人の動作のスポッティング認識に基づくエージェントのための動きモデル自動獲得

Automatic Acquisition of Motion Models for an Agent Based on Spotting Recognition of Human Motions

岡　隆一[1]　松崎　隆[1]

Ryuichi Oka[1],　Takashi Matsuzaki[1]

[1]会津大学　コンピュータ理工学部

[1]The University of Aizu, School of Computer Science and Engineering

Abstract: We propose a method for automatic acquisition of motion models based on so-called Time-Space Continuous Dynamic Programming which performs automatic recognition of a reference time-space pattern as a motion model from a time-varying image without segmentation. A set of primitive reference patterns are used as an initial stage of the model and then the recognized reference patters will create new reference patterns for adapting the situation by generalization and combination of them.

はじめに

人間とコンピュータのインタラクションの媒体の種類は多く考えられるが、その中にあって動画像で捉えられる動きの役割は大きい。その最大の理由は、人間の動きに込められる意図の表現の豊富さと相手のエージェントへの指示性がインタラクションに欠かせないものが多いためである。また、人間の動きが良好に動画像から識別できる技術は、同時に人間に限らず、ものの動きの識別も扱いうるものとなるからである。しかし、従来の技術では、それを実現するに際し、単純であってかつ実世界でロバストに機能する有効なものが数少ない状況であった。文献[1]では様々な従来法のよい紹介がなされている。

このような状況の中で、われわれは近年、「時空間連続ＤＰ」という手法を開発してきた[2][3][4]。この手法は、図１に示されるようにピクセル系列を参照動画像としたとき、それを動画像中にセグメンテー

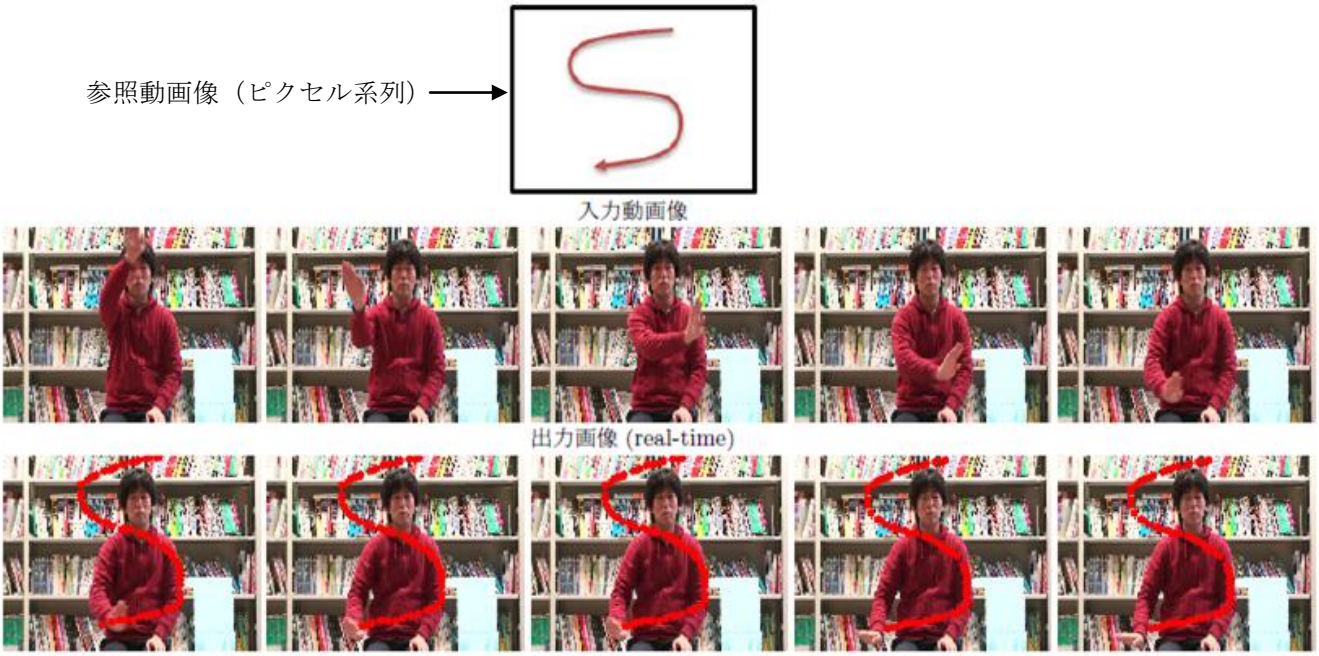


図１動作モデル（上図）と入力動画像（中図）スポッティング認識結果（下図）

ションフリーに識別するものである。この手法は「連続ＤＰ」[5]と呼ぶ、時系列の参照パターンを始終端のない時系列中にセグメンテーションを事前に行うことなく参照パターンに類似する部分を認識するアルゴリズムを、動画像のような時空間パターンが扱えるように拡張したものである。時空間連続ＤＰは様々な応用があるが、人間とエージェントとのインタラクションも重要な１つである。それを本論文で示す。特に、この手法を用いてエージェントが状況に応じて識別できる、あるいは識別されるべき一群の動きを自動獲得する方式を提案する。

動作の識別とインタフェース

エージェントやロボットが人間と良好なコミュニケーションを行うには、人間の動作を実世界で良好に識別できることが肝要である。適切なインタラクションの工学的設計は、人工知能や認知科学の支援を最終的には必要とするも、その前に生の動画像からどれだけの有用な動きが識別できるものであるかを追求する必要があると考える。すなわち、実環境での人間側の動作の、エージェント側による認識の正確さは、その後のインタラクションの基本設計に与える影響が大であると思われる。エージェント側の合成されたマルチメディアの出力の巧妙さに関心がいくことがあるが、認識が合成に比べ格段に技術的に困難であることは論をまたない。図２に示されるように、人間側の動きには多様なものがあり、これらを適切に認識することの困難さは文献[1]においても解説されている。

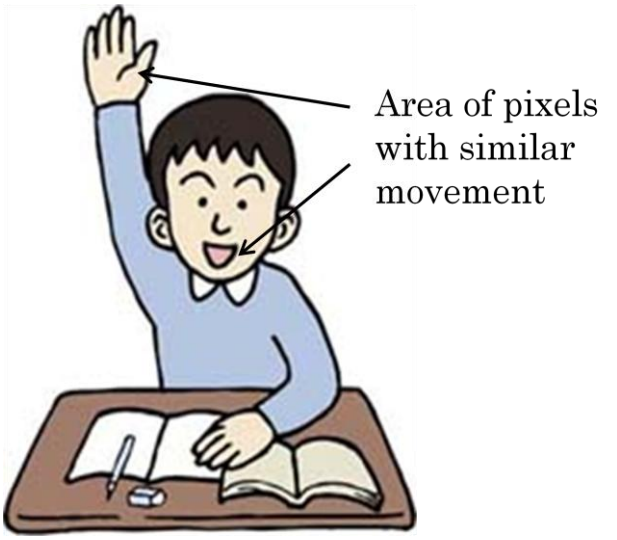


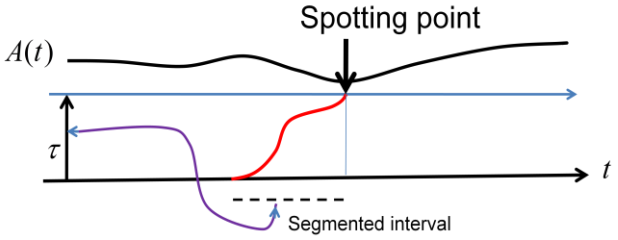
図２　動画像中の人間の動作である手の動きや唇の動きを良好に識別することがエージェントやロボットに求められる。

従来の動き認識の方法として大枠、２つに分類される[1]。それらは、動きの特徴抽出と特徴系列のマッチングである。このとき、従来法での特徴抽出は時空間パターンにおける局所的なもの(Kalman filter, Particle filter, ST-Interest Points, ST-Path, CHLAC, Volumetric Features)が主たるものである[1]。 マッチングは特徴ヒストグラムや特徴の時系列のＤＰやHMMなどによっている。従来法において、特徴が時空間パターンの局所的な領域で抽出されていることにより、オクルージョン、背景ノイズ、動きの変動、オブジェクトの出現や消滅、に脆弱であることが生じている。これらは上記の特徴の時系列によるマッチングでもその本格的な解消は困難であるように思われる。

我々は、上記の脆弱さをかなりの程度軽減できる時空間連続ＤＰという大域的な時空間パターン間のスポッティング認識ができるアルゴリズムを開発した[2][3][4]。この方式を以下簡単に説明し、それを用いて人間の動きのモデルの自動獲得の方法を提案したい。時空間連続ＤＰは連続ＤＰ[5]に空間パラメータを導入して拡張したもので、連続ＤＰの理解が前提となる。そこでまず連続ＤＰを説明する。

連続ＤＰ

連続 DP とは、区間時系列パターンを、別の始終端のない時系列パターンの中に見出すもので、事前にいかなる切り出しも行わないでそれを実行する。これをスポッティングという。その様子が図３に示してある。



$\{f(\tau)\,|\,1 \le \tau \le \mathrm{T}\}$: reference

$\{g(t)\,|\,-\infty \le t \le +\infty\}$: endress input pattern

図３　連続 DP の説明図

このアルゴリズムは、以下のような評価関数の最適解を各時刻でうるものである。

$$D(t,T) = \min_{r}\{\sum_{\tau=1}^{T} d(r(\tau),\tau)\},\ \text{where}\ r(\tau) \le r(\tau+1), r(T) = t.$$

r is a family of function from τ to t

これは参照区間パターンの各点と入力時系列パター

ンの各点の間に最適対応関係を与えるものである。ここで d は以下では局所距離というものである。累積局所距離を各時刻であたえるアルゴリズムが漸化式で与えられている。

$$d(t,\tau) = \| g(t) - f(\tau) \|$$

$$D(t,\tau) = \min \begin{cases} D(t-2,\tau-1) + 2d(t-1,\tau) + d(t,\tau) \\ D(i-1,\tau-1) + 3d(t,\tau) \\ D(i-1,\tau-2) + 3d(t,\tau-1) + 3d(t,\tau) \end{cases}$$

$$A(t) = \frac{D(t,T)}{3T}$$

boundary condition :

$$D(t,\tau) = \infty, \quad t \le 0, \tau \notin [1,T]$$

時空間連続 DP

上述の時系列同士のスポッティングマッチングである連続 DP に空間パラメータを内部的に導入するアルゴリズムを考える。それがここでいう時空間連続 DP とよぶものである。時空間動画像同士のスポッティングマッチングでは、２つのマッチングするべき動画像の１つを参照動画像と定義する。それを時間区間にわたるピクセルの１つの時空間系列とする。

$$Z(\xi(\tau),\eta(\tau)), \tau = 1,2,..,T$$

ここで、$(\xi(\tau),\eta(\tau))$ は座標点を示し、Z はそこにおけるピクセルの値である。この参照動画像を入力の動画像中に、開始時空間点と終了時空間点を任意として、類似ピクセル系列を抽出するものが時空間連続 DP である。

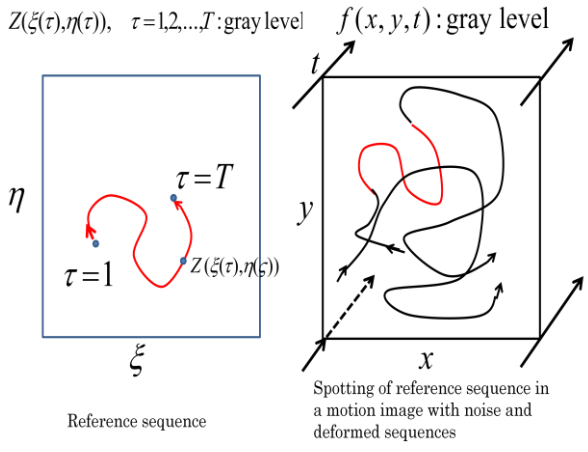


図４　参照動画像（左）と入力動画像（右）

このアルゴリズムを以下に示す。

$f(x,y,t)$: gray level motion image

$(1 \le x \le M,\ 1 \le y \le N), 1 \le t < \infty$

$Z(\xi(\tau),\eta(\tau)), 1 \le \tau \le T : (1 \le \xi(\tau) \le M,\ 1 \le \eta(\tau) \le N)$, gray level reference pattern

$v_x(\tau) = \xi(\tau) - \xi(\tau-1)$: difference of ξ in trajectory

$v_y(\tau) = \eta(\tau) - \eta(\tau-1)$: difference of η in trajectory

ここで局所距離は以下で定義される。

$d(x,y,\tau,t) = \| Z(\xi(\tau),\eta(\tau)) - f(x, y, t) \|$: local distance

次に、局所距離の和を３つの変数について最適化したものを、

Evaluation function :

$$S(x,y,T,t) = \min_{p,q,r} \left\{ \sum_{\tau=1}^{T} d(p(x,y,\tau), q(x,y,\tau), \tau, r(\tau)) \right\}$$

where $r(\tau) \le r(\tau+1), \quad r(T) = t.$

と記す。このとき、最適化の計算をＤＰで行うために４次元の空間が用いられる（図５）。

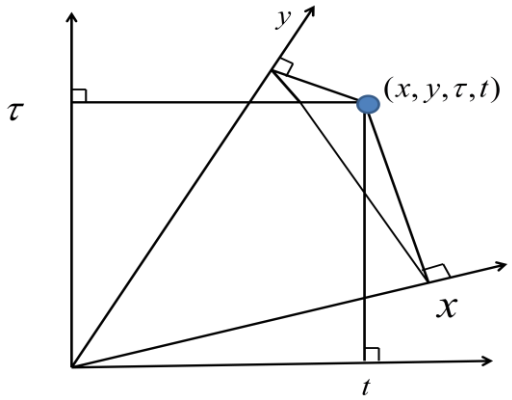


図５　最適化をＤＰで行うための４次元演算空間

DP の漸化式と境界条件は以下のものとなる。

$S(x,y,1,t) = 3d(x,y,1,t)$:

for $2 \le \tau \le T$;

$S(x,y,\tau,t) =$

$$\min \begin{cases} S(x - v_x(\tau), y - v_y(\tau), \tau-1, t-2) \\ \quad + 2d(x,y,\tau,t-1) + d(x,y,\tau,t); \\ S(x - v_x(\tau), y - v_y(\tau), \tau-1, t-1) + 3d(x,y,\tau,t); \\ S(x - v_x(\tau) - v_x(\tau-1), y - v_y(\tau) - v_y(\tau-1), \tau-2, t-1) \\ \quad + 3d(x - v_x(\tau), y - v_y(\tau), \tau-1, t) + 3d(x,\ y, \tau, t); \end{cases}$$

$S(x,y,\tau,t) = \infty, d(x,y,\tau,t) = \infty$ if $(x,y) \notin [M,N],\ t \le 0, \tau \notin [1,T]$

このとき、用いた傾斜制限は図６で示されるように４次元空間で構成される。また、出力である参照動画像をスポッティングする時空間点は以下の式で表される。

$$(x^*(t), y^*(t)) = \arg\{ \min_{\substack{(x,y)\in\{local \\ area\}}} \left\{ \frac{S(x,y,T,t)}{3T} \le h \right\} \}$$

上記の式で、(x, y) に関する最小値の計算を local area で行うが、その意味は、複数の類似する時空間軌跡がある場合、複数の局所領域における最小値でそれぞれがスポッティングされることを示す。

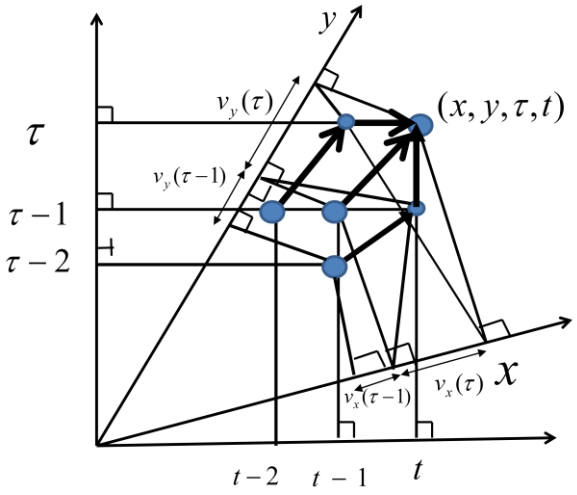


図６　４次元空間における傾斜制限

時空間連続 DP では１つ、または複数の参照動画像に入力動画像が適用される。したがって、事前に参照動画像群を設定することが前提になっている。図１では、S 字型の参照時空間パターンが入力の動画像パターン中にスポッティング認識されている。

時空間連続ＤＰにおける時間正規化と空間正規化

前述した時空間連続ＤＰの漸化式によるマッチングにおける時間正規化がどのようになされているかを説明する。時空間連続ＤＰの漸化式は右辺の３つの値のうちで最小値を選択することで、左辺の更新がなされる。図６に３つの選択の候補のパスが示されているが、これらは時空間パターンと参照パターンの間で 1/2 から 2 倍の時間軸における非線形対応が許されていることが示されている。ここで時間正規化のみを示しているが、空間的にも時間正規化と同時的にそのサイズが 1/2 から２倍の非線形変形を許すアルゴリズムは文献[3]に示されている。

動作モデルの自動獲得方式

エージェントやロボットが様々な状況に応じて人間と対話する際に、その状況を支配する、対話する相手の人間が振る舞う動作の集合のモデルを自動的に獲得できれば望ましいといえる。この獲得される動作モデルとは、**エージェントやロボット側が時空間連続ＤＰの処理を発動する際に直接的に用いられると同時に、対話を成り立たせる語彙の一部分としても用いられる。**この動作集合の表すモデルは、その状況において示される人間の動作を実際に動画像中から識別されるものである。つまり、単に必要な機能が記述されているという段階のものではない。すなわち、獲得されている動作モデルは、実世界で人間がそれらの動作を単独に、あるいは同時的に行った場合に、それらが実時間でほぼ確実に識別されることを保証しているということが肝要である。すなわち、これまでの実験[3][4]から、時空間連続ＤＰにおいて、始終端をもつピクセル系列である参照動画像は、その変形やオクルージョンを許して、また、背景のノイズに頑強に、時空間においてセグメンテーションフリーで入力の動画像中で識別できるものとなっている。ここにおいて、動作モデルの自動獲得問題とは、上述の機能をもつ時空間連続ＤＰの参照動画像パターン集合の自動的な獲得問題であると定義しよう。この参照動画像集合の自動獲得の方式を図７に提案する。以下、図７における処理についての説明を行う。

１）エージェントやロボットがいま遭遇している人物とのこの状況において、動作を介して良好なコミュニケーションを行うとする。
２）その際、事前には少ない一般的な動作の情報しかないとする。今の場合、図８で示されるような一般的と思われる数少ない時空間連続ＤＰの参照動画像（ピクセル系列）しかもたないという状況であるとする。
３）エージェントやロボットは対話の相手の動作をビデオカメラで捉え、初期に用意された参照動画像パターンによる識別を時空間連続ＤＰにより試み、その際、右上から左下へ直線運動をする動作を識別したとする。
４）エージェントはこの人物は直線の運動動作をすると想定し、多くの異なった方向をもつ直線運動の参照動画像を作成し、動きモデルの集合に加える。これは汎化による動きモデルの作成に相当する。

5）次に、もし円運動と直線運動の組み合わせがこの人物の意味をもつ動作であることがコミュニケーションを通じて判明したら、その組み合わせ動きを１つの参照動画像としてモデルに追加・登録する。これにより汎化とともに動作モデル集合の更新がなされる。これらによって作られた動作集合の例を図９に示す。

6）以下、既に登録されている参照動画像系列集合を用いて、上記の操作を繰り返す。この際、状況の把握に適さない参照動画像は削除される

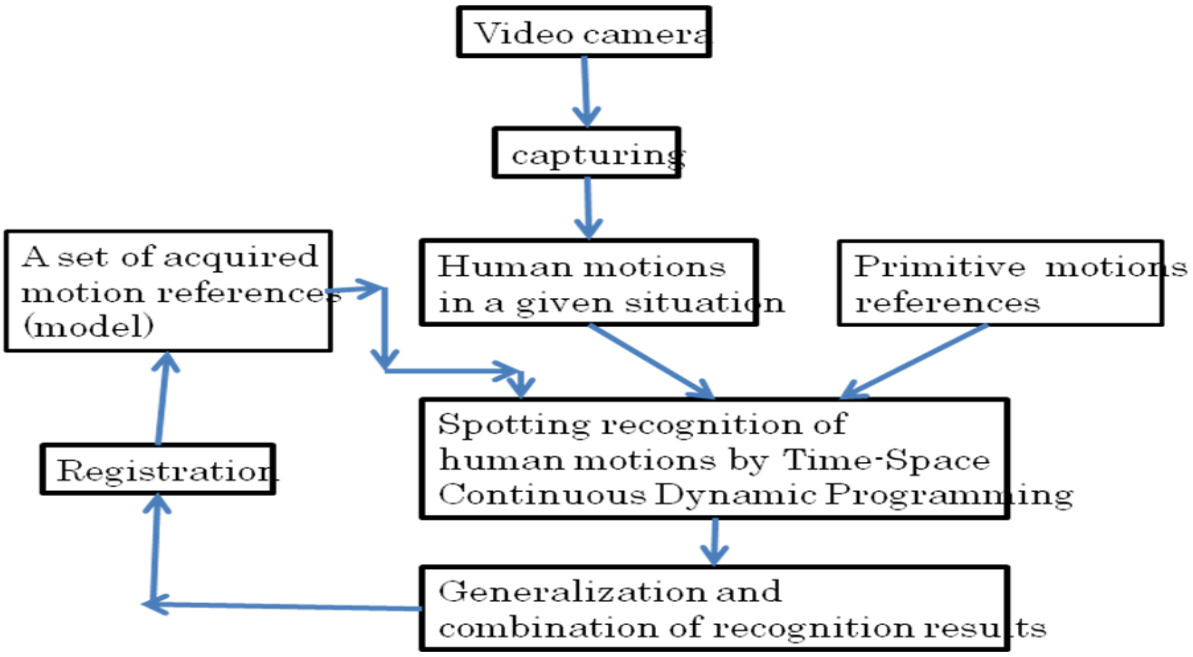


図７動作モデルの時空間連続ＤＰを用いた自動獲得方式

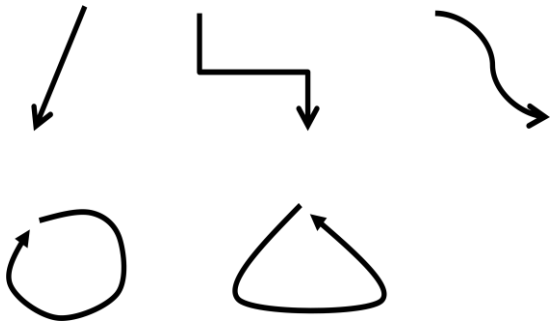


図８　初期に設定されるプリミティブ動作モデル集合の例

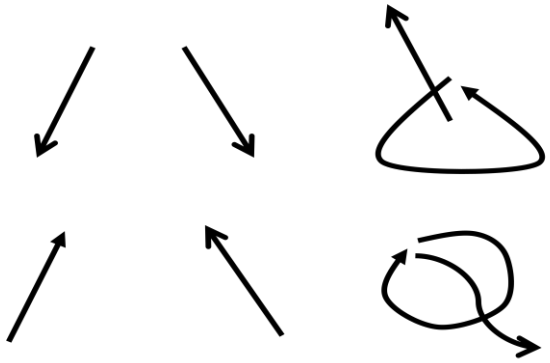


図９　状況に応じて獲得された動作モデル集合の例

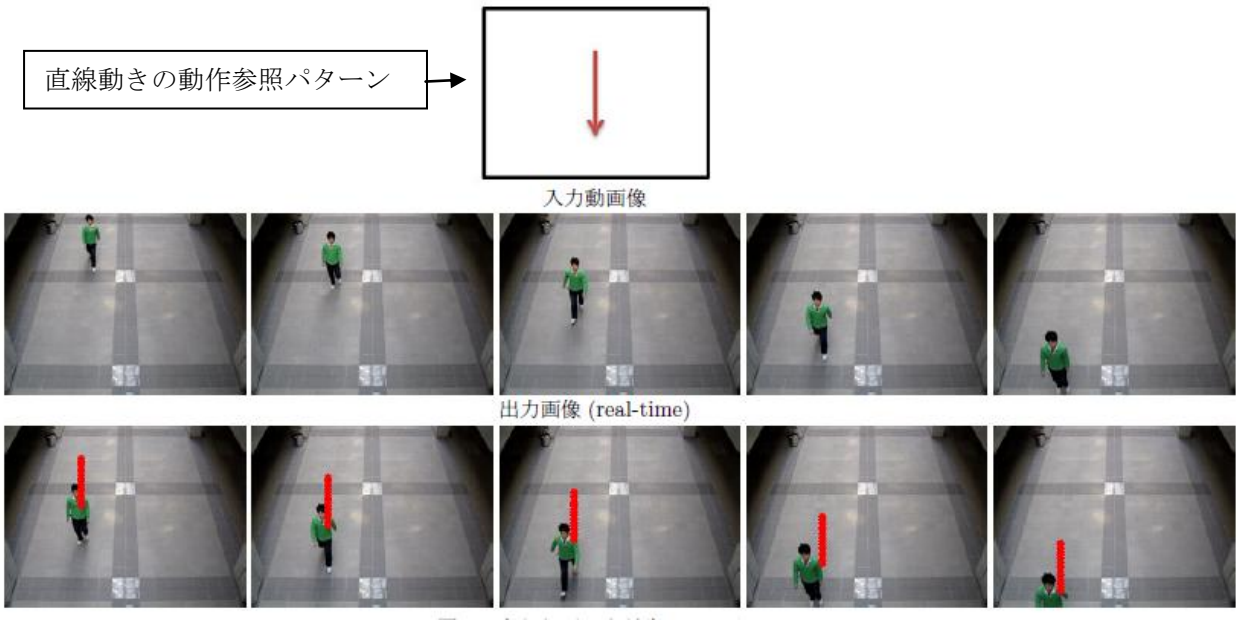


図１０　直的動作の時空間連続ＤＰによるスポッティング識別

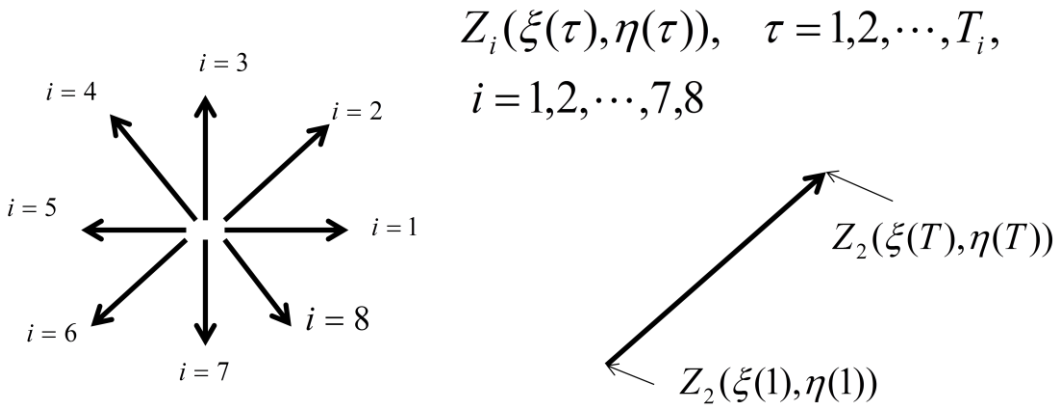


図１１　現在の状況（図 10 で例が検出される）の汎化によって 8 方向の直的動作の参照画像の構成または動き集合としての確定を行う.

簡単な実例を図１０で示す。いま、ある状況において、下向きの直線の動きのみの参照動画像を事前にもつ時空間連続 DP を想定し、長く継続している動画像中から、その下向きの動きの動きを図１０のようにスポッティング認識したとする。また、このスポッティング認識が頻繁に生じたとする。対象の長い動画像には様々な動きが存在していたと思われるが、これを時空間連続 DP で認識したとき、下向き直線的動きの参照動画像しかもっていないとき、それのみが検出されることになる。そのとき、いまこの下向きの動きが頻繁に検出される状況では、これまで参照動画像を保持していないため検出できなかったが、他の方向をもつ直線の動きも存在するという汎化の概念を適用する。すなわち、図１１で示されるような 8 方向の直線の動きの参照動画像を作成する。そして、この 8 方向の参照動画像をもって動画像に時空間連続 DP を動作させる。それによって実際に検出される直線的な動きを検証する。この検証によって実際に採用される動きの集合を確定する。

様々な参照動画像モデルの作成検証基準

上述の動作モデルの作成で重要な問題は、汎化や組み合わせによって構成される参照動画像パターンが、意味のある新規な動作として登録するための判断基準である。この判断基準に様々ものがあると考えられる。それらの例を述べると、以下に箇条書きされるものとなる。

１）すでに時空間連続ＤＰが保持している参照動画像集合の中から、実環境で識別ができた動作があった場合、エージェント側による合成音声で、それが重要であることを確認することである。例えば、「あなたは腕を左に下ろしましたが、それは何か意味ありますか？」などの質問をすることです。これへの応答に「はい」や「いいえ」などが含まれているか否かで判断する。

２）すでに時空間連続ＤＰが保持している参照動画像集合の中から、識別できる動作があったおき、それに基づいた行動をエージェントなりロボットが勝手に行動をとる。例えば、丸を描く動作を識別したら、ロボットの手を用いて自分で丸を描いてみる。それにより、人物の方でこのロボットは丸を認識できるものということを理解し、丸を使ったコミュニケーションを人物側が自発的に行うことを期待する。一般に、人間側はエージェントなりロボットが何を理解できるかを知りたがっているので、エージェントやロボット側が識別できるものを人間の動作から認識して、それを単に示すことに意味があるといえるのではなかろうか。

３）その他、動きではなくシーンの画像理解の結果と結び付けての応答によって判断する。

むすび

本稿では、われわれがすでに提案している時空間連続ＤＰというアルゴリズムを用いて、エージェントやロボットと人間との動作によるコミュニケーションを良好に行うための方式を示した。人間とロボットの間のコミュニケーションでは、ロボット側の応答の巧妙さ、高度さを求める方向とは別に、実世界から直接得られる動画像から人物の動作の識別をより良好にする方向にも意味があることを強調したい。本稿で提案して方式においては、時空間連続ＤＰが極めて簡単なアルゴリズムであり、動作モデルの自動獲得方式も極めて簡単なものであり、これらの実装作業には負担のかからないものである。本稿では主に短時間の動作に議論を限ったが、本提案方式は長時間の動作や行動にも容易に拡張できるものである。

謝辞

本研究の一部は科研費基盤研究(C)課題番号 23500220 の助成による。

参考文献：

[1] 特集：人間を理解するための ICT 技術, 電子情報通信学会誌, Vol.95, No.5, pp.371-465 (2012.05).

[2] 岡　隆一: 連続ＤＰによる画像処理, 電気学会, 情報処・次世代産業システム合同研究, IP-12-013, IIS-12-055, pp.65-70, (2012.03).

[3] 岡　隆一:時空間連続ＤＰにおける時間変形と空間変形およびオクルージョンへの頑健性, 電気学会, 情報処理・次世代産業システム合同研究, IP-12-25, IIS-12-67, pp.57-64 (2012.08).

[4] 松崎　隆, 矢口勇一,岡　隆一: セグメンテーションフリーかつオクルージョンに頑健な物体の動き認識とトラッキング, IS1-36, MIRU2012 (2012.08).

[5] Ryuichi Oka: Spotting Method for Classification of Real World Data, The Computer Journal, Vol.41, No.8, pp.559-565 (1998).