、複数のモダリティデータで発生するイベントの重なりからロボットが推定するべきインタラクション状態の変化点をマルチモーダルイベントと呼ぶ。ここでインタラクションコンテキストとは、インタラクションパターンを規定する文脈情報のことであるこのマルチモーダルイベントを検出し、それに応じてユーザとのインタラクションの状態を推定した上でロボットが適切な動作を生成することで、非言語情報を用いたユーザとロボットとの自然なインタラクションの実現を目指す。

そこで、本研究では複数人のインタラクション行為から得られるユーザの表出するマルチモーダルな非言語情報とロボットの動作系列データから教師なし学習と教師あり学習の両方を用いて頻出するパターンやイベント、マルチモーダルイベントの検出を行い、インタラクションの状態を推定した上で、それに応じてロボットの動作を決定するシステムを提案する。

2. 関連研究と本研究の貢献

Mohammad ら [3] は、ユーザのジェスチャとロボットの動作の対を教師なし学習で獲得するシステムを提案した。Learner ロボットが Operator のジェスチャと Actor ロボットの動作を観察し、得られた時系列データから頻出パターンを検出し、ジェスチャと動作の対を獲得する。

頻出パターンを検出するアルゴリズムには Robust Singular Spectrum Transform （RSST）アルゴリズム [6]、Distance Graph Constrained Motif Discovery（DGCMD）アルゴリズム [5]、Granger-causality が用いられている。本研究では、頻出パターンを検出する基本アルゴリズムに RSST アルゴリズムと DGCMD アルゴリズムを用いる。

本研究の貢献としては、[3] では移動系ジェスチャのみを用いた迷路タスクだったのに対し、撮影タスクという、迷路タスクより複雑なタスクの中で起こる多種のジェスチャインタラクションを対象とした点や、[3] ではユーザ１人とロボットとの一対一のインタラクションであったのに対し、複数人のユーザとロボットとのインタラクションを対象とし、複数のモダリティを統合したイベントの変化も検出可能となり、インタラクション状態を推定した上でユーザのジェスチャを認識してロボットが動作をするという、自然なインタラクションを実現するための階層的システムを提案している点が挙げられる。

また、システムの改良点として、[3] では DGCMD アルゴリズムにおいて検出された頻出パターンの適切なクラスタリングが行われていなかったため、検出されたパターンに再度、階層的クラスタリングを行うことでより高精度に検出したパターンを分類することが可能となった。

3. システムの全体像

3.1 システム概要

最初に、ユーザのジェスチャデータとロボットの動作データからそれぞれ頻出パターンの検出を行う。ポインティングや視線といった、人によってその非言語情報の持つ意味が変わらないものの頻出パターン検出は、教師データをテンプレートとして用いる、テンプレートマッチングを行う。それ以外のジェスチャの検出については、教師なし学習を用いて行う。この教師あり学習と教師なし学習を統合したシステムにより複数のモダリティから頻出パターンを得る。

タスクの観測から得られた多次元時系列データを入力とし、RSST アルゴリズムで時系列データの変化点を検出する。その検出された変化点を元に、DGCMD アルゴリズムで頻出パターンを抽出する。[3] との相違点は、本システムでは頻出パターンを抽出した後、Dynamic Time Warping（DTW）による階層的クラスタリングを行い、再度頻出パターンのクラスタリングを行うことで、より適切なパターンのクラスタリングが可能となっている。

連絡先: 坂本佳愛 sakamoto@ii.ist.i.kyoto-u.ac.jp

ジェスチャと動作の頻出パターンが検出されたら、検出された時刻の近いものどうしを関連付ける。こうして確率ネットワークを生成する。

さらに、パターンの組み合わせからマルチモーダルイベントを検知し、インタラクションの状態の変化点を検出する。そこから今現在のインタラクションの状態を推定することができる。こうして、より自然なインタラクションを実現するための、マルチモーダルデータを用いたロボットの動作生成コントローラを獲得する。システムのフローを図 1 に示す。

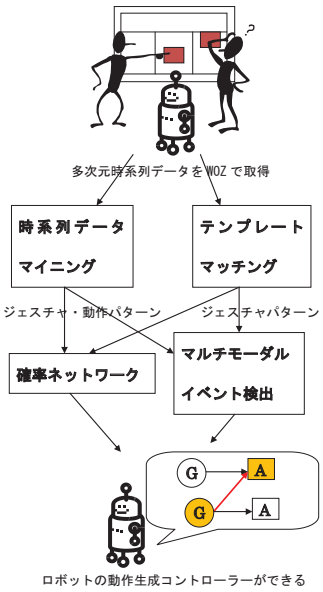


図 1: システムフロー

3.2 タスク

センサールーム内にポスターを設置し、その前で被験者 2 人とロボット 1 体との撮影タスクを行う。ロボットの位置は固定し、操作して動かすのはカメラ部分のみである。今回は被験者の向きや視線、ジェスチャといったモダリティに対して動作するロボットシステムの構築を目指すため、被験者 2 人はポスターの前をほとんど動かないようにして、もう 1 人の被験者と話したり、ロボットに撮影して欲しいポスター上の場所を自由なジェスチャで指示したりする。2 人が同時にロボットに指示することはない。ロボットは WOZ で操作され、被験者のジェスチャに対して正解の動作を行う。図 2 に撮影タスクの様子とモーションキャプチャセンサー取り付け位置を示す。[3] で行われた迷路タスクと比較すると、ジェスチャが連続的に現れないため、頻出パターンの検出が困難になる可能性がある。また、より複雑で多種のジェスチャが現れるタスクとなっている。

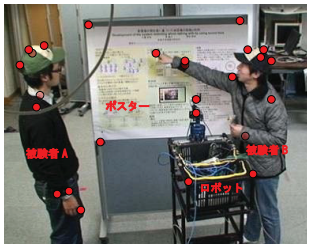


図 2: 撮影タスク

4. システムに用いたアルゴリズム

4.1 頻出パターン発見の基本アルゴリズム

頻出パターンの発見には、Robust Singular Spectrum Transform（RSST）アルゴリズム [6] と Distance Graph Constrained Motif Discovery（DGCMD）アルゴリズム [5] を用いた。

RSST は連続データの変化点を検出するアルゴリズムである。ある時刻 T の前後の変化量を計算する場合、t より前の観測系列データからハンケル行列 $H(t)$、後の観測系列データからハンケル行列 $G(t)$ を形成する。

$$H(t) = [seq(t-n), ..., seq(t-1)] \tag{1}$$

$$G(t) = [seq(t+1), ..., seq(t+n)] \tag{2}$$

$H(t)$ について特異値分解を行い、最適な l 個の特異ベクトルを求める。ここでは l は累積寄与率に基づいて算出される。

$$H(t) = U(t)S(t)V(t)^T \tag{3}$$

ここで $S(i-1,i-1) \leq S(i,i) \leq S(i+1,i+1)$ である。これは t より前の観測系列データ群の部分空間を求めていることに相当する。同様に $G(t)$ の最適な l 個の固有ベクトルを見つける。

$$G(t)G(t)^T u = \mu u \tag{4}$$

$$\beta_i = u_i, i \leq l_f and \lambda_{j-1} \leq \lambda_j \leq \lambda_{j+1} for 1 \leq j \leq w \tag{5}$$

求められたベクトルと部分空間を線形合成し、ノイズを減らすためのフィルタリングを行った後、変化度を調べる。こうして連続時系列データの変化点データを得る。

$$\alpha_i(t) = \frac{U_l U_l^T \beta_i(t)}{||U_l U_l^T \beta_i(t)||}, i \leq l_f \tag{6}$$

$$cs_i(t) = 1 - \alpha_i(t)^T \beta_i(t) \tag{7}$$

DGCMD は RSST で得られた変化点のデータを制約データとして、頻出パターンを検出する。時系列データ上で、制約データの極大値の前の部分のパターンを最短パターン長 l_{min} で切り出す。全てのパターン同士の距離行列を DTW を用いて作成し、閾値より小さいものだけを残して距離グラフを作成する。そのグラフで結ばれたパターン同士を最長パターン長 l_{max} になるか、分散が大きくならない範囲まで前後に伸ばしていき、頻出パターンを決定する。

4.2 階層的クラスタリングを用いたクラスタリングの改善

DGCMD アルゴリズムでは、頻出するパターンは発見できているものの、異なるクラスタに分類されるべきパターンが同じクラスタに分類されているなど、適切なクラスタリングが行えていないという問題があった。最短パターン長 l_{min} の長さのパターン同士で比較する際には、類似度に基づいて適切にクラスタリングされているが、パターン長を伸ばしていく過程で、後にはほとんど伸ばされず、前の方に伸ばされ、その結果として類似度が低くなっても類似度に基づいてクラスタリングされておらず、類似度の低いパターン同士が同じクラスタに分類されてしまうことが原因であった。

この問題点に対し、本研究ではパターンを発見した後、さらに発見された全てのパターンに対して DTW を用いた階層的クラスタリング（Ward 法 [4]）を用いて再度クラスタリングを行い、クラスタを分割することでアルゴリズムの改善を行った。DGCMD アルゴリズムでは、RSST アルゴリズムで検出されたデータ変化点を元に、それより前の部分から頻出パターンを検出し、まずパターンを前に伸ばせるところまで伸ばしてから、後に伸ばすようになっている。ここで類似度を算出するためのパラメーター設定により、類似度が低くなる場合でも最長パターン長 l_{max} になるまで前に伸ばされてしまう。それか

ら後に伸ばす際に、すでに限界まで伸ばされているため類似度が低くならない場合でも後には伸ばされていなかった。

そこで再度階層的クラスタリングを行う際にはそのパターンを後へ伸ばし、その伸ばしたパターン同士を DTW を用いて類似度を算出し、それに基づいて階層的クラスタリングをすることで DGCMD アルゴリズムだけでは検出できなかった、変化点をまたぐようなパターン同士を高精度に検出することができ、適切なクラスタリングをすることができる。図 3 に、従来手法では同じクラスタに分類されていたパターンが提案手法により適切にクラスタリングされた例を示す。

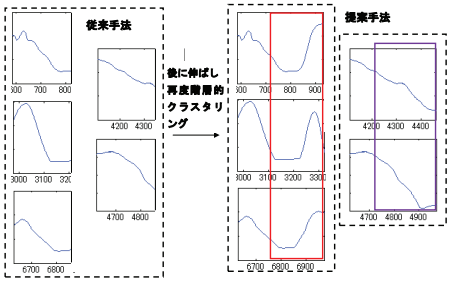


図 3: 改良されたパターンのクラスタリング

4.3 ジェスチャパターンと動作パターンの関連付け

ジェスチャパターンと動作パターンを検出したら、それぞれのジェスチャパターンと動作パターンの関連付けを行う。検出されたジェスチャパターンそれぞれに対し、ジェスチャパターン開始時刻付近にある全ての動作パターンのリストを関連付け、ジェスチャパターンのすぐ後にある動作パターンの密度を求める。その密度を元にジェスチャパターンと動作パターンの確率ネットワークを形成する。

4.4 テンプレートマッチングによる視線・ポインティングの方向検出による検出

ポインティングジェスチャの発見は以下のように行う。人差し指と手首に取り付けられたマーカーからポインティングの方向ベクトルを取得し、その直線とポスター平面の交点を検出し、1 秒以上交点が動かない場合は静止ポインティングであるとして検出した。

視線方向は、ポスターを見ているとき、もう一人の被験者を見ているとき、ロボットを見ているときを検出する。頭の中心と前に取り付けられたマーカーを通る直線を視線方向とする。この直線と、頭の中心とロボットのカメラ部分の後ろ側に取り付けたマーカーを通る直線の角度を求め、適切な閾値１、閾値２を求めた。角度が閾値１より小さいときはロボットを見ている、閾値１より大きく閾値２より小さいときはもう一人の被験者を見ている、閾値２より大きい時はポスターを見ている、というようにして検出した。

4.5 マルチモーダルイベントの検知

この実験におけるインタラクションの状態は１：被験者 A とロボットとのインタラクション、２：被験者 B とロボットとのインタラクション、３：被験者 A と被験者 B のインタラクションの３つである。

視線方向の検出結果から、被験者 A が被験者 B を見ている、かつ、被験者 B が被験者 A を見ているというイベントが発生した時、インタラクション３に変化したと検知される。

視線方向の検出と頻出ジェスチャパターンの結果から、被験者 A がロボットまたはポスターを見た後、数秒以内に A がジェスチャを行っていればインタラクション１に、同様に被験者 B がロボットまたはポスターを見た後に B がジェスチャを行っていればインタラクション２に変化したと検知される。

5. 評価実験

5.1 取得データと正解データの作成

実験は約 5 分 24 秒間行われ、38983 フレームの 78 次元のインタラクションデータを取得した（1 秒につき 120 フレーム）。また、今回の実験では時系列データ中の頻出パターンの正解データを手動で作成した。正解データの作成にはインタラクションデータ総合分析ツール iCorpusStudio[2] を用いた。今回は iCorpusStudio のラベル作成機能を用い、ビデオデータを元に頻出パターンが表出している箇所にラベルを付けた。また、マルチモーダルイベントの変化にもラベルを付け、正解データを作成した。本実験では、被験者 A のジェスチャ、被験者 B のジェスチャ、A の静止ポインティングとその位置、B の静止ポインティングとその位置、A の視線方向、B の視線方向、インタラクションの状態についてラベル付けを行った。

5.2 実験結果

5.2.1 教師なし学習による頻出パターン検出

今回発見できた頻出パターンは、被験者 A は、ポスターに沿って手を上から下へ動かす down、同じく下から上へ動かす up、撮って欲しい範囲の周りを人差し指で円を描く用に示す round、意味のないジェスチャunknown1〜3 の 6 種類だった。被験者 B は、down、round、unknown の 3 種類だった。動作に関しては、カメラを右に動かす right、左に動かす left、上に動かす up、下に動かす down、意味のない動作 unknown1〜5 の 9 種類だった。

それぞれの検出されたパターンついて、正解データと比較し個数に基づいて適合率及び再現率を求める。正解データと検出されたパターンの許容誤差 ϵ を定め、２つのフレーム差が ϵ 以内なら正解と判定する。また、正解データでは１つのパターンだが実験結果では複数のパターンとして検出されたものについては、複数のパターンで１つのパターンと数えることにする。表 1 にジェスチャパターンの適合率と再現率を、表 2 に動作パターンの適合率と再現率を示す。

表 1: ジェスチャパターンの適合率と再現率

被験者	ジェスチャ	適合率	再現率
A	ジェスチャdown	0.27	0.6
A	ジェスチャup	0.36	1
A	ジェスチャround	0.8	0.8
B	ジェスチャdown	0.5	0.57
B	ジェスチャround	0.5	0.2

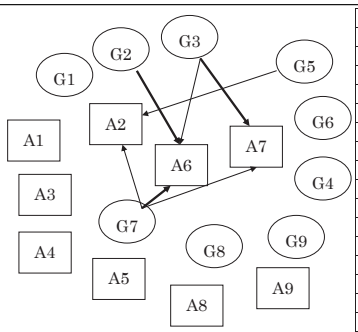
表 2: 動作パターンの適合率と再現率

動作	適合率	再現率
動作 down	0.85	0.78
動作 up	0.69	0.82
動作 left	1	0.59
動作 right	1	0.57

ジェスチャdown の適合率が低いのは、ポスター上を指し示していた後に手を下ろす動作もジェスチャdown として検出されたものが多かったからである。

検出されたパターンを元に、4.3 節で述べた方法でジェスチャパターンと動作パターンの関連付けを行った。図 4 に生成された確率ネットワーク図を示す。矢印が太いほど確率が高くなる。

意味のない動作 unknown はどのジェスチャとも関連付けられていないことが分かる。B のジェスチャdown が動作 down より高い確率で動作 up に関連付けられたことに関しては、ジェスチャdown の適合率が低いことからも分かるように、ジェス



記号	被験者	パターン種類
G1	A	round
G2	A	up
G3	A	down
G4	A	unknown1
G5	A	unknown2
G6	A	unknown3
G7	B	down
G8	B	round
G9	B	unknown
A1	-	left
A2	-	right
A3	-	unknown1
A4	-	unknown2
A5	-	unknown3
A6	-	up
A7	-	down
A8	-	unknown4
A9	-	unknown5

図 4: 確率ネットワーク

チャup がジェスチャdown として誤って分類されてしまったため、そのジェスチャの直後に現れた動作 up と関連付けられたのが原因である。

5.2.2 テンプレートマッチングによる頻出パターン検出

4.4 節で述べた方法で、静止ポインティング（被験者 A のみ）と視線方向（ロボット、もう一人の被験者、ポスター）の検出を行った。5.2.2 節と同じ方法で適合率及び再現率を求める。表 3 に静止ポインティングの適合率と再現率を、表 4 に視線方向の適合率と再現率を示す。

表 3: 静止ポインティングの適合率と再現率

	適合率	再現率
ポインティング	0.67	0.5

表 4: 視線方向の適合率と再現率

被験者	視線方向	適合率	再現率
A	ロボット	0.5	0.43
A	被験者	0.35	1
A	ポスター	0.6	0.71
B	ロボット	0.57	0.76
B	被験者	0.28	0.92
B	ポスター	1	0.6

4.4 節で、１秒以上静止している場合のみ静止ポインティングと見なして検出すると述べたが、実際には１秒未満の静止ポインティングも起きていた。しかし、１秒未満の静止ポインティングも検出しようとすると他の静止ポインティングでないジェスチャも検出されるため今回は１秒以上と定めた。１秒未満の静止ポインティングを無視した場合、再現率は１となる。また、表 4 より、テンプレトーマッチングを用いて視線方向の検出が可能なことが分かった。

また、ポスターの領域を４分割（左上、右上、左下、右下）し、静止ポインティング時にどの領域を指し示しているか調べ、ポインティング開始時のカメラが見ている領域とポインティング開始時刻付近で起きたロボットの動作との関係を調べた。静止ポインティングは全部で６回検出され、そのうちロボットの動作と関連付けられたのは５回であった。表 5 にその結果を示す。

今回は静止ポインティングの回数が少なかったため、ポインティング位置とカメラ位置の全ての組み合わせを網羅できず、ロボットの動作との関連性が分かりにくいが、この方法で関連性を調べることができることが分かった。

5.2.3 マルチモーダルイベント検出によるインタラクション状態の推定

4.5 節で述べた方法でマルチモーダルイベントを検知し、それを元に各時刻でのインタラクションの状態を推定した。今回の実験ではインタラクションの状態は３種類見られ、それぞれ正解データと比較し、秒数に基づいて適合率及び再現率を求める。表 6 にインタラクション状態の適合率と再現率を示す。これより、マルチモーダルイベントの検出によりインタラクション状態の推定が可能なことが分かった。

表 5: カメラ位置、ポインティング位置とロボット動作

番号	ポインティング位置	カメラ位置	ロボット動作
1	左上	右下	左
2	左上	右上	上
3	右上	左下	右
	右上	左下	上
4	左下	左上	下
5	左下	左上	下

表 6: インタラクション状態の適合率と再現率

インタラクション状態	適合率	再現率
A とロボット	0.78	0.2
B とロボット	0.77	0.33
A と B	0.27	0.85

以上の結果から、時系列データマイニングによるジェスチャ検出、テンプレートマッチングによるイベント検出、確率ネットワークの作成、マルチモーダルイベントの検出によるインタラクション状態の推定が提案システムで可能であると分かった。

6. まとめ

本論文では、マルチモーダルデータから時系列データマイニングとテンプレートマッチングを用いて頻出パターンの検出、インタラクション状態を推定し、ロボットの動作生成コントローラを作成するシステムを提案した。

評価実験により、撮影タスクにおいてもジェスチャと動作のパターンの対を得られることが分かり、マルチモーダルパターンの組み合わせからインタラクション状態の変化を検知できることが分かった。

今後の課題として、よりインタラクションコンテキストの理解に近づくよう、過去のインタラクション、言語情報、現在の環境といった情報も含めて動作生成コントローラを作成することが挙げられる。

参考文献

[1] 西田豊明, インタラクションの理解とデザイン, 岩波書店, November 2000.

[2] 來嶋宏幸, 坊農真弓, 角康之, 西田豊明, マルチモーダルインタラクション分析のためのコーパス環境構築, 情報処理学会研究報告, vol.2007, no.99, pp.63-70, 2007.

[3] Y. Mohammad, T. Nishida, S. Okada, "Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction, " IROS 2009.

[4] Cosma Shalizi, "Distances between Clustering, Hierarchical Clustering, " http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

[5] Y. Mohammad, T. Nishida, "Constrained motif discovery, " International Workshop on Data Mining and Statistical Science (DMSS2008), September 2008.

[6] Y. Mohammad, T. Nishida, "Robust singular spectrum transform, " Proceeding of the 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems IEA-AIE 2009.