2S-03

覚醒度と感情価に基づく音楽による画像スタイル変換

神庭 有花 † 田中 啓太郎 † 平田 明日香 † 森島 繁生 ††
† 早稲田大学 †† 早稲田大学理工学術院総合研究所

1. はじめに

本稿では，音楽から想起される感情に基づく画像のスタイル変換を扱う．近年の SNS の普及に伴い，個人で撮影した動画像の編集や背景音楽（BGM）の挿入は，広く一般的な作業となっている．動画像と BGM の印象には一貫性が求められる一方，視覚効果の追加と BGM の挿入は独立した工程である場合が多い．そのため，これらを効率よく同期し，かつ統一感のある印象に基づいて編集可能な手法が求められる．

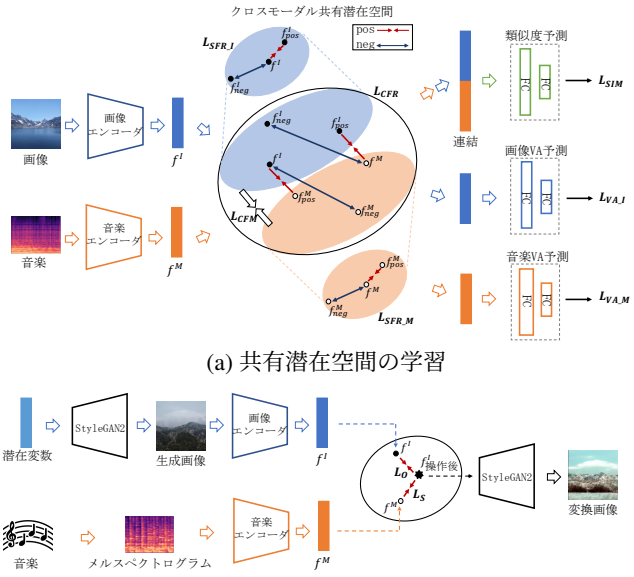
このような編集に向けた試みとして，音響信号を用いた画像のスタイル変換手法が提案され始めている．C. Lee ら [1] は，画像のスタイル変換で通常必要とされるスタイル画像を，入力音楽を可視化した画像で代替する手法を提案した．彼らは西洋音楽と西洋絵画のデータセットにおいて，両者の制作年代を共有ラベルとして用い，音楽情報を反映した画像の生成器の学習を行なっている．しかし，このアプローチでは，一つの共有ラベルで指定される音楽と画像の対象範囲が広すぎるため音楽と可視化画像の印象が乖離し，また時間変化する音楽の局所的なニュアンスを考慮できないという問題点がある．

ごく最近には S. H. Lee ら [2] によって，音楽の可視化を経由せず，音響信号をもとに直接画像を変換する手法が提案されている．これは，テキスト入力による画像編集が可能な StyleCLIP [3] の手法を，音響信号に拡張したものである．事前学習済みのテキストと画像の潜在空間において，テキストによる注釈のついた音響信号の埋め込みを学習し，テキスト，画像，音響信号の 3 種類のモダリティの関係性を獲得する．彼らの手法では，テキスト表現の容易な環境音の場合には音響信号の特徴を適切に画像に反映できる一方，テキスト表現の困難な楽曲の一部の場合にはその特徴が適切に反映されない．また，音楽の時間変化は依然考慮されていない．

本研究では，覚醒度と感情価に基づく潜在空間を介して，音楽音響信号をもとに画像を変換する手法を提案する．覚醒度と感情価は感情を連続的に表現するモデルの一つで [4] である．この値を用いた距離学習によって，音楽音響信号と画像から想起される感情を反映した共有潜在空間を獲得する．また獲得した共有潜在空間上で画像の潜在ベクトルを音楽音響信号の潜在ベクトルに近づけることにより，スタイル変換後の画像を獲得する．評価実験によって音楽音響信号による画像のスタイル変換における，本手法の有用性を確認する．

2. 提案手法

提案手法は，図 1(a) に示す画像と音響特徴量の共有潜在空間の学習段階と，図 1(b) に示す音楽による画像のスタイル変換段階で構成される．共有潜在空間の学習段階は，覚醒度と感情価に基づき画像と音楽をマッチングする手法 [5] に倣う．具体的には，覚醒度と感情価に基づく共有潜在空間での距離学習と，画像および音楽の覚醒度と感情価の類似度予測器の学習，各値の予測器の学習を同時に行う．本研究では，事前に覚醒度と感情価の注釈がつけられた音楽のデータセット DEAM [6] と画像のデータセット NAPS [7] 及び EMOTIC [8] を用い，覚醒度と感情価の値は $[0, 1]$ に正規化したものを用いる．



(a) 共有潜在空間の学習

(b) 音楽による画像のスタイル変換

図 1: 提案手法のモデル

学習段階ではまず，画像と音響特徴量を同時にモデルに入力する．ResNet50 に基づくエンコーダに画像を入力して 512 次元の特徴量 f^I を，ResNet18 に基づくエンコーダに入力音響信号のメルスペクトログラムを入力して 512 次元の特徴量 f^M を，それぞれ得る．同時に入力する画像と音楽の組み合わせは IMEMNet データセット [5] に依拠する．次に，画像と音響特徴量のクロスモーダル距離学習，画像単独の距離学習，音響特徴量単独の距離学習を同時に行う．クロスモーダル距離学習に用いる損失関数は，画像と音響特徴量の覚醒度と感情価の類似度を潜在空間内の距離に反映する L_{CFR} と，画像の潜在空間全体と音響特徴量の潜在空間全体を近づける L_{CFM} から成る．L_{CFR} と L_{CFM} は次式で定義される．

$$L_{CFR} = \sum_{i=1}^{N} \left\{ \log \frac{D(f^{I_i}, f^{M_i})}{D(f^{I_i}, f^{M_j})} - \log \frac{S(I_i, M_i)}{S(I_i, M_j)} \right\}^2 + \sum_{i=1}^{N} \left\{ \log \frac{D(f^{M_i}, f^{I_i})}{D(f^{M_i}, f^{I_j})} - \log \frac{S(M_i, I_i)}{S(M_i, I_j)} \right\}^2 \quad (1)$$

$$L_{CFM} = \sum_{i=1}^{N} \max(0, ||f^{I_i} - f^{M_i}||_2 - \alpha) \quad (2)$$

なお，$D(\cdot)$ は 2 乗ユークリッド距離を，$S(\cdot)$ は $1 \leq i \leq n$

Music-Driven Image Style Transfer Based on Valence and Arousal: *Yuka Kaniwa†, Keitaro Tanaka†, Asuka Hirata†, and Shigeo Morishima†† (†Waseda University, ††Waseda Research Institute for Science and Engineering)*

と $1 \leq j \leq m$ に対して次式で定義される類似度を表す．

$$S(I_i, M_j) = \exp(-\frac{||y^{I_i} - y^{M_j}||_2}{\sigma_n^m}) \quad (3)$$

また，α は画像と音響特徴量の潜在空間上で許容する距離を表す．y^{I_i},y^{M_j} はそれぞれ画像 I_i，音楽 M_j の覚醒度と感情価の値であり，σ_n^m は IMEMNet データセット [5] の訓練データの，すべての組み合わせに対するユークリッド距離の平均である．画像と音響特徴量の個別の潜在空間における距離学習では，次式で定義される特徴量比損失 L_{SFR_I}，L_{SFR_M} を用いる．

$$L_{SFR\_I} = \sum_{i=1}^{n} \left\{ \log \frac{D(f^{I_i}, f^{I_j})}{D(f^{I_i}, f^{I_k})} - \log \frac{D(y^{I_i}, y^{I_j})}{D(y^{I_i}, y^{I_k})} \right\}^2 \quad (4)$$

$$L_{SFR\_M} = \sum_{i=1}^{m} \left\{ \log \frac{D(f^{M_i}, f^{M_j})}{D(f^{M_i}, f^{M_k})} - \log \frac{D(y^{M_i}, y^{M_j})}{D(y^{M_i}, y^{M_k})} \right\}^2 \quad (5)$$

画像と音楽の覚醒度と感情価の類似度予測器，および各値の予測器の学習では，各予測器の学習における損失関数 L_{SIM}，L_{VA_I}，L_{VA_M} として平均二乗誤差を用いる．高精度かつ円滑な学習を可能にするため，学習では上述の損失関数 L_{CFR}，L_{CFM}，L_{SFR_I}，L_{SFR_M}，L_{SIM}，L_{VA_I}，L_{VA_M} の総和 $\mathcal{L}_{total}$ を最適化する．

画像のスタイル変換段階は，S. H. Lee ら [2] に倣う．彼らが画像と音響特徴量の共有潜在空間をテキストを介して獲得しているのに対し，提案手法では覚醒度と感情価を用いて上述のように獲得した共有潜在空間を用いる．まず，共有潜在空間の学習部分で得た画像と音響特徴量のエンコーダに，潜在コード w_o から StyleGAN2 [9] を用いて生成した画像と，入力音楽のメルスペクトログラムをそれぞれ入力し，両者の潜在表現から潜在コード w_m を得る．w_m をさらに StyleGAN2 の入力とすることで，音響特徴量によって変換された画像を得る．次に，類似損失 L_S を次式で定義する．

$$L_S = 1 - d_{cos}(G(w_m), f^M) \quad (6)$$

ただし $d_{cos}(G(w_m), f^M)$ は，w_m から生成された画像の潜在ベクトル $G(w_m)$ と入力音響特徴量の潜在ベクトル f^M のコサイン類似度を表す．また，入力画像の固有性を保持するための損失 L_O を次式で定義する．

$$L_O = ||w_o - w_m||_2 \quad (7)$$

最後に，L_S と L_O から成る変換損失 $\mathcal{L}_{transfer} = L_S + \lambda * L_O$ を最小化する w_m を求めることで，音楽音響信号の感情の反映と元の入力画像の固有性の保持を両立した変換画像を得る．なお，λ はどの程度元の画像の固有性を保つかを制御するハイパーパラメータである．

3. 評価実験

3.1 実験条件

本稿では，制作年代を用いて音楽の可視化を経由する手法 [1]，StyleCLIP の潜在空間を音響信号に拡張した手法 [2]，提案手法の三つの手法を，音楽のもたらす印象の違いをいかに反映できているかという観点で定性的に比較する．同一の入力画像に対し，入力する音楽と画像の変換手法を変更した．入力音楽は IMEMNet のテストデータから，感情価が大きい楽曲 2 曲，感情価が小さい楽曲 2 曲の計 4 曲を使用した．

表 1: 音楽音響信号による操作結果の比較

アーティスト, (アルバム,) 楽曲	(覚醒度，感情価)	入力画像	手法別の出力画像		
			音楽可視化経由	StyleCLIP 拡張	提案手法
Goto80 and the Uwe Schenk Band, Volksing Ferretismico	(0.828, 0.807)				
Calypso, netBloc Vol. 02: DRM killed the music-product machine, Copacabana Palace Hotel	(0.392, 0.617)				
Creepoid, Old Tree	(0.746, 0.283)				
Angels in America, (E.M.S.,) II	(0.252, 0.227)				

3.2 実験結果

表 1 に，(覚醒度，感情価) の値をもつ音楽を入力した結果を示す．上 2 列は感情価の高いポジティブな音楽を，下 2 列は感情価の低いネガティブな音楽を，それぞれ入力した結果である．出力結果の色合いに着目すると，4 曲の異なる入力に対して最も印象の異なる出力を得られたのは提案手法，次いで音楽可視化経由の手法である．音楽可視化経由の手法は 2 列目から 4 列目の，StyleCLIP 拡張の手法は 1 列目と 3 列目の，出力結果が類似している．それに対して提案手法では，上 2 列と下 2 列では色味が，またそれぞれでは色の濃さが異なっており，手法の有効性がうかがえる．入力音楽の覚醒度と感情価に照らすと，覚醒度が色の濃淡に影響し，感情価が出力画像の色味に影響していることがわかる．ただし，覚醒度や感情価が近い場合は，出力結果にあまり相違がなかった．

4. おわりに

本稿では，覚醒度と感情価に基づく潜在空間を介して，音楽音響信号をもとに画像のスタイル変換をする手法を提案した．提案手法を含む 3 手法の出力を比較した結果，提案手法は音楽のもたらす印象の違いの大枠を反映して画像を変換できることが確認された．今後は，覚醒度と感情価が近い音楽の印象を画像のより微細な差異として反映させる．また，入力音楽の時間変化を取り入れた，より精密な画像操作が可能な枠組みへの拡張を目指す．

謝辞 本研究は，JSPS 科研費 (19H04137, 21H05054, 22J22424) の補助を受けています．

参考文献

[1] C. Lee *et al.*: “Crossing You in Style: Cross-modal Style Transfer from Music to Visual Arts,” *ACM MM*, 3219–3227, 2020.

[2] S.H. Lee *et al.*: “Sound-Guided Semantic Image Manipulation,” *CVPR*, 3377–3386, 2022.

[3] O. Patashnik *et al.*: “StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery,” *ICCV*, 2085–2094, 2021.

[4] H. Schlosberg *et al.*: “Three dimensions of emotion,” *Psychological Review*, 81–88, 1954.

[5] S. Zhao *et al.*: “Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space,” *ACM MM*, 2945–2954, 2020.

[6] A.Alajanki *et al.*: “Benchmarking music emotion recognition systems,” *PLOS ONE*, 835–838, 2016.

[7] A.Marchewka *et al.*: “The Nencki Affective Picture System (NAPS): introduction to a novel, standardized, wide-range, high-quality, realistic picture database,” *Behavior Research Methods*, 596–610, 2014.

[8] R.Kosti *et al.*: “Context based emotion recognition using emotic dataset,” *PAMI*, 2755–2766, 2019.

[9] T.Karras *et al.*: “Analyzing and improving the image quality of stylegan,” *CVPR*, 8110–8119, 2020.