

投稿データを利用した動画投稿のためのBGM推薦システム*

佐久間 廉[†], 伊藤 克亘[‡],

1 はじめに

近年 YouTube や TikTok など大規模動画共有サイトの普及により、自身が製作した動画を投稿する機会が増えている。しかし、動画編集初心者にとって、膨大な楽曲候補の中から求める楽曲を探し出す作業には、多大な労力と時間を必要とする。そのため、動画に合った音楽を推薦するシステムはいくつか研究がされているが、その動画を視聴した側からの客観的な評価を利用していることが多い。本研究では、投稿者の観点から付与したいBGMを検索するために実際に動画投稿サイトに投稿されている動画と付与されているハッシュタグなどのテキストデータを抽出し、回帰分析を行い対応した楽曲の音響特徴量を推定しデータベースから検索を行う。

2 関連研究

従来のBGM推薦システムでは入力した動画の色特徴量や動きの特徴量をもとに合った曲を出力することを考えられている場合が多く、評価としても客観的に動画と音楽がマッチしているかについて議論している場合が多い。この方法では動画制作者は動画に適している曲の中からでしか選曲することしかできず、動画の雰囲気を変えるためにBGMを付与することが出来ない。本研究では実際に投稿サイトで用いられるようなハッシュタグを動画と同時に入力し、ユーザーの意向に沿ってBGMを推薦できるようなシステムを考えた。

3 提案手法

3.1 概要

提案手法のBGM検索システムは大きく分けて3つの処理段階で構成される。

- (1) 特徴量抽出:ユーザーが入力したBGMを付与したい動画とそれに付与するハッシュタグや説明文のテキストデータから特徴量を抽出する。
- (2) 音響特徴量の推定:前行程で取得した特徴量から回帰分析を行うことで、その動画とテキストに対応した楽曲の音響特徴量を推定する。
- (3) 類似曲の検索:データベースにある楽曲の音響特徴量と推定した特徴量を照合し、類似曲を検索しその曲を推薦曲リストとしてユーザーに出力する。

3.2 楽曲の音響特徴

この研究で使用する楽曲の音響特徴を表1に示す。特徴量はそれぞれフレームの平均を使用した。特徴量を分析するためにはMIRtoolbox[5]を使用する。

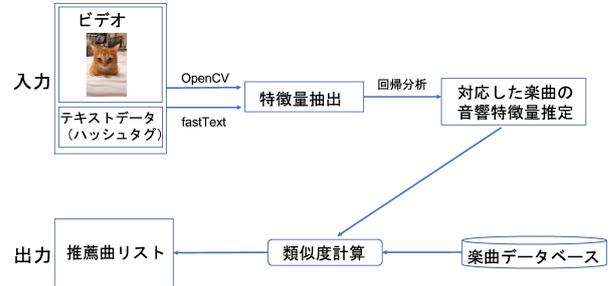


図 1. システム概要

表 1. 使用する音響特徴量

RMS	音響信号の実効値信号の物理的な強度
ZeroCross	音響波形の0との交差回数を表す ノイズの量を表す指標
スペクトル セントロイド	パワースペクトルの重心点の周波数 音響信号の明るさに関係する
スペクトル ロールオフ	低周波数帯から全体の85% を占めるエネルギー量を指す周波数の値
スペクトル フラックス	時間の経過に対する スペクトルの変動
テンポ	楽曲の速さに関係する
ブライトネス	1500Hz以上の周波数 のエネルギーの割合を示す

3.3 動画の特徴量

動画からは色に関する特徴量と動きに関する特徴量合わせて17次元の特徴量を抽出する。色に関する特徴量としてカラーヒストグラムを用いる。OpenCVを用いて12色(黒、灰色、白、茶色、赤、オレンジ、黄色、緑、水色、青、ピンク、紫)へ減色処理を行い、各色の画素数を集計する。得られたヒストグラムの数値から各色の画素数の平均を求め、これを動画全体に対する平均の色の割合とみなし、12次元の特徴量ベクトルとする。

また、動画の動きの特徴量として動画をフレームごとに分割し、OpenCVを用いてオプティカルフローを求める。そのオプティカルフローを構成するベクトル群の速度・角度を集計する。そして、速度の平均、分散速度のヒストグラム上で度数が最大となる階級値角度の分散、角度のヒストグラム上で度数が最大となる階級値各特徴量の全体の平均を求め、合計5つの動きの特徴量を5次元の特徴ベクトルとする。

3.4 テキストの特徴量

テキストから特徴量を抽出するためにはfastTextを用いる。fastTextはword2vecをベースとした単語の分散処理でword2vecよりも高速に処理できる。fastText公式が配布している日本語の学習済みモデルを使用した。単語は300次元のベクトルに変換できる。投稿データに複数の単語が含まれている場合、平均の数値を主

*: BGM recommendation system for video contributors using posted data Ren Sakuma (Hosei Univ.) et al.

[†]法政大学 情報科学部

[‡]法政大学 情報科学部

成分分析によって次元削減を行い、これをテキストの特徴量とした。

4 評価

学習用データセットは実際に動画を投稿するユーザーが動画に対してどのようなBGMを付けて、ハッシュタグなどの情報を付け加えるのかを分析するために、動画投稿サイトTikTokに投稿されている動画から収集した。TikTokの曲検索APIにしたがって、同じBGMを付けている動画を調べることが出来るのでBGMを100曲選び、1曲につき30本、計3000本の動画を分析に用いた。BGMは使用されている動画数が1000本を超えるものから選択し、「オリジナル曲」というタイトルのBGMは特定の動画に合わせてユーザーが作ってアップロードしたものが多く、他の動画にうまく一般か出来ない可能性を考慮して除外した。

動画の長さは全て1分以下であり、動画の内容として”踊ってみた”や”歌詞動画”のような音楽に依存している物も除外した。

楽曲データベースは動画制作者に向けて推薦する楽曲は、YouTubeが提供しているオーディオライブラリ約4000曲を使用する。楽曲はタイトル、ジャンル、モード、アーティスト名、帰属情報、時間(秒)の情報を持っており、それぞれフィルタをかけることが出来る。

楽曲の音響特徴量の推定のために重回帰分析を用いて回帰係数を求め、入力した動画の特徴量とテキスト特徴量からそれに対応した楽曲の音響特徴量を推定する。回帰の式を以下に示す。説明変数 X は動画の特徴量とテキスト特徴量、目的変数 Y は楽曲の音響特徴量である。

$$y_i = a_{i0} + a_{i1}x_{i1} + a_{i2}x_{i2} + a_{i3}x_{i3} + \dots + a_{in}x_{in} \quad (1)$$

音響特徴量を用いた類似曲検索には近似最近傍法のアルゴリズムを使用できるpythonのnmslibモジュールを用いた。

本研究の有用性を検証するために評価用データセットを用いた性能評価を行う。評価用データセットから動画とテキストデータを入力し、元々ついていたBGMが推薦曲リストに出力されるかを検証する。推薦曲のリストは100位までを出力する。

4.1 結果

評価用データセットを用いた実験の結果は10曲のうち2曲のみが推薦曲のリストに出力され、それ以外は100曲のうち出力されない圏外という結果になった。出力された動画3と動画4はどちらも40位として出力された。出力された推薦曲は学習用データセットで見られたような楽曲が出力された。

4.2 考察

上位に検索結果を出力できた動画3はアニメの切り抜き動画であり、余白部分に黒色が多く使われていて、動画4はお菓子作りの動画で鮮やかな色を多く使われていたため、色に特徴のある動画は上手く予測できる

のではないかと考えられる。今後、多変量回帰の推定の精度を上げるためにはサンプル数を増やすことで改善することが出来るのではないかと考えられる。今回、投稿されている動画の中から音楽に依存していないものを選択する作業を手動で行ったため、非常に時間的にコストがかかった。この作業を自動化して、サンプルを多く確保できるように改善していきたい。

また、動画の動きの特徴量としてオプティカルフローを使用した。定点カメラで調理を行う動画など正しく特徴量を求められず動画のテンポとは合わない遅い曲しか推薦されない場合があるため、他の特徴量を使用することを検討したい。動画を収集する中で投稿されている動画にはペットを映したり、食べ物を食べていたり明るい印象の動画が多く、悲しかったり、暗いような印象の動画はあまり見なかった。動画の印象の偏りがあるため、それに応じてデータベースを明るい曲調のものを多くし、暗い印象のものを少なくバランスを捕ることでよりユーザーが求めている楽曲を推薦できるシステムに改善できるのではないかと考えられる。

5 おわりに

本研究では動画投稿サイトのデータを利用し、動画投稿者に向けたBGM推薦システムの実装を行った。推薦の性能を向上させるために使用する特徴量を再考することと、回帰の処理を工夫するなど改良が必要であると考えられる。また、今後実用化をするためにシステムのUIの実装を考えていきたい。

参考文献

- [1] Philippe Mulhem, et al., "Pivot Vector Space Approach for Audio-Video Mixing," IEEE Multimedia 2003, Vol.10, No.2, pp.28-40, 2003.
- [2] oote J et al., "Creating music videos using automatic media analysis," Proceedings of ACM multimedia, New York, pp.553-560, 2002.
- [3] T. Yoshida, T. Hayashi, Otopittan: a music recommendation system for making impressive videos, Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM), IEEE, 2016, pp. 395-396.
- [4] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161- 1178
- [5] O. Lartillot and P. Toivainen. MIR in matlab (II): A toolbox for musical feature extraction from audio. In Proceedings of 5th International Conference on Music Information Retrieval, 2007.
- [6] 岩宮眞一郎:音楽と映像のマルチモーダル・コミュニケーション,九州大出版会(2011).
- [7] 大野直紀, 土屋駿貴, 中村聡史, 山本岳洋,"独立した音楽と映像に対する印象評価と音楽動画の印象の関係性に関する研究", 情報処理学会論文誌, Vol.59, No.3, 929-940(Mar.2018)
- [8] 本颯太, 奥健太,"楽曲-景観データに基づく音響特徴量の分析", DEIM Forum 2018, P2-4
- [9] 熊本忠彦, 太田公子,"印象の基づく検索のための印象語選定法の提案", 情報処理学会論文誌 44(7), 1808-1811, 1003-07-15