情報処理学会論文誌 Vol.59 No.3 929–940 (Mar. 2018)

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

大野 直紀[1,a)] 土屋 駿貴[1] 中村 聡史[1] 山本 岳洋[2]

受付日 2017年6月11日, 採録日 2017年12月8日

概要：音楽動画の印象に基づく検索や推薦，音楽動画の類似判定のためには，音楽動画の印象推定に関する技術が必須となる．しかし，音楽に対する印象評価や映像に対する印象評価に関する研究は多数なされている一方で，音楽と映像が組み合わされた音楽動画に対する印象評価の研究は十分になされていない．我々は，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を調べるため，「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た．また，それらを分析することで，音楽と映像の印象評価の組合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．またデータセット内の音楽動画の音楽と映像を任意に合成した音楽動画を生成し，印象評価を行ってもらうことで，音楽印象と映像印象の組合せが音楽動画の印象とどのように関係しているのかの分析を行った．その結果，音楽と映像の印象を組み合わせることによる印象推定の可能性があること，また各印象によって印象の組合せ方が異なることを明らかにした．

キーワード：音楽動画，印象推定，音楽，映像，可視化

A Study of the Relationships between Music-impression, Visual-impression and Music Video Clip's Impression

Naoki Ono[1,a)] Shunki Tsuchiya[1] Satoshi Nakamura[1] Takehiro Yamamoto[2]

Received: June 11, 2017, Accepted: December 8, 2017

Abstract: It is necessary to estimate impression of video clips in order to realize impression based video search and video recommendation system. There are many researches to estimate music impression and visual impression by analyzing them. However, there is no good method to merge music impression and visual impression. In this paper, we generate two types of dataset. One dataset consists of music video clip with evaluation score in each media type and each impression type. The other dataset consists of evaluation score of music video clip by synthesizing movie in each impression type. We use these dataset in order to clarify the relationship of music impression, visual impression and music video clip's impression.

Keywords: music video clip, music, movie, estimating impression, visualization

1. はじめに

コンテンツ制作支援システムの普及や発展により，誰でも楽曲や動画を創作することが容易になった．また，大規模動画共有サイトの普及により，多くのアマチュア作曲者や動画制作者が創作したコンテンツを発表する場ができ，他者が容易に閲覧することが可能となった．これにより，人々が接することのできる音楽動画は非常に増加したといえる．なお，本研究では音楽が主としてありながらも，その音楽と時間的に同期して映像が提示されるものを「音楽動画」と呼ぶ．

音楽動画の増加に関して顕著な例が，初音ミクをはじめとする VOCALOID を使用したものであり，日本最大の動画

[1] 明治大学院先端数理科学研究科
Meiji University, Nakano, Tokyo 164–8525, Japan
[2] 京都大学院情報学研究科
Kyoto University, Kyoto 606–8501, Japan
[a)] kas.naoki.0212@gmail.com

共有サイトであるニコニコ動画[*1]においてはVOCALOIDを用いた音楽動画が数多く存在している（2016年1月1日時点で約39万件）．また，アマチュア作曲者および動画作成者が投稿した音楽動画のうち，100万回以上再生されているものも近年では単位月あたりで減少しているものの，多数存在している．これはニコニコ動画という動画投稿および共有基盤の存在と，VOCALOIDなどのコンテンツ制作支援システムによってコンテンツの制作が容易になったことの影響が大きいが，それに加え「歌ってみた」「踊ってみた」などに代表される二次創作が広まったことも大きく寄与していると考えられる．

人々がアクセスできる音楽動画の数が増加した一方で，音楽動画を探すための検索手段は多様ではない．たとえば，ニコニコ動画では音楽動画名やアーティスト名，タグといったテキスト情報に対するキーワード検索や，再生数や投稿日による動画のソートなどの方法でしか検索を行うことができない．

ユーザが音楽動画を検索する際に作者名やタイトルといったキーワードを思いつくことができない場合や，ユーザが求める気分にあった未知の音楽動画を検索する場合において，「泣きたい気分なので悲しい音楽を探したい」や「気分を昂ぶらせたいのでかっこいい音楽を探したい」といったようにユーザは求める音楽動画を雰囲気や印象といった曖昧な情報で表現することになる．しかし，「悲しい音楽」「かっこいい音楽」とキーワードを絞ったとしても，音楽動画の説明文自体に雰囲気や印象などの情報がテキストとして含まれていることは少なく，テキストマッチでの検索は難しい．また，こうしたサービス上では，ユーザが印象タグを付与して他者の検索に役立てることが可能であるが，タグがコンテンツに付与される割合はニコニコ動画では5% [9] であり，また，音楽の投稿，共有に関して利用されるSoundCloud[*2]であっても，タグは投稿者のみが付与できるものであり，そのほとんどがジャンルに関するものとなっているため，現状では検索に利用するには不十分であると考えられる．

また，近年音楽コンテンツの主流プラットフォームになりつつある定額制音楽配信サービスには，Spotify[*3]やLINE MUSIC[*4]，Google Play Music[*5]，Amazon Prime Music[*6]，Apple Music[*7]など多くの種類が存在するが，これらのサービスではユーザが単一の楽曲に対してタグを付与することができない．また，これらのサービスでは，「テンションの上がる洋楽集」といったようにあるユーザの印象に即したプレイリストを作成，共有することで，印象に基づいた音楽群を取得することが可能であるが，これはユーザ1人の主観的な印象に左右されるため，プレイリスト名とプレイリスト内の楽曲が一致しない，また単一の楽曲に対しては使用できないという問題があり．現状では印象や雰囲気で検索を行うことは難しい．

こうした問題を解決するため，音楽情報検索の分野では，楽曲の聴取を通してユーザが受ける主観的な印象を推定する研究が多数行われている [6], [9], [13]．ここで，主観的な印象に基づく検索とは，ユーザが楽曲を聴いて受ける印象に合うように，「人気のある切ない音楽」や「元気の出る印象を受ける動画」といった，主観的な印象語をクエリに含んだ楽曲の検索を可能とする手法のことである．

このような主観的な印象に基づく検索が可能となれば，VOCALOIDを用いた音楽動画のような，比較的新しく，ユーザ自身が好むアーティストやジャンルがまだはっきりとしていないドメインにおける音楽動画を探しているユーザへの検索手段となる．また，既存のドメインにおいても，推薦手法の1つとして雰囲気が類似している音楽動画を推薦することが可能になり，アーティスト，ジャンルに縛られない動画推薦が可能になる．さらに，これまでにない新しい観点からの検索手段を提供することができ，ユーザが未知かつユーザの求めている雰囲気に合った音楽動画を検索することができる．

本研究では，主観的な印象に基づく音楽動画の検索を実現するため，音楽動画を対象とし，音楽動画に対する主観的印象と，その音楽動画の音楽のみや映像のみに対する主観的印象との関係の分析に取り組む．音楽自体に対する印象 [6] や画像に対する印象推定に関する研究 [14]，また動画全般に対する印象の推定に関する研究は [11] あるものの，音楽動画に対する印象推定を行っている研究は少ない．山本らは楽曲動画に付与された視聴者のコメントから楽曲動画の印象を推定する手法を提案している [9], [12] が，この手法を用いるためには音楽動画に十分な量の視聴者のコメントが付与されている必要がある．また，音響特徴量や映像の特徴量を用いて音楽動画の印象を推定する研究も存在する [15], [16] が，音楽の印象と映像の印象の関係性に着目して詳細な分析を行っている研究はない．

音楽動画に対する印象と，それを構成するメディア（音声，映像）に対する印象の関係を明らかにすることができれば，音楽や映像に対する既存の印象推定技術を用いて，音楽動画の印象を推定することが可能になると考えられる．また，これらの関係を明らかにすることは，音楽動画を創作するユーザにとっても，視聴者にある印象を持ってもらうために音楽や映像をどのように作成するべきかの判断に役立つと考えられる．

音楽や映像といったメディアを融合した際にどのように印象評価が変化するかという点に関する研究としては，静

[*1] http://www.nicovideo.jp/
[*2] https://soundcloud.com/
[*3] https://www.spotify.com/
[*4] https://music.line.me/
[*5] https://play.google.com/store/music
[*6] https://www.amazon.co.jp/gp/dmusic/promotions/AmazonMusicUnlimited
[*7] https://www.apple.com/jp/apple-music/

止画と音楽の組合せによる印象の変化などが検証されている [4]．また，映像と音楽の組合せでは視覚刺激による影響が大きいことなどが明らかにされている [5]．さらに，音楽動画の印象推定に関する研究としては，音楽動画に付与されたソーシャルコメントを用いて印象を推定する研究 [9], [12] などが存在するが，音楽動画における，音楽や映像といったメディアに対する印象と音楽動画そのものに対する印象との関係を分析した研究は著者らの知る限り存在しない．

一方，我々はこれまでの研究において，ニコニコ動画上の音楽動画 500 件の音楽動画全体（動画の最初から最後まで）に対する印象評価データセットの作成を行った [1]．しかし，このデータセットは音楽動画の最初から最後までに対する評価を行っているものであり，音楽と映像がセットになったものを評価してもらっているため，印象評価のスコアが音楽動画のどの部分のどのメディアに対する印象を意味するものなのかを明らかにはできていなかった．さらに，音楽動画の各メディアの印象がどのように組み合わさって音楽動画の印象になっているかの詳しい解明には至っていなかった．そこで，まず音楽動画の一部分に限定して評価を行ってもらうことにより，音楽動画のメディアごとの印象の組み合わさり方が明確に分析できると考えた．

そこで本研究では，先述のデータセットで対象とした 500 件の音楽動画について，サビ部分の 30 秒間を「音楽のみ」「映像のみ」「音楽動画」の 3 種類に分け印象評価を行ってもらう．そうして得られたデータをもとに分析を行い，音楽，映像から受ける印象の組合せによる音楽動画の印象推定の可能性を検討する．

また，音楽の印象と映像の印象の組み合わさり方についての調査として，先述の研究 [1] で構築したデータセットで対象とした 500 件の音楽動画のサビ部分の「音楽のみ」「映像のみ」をランダムに組み合わせ，音楽動画を自動生成する．こうして生成された音楽動画に対しユーザに印象評価を行ってもらうことで，「音楽のみ」と「映像のみ」の印象評価が組み合わさることにより音楽動画から受ける印象評価がどのように変化するのかということを分析可能とする．

本研究による貢献は下記のとおりである．

- 500 件の音楽動画について，音楽のみ，映像のみ，音楽と映像の組合せという 3 種のメディアタイプのサビ部分印象評価データセットを構築した．
- 音楽動画全体と音楽動画のサビ部分では印象評価が大きく異なり，また，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できる可能性があることを明らかにした．
- 音楽，映像から受ける印象と音楽動画から受ける印象の関わり方を可視化し，音楽動画はすべての印象において音楽，映像のどちらからも影響を受けているが，印象よって各メディアからの影響のされやすさが違うこと，また音楽と映像を任意に組み合わせると音楽からの影響を受けやすい印象が多くなることを明らかにした．

2. 関連研究

音楽情報処理の分野では，ユーザの検索を支援するために，楽曲の印象の推定や印象にまつわる楽曲検索に関する研究が多数行われている．

2.1 楽曲の印象モデル

楽曲の印象の表現方法については，様々なアプローチが提案されている．MIREX [*8] では，印象を表す形容詞をクラスタリングすることで，印象を 5 つのクラスに分割し，印象推定のタスクに用いている．また，楽曲のみを対象としたものではないが，楽曲の印象推定にも広く用いられるモデルとして，Russel が提案した Valence-Arousal 空間がある [7]．Valence は快–不快を表す次元，Arousal は覚醒–鎮静を表す次元であり，印象をこの 2 つの軸で表現するという考え方である．

これらの研究のほかにも，印象による検索を行うため，ユーザの検索ニーズに合わせた印象語を選定する手法なども行われている [10]．このように印象語を選定し，その結果を用いて様々なメディアの印象推定が行われている．

2.2 楽曲の印象推定

楽曲の印象推定に関する研究は，音楽情報検索の分野において，近年特に取り組まれている．それらの研究では，音響特徴量をベースとした印象の推定が数多くなされている．また，近年では音響特徴量に加え楽曲の歌詞情報を利用した印象推定手法の提案 [8] もなされている．一方，楽曲の音響的特徴に依らない印象推定手法として，楽曲に付与されたタグやコメントによる印象推定 [9] も行われている．

このように，楽曲の印象を推定する手法はいくつか提案されているものの，それを音楽動画の印象の推定に使用した研究はない．本研究は，音楽の印象推定技術と音楽動画の印象推定技術の橋渡しになると考えられる．

2.3 メディア間の印象の差異

音楽動画をはじめとするマルチメディア情報での印象に関する研究として，各メディアから受ける印象の違いに関するものがある．佐藤らの研究 [5] では，音楽と静止画では音楽が，音楽と映像では映像から受ける印象が強いことが分かっている．また，長谷川らは静止画と音楽の印象の類似はユーザの好みのジャンルに影響されることを明らかにしている [4]．しかし，どちらの研究も音楽動画のコン

[*8] http://www.music-ir.org/mirex/wiki/MIREX_HOME

テンツを扱っていない，比較するコンテンツの件数が少ないという問題があり，音楽動画の印象の関係性の解明にはなっていないと考えられる．

また，音楽と映像の印象の変化に対するものとしては，書籍「音楽と映像のマルチモーダル・コミュニケーション」[13] がある．ここでは，音楽と映像を単独で提示した際と，音楽と映像を同時に提示した際の印象の変化を可視化している．しかし，実験に使用したサンプルの数が 20 種類と少ない点，音楽動画に特化したものではない点，音楽と音楽動画の関係性についての分析はなされているものの，映像と音楽動画の影響に関する分析がなされていない点で本研究とは異なっている．

2.4 音楽動画の印象推定

また，音楽動画の印象推定に関する研究としては，ニコニコ動画で付与されるソーシャルコメントを用いた印象推定を行った研究がある [9], [12]．また，音響特徴量や，映像の特徴量を用いて動画への印象の付与を行っている研究も存在する [15], [16]．また，音楽動画ではない，一般的な動画に対する印象推定を発話の特徴や映像の特徴などを用いて印象を推定している研究も存在する [11]．これらは音楽動画に対するものではなく，一般的な映像に対するものであるため，本研究とは立ち位置が異なるものの，このように，印象に基づく音楽動画の検索の実現に向けて，様々な研究が行われている．

3. データセット構築

音楽と映像から受ける印象の組合せによる音楽動画の印象推定の可能性を明らかにするため，ある音楽動画コンテンツと，ある音楽動画を音楽，映像の 2 つのメディアに分離することで得られる音楽のみのコンテンツ，映像のみのコンテンツの 3 種類に対する印象評価データセット[*9]を構築する．

評価対象は，ある音楽動画のサビ部分と，音楽動画のサビ部分を単体の音楽と映像に分離したもの 2 種類の計 3 種類を評価対象とした．

ここで評価対象とする音楽のみのコンテンツ，映像のみのコンテンツは，文献 [1] で用いられた 500 件の音楽動画（動画共有サイト「ニコニコ動画」上に投稿された音楽動画のうち，タグ「VOCALOID」が付与された動画の 2012 年 8 月時点で再生数が多い動画上位 500 件）をサビ開始から 30 秒の部分を抽出し，それらを音楽のみ，映像のみに分離したものである．

また，音楽と映像の組合せによって，音楽動画の印象はどのように変化するのかを明らかにするため，音楽動画のサビ部分を音楽のみ，映像のみに分離し，ランダムに組み合わせることで作成した 250,000 件の音楽動画のうち，ランダムに抽出した 500 件の音楽動画も評価対象とする．本研究では，ここで評価される音楽動画を「合成音楽動画」とする．合成音楽動画について，テンポの修正などの処理は行わなかった．

なお，本研究では「音楽のみ」「映像のみ」「音楽動画」に対する印象データセットを「印象評価データセット」，「合成音楽動画」に対する印象データセットを「印象変化データセット」とする．

以降，サビ区間の検出方法，評価対象とする印象軸，印象評価のインタフェース，印象評価の手続き，作成されたデータセットについての基礎検討について述べる．

3.1 サビ区間の検出

本稿では，音楽動画のサビ区間を評価対象として印象評価を行う．これは，多くの楽曲においてサビが盛り上がる部分であるため，今回はサビ区間を対象とした．

しかし，今回評価対象とした音楽動画が投稿されている大規模動画投稿サイト「ニコニコ動画」には，どこからがサビの区間なのかといった情報が付与されているわけではない．そのため，500 曲のサビ区間を検出する必要がある．そこで本稿では，後藤のサビ区間検出手法 RefraiD [2] を用いる．

この RefraiD は，楽曲中の様々な繰り返し区間をグルーピングすることで繰り返し区間の集合を求め，それぞれの集合ごとに「サビらしさ」を評価し，最終的に「サビらしさ」が高い集合をサビ区間として選択する．RefraiD では，このサビ区間として検出された区間集合中の区間に対して，それがどれくらいほかの区間と似ているかという値を算出できるので，本稿ではそれを各サビ区間の信頼度スコアとみなす．

そして，このスコアを用いて，サビ区間集合の中でもサビらしい区間を求める．そのうえでそのサビらしい区間の開始場所の 5 秒前から 30 秒間を評価対象として抽出した．

ここで，サビの開始場所が音楽動画の開始時間 5 秒未満と推定された場合は，音楽動画の開始から 30 秒間を抽出した．なお，ここでサビとして検出されたタイミングの 5 秒前からを抽出対象とした理由は，サビに入る少し前の部分からサビへの変化も重要であると考えたためである．

また，今回対象とした音楽動画はサビが 25 秒以下のものがほとんどであり，また著者らで 500 件の音楽動画を実際に視聴し確認を行ったが，サビが誤検出された事例はなかった．

こうして得られた音楽のサビ部分に該当する音楽動画と音楽のサビ部分に該当する映像をそれぞれ取得し評価対象とした．

[*9] http://nkmr.io/mood/

表 1 8 つの印象軸
Table 1 Eight impression axis.

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 気がめいる，哀愁の
C4（激しい）	アグレッシブな, 激しい, 興奮させる 感情的な，感情あらわな
C5（滑稽）	滑稽な, ユーモラスな, おもしろげな 奇抜な，気まぐれ，いたずらっぽい
C6（かわいい）	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

3.2 印象軸

本研究では，我々の過去の研究 [1] と同様に，音楽動画に対する印象として，音楽情報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 らの Valence-Arousal 空間を参考にした．ここで，MIREX では，5 つの印象クラスが用いられているが，これまでの研究 [1] により，ニコニコ動画上では「かわいい」と感じる楽曲やそれに関するタグが多く存在することが分かっているため，本研究でもこれまでの研究の評価にのっとり，MIREX の 5 クラスに加え，可愛らしさを表す印象クラスを加えた 6 軸と，Valence-Arousal に関する 2 軸の合計 8 軸を評価の収集対象とした．ここで，本研究は音楽と映像の関わり方，また音楽と映像の組合せによる音楽動画の印象の推定の可能性を示すための研究であるため，各印象軸についての独立は考慮しない．本研究で用いた 8 つの印象軸は，**表 1** に示すとおりである．表中の「印象クラス名」は，著者らが便宜上付与した，印象を表すラベル名である．また，「印象を表す形容詞」は，データセット構築において評価者から評価値を収集する際に，その印象クラスを表現するために用いた形容詞を表す．C6 については，「かわいい」の類義語を集めた．また Valence-Arousal についても，既存研究を参考に著者らが日本語に直したものを用いた．

3.3 印象評価インタフェース

図 1 に評価データ収集に用いたウェブインタフェースを示す．

図にあるように，評価者は各コンテンツを視聴し，その音楽動画に対する印象を，以下に示す形で付与する．

図 1 評価用ウェブインタフェース
Fig. 1 Web interface for evaluation.

- **C1-C6 の印象クラス：**表 1 に示した形容詞，形容動詞群に対する 1（まったくそう思わない）〜5（とてもそう思う）の 5 段階のリッカート尺度
- **Valence：**−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）の 5 段階のリッカート尺度
- **Arousal：**−2（穏やか，消極的な，弱気な）〜+2（激しい，積極的な，強気な）の 5 段階のリッカート尺度

なお，音楽と映像をランダムに組み合わせたものに対する印象変化データセットの構築では，「違和感を覚える」という項目も用意し，「感じる」「感じない」の 2 値でユーザに評価してもらった．

3.4 データセット構築

2015 年 3 月 26 日から 2016 年 6 月 18 日にかけて，対象動画の印象に対する評価データを収集した．印象評価データセット構築の協力者は大学生と著者を含む 21 人，印象変化データセット構築の協力者は大学生計 4 人であった．ここで，協力者の人数差，また評価者の個人差による影響などが考えられるが，500 件という多くの件数を用いて分析を行うため，傾向は間違ったものにはならないと考え，今回は考慮しないこととする．

データセット構築者には，評価対象であるコンテンツを視聴し，3.3 節で述べた印象評価用のウェブインタフェースを用い，対象コンテンツに対する印象を評価してもらった．これにより，500 件の音楽動画のサビ部分の「音楽のみ」「映像のみ」「音楽動画」，また 500 件の「合成音楽動画」に対して，最低 3 人以上の評価を収集し，それを平均したものを評価値とした．このとき，1 人の協力者が同一音楽動画の同一メディアの評価は行わないようにした．

結果として，文献 [1] より 500 件の「音楽動画全体」（音楽動画の最初から最後まで），音楽動画のサビ部分に対して，「音楽のみ」「映像のみ」「音楽動画」，500 件の「合成音楽動画」に対する印象評価値がデータとして存在するこ

表 2 印象評価値 0.5 以上のメディアの件数
Table 2 Number of media with impression evaluation value 0.5 or more.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽	205	186	91	107	91	115	189	238
サビ映像	44	83	192	65	114	112	100	144
サビ音楽動画	135	142	140	90	119	138	157	200
合成音楽動画	227	163	46	82	37	121	212	192

表 3 印象評価値 0.5 以下のメディアの件数
Table 3 Number of media with impression evaluation value −0.5 or less.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽	98	146	277	253	255	270	107	89
サビ映像	312	319	190	308	276	267	163	213
サビ音楽動画	157	217	224	267	235	263	107	139
合成音楽動画	122	173	342	283	408	243	83	110

とになる．

なお，本研究では便宜的に，各印象ベクトルのうち，サビ部分の音楽に対するものを「サビ音楽」，サビ部分の映像に対するものを「サビ映像」，サビ部分の音楽動画に対するものを「サビ音楽動画」，文献 [1] の研究で求められている音楽動画の最初から最後までに対するものを「フル音楽動画」と表記する．

3.5 データセットに対する基礎検討

本研究では，3.4 節で得られたデータセットをメディアごとに高評価のものと低評価のものに分け，それぞれに対して分析を行う．そのため，本節では，得られたデータセット内の各印象軸について印象の高低による件数の偏りを調査し，データセットに対する基礎検討を行う．

表 2，**表 3** は，各印象軸・各メディアタイプでの音楽動画，また合成音楽動画の件数を各評価ごとにまとめたものである．

表 2，表 3 より，印象クラスによって件数に差があることが分かる．しかし，全体として見ると極端な差とはなっていないと考えられるため，これらのデータセットを用いて分析を行う．

4. 音楽動画の印象推定可能性

本章では，3 章で得られた「印象評価データセット」の分析を行うことで，音楽から受ける印象と映像から受ける印象を組み合わせることでの音楽動画の印象推定可能性を検討する．

表 4 コサイン類似度 0.8 以上の音楽動画の割合
Table 4 Percentage of music video clips with a cosine similarity of 0.8 or more.

比較するメディアタイプ	0.8 以上の割合
サビ音楽動画とサビ音楽	0.388
サビ音楽動画とサビ映像	0.386
サビ音楽とサビ映像	0.245
サビ音楽動画とサビ音楽映像平均	0.496
サビ音楽動画とフル音楽動画	0.101

4.1 分析のためのデータの補正

まず C1 から C6 のデータは，1〜5 の 5 段階，Valence および Arousal では −2 から +2 の 5 段階で評価を入力してもらっていた．分析にあたり，両者のデータの最小値と最大値をそろえるため，C1 から C6 のデータは 1〜5 までの数値を，−2〜+2 までの数値へと変換してデータ分析に用いた．

次に，各音楽動画に対するデータセット構築者による各軸に対する評価の平均値を，その音楽動画の印象ベクトルとする．つまり，各音楽動画は，8 軸の印象ベクトル値を持つものとなる．

また，メディア間の比較の際，音楽と映像の組合せと「音楽動画」の関係を調べるため，「サビ音楽」「サビ映像」の印象ベクトルの平均を求めた音楽と映像の平均に関する印象ベクトルを作成し，これを「サビ音楽動画」との比較に使用した．本研究では，便宜的にこれを「サビ音楽映像平均」とする．

4.2 異なるメディア間での印象の相違

各音楽動画コンテンツに対する，あるメディアタイプでの 8 軸の印象評価が，ほかのメディアタイプでの印象評価と一致しているかどうかを調べるため，メディア間の類似度をコサイン類似度で計算する．なお，コサイン類似度で比較するため，どの印象軸についても特徴が出ていない音楽動画を排除する必要があるが，データセットにそのような音楽動画は存在しなかった．

表 4 は，各メディア間での印象値のコサイン類似度が 0.8 以上となった音楽動画の割合を示したものである．

また，**図 2** は，各メディアタイプでのコサイン類似度が閾値以上になった音楽動画の割合を示した図である．ここで，閾値は 0.1 から 0.9 までの間を 0.1 刻みで表示した．

表 4 を見ると，すべてのメディアタイプの比較において，コサイン類似度での一致率が 0.8 以上の音楽動画の割合が 0.5 以下となっている．このことから，それぞれのメディアから受ける印象は食い違っていることが分かる．特に，「フル音楽動画」と「サビ音楽動画」については，ほか

のメディアどうしでの比較よりも割合が非常に小さい値となっている．これは，図 2 を見ても，閾値が 0.5 以上になるとフル音楽動画とサビ音楽動画で一致している件数の割合は他のメディアタイプとの比較よりも低くなっていることからも明らかである．

この原因として，序盤では悲しい雰囲気であったが，サビでは元気の出るような雰囲気に変化していく音楽動画などにおいて，「フル音楽動画」と「サビ音楽動画」で印象が大きく乖離してしまうことが原因である可能性がある．これより，メディアどうしの比較よりも，音楽動画の部分によって受ける印象が大きく違ってしまっていることが分かる．

また，図 2，表 4 より，「サビ音楽動画」から受ける印象と「サビ音楽映像平均」から受ける印象は，「サビ音楽動画」と「サビ音楽」，「サビ音楽動画」と「サビ映像」それぞれとの比較結果に比べ，似通った印象となることが分かった．

サビ音楽とサビ映像から受ける印象がそれぞれ類似していないにもかかわらず，サビ音楽動画とサビ音楽映像平均の評価値が似通っている点はとても興味深い．つまり図 3 のように，音楽と映像がベクトルとして違う方向を指しているが，ベクトルの平均は，そのメディアを融合したものに類似するということになる．このことより，印象ごとに組合せ方は異なるものの，音楽から受ける印象と映像から受ける印象を組み合わせることで音楽動画の印象推定が可能であると考えられる．

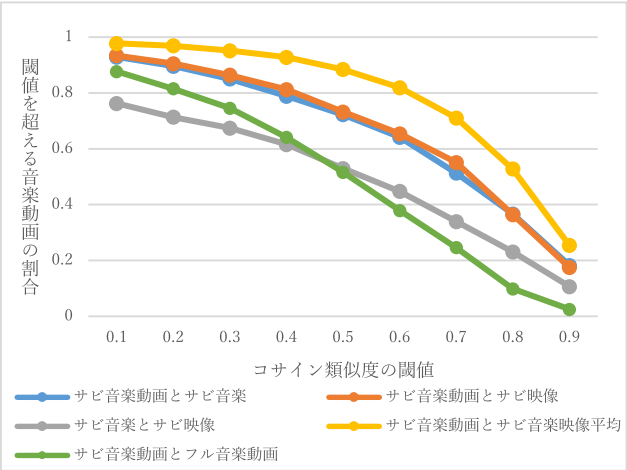

図 2 コサイン類似度が各閾値を超える割合

Fig. 2 Proportion that the cosine similarity exceeds each threshold.

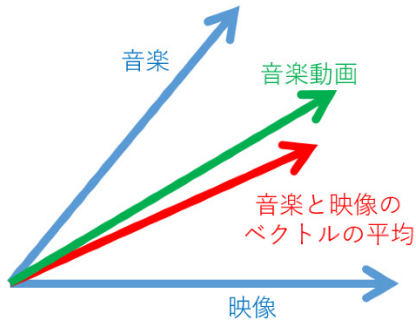

図 3 音楽と映像を合成すると音楽動画の評価に近づく？

Fig. 3 Do you approach the evaluation of music video clips when synthesizing music and movie?

5. 各印象のメディア間での変化

本章では，3 章で得られた「印象変化データセット」「印象評価データセット」の各印象，各メディアに対する詳しい分析を行うことで，音楽動画の印象がどのメディアからどのような影響を受けるのかを明らかにする．その際，4.1 節と同様のデータ補正を行ったのちに分析を行った．

5.1 印象変化データセットを用いた印象評価値の分析

本節では，音楽のみ，映像のみがどのように組み合わさって音楽動画の印象になっているのかを分析する．その際，合成音楽動画の印象評価値が −0.5 以下のものを低評価群，印象評価値が +0.5 以上のものを高評価群として分析を行う．

分析手法としては合成音楽動画の各評価群のうち，合成元の音楽，映像がどの評価群に属していたかを調べた．図 4 は，印象変化データセットをもとに，縦軸を映像の印象評価値，横軸を音楽の印象評価値とし，そこから生成された音楽動画を評価群ごとに色別にプロットしたものを各印象で表示したものである．ここで，高評価群は赤い丸で，低評価群は青いバツ印で，どちらに属さないものを緑の十字で表示した．

また，表 5 は，「合成音楽動画」の評価値と「音楽のみ」「映像のみ」のそれぞれについて各印象の相関係数を表示したもの，表 6，表 7 は散布図の各象限に属する合成音楽動画の件数を，音楽動画が高評価の場合と音楽動画が低評価の場合のそれぞれで表示したものである．

表 5 合成音楽動画とサビ音楽，サビ映像の相関係数（印象変化データセット）

Table 5 Correlation coefficients of synthetic music video clips and chorus part music, chorus part movie (Impression change dataset).

	サビ音楽	サビ映像
C1(堂々)	0.5239	0.0059
C2(元気が出る)	0.5581	0.2872
C3(切ない)	0.5944	0.2316
C4(激しい)	0.6852	0.1748
C5(滑稽な)	0.3598	0.3497
C6(かわいい)	0.4096	0.5206
Valence	0.1346	-0.0754
Arousal	-0.0762	-0.1129

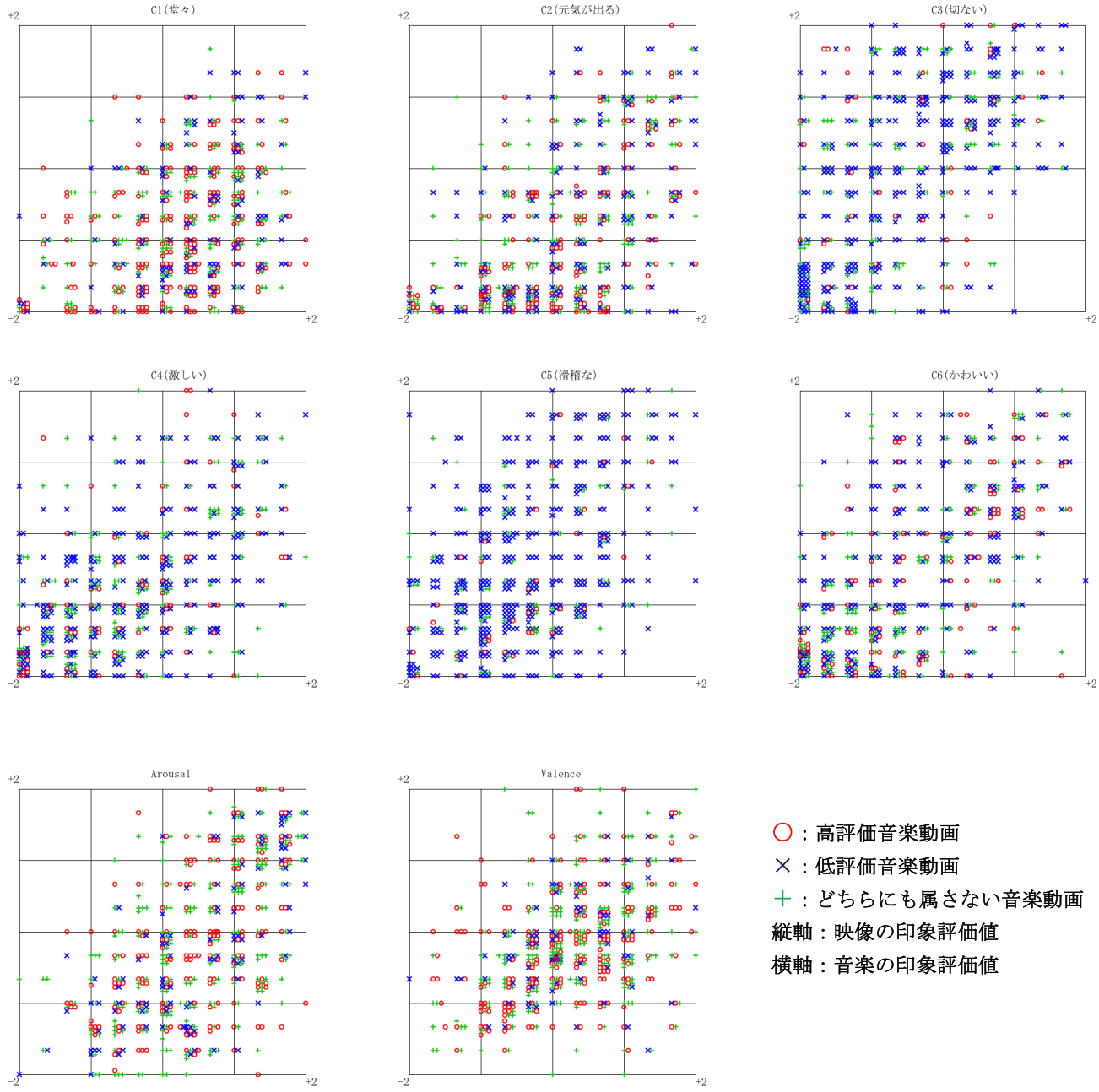

図 4 印象変化データセットを用いた分析の結果
Fig. 4 Results of analysis using impression change dataset.

5.1.1 印象軸ごとの傾向の分析

本項では，印象変化データセットにおいて，どの印象ではどのメディアタイプが特に影響を及ぼしているのかといった，印象軸ごとの詳しい分析を行う．

図 4，表 6，表 7 のうち，C2（元気が出る），C3（切ない），C4（激しい）では，音楽の印象評価が正の値である場合，高評価群に属する音楽動画の数が低評価群に属する音楽動画の数よりも多く，音楽の印象評価が負の値である場合，低評価群に属する音楽動画の数が多いことが分かる．また，表 5 より，C1（堂々），C2（元気が出る），C3（切ない），C4（激しい）の印象では「サビ音楽」と「合成音楽動画」の評価値の相関係数が 0.4 よりも高くなっていることが分かる．これらのことより，C1，C2，C3，C4 では，音楽の印象評価値に影響されやすい傾向があることが分かった．

図 4，表 6，表 7 のうち，C5（滑稽な），C6（かわいい）では，ほかの印象軸と比べ，音楽の印象評価値が負の値であり，映像の印象評価値が正の値である場合，音楽動画が高評価群に属している音楽動画の件数が多い．また，表 5 より，C1〜C4 では「合成音楽動画」と「サビ音楽」の相関係数が「サビ映像」と比較して高くなっていたが，それに対し，C5（滑稽な），C6（かわいい）では相関係数が同

表 6 高評価合成音楽動画の内訳
Table 6 Breakdown of high rating synthetic music video clips.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽が正の値かつサビ映像が正の値	47	30	12	14	9	36	85	82
サビ音楽が正の値かつサビ映像が負の値	104	68	7	19	4	14	49	56
サビ音楽が負の値かつサビ映像が正の値	8	3	13	5	2	13	23	9
サビ音楽が負の値かつサビ映像が負の値	64	59	13	44	21	54	51	41
合計	223	160	45	82	36	117	208	188

表 7 低評価合成音楽動画の内訳
Table 7 Breakdown of low rating synthetic music video clips.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽が正の値かつサビ映像が正の値	32	49	105	51	87	59	36	47
サビ音楽が正の値かつサビ映像が負の値	61	62	13	58	64	30	17	30
サビ音楽が負の値かつサビ映像が正の値	5	3	86	24	68	36	9	2
サビ音楽が負の値かつサビ映像が負の値	20	52	130	139	176	113	19	26
合計	118	166	334	272	395	238	81	105

程度になっている．これらのことより，C5（滑稽な），C6（かわいい）の印象軸に関しては，ほかの印象軸よりも映像の印象評価が音楽動画の印象評価に大きく関わっている傾向があることが分かった．

図 4 のうち，Valence，Arousal では，音楽動画の高評価群，低評価群ともに散らばっていることが分かる．また，表 5 より，「合成音楽動画」と「サビ音楽」「サビ映像」それぞれの相関係数も非常に低くなっていることが分かる．これより，Valence，Arousal では音楽の印象と映像の印象に直接の相関がなく音楽動画の印象になっていると考えられる．

5.1.2 結果

これらの結果より，C1～C6 までは音楽動画の印象と音楽の印象，音楽動画の印象と映像の印象に関しては，各印象軸で異なる相関関係が存在すること，Valence，Arousal に関しては，相関関係が見受けられないことが分かった．

また，結果として，任意に作成した合成音楽動画では，映像に影響される印象軸よりも音楽に影響される印象軸の方が多いことが明らかになった．これより，音楽動画の印象を音楽と映像の印象評価から推定する際には，音楽の印象に重みをつけることが必要であることが分かった．

また，音楽動画を制作する際には，音楽に重点を置くことで印象を伝えるのが容易になると考えられる．

また，評価項目のうち，「違和感を覚える」の項目では，違和感を覚えると評価された音楽動画は，500 件のうち 15 件程度であり，ランダムに作成した音楽動画であっても違和感は少ないことも分かった．

5.2 印象評価データセットを用いた印象評価値の分析

本節では，ランダムに生成した音楽動画と，オリジナルの音楽動画それぞれから受ける印象の違いを明らかにするために分析を行う．具体的に，我々の過去の研究 [12] で構築した，500 件の音楽動画のサビ部分を音楽のみ，映像のみに分離したものに対する印象評価データセットを用いて，4.1 節と同様の補正を行ったデータで比較を行った．

表 8 サビ音楽動画とサビ音楽，サビ映像の相関係数（印象評価データセット）
Table 8 Correlation coefficients of chorus part music video clips and chorus part music, chorus part movies (Impression evaluation dataset).

	サビ音楽	サビ映像
C1(堂々)	0.5323	0.4993
C2(元気が出る)	0.6437	0.7064
C3(切ない)	0.6736	0.7148
C4(激しい)	0.6335	0.5478
C5(滑稽な)	0.5778	0.6284
C6(かわいい)	0.6837	0.7711
Valence	0.6387	0.4812
Arousal	0.6771	0.7272

図 5 は，印象評価データセットをもとに，縦軸を映像の印象評価値，横軸を音楽の印象評価値とし，そこから生成された音楽動画を評価群ごとに色別にプロットしたものを各印象で表示したものである．ここで，高評価群は赤い丸で，低評価群は青いバツ印で，どちらに属さないものを緑の十字で表示した．また，**表 8** は，「サビ音楽動画」の評価値と「サビ音楽」「サビ映像」のそれぞれについて各印象の相関係数を表示したもの，**表 9**，**表 10** は散布図の各象限に属するサビ音楽動画の件数を，音楽動画が高評価の場合と音楽動画が低評価の場合のそれぞれで表示したもので

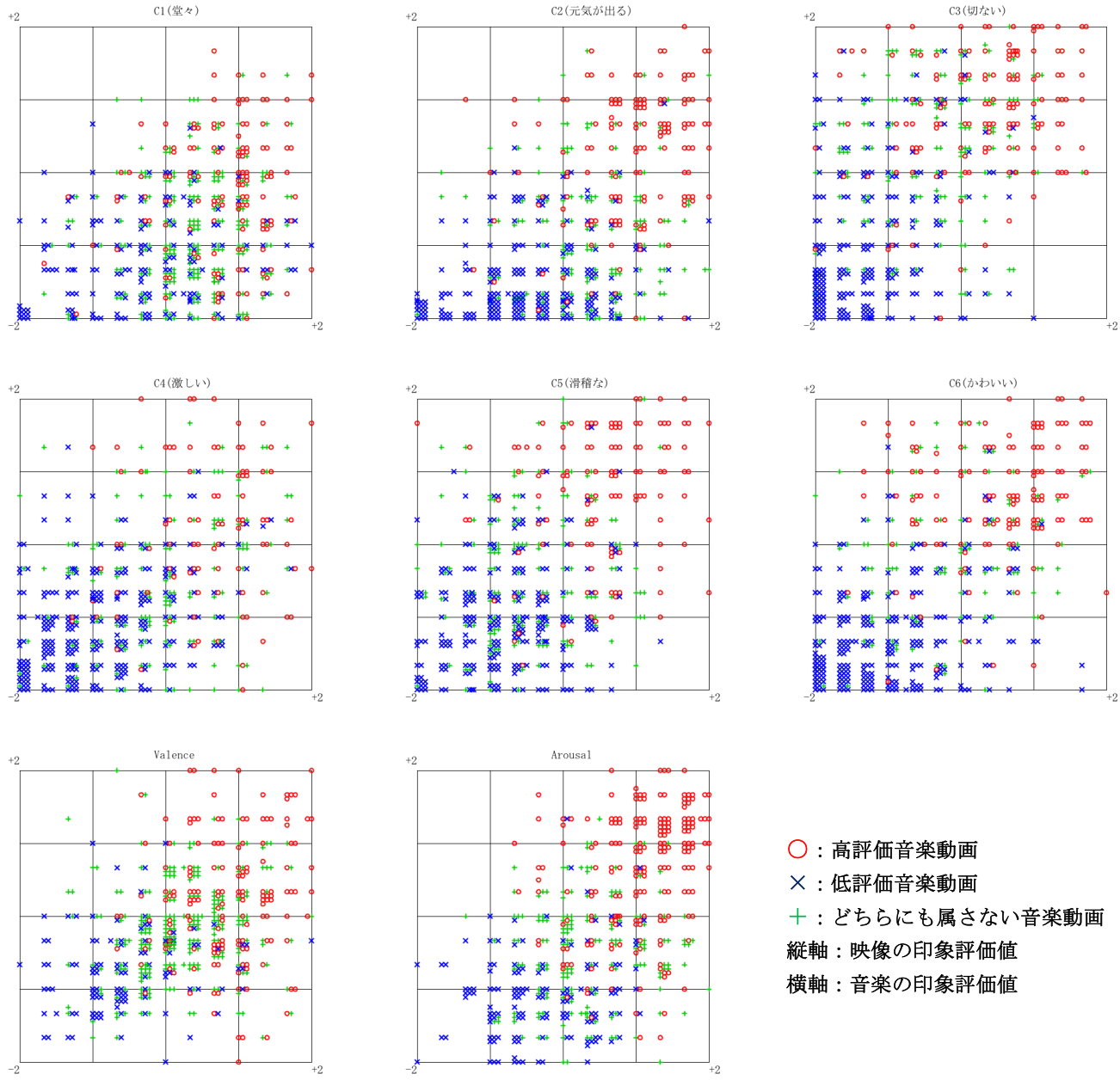

図 5 印象評価データセットを用いた分析の結果
Fig. 5 Results of analysis using impression evaluation dataset.

表 9 高評価サビ音楽動画の内訳
Table 9 Breakdown of high evaluation chorus part music video clips.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽が正の値かつサビ映像が正の値	64	86	93	47	79	93	103	154
サビ音楽が正の値かつサビ映像が負の値	57	43	5	25	16	11	43	36
サビ音楽が負の値かつサビ映像が正の値	4	6	36	8	18	28	6	8
サビ音楽が負の値かつサビ映像が負の値	10	7	6	10	6	6	5	2
合計	135	142	140	90	119	138	157	200

ある．

5.2.1 データセット間の印象評価の相違

本項では，印象評価データセット分析の結果と，印象変化データセットの分析の結果での相違についての詳しい分析を行う．

印象変化データセットを用いた分析の結果では，音楽動画の印象に音楽が大きく影響しているという結果になった一方で，図 5 ならびに表 8 より，音楽動画の印象に対して

表 10 低評価サビ音楽動画の内訳
Table 10 Breakdown of low evaluation chorus part music video clips.

	C1	C2	C3	C4	C5	C6	V	A
サビ音楽が正の値かつサビ映像が正の値	7	4	11	12	10	10	5	8
サビ音楽が正の値かつサビ映像が負の値	59	67	9	36	33	31	21	48
サビ音楽が負の値かつサビ映像が正の値	4	3	41	16	25	12	13	3
サビ音楽が負の値かつサビ映像が負の値	87	143	163	203	167	210	68	80
合計	157	217	224	267	235	263	107	139

音楽と映像が同程度に影響を及ぼしていることが分かる．図 5 をみると，すべての印象において，音楽の印象と映像の印象が同程度の影響を及ぼして音楽動画の印象になっている傾向があることが分かる．その一方で，印象ごとに影響の仕方が変わっていることも分かる．図 5，表 9 のうち，C3（切ない），C5（滑稽な），C6（かわいい）では，映像の印象評価値が正の値の場合，印象評価値が高評価になっている音楽動画の件数が図 4 での結果と比べ，多くなっていることが分かる．また，図 5，表 9，表 10 の C1 や C2 では音楽の印象評価が高く，映像の印象評価が低くても音楽動画での印象評価が高評価になっているものが多い．これより，C1（堂々），C2（元気が出る）は音楽に影響を受けやすい印象であると考えられる．このように，各印象軸で違った影響の受け方をしていることより，実際の印象推定では，各印象軸で違った組合せ手法を考える必要があると考えられる．

また Valence，Arousal に関して，表 5 では相関がみられなかったのに対し，表 8 では，音楽と映像それぞれに対して相関している．また，C3 における結果では，合成音楽動画では音楽に影響を受ける傾向があったが，サビ音楽動画では映像に影響を受けやすくなっている．

図 4，図 5 のどちらの結果でも，C5 に関しては，比較的音楽動画の印象評価値が低評価となっているものの件数が多い．これは，C5 という印象が音楽動画というコンテンツにおいて伝えるのが比較的困難な印象であるからであると考えられる．また，図 4 では，図 5 に比べ，音楽動画の印象評価が低評価となっているものが多いことが分かる．

5.2.2 結果

これらの結果より，音楽動画は印象軸ごとにそれぞれ異なった音楽と映像の組み合わさり方をしていることが分かったが，サビ音楽動画と合成音楽動画でも異なった組み合わさり方をしていることが分かった．

この違いは，それぞれのデータセット内の音楽動画の作成方法の違いによって起きたと考えられる．つまり，映像制作者と音楽制作者間での意思の取り合いができている場合と，音楽と映像をランダムで組み合わせて制作したような，制作者の意図を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る．

この結果より，音楽動画の印象推定として，音楽動画の制作者たちが意図して制作したものなのか，音楽動画をランダムに組み合わせて制作したものかという点に関して考慮する必要があることが分かった．

6. まとめ

本研究では，500 件の音楽動画のサビ部分に対し，「音楽動画」「音楽のみ」「映像のみ」の 3 タイプのメディアに分離したものに対して印象評価データセットを作成し，分析を行った．また，音楽と映像を任意に組み合わせた音楽動画を作成し，それに対する印象変化データセットもあわせて構築した．またそれについて，音楽の印象，映像の印象と音楽動画の印象の関わり方について可視化ならびに分析を行った．

その結果，音楽動画全体とサビ部分では受ける印象が大きく食い違うこと，また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できる可能性があることを明らかにした．また，音楽のみでの印象と音楽動画の印象，映像のみの印象と音楽動画の印象はそれぞれの印象軸ごとに異なった相関がある傾向を明らかにした．

今後は，サビ部分だけでなく楽曲全体での印象の組み合わさり方の分析を行うとともに，本研究で明らかにした傾向をもとに，音楽の印象と映像の印象の組合せによる印象推定手法の提案を行っていく予定である．

謝辞 本研究の一部は，JST CREST，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：楽曲動画印象データセットの作成とその分析，ARG 第 2 回 Web インテリジェンスとインタラクション研究会 (2013).
[2] 後藤真孝：SmartMusicKIOSK：サビ出し機能付き音楽試聴機，情報処理学会論文誌，Vol.44, No.11, pp.2737–2747 (2003).
[3] 大出訓史，今井 篤，安藤彰男，谷口高士：音楽聴取における “感動” の評価要因—感動の種類と音楽の感情価の関係，情報処理学会論文誌，Vol.50, No.3, pp.1111–1121 (2009).
[4] 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).

[5] 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, No.7, pp.1053–1057 (2001).
[6] 熊本忠彦，太田公子：印象に基づく楽曲検索システムの設計・構築・公開，人工知能学会論文，Vol.21, pp.310–318 (2006).
[7] Russell, J.A.: A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, Vol.39, No.6, pp.1161–1178 (1980).
[8] 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ビジュアル複合情報処理，pp.1–5 (2013).
[9] 山本岳洋，中村聡史：視聴者の時刻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).
[10] 熊本忠彦，太田公子：印象に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，Vol.44, No.7, pp.1808–1811 (2003).
[11] Morency, L.-P., Rada, M. and Payal, D.: Towards multimodal sentiment analysis: Harvesting opinions from the web, *Proc. 13th International Conference on Multimodal Interfaces*, ACM (2011).
[12] 土屋駿貴，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に関する考察，情報処理学会研究報告グループウェアとネットワークサービス，Vol.96, No.3, pp.1–6 (2015).
[13] 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，九州大学出版会 (2011).
[14] Machajdik, J. and Hanbury, A.: Affective image classification using features inspired by psychology and art theory, *Proc. 18th ACM International Conference on Multimedia*, ACM (2010).
[15] Ashkan, Y. et al.: Multimedia content analysis for emotional characterization of music video clips, *the EURASIP Journal on Image and Video Processing*, No.26 (2013).
[16] Acar, E. et al.: Understanding Affective Content of Music Videos through Learned Representations, *Proc. MMM2014*, pp.303–314 (2014).

大野 直紀 （学生会員）

1995 年生．2017 年明治大学総合数理学部先端メディアサイエンス学科卒業．現在，同大学大学院先端数理科学研究科博士前期課程在学中．音楽動画の印象に関する研究や人間の視覚に関する研究活動に従事．学士（理学）．

土屋 駿貴

1993 年生．2017 年明治大学総合数理学部先端メディアサイエンス学科卒業．現在，同大学大学院先端数理科学研究科博士前期過程在学中．音楽動画の印象推定や料理行為における個性等の研究活動に従事．学士（理学）．

中村 聡史 （正会員）

1976 年生．2004 年大阪大学大学院工学研究科博士後期課程修了．同年独立行政法人情報通信研究機構専攻研究員．2006 年京都大学大学院情報学研究科特任助手，2009 年同特定准教授，2013 年明治大学総合数理学部先端メディアサイエンス学科准教授，現在に至る．サーチとインタラクションや，ネタバレ防止技術，平均手書き文字等の研究活動に従事．ヒューマンインタフェース学会等の会員．博士（工学）．

山本 岳洋 （正会員）

京都大学大学院情報学研究科社会情報学専攻助教．2011 年京都大学大学院情報学研究科博士課程修了．博士（情報学）．情報検索におけるユーザインタラクションやユーザ理解に関する研究に従事．日本データベース学会，ACM 各会員．