文脈を取り組みながら行動を言語化するクラウドコンピューティング

Generation of Situation-Dependent Description of Human Behavior on Cloud-Computer

高野　渉（Wataru　Takano, Ph. D.）
大阪大学　数理・データ科学教育研究センター
特任教授
(Specially Appointed Professor, Osaka University, Center for Mathematical Modeling and Data Science)
IEEE,　日本ロボット学会　他
受賞：日本機械学会 Robomech 表彰(2019 年)、
IEEE IROS 2018 Best paper award finalist (2018 年)　他
著書：“Human Motion Reconstruction,” Springer Handbook of Robotics(2016 年)　他
研究専門分野：ロボティクス　知能情報学

あらまし

高度な知能の根源は言語である。実世界は連続値の情報に満たされている。その中から重要な情報を取捨選択し、分節・類型化を通じて抽象度の高い記号が獲得される。これによって、高次な推論が可能となるのである。実世界事象の中で、特に人間行動に焦点を当て、人間の身体運動を分類する計算と自然言語処理の数理モデルの類似性に着目して、数学的整合性良く運動を記号・言語化する人工知能の枠組みを提案してきた。本研究では、身体運動の周辺環境情報を行動の文脈として取り入れることによって、身体運動の計測データを詳細かつ正確に言語表現へ変換する計算基盤を構築した。身体運動と環境の各分類結果の組と単語の連想関係を統計的に表現する数理モデル（センサーデータと単語の関係に関する知識）と文章中の単語の並びを統計的に表現する数理モデル（単語の並びに関する知識）を接続する知能の枠組みである。センサを通じて取得された人間全身運動と環境の画像を合わせた行動データから関係の強い単語を出力し、さらに単語の並びに関する知識と照らし合わせながら、単語を整列することによって文章を作文する。環境情報を取り入れることによって行動データから出力される単語の正確性を上げるとともに、単語の並びに関する知識を活用して、文章らしさを保持した文章が生成可能となった

１．研究の目的

人間の身体運動の計測技術とロボティクスの運動学・動力学アルゴリズムの発展は、正確な人間の全身運動データを与えようとしている。これら身体運動データを活用することによって、大自由度・複雑なヒューマンロボットの運動学習、CG アバターを操るゲームコントローラー、人間行動の意図・目的を推論する支援システムへと応用技術は多岐な広がりを見せている。

膨大な人間の運動データが計測・蓄積されつつある。計測した運動データをそのまま記憶しておくのでは、必要なデータを記憶から探し出すことの一例を挙げても、どのように必要なデータにたどり着けばいいのか、何を検索の手掛かりにしたらいいのかと想像すれば、その再利用性が乏しいことが容易に想像がつく。膨大な人間の身体運動データを類別・構造化して、その構造と言語表現を巧みに接続することによって、運動データの記憶から必要な動きを呼び起こし、それらを再利用することができる。

人間の行動とは、身体を動かして外部環境に働きかけて変化を与え、それを知覚して次の動きを生成することの連続である。身体運動だけでは人間の行動を語ることはできず、その周辺の外部環境と合わせって、動きの背後にある行動の意図や目的などが浮き上がってくる。身体運動だけでなく、周辺環境のデータを組み合わせた行動データの類別・構造化を経て、行動を言語化する技術の基盤を構築することが求めらる。

２．研究の背景

身体運動を力学系や確率・統計などの数理モデルによって学習する知能の枠組みは提案されてきている[1][2]。身体運動は関節角の時系列であり、連続値データにて表現されている。それらを数理モデルにて学習することは、運動パターン毎を数理モデルにて代表させることであり、各数理モデルは運動を離散化した記号表現とみなすことができる。各数理モデルは、モデルパラメータとして記述されているため、各数理モデルが何を意味するのか、どのような運動を表現するのかを直感的に理解することはできない。人間が理解して、容易に使用するために、運動の記号を言語表現に

文脈を取り組みながら行動を言語化するクラウドコンピューティング

Generation of Situation-Dependent Description of Human Behavior on Cloud-Computer

変換する必要がある。上述のような運動の記号と言語表現を結びつける枠組みは提案されている [3]。

近年の深層学習の発展は著しく、コンピュータビジョンや自然言語処理にて華々しい研究成果が報告されている[4]。深層学習の枠組みの一つとして、機械翻訳にて利用されている sequence to sequence モデルがある。これは、単語を入力、文章中において次に出現する単語を出力とするニューラルネットワークであり、入力された単語の履歴情報を内部状態として保持することを特徴としている[5]。これによって、ある文章（単語列）から別の文章を生成することが可能となる。このようなデータ列を変換する機能が、身体運動から文章を生成する計算に利用されている[6]。

しかし、上述のような身体運動の記号化・言語化研究では、身体運動の周辺環境情報を活用して、人間の行動データから詳細かつ正確な文章を生成する研究には至っていない。本研究では、人間の身体運動の記号化や深層学習の画像認識処理を組み合わせて、環境情報を文脈として取り込みながら、人間の行動を言語化する計算論を提案する。

３．研究の方法

3.1 運動と環境の類別処理

人間の行動は、身体運動と周囲環境の組わせとして表現される。身体運動は、複数の赤外線カメラにて被験者に貼られてマーカーの位置を計測する光学式や、並進加速度・角速度を計測するセンサを被験者に取り付けて動きを計測する慣性センサ式モーションキャプチャシステムを利用して取得することができる。計測された位置データに整合するように人体骨格モデルの関節角を推定するのが、ロボティクスの運動学計算である。推定された全身の関節角データの時系列を統計モデルの一種である隠れマルコフモデルによって学習する。各隠れマルコフモデルが運動記号に相当し、観察した運動を記号として認識することができる。

周囲環境をビデオカメラにて撮影する。撮影された画像から周囲環境を類別する深層ニューラルネットワークを構築する。ニューラルネットワークは、物体認識で広く利用されている light-weight 畳み込むニューラルネットワーク(Mobile Net)を採用した[7]。入力は画像データであり、撮影された場所が各環境・場面である確率値を出力する。最も確率が高い場面として、環境が認識される。

3.2 運動と環境データから文章生成

連続情報である運動・環境は、運動の記号や場面として離散化された。この離散表現の運動・場面と自然言語を統計モデルによって結び付ける枠組みを考案する。図１にその概略図を示す。提案する枠組みは２つの機能（意味論的機能、統語論的機能）から構成されている。

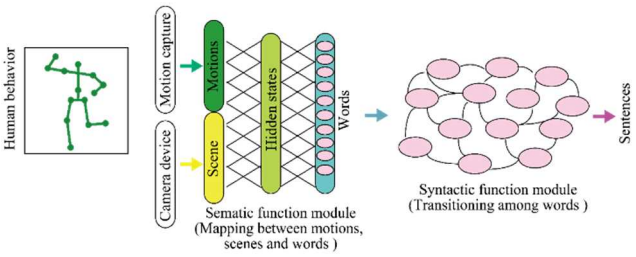

図１　運動と場面の識別・推定から文章の生成。運動と場面から単語を連想し、その単語を並び替えて文章を作成する。

意味論的機能は、運動・場面の離散表現と文章を構成する単語の連想構造を数理モデル化して実装する。この数理モデルのパラメータは、運動・場面の組み合わせから各単語が生成される確率値である。パラメータは、学習データとして与えられた行動データの運動・場面とそれに付与された文章中の単語が再現される確率が最大になるように最適化して導出することができる。

統語論的機能は、文章中の単語の並びを統計数理モデルとして実装される。この数理モデルのパラメータは、ある単語の次に各単語が出現する確率値であり、このパラメータは、学習データの文章が再現される確率が最大となるように最適化して導出することができる。

意味論的機能と統語論的機能を用いて行動データから文章を生成するアルゴリズムを開発する。行動データは、運動と場面の離散値として表現される。これら運動と場面の組み合わせから、意味論的機能を通じて各単語が生成される確率値を計算することができる。

文脈を取り組みながら行動を言語化するクラウドコンピューティング

Generation of Situation-Dependent Description of Human Behavior on Cloud-Computer

生成された単語を並べたとき、その並びの文章らしさが統語論的機能を通じて計算することができる。行動データから関係の深い単語を生成し、それらを整列させて文章らしさの高い単語の並びを探索して、行動を表現するに最適な文章を発見することができる。この探索計算にはダイクストラ法を用いるなどして、効率的に最適解を見つけることができる。

上述のような行動データから言語を生成する処理をクラウドコンピュータ（Amazon AWS）にて計算する基盤を整備した。計測した運動データと撮影した画像をクラウドサーバーに転送し、その中で運動や場面の推定、および言語生成を行う。生成された文章をクライアント計算機に送る機能を備えたシステムである。

3.3 実験結果

約 194 時間にわたり 13 人（男性 8 人、女性 5 人）の全身運動の計測と周囲環境の撮影を行った。図 2 にその一例を示す。Yahoo 利用者が行動の動画を閲覧して、それに説明文を付与するマイクロタスクをクラウドソーシングにて掲載し、膨大な説明文を収集した。収集した文章の総数は約 62,000 個であった。

図２．計測した行動の一例

行動を計測した場所は、運動場、和室、台所、居間、ロビー、ラウンジ、会議室、事務室、休憩室、洋室、研究室、その他である。撮影した動画を、上述の 11 種類に分類したデータセットを作成し、動画から場面を推定するニューラルネットワークを設計した。データセットのうち、85%を学習データとして活用し、残り 15%をテストデータとして利用して、場面推定の精度を検証した。ニューラルネットワークの学習については、全結合パラメータを更新する方法と最終層間の結合パラメータだけを更新する方法（転移学習）を採用し、推定精度を評価した。全結合パラメータを更新する学習則を用いた場合、場面の推定精度は 99%であった。また、転移学習を採用した場合、場面の推定精度は 97%であった。本実験条件では、あらかじめ計算済みのモデルパラメータを再利用し、出力層のパラメータだけを更新するのでも、十分な場面推定能力が獲得できることを確認した。

次に上述の場面推定結果と運動識別結果を組み合わせて文章を生成する実験を行った。場面推定の場合と同様に、行動データの 85%を学習データに、残り 15%をテストデータに利用した。本実験では、1 つの行動データに対して、確率が高い文章を 10 個生成している。図 3 に行動データと生成された文章の一例を示す。

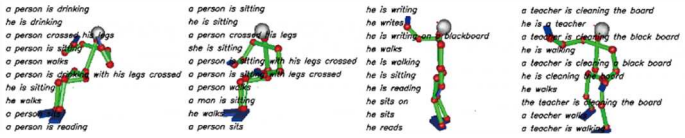

図３．行動データから文章生成の実験結果

例えば、「座りながら飲む」行動データに対して、最も確率が高い文章として “a person is drinking” が生成されている。また、「座りながら足を組む」姿勢から “a person crosses his legs” が第 3 番目の文章候補として生成されている。さらに、それらを複合した “a person is drinking with his legs crossed ” という文章も生成されていることが確認できる。「座りながら作業する」行動データに対して、最も確率が高い文章は “a person is sitting” であり、第 2 番目の文章は、主語が異なる “he is sitting” であった。また、第 3 番目候補として、「足を組む」動作を表現した文章 “a person crosses his legs” が生成されている。さらに、「座る」と「足を組む」の両方を言い表す文章 “a person is sitting with his legs crossed” が生成されている。会議室にて講義をしている「書く」動作に対して、最も確率が高い文章として、“he is writing” が生成されている。さらに、会議室の場面も考慮して、「黒板」という単語への連想を広げて、文章 “he is writing on a blackboard” が生成されている。同様に、会議室にて講義をしている「消す」動作に対して、文

章 “he is cleaning the board” が生成されている。このように、行動に対して正しい文章が生成されていることが確認できる。

生成された文章の正しさを客観的に評価するために、生成された文章と人手によって付与された正解文章の単語 N-gram の一致度合いを数値化した。数値指標として、BLEU (BiLingual Evaluation Understudy) スコアを用いた。BLEU スコアは、0 から 1 の数値を取り、数値が高いほど、2 つの文章はよく一致していることを示す。2 つの文章が完全に一致した場合は 1 を取ることになる。第 1 番目の候補として生成される文章の平均 BLEU スコア値は、0.55 であり、第 2 番目、第 3 番目の候補文章の BLEU スコア値は、0.50、0.40 であった。生成される確率値が高い文章ほど、正確文章と良く一致する傾向があることが確認できた。

４．将来展望

本研究では、人間の身体運動を計測する装置として、光学式もしくは慣性センサ式モーションキャプチャを利用している。これらモーションキャプチャシステムを利用する場合、被験者にマーカーやセンサを貼り付ける計測準備の労力や計測に特化したソフトウェアの使用という制約がある。そのため簡易に行動を計測することができない問題がある。昨今の画像処理技術を駆使して、画像中の関節位置を検出して、全身の姿勢を推定する処理と組み合わせて、人間の行動を言語化するシステムを開発している。これによって、汎用のカメラで行動を撮影するだけで、簡易に身体運動と環境のデータを取り込んで、それを言語化することが可能となる。

おわりに

本研究では、人間の身体運動から言語を生成する計算処理システムを周辺の環境情報を活用できるように拡張することで、詳細に行動を言語できる計算基盤を構築した。身体運動と環境を識別し、その識別結果の組み合わせから単語を連想する計算と、それら単語を整列して文章らしさの高い単語の並びを探索する計算から行動の言語化を実現できる。大規模な行動データを用いて、身体運動と環境データから正しく文章が生成できることを実験によって示した。

参考文献

[1] J. Tani, M. Ito, “Self-organization of behavioral primitives as multiple attractor dynamics : A robot experiment,” IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans, 33(4):481–488, 2003.

[2] W. Takano, Y. Nakamura, “Real-time Unsupervised Segmentation of human whole-body motion and its application to humanoid robot acquisition of motion symbols,” Robotics and Autonomous Systems, Vol.75, PartB, pp.262-272, 2016

[3] W. Takano, Y. Nakamura, “Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions,” International Journal of Robotics Research, Vol.34, No.10, pp.1314-1328, 2015

[4] Y. LeCun Y. Bengio, G. Hinton, “Deep learning,” Nature, Vol. 521, pp.436–444, 2015

[5] I. Sutskever, O. Vinyals, QV. Le, “Sequence to sequence learning with neural networks,” Advances in Neural Information Processing Systems, pp. 3104–3112, 2014

[6] M. Plappert, C. Mandery, T. Asfour “Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks,” Robotics and Autonomous Systems, 109:13–26, 2018.

[7] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017

この研究は、平成２８年度ＳＣＡＴ研究助成の対象として採用され、平成２９～令和元年度に実施されたものです。