J-020

スタイル変換による VAE を用いた感情表出動作生成の提案
An Emotional Motion Generation Method Using VAE with Style Transfer

白井 真歩[1)]　原田 誠一[1)]　佐久間 拓人[1)]　加藤 昇平[1)]
Maho Shirai　Seiichi Harata　Takuto Sakuma　Shohei Kato

1　はじめに

近年，ヒューマノイドロボットが人間に介護支援や案内などの多様なサービスを提供する機会が増加している．サービスを提供する上で，ロボットと人間の間にはコミュニケーションが発生する．ロボットと人間が円滑なコミュニケーションを取るためには，ロボットの感情表出が必要であると考えられる．ロボットの感情表出手段には，話し言葉による言語情報や，ジェスチャーや表情などによる非言語情報を用いる方法がある．

感情表出に用いる非言語情報として，表情が挙げられる。松井ら [1] は、ifbot [2] を用いて、基本的な表情から類似した新しい表情を自動生成する手法を提案した．ifbot は目の色やまぶたと顔の動きにより様々な感情表現が可能であるが，現在普及しているロボットのなかには顔表情が動かないロボットや動かすことのできる可動部位が少ないロボットが存在する．また顔の動きより動作の方が振幅が大きいため，より分かりやすく感情を表出することができると考える．そのため本研究では非言語情報として身体動作による感情表出に着目する．

画像生成分野では，ある画像の画風や質感を表すスタイル情報を抽出し，他の画像の原型を残したまま画風や質感を合成するスタイル変換について研究されている [3]．また，動作生成分野でも動作の内容（コンテンツ）を保持したまま動作の様子（スタイル）を付与するためにスタイル変換を取り入れる研究がされている [4]．

本研究では，コンテンツ動作の関節軌道を入力するコンテンツエンコーダとスタイル動作の関節軌道を入力するスタイルエンコーダを用い，それぞれの出力をデコーダに入力し訓練する．同時に，動作に付随する感情ラベルをエンコーダ，デコーダそれぞれに入力することによってコンテンツを保持したまま感情を表出した動作生成を目指す．

2　関連研究

2.1　感情を表出する動作生成

Marmpena ら [5] は Valiational Auto-Encoder (VAE) の潜在空間をサンプリングすることで新規の動作を生成する手法を提案した．実験ではまず，入力データを動作の関節角度と付随する Valence 値として VAE を学習した．次に，3 次元潜在空間上をホーントーラス型にサンプリングした．サンプリング半径の大きさが Arousal を表すと仮定して半径（低・中・高）を 3 種類用意し，入力時に Valence 値を 3 種類（低・中・高）加えることで，全 216 種の動作を生成した．感性評価実験では，Valence 値と Arousal 値を指定した動作を被験者に示し，動作がどの程度 Valence 値と Arousal 値を表現しているか評価した．結果，高 Valence 値と低 Arousal 値以外を表現する動作の表現力が高いことが示された．このことから，指定した感情を表現できていると考えられる．

しかし，入力動作に対して出力動作は乖離しており，入力動作の内容を保持していないことが問題として挙げられる．

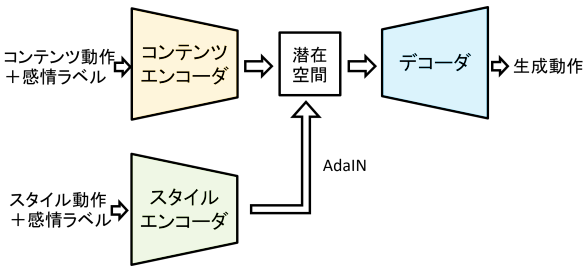

図 1: 提案モデル

2.2　スタイル変換による動作生成

深澤ら [6] は Content conditioned style encoder (COCO) [4] を導入し，パンチとキックなど類似度の低い動作間での多種類動作のスタイル変換を可能とした．入力動作のうち，コンテンツ動作はコンテンツエンコーダに入力し，コンテンツエンコーダを通してスタイルに依存したコンテンツ特徴量を抽出する．スタイル動作はコンテンツ動作とともに COCO に入力し，COCO を通してコンテンツ条件付きのスタイル特徴量を抽出する．抽出されたスタイル特徴量は全結合層を通して AdaIN [7] のパラメータへと変換され，コンテンツ特徴量とともにデコーダに入力される．デコーダでは AdaIN によってスタイル変換をして合成動作を得ることができる．実験ではコンテンツ動作として Sexy Punch のラベルが付随した動作，スタイル動作として Angry Kick のラベルが付随した動作を入力することで Punch の動作が保たれたまま Angry のスタイルが反映されることの検証を試みた．しかし，Punch 動作が保たれていることが確認されているが Angry というスタイル要素の反映は確認できていない．

3　提案手法

図 1 に提案手法の概観を示す．本研究の最終目的は動作内容を保持したまま感情を表出する動作を生成することである。そのためには，VAE によって動作の持つ特徴量（コンテンツ特徴量）を抽出し，抽出した潜在空間上に動作の持つ感情の特徴量（スタイル特徴量）を反映させることが必要である．コンテンツ特徴量を抽出するためのエンコーダをコンテンツエンコーダ，スタイル特徴量を抽出するためのエンコーダをスタイルエンコーダとしてエンコーダを 2 つに分けることで，それぞれの特徴量を抽出する．AdaIN を用いてコンテンツエンコーダから出力された潜在変数の平均・分散を，スタイルエンコーダから出力された潜在変数の平均・分散に線形変換する．これによりスタイルエンコーダで抽出した特徴量をコンテンツエンコーダで抽出した特徴量に反映でき

1）　名古屋工業大学 大学院工学研究科 工学専攻 情報工学系プログラム
Computer Science Program, Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology

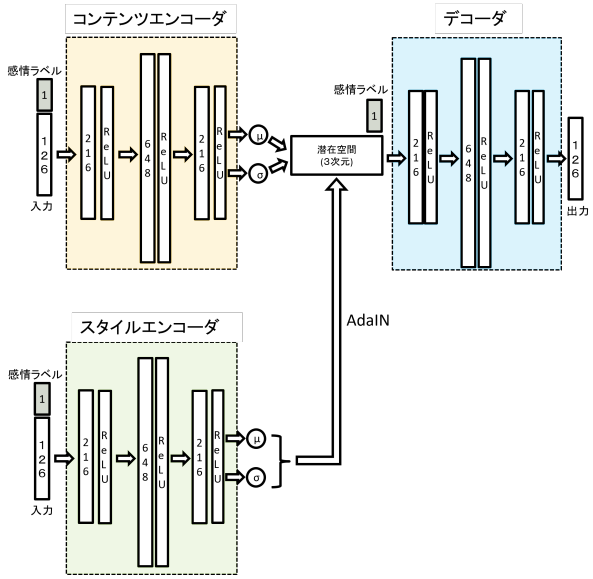

図 2: モデル詳細

表 1: 動作ラベル一覧

bow	walk	run	bye
guide	dash	byebye	walk-left
walk-back	walk-right	respond	slash
dance	punch	call	kick

る．また，今回使用する動作データは時系列データとして扱わず，ある瞬間の関節角度を入力とする．同時に感情ラベルを入力することで感情ごとの特徴量を抽出する．

動作生成時には，学習したモデルのコンテンツエンコーダに感情を持たない動作（コンテンツ動作）の関節角度と感情ラベルを入力し，スタイルエンコーダに感情を持つ動作（スタイル動作）の関節角度と感情ラベルを入力する．以上により，コンテンツを保持しながら，入力したスタイルを表出する動作を生成する．

3.1 モデル構成

図 2 にモデルの詳細を示す．コンテンツエンコーダとスタイルエンコーダは共に 216, 648, 216 ユニットの全結合層と，各全結合層の後に活性化関数である ReLU 層を持つ．デコーダは 216, 648, 216 ユニットの全結合層と，各全結合層の後に ReLU 層を持ち，出力層に tanh 関数を使用する．潜在空間は 3 次元とし，潜在変数はガウス分布に従うと仮定する．損失関数については，平均二乗誤差と KL ダイバージェンスの和を用いる．今回は損失関数の KL ダイバージェンスに係数 β を用いて重みづけをする β-VAE [8] を用いる．本実験では $\beta = 0.001$ とする．最適化手法には学習率を 0.0001 に設定した Adam optimizer [9] を使用する．AdaIN によりコンテンツエンコーダから出力された潜在変数の平均と分散をスタイルエンコーダから出力された潜在変数の平均と分散に置き換える．式（1）に AdaIN の式を示す．ただし x はコンテンツエンコーダから出力された潜在変数，$\mu_x^{content}$ はその平均，σ_x は x の分散を表し，$\mu_y^{content}$ はスタイルエンコーダから出力された潜在変数の平均，σ_y はその分散を表す．

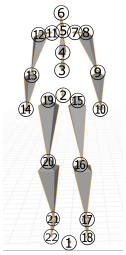

図 3: 関節名の対応図

表 2: 関節名一覧

1	joint_root	2	hip	3	spine
4	chest	5	neck	6	head
7	shoulder_L	8	upperArm_L	9	lowerArm_L
10	head_L	11	shoulder_R	12	upperArm_L
13	lowerArm_R	14	hand_R	15	upperLeg_L
16	lowerLeg_L	17	foot_L	18	toe_L
19	upperLeg_R	20	lowerLeg_R	21	foot_R
22	toe_R				

$$AdaIN(x, y) = \sigma_y\left(\frac{x - \mu_x^{content}}{\sigma_x}\right) + \mu_y^{content} \quad (1)$$

また，スタイル動作の影響を大きく受けコンテンツ動作が保持されないことを防ぐために，以下のように AdaIN に重み α を付ける（式（2））．本実験では $\alpha = 0.3$ とする．

$$x = x(1 - \alpha) + AdaIN(x, y)\alpha \quad (2)$$

4 実験

4.1 データセット

コンテンツ特徴量を抽出するため，感情ラベル付きモーションデータが必要である．本研究では，Bandai-Namco-Research-Motiondataset [10] を利用する．本データセットは 16 種のコンテンツ（日常動・格闘・ダンスなど）を感情を含む 15 種のスタイルで表現した総計 36,673 フレームを持つ”Bandai-Namco-Research-Motiondataset-1”と，10 種類のコンテンツを性別や年代などの 7 種のスタイルで表現した総計 384,931 フレームを持つ”Bandai-Namco-Research-Motiondataset-2”の 2 種類から構成されている．本研究では，感情のスタイルを含むため”Bandai-Namco-Research-Motiondataset-1”を利用する．表 1 に 16 種類の動作ラベルを示す．15 種のスタイルの中で，感情を表すスタイルは「angry」「happy」「normal」「sad」「tired」の 5 種類である．動作は 30fps のキーフレームで表され，各キーフレームは 22 関節のそれぞれ x,y,z 軸に対する位置座標と角度の値が計 132 次元で記述されている．図 3 に関節位置，表 2 に対応する関節名を示す．

4.2 データ前処理

まず，データの精度向上のため，基準の関節以外の関節の位置座標を除去した．本データセットの関節角度の記述は [−180°, 180°] で表されている．そのため，関節角度が 180° から −179° に推移する場合，動作は連続であるが関節角度は非連続値で構成されている．学習時に本データセットの角度表記のまま入力すると，ある関節角

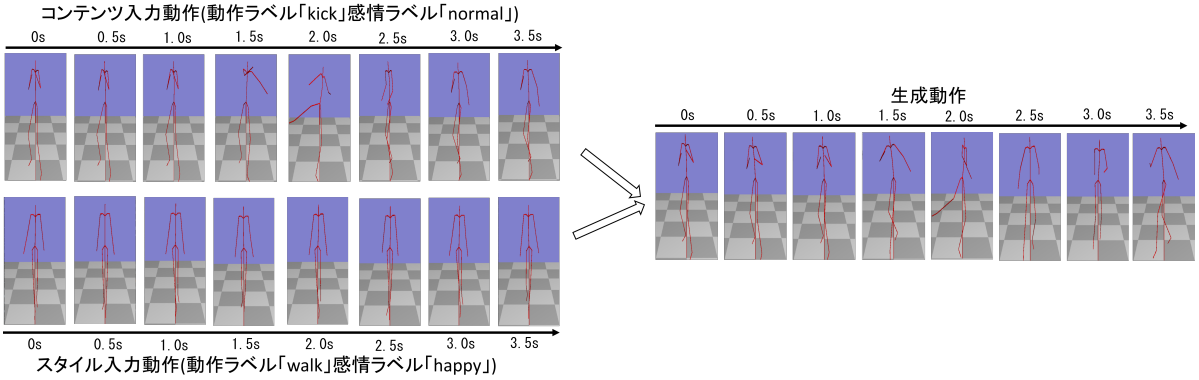

図 4: 感情ラベル「haapy」を付与した生成動作

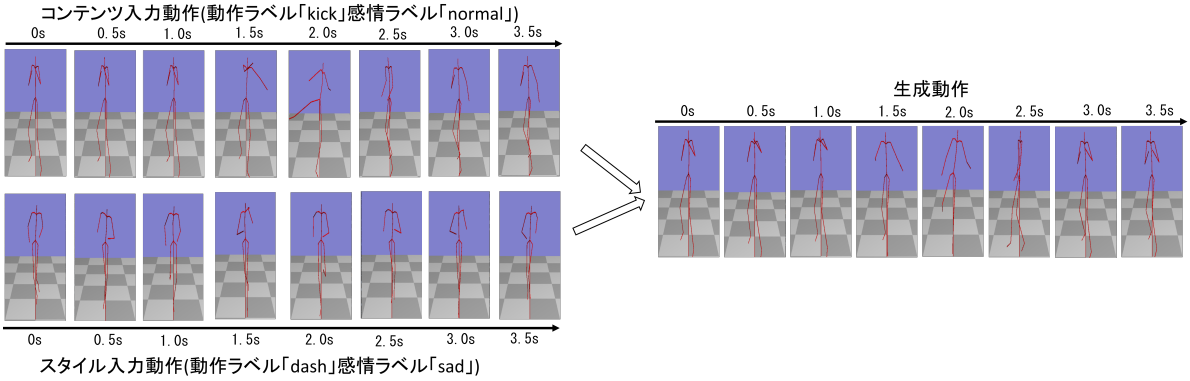

図 5: 感情ラベル「sad」を付与した生成動作

度が非連続値となるタイミングで，必要以上にその関節における損失が大きくなり効果的な学習を損ね，学習精度が低下する．関節角度の非連続値を回避するために，各関節角度を単位円上の極座標から直交座標に変換することで連続性を保持した．変換後，入力次元は 132 次元となった．

4.3 実験設定

コンテンツエンコーダとスタイルエンコーダの学習精度が偏ることを防ぐために各エンコーダに入力する動作データのフレーム数を均等にした．コンテンツエンコーダに感情ラベル「normal」47 動作の合計 8139 フレーム，スタイルエンコーダに感情ラベル「angry」「happy」「sad」「tired」の 4 種類各 11 動作の合計 8081 フレームを入力し，7000 エポック学習する．

4.4 FID 評価実験

Frechet Inception Distance (FID) [11] とは生成画像と現実の画像の類似度を評価する指標として用いられる．本実験では，生成動作の姿勢と入力動作の姿勢の類似度を評価する指標として用いる．FID 値が小さいほど生成動作が入力動作に類似した動作を生成していることを表す．提案モデルに使用した VAE とは異なる，エンコーダ・デコーダを各 1 つ持つ VAE にコンテンツ入力動作と生成動作の関節角度をそれぞれ入力したときの潜在空間上の分布を FID によって比較する．式（3）に FID の式を示す．ただし μ_x^{VAE} はコンテンツ入力動作の潜在空間上分布の平均，Σ_x はコンテンツ入力動作の潜在空間上分布の共分散を表し，μ_y^{VAE} は生成動作の潜在空間上分布の平均，Σ_y は生成動作の潜在空間上分布の共分散

表 3: 感情ラベルごとの FID(「normal」との比較)

	angry	happy	sad	tired
bow	0.0389	0.213	0.0249	0.0403
bye	0.0877	0.0983	0.00270	0.0520
byebye	0.396	0.631	0.0891	0.389
dash	0.100	0.199	0.0727	0.0825
guide	0.163	0.0324	0.119	0.154
kick	0.252	**0.117**	**0.906**	0.138
punch	0.0556	0.109	0.0428	0.115
respond	0.231	0.749	0.663	0.334
run	0.0503	0.0911	0.0419	0.173
slash	0.430	0.685	0.148	0.334
walk	0.0403	0.0455	0.0522	0.0425
walk-back	0.0349	0.0870	0.0157	0.0199
walk-left	0.264	0.71	0.206	0.296
walk-right	0.126	0.0215	0.0750	0.0196
average	0.162	**0.270**	0.172	0.156

を表す．

$$FID = \|\mu_x^{VAE} - \mu_y^{VAE}\|^2 + Tr(\Sigma_x + \Sigma_y - 2\sqrt{\Sigma_x \Sigma_y}) \quad (3)$$

同じ動作ラベルの中で，コンテンツ入力動作 (感情ラベル「normal」) と，様々なスタイル入力動作から生成された動作間の FID 値の平均を感情ラベルごとに求めた．

5 実験結果と考察

5.1 動作生成

図 4，図 5 に入力動作と生成動作の一例を示す．図 4 はコンテンツ動作として動作ラベル「kick」感情ラベ

ル「normal」を入力し，スタイル動作として動作ラベル「walk」感情ラベル「happy」を入力したときの生成動作である．図 5 はコンテンツ動作として動作ラベル「kick」感情ラベル「normal」を入力し，スタイル動作として動作ラベル「dash」感情ラベル「sad」を入力したときの生成動作である．各動作は再生時間 3.5 秒，0 秒から開始し 1 コマ 0.5 秒で合計 8 コマで切り抜いたものである．

5.2 感情ごとの FID 値比較

各コンテンツ入力動作と生成動作の感情ラベルごとに FID 値を比較する．表 3 に感情ごとに算出した FID 値を示す．表 3 から動作ラベル「kick」は感情ラベルごとの FID 値の差が最も大きいことが分かる．図 4 に「kick」の中で最も FID 値が小さい動作（「happy」を付与）を示し，図 5 に最も FID 値が大きい動作（「sad」を付与）を示す．図 4，5 の生成動作を比較すると，「sad」の関節角度は「happy」に比べ入力動作から変化し，例えば 2.0s で関節 20 の lowerLeg_R（図 3・表 2 参照）の関節角度差が大きくなっていることが確認できる．以上のように 2 つの動作の変化量の違いが読み取れ，FID 値の大小と一致している．しかし，動作ラベルの FID 値の平均を感情ラベルごとに比較すると「happy」の値が最も大きい．これは「happy」の付与は動作に与える影響が他の感情に比べて大きいことを示している．「kick」において，「happy」の FID 値が小さい原因として動作の性質の違いが考えられる．データセットにおいて「happy」の感情ラベルがついた動作は関節の振幅が大きく，「sad」の感情ラベルがついた動作は関節の振幅が小さくなる傾向があり，動作の大きさは感情と関係性があると考えられる．「kick」は元々の動作が大きく，「normal」の状態でも「happy」の性質が強く，感情を付与した際の変化が小さくなったと考えられる．動作の持つ性質によって感情ラベルを付与した際の変化に差があり，FID 値の大小による感情付与の成否を確認できないため，感性評価実験によって感情の表出を確認するべきである．

6 今後の展望

今後の展望として第一に，生成動作が目的を満たしているか評価実験を実施することが挙げられる．生成動作が入力した感情を表出しているか，入力した動作内容を保持しているかを確かめるために，実験協力者に実際に生成した動作を示し，評価を得るための感性評価実験を実施する．今回は骨格モデルで動作を表示しているが，他の 3D モデルを用いることで動作の視認性を高める必要がある．また，提案手法の有用性を示すために他の動作生成手法との比較実験を実施する．スタイル変換を用いた動作生成手法や他の生成モデルを用いた手法に今回使用したデータセットを入力して同様に感性評価実験を実施する．

第二に，学習時に動作の時系列を考慮することが挙げられる．今回は VAE を用いていることから動作データを時系列データとして扱えない．時系列データとして扱うことで動作の速度や前後のつながりを考慮することができ，学習時により詳細な感情の特徴量を抽出することができると考える．

7 まとめ

本研究の目的は動作内容を保持したまま感情表出動作を生成することである．そのために VAE の潜在空間上に動作の持つ動きの特徴量と感情の特徴量を抽出することが必要と考え，スタイル変換を用いることで感情を表出する動作を生成する方法を検討した．学習時にはコンテンツエンコーダに感情ラベル「normal」の動作の関節角度と感情ラベルを入力し，スタイルエンコーダに感情ラベル「angry」「happy」「sad」「tired」の動作の関節角度と感情ラベルを入力した．AdaIN を用いてスタイルエンコーダから抽出した特徴量をコンテンツエンコーダから抽出した特徴量に反映させた．その特徴量をデコーダに入力することで入力動作内容を保持したまま感情を付与することを試みた．実験では，生成動作と元動作間の特徴量距離を確認するために FID 値を算出した．感情ごとに FID 値を求めた結果，感情ラベルごとに全体の FID 値の平均は「happy」が最も大きいことが分かった．このことから「happy」の付与は動作に与える影響が他の感情に比べて高いことが示された．一方で一部の動作ラベルでは「happy」の FID 値が他の感情ラベルと比較し小さいことが確認された．

謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP24H00741），ならびに，国立研究開発法人情報通信研究機構委託研究の助成により行われた．

参考文献

[1] 松井裕紀, 加納政芳, 加藤昇平, 伊藤英則：Simple recurrent network を用いた感性ロボットのインタラクティブ表情表出, 日本ロボット学会誌, Vol. 28, No. 3, pp. 360–368 (2010).

[2] Kato, S., Ohshiro, S., Itoh, H. and Kimura, K.: Development of a communication robot Ifbot, in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, Vol. 1, pp. 697–702IEEE (2004).

[3] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium, *Advances in neural information processing systems*, Vol. 30, (2017).

[4] Saito, K., Saenko, K. and Liu, M.-Y.: Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder, in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 382–398Springer (2020).

[5] Marmpena, M.: Data-driven emotional body language generation for social robotics, *arXiv preprint arXiv:2205.00763*, Vol. 2205.00763, No. 3, pp. 1214–1223 (2022).

[6] 深澤康介, 森島繁生他：スタイル変換を用いた多様な動作合成研究, 第 84 回全国大会講演論文集, Vol. 2022, No. 1, pp. 251–252 (2022).

[7] Huang, X. and Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization, in *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510 (2017).

[8] Higgins, I.: beta-vae: Learning basic visual concepts with a constrained variational framework, *International conference on learning representations*, Vol. 57, No. 1, pp. 1–15 (2017).

[9] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv e-prints arXiv:1412.6980*, Vol. 57, No. 1, pp. 1–15 (December 2014).

[10] Makito, K. and al, et C.-C. L.: Motion Capture Dataset for Practical Use of AI-based Motion Editing and Stylization, *arXiv preprint arXiv:2306.08861* (2023).

[11] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium, *Advances in neural information processing systems*, Vol. 30, (2017).