1G3-4

時系列マルチモーダルデータマイニングを用いたロボットの撮影行動則の獲得

The Acquisition of the Robot Filming Behavior Rules by Time Series Multi-modal Data Mining

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

[*1]京都大学大学院情報学研究科知能情報学専攻
Graduate School of Informatics Kyoto University

[*2]東京工業大学大学院総合理工学研究科知能システム科学専攻
Department of Computational Intelligence and Systems Science Tokyo Insutitute of Technology

In this research, we propose a robot system which can recognize gesture and action patterns from multiple motion sensors and implement the filming task corresponding to them. When we want the robot to film our working, the robot should move automatically to some degree and moreover act corresponding to their directions.

To implement it, first we do the filming tasks by Woz. We can get multi-modal multi-dimensional time series data from sensors. The robot can learn the action which peculiar to the task and direction patterns by time series data mining. Two obtained learning result can be integrated based on the changing points of interaction states and we can get the robot action generation controller. In this paper we do the learning of direction patterns. As a result, detected gestures are clustered by 56.6% purity and 4 kinds of gestures of all 7 kinds are associated with corresponding actions.

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]。特にユーザの動作に基づいて作業の記録を行う撮影ロボットにおいては、空間的な情報を直感的に伝えることのできる非言語情報でのインタラクションが重要となる。本研究ではユーザとロボットで撮影タスクを行い、ジェスチャや視線、姿勢、位置関係といったマルチモーダルデータをもとに、ロボットがユーザとのインタラクションの状態を推定し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。

また、例えば手芸ノウハウを撮影する際には、作業中は作業者の手元にズームインし、ひとつの作業が終わるとズームアウトして全体像を映す、というようなタスクに固有の撮影技術を、ユーザの指示がなくても行うことが望ましい。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タスクの学習と指示パターンの学習という 2 つの学習を行うことで人間と自然なインタラクションを行うことのできるロボットの動作生成コントローラを獲得するシステムを提案する。

本論文では提案したシステムのうち、指示パターンの学習に関する実験を行い評価する。実験結果をもとに、提案システムが撮影タスクに適用できるか検討し議論することが本論文の目的である。撮影タスクを設定し指示パターンの学習を行った結果、検出されたジェスチャは purity56.6%の精度でクラスタリングされ、また指示者が使用した 7 種類のジェスチャのうち 4 種類はそれに対応する正しいロボットの動作と関連付けられた。

連絡先:　坂本佳愛　京都大学大学院情報学研究科
sakamoto@ii.ist.i.kyoto-u.ac.jp

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。ここで Actor ロボットは Woz（Wizard of Oz）で操作されており、Operator のジェスチャに対して正解の動作を行う。ここではナビゲーションタスクを行い、ユーザとロボット 1 対 1 でインタラクションを行った。

[6] では、[3] で用いられた Robust Singular Spectrum Transform (RSST) アルゴリズム [5]、Distance Graph Constrained Motif Discovery (DGCMD) アルゴリズム [4] といった頻出パターン検出アルゴリズムを用いて、ユーザ 2 人とロボット 1 体とのポスター撮影タスクに適用した。DGCMD アルゴリズムの改善を行い、クラスタリングの精度を向上させた。さらに、複数のモダリティのパターンの組み合わせからインタラクション状況の変化点をマルチモーダルイベントとして規定し、これを検出することでロボットがルールベースでインタラクション状況を推定した上で動作を生成するシステムを提案した。

本研究では、[6] の枠組みを用いて撮影タスクに適用する。[6] の撮影タスクではポスターという平面のものを対象としていたが今回はトルソーという立体物を対象とし、カメラ部分だけでなくロボット自体も動くタスクとなっており、より多種のジェスチャが観測されるタスクになっている。また、本研究ではタスクの学習と指示パターンの学習という 2 つの学習を行う。タスクの学習では、あらかじめ撮影タスクに必要な基本動作を獲得し、指示パターンの学習では、ユーザのジェスチャに応じたロボットの行動則を獲得する。この 2 つの学習の結果を統合することで自然なインタラクションの実現を目指すのが本研究の最終目標である。

3. システムの全体像

3.1 システム概要

本システムでは、タスクに固有な撮影技術と、ユーザの自由で多種なジェスチャの両方を学習するので、タスクの学習と指示パターンの学習という 2 つの学習を行う。手順としては、まず Woz を用いてロボットを操作してユーザと撮影タスクを行う。このタスクでセンサーを通じて得られたマルチモーダルな多次元時系列データからタスクの学習と指示パターンの学習との 2 つの学習を行う。タスクの学習は、タスクに依存する作業者の動作とそれに対するロボットのアクションの対を獲得するために行われる。指示パターンの学習では、指示者のジェスチャとそれに対するロボットのアクションの対を獲得するために行われる。これらの学習の結果を、マルチモーダルイベントによるインタラクション状況の推定に基づいて統合し、ロボットの動作生成コントローラを得る。図 1 に、システム概要を示す。

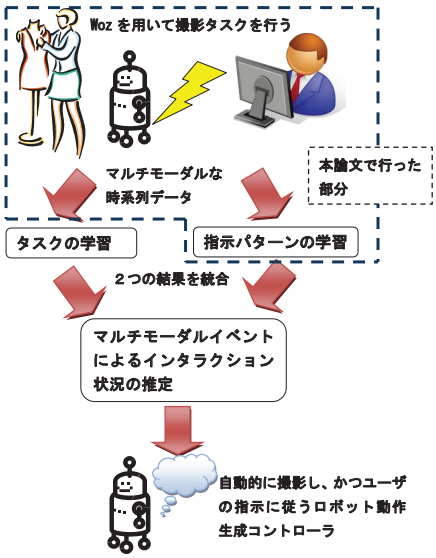

図 1: システム概要　2つの学習を用いて行動則獲得

3.2 タスクの学習

タスクの学習では、タスク作業者とロボットの動作を観測し、ロボットは撮影の基本的な動作を学習することとなる。具体的には、作業者の手元を追うのを基本動作とし、ひとつの作業が終わると全体像を撮るためにズームアウトする、などといった撮影の一連の動作を獲得する。作業者に取り付けたセンサーから得られる時系列データをもとに HMM などを使用して教師あり学習を行い、作業者の動作とロボットの動作のパターンの対を獲得し、前述のような撮影の基本動作を学習する。

3.3 指示パターンの学習

指示パターンの学習では、[3] で提案された RSST アルゴリズムと DGCMD アルゴリズムをベースに用いて頻出パターンの検出を行う。具体的には、「左に回れ」といったジェスチャが見られたらロボットは左に回り、「カメラを下に」といったジェスチャが見られたらカメラを下に動かすといったような動作を獲得する。指示者に取り付けたセンサーから得られる時系列データをもとに前述の RSST アルゴリズムや DGCMD アルゴリズムを用いて教師なし学習を行うことで、指示者のジェスチャとロボットの動作パターンの対を獲得する。図 2 に、センサーから得られたデータの処理の流れを示す。

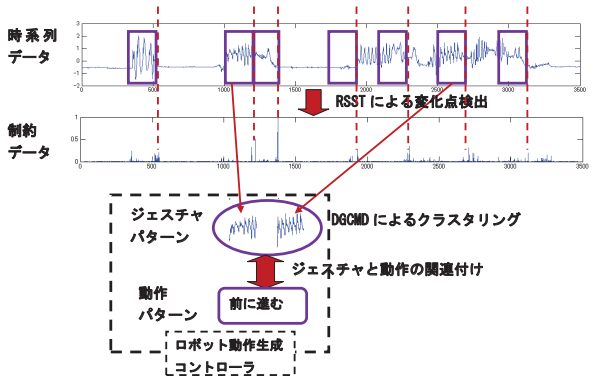

図 2: センサーから得られた時系列データの処理フロー

RSST アルゴリズムは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 の前の部分と後ろの部分を比較し、変化度を調べる。これを時系列データの最初から最後まで適用し、時系列データの変化点データを得る。

DGCMD アルゴリズムは、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アルゴリズムである。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切り出し、切り出された全てのパターン同士の距離行列を DTW を用いて作成する。その後、閾値より小さいものだけを残して距離グラフを作成し、そのグラフで結ばれたパターン同士を、分散が大きくならない範囲まで前後に伸ばし、頻出パターンを決定する。

3.4 マルチモーダルイベントによるインタラクション状況推定

最終的には、タスクの学習と指示パターンの学習の結果を統合してロボットの動作生成コントローラを得る。ロボットは基本的にはタスクの学習の結果に従い、作業者の動作に応じて動くが、指示者がジェスチャを行った場合は、指示パターンの学習の結果に従ってそのジェスチャに応じて動く。しかし、実際にはそのジェスチャがロボットに向けられたものではない可能性も考えられる（指示者が作業者と話している際に出たものなど）。そこで、ジェスチャだけでなく視線、姿勢、位置関係といった複数のモダリティデータからインタラクション状況の推定を行った上で、ロボットは動作を生成する。ここで複数モダリティのパターンの組み合わせから定義されるインタラクション状況の変化点をマルチモーダルイベントと呼ぶ。例えば、指示者がロボットの方を見ながらジェスチャを出した場合は、視線方向「ロボット」とジェスチャ「前進」というパターンの組み合わせから指示者とロボットとのインタラクションであると判断し、ロボットはジェスチャに応じた動作を行う。図 3 にマルチモーダルイベントの概念図を示す。

4. タスク・環境設定

4.1 タスク内容

本研究では、ユーザが T シャツの飾り付けを行い、ロボットがその様子を撮影するタスクを題材とする。タスクは作業者、指示者、撮影ロボットで行う。作業者はトルソーに着せた T シャツに飾り付けを行う。指示者は撮影したい箇所をロボットに対して指示する。ロボットは指示者の指示や作業者の動作に応じてムービー撮影をする。作業者は指示者の先生のような

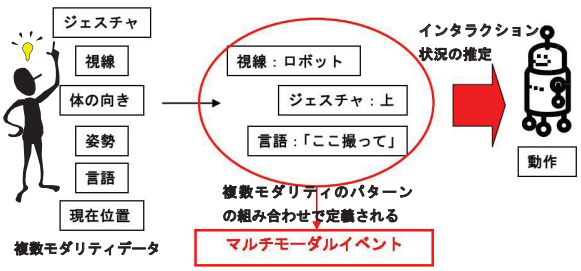

図 3: マルチモーダルイベントの概念

存在で、指示者は先生の作業を見て記録しておきたいポイントなどを、ロボットに指示して撮影させる。ロボットに対する指示はハンドジェスチャのみを用いる。図 4 に撮影タスクの様子を示す。

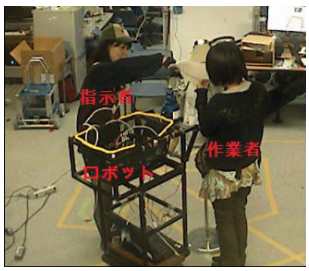

図 4: 撮影タスク T シャツ飾り付け

4.2 ロボット

今回使用したロボットは、図 4 にも示されているようなカートロボットを用いた。ロボット本体上部にパンチルト制御可能なネットワークカメラが取り付けられ、下部には車輪駆動部がある。Woz 操作者はネットワークカメラからの映像を見ながら、カメラの角度やロボットの移動を制御することができる。

4.3 使用したセンサー類

IMU-Z は 3 軸加速度、3 軸角速度、3 軸地磁気の 9 軸モーションセンサを搭載し、Bluetooth で通信できるセンサーである。今回の実験では指示者の右手と背中に取り付け、右手に取り付けたセンサーの加速度をジェスチャの時系列データとして本システムに使用した。

MAC3D はリアルタイム光学式モーションキャプチャシステムである。指示者の被っている帽子、ロボット本体とカメラ部、トルソーにそれぞれマーカーを取り付け、位置関係やカメラの角度、指示者の頭部方向等を得ることができる。今回の実験分析ではこのセンサーのデータは使わないが、マルチモーダルイベントの検出の際に複数モダリティの時系列データを取得するために使用する。

5. 指示パターン学習のための実験

今回の実験では、提案したシステムのうち指示パターンの学習のみを行う。これにより、T シャツの飾り付けを撮影するタスクにも提案システムの学習方法が適用可能か検討する。

5.1 取得データと正解データの作成

T シャツの飾り付けを指示者の指示に従って撮影するタスクを約 9 分間行ったところ、指示者の右手に取り付けた IMU-Z センサーから 26971 フレームの時系列データを取得した（1 秒につき 50 フレーム）。

また、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。iCorpusStudio のラベル作成機能を用い、タスクのビデオデータを元にして指示者の頻出パターンが表出している箇所にラベルをつけていき、正解データの作成を行った。

5.2 実験結果

今回の実験で見られた指示者が用いたジェスチャパターンは、表 1 に示すとおり D1〜D7 の 7 種類であった。ロボットの動作も、これらのジェスチャに対応して 7 種類の動作が見られた。これらを A1〜A7 とする。

表 1: タスクで見られた指示者のジェスチャ

D1	左に回転させる
D2	右に回転させる
D3	カメラを近寄らせる
D4	カメラを遠ざける
D5	カメラの角度を下げる
D6	カメラの角度を上げる
D7	前進させる

本システムを用いて検出されたジェスチャパターンは 20 種類であった。これを G1〜G20 とする。それぞれの検出されたパターンとラベル付けされた正解データを比較し、正解データのラベルと半分以上時間が重なっているものを正解と判定し、その個数の一覧を表 2 に示した。Non は正解と判定されなかったパターンの個数である。これらの結果からクラスタリング精度の評価尺度のひとつである Purity を求めると、56.6%となった。

表 2: ジェスチャパターンと正解データの比較

	D1	D2	D3	D4	D5	D6	D7	Non	合計
G1	0	0	0	1	0	0	1	0	2
G2	0	1	0	0	0	0	1	0	2
G3	0	1	1	0	0	0	0	0	2
G4	0	0	0	0	0	1	1	1	3
G5	0	0	0	0	0	0	0	3	3
G6	0	0	0	0	1	0	0	4	5
G7	1	0	0	0	0	0	4	1	6
G8	3	2	0	0	0	0	0	0	5
G9	1	0	0	0	0	0	0	0	1
G10	0	1	0	0	0	1	0	2	4
G11	0	0	0	0	0	0	0	4	4
G12	0	0	0	0	1	0	0	3	4
G13	1	0	0	0	0	0	0	2	3
G14	1	0	0	0	1	0	0	2	4
G15	2	0	0	0	0	0	1	1	4
G16	1	0	1	0	0	0	1	2	5
G17	0	0	5	0	0	0	1	2	8
G18	0	0	0	0	1	0	0	5	6
G19	2	0	0	3	0	0	0	0	5
G20	2	0	0	3	0	0	0	2	7
合計	14	4	7	7	4	2	10	34	83

さらに、検出されたパターンを元に、指示者のジェスチャパターンとロボットの動作パターンの関連付けを行う。ロボット

の動作パターンは、Woz の操作履歴から自動的にラベル付けされたものを用いる。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から遅延 s 秒以内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後にある動作パターンの密度を求める。今回は正解データのジェスチャパターンと動作パターンの間の遅延を平均したものから考えて $s = 3$ とした。その密度が閾値より大きいもののみを用いて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。G5、G6、G9、G10、G11、G12、G18、G20 はどの動作パターンとも関連付けられなかった。G1、G2、G3、G4、G13、G14、G15、G16 はばらつきが大きく密度が低くなった。密度が閾値より大きくなった G7、G8、G17、G19 のみ図 5 に確率ネットワーク図として示す。

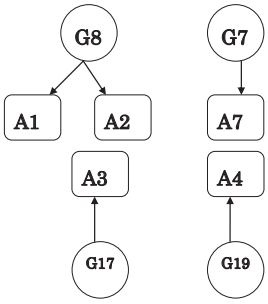

図 5: 確率ネットワーク

以上の結果から、ジェスチャ検出の精度は決して良くないが、確率ネットワークを形成すると、ロボットの動作生成コントローラとして十分使えるものを得られることが分かった。しかし、A5 と A6 がどのジェスチャとも関連付けられなかったり、A1 と A2 は回転方向が逆の関係にあるのに同じジェスチャに関連付けられたりしている。

5.3 結果に対する議論

今回ジェスチャデータとして加速度のみを用いて頻出パターン検出を行ったが、IMU-Z センサーは地磁気や角速度もセンシングできるので、それらを統合して用いることでクラスタリング精度を改善できる可能性がある。実際に地磁気データを用いて、同じように指示パターンの学習を行うと、クラスタリング精度は加速度データほど良くないが、加速度の結果と比較すると、加速度で検出できなかったジェスチャが検出されていることや、加速度でも地磁気でもジェスチャが検出された箇所は意味のあるジェスチャが表出している可能性が高いことが分かった。よって、モーションセンサの複数種類の値から、よりクラスタリング精度を高める方法を今後検討する必要がある。

また、MAC3D センサーで得られたジェスチャ以外のマルチモーダルデータを用いて、よりジェスチャ検出の精度を上げられる可能性も考えられる。例えば指示者の頭部方向に着目すると、タスク中の大部分はロボットの方を向いているが、トルソーや指示者を向いているときは意味のあるジェスチャが表出していなかった。つまり、ロボットの方を向いていない区間でジェスチャが検出されたとしても、意味のないジェスチャとして除外することができ、結果としてクラスタリング精度を上げることができる。マルチモーダルデータによるクラスタリング精度の向上は今後検討する予定である。

6. まとめ

本研究では、ユーザとロボットで撮影タスクを行い、ジェスチャや視線、姿勢、位置関係といったマルチモーダルデータをもとに、ロボットがユーザとのインタラクションの状態を推定し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を実現するロボットシステムの提案を行った。Woz で撮影タスクを行ってマルチモーダルな時系列データを獲得した後、タスクの学習と指示パターンの学習という 2 つの学習を行い、その学習結果をマルチモーダルイベントに基づいて統合し、自然なインタラクションのできるロボットの動作生成コントローラを獲得するのが本研究の最終目標である。本論文では、この提案したシステムのうち指示パターンの学習の部分を取り上げ、今回設定した撮影タスクにも [?] の枠組みが適用できるか実験を行ったところ、Purity56.6%でジェスチャのクラスタリングが行われた。また、指示者が使用した 7 種類のジェスチャのうち 4 種類はそれに対応する正しいロボットの動作と関連付けられた。

今後の課題としては、前節でも述べたジェスチャ検出精度の改善が挙げられる。また、2 つの学習結果を統合して動作生成コントローラを獲得するという最終目標に向けて、指示学習だけでなく、タスクの学習やマルチモーダルイベントによるインタラクション状況推定を行い、それらの結果を統合することで自然なインタラクションを可能とするロボットシステムを構築する予定である。

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

[5] 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.

[6] 坂本佳愛, 岡田将吾, 西田豊明, 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, 第 24 回人工知能学会全国大会 (2010).