2Q-01

動画特徴量からの印象推定に基づく動画 BGM の自動生成

清水 柚里奈† 菅野 沙也† 伊藤 貴之† 嵯峨山 茂樹†† 高塚 正浩 †††

† お茶の水女子大学大学院 理学専攻 情報科学コース 〒112-0012 東京都文京区大塚 2-1-1
†† 明治大学 総合数理学部 先端メディアサイエンス学科 〒164-8525 東京都中野区中野 4-21-1
††† シドニー大学 工学部 情報科学学科 J12 School of IT Building, University of Sydney 2006 NSW Australia

1. 概要

近年，デジタルカメラやスマートフォンの普及により，写真や動画を撮影する機会が増え，またその撮影したものを SNS サイトに投稿することで，多くの人々と共有して楽しむようになった．その際に，ただ撮影したものを投稿するのではなく，撮影映像に BGM を付与するなどの動画編集も行うようになってきた．しかし動画編集では一般的に，動画に合った音楽を自分で探したり，動画の長さに合うように音楽を調整したり，といった手間とスキルが必要となる．

我々は，動画の印象に合った楽曲を自動付与することを目標として，動画特徴量からの印象推定結果に基づいた楽曲生成手法を提案している[1]．本報告ではその改良についていくつかの技術要素を述べる．具体的には動画特徴量の抽出，印象評価のためのユーザインタフェースの改良，印象値の回帰手法の再考，メロディの音色選択，の各技術要素について述べる．

2. 関連研究

ビデオに BGM を付与させる研究として，映像の動きと同期する部分を楽曲から抽出し動画へ付与する研究[2]や音楽分析アルゴリズムに基づいてホームビデオの音楽ビデオを自動で生成するシステム[3]などが挙げられる．しかし，[2]では映像の動きだけを考慮して楽曲生成がされており，映像の内容や雰囲気に対しての考慮はされていない．また[3]では動画の内容を予めユーザが指定した上で楽曲生成がなされており，動画解析処理は自動化されていない．

3. 提案手法

本手法は大きく分けて 4 つの処理段階で構成される．具体的には，

(1)動画特徴量：色分布・動き分布の特徴量抽出
(2)音楽特徴量：メロディ・リズムの特徴量抽出
(3)学習：動画，メロディ・リズムの印象の関係性算出
(4)楽曲生成：ユーザの印象に合った楽曲生成

の 4 段階である．詳細について以下に論述する．

3.1 動画特徴量

現時点での我々の実装では，色分布，動き分布の 2 種類の低水準特徴量と印象との関係を学習している．

色分布，動き分布の特徴量抽出に関しては，文献[1]の 3.1.1 項，3.1.2 項にて記載されている．以前の実装では，色分布や動き分布を一定時間ごとに抽出していた．この抽出を改良するために現在の実装では，動画変換ライブラリ FFmpeg (http://www.ffmpeg.org/) を用いて入力動画像を「I フレーム」と呼ばれるキーフレームごとに分割し，分割された各動画に対して画素値およびオプティカルフローを求めることにした．色分布および動き分布に関する特徴量抽出のためのそれ以降の処理に関しては，文献[1]の 3.1.1 項，3.1.2 項にて記載されている通りである．映像中の重要なキーフレームを単位として特徴量を抽出することで，より動画の特徴を捉えることができる．

3.2 音楽特徴量

現時点での我々の実装では，メロディとリズムを別々の素材として用意し，これらに対して文献[1]の 3.2.1 項，3.2.2 項に記載された音楽特徴量を算出している．

3.3 学習

続いて本手法では，動画特徴量とそれに対する各ユーザの印象の関係，またリズム・メロディの音楽特徴量とそれに対する各ユーザの印象の関係を学習する．

3.3.1 ユーザ印象評価

まず予め用意したサンプル動画，サンプルリズム・メロディを評価する際に使用する感性語対を決定する．現在採用している感性語対は文献[1]の 3.3.1 項に記載されている．そして各ユーザにサンプル動画を閲覧してもらい，またサンプルメロディ・サンプルリズムを聴取してもらい，記載されている感性語への適応度を 6 段階評価で回答してもらう．しかし，ユーザに回答してもらった際に評価システムに関して以下のコメントを頂いた．

- 評価基準がないと判断が難しい
- 直前・直後に鑑賞したサンプルと比較してしまう

このような問題点を解決するために現在の実装では，評価対象となる動画，楽曲を順に並べてもらうことで相対評価を可能とする評価システム（図 1 参照）を開発した．

図 1：ユーザ評価システム

このシステムの使い方は以下の通りである．ユーザに鑑賞してもらったサンプル動画・メロディ・リズムを，文献[1]の 3.3.1 項に記載されている感性語への適応度順に，スライダーを用いて並べる．そして，そのスライダーの位置から各動画・メロディ・リズムの印象値を算出する．このようにして，各ユーザの印象値を収集する．

3.3.2 動画・楽曲特徴量からの印象学習

3.1 項，3.2 項で示した動画・楽曲特徴量から印象値を推定する．文献[1]の 3.3.3 項，3.3.4 項では，動画・音楽特徴量から印象値の推定に重回帰分析を採用していた．

Automatic music generation for movies based on impression estimation from image features
†Yurina Shimizu, Saya Kanno and Takayuki Itoh, Ochanomizu University
††Shigeki Sagayama, Meiji University
††† Masahiro Takatsuka, The University of Sydney

しかし，重回帰分析では教師信号が少ないと誤差が大きくなる，また非線形の特性を有する応答値を適切に推定できないという制約がある．本手法ではあらかじめユーザにサンプル動画・リズム・メロディを評価してもらい，それをもとに算出した印象値を学習データとして扱うため，ユーザの負担の観点から多くの学習データを集めることは期待できない．また文献[1]の 3.3.2 項で既に，色分布の特徴量と感性語の間で線形の相関関係が見られず，重回帰分析を適用できなかったことを指摘している．これらの問題を解決するために現在の実装では，教師なしのニューラルネットワークアルゴリズムである SOM (Self Organizing Map) を適用している．

3.3.3 SOM の精度評価実験

SOM を利用するにあたって，これまで使用していた重回帰分析と SOM を比較して精度検証を行った．

[実験]文献[1]の 3.3.1 項で収集した，あるユーザ 1 名のメロディ 15 曲分の学習データを用いる．その中からランダムに選んだ 10 曲分の学習データを重回帰分析，SOM に適用し，残りの 5 曲分のデータに対して印象値計算を行う．そして，重回帰分析，SOM によって計算された印象値と実際の回答値の乖離度を以下の式を用いて求める．

$$乖離度 = \frac{\sum_{i=1}^{5} \left| 実際の回答値 X_i - SOM / 重回帰分析による計算値 Y_i \right|}{5}$$

この **[実験]**を 6 回繰り返し，実際の回答値と重回帰分析，SOM で計算した値との乖離度の平均を計算した．結果の例を表 1 に示す．

表 1：実際の回答値との乖離度平均

	明るい-暗い	さわやか-激しい	落ち着いた-元気な
SOM	0.381754758	0.399625325	0.493688417
重回帰	7.637503328	6.393207412	16.53777944

この実験から，重回帰分析より SOM の方が，少ないデータ数であるにもかかわらず印象推定結果の誤差が小さいことがわかった．そこで今後は，ユーザ評価結果の与えられていない動画，楽曲に対して，色分布，動き分布，メロディ，リズムの印象値を SOM により推定する．

3.4 楽曲生成

3.4.1 メロディ・リズムの合成

次に楽曲の素材となるメロディとリズムを選出し，合成する．3.3.2 項で算出した動画の印象値と，メロディ・リズムの印象値を比較して，ユークリッド空間上で最も距離の近いメロディ・リズムを動画の印象に沿った楽曲の素材とする．そしてこの選出したメロディとリズムを組み合わせて楽曲を生成する．続いて生成した楽曲にコード進行を加える．さらに，動画の再生時間に合うように小節数やテンポを設定する．

3.4.2 音色の付与

現在の実装では，生成した楽曲のメロディの音色を自動選択する機能も設けている．音色を選択するにあたって，24 個の各楽器の音色と 7 色からなる色彩の相関関係を数値として表している文献[5]を参考にした．これをもとに，楽曲を付与したい動画の色分布と近い色分布をした楽器の音色を cos 類似度推定法により求め，メロディの音色として付与する．

以上によって生成された楽曲と動画を合成することで，動画に BGM を付与する．

4. 実行結果と考察

本手法で使用するメロディには自動作曲システム Orpheus[6]を利用して作成した 30 パターンを用意し，リズムには文献[1]で使用した 21 パターンを用意した．このうちメロディ 15 種類，リズム 10 種類を学習用のサンプルメロディ・サンプルリズムとした．また動画は 1 分以内の 11 種類の動画をサンプルビデオとして用意した．

本実験ではユーザ A とユーザ B の各々に対してユーザ印象評価を依頼し，この結果をもとにしていくつかの異なるジャンルの動画に対して楽曲生成を行った．以下の 2 種類の動画に対して楽曲を付与した結果を表 2 に示す．

動画 1：人がいない夕暮れの海辺の様子
動画 2：犬が草むらを元気に走っている様子

表 2：動画 1,2 の楽曲生成を行った結果

	ユーザ A	ユーザ B	音色
動画 1	Melody15.mid Rhythm18.mid	Melody5.mid Rhythm16.mid	ハーモニカ
動画 2	Melody17.mid Rhythm6.mid	melody14.mid rhythm21.mid	トランペット

ユーザ A とユーザ B では異なる楽曲素材が選ばれており，学習段階の影響によりユーザの印象の違いを考慮した楽曲が生成されていることがわかる．しかし動画 2 の切ない雰囲気の動画に対し，暗くどんよりとした楽曲が生成されてしまった．このことから，例えば，学習段階における改善や，動画および楽曲の特徴量の見直しなどが必要である．

5. まとめと今後の課題

著者らは動画から一定時間ごとに抽出した動きや色の動画特徴量から動画の印象を推定し，その結果に基づいて楽曲を生成する手法を提案している．本報告ではその改良についていくつかの点を述べた．

今後の課題として，学習段階における実験計画法を用いたサンプル動画・サンプル楽曲の選定や動画の内容把握や物体認識など，高レベルな動画特徴量の追加が挙げられる．また現段階では単純な音形で付与しているコードの弾き方を，リズムや曲調に合わせて変えることも検討する．

参考文献

[1] 清水柚里奈, 菅野沙也, 伊藤貴之, 嵯峨山茂樹, “動画解析・印象推定による動画 BGM の自動生成”, 第 7 回データ工 学と情報マネジメントに関するフォーラム (DEIM), F2-3, 2015.

[2] 小野佑大, 甲藤二郎, ”音楽のムード分類結果を利用したホームビデオへの BGM 付与支援システム”, 情報処理学会音楽情報処理研究会, Vol. 2011-MUS-89, 2011.

[3] Jun-Ichi Nakamura, Tetsuya Kaku, ”Automatic Background Music Generation based on Actor’s Mood and Motion”, The Journal of Visualization and Computer Animation, Vol. 5, No. 4, pp. 247-264, 1994.

[4] 高塚正浩, Ying Xin WU, ”球面 SOM のデータ構造と量子化誤差の考察およびインタラクティビティの向上”, 日本知能情報ファジィ学会誌, Vol. 19, No. 6, pp. 611-617, 2007.

[5] 赤井良行, 李昇姫, ”音色からイメージされる色彩の寒暖と音色構造の関係”, 日本感性工学会論文誌, Vol. 13, No. 1, pp. 221-228, 2014.

[6] 東京大学 大学院情報理工学系研究科 システム情報学専攻, 自動作曲システム Orpheus, http://www.orpheus-music.org/v3/