投稿データを利用した動画投稿者のための BGM 推薦システム

佐久間廉
Ren Sakuma
法政大学情報科学部コンピュータ科学科
E-mail: ren.sakuma.5j@stu.hosei.ac.jp

Abstract

In recent years, with the spread of large scale video sharing sites such as YouTube and TikTok, there are more and more opportunities to post videos that one has created. However, for a amateur video editor, it takes a lot of effort and time to find the desired music from a huge number of music candidates. For this reason, several systems have been studied to recommend music suitable for videos, but they often use objective evaluations from the viewers of the videos. In this research, we extract features from text data such as hashtags and perform regression analysis to estimate the acoustic features of the corresponding music and search the database. We also demonstrated the usefulness of the system by using an evaluation dataset to verify whether the BGM originally assigned to the input video and text data was output as a recommended song.

1 はじめに

映画、アニメーション、ドラマなど映像メディアは映像だけでは成り立たず、ほとんどが音を伴っている。特殊な効果音や音楽は映像が視聴者に与える印象を高められ、 例えば映画やドラマでは、別れのシーンに悲しい音楽を流すな ど、シーンのムードに合った BGM を使用することで、シーン を効果的に演出できる。映像に BGM を付与するこ とによって他にも様々な効果 (時代や世界観を表現、場面の確 立、映像体験の強調) が得らるため、映像作品 において BGM は重要な役割を担っている。

また、YouTube や TikTok など大規模動画共有サ イトの普及により、製作した動画を誰でも気軽にイン ターネットに投稿する機会が増えているが、その評価基準は動画視聴者しかし、動画編集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，1 曲 1 曲試聴しながら，試行錯誤を繰り 返す必要があるため、多大な労力と時間を必要とする.

また、動画と音楽の調和による BGM 推薦の研究は多く行われているが、その評価は動画を見た視聴者のものであり、動画制作者の印象を考慮している研究はあまり見られない。そのため、本研究では、動画制作者が動画に BGM を付けることによって意図した印象を付けられるように、実際にインターネットに投稿されている動画の説明文やハッシュタグを分析し、BGM データベースから検索することで、動画編集の初心者でも製作した動画に簡単に BGM を 選定できる様なシステムを考えることとした。

Supervisor: Prof. Katunobu Itou

2 関連研究

2.1 楽曲推薦システム

映像編集において、BGM 付与を支援する研究が多く報告されている。[1] は、映像編集のルールに基づき、映像と音楽の特徴量を、ダイナミクス、ピッチ、モーションの 3 つのカテゴリ分け、それぞれのカテゴリにおいて、映像の特徴と相関の高い特徴を持つ楽曲を BGM として推薦する手法を提案している。

[2] は楽曲構造の境界に合わせて、映像の動きや輝度のちらつきが少ないショットを切り貼りすることでミュージックビデオを自動で生成する手法を提案している.

また、楽曲検索の研究については, 感性ムードタグを用いた楽曲検索の研究がある。この手法は BGM データベースサイトでも使われていて、”明るい”や”ハッピー”などの楽曲にタグ付けされたキーワードで検索することによって楽曲を絞り込む。しかし、キーワードに当てはまらない楽曲は検索することが出来ないうえ、同じキーワードが当てはまる楽曲でもその微妙なニュアンスの違いを考慮して曲を検索することが出来ないという問題点がある。

人の感性を表現する AV 空間を用いてビデオに BGM 推薦を行っているのが [4,5] である。AV 空間は Arousal 軸と Valence 軸からなる二次元平面であり、Arousal は感情の興奮の度合い、Valence は感情のポジティブ-ネガティブ度合いを表す値である。この手法では 30 次元の映像特徴量、20 次元の音響特徴量をあわせた特徴ベクトル次元数を減らすためにを主成分分析し、トレーニングデータの AV 値と特徴ベクトルのペアから重回帰分析によって楽曲推薦を行い、ユーザーが望むような印象を動画に付与出来たか五段階評価で平均 3.5 標準偏差 0.79 という結果を得た。しかし、重回帰分析で AV 値を推定する点とユーザーが表現したい感情を正しく AV 値として入力することが難しいことが問題点である。

2.2 映像と音楽の関係

音楽と映像の印象がどのように音楽動画の印象に影響するのか調べたのが [6] である。「音楽のみ」「映像のみ」「音楽動画」の 3 つの関係性に着目し、これらに対する 8 印象軸の印象評価データセットを構築し、それらを分析することで音楽と映像の印象評価の組み合わせによる音楽動画の印象推定の可能性について検討を行った。

この研究の中で、音楽動画に対して、「音楽のみ」「映像のみ」の印象ベクトルよりも音楽と映像の印象ベクトルの平均の方がコサイン類似度が 0.8 以上の割合が高い結果が得られたため、音楽の印象と映像の印象を組み合わせることで音楽動画の印象推定が出来る可能性があることを示した。

また、音楽と映像をランダムに組み合わせたでも同じように

評価実験を行い、映像作者と音楽製作者間での意思の取り合いが出来ている場合と、音楽と映像をランダムに組み合わせて制作したような、制作者の意思が反映できていない音楽動画では印象の受け方が変わってくることが分かった。

3　投稿データを利用した BGM 推薦システム

3.1　概要

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制作した。

本研究の提案手法である 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。

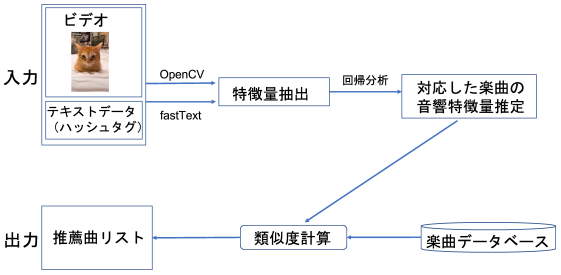

図 1. システム概要

3.2　楽曲の音響特徴

画像に合わせた音楽に関する先行研究 [12] で用いられている 7 次元の特徴量を本研究でも使用する。**RMS**

音響信号の実行値を表す。音響信号の物理的な強度と関連する。算出式は次式である、n は解析窓内のサンプル数、x_i は第 i 番目のサンプル値を示している。

$$y_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2} \tag{1}$$

ゼロ交差 (ZeroCross)

音響波形の振幅数が 0 との交差回数を表す。ノイズの量を表す指標として知られている。算出式は次式であり、x_i は第 i 番目のサンプル値を表しており、t は解析窓の開始時刻、t' は解析窓の終了時刻を示している。θ は交差の有無を表しており、交差した場合に 1 を、そうでない場合に 0 の値を出力する。

$$y_{zc} = \sum_{i=t}^{t'} \theta \begin{cases} 1 & (x_i x_{i-1} < 0) \\ 0 & (x_i x_{i-1} > 0) \end{cases} \tag{2}$$

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を表す。音響信号の明るさに関係する指標である。算出式は次式であり、x は周波数を示し、$f(x)$ は x の時のパワーの値を示す。

$$y_{sc} = \int x f(x) dx \tag{3}$$

スペクトルロールオフ

パワースペクトルにおいて、低周波数帯域から全体の 85 % を占めるエネルギー量を指す周波数の値を表す。音響信号の明るさに関係する指標である。算出式は次式であり、$M_t[\mathrm{k}]$ は t 番目の窓をフーリエ変換して得られるスペクトルにおける、k 番目の周波数ビンにおける振幅値を示す。

$$y_{sro} = 0.85 * \sum_{k=1}^{K} M_t[k] \tag{4}$$

スペクトルフラックス

スペクトルフラックスは、時間の経過に対するスペクトルの変動の尺度である。算出式は次式であり、s_k はビン k においてのスペクトルの値を表し、b_1, b_2 はバンドのエッジ,P はノルムタイプを示す。

$$y_{sf} = \left(\sum_{k=b_1}^{b_2} |s_k(t) - s_k(t-1)|^P \right)^{1/P} \tag{5}$$

テンポ

オンセット時刻から周期性を検出することで得られる値を表す。楽曲の速さに関連する。

ブライトネス

カットオフ周波数を 1500Hz とした高周波数のエネルギーを計算することで求められる。楽曲の明るさに関係している。

3.3　動画の特徴量

[12] では画像から受ける印象には個人差があり、色など低次元な特徴から印象を受ける人と、被写体などの高次元の意味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。この研究での実装では色分布に関する特徴量と動き分布に関する 2 種類の低次元特徴量合わせて 17 次元の特徴量を抽出する.

色分布の特徴量

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、緑、紫、オレンジ、赤、青、黄色、水色、ピンク）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。

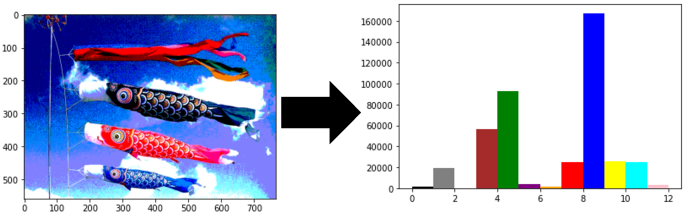

図 2. 画像からカラーヒストグラムを抽出

動き分布の特徴量

映像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、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、それぞれの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。

図 3. オプティカルフローの検出

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。

表 1. 投稿された動画に付与されたテキスト情報の一部

動画 1	赤ちゃん	犬のいる暮らし	トイプードル	甘えん坊
動画 2	お父さんと息子	赤ちゃん	お尻	たまらん
動画 3	風邪	赤ちゃんのいる生活	熱	冷えピタチャレンジ
動画 4	部屋汚い	いい人	末期	
動画 5	次男	子供のいる暮らし	こどものいる暮らし	平和な日常
動画 6	チワワ	動物コレクション	犬のいる暮らし	子犬
動画 7	コロナに負けるな	子供のいる暮らし	うちの子が可愛すぎる	
動画 8	ぶく顔	赤ちゃん	おもしろ	女の子ベビー
動画 9	マンマ	子供のいる暮らし	おすすめ	怒る
動画 10	娘	可愛い	女の子	癒し

投稿されている動画のハッシュタグには流行り言葉やアニメのタイトル、商品名などモデルには対応していない単語が多く見られた。そのような単語をベクトル化するために wikipedia の API を使用し、その単語の要約文を取得し、その要約文を MeCab を用いて名詞を抜き出しその平均のベクトルを対応していない単語の単語ベクトルとして処理を行った。対応していない単語をベクトル化しその類似語を検索した結果似たような単語が求められた。(表 2)

表 2. 対応していない単語のベクトルの類似度から類似語を検索

ジャニーズの検索結果	
アズリードカンパニー	0.6251
エクセルヒューマンエイジェンシー	0.6200
バルドエージェンシー	0.6059
ジョビィキッズプロダクション	0.6016
ホーリーピーク	0.6015
アールグルッペ	0.6014
スターレイプロダクション	0.6010
スリーファンキーズ	0.5995
ベリーベリープロダクション	0.5975
バンビプロモーション	0.5959

3.5 楽曲の音響特徴量の推定

求めた特徴量を主成分分析により色特徴、音響特徴量それぞれ 2 次元に次元削減処理を行い、scikit-learn を用いて、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。

$$y_i = a_{i0} + a_{i1}x_{i1} + a_{i2}x_{i2} + a_{i3}x_{i3} + \cdots + a_{in}x_{in} \quad (6)$$

推定した音響特徴量 y_i を用いて、楽曲データベースから類似曲を推薦曲リストに出力する。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。

3.6 データセット

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。ペットに関するタグを持った 610 動画を収集し、そこから 10 動画を無作為に選びそれを評価用のデータセットとして用いた。動画はいいね数の多いものから条件に合うものを選択した。また、「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化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。

図 4. TikTok から収集した動画の一部

動画制作者に向けて推薦する楽曲のデータベースとして 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、ムード、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られる。選択できるジャンルとムードは以下に示す (表 3, 表 4)。

表 3. ムード

怒り
明るい
穏やか
暗い
ドラマチック
ファンキー
ハッピー
インスピレーション
ロマンチック
悲しい

表 4. ジャンル

オルタナティブ & パンク
アンビエント
子供向け
映画
クラシック
カントリー & フォーク
ダンス & エレクトロニック
ヒップホップ & ラップ
ホリデー
ジャズ & ブルース
ポップ
R& B & ソウル
レゲエ
ロック

4 システムの性能評価

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。

4.1 評価実験

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。推薦曲リストは上位 100 曲までを出力する。データ数が 100 から 600 まで 100 ずつ増やし、それぞれ実験を 10 回行った。

4.2 結果

各データ数の実験結果を表 5 に示す。最も出力された割合が高かったのは 500 の場合で、最高順位は 300 の場合だった。順位が高く出力された動画に付けられていたハッシュタグには”サロン”、”トリミング”、”眠気”、”日常”、”動物園”、”作戦”、”コメディー”などが見られた。動画の印象は動きが激しいものではあまり見られなかった。反対に出力されない動画はデータセットの中でも多くの物が付けられている”可愛い”、”猫”、”犬”といったハッシュタグを持っていた。

推薦曲リストには動きの速い動画に対してはテンポの速い曲が出力され、赤や黒色の割合の多い動画には YouTube オーディオライブラリのジャンル (表 3,4) でロックを持っている曲が多く、白や茶色の割合の多い動画にはハッピー、明るいといったジャンルの曲が多く出力された。本システムでは元々ついている曲を正解とした場合の精度は高い結果を得ることが出来なかったが、推薦曲リストの上位 5 曲の中でその動画につけたいと思える物を出力することはできた。

表 5. 実験結果

データ数	100	200	300	400	500	600
平均順位	42.7 位	42.2 位	39.5 位	47.8 位	34.7 位	34.0 位
出力された割合	4%	4%	7%	8%	9%	7%
最高順位	16 位	36 位	8 位	36 位	21 位	13 位
決定係数	0.8123	0.7237	0.6240	0.5674	0.5083	0.4564

4.3 考察

データ数が 100 と 200 では学習に使用したデータが少ないため、過学習を起こし、未知のデータに適合することが出来ないかったため精度が落ちたと考えられる。

また、順位が高く出力される動画にはハッシュタグに被写体の意味や状態をを表すような情報が付けられていた場合が多くみられた。これは使用した動画特徴量では表しきれない高次元の特徴量を補完することが出来ていたためだと思われる。そのため、学習用のデータセットとして多く情報を含む動画を厳選して集めるか、物体検出を行い被写体の名前など高次な特徴量を学習に用いることで推薦の精度が上がると考えられる。

反対に推薦リストに出力されない動画には多くの動画に付けられているハッシュタグを持っている場合が多かった。これによって特徴量の差別化が難しくなり、回帰の精度が落ちてしまったと思われる。

データセットを作成する際、投稿されている動画の中から”ダンス動画”や”歌詞に合わせた動画”などの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、非常に時間がかかった。

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が、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と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、複雑な分布にも対応させるためニューラルネットワークを用い、回帰の処理を改良させる必要があると考えられる。また、今後実用化をするためにシステムの UI を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] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161- 1178

[4] 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.

[5] 小野佑大、石先広美、帆足啓一郎、小野智弘、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”、情報処理学会研究報告 Vol.2011-MUS-89 No.16

[6] O. Lartillot and P. Toiviainen. MIR in matlab (II): A toolbox for musical feature extraction from audio. In Proceedings of 5th International Conference on Music Information Retrieval, 2007.

[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] Jacek Grekow, ”Music Emotion Maps in Arosal-Valence Space”, CISIM 2016: Computer Information Systems and Industrial Management pp 697-706.

[10] 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

[11] 大山喜冴、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,No3,pp,126-135,2007

[12] Chien-Liang Liu and Ying-Chuan Chen. Background music recommendation based on latent factors and moods. Knowledge-Based Systems, 159:158?170, 2018.