ジェスチャ動作の動画像からのスポッティング認識について

岡 隆一[†,☆] 西村拓一[†,☆☆] 矢部博明[†,☆☆☆]

人間のジェスチャ動作をビデオカメラで動画像としてとらえ，その意味するカテゴリ名をコンピュータで同定する手法を述べる．この手法には複数あるが，それらはジェスチャが行われる様々な状況に応じて個々に用いられる．状況の違いとは，ジェスチャが１人の人物によってなされているか，複数人でなされているか，戸惑いをともなうジェスチャであるのか，あるいは身体の全体を使う動作であるか，などである．さらに，ジェスチャを行う人物がカメラの正面に向いているかどうかも状況の違いとなる．本稿では，ジェスチャ認識の結果を，他のメディアである音声の認識結果と組み合わせて，CGや音声合成でユーザと対話する実時間システムを構成するためのアーキテクチャについても述べる．

On Spotting Recognition of Gesture Motion from Time-varying Image

Ryuichi Oka,[†,☆] Takuichi Nishimura[†,☆☆] and Hiroaki Yabe[†,☆☆☆]

This paper describes some methods for recognizing human gestures from a time-varying image captured by a single or multiple video cameras. Each method is suitable to recognize human gestures performed in a different situation. The situations include the case of a single person facing a camera and the case of multiple persons captured by an omni-view camera and so on. The paper describes an architecture to realize a real-time dialogue system consisting of speech recognition, task model, CG output and speech synthesis output modules which cooprerate with gesture recognition module.

1. はじめに

人間にやさしいコンピュータの開発が望まれている．この開発の一環として，人間にとって負担のより少ない形でコンピュータと対話できるシステムの実現に向けて多大の研究努力が続けられている．本稿での主題である人間のジェスチャ動作の認識は，コンピュータ側で実現してほしい機能であり，その実現によって人間とコンピュータの間でのより親和性のある対話システムの構築が期待できる．そこで，まず人間とコンピュータの対話の全体像を概観し，その中でジェスチャの位置を特徴付けたい．

対話する人間とコンピュータの間は物理的に離れており，対話を媒介する情報（これを「対話メディア」あるいは単に「メディア」と呼ぶ）はこの物理的に離れた空間を，光や音などを介して伝達される．この対話場面で，人間は生物として視聴覚という生の感覚でメディアを知覚し，また身体や発話機能によってメディアを生成する．そのとき人間は，通常，目的達成のための対話以外に余分と思われる負荷の少なさによって，対話の心地よさや便利さや自然さを判断することになる．この「余分の負荷」には，たとえばキーボードを打つという動作をすることや，身体的な動きで十分意思を表現するほうが適切である場合に言語的な表現を強制されること，などがある．また，音声発声で応答するときの語彙の制約や発声タイミングへの制約，ジェスチャで応答するときの語彙の制約，ジェスチャ動作を行う際の動作可能範囲の物理的な制約などもそれに含まれる．これらの「余分な負荷」がどの程度少なくなっているかに，対話システム構築のための技術の高度さが判断される基準が適用される．また，この高さを獲得することが，今後の技術の発展の方向につながるといえる．

さて，上記の「対話メディア」にはどんなものがあるかをあげよう．それらは，「文字」，「音声波形」，「静止画・動画」の３つとしてよい．いずれも人間側は，

† 新情報処理開発機構
Real World Computing Partnership
☆ 現在，会津大学
Presently with University of Aizu
☆☆ 現在，産業技術総合研究所
Presently with Advanced Industrial Science and Technology
☆☆☆ 現在，シャープ株式会社
Presently with Sharp Inc.

これらのメディアの生成段階でなんらかの意味を込めることになり，他方コンピュータ側も内部処理の結果として，対話の進行のためにこれらのメディアを通じて人間側に意味を伝える．

対話における人間側の「文字メディア」の扱いは，自らが出力するものとしてはキーボードを「打つ」ことであり，入力されるものとしては，コンピュータの画面に表示された文字を「読む」ということであり，いずれもきわめて知的な処理が要求されている．コンピュータ側の扱いは，人間側の入出力時点においてなされるきわめて知的な処理の結果，コンピュータ内に取り込まれた「文字表記データ」の利用がきわめて容易にできることになる．対話における「音声波形」の人間側の扱いにおいては，出力である音声発声という作業は知的な動作であるが，その表現が物理的波形であるということで，それを受け取るコンピュータ側にとっては，波形から文字へという変換を行うことが（現在の記述水準では）困難な課題となっている．一方，コンピュータ側では，受け取った波形を部分的に文字に変換したとしても，対話を自然に行わせるテキスト（記号列）の生成は困難な課題であるが，いったん文字になった段階からの音声波形への変換（合成音声）は比較的簡単である．人間側による合成音声の認知は，知的作業であるが容易に実行できる．

「静止画・動画」のメディアは，対話においてはどのような扱いになるであろうか？ 人間側では，静止画メディアの利用には，たとえば，図形を描くことでその意味されるものを表現することができる．一方，コンピュータ側では，その意味されるものを記号化することは（現在の技術水準では）困難である．コンピュータ側では，記号ラベルと静止画像との対応がついていれば，記号から静止画を検索し，それを人間側に表示することは容易である．人間がその提示されたものを理解することも容易である．動画は，人間のジェスチャや身体動作をカメラでとらえることで「対話メディア」として扱われることになる．コンピュータ側は，記憶した動画像を再生したり，人間の身体動作を模擬したりできるハードウェアを持つことにすると，対話の応答をそのハードウェアを使って人間と類似したジェスチャで表現することもできる．人間側でのその理解は比較的容易である．

さて，対話メディアとなっている人間のジェスチャ動作を観測した動画像から，人間のジェスチャで意味されるものをコンピュータ内で記号で記述する（ジェスチャ認識）ことを考える．このとき，コンピュータ内の記号処理と，人間のジェスチャで示された意味を，より直接的に結び付けることを容易ならしめるということで，ジェスチャ認識技術は意味を持つことになる．対話メディアの動画像で表現されたジェスチャは，その表現自体が記号ではなく，またその意味されるものもすべて記号で表現し尽くすことも不可能であるので，「ノンバーバル」なメディアとして，その特徴を積極的に利用するということも考えれる[1),2)]．すなわち，ジェスチャ認識は，カテゴリ記号を同定するだけでなく，ノンバーバルな情報として，たとえば「こんな形」というように明確にカテゴリできないものや,「この程度」という程度表現や，様々な「OK」という意思表現の同定を含んでいる．さらに，ジェスチャには，たとえば，その意味される内容には「戸惑い」があると見てとれるような場合，そのノンバーバルな内容を取り出すことも課題とすることができる．これらの様々なノンバーバルな情報もジェスチャ認識の一部をなしている．

このような意思表現手段としての特徴を持つジェスチャの認識の応用場面としては，たとえば，複数の人が同時に意思表現（賛成や反対など）をしている状況の把握，言葉で明確に言うこともないような意思（もちろん賛成ですよ，などの同意のうなずきなど）の把握，音声で話している内容の強調の度合い（断固としてそれを言っているのかなど）の把握，などに用いるとその有効性が出てくる．これらは他のメディアでの把握が困難であるといえよう．

上記のような役割を持つジェスチャなどの身体動作を，コンピュータと人間の対話システムの中で，マルチモーダル情報の表現手段の１つとして位置付ける研究がさかんになってきている[3)~7)]．音声などと相まってマルチモーダルな情報による対話システムは，コンピュータ利用の標準的な機能として近い将来，電子秘書，共同作業支援システム，ゲームなどに利用されることになると思われる．通常の場合，ジェスチャ動作の認識とは，１つの動作を表すラベルが１つの区間動画像に対応するとき，未知の動画像系列のなかにそれを見出すという定義がされている．ジェスチャ動作を表す特徴パラメータの時系列を動画像から抽出する以外に，データグローブを手に装着して，これによる測定データを特徴パラメータ時系列とする方法もある[8)]．しかし，本稿では，ジェスチャ動作を動画像として観測されたものから認識する方式のみについて述べるものとする．その理由は，ビデオカメラによるジェスチャの観測は最も簡便な方法であり，これによって認識できるものが，ジェスチャ認識研究の最も基本となる方法として確立されなければならないと考るからで

図 1 パソコンのカメラと対面して人物のジェスチャ認識．パソコン画面の右上に小型のビデオカメラがついている．下図はコンピュータに指示できるジェスチャの種類の例を示す[9]．それぞれの指示内容を表すジェスチャは区間動画像であるが，ここではその中の特徴的な 1 フレーム画像のみを示している．

Fig. 1 Gesture recognition for a person facing a single camera attatched a note personal computer. The categories of gesture are shown at the bottom of figure. Each gesture category is corresponding an interval of motion image. A single frame of image in the interval is shown at the figure.

ある．

人間のジェスチャなどの身振りや動作をコンピュータで認識処理するとき，最も問題となるのは人間の動作をする場所と，それをコンピュータへ取り込むためのカメラの種類とそれが置かれる位置との関係である．たとえば，カメラと人物の位置関係によって，人物の動きの範囲や動作の種類が制約されるからである．この制約によって，ジェスチャによる対話の内容も異なってくるといえる．本稿では，はじめに上記の位置関係について 3 つの場合に分け，身振りや動作の認識システムを述べる．次に，あるジェスチャに「とまどい」があるときのジェスチャの変動にどう対処すればよいかなどについて述べる．最後に，ジェスチャ認識や音声認識を統合して実時間で応答する対話システムを，たとえば，ノート型のパソコン上で簡単に構築するためのアーキテクチャについて述べる．

2. カメラと向き合うジェスチャ動作

ノート型のパソコンなどに装着できるカメラが市販のものである場合，カメラは通常固定されていて，そ

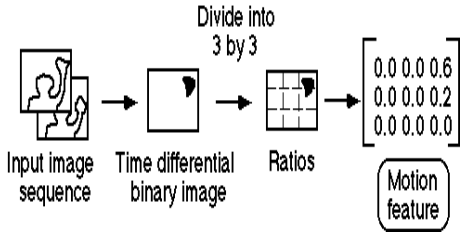

図 2 ジェスチャ認識のための動画像からの特徴抽出法．ジェスチャ認識処理にかかる動画像のフレーム特徴は 9 次元のベクトルでよいことに注目されたい．

Fig. 2 The fetaure extraction process used in our gesture recognition system. The 9 dimensional vector of feature is enough to represent a frame time of motion image for gesture recognition.

のカメラでとらえることのできる視野も限定されている．したがって，許される人間の動作はその固定カメラの視野の中に入ることが前提となる．その視野の中に入る場合は，動作者をトラッキングすることもできる．固定カメラを前提にする場合，**図 1** の上図に示されるようにカメラと向きあってのジェスチャが認識処理されることになる．このとき，認識の対象とされるジェスチャの種類については，図 1 の下図に示されているカテゴリとなることが通常であり，種類の数としては 40 程度となる[9]．

さて，ここで用いている動画像からの特徴について述べる．**図 2** に示すように明暗動画像のフレーム間での差分 2 値画像を作成し，全体を 3×3 へと領域を分割し，各領域における 2 値の Pixel の割合を計算して，結果として 9 次元のベクトルをもって 1 フレームの特徴としている[10]．この 9 次元の特徴ベクトルを作成するに必要なもとのフレーム（明暗レベルを持つ）の分解能がどの程度かを実験的に調べると，12×12 程度あれば，先に示した 40 カテゴリ程度のジェスチャを対象にして認識性能を落とさないということが判明している．この結果から 2 つのことがいえる．1 つは，動画像系列でジェスチャを認識するために必要な 1 フレームの画像から抽出すべき特徴の量はきわめて少ないもので十分であるということ，2 つは，フレームの特徴量の少なさは，いまの場合，認識のロバスト性を獲得するために有利に作用している，ということである．後者についていえば，ジェスチャを行っている人物の画像中での位置と，参照パターン中での人物の位置の違いで生じるシフト・ノイズについて，この 3×3 に圧縮したものがほぼ同じであれば許容されることになる．この許容の程度は，もとの画像フレームの Pixel 分解能に依存するが，実験的に 12×12 までは認識率の低下がほとんどないということから，9 分割の各領

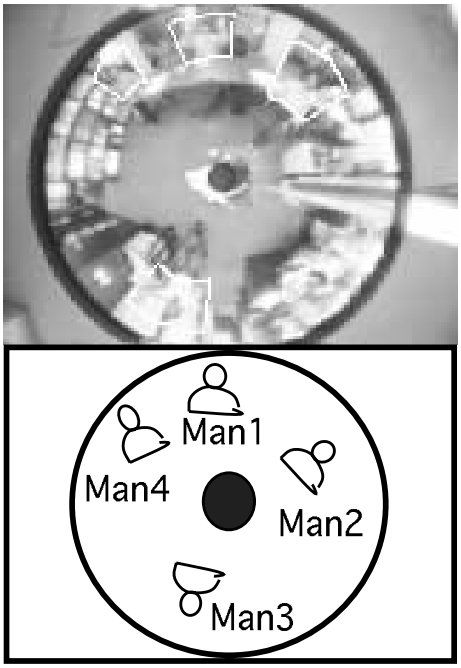

図 3 4 人の人物動作をとらえた HyperOmni Vision[11]の画像（160 × 120）．中心にビデオカメラがあり，その周りを人物が取り囲んでいる．カメラから人物が遠くに離れると 1 人あたりの画像部分は少なくなる．上図に写っている人物は 4 人であるが，それぞれの位置をはっきりと示すと下図のようになっている．

Fig. 3 Four people in the motion image by the called HyperOmni Vision. Each frame image has 160 × 120 pixels. People are around the camera. In proportion to the distance between the camera and a people, the size of image occupying a people reduces. The clear boundary lines to show the four people is shown in the bottom figure.

域では 4 × 4 の Pixel 領域における 2 値化差分特徴の占める割合が変化しない程度の許容性があるということになる．後述するように，この各領域における割合の値と入力と参照パターンのフレーム特徴間の距離の累積値によって認識が行われるので，認識のロバスト性はこの累積値に基づくことになる．

3. カメラの周りの多人数のジェスチャ動作

近年，ロボットの視覚センサとして開発された全方位視覚センサ（OMNI VISION）[11]がジェスチャ認識システムにおける観測装置として使用されてきている[10]．この視覚センサを使うかどうかは，ジェスチャを同時に何人のユーザで，またどれだけカメラと離れてジェスチャを行ってよいかという条件によって決められることができ，インタフェースの設計上きわめて興味深い．

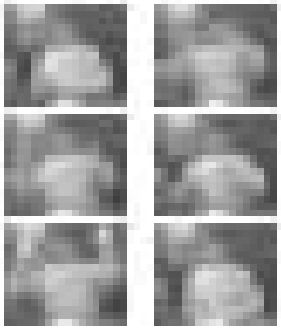

図 4 図 3 の枠内でとらえる動画像の例．ジェスチャ “バンザイ”（3 フレームごと，18 × 15）はこの程度の分解能となる．この程度の分解能からのフレーム特徴でも認識の性能の低下はほとんどない．

Fig. 4 A sample sequence of motion image at a small area of figure 3. The sample frame of “BANZAI” gesture has a small size of pixel image 18 × 15 which is enough to be correctly recognized.

この視覚センサの利用における条件の例を考えよう．**図 3** の上図は，視覚センサの周辺に 4 人の人物が写っている状況を示している．図 3 の上図で人物が写っている領域が白抜きの線で示されている．ここで写っている 4 人の人物をより分かりやすく表したものが図 3 の下図である．いま，この 4 人の人物の中で視覚センサから最も離れた人物 1 までの距離は約 4 m で，そのときの人物の画像のサイズ（pixel）は，18 × 15 となっている．最近の研究では，この程度の分解能でジェスチャの観測を行った場合でも動画像からのジェスチャのスポッティング認識ができることが報告されている（**図 4**）[10]．したがって，カメラでとらえられた動画像が 256 × 256 の pixel 画像であれば 1 人の人物が 20 × 20 の pixel 画像に内に入いるとすると，約 12 × 12 = 144 人の人物群が 4 m 程度カメラから離れて並んでいる（階段教室などに）状況でも，それらの人々が同時に異なった動作をしても認識できることになる．また，4 m より近くに人物がいれば，それらの人物が占有する pixel の領域が大きくなるので，同時に認識できる人物の数が比例して少なくなることになる．

このような，カメラによって多人数の動作者のジェスチャを認識できる場合でも，各動作者はカメラの方を正面に見ていることが前提となる．正面を向いている場合の区間動画像が参照パターンとして登録されるからである．また，複数の動作者が移動する場合，原理的にはそれらの複数の人物をトラッキングすればよいのであるが，これには技術的に困難な課題が十分解決されていない状況にある．

4. 自由動作の中からの登録動作のスポッティング認識

現在の技術では，人物が行う任意の動作すべてをコンピュータが認識できる状況にない．そこで，認識の対象としたい動作をあらかじめ登録しておくか，あるいはその場で登録するということをしなくてはならない．登録は，たとえば意味は「ダメ」という 1 つでも，動作が人によって大きく異なる場合などに必要である．さて，登録した動作を認識させようとするとき，登録した動作のみの動作をするというのもユーザにはきわめて不自由なものである．音声であればポーズとしての無音区間をとれば，発話の区切りが表現できるが，動作をカメラで観測されるとき，どこが始まりでどこで終わりであるという情報を静止動作で与えるのは人間にとって不自由である．また，認識させたい動作の前後にその動作につながるいろいろな動作がくることもよくあることである．そこで，人物にはいかなる動作への制約も課さないで，自由に動作をさせ，その自由な動作の流れの中に，登録した動作が出現したときのみ，それを識別して認識するという認識方式を採用すると動作する側はわずらわしさから開放される．このようなことを可能にする認識方式はスポッティング方式といわれるものである．スポッティングとは登録されているジェスチャや動作の時系列パターンの切り出しと認識を同時に実行するものである．このスポッティング法としてよく知られているものが，連続 DP [12),13)]であり，高橋らによって動画像からのジェスチャ認識に最初に応用されて以降様々に用いられている[9),10),14)~20)]．この連続 DP によるジェスチャのスポッティング認識の手法を述べよう．この連続 DP は各種の発展形があり，それぞれ様々な異なった機能の実現に対応している．たとえば，5 章で述べる「戸惑いのあるジェスチャの認識」のための連続 DP や 6 章で述べる「カメラの方向から自由なジェスチャ認識」のための連続 DP などである．

4.1 連 続 DP

入力画像系列に対し同様の特徴抽出処理を施し，あらかじめ作成してある標準パターンとの距離をスポッティング整合方式により計算し，認識結果をフレームごとに出力する．標準パターンは人間の動作を表現するモデルであり，始点および終点の定まった特徴ベクトル系列として表現される．標準パターンはシステムに認識させたい動作の数だけ作成する．その長さはそれぞれ異なる．以上の準備のもとに，スポッティング整合処理を施す．スポッティングのためのマッチング

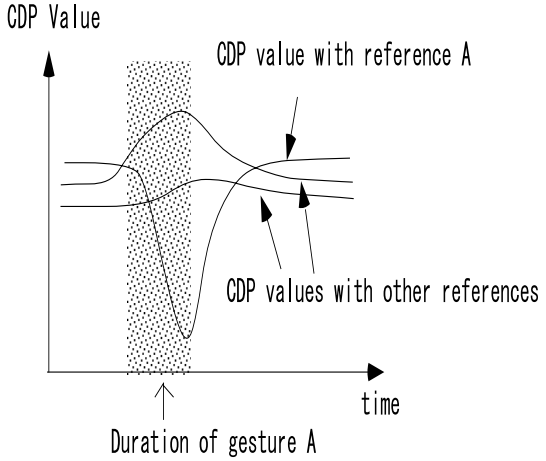

図 5 3 つの登録ジェスチャについての連続 DP の出力特性．一連の勝手な自由動作の中で登録ジェスチャ A がなされたときのみ，その時系列パターンにへこみ（dip）が発生する（図の「Reference A」）．登録以外のジェスチャには応答しないことに都合のよさがある．

Fig. 5 The property of three CDP output streams each of which is corresponding to a stored reference sequence of a gesture category. A local dip appears only when the one of registered gestures reaches the end of gesture in an eldless stream of movement including gestures.

とは，始終点の定まっていない系列のある時点がもう一方の系列の終点に対応すると仮定し，それ以前の部分の最適対応を求める方法で，系列パターンとその区間の判定を同時に行うことができる．スポッティング認識を実現する具体的アルゴリズムとして連続 DP [12),13)]を用いたマッチングを用いる．連続 DP の出力は登録ジェスチャの個数の時系列となっている．たとえば 3 つのジェスチャを登録したとき，**図 5** のように 3 つの時系列が得られる．そのとき，ユーザが登録ジェスチャと類似のジェスチャをし終わったときのみ，該当する登録ジェスチャに対応する連続 DP の値が下がり終わり上昇に転じる．そこで，この dip を検出すれば該当ジェスチャの認識ができることとなる．連続 DP の値はたえず出力しているので，ジェスチャはいつ始めてもいいし，またいつ終わってもよい．さらに登録ジェスチャ以外のものを入力しても反応，すなわち dip ができないのでそれらは無視される．これらのことがユーザに自由動作を許し，自由動作に挟まれた登録ジェスチャのみが識別されるのでユーザにとって便利であることになる．

1 つの標準パターン Z を，標準動作をとらえた T フレームの動画像から得られる特徴ベクトル z_τ の系列

$$Z = \{z_\tau | 1 \leq \tau \leq T\} \quad (1)$$

で表す．ここで，特徴ベクトル z_τ はその次元数を N

とすると

$$z_\tau = (z_\tau(1), z_\tau(2), \cdots, z_\tau(N)) \quad (2)$$

である．入力画像からも同様な特徴ベクトル系列 $u_t(0 \le t < \infty)$ が連続的に得られる．このとき，u_t と z_τ との局所距離を $d(t,\tau)$ と表記する．この $d(t,\tau)$ の定義の一例を以下に示す．

$$d(t,\tau) = \frac{1}{N}\sum_{k=1}^{N}(u_t(k) - z_\tau(k))^2. \quad (3)$$

ここで，入力，標準パターンの時間軸をそれぞれ t，τ と区別する．

さらに，点 (t,τ) を終点とした標準パターンと入力系列との累積距離を $S(t,\tau)$ で表す．連続 DP では $S(t,\tau)$ を以下のような漸化式で更新する．

Initial Condition：

$$S(-1,\tau) = S(0,\tau) = \infty \quad (4)$$

Iteration (t=1,2,...)：

For $\tau = 1$

$$S(t,1) = 3 \cdot d(t,1) \quad (5)$$

For $\tau = 2$

$$S(t,2) = \min \begin{cases} S(t-2,1) + 2 \cdot d(t-1,2) + d(t,2) \\ S(t-1,1) + 3 \cdot d(t,2) \\ S(t,1) + 3 \cdot d(t,2) \end{cases} \quad (6)$$

For $3 \le \tau \le T$

$$S(t,\tau) = \min \begin{cases} S(t-2,\tau-1) + 2 \cdot d(t-1,\tau) + d(t,\tau) \\ S(t-1,\tau-1) + 3 \cdot d(t,\tau) \\ S(t-1,\tau-2) + 3 \cdot d(t,\tau-1) + 3 \cdot d(t,\tau) \end{cases} \quad (7)$$

となり，出力は，

$$A(t) = \frac{1}{3 \cdot T} S(t,T) \quad (8)$$

となる．

ここで，$S(t,T)$ は，

$$S(t,T) = \min_{1 \le \tau \le T,\ \ t \ge \beta(\tau), \beta(\tau+1) \ge \beta(\tau)} \sum_{\tau=1}^{\tau=T} d\left(Z(\tau), f(t - \beta(\tau))\right). \quad (9)$$

という最適値に対応している．連続 DP では，各標準パターンが持つその出力 $A(t)$ 群の中で閾値以下で Dip をなす時刻のものがスポッティング認識されたジェスチャのカテゴリとされる（図 5）．Dip が検出されたときに，その標準パターンと対応するものが入力パターン系列中に区間として決まることになる．認識と入力のセグメンテーションが同時的に行われるという意味でセグメンテーション・フリーの認識，すなわちスポッティング認識が実行されるという．各時刻 t において連続 DP の出力値を与える探索範囲が**図 6** に示してある．ここで縦軸は標準パターンの時間軸を示し，横軸は入力の動画像系列の時間軸を示し，$\tau = 1$ から $\tau = T$ に至るパスが標準パターンと入力パターンとの対応関係を示す．このパス上の局所距離の和が連続 DP の時刻 t における出力値となる．このパスは (t,τ) 平面において，単調増加となるものであり，認識される標準パターンの特徴系列に対応する部分入力パターンの系列の間では伸縮はあっても順序が逆転することはない．このような方式によるジェスチャ認識の性能の例として，**表 1** のものがある．これは図 4 で示された 4 人の人物によるジェスチャの認識結果である．

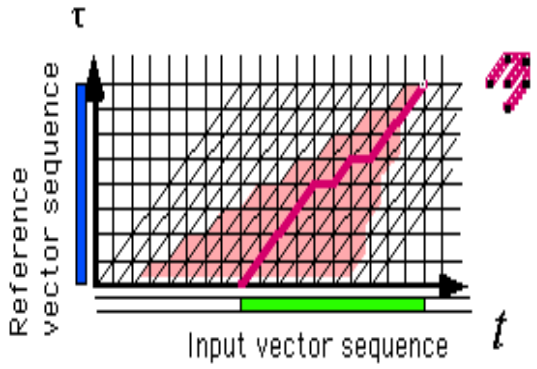

図 6 連続 DP 標準型における探索範囲
Fig. 6 The search area in the time-space plane for obtaining the optimal matching by the standard type of Continuous Dynamic Programming.

表 1 図 4 各人物のジェスチャが 8 つのジェスチャをしたときの認識率
Table 1 Recognition results of 4 people in a motion image.

	Man1	Man2	Man3	Man4
AFC(%)	83	88	82	83

5. 戸惑いのあるジェスチャの認識

人間のジェスチャは，同一動作であっても途中で戸惑ったり考えて止まったりすることがある．この戸惑っている動作は，時と場合によって無数に変化すると考えられる．先に述べた連続 DP 標準型では，このような戸惑っている動作は認識対象としていなかった．すなわち，1 つの標準パターンに時系列として類似しているものとして，非線形の伸縮を許すというものであった．戸惑いのあるジェスチャは非線形の伸縮とい

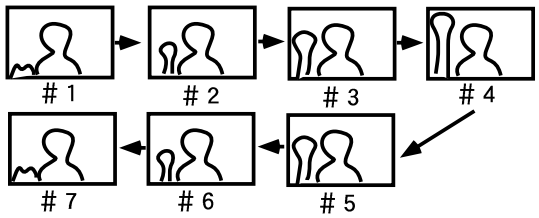

図 7 ジェスチャ“手をあげる”のスナップショット
Fig. 7 Snapshot of gesture “Raise hand”.

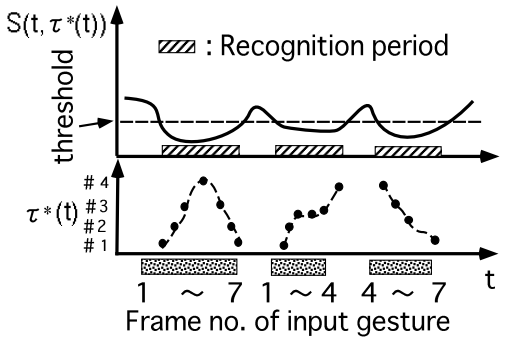

図 8 Non-monotonic 連続 DP による“手をあげる”の部分変形動作 3 種についての認識（標準パターンは図 7 のフレーム＃1～＃4）.
Fig. 8 Recognition of three partial “Raise hand” gestures using Non-monotonic CDP. (Standard pattern is frame #1 to frame #4 of Fig.7).

う範囲を超えるものといえ，それらを連続 DP 標準型で認識しようとすると，多くの標準パターンが必要となり非効率的である．いま，あるジェスチャを 1 つの標準パターンで表すとするとき，そのジェスチャ動作に戸惑いが生じたときを以下のもので表せるものであるとしよう．すなわち，あるジェスチャの標準パターン中のから，任意の部分区間を順方向だけでなく逆方向や静止している動作の組合せで作成されるものであるとしよう．

ここで，本手法の特長を説明するため，1 つ標準パターンを持った認識システムを考える．この標準パターンとしては**図 7** のような 7 枚のフレームからなるジェスチャ“手をあげる”とする．

このとき，ジェスチャ“手をあげる”のフレーム＃1～＃4 を標準パターンとし，フレーム＃1～＃7，＃1～＃4，＃4～＃7 の連続したジェスチャを入力する．すると，入力系列と標準パターンとの累積距離 $S(t,\tau^*(t))$ は，**図 8** のように変化する．Non-monotonic 連続 DP では，累積距離 $S(t,\tau^*(t))$ があるしきい値以下になった場合に，この標準パターンであると認識される．この図 8 では，標準パターンにジェスチャ“手をあげる”のフレーム＃1～＃4 しか含まれていないのにフレーム＃5～＃7 も認識されている．これは，図 7 の＃1 と＃7，＃2 と＃6，＃3 と＃5 のフレームどうしが似ているためである．このように，Non-monotonic 連続 DP では，図 8 のようにフレーム＃1～＃7 のジェスチャを入力した場合だけでなく，戸惑っている動作（フレーム＃1～＃4，＃4～＃7）も 1 つの標準パターンで認識可能である．

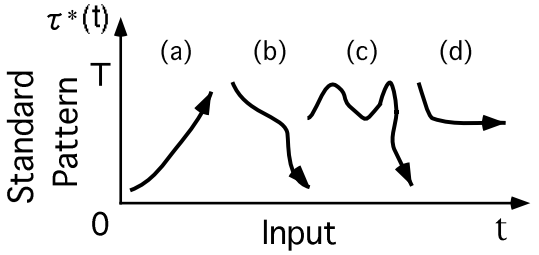

図 9 Non-monotonic 連続 DP によるある動作の部分的に変形した動作 4 種の認識
（(a) 順方向動作，(b) 逆方向動作，(c) 戸惑い動作，(d) 静止した動作）
Fig. 9 Recognition of four partial gesture using Non-monotonic CDP. ((a)Normal gesture, (b)Reverse gesture, (c)Hesitated gesture, (d)Stop gesture)

じつは，このとき Non-monotonic 連続 DP では，ジェスチャ“手をあげる”を認識しているのではなく，“片手を上下に動かす”動作を認識しているといえる．ほかにも，例として“両手を上下する”動作，“片手を左右に振る”動作などを認識するための標準パターンが考えられる．複数の標準パターンがある場合，$S(t,\tau^*(t))$ からしきい値を引いた値が最小となる標準パターン $l^*(t)$ を認識結果とする（ただし，この最小値が正の場合には認識結果は出力されない）．これは，ジェスチャの大分類を行ったものといえる．

さらに，図 8 の下方のようにマッチングしたフレーム番号 $\tau^*(t)$ も得られ，この変化からジェスチャの細分類が可能となる．**図 9** には，ある動作が部分的に変形した動作として，標準パターンに対して (a) 順方向，(b) 逆方向，(c) 戸惑い，(d) 静止した動作の 4 種類についての認識の様子を示した．これらの動作は，標準パターン中のどの部分区間で行われても認識できる．このような認識を従来の連続 DP で行おうとすると，多くの標準パターンを用意する必要が生じ非効率的である．したがって，このような認識が目的の場合には，Non-monotonic 連続 DP [19]が有用である．

また，本手法では戸惑い動作だけでなく，“少し大きい”，“非常に大きい”などの連続的な程度を示すジェスチャの認識も同じ枠組みで認識可能である．たとえば，両手を広げると“大きい”，狭めると“小さい”を

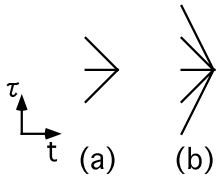

図 10 Non-monotonic CDP の傾斜パターン例
(a) $m \in \{-1, 0, 1\}$, (b) $m \in \{-2, -1, 0, 1, 2\}$
Fig. 10 Examples of local path of Non-monotonic CDP.

示すものとする．このとき，標準パターンを両手を狭めた状態から広げるまでのフレーム区間とすればよい．これにより，大小の程度を連続的に認識することができる．このように，本手法は戸惑い動作だけでなく，程度を示す動作をも同じ枠組みで認識できるという特徴がある．そのために，従来の連続 DP に局所的遷移において非単調性を導入するが，これが導入された連続 DP を「Non-monotonic 連続 DP」[19]と呼ぶ．この手法により，戸惑っているジェスチャだけでなく，“少し大きい”,“非常に大きい” などの連続的な程度を示すジェスチャの認識も可能となる．

Non-monotonic 連続 DP では $S(t,\tau)$ を以下のような漸化式で更新する．

初期条件 $(t = 0)$：

$$S(0,\tau) = d(0,\tau). \quad (1 \le \tau \le T) \tag{10}$$

漸化式 $(1 \le t)$：

$$\begin{aligned} &S(t,\tau) = \alpha \cdot d(t,\tau) \\ &+(1-\alpha) \cdot \min_{m\in\{-1,0,1\}} S(t-1,\tau+m). \\ &(1 \le \tau \le T) \end{aligned} \tag{11}$$

ここで，α は正規化係数 $(0 \le \alpha \le 1)$ である．式を簡単にするために，入力の系列は標準パターンと比べて -1〜1 倍の伸縮があってもマッチング可能であるとした．これは，**図 10** (a) のような傾斜パターンを採用していることになる．しかし，式 (11) の m の範囲を変えれば図 10 (b) のような様々な傾斜パターンを設定できる．

ここで整数 $p_0, p_1, \cdots, p_t$ を以下のように定義する．

$$\begin{aligned} &p_t = \tau, |p_k - p_{k-1}| \le 1 (k = t, t-1, \cdots, 1) \\ &\qquad\qquad and \\ &1 \quad \le p_k \le T (k = 0, 1, \cdots, t) \end{aligned} \tag{12}$$

このとき，式 (10)，式 (11) の漸化式は次式のように変形できる．

$$\begin{aligned} &S(t,\tau) \\ &= \min_{\{p_{t-1},p_t\}} \{\alpha \cdot d(t,\tau) \\ &\quad +\alpha \cdot (1-\alpha) \cdot d(t-1, p_{t-1}) \\ &\quad +(1-\alpha)^2 \cdot \min_{m\in\{-1,0,1\}} S(t-2, p_{t-1}+m)\} \\ &= \min_{\{p_{t-2},p_{t-1},p_t\}} \{\alpha \cdot d(t,\tau) \\ &\quad +\alpha \cdot (1-\alpha) \cdot d(t-1, p_{t-1}) \\ &\quad +\alpha \cdot (1-\alpha)^2 \cdot d(t-2, p_{t-2}) \\ &\quad +(1-\alpha)^3 \cdot \min_{m\in\{-1,0,1\}} S(t-3, p_{t-2}+m)\} \\ &= \cdots \\ &= \min_{\{p_0,p_1,\cdots,p_t\}} \{\sum_{k=1}^{t} \alpha(1-\alpha)^{t-k} \cdot d(k,p_k) \\ &\quad +(1-\alpha)^t \cdot d(0,p_0)\}. \\ &\qquad\qquad (1 \le t) \end{aligned} \tag{13}$$

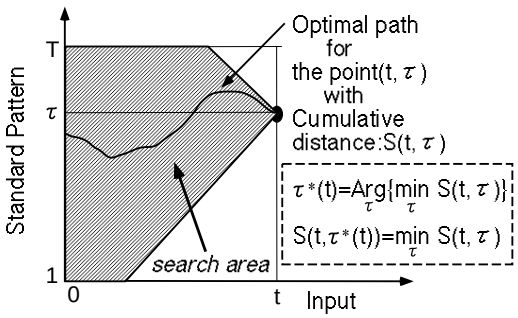

図 11 Non-monotonic 連続 DP におけるパスの探索範囲
Fig. 11 Path search area of the Non-monotonic CDP matching.

つまり，Non-monotonic 連続 DP では，図 10 (a) のように (t,τ) において $(t-1,\tau-1)$，$(t-1,\tau)$，$(t-1,\tau+1)$ の各点から局所最適パスがとられ，**図 11** の実線のように (t,τ) 平面での最適パスの τ が t に関して単調に増加するものとはなっていない．この意味により，ここで提案するものを「Non-monotonic 連続 DP」と呼ぶこととする．

「Non-monotonic 連続 DP」の有効性を「連続 DP」との比較で議論する．連続 DP はその標準パターンが時間的な単調性を保持し，入力系列との整合も入力側の単調性を保持するという制約がある．「Non-monotonic 連続 DP」は，標準パターンは単調性を保持していても，入力側には非単調性を許している．したがって，同一の単調性を持つ標準パターンについて，最適整合で整合する入力パターンの数は「Non-monotonic 連続 DP」の方が多いといえる．理論的には，この非単調性で整合をとりうる入力パターンの数は，任意の部分区間の逆方向を許すということで，その種類はその組合せの数だけとなる．いま，2 つの標準パターンがあるとすると，それぞれの扱いうる非単調性を許す入力系列の集合には重なる部分が存在する．たとえば，「右手を左から右へ動かす」という動作が認識しうる

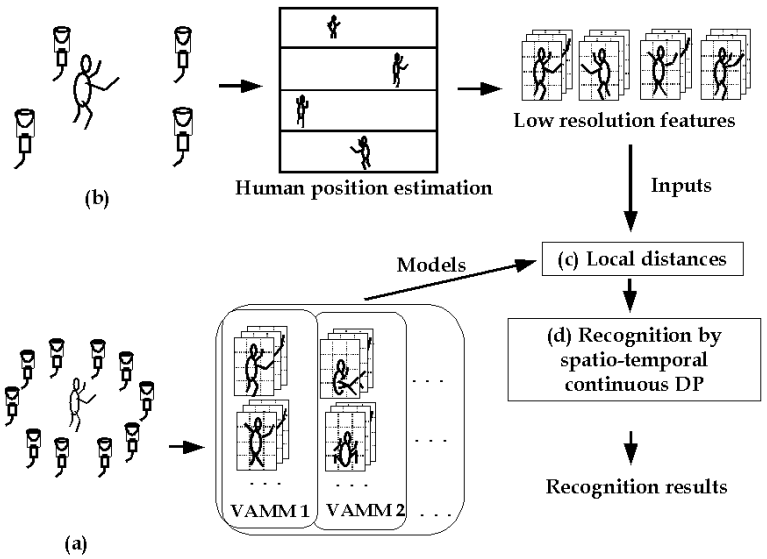

図 12 カメラから離れた場所で人物が動作をする (b) ものを認識するシステムの構成．この場合，動作のモデルは多数のカメラの動画像からつくられる (a)．

Fig. 12 Overview of the system for recognizing movements (b) of a person who takes a distance from the camera where each movement is made of many reference patterns captured by many cameras (a).

非単調な入力には,「右手を右から左へ動かす」という動作を含む．したがって，2 つの標準パターンを「右手を左から右へ動かす」と「右手を右から左へ動かす」とすれば，扱いうる非単調性を持つ入力の集合は多くの重なりを持つということになる．したがって，これらのことを区別するには,「Non-monotonic 連続 DP」値で局所的な整合の最大値によって判断するのではなく，その対応の部分的な整合部分を明示的に調べることで，これらの区別をすることが求められる．結論的には,「Non-monotonic 連続 DP」の方が,「連続 DP」より多様な変化系列パターンを整合できるということができる．ただし，その整合値の計算手法が式 (13) で前者が計算され，局所距離に付加される重みが現時点よりさかのぼるとより減衰するので，それが一様であるとする後者の「連続 DP」の方がより自然なものといえ，この違いが問題となる状況も皆無ではないといえる．

6. カメラから離れての動作

人間が部屋内で移動し，かつジェスチャでコンピュータと対話する状況を考えよう．人物はジェスチャ動作している最中には移動しないとする．また，ジェスチャは先の全方位画像を複数使ってとらえられているものとする．このような状況で人物の動作を理解するシステムも開発されている．そのシステムの構成が**図 12** に示されている．このシステムでは，固定された複数の全方位カメラで動作人物を観測し，それぞれのカメラの動画像から動作部分を表している部分を長方形に切り出し，特徴系列を抽出し，抽出された各動画像特徴系列について，多数の全方位カメラから得られる時空間パターンとしての動作モデル（1 つの動作に 1 つの時空間パターンが対応）との間で，非線形最適整合処理（時空間連続 DP）が行われる[20]．このとき，どのような動作が認識の対象になるかについては，**図 13** にその例が示されている．これらはジェスチャというより身体の全部を使う動作といえるものである．このシステムでは，動作者は特定のカメラ方向を向かなければならないという制約がない．

6.1 時空間連続 DP

時空間連続 DP は，連続 DP の 1 つの拡張方式である．拡張の内容とは，通常の連続 DP では累積距離値が定義される空間が入力とモデルの時間軸，つまり (t, τ) の 2 次元内であるものを，さらに，空間方向の次元を拡張することである．本報告では，人物の方向 i_θ（1 次元）を空間方向の次元として拡張した場合について**図 14** を用いて説明する．従来の連続 DP では，累積距離値が定義される空間軸は図 14 (a) のようにモデルのフレーム軸 τ と入力フレーム軸 t のみであり，したがって，図 14 (a) 上方の 3 個の局所パスのみが用いられている．一方，人物の方向変化を許容するため

図 13 カメラ方向に自由でも認識される 10 種類の動き
Fig. 13 Ten motion behaviors for recognition experiments.

に，図 14 (b) 上方のように i_θ 軸の変化を許容する局所パスを設定する．そこで，i_θ の変化を許容するパスと時間方向の変化を許容するパスを融合し，図 14 (d) に示す 9 個の局所パスを採用することとする．これによって，人物が動作を行う場合，モデルと比べて時間方向に $\frac{1}{2}$〜2 倍の伸縮を許容するだけでなく，人物の方向についても入力 1 フレームあたり最大 $2\pi/N_\theta$ の角度変化を許容することになる（図 14 (c)）．さらに大きな角度変化を許容したい場合は，同様に空間方向にシフトした点からパスを張ればよい．

以下，本報告で用いた時空間連続 DP の定式化を行う．点 (i_θ, τ, t) を終点としたモデルと入力系列との累積距離を $S(i_\theta, \tau, t)$ で表す．時空間連続 DP では $S(i_\theta, \tau, t)$ を以下のような漸化式で更新する．

境界条件 $(1 \le i_\theta \le N_\theta, 1 \le \tau \le T, 0 \le t)$：

$$S(i_\theta, \tau, -1) = S(i_\theta, \tau, 0) = \infty. \quad (14)$$

$$S(i_\theta, 0, t) = \infty. \quad (15)$$

漸化式 $(1 \le t)$：

$$S(i_\theta, 1, t) = 3 \cdot d(i_\theta, 1, t). \quad (\tau = 1) \quad (16)$$

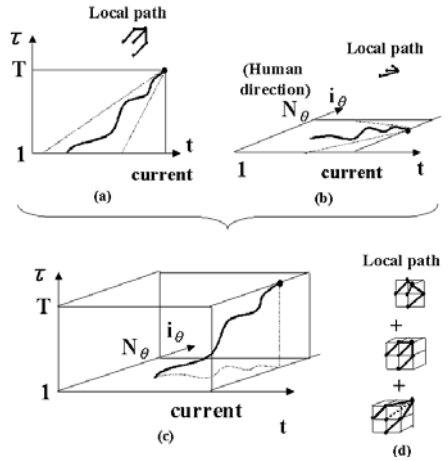

図 14 時空間連続 DP（人物の方向表現のための 1 次元が拡張されている）
t：入力のフレーム番号，τ：モデルのフレーム番号，i_θ：人物の方向に関するパラメータ．(a) 従来の連続 DP (b) 人物の方向変化への対応 (c) 時空間連続 DP (d) 時空間連続 DP の局所パス．
Fig. 14 Spatio-temporal Continuous DP. (a)Conventional continuous DP, (b)Cope with the change of human direction, (c)STCDP, (d)Local path of STCDP (When the space represents the direction of human which has one dimension).

$$S(i_\theta, \tau, t) = \min \begin{cases} S(i_\theta, \tau-1, t-2) + 2 \cdot d(i_\theta, \tau, t-1) \\ \quad + d(i_\theta, \tau, t) \\ S(i_\theta, \tau-1, t-1) + 3 \cdot d(i_\theta, \tau, t) \\ S(i_\theta, \tau-2, t-1) + 3 \cdot d(i_\theta, \tau-1, t) \\ \quad + 3 \cdot d(i_\theta, \tau, t) \\ S(\hat{\mathrm{i}}_\theta, \tau-1, t-2) \\ \quad + 2 \cdot d(\hat{\mathrm{i}}_\theta, \tau, t-1) + d(i_\theta, \tau, t) \\ S(\hat{\mathrm{i}}_\theta, \tau-1, t-1) + 3 \cdot d(i_\theta, \tau, t) \\ S(\hat{\mathrm{i}}_\theta, \tau-2, t-1) \\ \quad + 3 \cdot d(\hat{\mathrm{i}}_\theta, \tau-1, t) + 3 \cdot d(i_\theta, \tau, t) \\ S(\tilde{\mathrm{i}}_\theta, \tau-1, t-2) \\ \quad + 2 \cdot d(\tilde{\mathrm{i}}_\theta, \tau, t-1) + d(i_\theta, \tau, t) \\ S(\tilde{\mathrm{i}}_\theta, \tau-1, t-1) + 3 \cdot d(i_\theta, \tau, t) \\ S(\tilde{\mathrm{i}}_\theta, \tau-2, t-1) \\ \quad + 3 \cdot d(\tilde{\mathrm{i}}_\theta, \tau-1, t) + 3 \cdot d(i_\theta, \tau, t). \end{cases} \quad (17)$$

$(2 \le \tau \le T)$

ただし，$\hat{\mathrm{i}}_\theta = i_\theta - 1 (i_\theta > 1)$，$\hat{\mathrm{i}}_\theta = N_\theta (i_\theta = 1)$，$\tilde{\mathrm{i}}_\theta = i_\theta + 1 (i_\theta < N_\theta)$，$\tilde{\mathrm{i}}_\theta = 1 (i_\theta = N_\theta)$ とした．時空間連続 DP の出力 $A(i_\theta, t)$ は，重みの和 $3 \cdot T$ で正規化して $A(i_\theta, t) = \frac{1}{3 \cdot T} S(i_\theta, T, t)$ と定める．

さらに，モデルが L 個存在するとし，それぞれの連続 DP の出力を $A_\ell(i_\theta, t) (1 \le \ell \le L)$，しきい値を $h (0 \le h \le 1)$ とする．認識結果は，マッチングしたモデルのカテゴリ番号 $\ell^*(t)$ とそのときの人物の方向 $i_\theta^*(t)$ であり，以下の式で判定する．

$$\{\ell^*(t), i_\theta^*(t)\} =$$

$$\begin{cases} \mathrm{Arg}[\min_{\ell}\min_{i_\theta}\{A_\ell(i_\theta,t)\}] \\ \quad \text{if } \exists\ell \quad \text{so that} \quad A_\ell(i_\theta,t)\le h \\ \mathrm{null} \qquad \text{otherwise} \end{cases} \tag{18}$$

ここで，Arg は引数 $\{\ell, i_\theta\}$ を返す関数，null は空のカテゴリを表す．図 13 などを認識する場合，認識率が高くなるよう人手でしきい値 h を設定する．

時空間連続 DP の性能を述べる[20]．図 13 で示された 10 種類の動作をそれぞれ 16 個のカメラでとり，時空間標準パターンを作る．次に，同じく 10 種類の動作を同 1 人物がそれぞれ 10 回行い，これをテストパターンとする（評価用データ数 100）．ここで，動作者が場所を動かないで動作する場合と，場所を移動しながら動作をする場合，の 2 つの場合のデータを作成した（全部で，評価用データは 200 となる）．また，認識時に，1 つのカメラからの映像のみを使った場合と，4 つのカメラを使ってこれらを統合した場合の 2 つの異なった場合ついて認識実験を行った．その結果，人物が移動しない場合，1 つのカメラでの認識率は 83%，4 つのカメラを使う場合は，100%の認識率であった．一方，人物が移動しながら動作した場合は，1 つのカメラでは，62%，4 つのカメラでは，82%の認識率であった．

なお，時空間連続 DP は，上記の場合以外にもいろいろ応用が考えられており，その 1 つに，人が口ずさむメロディによって音楽情報をスポッティング検索する応用がある[21]．この応用の場合，検索対象の音楽データベースは音程を空間軸に持つ時空間パターンとなり，これが標準パターンとなる．一方，メロディの音程の相対変化の特徴系列が入力となる．時空間連続 DP は，時空間標準パターンを対象に，入力時系列についてスポッティング処理を行うこととなる．

7. 実時間対話システム設計のアーキテクチャ

ジェスチャ認識を実現するソフトウエアは 1 つの機能モジュールとなる．このモジュールは対話システムの中に組み込まれて実際の役目を果たすことになる．対話システムは通常他のモジュールも組み込んだものとなっており，これらをどう組み合わせるかのアーキテクチャを必要としている．いま，1 つの対話システムが，ジェスチャ認識モジュール，音声認識モジュール，対話タスクモデル，出力を担うものして音声合成モジュールと CG 合成モジュールという，合計 6 つのモジュールからできているとしたときの組み上げアーキテクチャの構成例を概念的に示したものを**図 15** に示す．これを「実時間完結原理」と呼ぶ[22]．これは，

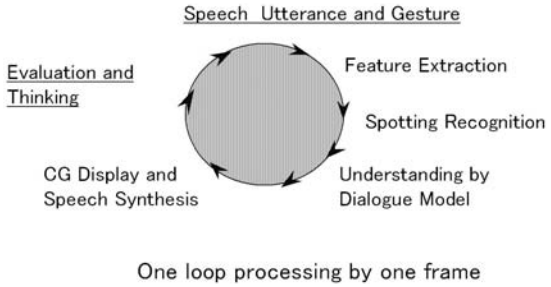

図 15 フレーム同期実時間対話システム構築のための「実時間完結原理」

Fig. 15 Realtime Completion Principle for realizing a frame-wise synchronized and realtime human computer dialogue system.

音声やジェスチャの認識がフレームごとにスポット的に実行されるという認識アルゴリズムの特性を他のモジュールの動作特性に拡大させたものである．具体的には，セグメンテーション・フリーに動作する音声やジェスチャの各フレームに同期して，他のモジュールもフレームごとに最終出力するように構成することである．図 15 に示すように，ジェスチャや音声の時間が 1 フレーム進行するとき，すべてのモジュール最終出力を出すものになっている．したがって，対話する人間側では，音声でいえば，いかなる時点においてもフレーム間隔である 10 msec 前までの入力の結果がシステムから応答することとなり，また，ジェスチャでは，いかなる時点においてもフレーム間隔である 100 msec 前までの入力にシステムからの応答があるということである．このようなシステムでは，人間側が思考している時間に応答が追随することができ，その対話の臨場感がきわめて強いものとなる[22]．出力形態としては音声合成と CG である．CG は実時間応答の追随性はきわめて高いが，合成音声の場合，合成音声の発声が終わるまで待つ必要はないが，最低その理解に要する時間は人間側に聞くということが必要になる．

図 15 で示したアーキテクチャをノート PC（233 MHz）上で実装した場合，1 つの CPU の計算スケジュールをどのように行っているかの例を**図 16** に示す．このノート PC で，キャプチャされる動画像は 10 フレーム/秒であることより，フレーム間隔は 100 msec となり，音声についての分析フレーム間隔は 10 msec とされている．この 2 種のフレーム間隔を考慮して，6 つの機能モジュールを 1 つの CPU による計算スケジュールを示したのが図 15 である．これは，縦軸に処理の優先順位を示し，横軸にその優先順位に基づく処理の内容を示す．音声のフレーム間隔の 10 msec の内の 4 msec の時間で，波形が分析さ

れ，特徴抽出とスポッティング認識が行われ，その結果によってタスクのネットワークモデルであるオートマトンの状態を更新する．上記の音声のフレーム間隔 10 msec で残っている 6 msec を 3 回分である 18 msec の時間で動画の 1 フレーム入力についての特徴抽出とジェスチャのスポッティング認識処理を行いその結果によってタスクのネットワークモデルの状態の更新を行う．これが実行されるのは，全体が 100 msec の中で 30 msec が経過したときである．この間，音声は，10 msec ごとにスポッティング認識とオートマトンの更新が 3 回行われる．その後，31 msec から 100 msec まで，7 回の音声のスポッティング認識が行われるが，各 10 msec のうちの 4 msec であるので，残り 6 msec の 7 回分，42 msec の CPU 処理時間が，更新されたタスクネットワークの状態に基づいて，ユーザに応答を音声合成と CG で示すために用いられる．この出力の機会は，100 msec の中で，音声で 10 回，ジェスチャで 1 回のオートマトンの最大 11 回の状態更新について，時間の早い順に選択される．更新したオートマトンを 42 msec の CPU 時間で表示できないときは，次の出力のために与えられる CPU 時刻である 30 msec あとにまわされる．しかし，通常の場合，100 msec の時間の進行に，42 msec の出力のための CPU 処理時間に間に合わないということは実際上ほとんど生じることはない．また，この実装は，CPU が 233 MHz のノートパソコン上で実装した場合[23]の CPU の計算スケジュールであり，CPU の速度がより高まれば，この対話の処理にかかわる計算負担はより小さくなることは明らかであり，対話以外のデータベースの検索処理に計算資源をより多く使えることとなる．

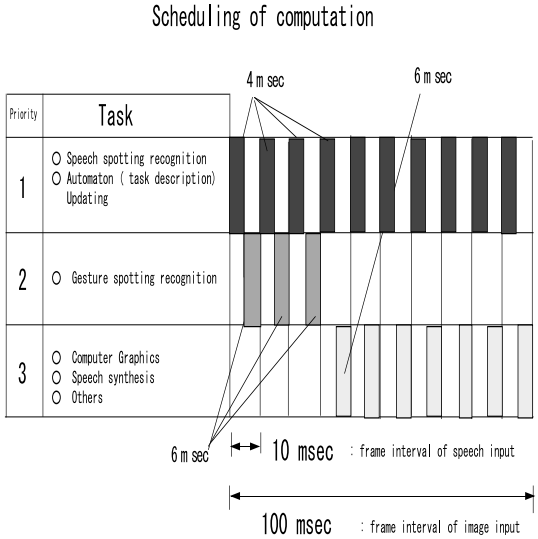

図 16 フレームごとに実時間処理を単一の CPU において実行するための計算負荷の配分スケジュール
Fig. 16 Time chart for scheduling the computation by a single CPU.

8. タスク記述ネットワークと応答出力

フレームごとの音声やジェスチャのスポッティング認識出力から対話システムがユーザに応答する出力がどのようになされるかを述べる．まず，ノードとアークからなるネットワークでタスクのモデルを考える

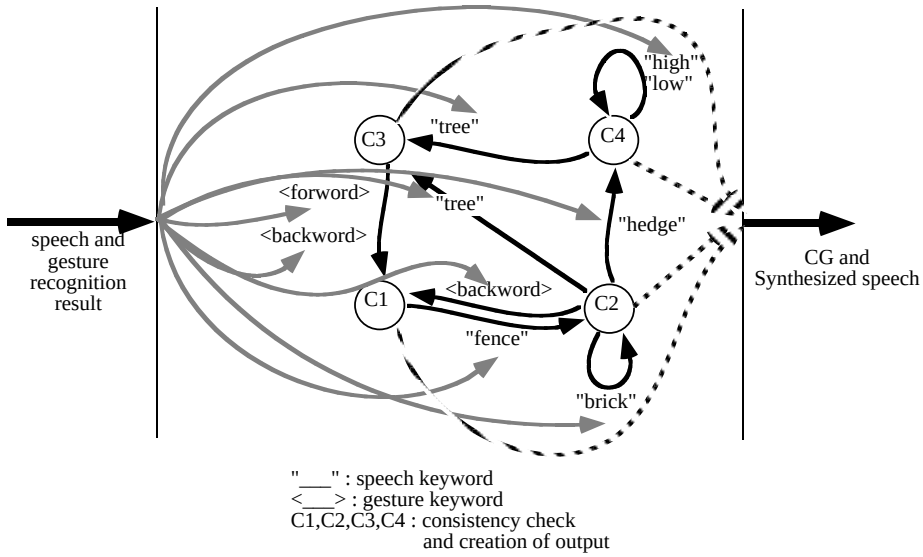

図 17 ネットワークモデルで記述されたタスクとそこにおける入出力の関係
Fig. 17 Network model of the task and its relation to I/O of the system.

（**図 17** 参照）．ここで，各ノードは自らのノードの状態のみで CG や合成音声によって出力が構成できる機能を持つとする．ノードに入ってくるアークにはジェスチャ認識や音声認識の語彙のラベルを付加してあり，すべてのアークのラベルには，フレームごとに音声やジェスチャのスポッティング認識結果が入力し，自らのラベルと一致するかどうかが判断される．各ノードはその時刻までの履歴とその時刻のアークを通じての入力との矛盾の有無をチェックし，そのノードの出力尤度を計算する．尤度の最も高いものが，CG と合成音声により外部に出力する．この出力後，すべてのノードの状態，次のフレームの入力を受け入れるべく状態を更新する．ノードの状態の記述は,「家の設計における対話モデル」の例の場合，2 次元のテーブルとされ，ジェスチャや音声認識の語彙が受け入れの可否が周辺の項目によって決められているものとなっており，そのテーブルの記述自体が CG 画像としての表示しやすいものとなっている[24]．また，合成音声の出力は，前回出力した CG 画像との変化分をユーザに伝えるものとなっている．図 17 では，アークで ”____ ” で付加ラベルを示したものが，音声認識によるもので，<____> はジェスチャ認識によるものであることを示している．複数人のユーザによる対話では，たとえば，

A：塀はレンガの方がいいのでは？

B：私は塀は植え込みの方がいい．

という会話では，話者 A の「塀」と「レンガ」と「いい」という単語が音声からスポッティングされ，塀がレンガとなった家の様子が CG で表示される．次に，「塀」と「植え込み」と「いい」が音声スポッティングされると，塀が植え込みに変わった CG が表示され，合成音声で,「塀が植え込みに変わりました」ということが出力される．

また，音声とジェスチャを交えて,「木はこちらへ」と音声で,「こちら」のところをジェスチャで右の方へ動かす動作をすると，音声で,「木」をスポッティング認識し,「右の方」をジェスチャでスポッティング認識すると，CG では，木が右に動いた様子が示され，合成音声で「木が右の方へ移動しました」ということが発声される．

このような対話のタスク記述のネットワークを手動でつくるのは，その語彙やネットワークが大規模になったとき困難であるので，対話の事例を与えて自動的にネットワークで構成する方式も提案されている[25]．

さらに，新しいジェスチャの種類をその場で登録できるように利便性を高める方式も提案されている[15]．

9. おわりに

本稿では，筆者らがこれまで開発してきた，動画像で観測された人物のジェスチャを認識する方式を述べてきた．ジェスチャ認識は人間とコンピュータの対話の中で用いられることが多いことから，このジェスチャ認識を組み込み音声認識と統合したシステムの設計のためのアーキテクチャとその上で働く対話の例も簡単に述べた．ジェスチャは言語的な意味を表す側面と，戸惑いの状態や，ものごとの程度や，指し示す物理的方向など，音声のような言語的なメディアに比べて，非言語的表現を多く持っているメディアである．そのため，対話システムで，テキストや音声の言語的な表現と相補的な役割を果たさせる場合に有効である．また，ジェスチャ認識の応用としては，手話認識の一翼を担うという側面が強調されるが，それ以外でも多くの利用される場面がある．たとえば，音声入力を行う場合，マイクとユーザが物理的に近くなくてはならないという条件が満たされない，移動ロボットを離れて操作するなどの場合に使われることでも有効な場合がある．あるいは，ダンスなどで身体動作をすること自体に意味があり，その際にコンピュータからのユーザの動作に的確に応答することが，たとえばダンスの技術の向上に役に立つなど，意味がある場合などにも有効である．このように人にとって日常的に使っている身体動作であるジェスチャは，コンピュータとの対話の中に自然に使われる場面は多いと思われるが，現在のジェスチャ認識の水準はきわめて不十分であるといえる．たとえば，ジェスチャの中には，手先の小さな動きに意味をこめたり，表情とのつながりでジェスチャの意味が変化する場合など，人間の大きな動作と小さな動作が同時に相互に関係して重要な意味を伝えている場合などが多い．これらの動作を認識する技術は現在確立されていない．これからのいっそうの技術開発が待たれるところである．

謝辞 本研究の遂行にご協力いただいた高橋勝彦氏関進氏，小島浩氏，長屋茂喜氏，向井理朗氏，櫻井茂明氏，高橋裕信氏ならびに新情報処理開発機構島田潤一研究所長の方々に深謝します．また，アルゴリズム実装化に協力いただいた（株）メディアドライブの松村博氏，ほか皆さんに深謝します．

参考文献

1) 黒川隆夫：ノンバーバルインタフェース，オーム社 (1994).
2) Ekman, P. and Friesen, W.V.: The reportoire

of nonverval behavior-categories, origins, usage, and coding, *Semiotical*, pp.49–98 (1969).
3) 石井浩史，望月研二，岸野文郎：人物像合成のためのステレオ画像からの動作認識法，信学論（D-II），J76-D-II, 8, pp.1805–1812 (1993).
4) Darrell, T. and Pentland, A.: Space-Time Gestures, *Proc. IJCAI'93 Looking at People Workshop* (Aug. 1993).
5) Yamato, J., Ohya, J. and Ishii, K.: Recognizing Human Action in Time-Sequential Images using Hidden Markov Models, *Proc. CVPR'92*, pp.379–387 (Jun. 1992).
6) 大和淳司，大谷 淳，石井健一郎：隠れマルコフモデルを用いた動画像からの人物の行動認識，信学論（D-II），J76-D-II, 12, pp.2556–2563 (1993).
7) *Proc. 4th International Conference on Automatic Face and Gesture Recognition*, Grenoble, France (March 2000).
8) Sagawa, H., Ando, H., Koizumi, A., Iwamura, K. and Takeuchi, M.: Sign Language Recognition and its Application, *Proc. 2000 Real World Computing Symposium* (*RWC 2000*), pp.143–146 (2000).
9) 西村拓一，向井理朗，岡 隆一：白黒動画像からの形状特徴を用いたジェスチャのスポッティング認識システム，信学論（D-II），J81-D-II, 8, pp.1812–1821 (1998).
10) 西村拓一，向井理朗，野崎俊輔，岡 隆一：低解像度特徴を用いた複数人物によるジェスチャの単一動画像からのスポッティング認識，信学論（D-II），J80-D-II, 6, pp.1563–1570 (1997).
11) 山澤一誠，八木康史，谷内田正彦：移動ロボットのナビゲーションのための全方位視覚センサー HyperOmni Vision の提案，信学論（D-II），J79-D-II, 5, pp.698–707 (1996).
12) 岡 隆一：連続 DP を用いた連続音声認識，音響学会音声研資，S78-20, pp.145–152 (1978-06).
13) 速水 悟，岡 隆一：連続 DP による連続単語認識実験とその考察，信学論（D），J67-D, 6, pp.677–684 (1984).
14) 高橋勝彦，関 進，小島 浩，岡 隆一：ジェスチャ動画像のスポッティング認識，信学論（D-II），J77-D-II, 8, pp.1552–1561 (1994).
15) 西村拓一，向井理朗，野崎俊輔，岡 隆一：動作者適応のためのオンライン教示可能なジェスチャ動画像のスポッティング認識システム，電子情報通信学会論文誌 D-II，Vol.J81-D-II, No.8, pp.1822–1830 (1998).
16) Nagaya, S., Itoh, Y., Endo, T., Kiyama, J., Seki, S. and Oka, R.: Information Integration Architecture for Agent-Based Computer Supported Cooperative Work System, *IEICE Trans. Information and Systems*, Vol.E81-D, No.9, pp.976–987 (1998).
17) 向井理朗，西村拓一，高橋裕信，遠藤 隆，中沢正幸，松村 博，岡 隆一：マルチモーダルヒューマンインタフェースのノートパソコンへの実装，電子情報通信学会，PRMU-98-70, pp.69–75 (1998).
18) Nishimura, T., Yabe, H. and Oka, R.: A Method of Model Improvement for Spotting Recognition of Gestures Using an Image Sequence, *New Generation Computing*, Vol.18, No.2, pp.89–101 (2000).
19) 西村拓一，野崎俊輔，向井理朗，岡 隆一：連続 DP への非単調性導入によるジェスチャ動画像からの戸惑い動作のスポッティング認識電子情報通信学会論文誌 D-II，Vol.J81-D-II, No.1, pp.18–26 (1998).
20) 西村拓一，十河卓司，小木しのぶ，岡 隆一，石黒 浩：動き変化に基づく View-based Aspect Model による動作認識，電子情報通信学会論文誌 D-II，Vol.J84-D-II, No.10, pp.2212–2223 (2001).
21) 橋口博樹，西村拓一，張 建新，滝田順子，岡 隆一：モデル依存傾斜制限型の連続 DP を用いた鼻歌入力による楽曲信号のスポッティング検索，電子情報通信学会論文誌 D-II，Vol.J84-D-II, No.12, pp.2497–2488 (2001).
22) 岡 隆一：脳機能実現の超並列アーキテクチャ，電気学会誌，Vol.115, No.12, pp.786–789 (1995).
23) Mukai, T., Nishimura, T., Nagaya, S., Kiyama, J., Kojima, H., Itoh, Y., Seki, S., Takahashi, T. and Oka, R.: Multi-Modal and Realtime Dialogue Through Gesture-Speech Interface on Personal Computer, *Proc. 1997 Real World Computing Symposium* (*RWC '97*), pp.1–7 (1997).
24) 岡 隆一，伊藤慶明，木山次郎，張 建新：概念スポッティングのための画像オートマトン，日本音響学会平成 7 年度春季研究発表会，講演論文集，pp.67–68 (1994).
25) 櫻井茂明，岡 隆一：対話タスクモデルのサンプル単語時系列からの自己組織化，電子情報通信学会論文誌 D-II，Vol.J83-D-II, No.2, pp.827–839 (2000).

(平成 13 年 12 月 17 日受付)
(平成 14 年 3 月 8 日採録)

(担当編集委員 中村 裕一)

岡　隆一

昭和 20 年生．昭和 45 年東京大学大学院工学系研究科計数工学専攻修士課程修了．同年電気試験所（現，産業技術総合研究所）入所．パターン認識の研究開発に従事．平成 5 年より新情報処理開発機構へ出向．平成 14 年 3 月産業技術総合研究所に帰任．平成 14 年 4 月会津大学に着任．ジェスチャ認識，音声認識，移動ロボット，動画像・静止画・音声・音楽・音響・テキストの統合検索の方式に関する研究に従事．工学博士．電子情報通信学会，音響学会，人工知能学会，AVIRG，IEEE 各会員．

西村　拓一

昭和 42 年生．平成 4 年東京大学大学院工学系研究科計数工学専攻修士課程修了．同年 NKK（株）入社．X 線，音響・振動関係の研究開発に従事．平成 7 年より技術研究組合新情報処理開発機構つくば研究センタに出向．平成 10 年 NKK（株）復帰．平成 11 年技術研究組合新情報処理開発機構つくば研究センタ主任研究員．平成 13 年産業技術総合研究所サイバーアシスト研究センター研究員．時系列パターンの検索および認識，情報提供システムに関する研究に従事．工学博士．電子情報通信学会，人工知能学会各会員．

矢部　博明

昭和 44 年生．平成 7 年東京大学大学院工学系研究科機械情報工学専攻修士課程修了．同年シャープ（株）入社．平成 10 年より新情報処理開発機構出向．ジェスチャ認識，時系列データの自己組織化の研究開発に従事．平成 13 年シャープ（株）帰任．