第 20 回 画像の認識・理解シンポジウム

LSTMを用いたシーンの変化を伴う物体使用手順の記述と想起

矢野 将基[1,a)]　池上 貴之[1]　松尾 直志[1]　島田 伸敬[1]

1. はじめに

近年、少子高齢化に伴う労働力の不足により、介護分野などでロボットの活躍が期待されている。また、各種センサの高性能、低価格化により、ロボットはより多くの情報を取得できるようになっている。これらの点から、ロボットは活躍の幅をさらに広げていくと考えられる。

一方、人間が日常的に扱う物体は同じカテゴリに分類される物体であっても大きさや形が無数に存在する。そのため、物体一つ一つの操作方法を人間が手動で開発するのは負担が大きく、現実的ではない。この問題を解決するための研究として、人間が物体を操作している様子を観察することにより、ロボットがその操作を獲得する研究 [1] などがある。しかし、これらの研究のうち、操作中にロボットや物体の状態が期待と異なった場合の対処について考慮している研究はほとんどない。

また、室内の様子を撮影し続け、人間による動作を検知したり認識したりする研究 [2][3][4] は数多く行われている。撮影された人物と物体の様子から、人物がその物体を操作するときの手順を自動でモデル化し、ロボットの動作プログラムに活用することができれば、プログラム開発にかかる負担を軽減することができる。

本研究では、物体の形状や状態が次々に変化していく操作の様子をロボットが観察することにより、その動作を自動で獲得できるようにすることを目標に研究を行っている。また、ロボットは物体の状態が常に期待する状態であるかどうかを確認し、次の動作の判断を行う。このとき、物体の状態が期待通りでない場合、期待した状態に変更するための動作をロボットが自動で生成し、実行できるようにすることを目指す。ロボット自身が状態変更動作を実行できることにより、物体の初期状態や操作の途中経過に左右されることなく操作の達成を実現できると考えられる。

本稿では、物体を使用するときの人物動作と空間内のシーンの変化の共起性を記述し想起する枠組みを提案する。人物特徴には各関節の位置 (スケルトン)、シーン特徴には深度画像をそれぞれ用い、これらを組み合わせてシーンに対する人物動作の Long Short-Term Memory(LSTM)[5] モデルを学習する。学習したモデルに現時刻における人物とシーンの状態を入力することで、次の人物の行動を想起する。また、初期状態から到達目標状態までのシーン変化の LSTM モデルを学習する。モデルにシーンの初期状態および到達目標となるシーン状態を入力し、到達シーンに至るまでのシーン変化を想起する。

[1] 立命館大学
[a)] yano@i.ci.ritsumei.ac.jp

2. Auto-Encoder を用いた空間の状態の記述

シーンの状態を用いて LSTM モデルを学習するために、状態を低次元のベクトルとして記述する。本研究では Sparse Auto-Encoder(SAE)[6] を用いて、シーンの状態を記述する。SAE はよく似た特徴を持つ画像同士を似たベクトルで、異なる特徴を持つ画像同士は全く異なるベクトルで表現する。

図 1 に今回学習する SAE のネットワーク構造、図 2 に SAE の学習に使用する画像の一部を示す。モデルの入出力は $32 \times 32 \times 1$ch の画像であり、中間層 (エンコード結果) の次元数は 50 次元である。

学習画像にはシーンの深度画像を使用する。本研究では、シーンの状態に対する人間の動作を記述する。このとき、シーンの状態を表現するための情報として、RGB(色) または距離 (形) が挙げられる。形のバリエーションは色に比べると少ないため、学習に必要なサンプル数を抑えることができると考えられる。

学習画像の枚数は 6500 枚であり、図 2 に示す 7 種類の状態が含まれている。今回は問題を簡単化するため、単一の視点から撮影した距離画像を用いる。第 3 節では全身を使用する動作を対象として LSTM モデルを学習するため、人物の全身が撮影できる位置にカメラを固定して空間内の撮影を行う。

学習時に最小化する誤差関数を式 1 に示す。ここで、I はモデルへの入力画像、$E(I)$ は入力画像をエンコードした結果、$D(E(I))$ は入力画像をエンコードし、さらにデコードした結果を表す。また、β、λ は正則化係数である。

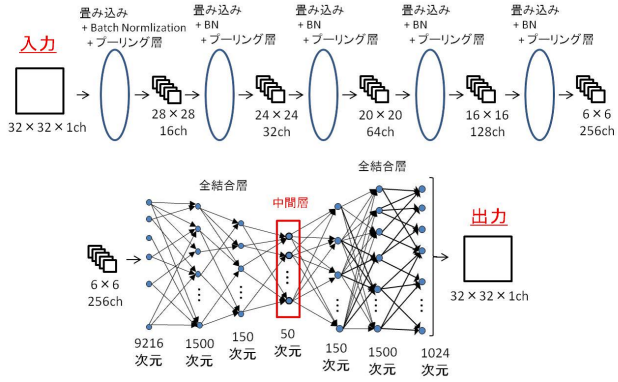

図 1　SAE のネットワーク構造

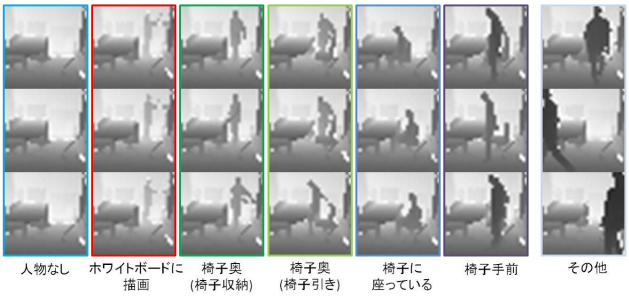

図 2　SAE の学習画像

$$\beta\|I - D(E(I))\|_2^2 + \lambda\frac{\|E(I)\|_{L1}^2}{\|E(I)\|_{L2}^2} \qquad (1)$$

SAE への入力画像とエンコード、デコードを行った結果の画像 (復元画像) の比較を図 3 に示す。いずれの例に関しても復元画像は入力画像を概ね再現できているといえる。

図 4 に SAE によるエンコード結果を主成分分析で 10 次元に圧縮し、その第 1 主成分と第 2 主成分をプロットした結果を示す。似た特徴を持つ画像同士が近くにプロットされ、異なる特徴を持つ画像同士が離れてプロットされている。これらの点から、SAE は期待した通り学習できていると考えられる。

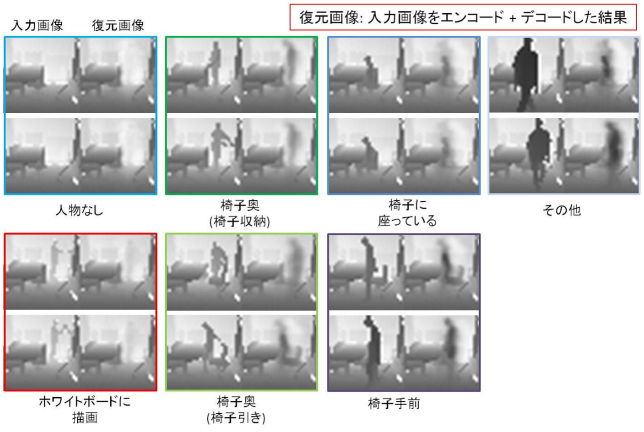

図 3　SAE への入力画像と復元結果の比較

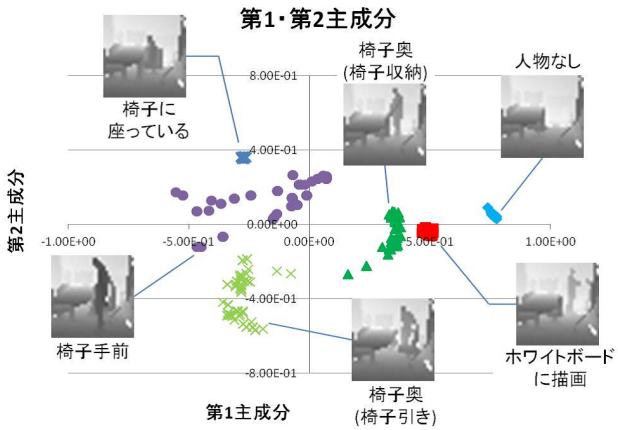

図 4　SAE によるエンコード結果の主成分分析結果

3.　LSTM を用いた人物動作の記述と想起

図 5 は次時刻の動作を想起するためのモデルである。モデルの入力は現時刻における人物のスケルトン (3 次元×25 関節=75 次元) とシーン特徴ベクトル (50 次元) を合わせた 125 次元のベクトルであり、モデルの出力は 1 時刻後のスケルトン (75 次元) である。モデルの中間層は 2 層であり、各層の次元数は 400 次元である。入力としてシーン特徴を使用することにより、椅子があれば座り、なければ通り過ぎるなど、周囲の状況に応じた動作の想起が可能になると考えられる。

図 5 の通り、スケルトンの想起結果をさらに 1 時刻後の想起のための入力として使用している。また、時刻 t におけるシーン特徴として、実際に観測された深度画像を第 2 節で学習した SAE に入力したときのエンコード結果を使用する。

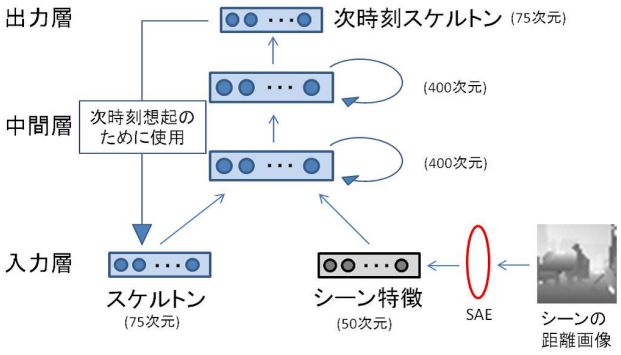

図 5　次時刻スケルトンの想起モデル

モデルの学習には、5314 時刻分のスケルトンとシーン特徴の組からなるデータセットを使用する。このデータセットは、図 6 に示すような動作を複数回含んでいる。

図 7 にモデルを用いて椅子を使用する動作を想起した結果を示す。図に示された青い点は各関節の位置を表わしている。

図 6　LSTM 学習のためのデータセットに含まれる動作の例

想起開始時において、椅子は机の下に収納されており、人物は椅子のそばに立っている。図 7 の上段は椅子を引く動作を行い、想起開始から約 40 時刻後に椅子が引けた場合の想起結果である。椅子を引いた後、椅子に座るという動作が想起できている。

一方、図 7 の下段は椅子を引く動作を行ったものの、椅子を引くことに失敗した場合の想起結果である。想起開始から終了まで、椅子のそばに立って椅子を引く動作を繰り返すという想起結果になっている。このことから、空間の実環境に応じて動作の想起ができていると考えられる。

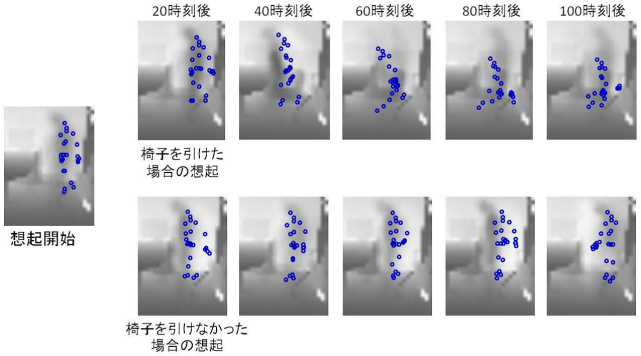

図 7　学習したモデルによるスケルトンの想起結果

4. LSTM を用いた到達目標までのシーン変化の想起

現在の状態から到達目標状態に至るまでのシーン変化を想起するために、図 8 に示す LSTM モデルを構築、学習する。モデルの入力は現時刻のシーン特徴 (シーン画像を第 2 節で学習した SAE に入力したときのエンコード結果) と到達目標となるシーン特徴 (各 50 次元)、目標到達までの時間 (1 次元) を合わせた 101 次元のベクトルである。モデルの出力は 1 時刻後のシーン特徴の想起結果 (50 次元) であり、これをさらに 1 時刻先の想起のための入力として使用する。モデルの中間層は 2 層であり、各層の次元数は 400 次元である。

モデルの学習には 5314 時刻分のシーン特徴を使用する。なお、この学習データはスケルトン想起モデルの学習に使用したものと同じデータである。目標状態には現在の状態から 50 時刻後の実測データを使用する。

各時刻ごとの誤差を式 2 に、最小化する誤差関数を式 3 に示す。式 2 の第 1 項は 1 時刻後の想起精度に関する項、第 2 項はシーケンスの一致精度に関する項、第 3 項は目標

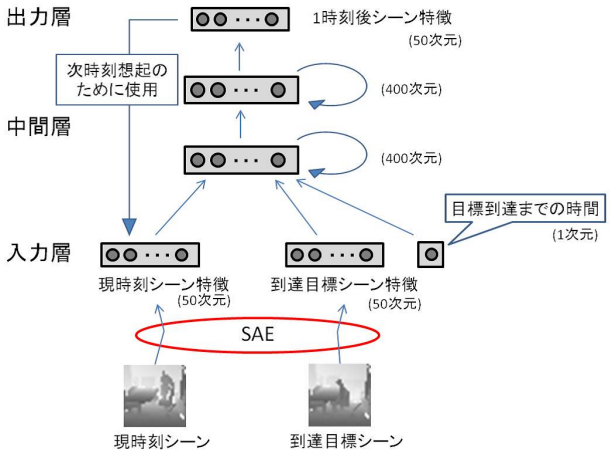

図 8　シーン変化の想起モデル

状態との一致精度に関する項である。ここで、d_t は時刻 t における実測データ、y_t は時刻 t を想起した結果を表す。また、α、β、γ は正則化係数である。

$$E(t) = \alpha(d_{t+1} - y_{t+1})^2 + \beta\left(\frac{1}{N-2}\sum_{i=2}^{N-1}(d_{t+i} - y_{t+i})^2\right) + \gamma(d_{t+N} - y_{t+N})^2 \quad (2)$$

$$E = \sum_{t=0}^{T} E(t) \quad (3)$$

図 9 に想起開始時から 50 時刻先を目標とし、目標までのシーン変化を想起した結果を示す。開始時は椅子から立ちあがった直後の状態であり、目標とする状態は椅子を机の下に収納し、椅子のそばに立っている状態である。図の上段がモデルによる想起結果であり、下段が実際の観測である。これらを比較すると、30 時刻後の想起結果は実際の観測とは異なっているが、それ以外の時刻における想起結果は概ね一致している。

上記の想起に関する各時刻の深度画像の想起誤差を図 10 に示す。人間が椅子の背もたれ側への移動を開始した 30 時刻後の前後は誤差が大きく、29 時刻後には 1px 当たり、平均 11.9cm の誤差が生じた。想起結果は実測よりも人物の移動開始が 3 時刻ほど早かったため、移動前後の想起誤差が大きくなったと考えられる。一方、それ以外のほとんどの時刻に関しては 1px 当たりの平均誤差は 5cm 未満であり、全時刻分の 1px 当たりの平均誤差は 4.39cm であった。

5. おわりに

本稿では、物体を使用するときの人物動作とシーンの変化の共起性を記述し想起する枠組みを提案した。人物の動作とそれによって生起するシーン変化を Long Short-Term Memory(LSTM) モデルを用いて記述し、人物とシーンの状態から人物動作の想起を行った。実際に椅子を使用する様子を対象として動作の想起を行い、椅子の状態に応じた

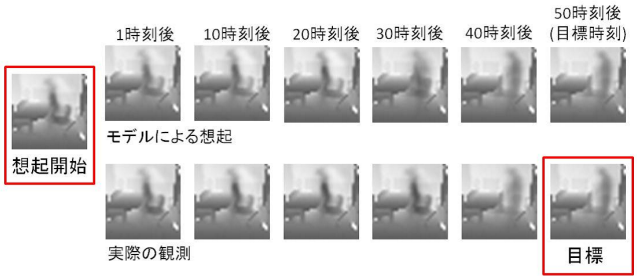

図 9　目標までのシーン変化の想起結果

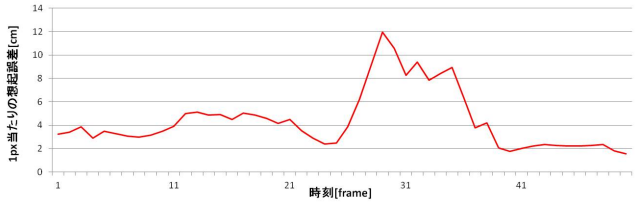

図 10　各時刻ごとの想起誤差

想起ができていることを確認した。また、現時刻におけるシーン特徴から目標となるシーン特徴に至るまでのシーン変化の想起モデルを学習し、想起を行った。今後は、現時点における人物とシーンの状態から人物動作を想起するだけでなく、その動作に伴うシーンの変化を想起するように LSTM モデルを拡張することを目指す。

謝辞　本研究は JSPS 科研費 24500224, 15H02764 の助成を受けたものです。

参考文献

[1] Lee, Kyuhwa, et al. ”A syntactic approach to robot imitation learning using probabilistic activity grammars.” Robotics and Autonomous Systems 61.12 (2013): pp. 1323-1334.

[2] 川本 祥悟, 池上 貴之, 川北 真也, 寺西 研翔, 島田 伸敬. “階層型イベント検知に基づく人と物の関わりのロギングシステム”, 第 18 回画像の認識・理解シンポジウム (MIRU2015), SS5-37, 2015

[3] 森武俊. “1. 生活支援のためのセンサデータマイニング:「みまもり工学」 への展開 (< 小特集> 安全・安心社会実現のためのセンサデータマイニング応用).” 電子情報通信学会誌 94.4 (2011): pp. 276-281.

[4] 入江耕太, 若村直弘, 梅田和昇. “ジェスチャ認識に基づくインテリジェントルームの構築.” 日本機械学会論文集 C 編 73.725 (2007): 258-265.

[5] Hochreiter, Sepp, Jürgen Schmidhuber. “Long short-term memory”, Neural computation 9.8 (1997): pp. 1735-1780.

[6] Tadashi Matsuo, Nobutaka Shimada. “Construction of Latent Descriptor Space of Hand-Object Interaction”,The 22nd Joint Workshop on Frontiers of Computer Vision (FCV2016): pp. 117-122,