Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

主観特徴と物理特徴の融合による音楽動画印象推定手法の検討

上西隆平[†1]　阿部和樹[†1]　大野直紀[†1]　土屋駿貴[†1]　中村聡史[†1]

概要：音楽と映像が一体となった動画コンテンツ（音楽動画）の検索方法の 1 つとして，近年印象に基づく検索に関する研究が盛んに行われている．この印象に基づく検索を実現するためには，音楽動画自体の印象を推定し，付与することが必要であるが容易ではない．我々はこれまでの研究において，音楽動画の音楽特徴，映像特徴，ソーシャルコメント特徴の 3 つを用いて音楽動画の印象推定を行なってきた．本研究では，その中でも特に映像とソーシャルコメントの特徴に着目し，各特徴の最適な抽出方法について検討する．また，その特徴量を用いることで音楽，映像，音楽動画のそれぞれについて印象の推定精度の改善を図る．

キーワード：印象推定，音楽動画検索，印象検索，映像印象，ソーシャルコメント

1. はじめに

音楽や映像などのコンテンツを個人で容易に制作できるようになり，それらを配信する YouTube やニコニコ動画といった動画共有サービスが普及した．それに伴い，これらのサービスを利用するユーザの数も増加し，音楽動画の数が飛躍的に増加している．しかし，音楽動画の数が増加する一方で, 音楽動画を検索するための手法は限られており，現状では曲名やアーティスト名など，音楽動画に対する事前情報を用いた検索が主流である．そのため，十分な知識がない限り未知の音楽動画を探し出すことが困難となっている．

未知の動画への到達を支援するため， YouTube ではユーザの閲覧した動画を元に，次の動画を推薦する機能が備わっている．しかし，この機能は同じ投稿者の動画や，同じ動画を視聴した他ユーザの履歴を推薦するといった機能であり，ユーザの好みに沿っていない事が多い．また，ニコニコ動画ではランキング機能によって未知の動画を探すことができるが，この機能においても，再生数やコメント数などが上位のものしか提示されず, ユーザが求めている「埋もれた」動画を見つけることはできない．

このような問題を解決する方法の一つに,「楽しい気分になれる音楽動画」や「切ない気持ちになれる音楽動画」といった，ユーザの主観にもとづいた音楽動画の印象検索が考えられる．こうした検索が確立されれば， 音楽動画の検索における新たな選択肢の一つとなり，ユーザが未知の音楽動画を検索する際に利用できるばかりか，真に求めている音楽動画と出会う機会が増えると期待される．また，音楽動画の投稿者にとっても，今まで埋もれていた自身の音楽動画が視聴されることになり，コンテンツ作成のモチベーションを上げることができるなど，相互に良い影響を与えられることが期待される．

音楽動画の印象検索を実現するには，動画に付与されているソーシャルタグを利用することが考えられる. しかし, そのソーシャルタグに含まれる印象にまつわるキーワードは全体の 5%程度であり[1], 検索手段として十分ではない. そのため，音楽動画の印象検索を行うためには，音楽動画に対してユーザが抱くと予想される印象を推定する必要がある．また，音楽動画の印象推定が可能になると，印象検索ができるようになるだけでなく，印象にもとづく音楽動画の推薦なども可能になると期待される．

音楽動画の印象推定に関する研究としては，ソーシャルコメントを利用したもの[2]や，音楽特徴量を利用したもの[3]，音楽特徴・映像特徴・ソーシャルコメントの 3 つの特徴を組み合わせたもの[4]など様々なものがある．ここで我々の過去の研究[4]では，映像特徴について十分に考慮できておらず精度が高くはなかった．そのため，この特徴量の抽出方法を改良することで推定精度を向上できると期待される．また，[2]や[4]の研究ではソーシャルコメントを利用し，音楽動画の音楽・映像に対しても印象推定を行っているが，その際ソーシャルコメントをそのまま利用している点にも改良の余地がある．

そこで本研究では，映像ならびに音楽動画に対する印象推定の精度を向上させるために，映像特徴を用いた推定手法を改良する．また，ソーシャルコメントから音楽・映像に対するコメントを抽出し，それらを用いて音楽動画の音楽・映像それぞれに対する印象推定の実験及び，分析を行う．具体的には，すでに構築してある 500 個の音楽動画データセットを利用し以下の 2 つの実験を行う．

- 映像の印象推定において，最も精度が出る映像特徴（色特徴）の検討
- 音楽動画に付与されているソーシャルコメントを音楽に対するもの，映像に対するものに分離し，対応するコメントを用いた音楽及び映像の印象推定

†1 明治大学
Meiji University

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

2. 関連研究

印象の表現方法については，楽曲の印象をモデル化する方法の一つとして，Hevner の研究[5]がある．Hevner は楽曲に対する印象の分類を，8 つのクラスを用いて行っている．他にも Russel の研究[6]で提案されている Valence-Arousal 空間は，印象推定に関する研究で多く使われている．Valence-Arousal 空間とは Valence（快-不快）と Arousal（覚醒-鎮静）の 2 つの軸で構成された空間上で印象を表現する方法である．また山本ら[1]の研究では，音楽動画に対する印象を 8 つのクラスに分類し，音楽動画に対する印象の評価データを収集している．

音楽動画の印象推定としては，Esra ら[7]の研究が知られている．Esra らの研究では，音楽動画の音楽情報（メル周波数ケプストラ係数）と映像情報（RGB 値）を用いて感情ラベルの推定を行っている．その結果，映像情報よりも音楽情報を利用したほうが推定の精度が高いことがわかった．また，音楽情報と映像情報といった物理的な特徴量だけでなく，音楽動画に対するソーシャルコメントなどの主観的な特徴量を使った印象推定の研究も行われている．土屋ら[2]の研究では，音楽動画に付与されるソーシャルコメントを利用し，音楽動画の印象推定においてどのようなコメントの品詞が良いのか，特徴量はどのように取得したらよいのかといった点について検討している．また，阿部ら[4]は音楽動画に対し，音楽特徴・映像特徴・ソーシャルコメントの特徴を組み合わせた印象推定を行っている．結果として，音楽・映像・ソーシャルコメントの 3 つの特徴を組み合わせた際の印象推定の精度が最も高いことを明らかにしている．さらに，ソーシャルコメントの特徴を単独で利用した場合の精度に比べ，音楽特徴・映像特徴を単独で用いた場合の精度が低くなるという結果が出ている．

そこで我々は音楽動画に対する，音楽・映像・ソーシャルコメントを組み合わせた印象推定の精度を向上させるため，Esra や阿部らの研究の中で精度が低いとされていた映像特徴を用いた印象推定の精度の向上とソーシャルコメントから音楽や映像に対する情報を抽出し，その情報を元に印象推定を行う．

映像の印象推定を行う研究の 1 つとして，清水ら[8]は，映像から得られる色情報と動きの情報を特徴として印象推定を行い，そこから推定された映像の印象をもとに，最適な楽曲を映像に付与するという研究を行っている．また，映像ではないが静止画に対する印象を推定する研究は多数存在する．例えば上坂ら[9]は，人間が画像を認識する際に色とテクスチャによって印象を判断していることから，色特徴（Lab 色空間）と人間の目の特徴抽出機構を模した 3 点間コントラストによる特徴を用いて印象推定を行っている．

本研究ではこれらの研究と同様に，映像の色特徴に着目し，印象推定に最適な色特徴の組み合わせについて検討していくものである．

3. 印象評価データセット

本研究では，音楽動画の印象を推定する際，大野ら[10]の研究において構築した印象評価データセット[11]を利用する．このデータセットは，音楽動画のサビ部分 30 秒間において，ユーザが抱く印象を調査したものであり，音楽動画を「音楽」「映像」「音楽動画（音楽+映像）」の 3 つのメディアタイプに分け，それぞれから受ける印象が評価されている．評価対象の音楽動画は，ニコニコ動画にある「VOCALOID」タグが付与されている動画のうち，2012 年 8 月時点で再生数の多い上位 500 件を利用しており，データセットを構築する際，各メディアタイプで 少なくとも 3 名以上印象評価を行っているものである．

表 1　利用した 8 つの印象クラス

印象クラス名	印象を表す形容詞
C1（堂々）	堂々とした，どっしりとした， 心躍る，にぎやかな
C2（元気が出る）	元気が出る，楽しい気持ちにさせる， 陽気な，心地よい
C3（切ない）	切ない，悲痛な，ほろ苦い， 気が滅入る，哀愁の
C4（激しい）	アグレッシブな，激しい，興奮させる， 熱情的な，感情あらわな
C5（滑稽）	滑稽な，ユーモラスな，面白げな，奇抜な， 気まぐれな，いたずらっぽい
C6（可愛い）	可愛らしい，愛くるしげ， 愛おしい，かわいい
Valence	明るい気持ちになる,楽しい 暗い気持ちになる，悲しい
Arousal	激しい，積極的な，強気な 穏やか，消極的な，弱気な

評価に用いた印象クラスについては，音楽検索ワークショップである MIREX で利用されている 5 つの印象クラスと，Russel が提案した Valence-Arousal 空間に「可愛い」という印象を追加した計 8 つのクラスを利用している．8 つの印象クラスについてまとめたものを表 1 に示す．表中の「印象クラス名」は便宜上付与されている印象を表すラベル名であり，「印象を表す形容詞」は，評価者に評価してもらう際にその印象クラスを表現するために用いられたもの

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

である．

評価の方法は，C1 から C6 については 1（全くそう思わない）〜5（とてもそう思う），Valence については-2（暗い気持ちになる，悲しい）〜+2（明るい気持ちになる，楽しい），Arousal については-2（穏やか，消極的な，弱気な）〜+2（積極的な，強気な）の 5 段階で評価されている．

つまり，このデータセットは，500 件の音楽動画×3 つのクラスタイプ×8 つの印象クラスの，合計 12,000 件以上のデータからなっている．[11]ではこのデータが公開されている．なお，C1 から C6 については，Valence や Arousal と統一した処理を行うため，値を 3 減らして，値域を-2 から 2 に変換して扱う．

4. 映像特徴量による印象推定

我々は音楽動画に対する印象を推定するうえで，動画の映像特徴量と動画に付与されているソーシャルコメントに着目している．本章では音楽動画の印象推定に有効な映像特徴量についての検討を行う．

ここで，清水ら[8]の研究では，映像の色情報と動きの情報（オプティカルフロー）を用いて映像の印象推定を行っており，映像の色情報を用いた印象推定が特に有効であったことを明らかにしている．そこで，我々も清水らと同様に色特徴を利用して，音楽動画の映像に対する印象推定手法の改良を行う．

4.1 色情報の抽出手法

本研究では，映像から色情報を抽出する際，映像を一定間隔で画像として切り出し，それらから色情報の抽出を行う．ここで，画像を切り出す際の間隔に関しては阿部ら[4]の研究が検証しており，5 秒の間隔で画像を切り出した場合が，計算コスト面や精度面でも印象推定に最も有効な手段であり，清水らの手法でも 5 秒としていたため，本研究でも画像を切り出す間隔は 5 秒とした．次に，どの色情報を用いて特徴とするかについて，清水らは映像から一定間隔で切り出された各画像に対して，12 色（赤，橙，桃，黄，緑，青，水，紫，茶，白，灰，黒）への減色処理を行い，すべての画像の画素数を集計することでカラーヒストグラムを生成している．こうして得られた全画像の画素数を用いて，各色の平均の画素数を求め，それを映像全体に対する 12 次元の特徴ベクトルとしている．また阿部らは清水らと異なり，RGB それぞれを 3 諧調とした 27 色に減色処理を行い，27 次元の特徴ベクトルによって印象推定を行っている．

清水らの研究では 12 色，阿部らの研究では 27 色すべてを利用して印象推定を行っているが，これらの研究で使用した色が映像に対する印象推定において最適な色なのかという検証を行っていない．そこで我々は，組み合わせる色の数を減らしたうえで，色の組み合わせ自体を変えることにより，映像に対する印象推定において印象クラスごとに最適な色の組み合わせや，組み合わせ数を模索する．

その方法は以下の手順になる．

(1) 清水らが用いた RGB 色空間における 12 色（赤，橙，桃，黄，緑，青，水，紫，茶，白，灰，黒）を用意する．
(2) 全 12 色すべての組み合わせである 4,095 通り（$= {}_{12}C_1 + {}_{12}C_2 + \ldots + {}_{12}C_{11} + {}_{12}C_{12}$）の組み合わせを作成する．
(3) 映像から一定間隔で切り出された画像の各画素と，(2)で作成した色のリストにある色とのユークリッド距離を 1 つ 1 つ計算する．
(4) 最もユークリッド距離が近かったリスト内の色に画像の画素を合わせることで，リスト内の色の数に画像を減色する（リストが「赤橙桃灰黒」だった場合，画像がリスト内の色である「赤橙桃灰黒」の 5 色で表現されることとなる）．
(5) 一定間隔で分割された画像それぞれに(3)および(4)の処理を行ったものについて，画素数の平均を求める．

上記の手順によって得られたものについて正規化を行い，色の数の次元を持った映像特徴ベクトルとして扱う．

4.2 評価実験

本章で行う評価実験では，3 章で説明したデータセットの中で，映像に対する印象値のデータを用いる．8 つの印象クラスごとに，評価値が 1 以上の高評価群と，-1 以下の低評価群の動画のみを印象推定の対象とする．この際，データの偏りをなくすために，高評価群と低評価群の動画数を一致するように設定した．各印象タイプにおいて対象とした動画数を表 2 に示す．

表 2　映像の印象分類実験に使用した動画数

印象クラス	動画数	印象クラス	動画数
C1（堂々とした）	42	C5（滑稽な）	158
C2（元気が出る）	100	C6（かわいい）	154
C3（切ない）	280	Valence	112
C4（激しい）	96	Arousal	216

実験は，4.1 節で得られた特徴量を用いて，機械学習を行い，高評価群，低評価群に該当するものを，どの程度の確率で分類できるかによって推定精度の評価を行う．分類器としてはサポートベクターマシン（SVM）を用いた．具体的には高評価群を正例，低評価群を不例とし，それらを 5 分割し，そのうちの 4 つを SVM の訓練データ，残りの 1 つをテストデータとして交差検定（5-fold クロスバリデーション）を行い，全体の正解率を計算する．

4.3 結果と考察

表 3 は音楽動画の映像に対する印象推定を行った際の，8 つの各印象クラスと，その印象クラスにおいて正解率が高かった色の組み合わせ，そしてその正解率をまとめたも

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

表 3　正解率が高い上位 3 つの色の組み合わせとその正解率の平均

	C1	C2	C3	C4	C5	C6	V	A	C1~A 平均
1 位	赤橙緑桃 0.880	青赤緑黄水紫黒 0.870	紫水赤橙青桃 0.748	赤橙黄青 0.896	緑紫黒 0.718	白水赤橙黒桃 0.883	赤黄紫黒桃 0.820	赤紫青白橙緑黒水桃 0.810	赤橙緑紫黒桃 0.776
2 位	黄緑桃 0.875	橙赤緑黄水紫黒 0.840	緑黄紫黒赤橙青桃 0.747	赤橙緑桃 0.870	赤緑黄桃 0.706	白緑水赤橙黒桃 0.877	赤緑紫茶白 0.811	白橙緑黒水桃 0.810	赤橙青水紫黒桃 0.773
3 位	黄青水桃 0.860	白黄紫 0.840	水赤橙青桃 0.747	赤紫茶白灰黒 0.865	赤緑青紫 0.706	緑水赤橙黒桃 0.871	黄緑青白 0.802	紫青橙緑黒水桃 0.806	赤橙黄緑青水紫桃 0.772

のである．

また，本手法における印象クラスごとに最も精度が高かった正解率と，阿部らの手法である 27 色全てを用いた場合の正解率，清水らの 12 色全てを用いた場合の正解率を表 4 に示す．なお，ここで表 3 および表 4 における V は Valence，A は Arousal を示すものである．

表 4　比較対象の正解率

	C1	C2	C3	C4	C5	C6	V	A	平均
本手法	0.880	0.870	0.748	0.896	0.718	0.883	0.820	0.810	0.828
27 色	0.855	0.789	0.722	0.811	0.575	0.877	0.740	0.778	0.768
12 色	0.805	0.770	0.687	0.794	0.600	0.832	0.694	0.765	0.743

表 3 より，まず清水らの 12 色のものより阿部らの 27 色のものが，正解率が高くなっていることがわかる．一方，表 3 および表 4 より，3 色～9 色程度の色に減らしたものの方が，12 色や 27 色のものに比べ正解率が高いことと，印象ごとに効果がある色の組み合わせが異なっていることがわかる．

また，各印象クラスの正解率を見ていくと，C1（堂々とした），C2（元気が出る），C4（激しい），C6（かわいい）は 1 位の正解率が 0.85 以上となっており，こうした印象においては色による推定精度は十分な高さであることがわかる．その一方で，C3（切ない）や C5（滑稽な）の正解率は 0.75 以下と推定精度が低くなっており，こうした印象においては色による推定は難しいことがわかる．しかし，C5 では 0.718 と低い精度であるが，27 色や 12 色での C5 の精度は 0.6 以下となっているため，色を限定することが精度向上につながっていることがわかる．

次に，表 3 で推定精度が高かった上位 3 位の色について注目すると，印象クラスごとに共通して利用されている色が存在していることがわかる．具体的には C1 は「桃」C2 は「黄・紫」，C3 は「青・赤・桃・橙」，C4 は「赤」，C5 は「緑」，C6 は「水・赤・桃・橙・黒」，Arousal は「橙・緑・黒・水・桃」となっている．特に，C1，C2，C4，C5 は共通している色が限定されており，これらの色が推定に強く影響していると考えられる．また，C2（元気が出る）は黄色，C4（激しい）は赤色など，一般的なイメージと一致する色が選択されている．さらに，各印象クラスの正解率が高い上位 3 件の色の組み合わせにおける，最大の組み合わせ数は 8 つとなっている．

これらの結果から，音楽動画の映像について色特徴を用いて印象推定を行う際には，色の組み合わせを考慮することで正解率が高く出るという傾向があることが分かった．また，印象クラスごとに推定をするのが得意な色の組み合わせを使い分けることにより，正解率を向上させることができることや，印象クラスごとに重要な役割をもつ色が存在することが分かった．

5. コメントを用いた印象推定の検討

本章では動画に付与されているソーシャルコメントを利用した，印象推定について検討する．

先述の通り，ソーシャルコメントを利用した印象推定は既に行われており，土屋ら[2]や阿部ら[4]はソーシャルコメントを用いて，映像，音楽，音楽動画の 3 つのメディアタイプに対する印象推定を行い，コメントによる印象推定の有効性を示している．ここで，大野らの研究[10]では，音楽動画における音楽のみ・映像のみといったメディアタイプごとの印象には差異があることを明らかになっている．また，音楽動画に付与されているソーシャルコメントには，音楽に向けられたコメント，映像に向けられたコメント，それ以外に向けられたコメント（投稿者に対する期待や感謝のコメント，他者との対話や歌詞）などがある．そこで，そのコメントがどのメディアタイプに向けられたコメントであるかを判定し，音楽・映像それぞれについて推定において適切でないコメントを除去することが，精度向上において重要であると考えられる．

そこで本研究では，動画に付与されるソーシャルコメントを「音楽に対するコメント」および「映像に対するコメント」となるようにフィルタリングし，それぞれのメディアタイプにおける印象推定に利用する．

5.1 ソーシャルコメントの収集

本研究では，ソーシャルコメントを用いて印象推定を行うため，3 章で得られた印象評価データセットに存在する音楽動画に付与されたコメントを収集する．特に印象評価データセットの評価対象となった，音楽動画のサビ部分の

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

時間を参考に，サビ区間内に投稿されているコメントを抽出した．収集したコメントは合計 132,050 件である．

5.2 コメントのフィルタリング

収集したソーシャルコメントを「音楽に対するコメント」および「映像に対するコメント」となるようにフィルタリングする．

音楽・映像のそれぞれのメディアタイプに向けられたコメントのみを用いる場合，コメントの数が限られたものとなり，推定に有効なコメントまで失ってしまう可能性がある．そこで本手法では，異なるメディアタイプに向けられたコメントを「ノイズコメント」とし，それらを除外することでコメントのフィルタリングを行う．例えば「音楽に対するコメント」となるようにフィルタリングする場合，映像に向けられたコメントをノイズコメントとし，それらを除外する．

コメントをフィルタリングする手順を以下に示す．

(1) 著者らの協議により，各メディアタイプに向けられるコメントに含まれる可能性のある単語群（音楽であれば「歌」，映像であれば「イラスト」など）を生成する．なお，本手順において用意した単語群は表 5 の通りである．
(2) (1)の単語群を補強するため，今回収集したソーシャルコメントから Word2Vec を用いて類似単語を追加する．その際，Word2Vec のモデルについては，全 35 億のニコニコ動画のコメントを学習したもの[12]を利用する．
(3) コメントの中で，(1)および(2)によって得られた単語が含まれるコメントの場合，それらをノイズコメントとして検出する．
(4) 音楽・映像それぞれのメディアタイプにおけるノイズコメントを除去し，それらを「音楽に対するコメント」「映像に対するコメント」とする．

これらの処理によって，本研究で用いるデータセット内の音楽動画 500 件それぞれに対し，「音楽に対するコメント」「映像に対するコメント」を取得した．

表 5 フィルタリングに使用する単語群

音楽	映像
歌，聴，聞，音，曲，メロディ，BGM，耳，歌って，リズム，音圧，ドラム，イントロ，声	絵，サムネ，顔，イラスト，映像，きれい，綺麗，顔文字

5.3 単語ベクトルの作成

本研究では，ソーシャルコメントによる各印象に対する推定の正解率を算出するため，機械学習ができるように 5.2 節で生成されたソーシャルコメントから，特徴量を抽出する必要がある．本研究では土屋ら[2]や阿部ら[4]と同じく，コメントの形容詞に着目し，TF-IDF 値を特徴量として生成する．ここで TF-IDF 値の計算には，実験対象とする各動画のコメントをそれぞれ 1 つのドキュメントとして利用する．つまり，DF 値の分母は，各印象において実験対象とする動画数となる．1 つのドキュメントにおける各単語（形容詞）の TF-IDF 値を求め，単語の数だけのベクトルを持った特徴量を生成する．また，生成する際にドキュメント全体で出現頻度の高い 30 個の形容詞を対象とするため，30 次元の特徴ベクトルを持つものとする．形容詞の抽出は MeCab を用いて行った．

5.4 評価実験

本章で行う評価実験は 4.2 節で説明した映像の印象分類実験に使用したものに加え音楽の印象分類も行う．その際に印象評価の対象とする条件は 4.2 節と同様にして行う．表 6 に音楽の印象分類実験で対象とした，各メディア・印象タイプにおける動画数を示す．

表 6 音楽の印象分類実験に使用した動画数

印象クラス	動画数	印象クラス	動画数
C1（堂々とした）	108	C5（滑稽な）	96
C2（元気が出る）	164	C6（かわいい）	142
C3（切ない）	90	Valence	120
C4（激しい）	138	Arousal	82

実験における学習も 4.2 節と同様に行い，5-fold クロスバリデーションにて正解率を算出する．

5.5 映像に対するコメントを用いた印象推定結果

5.2 節で生成した，「映像に対するコメント」を用いた印象推定（F あり）と，何も処理を加えていないソーシャルコメントを用いた印象推定（F なし）の比較を行う．それぞれの推定における正解率を以下の表 7 に示す．

表 7 映像に対する分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
F あり	**0.890**	**0.900**	**0.805**	**0.838**	**0.708**	**0.890**	**0.757**	**0.889**	**0.835**
F なし	**0.810**	**0.870**	**0.787**	**0.764**	**0.695**	**0.876**	**0.730**	**0.879**	**0.801**

表 7 の通り，映像向けに投稿されているコメントを利用した場合（F あり），すべてのコメントを利用しているもの（F なし）に比べ，全クラスの正解率が上昇するという結果となった．C1（堂々とした）・C2（元気が出る）はどれも 0.03 以上精度が向上し，特に C1 クラスでは推定精度が 0.08 上昇している．これらのことから，C1・C2 の印象クラスにおいて，「音楽に対するコメント」を除外すると印象推定に良い影響を与える可能性があることがわかる．また，フィルター有り無しにかかわらず，どちらも平均正解率が 0.8 近くあるため，音楽動画の印象推定において，映像に対す

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

るコメントを用いた手法は有効であるといえる．

5.6 音楽に対するコメントを用いた印象推定結果

5.2 節で生成した「音楽に対するコメント」を用いた印象推定（F あり）と，何も処理を加えていないソーシャルコメントを用いた印象推定（F なし）の比較を行う．それぞれの推定における正解率を以下の表 8 に示す．

表 8　音楽に対する分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
F あり	**0.721**	**0.723**	**0.833**	**0.759**	**0.708**	**0.880**	**0.758**	**0.830**	**0.776**
F なし	**0.735**	**0.723**	**0.800**	**0.759**	**0.702**	**0.851**	**0.750**	**0.822**	**0.768**

表 8 より，音楽に対するコメントを利用して印象推定を行った場合，C3（切ない）・C6（激しい）の推定精度は向上しているが，他の印象クラスに注目すると正解率の変化が少ないことがわかる．特に C2（元気が出る）・C4（激しい）の印象においては精度に変化が現れなかった．

以上の結果より，C3・C6 においては，音楽と映像の印象が乖離しているために，映像に対するコメントを除去すると精度が向上すると考えられる．一方，C2・C4 などの正解率に変化が少ないのは，これらの印象クラスは 「音楽に対するコメント」ではなく，「音楽・映像どちらに言及しているかわからないコメント」が，付与されたコメントの大部分を占めているからだと考えられる．

5.7 考察

5.5 節のように映像に対する印象推定の場合は，「音楽に対するコメント」をフィルタリングすることが有効であった．同じように音楽に対しても「映像に対するコメント」をフィルタリングすることで精度が向上するのではないかと考え，5.6 節のように実験を行ったが正解率はあまり向上しなかった．これは「音楽に対するコメント」自体の数が少なく，音楽に対する印象推定として重要な特徴となっているのが「音楽・映像両方を対象としているコメント」となっているためだと考えられる．そのため，映像に対する印象推定した際の正解率（0.835）より，音楽に対する印象推定した際の正解率（0.776）が低いのだと考えられる．ここで，音楽に対するコメントが少ないのは，音楽動画を視聴しコメントする際，音楽に対しコメントすることが，映像に対しコメントすることよりも難しいためと考えられる．

6. 映像特徴・コメントを組み合わせた推定

本章では 4 章の結果である，各印象クラスにおける最適な映像特徴と，5 章の「映像に対するコメント」から抽出した特徴量を組み合わせて，音楽動画の映像に対し印象推定を行った．その結果を表 9 に示す．

表 9　映像特徴・コメントを組み合わせた結果

	C1	C2	C3	C4	C5	C6	V	A	平均
映・コ	**0.802**	**0.880**	**0.744**	**0.895**	**0.741**	**0.881**	**0.732**	**0.786**	**0.808**
映像	**0.880**	**0.870**	**0.748**	**0.896**	**0.718**	**0.883**	**0.820**	**0.810**	**0.828**
コメント	**0.890**	**0.900**	**0.805**	**0.838**	**0.708**	**0.890**	**0.757**	**0.889**	**0.835**

映像特徴とコメント特徴を組み合わせた結果，C1~Arousal の平均が低下してしまった．しかし，C5（滑稽な）クラスに関しては，映像・コメントをそれぞれ単体で利用した場合と比べ，精度が向上している．このことから，単純に映像特徴とコメント特徴を組み合わせて印象推定を行うのではなく，C1 ではコメント特徴，C5 では映像・コメント特徴，Valence では映像特徴を用いるなど，各印象クラスにおいて最適な特徴を用いることで，音楽動画の映像に対する印象推定の精度を向上できると考えられる．また，それぞれの特徴による印象推定結果を，それぞれに対する投票とみなし，多数決などを取ることによって印象推定を行うことも考えられる．

なお，今回の実験では映像の色情報を用いる場合と，コメントから抽出された単語を用いる場合とでは次元が大きく異なる（色の場合は 3~9 次元，コメントは 30 次元）．その違いが結果に影響を及ぼしていると考えられるため，今後はそうした特徴の統合についても検討を行っていく予定である．

7. まとめ

本研究では音楽動画に対する印象推定の精度を向上させることを目的とし，映像特徴とソーシャルコメントを用いた印象推定に注目して手法の検討を行った．

実験結果より，映像特徴を用いた印象推定では，映像から抽出した画像を 4,095 通りの色の組み合わせで減色し，それらの画素数を特徴として学習することで，印象クラスごとに最適な色の組み合わせを利用することが推定精度の向上に有効だということを明らかにした．また，ソーシャルコメントから「音楽に対するもの」と「映像に対するもの」をそれぞれ分離し，それらを用いて音楽動画を構成する音楽・映像それぞれのメディアタイプに対する印象推定の実験を行った．その結果，「映像に対するコメント」を利用した映像の印象推定は有効であったが，「音楽に対するコメント」を利用した音楽の印象推定は，従来の手法と比べ，変化が少なかった．

以上の結果を踏まえ，映像特徴とコメント特徴を組み合わせて，音楽動画の映像に対して印象推定を行った．その結果，推定精度を向上させるためには，印象クラスごとに最適な特徴量を用いる必要があることが明らかになった．

Vol.2017-CG-168 No.40
Vol.2017-DCC-17 No.40
Vol.2017-CVIM-209 No.40
2017/11/9

今後の課題として，今回は注目しなかった音響特徴量を用いて，音楽動画の音楽に対する印象推定の精度を向上させることがあげられる．また，ソーシャルコメントから音楽・映像に関するものを抽出する際パターンマッチを用いるのではなく，Doc2Vec を利用しソーシャルコメントを 1 つ 1 つベクトル化することでより正確にフィルターをかけた場合，印象推定の精度にどのような違いが出るか検討したい．また，映像については動きの情報を組み合わせて精度向上を目指すことも考えられる．

謝辞 本研究の一部は，ニコニコ動画コメント等データを利用したものであり，JST ACCEL（グラント番号 JPMJAC1602）の支援を受けたものである．

参考文献

[1] 山本岳洋,中村聡史: 視聴者の同期コメントを用いた楽曲動画の印象分類,情報処理学会論文誌, Vol.6, No.3, pp.66-72 (2013).
[2] 土屋駿貴, 大野直紀, 中村聡史, 山本岳洋: ソーシャルコメントからの音楽動画印象推定手法の提案, DEIM Form 2016 E3-3 pp1-7 (2016).
[3] 西川直毅, 糸山克寿,藤原弘将, 後藤真孝, 尾形哲也, 奥乃博: 歌詞と音響特徴量を用いた楽曲印象軌跡推定法の設計と評価, 情報処理学会研究報告, Vol.2011-MUS-91, No.7, pp.1-8 (2011).
[4] 阿部和樹，土屋駿貴，大野直紀，中村聡史，山本岳洋：音楽動画に対するソーシャルコメントと音響・映像特徴量を用いた印象推定手法の検討，グループウェアとネットワークサービスワークショップ（GN Workshop 2016），pp.1-7 (2016)
[5] Hevner, K.: Experimental studies of the elements of expression in music, The American Journal of Psychology, Vol.48, No.2, pp.246-268 (1936).
[6] Russell, James A.: A Circumplex Model of Affect, Journal of Personality and Social Psychology, 39(6), pp.1161-1178 (1980).
[7] Esra Acar, Frank Hopfgartner, and Sahin Albayrak: Understanding Affective Content of Music Videos Through Learned Representations, MMM2014, Vol.8325, pp.303-314 (2014).
[8] 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, DEIM Form 2015 F2-3 pp1-6 (2015).
[9] 上坂俊輔: 主観評価の過程で注目する領域・特徴の推定と画像の自動分類・検索への応用，中央大学大学院研究年報理工学研究科篇，第 42 号 (2012).
[10] 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からの音楽動画の印象推定手法, DEIM Form 2016 E3-5 pp.1-8 (2016).
[11] 楽曲動画印象データセット配布サイト（最終閲覧日 2017 年 10 月 17 日）http://nkmr.io/mood/
[12] ニコ動コメントコーパスで kaomoji2vec して顔文字をベクトル表現で扱う（最終閲覧日 2017 年 10 月 17 日）https://qiita.com/ryo_grid/items/0c18513ae42778248a9f