I-075

プロアクティブヒューマンインタフェースのための ジェスチャの早期認識に関する検討
Gesture recognition for proactive human interface

内田 誠一† 倉爪 亮† 谷口 倫一郎† 長谷川 勉† 迫江 博昭†
Seiichi Uchida Ryo Kurazume Rin-ichiro Taniguchi Tsutomu Hasegawa Hiroaki Sakoe

1. まえがき

筆者らが検討しているプロアクティブヒューマンインタフェースとは，次の２つの性質を持った人間-コンピュータシステム間のインタフェースである．

1. 実体を伴い，人間に対して物理的な働きかけが可能
2. 人間の行動意図を推定・予測して先回りする機能を具備

これら 2 性質はいずれも，様々な人にとって使い易いコンピュータシステムの枠組みの提供することを目的としている．性質 1 はインターフェースに身体性 [1] を持たせることで実世界との乖離が少ない状況での情報授受を目指すものである．一方，性質 2 の予測に基づく先回り機能は本研究の独創的な点の１つであり，本報告では特にこの性質に関して議論する．

予測に基づく先回りの効果を図 1 に示すプロアクティブインタフェースの一形態において例示する．これは円滑な遠隔コミュニケーションを目指したインタフェースである．ヒューマノイドがインタフェース本体であり，一方のユーザの行動 (ジェスチャ) を模倣によりもう一方のユーザに伝達する．このインタフェース系における予測に基づく先回り処理およびその利点は以下の通りである：

- 予測結果に従ってヒューマノイド側の動き制御を早期に開始させる．この先回り処理により，ヒューマノイドによる動作表現の時間遅れを低減できる．
- ユーザがジェスチャを途中で止めても，予測に基づいてヒューマノイドに引き続き動作を継続させる．これも一種の先回り処理であり，その結果，ユーザがジェスチャ全体を提示する必要が無くなる．

本報告では，これら先回り機能の実現手段として，ジェスチャの早期認識法を提案する．この早期認識とは，例えば両手が上がり始めた段階でそれがジェスチャパターン「万歳」の冒頭部であると認識するような処理である．この早期認識により今後両手を頭の位置程度まで上げると予測でき，それに応じて先回り処理ができる．

本手法では，従来のジェスチャ認識の枠組みに新しく論理的な評価基準を導入することで早期認識を実現する．後述するように，従来法ではジェスチャを開始して認識結果を得るまでに少なくとも標準パターン長の半分程度の時間遅れが発生する．この遅れはプロアクティブインタフェースの先回り機能実現の障害となる．そこで本手法では，現時点付近の入力パターンが唯一のジェスチャに特定できるか否かを論理的基準により評価する．特定できる場合には直ちに認識結果を出力することで，こうした遅れを最小化する．

†九州大学, Kyushu University

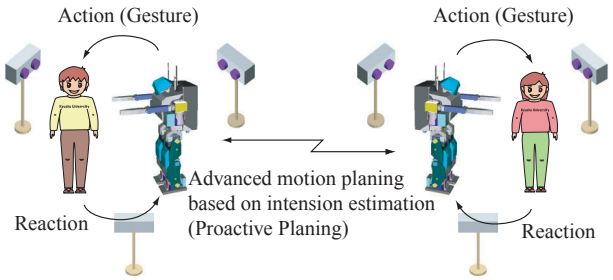

図 1: 人間型プロアクティブインタフェースによる遠隔地コミュニケーション

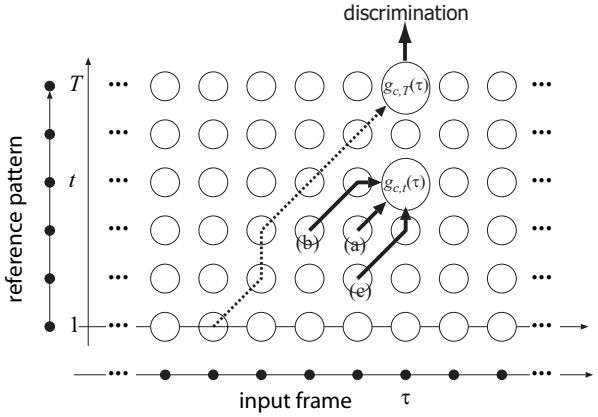

図 2: 連続 DP によるジェスチャのスポッティング認識

本手法は，DP マッチング法の一種である連続 DP[2] に基づく，ジェスチャのスポッティング認識法である．連続 DP はもともと音声認識用に提案されたが，ジェスチャ認識にも応用されている [3, 4, 5, 6]．ただし早期認識について特に注意した手法は無い．DP の確率的拡張である HMM もジェスチャ認識に利用されている (例えば [7])．しかし，HMM は基本的に始端 (もしくは終端) 固定であり，スポッテイング認識に使うためには，事前のセグメンテーションやジェスチャ無し区間のモデル化などが必要となる点で不利である．

2. ジェスチャの早期スポッティング認識

2.1 従来のスポッティング認識法

本節では，本手法の基礎となっている連続 DP[2] によるジェスチャのスポッティング認識について述べる．この連続 DP とは，始点・終点自由型の DP マッチングを

継続的に行う手法である．あらかじめジェスチャをセグメンテーションしておく必要がなく，さらにフレーム同期的な処理が可能であるため，実時間処理に向いている．

具体的には，従来法では以下の手順でスポッティング認識を行う．第 c ジェスチャの標準パターンを特徴ベクトルの時系列 $\boldsymbol{R}_1^c, \ldots, \boldsymbol{R}_t^c, \ldots, \boldsymbol{R}_T^c$ で表し，入力パターン (複数のジェスチャの時系列) も同様に $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_\tau, \ldots$ で表す．このとき従来法では，フレーム τ において，すべての t, c について次式に従い累積コスト $g_{c,t}(\tau)$ を計算する (図 2)．

$$g_{c,t}(\tau) = \min \begin{cases} g_{c,t-1}(\tau-1) + 2d_{c,t}(\tau) & \cdots(\mathrm{a}) \\ g_{c,t-1}(\tau-2) + 2d_{c,t}(\tau-1) + d_{c,t}(\tau) & \cdots(\mathrm{b}) \\ g_{c,t-2}(\tau-1) + 2d_{c,t-1}(\tau) + d_{c,t}(\tau) & \cdots(\mathrm{c}) \end{cases} \tag{1}$$

ここで $d_{c,t}(\tau)$ は現フレーム τ が標準パターン c のフレーム t に対応づけられた時の局所コストであり，$d_{c,t}(\tau) = \|\boldsymbol{R}_t^c - \boldsymbol{I}_\tau\|$ で定義される．記号 (a)-(c) については後述する．

累積コスト $g_{c,T}(\tau)$ はフレーム τ を終点とした場合の標準パターン c のコストになる．従って，現フレーム τ での認識結果 c^* は，

$$c^* = \operatorname*{argmin}_c g_{c,T}(\tau) \tag{2}$$

で与えられる (実際には経路長による正規化処理 [2] が入るが説明は省略する)．以上により入力パターンの区間 $[\tau', \tau]$ が標準パターン c に対応したというスポッティング認識が可能となる．なお，始点 τ' $(1 \leq \tau' < \tau)$ は，バックトラック処理を別途行うことで明示的に求められる．漸化式 (1) ならびに式 (2) は入力パターンのフレーム τ に同期して計算できるので，認識結果 c^* を時々刻々と出力できる．

図 2 の破線で示したマッチング経路からもわかるように，従来法では原理的にジェスチャ開始後 $T/2$ フレーム程度経たないと認識結果は出力されない (時刻 τ において発生したパスは，$\tau + T/2$ にならないと $t = T$ に到達しない)．すなわち必ず一定時間の遅れが発生する．このように，従来法は実時間処理に適した枠組みであるものの，早期認識が必要となるプロアクティブインタフェースにおいて用いるにはまだ不十分であると言える．

2.2 早期認識の実現

ジェスチャ開始後 $T/2$ フレーム経過していない過渡的な時点では，先行するジェスチャの混入などによりコスト $g_{c,t}(\tau)$ は一般に大きな値となる．早期認識とは，こうした過渡的なフレームにおいても尤もな認識結果を出力することと言える．本手法では，過渡的フレームにおいても，考え得るジェスチャが唯一に限定できれば，すなわち曖昧性がなければ，そのジェスチャを認識結果として出力する．

この曖昧性の有無を判定するために，DP 漸化式 (1) と同期して，次の論理値 (以下，サポートと呼ぶ)$h_{t,\tau} \in \{0, 1\}$ を補助的に計算する．

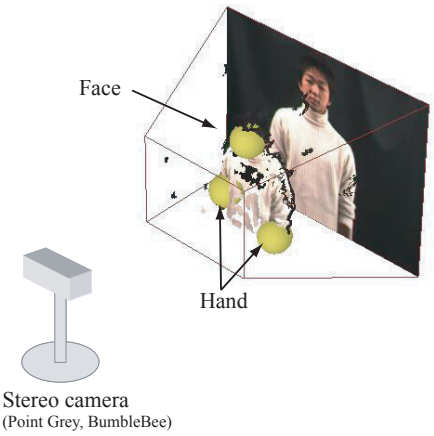

図 3: ステレオ視と肌色検出による両手と顔の 3 次元位置計測

$$h_{c,t}(\tau) = \begin{cases} h_{c,t-1}(\tau-1) \cap q(d_{c,t}(\tau)) & \text{if (a)} \\ h_{c,t-1}(\tau-2) \cap q(d_{c,t}(\tau-1)) \cap q(d_{c,t}(\tau)) & \text{if (b)} \\ h_{c,t-2}(\tau-1) \cap q(d_{c,t-1}(\tau)) \cap q(d_{c,t}(\tau)) & \text{if (c)} \end{cases} \tag{3}$$

記号 (a)-(c) は式 1 中の同じ記号に対応し，$g_{c,t}(\tau)$ の計算時にどの経路が選択されたかを示している．また $q(x)$ は x がある値 θ_1 以下 (以上) の時 1(0) を返す関数である．従って，$h_{c,t}(\tau)$ が 1 であるということは，$g_{c,t}(\tau)$ に至る最適経路上のすべての状態において標準パターンとほぼ整合していることを意味する．サポートは累積コスト $g_{c,t}(\tau)$ よりも厳しい基準にある．この厳しい基準の下で，もし現フレーム τ でこのサポートを 1 とできる標準パターン c が唯一であれば，曖昧性なく現在のジェスチャは c であると判断できる．以上を手順としてまとめると以下のようになる．

1. まず，フレーム τ の認識コスト $\min_c g_{c,T}(\tau)$ がある値 θ_2 以下であった場合，そのコストによる認識結果 (2) を信頼する．そうでない場合，過渡的なフレームと判断し，次のステップに進む．
2. そのフレーム τ の全てのサポート $\{h_{c,t}(\tau) \mid \forall t, \forall c\}$ を調査する．その結果，もし値が 1 のサポートが唯一の標準ジェスチャパターン c のみに存在した場合，曖昧性がないとして，現在の認識結果をそのジェスチャ c で置き換える．

3. 実験

本手法の先回り機能，すなわち本手法の早期認識能力を確認するために，予備的な実験を行った．

3.1 認識タスク

想定したジェスチャは「さようなら (bye)」「万歳 (banzai)」「指差し (point)」の 3 種である．「さようなら」は

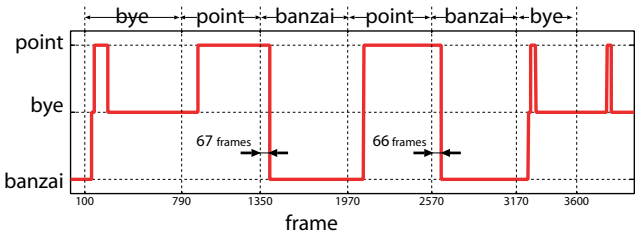

図 4: 本手法によるジェスチャ認識結果

右手を顔付近の高さに上げた後に左右に振る動作,「万歳」は両手を頭の高さまで同時に上げたり下げたりする動作,「指差し」は右手を顔付近の高さに上げた後に前後に振る動作である．このように「さようなら」と「指差し」の前半は同じ動作となっている．なお，いずれも開始・終了時には両手を下げた状態としたが，本実験ではこの事実を積極的に（例えばセグメンテーションに）使ってはいない．

まず 3 種のジェスチャのそれぞれについて，パターンを 4 個から 8 個程度，次節 3.2 で述べる方法で取得した．その後，各パターンの冒頭 300 フレーム分を標準パターンとして登録した (すなわち $T = 300$)．

認識対象である入力パターンも同様に次節 3.2 の方法で取得した．この入力パターンは複数のジェスチャからなる時系列である．ジェスチャ間の両手を下げた状態の時間は短いので，連続ジェスチャ列と言える．

3.2 ジェスチャデータの取得

認識の対象としたジェスチャデータは，人間の右手先および左手先の 3 次元位置からなる 6 次元特徴ベクトルの時系列として表現される．このデータは，(i) まずユーザの前方に置かれたステレオカメラ (Point Gray 社製, BumbleBee) により距離画像を計測 (30 フレーム/秒) し，(ii) 次に肌色検出により両手と顔部分を同定することで自動取得した．図 3 にその様子を示す．なお，アバタならびに実際にヒューマノイドを駆動することで取得したデータを観察した結果から，以上の手法で比較的精度よく 3 次元位置計測ができることがわかっている．

3.3 認識結果

図 4 は本手法によりある連続ジェスチャ系列を認識した結果である．図には手動で付与した各区間の正解ジェスチャも示している．

注目すべきは「万歳」のジェスチャが早期認識できている点である．具体的には，1 回目および 2 回目の「万歳」がジェスチャ開始後それぞれ 67 フレームおよび 66 フレームで検出されている点である．この結果は，認識の遅れを標準パターン長 $T = 300$ の半分よりも大幅に少なくできたことを意味している．これに対し，サポートを用いずコスト $g_{c,T}(\tau)$ によって (すなわち (2) に基づいて) 認識を試みた場合，この遅れは 1 回目および 2 回目の万歳でそれぞれ 118 フレーム，108 フレームであった．

一方,「指差し」については，ジェスチャ開始後，100 フレーム程度以上経ってから正しく認識されており，遅れが低減されていないことがわかる．ところで，3.1 節で述べたように,「指差し」と「さようなら」のジェスチャの冒頭部分は共に右手を顔の高さまで上げる動作であり，この段階では両者は本質的に区別できない．従って，本手法に限らず，いかなる方法においても早期認識を行うことは不可能であり，そのことが結果に現れていると言える.「さようなら」についても，冒頭部が「指差し」に誤認識された後に遅れて正しく認識されるのは同じ原因と考えられる．なお，曖昧であってもこれら 2 つのジェスチャのいずれかであると限定できていれば，右手を顔の高さまで上げる時点までは予測できていることになる．ヒューマノイドの動作表現の時間遅れを低減したい場合については，このような比較的短期の予測で十分な可能性もある．

4. まとめ

本報告では，プロアクティブヒューマンインターフェースにおける先回り処理の実現のために，早期認識が可能な行動（ジェスチャ）認識法を検討した．この早期認識とは，ジェスチャの開始から認識結果が出力されるまでの遅れを極力小さくすることに相当する．従来法では標準ジェスチャパターンの時間長の半分程度の遅れが生じる．これに対し本手法は，論理判定機構を新たに導入することで，現時点でのジェスチャの曖昧性の有無を判定し，それを用いて早期認識を行う．例えば，曖昧性が無い，すなわち現時点のジェスチャを特定してよいと考えられる場合，直ちにそのジェスチャを認識結果とする．この論理判定機構は，従来よりジェスチャのスポッティング認識に多用されている連続 DP 法に，容易に組み込むことが可能である．予備的な認識実験により，曖昧性がない場合については，従来法に比べてジェスチャを早い段階で正しく認識できるという，所期の効果が得られたことを確認した．

今後は，より大規模なタスクを用いて本手法の評価実験を行うとともに，早期認識結果によるインタフェースの駆動実験を行い，先回り処理の効果や誤認識と曖昧性の影響についても検証したいと考えている．

謝辞 本研究の一部は総務省戦略的情報通信研究開発推進制度の支援を受けた．

参考文献

[1] 岡田, 三嶋, 佐々木 (編), 身体性とコンピュータ, 共立出版, 2001.

[2] 岡, “連続 DP を用いた連続単語認識,” 日本音響学会音声研究会資料, S78-20, 1978.

[3] 高橋，関，小島，岡，“ジェスチャー動画像のスポッティング認識,” 信学論, vol. J77-DII, no. 8, pp. 1552–1561, 1994.

[4] 太田, 潮崎, 新井, “動的計画法に基づくマッチングによる運動認識,” 精密工学会誌, vol. 63, no. 6, pp. 812–818, 1997.

[5] 西村, 古川, 向井, 岡, “時系列パターン検索のための重み減衰型 Reference Interval-Free 連続 DP について,” 信学論, vol. J81-DII, no. 3, pp. 472–482, 1998.

[6] 呉, 木戸, 塩山, “ジェスチャ認識のための連続 DP の改良,” システム制御情報学会論文誌, vol. 14, no. 6, pp. 283–290, 2001.

[7] 山田, 山本, 酒井, 森園, 梅谷, “メンテナブルな人間／ロボット共存システムによるヒューマン・エラー・リカバリー,” 日本ロボット学会論文誌, vol. 21, no. 4, pp. 86–92, 2003.