2U-08

音響特徴量を考慮したミュージックビデオの色調編集手法

井上　和樹† 中塚　貴之† 柿塚　亮† 高森　啓史† 宮川　翔貴† 森島　繁生‡
早稲田大学† 早稲田大学理工学術院理工学総合研究所/JST ACCEL‡

1 はじめに

楽曲と動画が調和したミュージックビデオ(MV)を制作するためには，楽曲の雰囲気を考慮して動画を編集する必要がある．特に，動画の色調はユーザの印象に大きな影響を与えるため重要度が高い．しかし，動画の色調編集には高い動画編集技術と多大な労力を必要とする．例えば暗い雰囲気の楽曲を用いる場合，全体の色のバランスを保つために像の明度・彩度を下げる編集だけでなく，トーンカーブの調整や特定領域のマスク処理といった専門的かつ老両区を要する編集が求められる．そこで，我々はユーザの効率的な映像編集を支援するために「楽曲」と「動画」を入力とし，楽曲の雰囲気を考慮して動画の色調を自動編集する手法を提案する．

複数入力を用いて画像の色調を自動で編集する研究として Ali[1]らの研究がある．Ali らは画像及び ”anger”，”disgust”，"fear”，“joy”，”sadness”，”surprise”，”neutral”の 7 つの独立成分からなる感情ヒストグラムを入力として，ヒストグラムを再現するように入力画像の色調を自動で編集することを可能にした．しかし，本研究では楽曲の雰囲気に合わせて動画の色調を編集することを目的としており，楽曲を形容詞等に言語化することは非常に難しいため，Ali らの研究を我々の手法に適用することは難しい．

そこで我々は，入力楽曲から抽出された音響特徴量に基づいて動画の色調を編集する手法を提案する．概要を図 1 に示す．本手法では既存の MV をデータベースとして，MV における色調と楽曲の関係を深層学習によって計算する．ユーザは学習済み深層学習モデルに MV として使用したい動画と楽曲を入力することで，楽曲の雰囲気に合わせて色調編集が施された MV を得る．

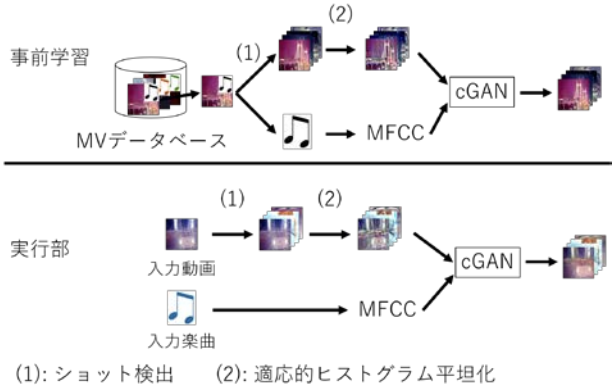

図 1　提案手法の概要

2 提案手法

2.1 MV データベース

MV のデータベースとして YouTube[2]に公開されている MV200 本を使用した．各 MV の動画に対して RGB ヒストグラム，ブロックマッチングを用いてショット検出を行い，各ショットの先頭フレームを抽出する．次に，各ショットにおける MFCC の平均値を計算し，先頭フレームと対応づける．

2.2 色調編集前の疑似画像生成

提案手法では既存の MV をデータベースとして MFCC と色調編集を学習するため，色調編集前後の動画が必要である．しかし，一般に公開されている MV はすでに色調編集が施されており，色調編集が施される前の動画を入手することは困難である．そこで，我々は既存の MV の RGB チャンネルそれぞれに適応的ヒストグラム平坦化を施した動画を，疑似的に色調編集前の動画として扱う．適応的ヒストグラム平坦化による変換写像は一意的に決定されるが，その逆写像は一意的に決定されないため，適応的ヒストグラム平坦化を施した動画を入力とすることで様々な種類の色調編集が施された動画の出力が可能であると考えられる．以下では，各ショットの先頭フレームを画像 A，画像 A に適応的ヒストグラム平坦化を施した画像を画像 B と呼ぶことにする．

2.3 色調編集と楽曲の関係の学習

特徴量ベクトルが付与された画像を他の画像

“Color tone editing method of music video considering acoustic features”
†Kazuki INOUE, Takayuki NAKATSUKA, Ryo KAKITSUKA, Hirofumi TAKAMORI, Shoki MIYAGAWA, Waseda University
‡Shigeo MORISHIMA, Waseda Research Institute for Science and Engineering / JST CREST

へと変換する手法の一つに条件付き敵対的生成ネットワーク(cGAN)がある．敵対的生成ネットワーク(GAN)は，データベースの画像を

図 2　提案ネットワーク

参照に画像を生成する Generator と，画像がデータベースに属する画像か Generator によって生成された画像かを判別する Discriminator の 2 つのネットワークからなる．cGAN とは GAN をベースとして Generator の入力に特徴量ベクトルを別に加えるネットワークなどを指す．提案手法のネットワークは Isola ら[3]が提案した cGAN をもとに構築した．Isola らは，白黒画像からカラー画像，エッジ画像からカラー画像など，様々な種類の変換が施された画像を入力として画像生成を行った．我々は Isola らの手法をもとに特徴量ベクトルとして MFCC を使用し，画像 B から画像 A を復元するようにネットワークを学習した．図 2 に提案ネットワークを示す．提案ネットワークでは，画像 A と画像 B を入力として Discriminator を学習する部分と，画像 B と MFCC を入力として Generator によって画像を生成し，この生成された画像と画像 A を入力として Discriminator を学習する 2 つの部分からなる．MFCC を加えることで楽曲の雰囲気を考慮して画像 B から画像 A への色調編集を学習させる．

2.4　MV の自動色調編集

実行時に上記で学習したネットワークを用いることで MV の自動色調編集を行う．使用したい動画と楽曲を用意し，事前学習時と同様にショット検出と各ショットにおける MFCC を計算する．次に，入力動画に対して適応的ヒストグラム平坦化を適用した動画(動画 B と呼ぶ)を用意する．動画 B の各フレームに対して同一ショットで計算された MFCC を付与し，上記で学習したネットワークに入力することで，動画の色調が編集されたフレームを得る．

3　結果と考察

図 3-1 に，図 3-2 に出力された動画を示す．

図 3-1　MFCC の有無による出力動画の比較(左：元動画，中央：MFCC 有，右：MFCC 無)

図 3-2　付加する楽曲による比較(左：元動画，元動画の MFCC を使用，death meal の MFC を使用)

テスト用 MV として音楽情報検索サイト last.fm[4]で音楽ジャンルが'pop'の MV を使用した．図 3-1 より，MFCC を使用せずに画像 A と画像 B を学習させた方が画像 B から画像 A への再現できていることが分かる．図 3-2 では，使用する MFCC としてテスト用 MV の元楽曲の MFCC を使用した場合と，last.fm で音楽ジャンルが'death metal'の楽曲を使用した場合の出力動画を比較している．図 3-2 より，付加する MFCC の音楽ジャンルによる再現画像の変化はほとんどないことが分かる．以上より，色調編集と対応させる音響特徴量として MFCC は適切ではなく，MFCC 以外の音響特徴量を使用する必要があることが分かった．

4　まとめと今後の課題

本研究では，既存の MV をデータベースとして cGAN を学習することで，任意の動画と楽曲の組み合わせを入力としたときの楽曲の雰囲気に合った動画の自動色調編集手法を提案した．今後の課題として，MFCC 以外の音響特徴量を使用した結果との比較を行いたい．また，生成動画と入力楽曲の雰囲気の一致度を評価したい．

謝辞

本研究は JST ACCEL(JPMJAC1602)の支援を受けた．

参考文献

[1] Ali et al. Emotional Filters: Automatic Image transformation for Inducing Affect, Proc. BMVC British Machine Vision Conference (2017).
[2] https://www.youtube.com/?hl=ja&gl=JP
[3] Isola et al. Image-to-Image Translation with Conditional Adversarial Network, Proc. CVPR The IEEE Conference on Computer Vision and Pattern Recognition (2017).
[4] https://www.last.fm/