全身運動から言語空間の構築と 運動の認識への応用

Language Space from Whole Body Motions and Its Application to Motion Recognition

高野 渉
Wataru Takano

東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
takano@ynl.t.u-tokyo.ac.jp, <http://www.ynl.t.u-tokyo.ac.jp/>

中村 仁彦
Yoshihiko Nakamura

(同 上)
nakamura@ynl.t.u-tokyo.ac.jp, <http://www.ynl.t.u-tokyo.ac.jp/>

keywords: stochastic model, motion symbol, language space, motion recognition

Summary

Language is a symbolic system unique to human being. Human understands a real world through this symbolic system, makes inference, and communicates with others. Language underlies human intelligence. This paper describes a novel approach to combining motion primitives with words, and extracting relations among the words based on the motion primitives. The motion primitives are symbolized by Hidden Markov Models (HMMs). The HMM is hereafter referred to as “motion symbol”. Observation of human whole body motions is automatically segmented into motion primitives, recognized by using the motion symbols and converted to a sequence of the motion symbols. A sequence of words is also manually assigned to the observation. The association between motion symbols and words can be derived from pairs of these sequences as probability parameters, and dissimilarities among the words can be extracted. Words are located in a multidimensional space so that distances between the words in the space can become as close as possible to the dissimilarities. Thus, “language space” is formed. The mapping of motion primitives onto the language space enables robots to understand human behaviors as words. The association between motion symbols and words can be also applied to generation of motion primitives from words. The validities of our proposed methods are demonstrated on a motion capture dataset.

1. は じ め に

人間は言語を使って円滑なコミュニケーションを実現している。このコミュニケーションでは、自分が表出する単語が実世界で指し示すものと、他人がその単語から連想するものが同じでなければならない。実世界には、見たり触れたりすることで感覚に到達する多様な情報が溢れている。単語と感覚情報の参照関係を効率よく習得するメカニズムに、身体を通じた行為が深く係わっているという説が唱えられている。

Donald は、人間は身体運動の系列をプログラムすることから言語を獲得したとする仮説を提唱している [Donald 91]。コミュニケーションにおいて言語を伝達する手段が、身振り手振りから音声に取って代わられることによって、身体はコミュニケーション手段から道具使用へとその利用目的が広がっていった。音声言語を獲得する以前では、言語が参照する実世界のモノを他者に伝えるには、そのモノを指さすことやその形に身体を似せる手段を取っていたであろう。ジェスチャ言語を獲得せずに、言語が指すモノを音声だけで伝えることは困難を極める。なぜな

ら、音声言語とそれを参照するモノに物理的な絶対的結びつきはなく、社会に依存した恣意的な結びつきがあるだけだからである [Saussure 66]。

Rizzolatti らによって発見されたミラーニューロンは、身体が言語の基盤であるとする考えを支持する [Rizzolatti 01]。マカクザルの前頭葉の特定の神経細胞は、腕を伸ばして物を掴む動きに対して極めて高い活動を示した。この神経細胞は、同じ人間の動きを観察したときにも反応した。この行為と知覚に同一の神経細胞が関与することは、単語と実世界の参照構造を他者と共有するのに大いに役立つ。また、このミラーニューロンは、音声言語処理を行うブローカ野に存在することも、身体・ミラーニューロン・言語が強く結び付くという推測を掻き立てる。

ロボティクス分野では、多自由度の身体運動の設計論として、手取り足取りロボットの身体を動かし、その運動軌跡を学習する Programming by Demonstration や、ヒューマノイドロボットと人間の身体の類似性を活用し、人間の動きをロボットの身体に写像して得られる運動軌跡を記憶する Imitation Learning が研究されてきている

[Argall 09, Breazeal 02]. 運動の軌跡は、力学系 [Okada 02, Ijspeert 03, Tani 03, Kadone 05] や統計モデル [Inamura 04, Billard 06, Asfour 06, Kulic 07] のパラメータとして抽象的に表現されることになる。これは、連続世界の運動軌道をパラメータ空間の点として離散的に表現したことに相当し、身体運動を記号として記憶している知能の枠組みである。身体運動の記号化を基盤として、Hidden Markov Model のパラメータを修正可能にした Parametric Hidden Markov Model による物体操作の運動認識・生成 [Herzog 08] や物体の視覚情報、運動知覚、言語の力学系を表現する複数のニューラルネットワークを結びつけることによって、言語を獲得する研究 [Sugita 05] へ拡張されている。

運動から言語を獲得する研究として、身体運動と単語の間の参照関係を抽出することで、単語の意味を抽出する方法も提案されている。身体内部の力覚と物体の変化をベクトルとして表現し、そのベクトル空間を分類することで動詞の概念空間を構築することができる [Hemerik 12]。ここでは、動詞とそれが指し示す概念領域の対応関係は人手によって付与される。動詞から唯一の概念が決定されることになり、言語の多義性を扱うまでには至っていない。また、運動記号と単語の関係を統計モデルによって表現することによって、単語から複数の運動記号が確率的に連想される枠組みも提案されている [Takano 07]。この単語と身体運動の参照関係は、身体に基づく言語の獲得やその構造解析へ適用することができる。自然言語情報処理の分野において、同じ文書中での単語の共起性に基づいて単語の隠れ状態や概念を抽出する Latent Semantic Analysis (LAS) [Deerwester 90], Probabilistic Latent Semantic Analysis (PLAS) [Hofmann 99], Latent Dirichlet Allocation (LDA) [Blei 03] が提案されている。これらは、同じ文書で用いられる 2 つの単語は類似した意味を含むという仮定から出発しており、単語が参照する実世界の情報は無視されている。

本論文では、身体運動と単語の参照構造を統計モデルによって学習することで、単語が各運動を指し示す確率を抽出する。この確率に基づいた統計的尺度を用いて単語間の距離を計算し、この距離構造を満足するように単語を多次元空間上に射影することで言語空間を構築する。それによって、身体運動から単語の関係性を抽出することが可能となる。また、身体運動を言語空間上に写像する計算を提案し、運動を言語として理解する方法を提案する。光学式モーションキャプチャにて計測した人間の全身運動データを用いて言語空間を構築し、運動を言語として理解する実験を通して提案する枠組みの有効性を示す。

図1 運動から言語空間の構築。運動データを HMM によって学習した運動記号と単語の参照関係を統計モデルによって抽出する。参照構造から単語間の距離を計算する。この距離構造を満足するように単語を多次元空間に布置することで言語空間を作成する

2. 運動と言語の参照構造から言語空間の構築

2.1 運動と言語の参照構造の学習

図1に示すように、運動と単語の参照構造を抽出する。その構造から単語間の関係性を多次元空間上に表現することで言語空間を構築する。

運動パターンデータを学習データとして、各運動パターンに対応する HMM のパラメータを最適化する。HMM は時系列データを入力とし、その時系列データがどのカテゴリに含まれるかを決定するための確率モデルである [Rabiner 89]。HMM は、第 i 番目の状態 q_i から第 j 番目の状態 q_j へ遷移する確率 a_{ij} の集合である状態遷移確率行列 $A = \{a_{ij}\}$ 、第 i 番目の状態が特徴量ベクトル x を出力する確率分布の集合 B 、および初期状態確率の集合 $\Pi = \{\pi_i\}$ の 3 組の集合 $\lambda = \{A, B, \Pi\}$ によって表される。HMM の学習は、非観測の状態遷移の分布を推定する E-step とその状態遷移分布に従って特徴量ベクトルの時系列 X が生成される確率が最大となるようにパラメータ集合 $\lambda = \{A, B, \Pi\}$ を最適化する M-step を繰り返す EM アルゴリズムを用いる [Rabiner 93]。このように、運動パターンデータは HMM のパラメータ集合として記号化され、この HMM を運動記号と呼ぶことにする。

運動データ O を、自動で運動セグメント $O[k]$ に分節化する [Takano 06]。

$$O = \{O[1], O[2], \dots, O[n]\} \quad (1)$$

自動分節化は、運動の時系列から次の全身姿勢の予測が困難な時点を運動パターンの境界として判定することで、運動セグメントを抽出し、その数 n は自動で決定される。予め運動パターンデータを学習した運動記号から、各運動セグメント $O[k]$ が生成される尤度 $P(O[k]|\lambda)$ を Forward アルゴリズムによって計算する。運動セグメン

図2 一連の運動データは、運動記号系列と単語系列として表現できる。各系列における運動記号の場所と単語の場所が結びついている確率（アライメント確率）と、特定の運動記号と単語が結びついている確率（翻訳確率）を抽出する。アライメント確率は系列間の位置の関係性、翻訳確率は記号と単語の種類間の関係性を示している

ト $O[k]$ は、この尤度が最大となる運動記号 $\lambda_{\mathcal{R}}[k]$ として認識されることになる。

$$\lambda_{\mathcal{R}}[k] = \arg \max_{\lambda} P(O[k]|\lambda) \quad (2)$$

この自動分節化と運動認識を運動データ O に適用することによって、運動を運動記号の系列 Λ として表現することができる。

$$\Lambda = \{\lambda_{\mathcal{R}}[1], \lambda_{\mathcal{R}}[2], \dots, \lambda_{\mathcal{R}}[n]\} \quad (3)$$

同じ運動データについて、人手で分節化して運動を指し示す単語を付与する。すなわち、運動データ O は、単語の系列 Ω として表現することもできる。

$$\Omega = \{\omega[1], \omega[2], \dots, \omega[m]\} \quad (4)$$

ここで、 ω は単語を表している。自動分節化と人手による分節化では運動パターンの境界が異なるため、運動記号系列に含まれる運動記号の個数 n と単語系列に含まれる単語の個数 m が必ずしも一致するわけではない。

運動記号系列と単語系列の対応関係を表現する問題を翻訳問題とみなし、図2に示すように、IBM 翻訳モデル [Brown 93] を用いて運動記号と単語の参照構造を抽出することにする [Takano 07]。IBM 翻訳モデルは、順番に少しずつ複雑になるモデル1からモデル5まで提案されている。

運動記号系列と単語系列の変換の基盤には、各単語がどの運動記号を指し示すかを表現する単語と運動記号間の参照構造、運動記号の並び、および単語の並びに関する語順の規則、そして語順の規則間を繋げる単語および運動記号の位置の対応関係が存在する。本研究で扱う運動記号系列および単語系列の語順は運動の時間に沿うものとして、複雑な語順規則を想定していない。運動記号の語順、単語の語順に関する規則を無視し、各運動記号と各単語の参照関係、および運動記号系列と単語系列内での運動記号と単語の位置の対応関係のみを用いて運動記号系列と単語系列の間の変換を行う。したがって、本

論文ではモデル2を採用する。翻訳モデル2は、単語 e が単語 f へ翻訳される確率 $t(f|e)$ 、文長 l の文 e における位置 i の単語が文長 m の文 f における位置 j の単語へ接続する対応付け確率 $a(i|j, m, l)$ によって構成される。翻訳確率として、単語 ω が運動記号 λ を参照する確率 $t(\lambda|\omega)$ が抽出できる。単語系列 Ω から運動記号系列 Λ が生成される確率が最大となる対応付け確率 $a(i|j, m, l)$ と参照確率 $t(\lambda|\omega)$ は、EM アルゴリズムによって求めることができる。

2.2 言語空間の構築

単語 ω が運動記号 λ を参照する確率 $t(\lambda|\omega)$ を用いて、各単語間の非類似度を Kullback Leibler 情報量として計算する。

$$d(\omega_i, \omega_j) = \sum_{k=1}^{N_{\lambda}} t(\lambda_k|\omega_i) \{\ln t(\lambda_k|\omega_i) - \ln t(\lambda_k|\omega_j)\}$$

$$D(\omega_i, \omega_j) = \frac{d(\omega_i, \omega_j) + d(\omega_j, \omega_i)}{2} \quad (5)$$

$d(\omega_i|\omega_j)$ は、単語 ω_i から単語 ω_j への Kullback Leibler 情報量である。式(5)では、 $d(\omega_i|\omega_j)$ と $d(\omega_j|\omega_i)$ を足し合わせて、対称性を満足する $D(\omega_i, \omega_j)$ を求めている。これを単語 ω_i と単語 ω_j の距離として定義する。

すべての単語間の距離を満足するように各単語を多次元空間上の点として布置する。すなわち、単語 ω_i の多次元空間上の点の座標を x_i として、以下の誤差関数が最小となるような空間上での単語の位置を定める。本論文では、Takane らが提案する多次元尺度法を用いる [Takane 77]。

$$T = \sum_{\forall i, j} \frac{(D(\omega_i, \omega_j)^2 - d_{ij}^2)^2}{4D(\omega_i, \omega_j)^2} \quad (6)$$

$$d_{ij} = \|x_i - x_j\| \quad (7)$$

誤差関数 T は、単語 ω_i の座標 x_i の4次多項式として表され、単語の最適な位置はニュートン・ラフソン法によって求めることができる。このように、多次元尺度法を通じて単語の非類似度に基づいた言語空間を構築する。

2.3 言語空間を用いた運動認識

人間は実世界の連続情報を知覚し、記号表現として抽象化することで、情報を圧縮し、計算量の少ない高速な推論を可能にしている。情報量が膨大な実世界とそれを記号化した離散世界を結ぶ情報処理はロボットの知能に求められる。本論文では、全身運動を言語空間上で表現することで、連続世界と離散世界の懸け橋となる情報処理を提案する。これは、ヒューマノイドロボットが人間の運動を観察し、言語として認識する知能の枠組みである。

図3に示すように、運動データを言語空間上の点へ射影する。人間の全身運動を分節化して得られる運動セグメントを運動記号として認識する。運動の認識は、式(2)

図3 運動データは自動分節化されて、運動セグメントの系列に分解される。各運動セグメントは、運動記号として認識され、言語空間上の点に写像される。運動が言語空間上の点の系列として表現される。

に示すように観察された運動セグメントを生成する尤度が最大の運動記号を見つけることによって実現できる。この認識によって得られる運動記号 $\hat{\lambda}$ を言語空間へ射影する計算は、運動記号 $\hat{\lambda}$ を確率 1 で生成する仮想単語 $\hat{\omega}$ を言語空間上の点として布置することに相当する。仮想単語 $\hat{\omega}$ が運動記号を参照する確率を以下のように与える。

$$t(\lambda_i|\hat{\omega}) = \begin{cases} 0, & \lambda_i \neq \hat{\lambda} \\ 1, & \lambda_i = \hat{\lambda} \end{cases} \quad (8)$$

この参照確率 $t(\lambda_i|\hat{\omega})$ を使って、仮想単語 $\hat{\omega}$ と全ての単語 ω の距離 $D(\hat{\omega}, \omega)$ を式 (5) に基づいて計算する。この距離構造を満足するように多次元尺度法によって、仮想単語 $\hat{\omega}$ の言語空間上での位置 $\hat{x}$ を算出する。このようにして、運動を言語空間上へ射影することができる。

3. 実験結果

3.1 人間の全身動作から言語空間の形成

被験者に貼り付けた 34 個のマーカーの位置を光学式モーションキャプチャシステムにて計測する。20 自由度の人間の身体モデルに基づいて、逆運動学計算を通じてマーカーの位置から関節角度 $(\theta_1, \theta_2, \dots, \theta_{20})$ およびルートリンク（体幹）の鉛直方向の高さ z 、水平方向の速度 $\dot{x}, \dot{y}$ 、ロール ϕ_r 、ピッチ ϕ_p 、およびヨーレート $\dot{\phi}_y$ を計算した。すなわち、身体運動は 26 次元の特徴量ベクトル $\mathbf{o} = [\theta_1, \theta_2, \dots, \theta_{20}, z, \dot{x}, \dot{y}, \phi_r, \phi_p, \dot{\phi}_y]^T$ の時系列として表現される。本実験では、野球に関する運動データを 537 個 $\{O^{(i)} : i = 1, 2, 3, \dots, 537\}$ 計測した。計測した運動データの総時間は 4088[s] であり、運動データの計測周波数は 30Hz である。

計測した運動データから自動分節化および競合学習を通じて 200 個の運動記号を獲得した [Takano 06]。各運動

図4 言語空間の次元数と、Kullback Leibler 情報量として計算された単語間距離と多次元空間上に配置された単語間から計測される距離の誤差の関係を示す。次元数を増やすにつれて、Kullback Leibler の距離構造を満足するように単語が空間上に布置されていることを示している。距離誤差は、10 次元空間で収束している。

データ $O^{(i)}$ は、運動認識処理を通じて運動記号の系列として表現される。また、同じ運動データに人手により単語の系列が付与される。運動データに付与された単語の種類は 61 個であった。表 1 は、運動データに付与された運動記号の系列と単語の系列の一例を示している。運動データ $O^{(1)}$ は、160 番目の運動記号 λ_{160} と 181 番目の運動記号 λ_{181} の系列として認識され、左打者が振る動作を表す単語 “left_swing” が付与されている。運動データ $O^{(5)}$ は、運動記号の系列 $\{\lambda_{116}, \lambda_{43}, \lambda_{118}, \lambda_{22}\}$ と認識され、左打者が振って走る動作を表す単語の系列 $\{\text{“left_swing”, “run”}\}$ が付与されている。

運動記号の系列と単語の系列から単語から運動記号を参照する確率 $t(\lambda|\omega)$ を最適化した。この参照確率を用いて、全ての単語間の距離を計算し、この距離構造を満足するような言語空間を構築した。図 4 に式 (5) によって求められる単語間の距離と多次元空間上に単語を配置した時の距離の平均誤差と言語空間の次元数の関係を示す。言語空間の次元数を増やすにつれて誤差は減少しており、10 次元で誤差が収束している。10 次元の言語空間を構築し、その第 1 軸と第 2 軸によって規定される平面に投影された言語空間を図 5 に示す。この図を見ると単語 “run”(走る動作を指す) と単語 “walk”(歩く動作を指す)、および単語 “raise_hands”(両手を挙げる動作を指す) と単語 “wave_hands”(両手を振る動作を指す) が互いに近くに配置されていることが確認できる。

表 2 は、距離が近い単語の組み合わせと、距離が遠い単語の組み合わせを示している。最も類似している単語の組は、単語 “raise_hands”(両手を挙げる動作を指す) と単語 “wave_hands”(両手を振る動作を指す) であった。2 番目に類似している単語の組は、単語 “raise_right_hand_running”(走りながら右手を挙げる動作を指す) と単語 “raise_right_hand_walking”(歩きながら右手を挙げる動作を指す) であった。3 番目に類似している動作は、単語 “raise_right_hand”(右手を挙げる動作を指す) と単語 “raise_right_hand_walking”(歩きながら右手を挙げる動作を指す) であった。最も類似している単語 “raise_hands”

表 1 運動データに表現する運動記号の系列と単語系列の組み合わせ． $O^{(1)}$ は、160 番目の運動記号と 181 番目の運動記号の系列として認識されている．また、この運動データには、人手によって単語 “lefty’s_swing” (左打者がスイングするを指し示す) が付与されている．運動記号系列と単語系列の組み合わせが、単語と運動記号の参照構造を抽出するための学習データとして利用される．

$O^{(1)}$	Motion Symbols	Word Lables
$O^{(2)}$	160 181	“lefty’s_swing”
$O^{(3)}$	160 43 181	“lefty’s_swing”
$O^{(4)}$	55 43 181	“lefty’s_swing”
$O^{(5)}$	116 43 118 22	“lefty’s_swing” “run”
$O^{(6)}$	68 43 118	“lefty’s_swing” “run”
$O^{(7)}$	17 43 50 22	“lefty’s_swing” “run”
$O^{(8)}$	85 4 103 22	“righty’s_swing” “run”
$O^{(9)}$	61 4 103 53	“righty’s_swing” “run”
$O^{(10)}$	85 4 121 22 38	“righty’s_swing” “run”
$O^{(11)}$	109 30 30 145 108	“run” “sliding” “stand_up” “standing”
$O^{(12)}$	34 30 30 79 108	“run” “sliding” “stand_up” “standing”
$O^{(13)}$	109 30 64 64 111 108	“run” “sliding” “stand_up” “brush” “standing”
$O^{(14)}$	89 109 4 64 64 108	“run” “sliding” “stand_up” “brush” “standing”
$O^{(15)}$	108 126 126 126 93 93	“standing” “walk”
$O^{(16)}$	28 174 174 174 44 44	“walk” “run”
$O^{(17)}$	124 93 44 44	“standing” “run”
$O^{(18)}$	35 53 53 44	“run”
$O^{(19)}$	115 174 144 95 44	“walk” “raise_right_hand” “run”
$O^{(20)}$	44 44 174 144	“run” “raise_right_hand” “run”

と単語 “wave_hands” , および 2 番目に類似している単語 “raise_right_hand_running” と単語 “raise_right_hand_walking” から各々運動データの時系列を生成した．各運動記号 λ である HMM が持つ運動の分布に従ってモンテカルロ法を用いて生成される一定長さの運動セグメント O_λ と、単語がその運動記号を参照する確率 $t(\lambda|\omega)$ の重み付け平均として、単語から運動データ O_ω を生成した．

$$O_\omega = \sum_{\forall \lambda} t(\lambda|\omega) O_\lambda \quad (9)$$

同一の単語が複数の運動に対応する言語の多義性が、単語から運動記号を連想する分布として表されている．これは、多義性の分布と運動記号が持つ運動軌道の分布に基づいて平均的な動きを生成していることになる．

図 6 に類似した単語から生成される全身の運動と右肩、左肩の関節角の時系列を示す．単語 “raise_hands” と単語 “wave_hands” から生成される動きの類似性、および単語 “raise_right_hand_running” と単語 “raise_right_hand_walking” から生成される動きが酷似していることが確認できる．また、単語 “raise_hands” と単語 “wave_hands” から生成される関節角度の平均誤差は 0.053rad であり、単語 “raise_right_hand_running” と単語 “raise_right_hand_walking” から生成され関節角度の平均誤差は 0.056rad であった．単語間の距離が最も大きい単語の組は、単語 “raise_left_hand_runnig” (走りながら左手を挙げる動作を指す) と単語 “sitting” (座っている動作を指す) であった．2 番目に離れている単語の組み合わせは、単語 “touch_walking” (歩きながらタッチする動作を指す) と単語 “hold” (抱える動作を指す) であった．図 6 に、これら単語から生成される運動データおよび全身運動を示す．単語 “raise_left_hand_runnig” 図 5 に示す言語空間上の単語を 20 個のクラスタに分類と単語 “sitting” から生成される運動のルートリンクの鉛

直軸方向の高さが約 0.4m 離れていることや、右肩関節角度が 0.2rad 程度離れていることが確認できる．また、生成された運動の関節角の平均誤差は 0.22rad であった．このように、単語が参照する運動の時系列データからも単語間の距離構造はそれが指し示す運動の大きさの違いを表現していることが確認できる．

また、類似した運動を参照する単語が言語空間上で近くに配置されているかどうかを検証した．言語空間上にある単語の分布を混合正規分布によって表現した．空間上の位置 x に単語が布置される尤度 $P(x)$ を

$$P(x) = \sum_{k=1}^K w_k \mathcal{N}(x; \mu_k, \Sigma_k) \quad (10)$$

として表す． w_k, μ_k, Σ_k は、混合正規分布を表現するパラメータであり、第 k 番目の正規分布の重み、平均ベクトル、共分散行列である． K は重ね合わせる正規分布の数である．混合正規分布のパラメータは、この分布から単語 $\omega_i (i = 1, 2, \dots, N)$ が空間上の位置 $x_i (i = 1, 2, \dots, N)$ に布置される尤度 $\phi = \prod_i P(x_i)$ を最大化するように最適化される．最適化の計算法として、EM アルゴリズムを用いる．単語の分類は、各正規分布から単語の位置 x が生成される確率が最大となる正規分布を求めることに相当する．

$$\begin{aligned} \mathcal{K} &= \arg \max_k P(k|x) \\ &= \arg \max_k \frac{w_k \mathcal{N}(x; \mu_k, \Sigma_k)}{\sum_l w_l \mathcal{N}(x; \mu_l, \Sigma_l)} \\ &= \arg \max_k w_k \mathcal{N}(x; \mu_k, \Sigma_k) \end{aligned}$$

した．各単語の位置に布置されたマークの色および形が

図5 単語が運動を参照する確率から単語間の距離を Kullback Leibler 情報量として計算する。この距離構造を満足するように、単語を多次元空間上の点として布置することで言語空間を構築する。10 次元の言語空間を第 1 軸と第 2 軸によって規定される平面上へ射影した空間を示す。空間上の各点が単語を示している。単語を 20 個のグループにクラスタ分類し、各マークの色・形はグループ毎に異なる。

クラスターのインデックを表している。表 3 は、各クラスターに属する単語を示している。15 番目のクラスターには、単語 “lefty’s_swing” (左打者がスイングする動作を指す) と単語 “righty’s_swing” (右打者がスイングする動作を指す) が含まれている。表 1 からわかるように、スイングの動作は一塁へ走る運動を伴うことが頻繁にある。そのため、単語 “lefty’s_swing” と単語 “righty’s_swing” はスイング動作を参照するだけでなく、走る動作も参照する。スイングの類似性と運動の時系列による文脈性を帯びた単語から運動の参照構造が形成される。したがって、この 2 つの単語は類似した単語として判断される。12 番目のクラスターには、右手を挙げる運動に関連する単語 “raise_right_hand”, “raise_right_hand_running”, “raise_right_hand_walking” が含まれている。6 番目のクラスターには、単語 “jump” (ジャンプする動作を指す), “raise_right_hand_held” (持ち上げられて右手を挙げる動作を指す), “raise_right_hand_jumping” (ジャンプしながら右手を挙げる動作を指す) が含まれている。この 3 つの単語が参照する動作は、動作主の体幹の位置が高いことが共通としており、その類似性から同じクラスターとして分類されている。11 番目のクラスターは、単語 “catch_ball” (ボールを捕球する動作を指す), “take_steps_backward” (捕球するために後ろに下がる動作を指す), “catch” (捕球する動作を指す), “catch_fly_ball” (高く上がったボー

ールを捕球する動作を指す) から構成されており、全ての単語がボールを捕る動作に関連している。これらの類似した動作を参照する単語が言語空間上で近い距離に配置されていることが確認できる。また、関節角の動きで見ると明らかに違うことが予測される単語 “stand_up” (立ち上がる動作を指す) と単語 “sitting” (座っている状態を指す) が 1 番目のクラスターに含まれている。座っている動作から立ち上がる動作へ遷移する一連の行動データに対して、単語系列 { “sitting stand_up” } が付与されており、両単語がともに「座っている」や「立ち上がる」に相当する運動記号を参照することがその原因と考えられる。

3.2 言語空間を通じた動作理解

人間の全身動作を言語空間上へ射影することによって、動作を言語空間上の軌跡として表現する。分節化された運動セグメントを運動記号として認識し、その運動記号を参照する仮想単語を言語空間上に求める計算として実現される。左打者がバットを振った後、一塁へ走り出す動作を、言語空間上へ射影した。図 7 に、射影された点の系列を示す。点の系列は、単語 “lefty’s_swing_pose” (左打者が振る構え動作を指す) 付近から出発し、単語 “lefty’s_swing” (左打者がスイングする動作を指す) と単語 “turn” (体の向きを変える動作を指す) を経由して、単語 “run” (走る動作を指す) へ到達している。同様に、「選手が飛び込んで、

表 2 距離が近い単語の組み合わせと距離が遠い単語の組み合わせを表示．

word1	word2	distance
raise_hands	wave_hands	0.0893
raise_right_hand_running	raise_right_hand_walking	0.0967
raise_right_hand	raise_right_hand_walking	0.1012
⋮	⋮	⋮
hold	next_batter's_bend	1.0822
touch_walking	hold	1.0892
raise_left_hand_runnig	sitting	1.0989

average distance 0.7184

standard deviation 0.2089

図 6 言語空間上で類似した単語から生成される運動(左図)と離れた単語から生成される運動(右図)の比較を示す．単語“raise_hands”と単語“wave_hands”から生成される運動は類似していることが確認できる．

立ち上がり，ボールを投げる」動作を言語空間上に射影すると，単語“dive”（飛び込む動作を指す）から出発し，単語“bend_down”（屈む動作を指す），単語“stand_up”（立ち上がる動作を指す），単語“throw_pose”（投げる構えを指す）を経由して，単語“throw”（投げる動作を指す）へ到達する軌道が得られる．このように，人間の全身動作が，言語空間上の点の軌跡として適切に射影されており，動作を言語空間で理解できていることが確認できる．これは，連続世界の運動を離散世界の言語へつなげる情報処理の基盤となり，構築した言語空間の一応用例である．構築した言語空間は，これまでのテキスト世界に閉じた単語間の繋がりではなく，実世界の事象を考慮した単語の関係性を抽出する技術であり，Web 検索・文書解析への展開も期待できる．

4. 結 論

本研究で得られた結論は以下の通りである．

- (1) 一連の運動を運動記号の系列と単語の系列として表現する．これら 2 つの系列の対応関係から単語が

運動記号を参照する確率を抽出した．単語が参照する運動に基づいて単語間の距離を計算し，この距離構造を満たすように単語を多次元空間上の点に布置することで，単語間の類似性を表現した言語空間の設計法を提案した．

- (2) 一連の運動を運動記号として認識し，その認識結果を言語空間へ射影する計算法を提案した．認識結果に対応する運動記号を確率 1 で参照する仮想単語を導入し，その参照確率を用いて他の単語との距離を計算する．この距離を満足するように仮想単語を言語空間上に布置することで，運動を言語空間へ射影できる．
- (3) 光学式モーションキャプチャシステムにて計測した 4088 秒の人間の全身運動データから言語空間を構築した．用いた運動記号の総数は 200 個，単語の総種類数は 61 である．言語空間の次元数は実験から 10 と定めた．言語空間上の単語をクラスタ解析することで，類似した運動を指し示す単語が空間上で互いに近くに配置されることを確認した．
- (4) 「左打者がバットを振って走り出す」一連の運動

表 3 Clustering analysis result

cluster	word
1	“stand_up” “sitting” “next_batter’s_warm_up”
2	“raise_left_hand” “throw_bat”
3	“standing” “turn” “walk”
4	“run_with_leaning_forward” “protest” “hold_person”
5	“run” “unite_shoulders_walking”
6	“jump” “raise_right_hand_held” “raise_right_hand_jumping” “wave_right_hand”
7	“dive” “bending” “take_stop_pose”
8	“run_with_hands_raised” “warm_up”
9	“dash” “raise_left_hand_running” “raise_hands_walking”
10	“slide” “wipe_the_leg”
11	“catch_ball” “take_steps_backward” “catch” “catch_fly_ball”
12	“touch_walking” “raise_right_hand” “raise_right_hand_running” “raise_right_hand_walking”
13	“bend_down” “throw”
14	“jump_and_catch” “wait_for_fly_ball”
15	“lefty’s_swing” “righty’s_swing” “touch_running” “next_batter’s_bend”
16	“raise_hands” “wave_hands”
17	“ready_to_dash” “ready_to_catch” “take_safe_pose” “touch_runner”
18	“first-baseman’s_catch” “throw_pose” “take_go_pose” “catch_ground_ball”
19	“shake_body” “applaud” “applaud_walking” “lefty’s_swing_pose”
20	“lefty’s_touch_walking” “shake_right_hand” “shake_right_hand_running” “throw_bat_with_both_hands”

データを言語空間上へ射影する実験を行った．射影して得られた点の系列は「構える」「バットを振る」、「体の向きを変える」「走る」の各動作を指し示す単語の付近を通ることを確認した．連続世界の運動を記号世界の言語空間上の軌道として理解する基盤技術であることを示した．

本研究では、全身運動を記号として扱い、その記号と動詞に相当する単語の関係を統計モデルとして表現している．実世界の情報として全身運動の表現形として関節角という単一のモダリティのみを考えており、単語の種類も動詞に限定せざるを得ない．しかし、運動は必ずしも関節角情報のみだけではなく、モノの操作などのように身体が働きかける環境をも包含するものである．運動を関節角のみならず視覚・力覚などのマルチモーダルな情報が混然一体となった身体表出として捉えることで、言語世界の単語を動詞から名詞・動作修飾詞へと拡張していく枠組みが必要である．

謝 辞

本稿に掲載した我々の研究成果は、科学研究費補助金基盤研究(S)(20220001)および独立行政法人科学技術振興機構さきがけ(10101610)の支援を受けて得られたものである．

◇ 参 考 文 献 ◇

- [Argall 09] Argall, B., Chernova, S., Veloso, M., and Browning, B.: A survey of robot learning from demonstration, *Robotics and Autonomous Systems*, Vol. 57, No. 5, pp. 469–483 (2009)
- [Asfour 06] Asfour, T., Gyarfas, F., Azad, P., and Dillmann, R.: Imitation learning of dual-arm manipulation task in humanoid robots, in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pp. 40–47 (2006)
- [Billard 06] Billard, A., Calinon, S., and Guenter, F.: Discriminative and Adaptive Imitation in Uni-manual and Bi-manual Tasks, *Robotics and Autonomous Systems*, Vol. 54, pp. 370–384 (2006)
- [Blei 03] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)
- [Breazeal 02] Breazeal, C. and Scassellati, B.: Robots that imitate humans, *Trends in Cognitive Sciences*, Vol. 6, No. 11, pp. 481–487 (2002)
- [Brown 93] Brown, P. F., Pietra, S., Pietra, V., and Mercer, R.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311 (1993)
- [Deerwester 90] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R.: Indexing by Latent Semantic Analysis, *Journal OF The American Society for Information Science*, Vol. 41, No. 6, pp. 391–407 (1990)
- [Donald 91] Donald, M.: Origin of the Modern Mind, in *Harvard University Press, Cambridge* (1991)
- [Hemeren 12] Hemeren, P. and Gardenfors, P.: A framework for representing action meaning in artificial systems via force dimensions, in *Artificial General Intelligence*, pp. 99–106 (2012)
- [Herzog 08] Herzog, D., Krueger, V., and Grest, D.: Parametric hidden Markov models for recognition and synthesis of movements, in *Proceedings of the British Machine Vision Conference*, pp. 163–172 (2008)
- [Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Analysis,

図7 人間の全身運動を運動記号の系列として認識する．その認識結果を言語空間上の点の系列として写像することによって，連続世界の運動を離散世界の言語空間の軌跡として表現することができる．

in *Proceedings of Uncertainty in Artificial Intelligence, UAI' 99*, pp. 289–296 (1999)

- [Ijspeert 03] Ijspeert, A. J., Nakanishi, J., and Shal, S.: Learning control policies for movement imitation and movement recognition, *Neural Information Processing System*, Vol. 15, pp. 1547–1554 (2003)
- [Inamura 04] Inamura, T., Toshima, I., Tanie, H., and Nakamura, Y.: Embodied symbol emergence based on mimesis theory, *International Journal of Robotics Research*, Vol. 23, No. 4, pp. 363–377 (2004)
- [Kadone 05] Kadone, H. and Nakamura, Y.: Symbolic Memory for Humanoid Robots Using Hierarchical Bifurcations of Attractors in Nonmonotonic Neural Networks, in *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2900–2905 (2005)
- [Kulic 07] Kulic, D., Takano, W., and Nakamura, Y.: Representability of Human Motions by Factorial Hidden Markov Models, in *Proceedings of the IEEE/RSJ 2007 International Conference on Intelligent Robots and Systems*, pp. 2388–2393 (2007)
- [Okada 02] Okada, M., Tatani, K., and Nakamura, Y.: Polynomial Design of the Nonlinear Dynamics for the Brain-Like Information Processing of Whole Body Motion, in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1410–1415 (2002)
- [Rabiner 89] Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, in *Proceedings of the IEEE*, Vol. 77, pp. 257–286 (1989)
- [Rabiner 93] Rabiner, L. and Juang, B. H.: Fundamentals of Speech Recognition, in *Prentice Hall Signal Processing Series* (1993)
- [Rizzolatti 01] Rizzolatti, G., Fogassi, L., and Gallese, V.: Neurophysiological mechanisms underlying the understanding and imitation of action, *Nature Reviews*, pp. 661–670 (2001)
- [Saussure 66] Saussure, F. D.: Course in General Linguistics,

McGraw-Hill Book Company (1966)

- [Sugita 05] Sugita, Y. and Tani, J.: Learning semantic combinatoriality from the interaction between linguistic and behavioral processes, *Adaptive Behavior*, pp. 33–52 (2005)
- [Takane 77] Takane, Y., Young, F. W., and Leeuw, J. D.: Nonmetric Individual Differences Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features, *Psychometrika*, Vol. 42, pp. 7–67 (1977)
- [Takano 06] Takano, W. and Nakamura, Y.: Humanoid robot's autonomous acquisition of proto-symbols through motion segmentation, in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pp. 425–431 (2006)
- [Takano 07] Takano, W., Yamane, K., and Nakamura, Y.: Capture Database through Symbolization, Recognition and Generation of Motion Patterns, in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3092–3097 (2007)
- [Tani 03] Tani, J. and Ito, M.: Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment, *IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans*, Vol. 33, No. 4, pp. 481–488 (2003)

〔担当委員：山川 宏〕

2012年12月14日 受理

著者紹介

高野 洩 (正会員)

1976年4月2日生．1999年京都大学工学部物理工学科卒業．2001年同大学大学院工学研究科精密工学専攻修士課程修了．同年三菱自動車工業株式会社入社．2006年東京大学大学院情報理工学系研究科知能機械情報学専攻博士課程修了．2006年同大学特任助教，2007年同大学助教，2009年同大学講師，現在に至る．2010年科学技術振興機構さきがけ研究「情報環境と人」領域研究員を兼任．博士（情報理工学）．IEEE，日本ロボット学会などの会員．ヒューマノイドロボットの統計的脳型情報処理，自動車の知能に関する研究に従事．

中村 仁彦

1954年9月22日生．1977年京都大学工学部精密工学科卒業．1982年同大学大学院博士課程退学，1982年同大学助手．1987年カリフォルニア大学サンタバーバラ校助教授，準教授．1991年東京大学工学部機械情報工学科助教授．現在同大学大学院情報理工学系研究科教授．工学博士．非ホロノミックロボット，脳型情報処理，ヒューマノイドやCGの力学計算，外科手術用ロボットなどの研究に従事．システム制御情報学会，計測自動制御学会，日本コンピュータ外科学会，IEEE，ASMEなどの会員．日本機械学会フェロー，日本IFTOMM会議委員長，IEEE Robotics and Automation Society, Distinguished Lecturer (2006-2007)．