オプティカルフローと音声トリガを用いた 手招き動作の認識手法

A Method for Recognition of Gestures Expressed by Cyclically Repeated Motion of the Hand Using Optical Flow and Voice Trigger

西村正典[†], 西川 敦[†], 小荒 健吾[†], 宮崎 文夫[†]

Masanori Nishimura[†], Atsushi Nishikawa[†], Kengo Koara[†] and Fumio Miyazaki[†]

あらまし 本報告では、人間の自然な手招き動作における身振りと言葉の時間的な関連性に注目し、手招き動作を撮影した連続動画のオプティカルフローから算出される手の移動速度と言葉による指示コマンドの発声タイミングを特徴量としてとらえ、音声をトリガとして用いることにより、指示者の意図する指示方向を自動認識する手法を提案する。

キーワード：身振り認識、手招き動作、オプティカルフロー、マルチモーダルインタフェース、音声トリガ

1. ま え が き

我々が日常生活で何気なく用いている身振り動作の代表的なものとして、指示対象となる人や機械の移動方向などを手の反復動作により表現する“手招き動作”がある。このような人間の自然な動作を非接触かつ無標点で自動認識する技術は、人にやさしいヒューマンインタフェースシステムを構築する上で極めて重要である。

手招き動作の解析に関しては、これまでにいくつかの研究がなされている¹⁾²⁾。これらの研究のほとんどは、反復動作がある速度パターンに従って行われるという仮定のもとに、手の3次元位置や速度を解析し、モデル化している。逆にいえば、ある決まった速度パターンを持つ身振りしか認識できないことになる。これでは、人間にとって使いやすいインタフェースとはいえない。人間の自然な身振りを認識するためには、コンピュータ側が身振りのパターンを指示するのではなく、人間が無意識に行う身振りからその特徴となる情報を抽出し、モデル化する必要がある。しかし、一般に、画像処理のみからこのような特徴量を検出することは容易でない。

本報告では、従来の視覚情報のみを用いた身振り認識手法に音声情報を付加することにより、特定の身振りコマンドではない自然な手招き動作の認識を試みる。提案手法の最大の特徴は、手招き動作による指示と言葉による指示の時間的な関連性に注目し、音声をトリガとして用いること

により手招き動作の指示方向を認識している点である。これにより、手招き動作の反復回数や速度、軌跡等の個人差および、指示言葉の曖昧さ（『こっち』等の方向を意味しない表現）に対してロバストに指示方向を認識できる。

2. 手招き動作における発声タイミングの測定

まず、人間が自然に行う手招き動作の指示方向と指示言葉の発声タイミングの関係を実験的に調査した。

本実験では、1台のCCDカメラを用いて身振り動画（濃淡画像列）を撮影する。画像のサイズは1枚あたり640×480画素、各点8ビットである。レンズの焦点距離は8mm（水平方向等価焦点距離912画素、垂直方向864画素）のものをを用いた。これらの画像入力およびその局所相関演算に基づくオプティカルフローを検出するための専用の画像処理装置として、富士通社製のカラートラッキングビジョンPCI版（TRV-CPW5³⁾）を利用した。本実験では、画面全体に横16個×縦14個のオプティカルフロー検出格子点を等間隔に設け、各フレーム毎に検出されたオプティカルフローの平均値をそのフレームでの手の移動速度とする⁴⁾。手の移動速度のサンプリング周期は2 frameである。なお、トラッキングビジョンの制御および認識処理を行うホストの計算機には1台のPC/AT互換機（Intel MMX Pentium, 166MHz, OS: Linux⁵⁾）を用いた。

被験者にはマイクを装着してもらい、音声をホスト計算機に入力する。ホスト計算機は、入力された音声に簡単なノイズ除去（閾値処理）を施した後、ある一定の音量のみを検出し、言葉の入力の有無を認識する。このとき、指示者の言葉のみをロバストに検出する必要があるが、本システムではバックに音楽を流した状態で、さらにすぐ隣で会

2000年3月1日受付, 2000年4月28日再受付, 2000年5月25日採録

[†] 大阪大学 大学院基礎工学研究科 システム人間系専攻 機械科学分野
(〒560-8531 豊中市待兼山町1-3, TEL 06-6850-6208)

[†] Graduate School of Engineering Science, Osaka University
(1-3, Machikaneyama, Toyonaka-shi, Osaka 560-8531, Japan)

図1 手招き動作の速度変化と発声の関係
The transition of hand motion velocity and the vocalization timing.

話する人がいる環境でも、指示者の発声のみを瞬時に検出できることが確認できている。

被験者6人に、手招き動作を左右それぞれ20回行ってもらった。1回の手招き動作での手の反復回数は特に指定せず、各手招き動作中に一度だけ『こっち』などの好きな言葉を発声してもらう。あくまで自然な身振り動作での発声タイミングを測定するため、被験者には実験の内容は説明しない。実験前には「発声はいつでも良い」「自然な手招き動作をして下さい」「常に振る方向を意識して下さい」などの指示を与える。

左方向へ手招き動作した場合の、振り方向の速度変化と発声タイミングの一例を図1に示す。図において、横軸は時刻(s)、縦軸は手の速度(pixel/frame)である。図中の斜線を施した部分は、言葉が検出された時間を示す。反復動作の速度は、ほぼ正弦曲線に従って変化していることがわかる。ここで、速度が負のとき手は右方向に移動していることを、速度が正のとき手は左方向に移動していることを意味する。

被験者毎に手の速度や振幅に個人差があるので、手の速度変化が正弦曲線に従っていると仮定し、実験により得られた各被験者毎の発声開始時刻を速度変化の1周期分を基準として正規化する。実際には、往復にかかる時間にばらつきがあるので、中間の節を0.5として半周期ごとに正規化する(図1参照)。以後、この正規化した値を発声タイミング t_s^* ($0 \leq t_s^* < 1$) と定義する。この定義より明らかなように、左右方向に手を反復させる手招き動作の場合、 $0 < t_s^* < 0.5$ のとき手は右方向に移動しており、 $0.5 < t_s^* < 1$ のとき手は左方向に移動していることになる($t_s^* = 0, 0.5$ のとき手の移動速度はゼロである)。右および左方向に手招き動作した場合の発声タイミングの解析結果を図2にヒストグラム表示する。

これより、右方向の手招き動作は $t_s^* = 0 \sim 0.6$ 、左方向の手招き動作は $t_s^* = 0.5 \sim 1.1(0.1)$ 付近に分布しており、ピークはそれぞれ $t_s^* = 0.3$ と $t_s^* = 0.8$ というように、

ちょうど正反対に出ているのがわかる。この結果より、ほぼすべて、手の移動方向が意図する指示方向と一致しているときに、声を発していることがわかる。つまり、声を発した瞬間の手の移動方向に着目することで手招き動作の指示方向を認識できると考えられる。しかし、この結果を見てもわかる通り、正弦曲線の節の付近で分布が重なり合っている。実際の身振り認識にこの情報を用いる場合、この分布の重なりが大きな問題(誤認識の要因)となる。次節では、この問題に対処する方法について述べる。

3. 認識結果の信頼性の評価

前節の最後で述べた問題点を踏まえ、ここでは、認識結果に“確からしさ(Certainty)” $C(0 \leq C \leq 1)$ という評価値を発声タイミング t_s^* の関数として与えることにより、認識結果の信頼度を評価することを考える。適用するアプリケーションに応じてある閾値以下の C 値をとる認識結果を除外することで、認識結果の信頼性を任意に変更できる。

“確からしさ”の評価関数は次の通りである。図2に示した右および左方向に手招き動作した場合の発声タイミングのヒストグラムの差の分布を正規化し、正弦曲線で近似し絶対値をとったものを評価関数とする(1式)。

$$C(t_s^*) = |\sin(2\pi(t_s^* - 0.05))| \quad (1)$$

この関数は、図2において、左右の発声頻度が同じとなる発声タイミング t_s^* で値がゼロ(節)となり、左右の発声頻度がピークになる発声タイミング t_s^* でそれぞれ値が1となり、 C 値が1に近づくほど、認識結果の“確からしさ”が高いことになる。なお、評価関数中の0.05という位相のずれは、発声タイミングの分布のピークと手の移動速度のピークとのずれ量に相当する。(1式から明らかなように、任意の t_s^* に対して $C(t_s^*) = C(t_s^* + 0.5)$ が成立する。そこで、便宜上、任意の t_s^* に対して $C(t_s^*) = C(\tilde{t}_s^*)$ が成立す

図2 発声タイミングの分布
Histogram of vocalization timing.

る等価発声タイミング $\tilde{t}_S^*$ ($0 \leq \tilde{t}_S^* < 0.5$) を次式で定義する.

$$\tilde{t}_S^* = \begin{cases} t_S^* & \text{if } 0 \leq t_S^* < 0.5 \\ t_S^* - 0.5 & \text{if } 0.5 \leq t_S^* < 1 \end{cases}$$

なお, ここでは, 左右方向の手招き動作を例に挙げて“確からしさ”を定義したが, 任意の方向に手を反復させる手招き動作についても全く同様の議論が成立することに留意されたい.

4. 手招き動作の認識手法

本章では, 前章までの結果を踏まえた上で, 音声トリガを利用した手招き動作の方向検出の具体的な手法を説明する.

前章で, 発声時の手の速度情報から手招き動作の方向検出が可能であることを説明したが, 現実には画像のサンプリング周期との兼ね合いから, 発声時の正確な手の速度を測定することはできない. そのため, 発声時の手の速度は発声の直前, 直後にサンプリングされた速度から推定する必要がある.

発声時の手の速度推定方法は, 発声直前, 直後の手の移動方向の違いから, 次の2通りに分けられる.

- (1) 発声の前後で手の移動方向が変化しない場合
- (2) 発声の前後で手の移動方向が逆転する場合

ここで, 発声の直前, 直後の時刻と手の移動速度をそれぞれ, $t_0, \mathbf{V}_0 = (u_0, v_0)^T$, $t_1, \mathbf{V}_1 = (u_1, v_1)^T$ とし, 発声を検出した時刻を t_S , 求める手の移動速度を $\mathbf{V}_S = (u_S, v_S)^T$ とする. このとき $\mathbf{V}_S = (u_S, v_S)^T$ を次式より推定する.

$$\mathbf{V}_S = \begin{cases} \frac{t_1 - t_S}{t_1 - t_0} \mathbf{V}_0 + \frac{t_S - t_0}{t_1 - t_0} \mathbf{V}_1 & \text{if } (\mathbf{V}_0^T \cdot \mathbf{V}_1) > 0 \\ \mathbf{V}_0 & \text{if } (\mathbf{V}_0^T \cdot \mathbf{V}_1) \leq 0 \\ & \text{かつ } \Delta t > 0 \\ \mathbf{V}_1 & \text{otherwise} \end{cases}$$

ここで,

$$\Delta t = \frac{t_1 |\mathbf{V}_0| + t_0 |\mathbf{V}_1|}{|\mathbf{V}_0| + |\mathbf{V}_1|} - t_S$$

上式は, 発声前後の速度の内積が正の場合, 前後の速度の直線補間値を発声時の手の速度とし, 内積が負の場合には手がその間で折り返したと判断し, 折り返し点(速度がいったんゼロになる点)を境に発声時の手の速度を発声前後の手の速度のうち, より近い方の速度で近似していることを意味する. Δt は, 折り返し点から発声時までの時間差を表す.

以上の手順により推定された手の移動速度方向 $\mathbf{V}_S = (u_S, v_S)^T$ を, 手招き動作の指示方向とする. 次に, “確からしさ” C の算出方法について説明する. まず, 上述の手順で求めた発声時の手の移動速度 $|\mathbf{V}_S|$ を, 手の移動速度の最大値 $|\mathbf{V}_{MAX}|$ で正規化し, $V_S^* (0 \leq V_S^* \leq 1)$ とする.

$$V_S^* = \frac{|\mathbf{V}_S|}{|\mathbf{V}_{MAX}|} \quad (2)$$

ここで, $|\mathbf{V}_{MAX}|$ の値は, 発声以前の手の速度履歴から算出しているが, 発声以前に十分な速度履歴が存在しない場合は, あらかじめ測定しておいた不特定の被験者の平均値を初期値として用いる. 次に, V_S^* から等価発声タイミング $\tilde{t}_S^* (0 \leq \tilde{t}_S^* < 0.5)$ を次式により求める.

$$\tilde{t}_S^* = \frac{\sin^{-1}(V_S^*)}{2\pi} \quad (3)$$

なお, 上式からは $\tilde{t}_S^*$ の値が2つ求まるが, 発声前後の速度勾配から適切な方を選択する. 最後に, 3章で定義した“確からしさ”の評価関数((1)式)に $t_S^* = \tilde{t}_S^*$ として代入し, “確からしさ” C を得る.

以上の手順で手招き動作の指示方向と, その“確からしさ” C を求める. 認識結果の信頼性を評価するために“確からしさ”の閾値 T を設け, $C \geq T$ となる場合のみ, その認識結果を“充分信頼できる結果”として採用し, それ以外の結果は認識結果として採用しない.“確からしさ”の閾値 T は適用するアプリケーションに応じて任意に選択できる.

5. 検証実験

本手法の有効性を検証するために, 次のような認識実験を行った.

5.1 実験手順

実験システムの概要を図3に示す. 被験者の正面1.5mにモニタを配置し, モニタ上部に被験者の身振り動作をとらえるカメラを設置する. モニタ上には, 被験者自身の映像が映し出されており, さらに認識結果等の情報が逐次表示される(図3(b)参照).

モニタ下部中央に図3(b)に示すように, ポインタを表示し, このポインタを被験者の手招き動作に対応させて指示方向に1ステップずつ動かす. ポインタが左右共に10ステップ進んだ所にゴールを設け, 被験者には, なるべく少ない回数でポインタを左右それぞれのゴールに誘導させるというタスクを与える. 最短で, 10回でゴールにたどり着き, 1回誤認識する毎に1ステップ後退することになるので余分に1回指示する必要が生じる. 身振り動作が認識されなかった場合は, ポインタは移動せず, 総指示回数だけが1増えることになる. 認識結果として, 目標達成にかかった身振り動作の回数と左右それぞれの方向に認識された回数, および認識されなかった回数をカウントする.

図3 実験システムの概要
Schematic representation of experimental system.

5.2 実験結果

検証実験を4人の被験者 A, B, C, D に対して行った。左右それぞれの方向について、“確からしさ”の閾値を $T = 0.3, 0.9$ の2通りに設定し、1人につき左右合わせて計4回行ってもらった。結果を表1~表5に示す。

実験結果からわかるように、閾値を0.3に設定した場合の誤認率は10.7%であるのに対して、閾値を0.9まで引き上げると誤認率は1.9%まで低減している。一方、認識率に注目すると、閾値が0.3の場合82.1%であるのに対し、閾値が0.9の場合には78.8%となっており、信頼性を重視し誤認識を最小限に抑えようとした結果、かえって認識率が低下していることがわかる。このように、“確からしさ”の閾値を大きくすると誤認率は低下するが、認識率もわずかながら同様に低下する場合がある。これは、誤認識の可能性がわずかでもある曖昧な結果を無視しているため、実際に正しい認識結果も除外されてしまう可能性があるからである。

表1 実験結果: 指示方向が“左”の場合 ($T = 0.3$)
Case that the human intention is “left” and $T = 0.3$.

	A	B	C	D	合計
左と正しく認識された回数	10	11	13	11	45
右と誤って認識された回数	0	1	3	1	5
認識されなかった回数	0	1	2	1	4
合計	10	13	18	13	54

表2 実験結果: 指示方向が“左”の場合 ($T = 0.9$)
Case that the human intention is “left” and $T = 0.9$.

	A	B	C	D	合計
左と正しく認識された回数	10	10	10	11	41
右と誤って認識された回数	0	0	0	1	1
認識されなかった回数	0	4	3	3	10
合計	10	14	13	15	52

表3 実験結果: 指示方向が“右”の場合 ($T = 0.3$)
Case that the human intention is “right” and $T = 0.3$.

	A	B	C	D	合計
右と正しく認識された回数	11	11	13	12	47
左と誤って認識された回数	1	1	3	2	7
認識されなかった回数	0	1	1	2	4
合計	12	13	17	16	58

表4 実験結果: 指示方向が“右”の場合 ($T = 0.9$)
Case that the human intention is “right” and $T = 0.9$.

	A	B	C	D	合計
右と正しく認識された回数	10	11	10	10	41
左と誤って認識された回数	0	1	0	0	1
認識されなかった回数	1	2	2	5	10
合計	11	14	12	15	52

表5 認識率と誤認率の比較
Comparison between recognition and mistaking rates.

(a) 認識率			(b) 誤認率		
	$T = 0.3$	$T = 0.9$		$T = 0.3$	$T = 0.9$
方向: 左	45/54	41/52	方向: 左	5/54	1/52
方向: 右	47/58	41/52	方向: 右	7/58	1/52
合計 (%)	82.1	78.8	合計 (%)	10.7	1.9

5.3 従来法との比較

提案手法では、不特定の人間の手招き動作に対して80%程度の認識率（誤認率2%以下）を達成することができた。一方、音声情報を付加しない従来法の場合を考えると、文献1)の手法では、90%以上の認識率で手招き動作の指示方向を認識している。文献2)には認識率の記述はないが、同程度の高い認識率が得られるものと推察される。また、筆者ら⁴⁾も、以前の研究においてオプティカルフローに基づく身振り認識法を提案し、この方法を用いて手招き動作の指示方向を不特定の動作者に対して90%以上の認識率（誤認率ゼロ）で認識することに成功している。このように、認識率および誤認率の数字を単純に比較する限り、提案手法の方が従来法よりも優れているとはいえない。しかし、実際には、従来法で認識できるのは必ずしも人間にとって自然な動作であるとは限らない特定の身振りパターンのみ（たとえば、左手をカメラに向け静止させて右手を往復させる²⁾、静止状態から片手を2回往復させる⁴⁾など）であり、従来法の数字（認識率）は、被験者が常にその動作パターンの特徴（動作開始時の状態、反復回数、手の位置・速度など）を意識して身振りを行った結果に過ぎないことに留意されたい。一方、提案手法では、上記のような動作の特徴を意識する必要は全くない。各人にとってごく自然な手招き動作を自然な発声とともにに行いさえすれば、平均で80%程度の認識率が得られる。この柔軟性が従来法との最大の相違点であり、提案手法の最大の長特でもある。

6. 移動ロボット誘導への適用

前章の検証実験では左右方向の指示のみを認識対象としたが、提案手法は左右以外の指示方向も認識できる。ここでは、指示方向がいろいろ変化する応用例として、提案手法を“パン・ティルト制御が可能なCCDカメラを搭載した車輪型移動ロボット⁶⁾”の誘導に適用した結果を示す。

実験に用いるロボット（図4(a)）は無線モデムを介してホスト計算機からの指示を受け、車輪の回転および搭載カメラの姿勢を制御することができる。搭載カメラにはトランスミッタが内蔵されており、ホスト計算機に接続されたカラートラッキングビジョンへの映像入力も無線化されている。システム構成の詳細は文献6)を参照されたい。

指示方向とロボットの移動方向の対応関係は、図4(b)に示す通りである。すなわち、手招き動作を行う動作領域と

(a) 移動ロボットの外観 (b) 指示方向とロボットの移動方向の関係

図4 手招き動作による移動ロボットの誘導
Mobile robot navigation using gestures.

図5 手招き動作によるロボット誘導の様子
A gesturer and a mobile robot that follows gestures.

表6 実験結果(移動ロボット誘導)
Experimental result (mobile robot navigation).

	$T = 0.3$	$T = 0.9$
指示方向が正しく認識された回数	11	10
指示方向が誤って認識された回数	1	0
どの方向とも認識されなかった回数	0	3
合計の指示回数	12	13

なる2次元平面上の指示方向をロボットが移動する床平面上に投影し移動方向とした。ロボットは1回の指示動作に対して、指示方向に一定距離進んだ後、停止し指示者の方向にカメラを向け、次の指示動作を待つという一連の動作を行う。手招き動作から指示方向が判別できない(“確からしさ”が閾値より低い)場合は、首(カメラ)を左右に振り指示者にその旨を伝え、もう一度指示動作を要求する。指示方向を誤認識した場合には、ロボットは誤って認識した指示方向に従って誤動作する。

以上の方法に従って、手招き動作によりロボットをスタート地点から折り返し点を通りゴール地点(スタート地点と同じ)まで誘導する実験を行った。実際に誘導している様子を図5に示す。また、“確からしさ”の閾値を $T = 0.3, 0.9$ とした場合の指示結果を表6に示す。

閾値が0.3の場合、誤認識を1回しているが、合計の指示回数は0.9の場合と比較して少なくなる結果となった。この結果だけからは一概にいえないが、このように信頼性を下げて常に何らかの認識結果を出力するようにした方が、かえって全体としてのパフォーマンスが向上することもある。しかし、誤認識が許されないようなタスクにおいては、信頼性をより重視したインタフェースが必要となる。

7. むすび

本報告では、手招き動作における指示方向と言葉の発声タイミングとの関連性に注目し、音声をトリガとして用いることで手招き動作の指示方向を自動認識する手法を提案した。本実験で得られた結果は次のようにまとめられる。

- (1) 人が自然に行う手招き動作では、その発声タイミングと指示方向との間に大きな相互関係がある。
- (2) 手招き動作において、発声時の手の移動方向を測定することにより、不特定の動作者に対して高い認識率(実験では80%程度)で指示方向を推定できる。
- (3) 認識結果に“確からしさ”という指標を導入することにより誤認識率を低減(実験では2%以下)することができる。

身振り動作と言葉とのタイミングに注目したこのアプローチは、人間同士の自然な身振り動作を的確にモデル化している。今後は、音声認識技術も積極的に取り入れることにより、様々なアプリケーションに適用可能なマンマシンインタフェースの構築を目指す。

〔文 献〕

- 1) 鳥毛明, 河野哲也: “画像処理を用いた手招き動作のリアルタイム認識”, 計測自動制御学会論文集, **29**, 9, pp. 1064-1069 (1993)
- 2) T. Kiriki, Y. Kimuro, and T. Hasegawa: “A 4-Legged Mobile Robot Control to Observe a Human Behavior”, Proc. IEEE Int. Workshop on Robot & Human Interaction, pp. 195-200 (1999)
- 3) 森田俊彦, 沢崎直之, 内山隆, 佐藤雅彦: “カラートラッキングビジョン”, 第14回日本ロボット学会学術講演会予稿集, pp. 279-280 (1996)
- 4) 西川敦, 大西映生, 西村正典, 平野敦士, 小荒健吾, 宮崎文夫: “局所相関演算に基づくオプティカルフローを用いた身振り動作の認識手法”, 情処学論, **40**, 8, pp. 3118-3133 (1999)
- 5) 松本吉央, 坂井克弘, 稲田哲也, 稲葉雅幸, 井上博允: “PCベースのハイパーマシン: 知能ロボットの汎用カーネル”, 第15回日本ロボット学会学術講演会予稿集, pp. 979-980 (1997)
- 6) 西川敦, 宮崎文夫: “オプティカルフローを用いた身振り認識と移動ロボット誘導への応用”, 画像ラボ, **11**, 2, pp. 21-27 (2000)

にしむら まさのり
西村 正典 1998年, 大阪大学基礎工学部機械工学科卒業。2000年, 同大学院基礎工学研究科システム人間系専攻修士課程修了。現在, 松下通信工業(株)に勤務。在学中は、動画像を用いた身振りインタフェースの研究・開発に従事。

にしかわ あつし
西川 敦 1995年, 大阪大学大学院基礎工学研究科物理系専攻修士課程修了。同年, 米国南カリフォルニア大学客員研究員。1996年, 大阪大学基礎工学部助手。1997年, 同大学院基礎工学研究科助手となり、現在に至る。ロボットの視覚、ヒューマンロボットインタフェースに関する研究に従事。博士(工学)。

こあら けんご
小荒 健吾 1997年, 大阪大学基礎工学部機械工学科卒業。1998年, 同大学院基礎工学研究科システム人間系専攻修士課程修了。現在, 同大学院基礎工学研究科システム人間系専攻博士課程に在学中。動画像からの人間の身体運動の計測と認識に関する研究に従事。

みやざき ふみお
宮崎 文夫 1979年, 大阪大学大学院基礎工学研究科博士課程中途退学。同年, 大阪大学基礎工学部助手。1986年, 助教授。1991年, 同教授。1997年, 同大学院基礎工学研究科教授となり、現在に至る。この間1987~1988年, 米国カリフォルニア大学客員準教授。ロボットの知能化に関する研究に従事。工学博士。