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

阿部和樹[†1]　土屋駿貴[†1]　大野直紀[†1]　中村聡史[†1]　山本岳洋[†2]

概要：音楽動画の印象に基づいた検索を実現するため，動画共有サイトに投稿された音楽動画を可愛い，切ないなどの印象によって分類する手法を検討する．音楽動画に対してユーザが付与するソーシャルコメント，音楽の音響特徴量，動画の映像特徴量の 3 つを音楽動画の特徴量ととらえ，印象分類に利用する．また，その分類精度について実験を行い，ソーシャルコメントによる推定精度の高さと，すべての特徴量を用いた推定精度の可能性を明らかにした．

キーワード：印象推定，音楽情報検索，音響解析，映像解析，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が容易に制作できるソフトウェアやサービスの登場と，コンテンツを他者と共有可能な YouTube やニコニコ動画といった動画共有サービスの隆盛によって，インターネット上で視聴可能な音楽動画の数は飛躍的に増大している．その一方で，音楽動画を検索するための手法は限られており，現状ではアーティスト名や曲名といったテキスト情報，ユーザが付与したタグ情報から検索する方法が主である．そのため，音楽動画を検索する際には事前知識が必要なことも多く，ユーザが求める音楽動画を探し出すことは容易ではない．

こうした検索における問題を解決する方法の一つが，音楽動画に対する印象を利用するものである．印象を利用した検索とは，「楽しい気分になる音楽動画」や「悲しい雰囲気の音楽動画」といった検索を可能とするものであり，音楽動画の検索における選択肢が広がるとともに，ユーザにとって未知の音楽動画を目にする機会が増えることが期待される．

印象に基づく検索の方法として，視聴者などによって付与されるソーシャルタグを利用することが考えられるが，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 5%程度であり[1]，検索手段として十分ではない．

こうした音楽動画の印象を推定するため，ユーザが視聴中の音楽動画の場面に対して付与するソーシャルコメントを利用した印象推定に関する研究に取り組んできた[2]．音楽については，音響的な特徴を利用した印象推定に関する研究が多数なされている[3]が，音楽と映像が融合されている音楽動画コンテンツの印象推定に関する研究は十分になされておらず，どのように印象推定をしていったらよいかなど明らかになっていない．そこで本稿では，これまでに構築した 500 個の音楽動画データセット[4]を活かし，以下の二つの実験を行う．

① 映像の印象推定についての検討
② 物理的特徴（客観的特徴）と主観的特徴の融合による印象推定に関する検討

①については，清水らの研究手法[5]と，他の手法をもとに音楽動画における映像的な特徴を利用した印象推定について画像の切り出し方も含めて検討する．次に②については，映像特徴量，音響特徴量に代表される物理的な特徴（客観的な特徴）と，ソーシャルコメントに代表される主観的特徴の両方を用いた印象推定について，SVM を用いた手法により検討する．また，物理的特徴と主観的特徴をどのように組み合わせると，視聴者にとってもっともらしいものとなるか明らかにする．

2. 関連研究

音楽情報処理の分野では，ユーザの検索支援を目的として，楽曲の印象推定に関する研究が多数行われている．

楽曲の印象を表現する方法については，多数のアプローチが存在する．楽曲の印象のモデル化に関する最も古いものとしては，Hevner の研究[6]がある．Hevner は楽曲に対する印象を 8 クラスの印象群を用いて分類を行っている．また，Russel が提案する Valence-Arousal 空間[7]は，楽曲の印象推定で広く用いられている．Valence-Arousal 空間とは，Valence（快-不快を表す次元），Arousal（覚醒-鎮静を表す次元）という 2 つの軸で張られる空間上で印象を表現するという考えである．

また音楽動画に関して，音楽のみや映像のみといった各メディアから受ける印象の違いに関する研究がなされている．佐藤らの研究[8]では，音楽と映像を比較した場合，映像から受ける印象が強いという結果を示している．また，長谷川らの研究[9]では，静止画が与える印象と音楽が与える印象の類似は，ユーザの好みのジャンルに影響されるこ

†1 明治大学
Meiji University.
†2 京都大学
Kyoto University

とを明らかにしている．

音楽の印象推定に関しては，楽曲が持つ様々な特徴から楽曲の印象を推定する研究が多数存在する[10]．これらの研究の多くは楽曲の音響信号に基づく特徴量を利用した手法であるが，近年では楽曲の歌詞を利用した手法も提案されている[3][11]．また，音楽動画の印象推定に関しては，ユーザが音楽動画に付与するソーシャルコメントを利用する研究がある[2]．これらは、得られるコメントを分析し，用いられるコメントの品詞や出現頻度から特徴量を抽出することにより推定に利用している．

山本らの研究[1]では，先に述べた音響特徴量による推定とソーシャルコメントを利用した推定を組み合わせ，それぞれ単独で行う推定との比較を行っている．実験の結果，音響特徴量に基づいた推定よりも，コメントを利用した推定の精度が高く，また，それらを組み合わせることでより高い精度の印象推定ができる可能性を示唆している．

楽曲に限らず，視聴者が映像から得られる印象に関する研究も近年行われている[5]．清水らは画像から得られる色情報と映像における動きの情報から印象を推定し，映像との同期に最適な楽曲を，その印象に基づき選択している．

以上のように音楽動画の印象推定に関しては，コメントによる特徴量，音響特徴量，映像特徴量のそれぞれを利用する手法が存在する．しかし，山本らはコメント特徴量と音響特徴量を組み合わせた手法を用いたものの，映像特徴量とともに利用し，それらを組み合わせた印象推定はなされていない．本稿では，これらコメント特徴量，音響特徴量，映像特徴量のすべてを用いて印象推定を行い，単独による推定との比較を行うことで推定に適した手法を検討する．

3. 印象評価データセット

本稿では，音楽動画の印象を推定する際に，[4]において構築した印象評価データセット（http://nkmr.io/mood/）を利用する．この印象評価データセットは，音楽動画のサビ部分においてユーザが受ける印象を調査したものである．音楽動画を音楽, 映像, 音楽動画（音楽と映像の組み合わせ）の 3 つのメディアタイプに分け，それぞれを 8 つの印象クラスによって評価したものである．音楽動画は，ニコニコ動画にある「VOCALOID」タグの付与された動画のうち，2012 年 8 月時点で再生数の多い音楽動画の上位 500 件を対象としている．評価は, 少なくとも 3 名以上が行っており，結果として，500 件×3 メディアタイプ×8 印象クラスの 12000 件のデータからなる．

評価する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 で用いられている 5 つの印象クラスと，Russel が提案した Valence-Arousal 空間[7]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．つまり，一つの音楽動画は，8 つの印象クラスについてそれぞれ印象評価値を持っている．8 つの印象クラスをまとめたものを表 1 に示す．

評価値は, 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（穏やかな，消極的な，弱気な）～+2（激しい, 積極的な, 強気な）の 5 段階で評価されている．

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

4. 映像特徴量抽出手法の検討

我々は音楽動画の印象推定に用いる特徴量として，音楽動画に付与されるソーシャルコメント, 音楽の音響特徴量, 動画の映像特徴量に着目している．これらの特徴量について，コメントによる印象推定[2]と音響特徴量による印象推定[3]は様々な手法が提案されているが，映像特徴量のみで音楽動画の印象推定を行う研究は少ない．そのため，本研究では予備調査として，印象推定に有用な映像特徴量を検討する．清水らの研究[5]では，映像の色情報と物体の動きの情報を用いて映像の印象推定を行っている．これらの実験の結果，特に映像の色情報を用いた印象推定が有効であることを明らかにしている．そこで，本研究においても同様に映像の色情報を用いることとする．また，映像からの画像の抽出方法としてどの程度がよいかも明らかになっていないため，これについても検討する．

4.1 色情報の抽出手法

色情報を基にした映像特徴量の生成は，映像を画像に分解し，各画像に対して色情報を抽出する方法を用いる．画像に対する色情報の抽出手法として，以下の 2 つの手法を候補とする．

① 手法 1 は，特徴的な色のピクセル数を計算する方法である．画像を RGB 色空間において特徴的な色の数まで減色し，各色のピクセル数を数えることでカラーヒストグラムを計算したものである．映像特徴量は，その色の数だけのベクトルを持つこととなる．画像ごとに処理を行うため，各色の平均と分散の値を計算することができる．

② 手法 2 は，手法 1 と同様に減色処理を行った後，色領域の面積によって特徴ベクトルを分割する，Color Coherence Vector （CCV）の手法を用いる．手法 1 と同じく，各色の平均と分散の値を持つ．

4.2 各手法の評価実験

各手法による映像特徴量を用いることにより，音楽動画の印象分類が可能かを検討するため，印象評価データセットを用いた評価実験を行う．

実験対象となる音楽動画は，3 つのメディアタイプの中でも各印象クラスにおいて映像のみに対する評価値の高いもの（高評価群）と評価値の低いもの（低評価群）を利用する．

実験は，音楽動画の映像特徴量を機械学習し，高評価群の音楽動画をどの程度分類できるかにより推定精度を測る．分類器としてサポートベクタマシン（SVM）を使用し，交差検定（5-fold クロスバリデーション）を行う．

手法 1，手法 2 のそれぞれを用いた印象推定において，その分類精度を比較する．また，それぞれの手法で得られる色情報の平均のみを利用する場合と，平均および分散の値を利用する場合も比較する．なお実験においてはアンダーサンプリングを実施し，高評価群と低評価群の数をそろえた。

4.3 結果

表 2 は，手法ごとに映像特徴量を抽出し，8 つの印象クラスにおいて実験を行った際の正例の適合率を示したものである．表において，①は手法 1 を，②は手法 2 を利用したものである．また，avg は平均の値，std は分散の値を，avg+std は平均と分散の両方を用いたものであり，それぞれについて推定精度を示している．

結果より，CCV を利用した手法は適していないことがわかる．また，「①avg」の場合に映像特徴量による印象推定の精度が，全体的に最も高い結果となった．以上より，音楽動画の映像特徴量抽出には，手法 1 における平均の値のみを用いることが適しているといえる．

表 2　映像特徴量による印象推定の分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
①avg	**0.730**	**0.725**	**0.712**	**0.754**	**0.725**	**0.793**	**0.571**	**0.692**	**0.712**
①avg+std	**0.720**	**0.745**	**0.707**	**0.727**	**0.725**	**0.784**	**0.569**	**0.679**	**0.707**
②avg	**0.629**	**0.725**	**0.701**	**0.701**	**0.647**	**0.772**	**0.539**	**0.676**	**0.673**
②avg+std	**0.629**	**0.728**	**0.703**	**0.701**	**0.642**	**0.770**	**0.549**	**0.671**	**0.674**
平均	**0.677**	**0.730**	**0.705**	**0.720**	**0.684**	**0.779**	**0.557**	**0.679**	**0.742**

4.4 画像抽出の時間間隔

先述の調査では，映像から一定の時間間隔における画像を抽出し，各画像の特徴量を見ることで映像特徴量としている．ここで，画像を抽出する時間間隔が，映像の印象推定に及ぼす影響についての調査を行う．

先述の評価実験と同様に，各手法において映像から画像抽出する際の時間間隔を変化させ，それぞれにおける印象の推定精度を比較する．例えば時間間隔を 5 秒とした場合，映像の 5 秒ごとにおけるフレーム（画像）を取り出し，それらの画像によって生成される映像特徴量によって，印象推定を行う．

時間間隔は 0.1 秒，1 秒，5 秒を設定する．対象とする音楽動画の時間は 30 秒に統一されているため，それぞれ 300 枚，30 枚，6 枚の画像を推定材料にすることとなる．ただし，各手法において色情報の平均または分散を算出しているため，枚数によるデータの不均衡は起こらないものとする．

評価実験の結果を表 3 に示す．結果の値は，C1～Arousal までの全印象クラスにおける推定精度より，その平均を計算したものである．表より，いずれの手法においても時間間隔による推定精度の変化は少なく，我々が設定した時間間隔の範囲では，推定精度に大きな影響はないといえる．

以上の結果より，映像特徴量の生成において，①の手法を用いて，映像の 5 秒ごとに抽出した画像群より映像特徴量を生成する方法が，コスト面でも精度面でも印象推定に最も有効な手段であるといえる．

表 3　時間間隔ごとの推定精度

	0.1 秒	1 秒	5 秒	平均
①avg	0.702	0.709	0.713	0.708
①avg+std	0.709	0.711	0.707	0.709
②avg	0.673	0.668	0.674	0.672
②avg+std	0.676	0.670	0.674	0.673
平均	0.690	0.690	0.692	0.690

5. 音楽動画からの特徴量抽出

本研究では，音楽動画から得られる特徴を数値として抽

出し，多クラス分類器にかけることで音楽動画の印象を推定する．音楽動画の印象推定に有用な特徴として，

（1） 音楽動画に付与されたソーシャルコメント
（2） 音楽から得られる音響的特徴
（3） 動画から得られる映像的特徴

の 3 つを利用する．（1）は，音楽動画を視聴するユーザが付与する主観的特徴，（2）（3）はユーザの意思が関与しない客観的特徴（物理的特徴）といえる．特徴量の抽出については，データセットと同様にサビ区間に対してのみ行うものとする．

5.1 単語ベクトルの生成

ソーシャルコメントから各印象推定の精度を考察するため，印象評価データセットの評価対象となったすべての音楽動画に付与されたコメントを収集する．そのうち，印象評価データセットで用いられた音楽動画のサビ部分の時間を参考に，サビ区間内に投稿されたコメントを抽出する．単語ベクトルの生成については，土屋ら[2]と同じく MeCab を用いて，コメントから形容詞を単語として抽出し，各単語の出現頻度を数えたものを音楽動画に対する単語ベクトルとする．

5.2 音響特徴量の抽出

音楽そのものから得られる音響特徴量の抽出については，楽曲分析において広く用いられている MARSYAS[12] を利用する．MARSYAS を利用することにより，スペクトル特徴量やメル周波数ケプトラム係数などの音楽から得られる 31 次元の特徴を抽出し，音響特徴ベクトルとして扱う．得られる特徴を表 4 に示す．

表 4　得られる音響特徴量

特徴量	次元数
スペクトル特徴量	3
メル周波数ケプトラム係数（MFCC）	13
クロマベクトル	14
ゼロクロッシング	1
合計	31 次元

5.3 映像特徴量の抽出

映像特徴量の抽出については，4 章の結果をもとに 5 秒ごとの画像を抽出し，各画像に減色処理を施す．扱う色については，三原色である RGB それぞれを 3 階調とし，3 の 3 乗である 27 色に静止画を減色している．また各色の画素数の合計から一つの静止画における画素数の平均を求めることで，映像全体に対する平均の色の割合とみなし，27 次元の映像特徴ベクトルとして扱う．

6. 評価実験

ソーシャルコメントと音響特徴量，映像特徴量を用いることにより，音楽動画の印象がどの程度推定可能か検討するため，4 章で利用した印象評価データセットを用いた評価実験を行う．本実験の目的は，音楽動画の特徴抽出において，どのような手法が有効であるのかを明らかにすることである．

6.1 検討手法

印象評価データセットにおいて，3 つのメディアタイプに対して 8 つの印象クラスごとに評価値が 4 以上（Valence, Arousal は 1 以上）の高評価群と，2 以下の低評価群（Valence, Arousal は-1 以下）の動画のみを対象とした．また，データの偏りがないように，高評価群と低評価群の動画数は一致するように設定した．

評価値は，音楽のみ，映像のみ，音楽動画といったメディアタイプごとに異なるため，それぞれで対象とする音楽動画も変化する．表 5 に各メディア・印象タイプにおいて対象とした動画数を示す．

表 5　印象分類実験に使用した動画数

	音楽	映像	音楽動画
C1（堂々とした）	108	42	150
C2（元気が出る）	164	100	206
C3（切ない）	90	280	172
C4（激しい）	138	96	108
C5（滑稽な）	96	158	162
C6（かわいい）	142	154	200
Valence	120	112	122
Arousal	82	216	186

音楽動画から得られる特徴量について，付与されたコメントによるコメント特徴量，音楽の音響特徴量，動画の映像特徴量，およびそれらの組み合わせを用いることによる分類精度の評価を行う．各メディアタイプについて，音楽のみについては音響特徴量とコメント特徴量，映像のみについては映像特徴量とコメント特徴量，音楽動画についてはすべての特徴量を利用して推定を行った．

6.2 実験設定

実験は，5 章で得られた特徴量を用いて機械学習を行い，高評価群に該当するものをどの程度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，低評価群を負例としてそれぞれを 5 分割し，そのうち 4 つを SVM の訓練データ，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，正例の適合率を計算する．

6.3 結果

以下の表 6〜8 は，メディアタイプごとに特徴量を抽出し，8 つの印象クラスにおいて実験を行った時の正例に関する適合率を示したものである．また，各印象クラスと使

用した特徴量ごとの適合率の平均値も表に示す．表において，comment はコメント特徴量，audio は音響特徴量，visual は映像特徴量を意味している．また，c+a は comment と audio の組み合わせ，c+v は comment と visual の組み合わせ，c+a+v は comment と audio と visual の組み合わせを意味している．さらに，V は Valence を，A は Arousal を意味している．

表 6　音楽のみに対する分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
comment	**0.660**	**0.710**	**0.804**	**0.677**	**0.681**	**0.819**	**0.649**	**0.660**	**0.708**
audio	**0.849**	**0.658**	**0.687**	**0.802**	**0.625**	**0.769**	**0.707**	**0.849**	**0.716**
c+a	**0.816**	**0.702**	**0.725**	**0.808**	**0.571**	**0.900**	**0.711**	**0.644**	**0.735**
平均	**0.775**	**0.690**	**0.738**	**0.762**	**0.626**	**0.829**	**0.689**	**0.647**	**0.720**

表 7　映像のみに対する分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
comment	**0.857**	**0.795**	**0.755**	**0.734**	**0.547**	**0.845**	**0.517**	**0.677**	**0.713**
visual	**0.730**	**0.725**	**0.712**	**0.754**	**0.725**	**0.793**	**0.571**	**0.692**	**0.716**
c+v	**0.900**	**0.826**	**0.744**	**0.744**	**0.565**	**0.902**	**0.540**	**0.711**	**0.742**
平均	**0.829**	**0.782**	**0.737**	**0.744**	**0.613**	**0.846**	**0.543**	**0.694**	**0.723**

表 8　音楽動画に対する分類精度

	C1	C2	C3	C4	C5	C6	V	A	平均
comment	**0.732**	**0.885**	**0.813**	**0.914**	**0.675**	**0.920**	**0.557**	**0.865**	**0.795**
audio	**0.777**	**0.623**	**0.619**	**0.767**	**0.550**	**0.711**	**0.694**	**0.666**	**0.676**
visual	**0.639**	**0.761**	**0.746**	**0.759**	**0.483**	**0.801**	**0.615**	**0.706**	**0.689**
c+a+v	**0.842**	**0.854**	**0.813**	**0.886**	**0.634**	**0.883**	**0.750**	**0.818**	**0.810**
平均	**0.748**	**0.781**	**0.748**	**0.832**	**0.586**	**0.829**	**0.654**	**0.764**	**0.742**

表 6 の音楽のみを対象とした実験の結果，C3（切ない）と C6（かわいい）は comment による推定精度が高く，C1（堂々とした）と C4（激しい）は audio による推定精度が高かった．また，comment と audio を組み合わせた手法は，C6 が 0.9 と最も高い推定精度であった．

次に映像のみを対象とした実験の結果，visual を用いた推定精度で 0.8 以上のカテゴリは存在しなかった．comment による推定は C1 と C6 の精度が 0.8 を超える値となった．comment と visual を組み合わせた推定では，C1 と C6 の精度が 0.9 以上となり，C2（元気が出る）も 0.8 以上の値となった．すべての手法において Valence（快-不快），Arousal（覚醒-鎮静）については，ほとんどが 0.7 以下と低い推定結果となっていた．

音楽動画を対象とした実験の結果，comment による推定は C2（元気が出る），C3，C4，C6，Arousal において精度が高く，特に C4 と C6 は 0.9 以上の推定の結果となった．audio による推定精度は 0.8 以下の精度ばかりであり，visual による推定は C6 のみ 0.8 以上の精度となった．comment, audio, visual のすべての特徴量を用いた推定では，C1, C2, C3，C4，C6，Arousal の推定が 0.7 以上となり，平均の推定も 0.8 以上と最も高い値となった．

6.4 ソーシャルコメントの分析

comment または comment と audio や visual を組み合わせた手法による推定精度が高いという結果より，音楽動画に用いられるソーシャルコメントについて分析する．

各印象における特徴的な単語を明らかにするため，推定実験に用いた単語（形容詞）の TF-IDF 値を計算する．ここで，8 つの印象クラスにおける高評価群，低評価群をそれぞれ 1 つのカテゴリとし，16 カテゴリ内の特徴語を算出する．ここで TF 値は，一つのカテゴリ内における全コメントから抽出された各単語の出現頻度である．DF 値は，ある単語がどの程度の数のカテゴリに含まれるかを数えるものとする．また，IDF 値は DF 値の逆数の対数である．なお，コメント数が多いため，TF および DF については，その単語が 20 回以上あらわれていないと出現していないものとして扱った．

表 9　TF-IDF 値が高い特徴的な単語

	高評価群	低評価群
C1	かわいいい，可愛，かわゆ， かわいいー，かわいー， かわゆい，よ～，かわいい～， かわい，でかい	怖い，多い，すごく
C2	かわいいい，可愛，かわゆ， かわいいー，かわいー，弱い， かわゆい，なう，かわい， かわいい～	こわい，美しい，怖い， かっこよ，こわ，良，すご， 多い，高い，やすい
C3	悪い，上手い，こいい，怖い， うまい，こわい，やすい， 欲しい，重い，凄い	かわいいい，可愛，おかしい， かわゆ，かわいいー， かわいー，かわゆい， かわいい～, かわい, 懐かしい
C4	早く，カッコイイ，こい， 怖い，凄い	かわいいい，ヤバイ，かわゆ， なう，かわいいー，かわいー， すごく
C5	おかしい，こわい，こわ， かわいいい，怖い，かわゆ， うまい，懐かしい， かわいいー，欲しい	かる，かわいいい，なう， やすい，よかっ，重い，イイ， すごく，いい，かっこいい
C6	かわいいい，可愛，切ない， かわゆ，かわいー，かわいい ー, 弱い, かわゆい, 可愛いー， なう	こわい，早く，怖い， カッコイイ，かっこよ，こわ， 良，多い，やすい，よかっ
Valence	早く，かっこよ，怖い，重い， 凄い	かわいいい，すごく
Arousal	かわいいい，可愛，かわゆ， かわいいー，良かっ， かわいー，かわゆい，おおー， よ～，かわい	かる，良，かっこよ， 素晴らしい，多い，怖い， うまい，よかっ，かわいいい， 重い

コメントに含まれる全ての単語（形容詞）について，各単語の TF-IDF 値を比較する．また，3 つのメディアタイプ（音楽のみ，映像のみ，音楽動画）によって実験に用いた単語の総数が異なるが，代表として音楽動画の推定に用いられた単語のみを対象とした．

分析において TF-IDF 値の高い上位 10 個の単語を表 9 に示す．印象クラスごとに比較した場合，それぞれの印象で異なる単語が選出されているもことがわかる．一方,「かわいい」などの複数の印象に共通して見られる単語も存在した．

7. 考察

実験により，comment，audio，visual の組み合わせによって各印象推定が変化することが分かった．

単独で推定した場合では，C6（可愛い）を comment によって推定した場合の精度が全体的に高く，C6 は comment のみを見ることによって推定が可能だと考える．また，C1（堂々とした）は comment よりも，audio または visual によって推定した場合の値が高く，いずれも comment と組み合わせることによって 0.8 以上の推定を出す結果となっている．よって，C 1 は客観的（物理的）特徴による印象推定が有効だといえる．

一方，C5（滑稽な）と Valence の推定は全ての手法において 0.8 以下の結果となった．これらの印象クラスは，ユーザに与える印象が薄いものであるか，我々の手法によっては十分に分類できない印象であると考えられる．

メディアタイプを音楽動画に注目した場合では，全体的に comment による推定が高く，印象推定では音楽動画に付与されるソーシャルコメントが大きく関係しているという結果となった．しかし，すべての特徴を用いた推定の平均は 0.81 と最も高い．これにより，comment，audio，visual のすべてを用いた推定は，さまざまな印象に対してある程度の精度の高い推定ができると期待される．

また，推定精度の高い comment（コメント特徴量）を分析した結果，それぞれの印象に沿った単語が多く選出された．こうした単語が高評価を導き出していると考えられる．また，高評価群と低評価群における特徴的なコメントの違いより，評価が高い音楽動画と低い音楽動画のコメントには大きな違いがあることもわかる．一方,「かわいい」のように多くのカテゴリに登場する単語も存在した．これは，単純に登場人物（例えば初音ミクなど）に対してかわいいと表現されているものがあり，結果的に評価に使いにくい単語になっているといえる．

8. まとめ

本稿ではコメント，音楽，映像の各特徴量を利用し，それぞれを単独で利用する場合と組み合わせた場合の推定精度の比較を行った．

映像特徴量によって印象推定を行った結果，映像から抽出した画像を特徴的な色に減色し，それらのピクセル数を数えることで，推定に有効な特徴量が取得可能なことを明らかにした．

コメント，音楽，映像の３つの特徴量による推定を比較した結果，コメントのみによる C4（激しい），C6（可愛い）といった一部の印象に対する推定精度が最も高いという結果となった．また，3 つの特徴量すべてを用いた推定の平均精度は高く，様々な印象に対して汎用的に活用できる可能性を示した．

今後の課題として，音響特徴量，映像特徴量の次元数の変化による推定精度の比較を行い，推定に適した特徴量生成を検討する必要がある．また実験において，印象クラスによってはコメント特徴量による推定精度が高いものもあれば，音響特徴量や映像特徴量による精度が高いものも存在した．そこで今後は，3 つの特徴量を利用する際に，印象によってはコメント特徴量に重み付けを行うなど，重要となる特徴量を設定することで推定精度が上がる可能性が考えられる．また今後は特徴量を組み合わせる方法をさらに詳細に検討することにより，推定精度の向上を図る．

謝辞

本研究の一部は，JST CREST，JST ACCEL 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] 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).
[5] 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).
[6] Hevner, K.: Experimental studies of the elements of expression in music, The American Journal of Psychology, Vol.48, No.2, pp.246-268 (1936).
[7] Russell, James A.: A Circumplex Model of Affect, Journal of Personality and Social Psychology, 39(6), pp.1161-1178 (1980).
[8] 佐藤淳也, 佐川雄二, 杉江昇: 音と映像の組み合わせによる主観的印象の変化,映像情報メディア学会誌,Vol.55, No7, pp.1053-1057 (2001).
[9] 長谷川優, 武田昌一: 好みの音楽ジャンルに着目した静止画と音楽の組み合わせに関する考察: − 個人の属性に着目した静止画と音楽に対する印象度の相互比較−, 日本感性工学会論文誌, Vol.11, No.3,pp.435-442 (2012).

[10] 絵本詩織, 糸山克寿, 奥乃博: 音響特徴量を用いた楽曲印象分布の推定,情報処理学会 76 回全国大会, pp.391-392 (2014).
[11] 舟澤慎太郎, 北市健太郎, 甲藤二郎: 楽曲推薦システムのための楽曲波形と歌詞情報を考慮した類似楽曲検索に関する一検討, 情報処理学会研究報告オーディオオビジュアル複合情報処理, pp.1-5 (2013)
[12] Tzanetakis, G. and Cook, P.: MARSYAS: A frameworkfor audio analysis, Organised sound, Vol.4, No.3, pp.169– 175 (1999).