動画像中の人の行動に対する言葉での説明への取り組み

†小林 瑞季　　†小林 一郎　　‡ Sergio Gudarrama
†お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース
‡ Electrical Engineering and Computer Sciences, UC Berkeley
†{kobayashi.mizuki, koba}@is.ocha.ac.jp, ‡sguada@eecs.berkeley.edu

1　はじめに

近年，大量の動画像データを取得することが容易になってきている．一方で，大量に収集したデータを有効に利活用出来ているとは言えない．例えば，監視カメラの動画像データに映る内容を把握するためには，全てを人目で見る必要があるが，データの多さに応じた時間を要してしまう．もし，大量の動画像データから特徴的なイベントを捉え，またそのイベントを言葉として表現することが出来たら，動画像データに映る内容を簡単に把握できるとともに，言葉で動画像中のイベントの検索も行うことができると考える．そこで本研究では，動画像中の人の行動を学習によって判別し言葉で表現することを目的とする．

本研究の課題に近い先行研究として，Percy らの研究 [1] が挙げられる．Percy らは，テキストと意味の対応関係を，より少ない教師情報で学習する手法の開発を行っており，事象の状態がデータベースのレコードとして与えられ，それを説明するテキストも同時に与えられていると仮定したとき，レコード内の各部とテキスト中のどのセグメント (区切りの部分) が対応しているかを機械学習により判別する手法を提案している．本研究においても，彼らの手法を参考にし，取得された動画像から得られる時系列データに対して，それを説明するテキストとの対応関係を学習する手法を提案する．

2　研究概要

本研究の概要を図 1 に示す．まず，動画像中に映る人の動作を Kinect[1] のライブラリを用いて捉え，その動きを時系列データとして取得する．取得された時系列データは SAX(Symbolic Aggregate approXimation) を用いて次元圧縮を行い，データベースに格納される．その後，人の動作を説明する文章とデータベース内に蓄積された時系列データの対応関係を学習することにより人の動作を説明するテキスト生成モデルを構築する．

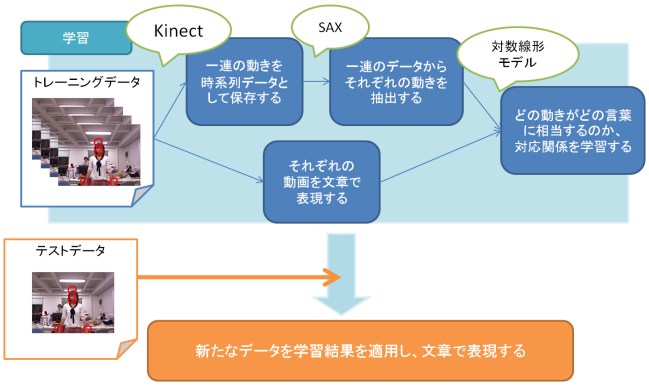

図 1: 研究概要

2.1　動画像取得

Kinect を用いて人物追跡を行った動画像データを取得する．Kinect には一般的な Web カメラでも取得できる RGB 画像に加え，深度センサーやマルチアレイマイクロフォンも搭載されているため，奥行きや音声も保存することができる．また，人物の関節の位置を推定する機能が標準ライブラリに搭載されており，人物の位置を関節単位で 3 次元データで推定することもできる．本研究では，RGB 画像と深度センサー，またそれらを用いた人物の関節位置推定も用い，RGB 動画像と人物の頭・肩の中心・右手・左手の 4 箇所の xyz 座標の時系列データを取得する (図 2 参照)．

2.2　時系列データ処理

2.1 節で取得された動画像データから人の各部の動作を示すの時系列データを取り出す．ここで SAX を用いて，時系列データを文字列に変換する．その文字列群から動作に該当する部分を抽出し，その後，学習

[1]http://www.microsoft.com/en-us/kinectforwindows/

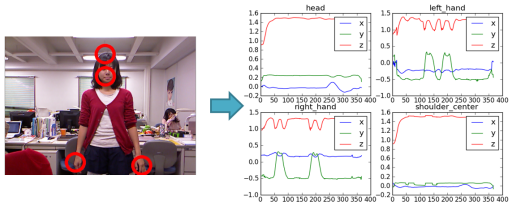

図 2: Kinect を用いた時系列データ取得

するのに適した形にデータを変換する処理を施す．処理に使われる各手法について，以下に詳細を述べる．

2.2.1 SAX

SAX(Symbolic Aggregate approXimation)[2] とは，時系列データの近似表現方法の 1 つで，時系列データを文字列に変換する方法である．SAX を行う際，まず PAA(Piecewise Aggregate Approximation) というデータ圧縮作業を行う．長さ n の時系列データ C を用いて，w 次元の空間ベクトル $\bar{C} = \bar{c_1}, \ldots, \bar{c_w}$ に変換すると仮定する．$\bar{C}$ の i 番目の要素は，式 (1) を用いて計算される．

$$\bar{C_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} C_j \qquad (1)$$

つまり，データを等間隔に w 個のフレームに分け，それぞれのフレーム内でのデータの平均をとることで，n 個ある時系列データを w 個の要素に簡約することができる．正規分布に従って，a,b,c…とアルファベットを割り振り，正規分布の各面積が等しくなるような分割線を定める．上で求めた平均値をこの分割線に従って文字に変換する (図 3 参照)．

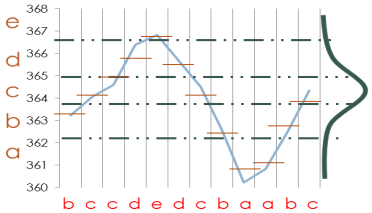

図 3: SAX 法による文字列変換

2.2.2 動きの抽出と圧縮

2.2.1 節の SAX によって変換して得られた文字列から動作とみられる個所を取り出す．ここでは，ある動画像データ中の全ての文字列において一つ前の文字から変化がなければ「動きがない」，変化があれば「動きがある」とみなす (図 4 参照)．

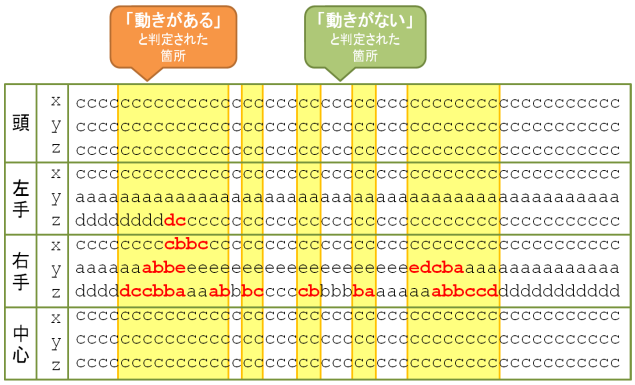

図 4: 動き抽出の例

その後，「動きがある」とみなされた個所の文字列を変化量 (図 5 中のアルファベットの下の数値) に変換し，圧縮する (図 6 参照)．これは同じ動作でも位置やスピードによっては文字列がある一定の間隔でずれたり文字列の長さが変化したりしてしまい，同じ動きとして学習されないためである．これにより，一定の間隔でずれてしまったものも長さが違うものでも，同じ動きとしてとらえることを可能とする．また，より特徴的な動作を抽出するために，圧縮された変化量うち最大の大きさが 2 未満を示す動きは取り除く (図 6 参照)．

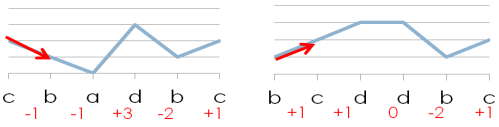

図 5: 文字列の変化量

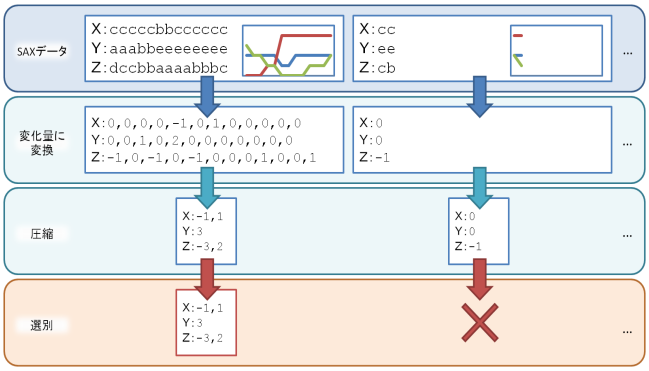

図 6: データの圧縮・選別の例

2.3 時系列データからの動作判別

本研究では，学習に対数線形モデルを用いる．2.2.2 節で処理された動作における人の各部の動きを示すデータ d と，その動作を説明する言語表現 y から，素性関数 $\phi_k(d, y)$ を式 (2) のように定義する．素性は式 (3) の様に表現され，観測された事例の下，構成される素性ベクトルを用い，式 (4) 中の重み $\mathbf{w}$ を学習し求めることで，データが与えられた下での各言語表現が選ばれる確率 $P(y|d)$ を求める (式 (4) 参照)．

$$\phi_k(d, y) = \begin{cases} \text{各部の変動量} & \text{(動作が観測された場合)} \\ None & \text{(それ以外)} \end{cases} \tag{2}$$

$$\phi_k(d, y) = (x_{head}, y_{head}, z_{head}, \ldots, z_{shoulder}) \tag{3}$$

$$P(y|d) = \frac{1}{Z_{d,\mathbf{w}}} \exp(\mathbf{w} \cdot \phi(d, y)) \tag{4}$$

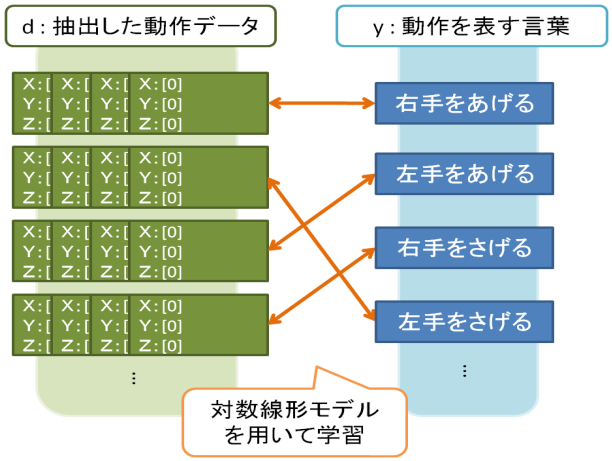

図 7: 動作データと動作を表す表現の対応関係の学習

3 実験

3.1 実験仕様

ここでは簡単に，以下の 8 つの動きを動作を判別，言語化することを目的とする．

- 左手を挙げる/下げる
- 右手を挙げる/下げる
- 両手を挙げる/下げる
- 頭を左に傾ける
- 頭を右に傾ける

これら 8 つの動きを様々なパターンで組み合わせた一連の動きに対し左手・右手・頭・肩の中心の 4 箇所の xyz 座標データを 20 個取得し，その内，学習用データを 15 個，テスト用データを 5 個として学習を行い正解率を測った．

3.2 実験結果・考察

20 個のデータをランダムに 15 個の学習用データと 5 個のテスト用データに分け正解率を測る実験を 5 回繰り返した．その結果，各回の正解率は以下の表 1 のようになった．また，学習によって言語化された一例を図 8 に示す．

実験結果より，5 回とも 0.8 以上の正解率を示しており，高い精度が得られていることが分かる．

表 1: 正解率

	1 回目	2 回目	3 回目	4 回目	5 回目
正解率	0.8	0.8	0.8	0.8	1.0

図 8: 実験結果の例

4 おわりに

本研究では，動画像中の人の行動を学習によって判別し言葉で表現する手法を提案した．具体的には，動画像中の人の動作を表す時系列データに対し，SAX による記号化，さらに圧縮をかけて得られたデータと，動作を説明する言語表現の対を素性とし，機械学習による動作判別のモデルを構築した．

現時点では，観測された動作に対して，動作を示す言語ラベルを付与しているだけとなっており，今後は，一連の異なる動作の説明や生成する言語表現のバリエーションを増やしていくつもりである．

参考文献

[1] Percy Liang, et al. Percy Liang, Michael I. Jordan, Dan Klein : Learning Semantic Correspondences with Less Supervision, ACL-IJCNLP, 2009.

[2] Lin, J. et al. Lin, J., Keogh, E., Lonardi, S. and Chiu, B. : A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, DMKD' 03, 2003.