

動画特徴量からの印象推定に基づく動画 BGM の自動素材選出

理学専攻 情報科学コース 清水柚里奈（指導教員：伊藤貴之）

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。

本研究では、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]。本論文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. 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、高水準情報の抽出

(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。

2.1 動画特徴量

画像から受ける印象には個人差がある。文献[2]の実験結果からは、色などの低水準な特徴から印象を受ける人と、被写体などの高水準な特徴から印象を受ける人に二分されることが示唆されている。そこで、本手法では、色分布、動き分布の 2 種類の低水準特徴量と、被写体の種類にもとづく高水準情報を抽出し、これらの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。

高水準情報に関しては、動画から認識されると思われる被写体のうち、最も視聴者の印象に残ると思われる 1 つの被写体を主観的に選び、そのキーワードを手動で付与している。

2.2 音楽特徴量

本手法では、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。

2.3 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。

2.3.1 ユーザ印象評価

まずあらかじ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、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。

2.3.2 動画の低水準特徴量・音楽特徴量からの印象学習

2.1, 2.2 節で示した動画の低水準特徴量・音楽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、動画の低水準特徴量・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。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、また非線形の特徴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、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。

2.3.3 被写体の印象推定

高水準情報である被写体のキーワードから印象値を計算するにあたって、word2vec (<https://code.google.com/archive/p/word2vec/>) を使用した。word2vec は単語の意味ベクトルを大規模なコーパスから学習することで、単語を定量化する。この word2vec を用いて、被写体のキーワードと本手法で使用している感性語間の cos 類似度を計算する。現時点では学習データとして、日本語版 Wikipedia を構成する文書群を用いている。

単語間の類似度を計算するにあたって、まず 3.3.1 項で選出した対となっている感性語（「明るい」「暗い」であれば「明るい」と「暗い」）それぞれと単語間との cos 類似度 X_i, Y_i を計算する。さらに、求めた値に対して以下の計算を行うことで、 i 番目の感性語対に対する印象値 a_i^{high} を求める。ただし、現段階で使用している学習データには「迫力のある」、「迫力のない」の感性語が存在しないため、この感性語対のみ「迫力」と単語間の cos 類似度を計算した値を印象値としている。

$$a_i^{high} = \cos \left(\frac{\cos^{-1} X_i}{\cos^{-1} X_i + \cos^{-1} Y_i} \pi \right)$$

2.3.4 動画の印象値計算

2.3.2 項で求めた色分布、動き分布の印象値と、2.3.3 項で求めた被写体の印象値から、動画の印象値を求める。2.1 節の冒頭でも述べたように、画像閲覧者には低水準な特徴から動画の印象を受ける人と高水準な情報から印象を受ける人に二分されることが示唆されている。このことから、ユーザにあらかじめ、どちらの印象に左右されるかを自己回答してもらった上で、以下の計算式を用いて BGM を付与したい動画の i 番目の印象値 a_i^{movie} を求める。ただし、 a_i^{low} は低水準特徴量から得られた印象値とし、 a_i^{high} は高水準情報から得られた印象値である。

$$a_i^{movie} = s a_i^{low} + (1.0 - s) a_i^{high}$$

現時点では、低水準な特徴に印象が左右されると回答した場合は、 $s=0.8$ とし、高水準な情報に印象が左右されると回答した場合は、 $s=0.2$ として重みづけ計算を行うことで、ユーザの特性を印象値に反映させている。

2.4 楽曲生成

2.4.1 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。2.3.4 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ることで楽曲を生成する。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。

2.4.2 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[4]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。

3. 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[5]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、普段動画を閲覧する際に色や動きといった低水準な特徴に印象が左右されると回答した被験者 A と、被写体のような高水準な特徴に印象が左右されると回答した被験者 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を生成した。以下の 2 種類の動画に対してメロディおよびリズムを選出した結果を表 1 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 1：動画 1,2 の楽曲生成を行った結果

	被験者 A	被験者 B	音色
動画 1	Melody23.mid Rhythm4.mid	Melody3.mid Rhythm9.mid	ハーモニカ
動画 2	Melody11.mid Rhythm21.mid	Melody27.mid Rhythm21.mid	トランペット

被写体のキーワードを動画 1 では海、動画 2 では犬として楽曲を生成した。この結果を被験者 A、被験者 B に評価してもらった。被験者 B からは、動画の雰囲気合っている楽曲が生成されたという意見があった。一方で被験者 A からは、動画 1 の穏やかな海の情景であるのに対し、スネアを多用した激しいリズム素材が選ばれている、という意見があった。この原因として、リズム素材が少ないために SOM によって最適な素材を導くことが難しかったか、あるいはリズムの特徴量を再考する必要があると考えられる。

4. まとめと今後の課題

本論文で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た。

今後の課題として、動画の低水準・高水準どちらの印象に左右されるかを自動判定すること、また動画の内容把握や物体認識があげられる。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] 大山喜彦, 伊藤貴之, ”DIVA:画像の印象に合わせた音楽自動アレンジの一手法の提案”, 芸術科学会論文誌, Vol. 6, No. 3, pp. 126-135, 2007.
- [3] 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ブ性の向上”, 日本知能情報ファジィ学会誌, Vol. 19, No. 6, pp. 611-617, 2007.
- [4] 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.
- [5] 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/>