Dynamic Pre-trainingを導入したDeep Neural Networkによる関節角時系列の予測

○杉浦孔明，是津耕司（情報通信研究機構）

1. はじめに

機械学習技術の進展および計算機が扱うデータ量の増大に伴い，大規模データの利活用が様々な分野で進められている．特に，動画像コンテンツは全世界のデータの半分を占めると言われ，コンテンツの利活用はロボティクスにおいても重要性が高い．また，姿勢推定の技術の進展とともに，動作理解のための情報が安価に入手できる環境が整備されつつある．[1] では，映画等の動画像を入力とした姿勢推定に対し Deep Learning を適用した精度向上が報告されている．

模倣学習分野においても動作理解は主要な研究課題であり，物体操作や全身動作の模倣，動作の言語化に関する研究が広く行われている [2–5]．動作の入力方法としてはマーカを使用して全身動作をキャプチャするものもあるが，Kinect 等の RGB-D カメラが広く用いられるようになってきた．これらの安価なデバイスで得られた動画像（およびウェブ上の動画像）においては，全ての骨格情報が観測可能であるとは限らない．すなわち，隠れた関節角を欠損値として扱うか，関節角の推定値を求める必要がある．

一般的な時系列予測問題を扱ったものは非常に多く存在する（例えば [6]）．予測問題における Deep Neural Network の構造を検討したものに [7] がある．[8] では，2 つの restricted Boltzmann machine からなる Deep Belief Network を用いた時系列予測手法が提案されている．一方，これまで種々の Deep Learning 手法が提案されているが，動作の予測に Deep Learning が適用された例はほとんどない．

Deep Learning において学習データの提示法を検討した研究としては，Curriculum Learning と呼ばれるアプローチがある [9]．Curriculum Learning では画像認識や言語モデルが議論の中心であるが，予測など他のタスクについても有効であることが示唆される．

このような背景から，本研究では Kinect 等の安価なデバイスで得られた関節角時系列の予測問題を扱う．提案手法では，時系列に特化した Pre-Training 手法を用い，動作予測に Deep Neural Network を適用する．実験に用いたデータセット（MSR Action3D Dataset [10]）の例を図 1 に示す．

本研究の独自性は以下である．

- 動作時系列の予測に対し，Dynamic Pre-Training(DPT) を導入した Deep Neural Network を適用した．

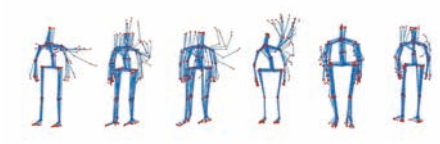

図 1　実験で用いたデータセットに含まれる動作の例．

2. Dynamic Pre-Training によるオートエンコーダの学習

本節では，[11] で提案した Dynamic Pre-Training について説明する．

Dynamic Pre-Training (DPT) は，Pre-Traning におけるオートエンコーダの学習を対象とする．いま，長さ D の時系列 $\mathbf{x} = \{x_1, \ldots, x_D\}$ が得られたとする．表記の都合上，特徴量は 1 次元であるものとする．ただし，実際には多次元の特徴量を扱う．

DPT では，入力時系列 $\mathbf{x}$ を順序を保ったまま η 個の部分時系列に分割する．分割された $j(= 1, ..., \eta)$ 番目の部分時系列 $\mathbf{z}_j$ は以下で与えられる．

$$\mathbf{z}_j = \{x_k | k = m(j-1) + i;\ i = 1, ..., m\} \qquad (1)$$

ここに，$j = 1, ..., \eta$ であり，$m(= D/\eta)$ は部分集合の要素数である．

各部分時系列は，反復回数 e に応じて変化する重要度 $w_j(e) \in [0, 1]$ が割り当てられる．重要度は $[0, 1]$ に含まれる実数であるものとする．$w_j(e)$ は以下のように更新される．

$$w_j(e) = \begin{cases} 1 & \text{if } j < c \\ e/\gamma - j + 1 & \text{if } j = c \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

ここに，$c = \text{ceiling}(e/\gamma)$，$\gamma = H/\eta$，$H$ は反復回数の最大値である．上記の $w_j(e)$ を用いて各部分時系列を重み付けして結合し，実際の学習に用いるサンプル $\tilde{\mathbf{x}}(e)$ を作成する．$\tilde{\mathbf{x}}(e)$ は以下で定義される．

$$\tilde{\mathbf{x}}(e) = \{w_1(e)\mathbf{z}_1, \ldots, w_\eta(e)\mathbf{z}_\eta\}. \qquad (3)$$

ここに，$e = 1, ..., H$ である．

3. 実験

3.1 実験 1: CATS ベンチマークによる検証

提案手法の有効性を検討するため，時系列予測の性能評価を行う．本研究では，時系列予測のベンチマークとして標準的に用いられている CATS [12] を用いる．CATS ベンチマークは，5000 フレームの人工データから 100 フレームの欠損値を予測するタスクである．図 2 に示すように欠損値は 20 フレーム連続しており，5 つの部分に分かれている．

ベースライン手法としては，Kuremoto et al. [8] らにより提案された手法を用いる．この手法は，2 種類の restricted Boltzmann マシンで構成され，ARIMA [13] などの予測モデルを上回る性能が示されている [8]．

表 1 に提案手法である DPT を導入した Deep Recurrent Neural Network による結果を示す．評価尺度として，CATS ベンチマークにおいて使用されている誤差の指標である E_1 [12] を用いた．表より，ベースライン

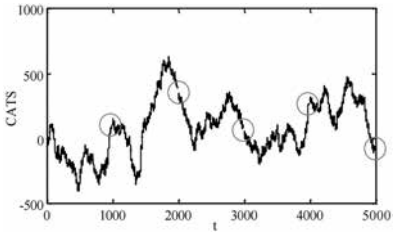

図 2 CATS ベンチマークにおける時系列．欠損値を丸で示す．

表 1　CATS ベンチマークによる性能評価

Method	Score (E_1)
DPT-DRNN (proposed)	1451
RBM (baseline)	1622

と比較し，本手法の誤差が小さいことがわかる．

3.2　実験 2: 動作の予測

次に動作予測に対する提案手法の評価を行う．評価において標準的なデータセットを用いることは重要であり，本研究では MSR Action3D Dataset [10] を用いる．本データセットは 10 人の被験者に 20 種類の動作を行なわせ，Kinect により収録したものである．各動作は平均 120 フレームほどであり，少なくとも 3 回動作が繰り返される．

本データセットにおいて，ほとんどの動作は左半身に関する動作であり，右半身の動きが予測に与える影響は少ないと考えられる．このことから，入力として左半身の関節のうち，肩，肘，手首に関する関節角を用いる．各関節 $\boldsymbol{j}_i$ の特徴量は，隣接する関節角を基準とした 3 次元相対位置であるものとし，時刻 t 時点で得られる入力として $\boldsymbol{j}(t-3), \boldsymbol{j}(t-2), ..., \boldsymbol{j}(t)$ を 36 次元の入力特徴量とする．出力としては，$\boldsymbol{j}(t+1), \boldsymbol{j}(t+2)$ を 6 次元の出力として予測するものとする．

評価において，学習セットとテストセットの分割は被験者を基準として行った．すなわち，学習セットとして被験者 4 名による動作を用い，テストセットとして学習セットに含まれない被験者 3 名が同じ動作を行ったものを用いた．つまり，モデルの予測にはその動作を行った被験者の情報は使われていない．

図 3 に定性的結果を示す．図は，データセット中の「動作 01（手を左右に振る動作）」に対して，提案手法を適用した結果である．図において，上図・中図・下図に手首の特徴量に対する x, y, z 軸の軌道を示す．図より，予測結果（Prediction）と真値（Observation）の間の誤差は少ないことがわかる．

次に定量的結果について述べる．評価の尺度として，以下で定義される二乗平均平方根 (RMSE) を用いる．

$$RMSE_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N} ||\boldsymbol{j}_i(t) - \hat{\boldsymbol{j}}_i(t)||^2} \qquad (4)$$

ここに，$\boldsymbol{j}_i$ および $\hat{\boldsymbol{j}}_i$ は，予測対象の真値および予測値を表し，N は総フレーム数である．予測の良好さについては種々の尺度があり得るが，本論文では RMSE が小さいことを予測精度が高いとみなす．実験の結果，$t+1$ の予測に対して $RMSE = 2.04$，$t+2$ に対して $RMSE = 2.29$ であった．

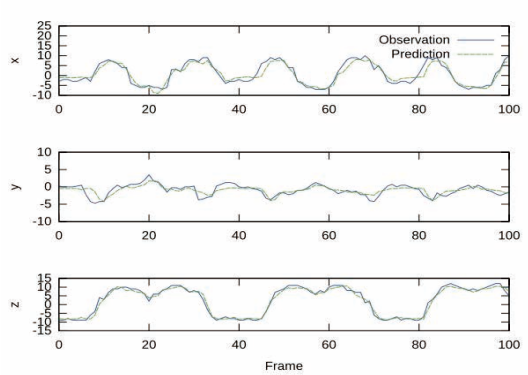

図 3　提案手法による手首位置の予測例．

4. おわりに

動作の予測は，ジェスチャインタフェースや，スポーツの動作解析，動作認識の性能向上など様々な応用が考えられる．本論文では，Dynamic Pre-Training を導入した Deep Neural Network による動作の予測手法について述べた．提案手法の評価のために，時系列および動作に関する標準データセットを用いた．実験の結果，ベースライン手法と比較して誤差を低減できることが示された．

謝辞

本研究の一部は，立石科学技術振興財団研究助成および JSPS 科研費 15K16074 の助成を受けて実施されたものである．

参　考　文　献

[1] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” Proc. CVPR, pp.1653–1660, 2014.
[2] K. Sugiura, N. Iwahashi, and H. Kashioka, “Motion Generation by Reference-Point-Dependent Trajectory HMMs,” Proc. IROS, pp.350–356, 2011.
[3] W. Takano and Y. Nakamura, “Statistically integrated semiotics that enables mutual inference between linguistic and behavioral symbols for humanoid robots,” Proceedings of the 2009 IEEE International Conference on Robotics and Automation, pp.2490–2496, 2009.
[4] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, “Embodied symbol emergence based on mimesis theory,” International Journal of Robotics Research, vol.23, no.4, pp.363–377, 2004.
[5] T. Ogata, M. Murase, J. Tani, K. Komatani, and H.G. Okuno, “Two-way translation of compound sentences and arm motions by recurrent neural networks,” Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System, pp.1858–1863, 2007.
[6] H. Cheng, P.-N. Tan, J. Gao, and J. Scripps, “Multistep-ahead time series prediction,” Advances in Knowledge Discovery and Data Mining, vol.3918, pp.765–774, 2006.
[7] S.F. Crone, M. Hibon, and K. Nikolopoulos, “Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction,” International Journal of Forecasting, vol.27, no.3, pp.635–660, 2011.
[8] T. Kuremoto, S. Kimura, K. Kobayashi, and M. Obayashi, “Time series forecasting using a deep belief network with restricted boltzmann machines,” Neurocomputing, vol.137, pp.47–56, 2014.
[9] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” Proc. ICML, pp.41–48, 2009.
[10] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” Proc. CVPR, pp.1290–1297, 2012.
[11] B.T. Ong, K. Sugiura, and K. Zettsu, “Dynamically Pre-trained Deep Recurrent Neural Networks using Environmental Monitoring Data for Predicting PM2.5,” Neural Computing and Applications, pp.–, 2015.
[12] A. Lendasse, E. Oja, O. Simula, and M. Verleysen, “Time series prediction competition: The CATS benchmark,” Neurocomputing, vol.70, no.13-15, pp.2325–2329, 2007.
[13] G.E.P. Box and G.M. Jenkins, Time series analysis: forecasting and control, Cambridge University Press, 1976.