情報処理学会論文誌　データベース　**Vol.6　No.3　148–158 (June 2013)**

視聴者反応と音響特徴量に基づくサムネイル動画の生成手法

中村 聡史[1,4,a)]　山本 岳洋[2,4]　後藤 真孝[3,4]　濱崎 雅弘[3,4]

受付日 2012年12月20日, 採録日 2013年4月8日

概要：本稿では，動画共有ウェブサイトにおいて日々アップロードされる膨大な楽曲動画について，ユーザがその動画に対して興味があるかどうかを短時間で判断する手段として，15 秒のサムネイル動画を自動生成する手法を提案する．ここでは，楽曲動画作成者と楽曲動画視聴者に注目し，楽曲動画のサビ検出技術と，視聴者の盛り上がり検出技術を使うことにより，サムネイル動画を自動生成する仕組みを実現する．また，評価実験により，組合せ手法の有効性とその特徴を明らかにする．

キーワード：視聴者反応，音響特徴量，サムネイル動画

Methods of Generating a Thumbnail Video Clip Based on Viewers' Responses and Audio Features

Satoshi Nakamura[1,4,a)]　Takehiro Yamamoto[2,4]　Masataka Goto[3,4]　Masahiro Hamasaki[3,4]

Received: December 20, 2012, Accepted: April 8, 2013

Abstract: The number of uploading video clips to video sharing Web sites has been explosively increasing. Then, it is not easy for users to find their preferable video clip. In this paper, we propose several methods to generate a fifteen-second thumbnail video clip from an original video clip based on analysis of viewer's responses and analysis of audio features. We clear the advantage and characteristics of the combination method with analysis of viewer's responses and analysis of audio features based on the evaluation test.

Keywords: viewer's responses, audio features, thumbnail video

1. はじめに

YouTube [*1]やニコニコ動画[*2]に代表される動画共有ウェブサイトが爆発的に成長している．YouTube では 2012 年 1 月時点で 1 秒あたりに動画長にして 1 時間分の動画が投稿され[*3]ている．また，ニコニコ動画でも，2012 年 11 月 20 日の我々の調査によると，1 日あたりに約 5,500 本の動画が，1 秒あたりに動画長にして約 47 秒分の動画が投稿されている（**図 1** 参照．2010 年 9 月 1 日から 2012 年 10 月 29 日までにニコニコ動画上にアップロードされたすべての動画の総再生時間を計算し，1 日 = 86,400 秒で割った値．土日および祝日にアップロードが集中する傾向が見て取れる）．YouTube は 2012 年 1 月の時点で 1 日に 40 億視聴，ニコニコ動画では 2012 年第 2 四半期に 1 日あたり 1 億視聴，平均視聴者数は 786 万人で，ユーザの平均視聴時間は 102.5 分であるという[*4]．

YouTube とニコニコ動画は様々に違う点があるが，その中でも顕著な違いは，ある動画と，その動画の視聴者によって投稿されるコメントとがどの程度密接になっているかという点である．YouTube では，一般的に動画全体へのコメントを投稿するようになっている．YouTube でも動画のある再生時間に対してコメント（注釈）を投稿するこ

1 明治大学
Meiji University, Nakano, Tokyo 164–8525, Japan
2 京都大学
Kyoto University, Kyoto 606–8501, Japan
3 産業技術総合研究所
AIST, Tsukuba, Ibaraki 305–8568, Japan
4 JST CREST
JST CREST, Chiyoda, Tokyo 102–0076, Japan
a) satoshi@snakamura.org

*1 http://www.youtube.com/
*2 http://www.nicovideo.jp/
*3 http://jp.techcrunch.com/archives/20120123youtube-reaches-4-billion-views-per-day/
*4 http://dic.nicovideo.jp/a/ニコニコ動画

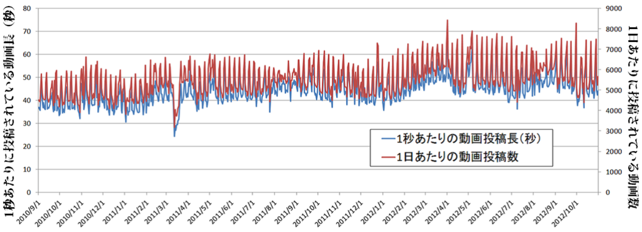


図 1　ニコニコ動画において 1 秒・1 日あたりに投稿されている量
Fig. 1　Variation of the quantity of uploaded video clips per day.

とも可能ではあるが，その再生時間軸へのコメント投稿は動画の頭出し機能の役目しか果たしていない．一方，ニコニコ動画では，コメントは動画を再生しながらその再生時間に対して投稿されるようになっており，投稿された再生時間にそのコメントがその動画上を流れるように提示される．それぞれのコメントの投稿された日時は異なっていても（投稿時間は非同期），動画の同じ再生時間に対して投稿されているコメントは動画上で一緒に流れる（再生時間軸で同期）ため，いつでも他者との盛り上がりを体験することができる．つまり，いつでも擬似的に感情を同期および共有することが容易なシステムであり，これが動画視聴体験をより豊かで面白いものにしているといえる．そこで本稿では，視聴者と動画との関係が深いニコニコ動画に注目する．なお，ニコニコ動画に類似したサービスもその後海外で登場しており，多くの視聴者と動画に対するコメントを集めている．今後こうしたサービスは世界中に広がると期待される．

さて，ニコニコ動画には日々膨大な量の質の高い動画がアップロードされているが，すべての動画を視聴するのは時間的に難しい．動画共有ウェブサイトにおいて好みの動画を探すには，キーワード検索を行うことによって動画集合を絞り込み，そこから目的とする動画をタイトルや説明文，タグなどの情報を頼りに探し出すか，デイリーまたはマンスリーランキングなどから選び出す．または，これまでの視聴履歴などを利用した推薦システムから動画を選ぶか，Blog などで紹介されているものを頼りに探すことが多い．こうした結果や推薦では，動画はタイトルや説明文，タグなどのテキスト情報と，サムネイル画像によって表示されている．

一般的なウェブページ検索や画像検索であれば，その結果が適合するかどうかを，提示されているテキストサマリやサムネイル画像から判断することができる．しかし，動画は時間的連続性のあるコンテンツであるため，タイトルや説明文，タグなどのテキスト情報とサムネイル画像だけでは動画の時間的連続性を表現することができず，その提示されているコンテンツがユーザ自身にとって求めるものかを判断することは困難である．

ユーザが，そのコンテンツを視聴するかどうかを判断するには，そのコンテンツの良さを短時間で伝える技術が必要となる．そこで我々は，その動画の魅力や雰囲気を音楽（音声）や映像とともに短時間で視聴および把握し，その動画自体を見るかどうかを判断することを可能とする，「サムネイル動画自動生成技術」を実現することを目的とする．

なお，本稿ではサムネイル動画自動生成の第 1 歩として，歌声合成技術 VOCALOID [8] を利用した楽曲動画に注目する．VOCALOID に基づくソフトウェア（「初音ミク」など）は，作詞や作曲ができるユーザに対して，思いどおりに歌ってくれる「歌い手」を提供する役割を果たしてきた．さらに，これまで楽曲作成に興味があったユーザだけでなく，これまで楽曲作成とは無縁であったユーザ層による創作活動まで活発にしてきた．また，ニコニコ動画ならではの N 次創作（他者の作成・投稿したものを引用しつつ利用し，別のコンテンツ創作を行う）文化 [6], [7] によって，現在では多くの人々の手によって膨大な数の楽曲が日々創作，公開されている [4]．このように，多くのユーザが楽曲動画を創作する世の中では，人手で 1 つずつアノテーションを付与したり，サムネイル動画を作ったりすることは困難である．

そこで本稿では，楽曲動画からサムネイル動画を自動生成するため，楽曲動画に対する視聴者の盛り上がりを推定する手法と，楽曲動画の自動理解技術を用いてサビ区間を検出する手法，そしてその両手法を融合した手法を提案する．また，評価実験を行うことにより手法の有用性を明らかにし，さらに実験結果を考察することにより，各手法の可能性について議論を行う．なお，本稿では，サムネイル動画は 15 秒と設定している．15 秒の設定理由は，テレビ CM の一般的な長さが 15 秒であることを理由にしている．

2.　提案手法

ニコニコ動画にアップロードされる動画は数分～数十分のものが大半である．サムネイル動画の 15 秒という時間はかなり限られて短いため，ある程度の内容把握を可能とする動画要約とは本質的に問題が異なる．

ここで，良いサムネイル動画に対して求められることは，そのサムネイル動画の生成元となったオリジナル楽曲動画の雰囲気や，良さ，盛り上がり，質の高さを表現できており，サムネイル動画を視聴したユーザに対してそのオリジナル動画を見たいと思わせることである．サムネイル動画をどのような目的で作成するかという点についてはいろいろ考えられるが，本稿では特に「オリジナル楽曲動画を見たいと思わせる，ユーザにとって質が高く感じるシーンを抽出すること」に注目する．

ここで，オリジナル楽曲動画を見たいと思わせるような盛り上がりを抽出する際，楽曲動画投稿者（楽曲動画製作者）視点，楽曲動画視聴者視点に立った場合，以下の 2 点が考えられる．

- 楽曲においてサビは一般的にメインで一番盛り上がる

部分である．つまり，楽曲動画投稿者（楽曲動画作成者）にとって，楽曲のサビにあたる部分は特に盛り上げようとしているシーンであり，聞いてほしいお薦めの部分であると考えられる．
- ニコニコ動画は，動画の再生時間軸でコメントを介して感情を共有できるサービスである．多くの視聴者が感情を顕わにしているシーンは，視聴者らにとって盛り上がるシーンであると考えられる．

つまり，楽曲動画のサビ区間や，楽曲動画視聴者が感情的に盛り上がっているシーンが，サムネイル動画として抽出する部分に適していると考えられる．

そこで以下では，まず視聴者が投稿したコメントからの盛り上がりの取得方法と，音響分析に基づく楽曲動画からのサビの検出手法について述べる．次に，サムネイル動画の自動生成手法として，楽曲動画の視聴者の反応を利用した手法，楽曲動画のサビを利用した手法，その 2 つを融合した手法を提案する．

2.1　視聴者の感情の時間的変化推定

先述のとおり，ニコニコ動画は時間的に非同期な視聴者と，あたかも一緒に視聴しているかのような体験を提供している．そのため，特に笑えるシーンでは他のユーザが投稿したコメントによりさらに笑いが増幅され，特に泣けるシーンでは他ユーザが投稿した悲しみのコメントによりさらに悲しみが増幅され，さらなる笑い/泣きのコメントを集める傾向がある．これはバラエティ番組などの映像コンテンツに観客の笑い声を人為的に付加したり，ドキュメンタリー番組に俳優の泣き顔をスーパインポーズしたりすることにより，視聴者の共感作用に働きかけ笑いや泣きの閾値を下げる効果に似ている．

Nakamura らのニコニコ動画上の 968,721 件の動画を対象とした調査 [11] によると，動画の最初と最後にコメントが集中する傾向があり，同様に動画の最初と最後に肯定的なコメントが集中することも分かっている．特に，動画の最後において肯定的コメントが集中的に投稿されることが分かっている．これは，視聴者らが動画の最後において，動画投稿者に対する感謝や今後への期待をコメントとして投稿しているためである．

上記の調査はすべての動画を対象としていたため，今回新たに上記データセットのうち，「音楽」および「BGM」とタグが付与された 4,926 件の動画について時間的なコメント量の変化をプロットしたものが**図 2** である．ここで利用した楽曲動画には，かなり動画長が長いものもあったため，後半に進むに従いコメント総量が減る傾向が見て取れるが，それを考慮しても最初と最後にコメントが集中するという，文献 [11] と類似した傾向になっていることが分かる．以上のことより，単純にコメント量だけを利用した手法では，動画の開始や終了の区間が抽出されてしまい，盛り上がるシーンを抽出できないことも多い．実際，楽曲動画の最初は歌が再生されるまでの無音区間や歌が始まる前のイントロが流れていることが多く，楽曲の最後は同じく無音区間やフェードアウトしていることが多かった．こうしたシーンはサムネイル動画としてはふさわしくない．

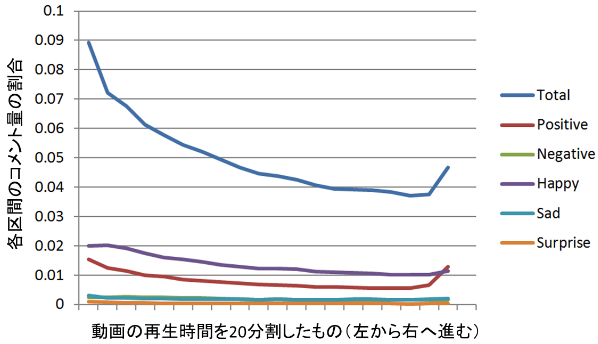


図 2　音楽動画に付与される時間ごとのコメント量の変化

Fig. 2　Changing the number of posted comments in each play-back period in music video clips.

そこで我々は，ニコニコ動画上の動画に対して投稿された視聴者のコメントから喜びや悲しみ，驚きといった感情表現を取得し，その度合いの時間的変化をとって感情がより増幅されているシーンを抽出することによって，視聴者にとってより盛り上がるサムネイル動画を生成できるのではないかと考えた．

コメントからの感情推定については，一般的な自然言語処理は困難である．そこで，我々が過去に構築した感情判定正規表現辞書 [11] を新しい語などに対応させ，パターンマッチにより判定している．ここでは，「笑った」「ワロタ（わらった→わらた→ワロタと変化）」「www（笑いの略称）」などのテキストだけでなく，顔文字の一部も正規表現で判定するようにしている．なお，文献 [11] では喜び，悲しみ，驚き，肯定，否定を分類していたが，特に肯定は楽曲の最初と最後に，期待や賞賛，感謝の意味で現れる傾向が強く，サムネイル動画自動生成においてはマイナスに働くと考えたため，本稿では喜び，悲しみ，驚きを統合して感情コメントとし，利用する．

コメントは動画の再生時間においてミリ秒を指定して投稿可能であるが，本稿では 1 秒を最小単位とする．また，単純に数の時間的変化をプロットすると，弾幕（あるシーンに対して決まりごとのようにいっせいに投稿されるコメント群）などに左右され，誤ったシーンが検出されることが多かったため，時間的変化についてはローパスフィルタをかけることにより平滑化を行う．結果として，**図 3** の左上のようなグラフが作成される．図 3 の縦軸は抽出された感情コメントの数を，横軸は動画の再生時間軸を表している．

2.2　楽曲動画からのサビ検出

先述のとおり，ニコニコ動画に日々アップロードされて

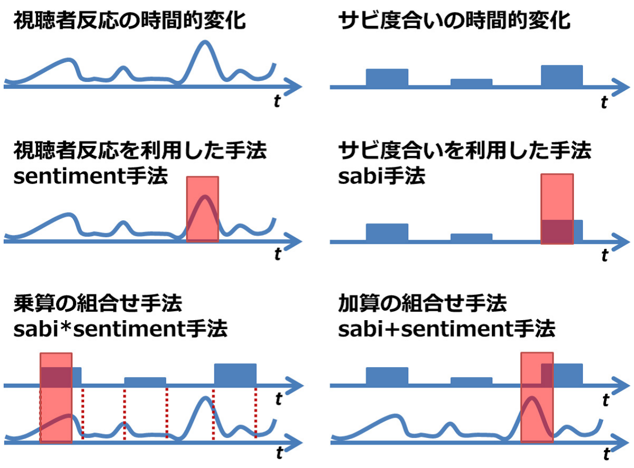


図 3　提案手法の視覚化
Fig. 3　Visualization of the behavior of each proposed method.

いる楽曲動画は，様々な人の手によって作詞，作曲されているものである．一方，こうした楽曲動画には，どこから楽曲が始まりどこで終わるか，サビや A メロ，B メロなどといったアノテーションが付与されているわけではない．

そこで本研究では，文献 [3] で提案されているサビ区間検出手法 RefraiD を用いる．RefraiD は，サビは楽曲中で最も多く繰り返されることが多いことに着目した手法である．RefraiD では，楽曲中の様々な繰返し区間の相互関係を調べることによって，楽曲中で繰り返されるすべてのサビ区間を網羅的に推定しようとする．実際には，音響信号の特徴量としてコードとメロディが反映されやすい 12 次元のクロマベクトルを求め，クロマベクトル間の類似度を利用することによって，全体の響きがある程度の区間類似していれば繰返し区間であると検出する．

RefraiD は，楽曲中の様々な繰返し区間をグルーピングして繰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」を評価する．ここで，集合内の繰返し区間の数が多く，しかも区間どうしが似ていて長いほど，その繰返し区間の集合がサビらしいとして，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．ただし，選択の際にはポピュラー音楽を前提としたヒューリスティクスも考慮する．本稿では，このサビ区間として検出された集合中のそれぞれの区間に対して，それがどれぐらい他の区間と似ているかをそのサビ区間の信頼度スコアと見なし，サビ区間集合の中でもよりサビらしい（信頼度が高い）区間を，そのスコアによって求める．

RefraiD が計算したサビ区間ごとの信頼度スコアに基づいて，図 3 右上のようなグラフが作成できる．ここで，図の縦軸の値が高いほどサビらしい区間となっている．

2.3　サムネイル動画自動生成手法

本稿では視聴者反応と音楽的特徴量に基づくサビ検出を利用した 4 つの手法を提案する．

- **sentiment 手法：**2.1 節で推定した楽曲動画に対する感情コメント（喜び，悲しみ，驚き）をスコアとし，そのスコアの時間的変化から盛り上がっているシーンを抽出し，その盛り上がっている部分を中心として 15 秒分をサムネイル動画として抽出する手法．
- **sabi 手法：**2.2 節で推定した楽曲動画のサビ区間集合の中から，その信頼度スコアに基づき最もサビらしい区間を推定し，その区間の最初から連続した 15 秒をサムネイル動画として抽出する手法．
- **sabi*sentiment 手法：**楽曲動画のサビ区間集合内で，どこが最も視聴者の反応を集めているかということを考慮し，最適なサビ区間を選び出す手法．実際には，まず楽曲内の各サビ区間への感情コメント量をそれぞれ求め，各サビ区間の信頼度とその区間に投稿されたコメント量を掛け合わせる．次に，それぞれのサビ区間のスコアを求め，スコアが最も高いものを視聴者が最も盛り上がっているサビとして推定する．さらに，その区間の最初から連続した 15 秒をサムネイル動画として抽出する．
- **sabi+sentiment 手法：**まず，楽曲動画に対して投稿された感情コメント量の時間的変化と，楽曲内のサビらしさの信頼度を計算し，それぞれの最大値で時間的変化の値を正規化する．次に，それぞれの数値を各再生時間で加算することで，スコアの時間的変化を作成する．そこからスコアが最も高い部分を中心として 15 秒分をサムネイル動画として抽出する．

図 3 は，各実験手法がそれぞれどこを切り出すかを例示した模式図である．図において赤の矩形で囲まれた領域が，サムネイル動画として抽出される部分である．

抽出では，まずすべての時間（秒）のスコアを計算し，その中でスコアが最大値となる時間 t を求める（ここで，同じスコアのものがある場合は前を優先する）．次に，その時間 t を必ず含むように $t+i-15$ 秒から $t+i$ 秒までのスコアの積分値を，i を 0 から 15 まで変化させてそれぞれ求め，そのスコアの積分値が最も高くなる i を求める．そのうえで，スコアの積分値が最も高くなる $t+i-15$ 秒から $t+i$ 秒までをサムネイル動画として抽出する．

3.　評価実験

どのサムネイル動画生成手法がユーザにとってオリジナル動画を視聴したいと思えるようなものになっているかを明らかにするため，ユーザベースでの評価実験を行った．本実験では動画共有ウェブサイト上ですでに高く評価されている，質の高い動画を評価実験の対象とする．理由としては，ある一定以上の評価をされている動画であれば，好き嫌いの差が少なく，サムネイル動画の質の高さが，そのオリジナル動画を見たいと思わせることにつながると考えたためである．

3.1　実験準備と手続き

評価実験のため，ニコニコ動画からデータセットを構築した．ここでは，ニコニコ動画から「VOCALOID」タグの付与された動画 186,987 本を収集し，これを VOCALOID を利用した楽曲を扱った楽曲動画であると見なし処理の基礎データセットとする．また，上記データセットの中で，1,000 件以上のコメントが視聴者によって投稿されている楽曲動画を対象とする．なお，実験開始段階で 1,763 本の楽曲動画について，下記に説明するベースラインの 2 手法も含めた 6 手法すべてによりダイジェスト生成ができたため，この動画群を評価実験用のデータセットとした．

ここで，ベースライン手法として下記の 2 手法を用意した．

- **middle 手法（ベースライン手法 1）**：まず楽曲動画の長さを求め，その楽曲の中央の 15 秒をサムネイル動画として抽出する手法．
- **comment 手法（ベースライン手法 2）**：楽曲動画に対するすべてのコメントの時間的変化から盛り上がっているシーンを抽出し，その盛り上がっているシーンを中心として 15 秒分をサムネイル動画として抽出する手法．

被験者には指定の URL にアクセスすることを依頼し，後の説明についてはすべて実験システム上で行う．被験者の実験における手続きは下記のとおりである．

(1) 被験者はまずウェブブラウザを利用して指定の実験ページの URL にアクセスする．
(2) ページにはユーザ名入力ボックスと，送信ボタン，実験に関する各種の説明と，実験には終わりがなくいつやめてもよいという説明がある．また，いつでも再開できることを記述している．被験者はユーザ名を入力し，送信ボタンを押すことで実験を開始（または中断後の再開を）する．
(3) 実験システムは処理済みの楽曲動画集合の中から無作為に 1 つの楽曲動画 ID を選択し，その楽曲動画に関して 6 手法それぞれについて生成されたサムネイル動画の URL を取得する．次に，この 6 手法で生成されたサムネイル動画を無作為に並べ替え，ユーザに提示する（オリジナルの楽曲動画や，どのサムネイル動画がどの手法で作成されたかなどの情報は提示しない）．また，このときにそれぞれのサムネイル動画の横に評価のためのラジオボタン（7 段階のリッカート尺度）を用意する．なお，ここでサムネイル動画の作成目的についても説明を行う．
(4) 被験者は提示されたサムネイル動画上の再生ボタンを押すことでそのサムネイル動画を視聴し，その後，そのサムネイル動画について 7 段階のリッカート尺度で評価を行う．
(5) 実験システムは，被験者がすべてのサムネイル動画を評価し終えたことを検知すると，「次へ」のボタンを押すことができるようにボタンの挙動を変更する．被験者が「次へ」のボタンを押すと (3) へと戻る．被験者は，自身のタイミングで実験を終了することが可能となっている．

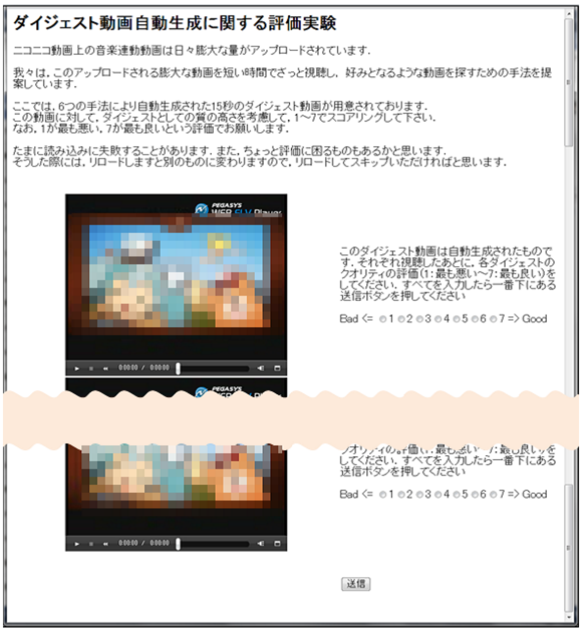


図 4　実験システム
Fig. 4　The evaluation system.

以上のプロセスにおいて，ユーザ名をハッシュ化したものをユーザ ID として保存し，評価の投稿日時，評価された楽曲動画 ID，それぞれの手法に対する評価スコアをログとして記録した．

実験のウェブインタフェースは**図 4** のとおりである．ページ上部に実験の目的が書いてあり，サムネイル動画を再生するプレイヤとそのサムネイル動画を評価するためのラジオボタンが左右に並んだものが，縦に 6 手法分並んでいる．また，ページ下部に送信ボタンが用意されている．なお，楽曲動画に対するタイトルや説明文などは表示しておらず，ニコニコ動画上で投稿者によって設定されているサムネイル画像（静止画）のみをここでは表示している．

3.2　実験結果

実験は，口頭またはメールによって第 1 著者の所属している研究室の学生および職員と，関連研究室の学生に依頼した．実験参加に関する報酬はなしとした．

評価結果のログによると実験の参加者は 12 人であった．1 つの楽曲動画を評価するには少なくとも 90 秒（1 手法あたり 15 秒 × 6 手法）の時間がかかるため，前処理として同一ユーザ ID が連続評価しているもので 90 秒以下のものを除こうとしたが，該当する評価結果はなかった．

評価した楽曲動画件数には被験者によってばらつきがあり，10 以下（合計 6 手法で 60 サムネイル動画以下）の被

表 1　評価実験の全結果の平均と分散
Table 1　Average and variance of evaluation results by each method.

手法	スコア平均	分散
middle	3.88	2.75
comment	3.31	3.77
sentiment	3.59	3.14
sabi	4.17	2.29
sabi*sentiment	4.12	2.84
sabi+sentiment	4.38	2.91

験者が 7 人で，29 以上（合計 6 手法で 174 サムネイル動画以上）の被験者が 5 人と，実験に対する取り組み方（参与度）に大きな違いが出た．10 本以下を評価した被験者については，1 本や 2 本の動画しか評価していない被験者も多数いた（1 本のみを評価した被験者が 4 人，2 本のみが 1 人など）．参与度の低い被験者はいろいろなパターンを試しておらず，評価が安定していないと考えたためある一定以上の楽曲動画（ここでは 29 本以上）に対して評価している被験者のみを評価の対象とした．なお，29 本以上評価している被験者は，少ない人で 1 日，多い人で 5 日程度をかけ評価を行っていた．以上の手続きにより，209 個の動画（1,254 個のサムネイル動画）を対象とした．以下では実験結果を示す．

実験結果（手法ごとのスコア平均，スコア分散）は**表 1** のとおりである．評価実験の結果より，まず **comment** 手法の精度が最も悪く，**middle** 手法や **sentiment** 手法より下回っていることが分かる．また，**sabi+sentiment** 手法がスコア平均としては一番いい結果になっていることが分かる．さらに，**sabi** 手法はスコア平均は最高にはならなかったものの，分散が低く同じようなスコアで安定していたことが分かる．**sabi*sentiment** 手法は，**sabi+sentiment** 手法より悪く，**sabi** 手法と比較しても差がないか，やや悪かったことが分かる．

図 5 は，各スコアについて手法がどのように分布していたのかを示している．この図からも，**comment** 手法が特に悪いことが分かる．また，**sabi** 手法と，**sabi+sentiment** 手法を比較すると，**sabi** 手法は評価値 4（中央値）においてピークが来ており，**sabi+sentiment** 手法は評価値 5（やや良い）にピークが来ていることが分かる．特に，**sabi** 手法は，7（最も良い）と評価されている数が最も少なく，**sabi+sentiment** 手法や **sabi*sentiment** 手法の半数でしかないことが分かる．なお，**sabi** 手法は，1（最も悪い）と評価されている数も最も少なく，ここからも **sabi** 手法が比較的安定していることが分かる．

各被験者がどの手法を最も高く評価していたかを調べるため，被験者ごとに手法の評価平均をとったものが**図 6** である．いずれの被験者においても，わずかながら **sabi+sentiment** 手法が最も良いスコアになっていることが分かる．

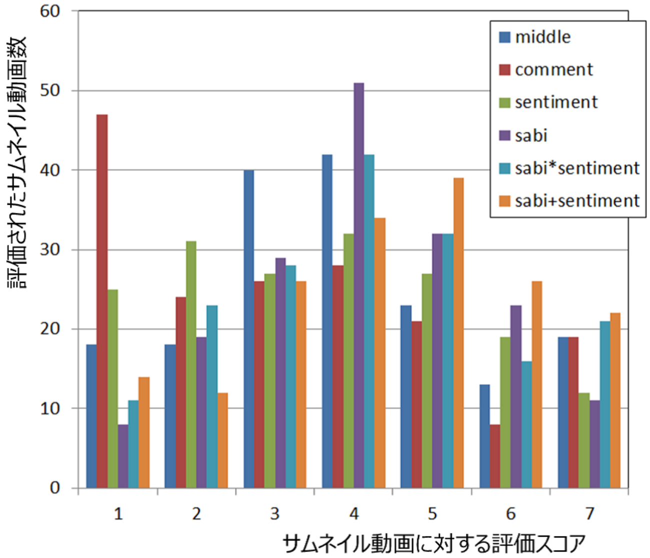


図 5　実験結果（スコアごとの数）
Fig. 5　The number of evaluated score in each method.

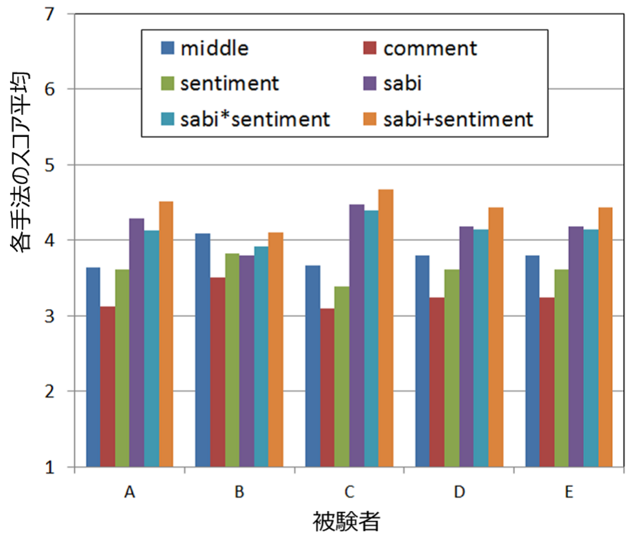


図 6　実験結果（被験者ごと）
Fig. 6　Comparison of each subjects' evaluation.

3.3　sabi 手法と sabi+sentiment 手法の開始時間の比較

音響特徴分析と視聴者反応分析を組み合わせた場合に，どういった特性の違いが出るかを明らかにするため，ここでは sabi 手法と sabi+sentiment 手法の詳しい比較を行う．ここでは，

$$\delta = (\text{sabi 手法開始秒}) - (\text{sabi+sentiment 手法開始秒})$$

としたときの，各手法のスコア平均を

- 開始時間の差が前後 2 秒ずれる程度であれば両者の間にはそれほど差がない，
- サムネイル動画の長さが 15 秒であるため，開始時間の差が前後 15 秒以内に収まっていれば，両者間にオーバラップがある，

ということを考え，**表 2** のように開始時間の差に基づき整理した．

表 2 sabi 手法と sabi+sentiment 手法の比較

Table 2 Comparing sabi+sentiment method with sabi method depending on the starting time of thumbnail video clip.

開始時間の差	動画数	sabi 手法	sabi+sentiment 手法
$\delta \leq -15$	62	4.40	4.37
$-15 \leq \delta < -2$	50	3.90	**4.69**
$-2 \leq \delta \leq 2$	48	4.06	4.04
$2 < \delta < 15$	7	4.00	4.00
$15 \leq \delta$	42	4.26	**4.55**

表 3 sabi 手法と sabi+sentiment 手法の詳細な比較

Table 3 Comparing sabi+sentiment method with sabi method depending on the starting time of thumbnail video clip.

開始時間の差	動画数	sabi 手法	sabi+sentiment 手法
$\delta < -25$	50	4.50	4.20
$-25 \leq \delta < -15$	11	4.09	**4.91**
$-15 \leq \delta < -10$	12	4.33	4.50
$-10 \leq \delta < -5$	19	3.63	**5.11**
$-5 \leq \delta < 0$	37	4.03	4.30

この結果より，$-15 \leq \delta \leq -2$ および $15 < \delta$ において **sabi+sentiment** 手法が **sabi** 手法を上回っていることが分かる．この点についてさらに分析を行うため，**sabi+sentiment** 手法が **sabi** 手法より後の部分を抽出している結果のスコアの平均を計算したものが**表 3** である．

この結果より，**sabi** 手法より 5〜10 秒程度前方を **sabi+sentiment** 手法がサムネイル動画として抽出している場合に良い結果になっていることが分かる．

4. 考察

評価実験の結果より，視聴者反応と音響特徴を利用したサビ検出の組合せである **sabi+sentiment** 手法が最も良い結果を示していた．また，視聴者反応によって最も良いサビを探すことを目的としている **sabi*sentiment** 手法はあまり効果的でないことが分かった．なお，表 1 や図 5 によると，**sabi** 手法は平均的で最低の評価は付けられにくいものの，最高の評価も付けられないことが分かる．サムネイル動画でその魅力を伝えるには，最高の評価または最高に近い評価が重要になってくると考えられる．そうした点で，**sabi+sentiment** 手法は可能性を秘めていると考えられる．**comment** 手法はスコアが最も悪い手法となっていたが，これは楽曲動画の最初と最後にコメントが集中する傾向が高く，そうしたシーンが抽出されてしまったことに起因している．実際，抽出されたサムネイル動画を確認すると，黒背景でスタッフスクロールのようなものが提示されていたり，楽曲のイントロで終わってしまったりしているものが多かった．

sabi 手法と，**sabi+sentiment** 手法との開始時間の比較（表 2 および 3）は，**sabi+sentiment** 手法がどういう状況において **sabi** 手法より良い結果になるかということを示唆している．特に，**sabi+sentiment** 手法が，**sabi** 手法より 5〜10 秒前の部分からサムネイル動画として抽出した場合に **sabi+sentiment** 手法の効果が現れているといえる．ここで，**sabi** 手法はサビの開始場所からをサムネイル動画として抽出しているため，**sabi+sentiment** 手法はサビの開始場所よりやや前方を抽出していることになる．このことから，サムネイル動画としては，サビの先頭から開始するものではなく，あるメロディラインからサビへと入るようなものを被験者が高く評価したということが分かる．実際に生成されたサムネイル動画を視聴したところ，そうした変化が心地良いものが多かった．また，サビに入る部分で映像がドラスティックに変化するものがあり，そうした変化も被験者が高く評価した部分であると考えられる．今後は，こうした点について詳細な比較分析を行っていく予定である．

表 2 によると，**sabi+sentiment** 手法が，**sabi** 手法より後の部分をサムネイル動画として抽出したときに高く評価される傾向も見て取れる．この点について実際に生成されたサムネイル動画を視聴することによって比較を行った．その中でも特徴的だったものは，**sabi** 手法と **sabi+sentiment** 手法でそれぞれ音楽的に類似した部分を抽出していても，**sabi+sentiment** 手法で抽出していた部分が，より視聴者に対して訴えかけるものであったという点である．**図 7** は，特に被験者が評価に差をつけた動画のそれぞれのシーンである．視聴者にとって訴えかけられるものであり，その点が高く評価されたと考えられる．今回，視聴者反応と音響特徴量しかサムネイル動画の自動生成に使っていないが，視聴者反応の大きさからその視聴者反応につながる映像特徴量を分析することによって，アピール度の高いシーン特有の映像特徴量がどのようなものかを明らかにできると期待している．そこで，今後は視聴者反応と音響特徴量に加え，映像特徴量の 3 者を組み合わせた手法を実現予定である．

今回，**sabi+sentiment** 手法が最も良い結果を示したものの，**sabi** 手法と比較して大きな差が見受けられなかった理由としては，下記が考えられる．

- **sabi** 手法と **sabi+sentiment** 手法が同じ結果であることが多く，差が被験者にとって分かりにくいものが多々あった．
- サビ検出手法が検出した結果のうち，サビとして弱いと判定された部分が楽曲の序盤や終盤に検出されており，そこに動画に対する期待や，動画投稿に対する感謝のコメントが集中していた．このときに，笑いに関するコメントも一定数以上現れていたため，この部分がサムネイル動画用に抽出されてしまった．

上記の問題のうち，前者は正規化をそれぞれの最大値で

図 7　sabi 手法と sabi+sentiment 手法で生成されたサムネイル動画のシーン比較

Fig. 7　Examples of scenes of each thumbnail video clip generated by sabi method and sabi+sentiment method.

表 4　評価実験のスコア平均の比較

Table 4　Comparison of the average of evaluation score in each method.

手法	3 手法独立の 64 件	重複がある 145 件
middle	3.70	3.96
comment	3.36	3.29
sentiment	3.69	3.54
sabi	4.19	4.16
sabi*sentiment	4.30	4.04
sabi+sentiment	**4.63**	4.28

行っていたため，差がでないことがあったということが考えられる．この点については，最大値で正規化するのではなく，総量で正規化する方法を今後検討予定である．ここで，2 手法間で生成したサムネイル動画にほぼ差がなかった（差が 1 秒以内）ものが，**sabi** 手法と **sabi+sentiment** 手法間では 46 件，**sabi** 手法と **sabi*sentiment** 手法間では 117 件，**sabi*sentiment** 手法と **sabi+sentiment** 手法間では 52 件であった．サビ検出に関する 3 手法のうち，いずれかの 2 手法間で生成したサムネイル動画がほぼ同じ（1 秒以内の差）は 145 件となっていた．そこで，サビに関する 3 手法の結果が異なっていた残りの 64 件の結果について，それぞれの手法の平均を計算したものが**表 4** である．この結果より，同じような動画が出力されない場合は，**sabi** 手法に比べ，**sabi+sentiment** や **sabi*sentiment** 手法の評価が高くなることが分かる．

一方，後者の問題は，コメントからの感情抽出に問題があると考えられる．今後は，感情の分類および感情判定辞書を精査し，精度向上を目指す予定である．また，今回は音響分析においてはサビ検出しか行っていないが，A メロや B メロ，コード進行などの情報を利用したより深い音響分析についても導入し，検討を行う予定である．

ベースラインとして考えていた **middle** 手法が悪くなく，**comment** 手法より良い結果を示したのは驚きであった．詳しく調べてみたところ，**middle** 手法が高く評価されていたのは，**comment** 手法や **sentiment** 手法が最初や最後などの音楽さえないシーンをサムネイル動画として出力している場合や，サビに関する 3 手法が類似した結果を出していて，**middle** 手法がまったく異なるシーンを抽出していた場合などであった．実際，先述の表 4 の結果は，**sabi** 手法，**sabi*sentiment** 手法，**sabi+sentiment** 手法のいずれか 2 手法以上が同じ結果である場合に，**middle** 手法が高く評価されたことを示唆している．

sentiment 手法は，サビが検出できないような楽曲動画や，そもそも音楽が主体ではない動画などで効果的に働くと期待される．**sabi** 手法はコメントがなくてもサムネイル動画を自動生成できるため，**sentiment** 手法と，**sabi** 手法は相補関係にある．しかし，コメントがほとんどなく，サビの検出に失敗することも考えられる．そうした際には，映像解析を行い，どのような映像特徴があるときに視聴者が盛り上がる傾向があるか，またどのような音響特徴になったときに視聴者が盛り上がる傾向があるかということを，これまでに高く評価された動画をもとに学習し，サムネイル動画自動生成において利用することが考えられる．

今回実現した手法は楽曲動画であったため，サビ検出を中心とした手法が有効に働いた．一方，サビが存在しない，またはサビ検出が困難なストーリー性のある動画の場合，サビ検出手法をサムネイル動画自動生成に利用することができない．サビがそもそも存在しない動画においては，**sentiment** 手法が効果的に働く可能性がある．こうした動画からのサムネイル動画自動生成手法については，今後検討予定である．

今回の研究では正規表現辞書を用いて感情の推定を行った．ここで，Twitter での短い投稿文にもニコニコ動画などで投稿されるコメントに似たような自然言語処理困難な書き込みが多くあることが知られており，Brody ら [2] は，「Cooooooooooollllll!!!!!!!!」などのような投稿から，その投稿文が肯定か否定かを認識する方法を提案している．感情の度合いはキーボードを押していた時間（ボタンを押し下げたまま何文字も入力している状態）に比例すると考えられる．そこで，「www」ではなく「wwwwwwww」をより笑いの強い表現として，「すげえ」より「すげえええええええ」をより驚きの強い表現として取得する手法に取り組む予定である．

ニコニコ動画に関するコミュニティでは，タグベースで動画集合を決め，1日あたりの再生数の増加率やコメントの増加率，お気に入り（マイリスト）の増加数に基づいてランキングスコアを決め，自動的にデイリーランキング動画を生成するソフトウェアが開発されている [13]．こうした自動ランキング生成ソフトウェアを利用して，VOCALOIDに関する動画のランキングは2008年2月よりアップロードされるようになり，動画の発掘に役立つため人気を集めている（2012年9月12日時点で，1,675個の動画が投稿されている）．なお，このソフトウェアで切り出す部分は，60秒から80秒程度と決め打ちとなっている．どの部分を切り出すかという点において，我々の手法は効果的であると考えられる．

サムネイル動画の応用範囲は上記のようなランキング自動生成以外でも，検索結果のランキングに利用することも考えられる．ランキングの結果では単に静止画が提示されているが，この画像をサムネイル動画に差し替えれば，マウスをホバーすることで手軽にサムネイル動画を視聴してオリジナル動画を把握できる．また，動画の推薦を行う際にテキストやサムネイル画像のみを表示するのではなく，短時間で視聴可能なサムネイル動画を提示すると，その動画を見るかどうかの判断に有効に働くであろう．さらに，Songrium [5] などのようなサービスでは，楽曲動画を最初から再生することが多いが，短時間で把握するには最も良い部分を再生するべきであり，そうした際にも本手法は有効である（なお，Songriumではユーザの待ち時間を短くするために意図的に最初から再生しており，搭載されたサビ出し機能を利用することにより，すぐにサビから聴くことが可能である）．

一方，今回明らかになったサビの開始位置より少し前からをサムネイル動画として抽出することが有効であることが多かったという点について，今後はどういった視聴者反応のときに，サビの何秒前から再生されると高く評価されるのかといったことを明らかにする予定である．上記の点を明らかにできると，膨大な楽曲動画群を分析することにより，どのような音響特徴，映像特徴の場合にサムネイル動画が高く評価されるのかという点を明らかにできると期待される．また，こうした技術を推し進めることにより，楽曲動画の各シーンに対するダイレクトな視聴者反応が存在しないYouTube上の楽曲動画のサムネイルの自動生成も行えると期待される．

5. 関連研究

Nakamuraら [11] はニコニコ動画の動画に対して付与されたコメントから「喜び」,「悲しみ」などの印象情報を抽出し，インデックスを作成することで，印象に基づく動画検索およびランキングを可能としている．また，印象の時間的遷移を可視化することによって任意の印象に関するシーンでの頭出しを可能にしている．しかし，サムネイル動画などダイジェストを生成することは考慮していない．

Miyamoriら [9] は，テレビ番組を視聴中に実況チャットシステムに集まって番組に関するチャットを行うことに注目し，テレビ番組のダイジェストを生成するためのインデックス生成手法を提案し，インデックスを利用した閲覧システムを実装している．ここでは，テレビ番組に連動して投稿されたチャットコメントから喜び，悲しみといったセンチメントを抽出し，そのセンチメントの時間的変化によってインデックスを作っている．また，その度合いに基づいて，シーンを選んで視聴することを可能としている．ただ，スポーツ番組の重要なシーン検出において実況チャットから抽出されるセンチメントが有効であることを示しているが，楽曲動画などには取り組んでいない．

青木ら [1] は，ニコニコ動画の動画に対して付与されるコメントの出現頻度を用いて動画内で最も重要な箇所の判別やサビの検出，映像の要約などを試みている．この手法では単純に再生時間あたりのコメント数を利用しているだけであり，コメントの感情分析などには踏み込んでいない．なお，今回提案手法との比較に利用したベースライン手法2は青木らの手法に近いコメントの数のみを利用するものである．

楽曲動画が対象ではないが，楽曲に対する音響的な分析に基づいて，楽曲の代表的な部分（サムネイルに相当）を1カ所音響信号として切り出す手法 [17], [18], [19] も提案されている．しかし，視聴者のコメントは考慮されていなかった．一方，視聴者のコメントではなく再生履歴に基づいて，各楽曲中で不特定多数の視聴者が多く再生している区間（再生頻度が高い区間）をサビ区間と見なして検出する手法 [20] も提案されている．

動画の要約の自動生成に関する取り組み [10], [16] は様々になされており，先述のような動画に投稿されたコメントを利用するものだけでなく，各種のアノテーションに基づくものや映像と音声の一貫性，映画文法を使うものなど様々である．しかし，動画の要約と，サムネイル動画の生成は本質的に異なる．動画の要約は，動画全体のトピックの網羅性を高めることを目的とすることが多いが，サムネイル動画の自動生成の場合は最も良いシーンを探すという，適合率を向上させることを目的とするものである．また，音楽の音響分析と視聴者のコメントを利用したようなものは存在しない．

Ueharaら [15] は，テレビ番組の実況チャットから，役者名を抽出し，その役者名が現れる時間はその役者が登場しているという仮説のもとに，役者の登場シーンをインデックス化し，役者名を検索キーワードとしたシーン検索を可能としている．佃ら [14] は，テレビ番組の実況チャットではなく，ニコニコ動画上の動画に対して投稿されたコメントからその人物の登場の度合いを推定し，登場の度合いの

時間的変化から人物の活躍度を推定することによって，あらたな動画検索を可能にする仕組みを提案および実装している．我々が今回取り組んでいる楽曲動画のサムネイル動画自動生成においては，登場人物の理解やシーンの理解はそれほど重要ではないため今回はこうした点には取り組んでいないが，ストーリー性のある動画などでは登場人物理解などが重要になると考えている．そうした点については，今後取り組む予定である．

6. おわりに

本稿では，動画共有ウェブサイトにおける楽曲動画に対し，短い時間でオリジナル動画の魅力を伝え，見るかどうかを判断することを可能にするようなサムネイル動画自動生成手法を提案した．また，音響分析によるサビ検出と，視聴者反応分析を組み合わせた手法が最も良い性能を示すことを明らかにした．

今回の実験はサムネイル動画の質を評価してもらっているだけであり，サムネイル動画を視聴してから，オリジナル動画を視聴した際にどのようなユーザの心境の変化があるかなどを調べることはできていない．今後はユーザがサムネイル動画を見た後と，オリジナル動画を見た後にそれぞれ評価してもらい，サムネイル動画とオリジナル動画にどの程度のギャップがあったかを検証する予定である．

本研究はサムネイル動画自動生成の第1歩である．楽曲動画とストーリー性のある動画では状況は異なってくると考えられる．また，今回は映像を考慮していないが，映像からもサムネイル動画として抽出すべき特徴量を取得できる可能性がある．今後は，こうした新たな取り組みも行っていく予定である．

なお，最も盛り上がるシーンが必ずしもサムネイル動画としてつねにふさわしいとは限らない．たとえば，ストーリー性のあるような動画の場合，最も盛り上がるのは視聴者の意表をつくようなシーンかもしれない．そうした部分のみを提示されるとネタバレになってしまいオリジナル動画視聴の楽しみが減退してしまうと考えられる [12]．今後は，サムネイル動画生成時のネタバレ防止についても取り組む予定である．

謝辞　本研究の一部は，文部科学省科学研究費補助金若手研究（A）（研究代表者：中村聡史，課題番号 23680006）と，科学技術振興機構 OngaCREST プロジェクトによるものです．ここに記して謝意を表します．

参考文献

[1] 青木秀憲，宮下芳明：ニコニコ動画における映像要約とサビ検出の試み，情報処理学会研究報告，2008-HCI-128/2008-MUS-75，Vol.2008, No.50, pp.37–42 (2008).

[2] Brody, S. and Diakopoulos, N.: Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.562–570 (2011).

[3] 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).

[4] 後藤真孝：初音ミク，ニコニコ動画，ピアプロが切り拓いた CGM 現象，情報処理（情報処理学会誌），Vol.53, No.5, pp.466–471 (2012).

[5] 濱崎雅弘，後藤真孝：Songrium：多様な関係性に基づく音楽視聴支援サービス，情報処理学会研究報告，Vol.2012-MUS-96, No.1 (2012).

[6] 濱崎雅弘，武田英明，西村拓一：動画共有サイトにおける大規模な協調的創造活動の創発のネットワーク分析—ニコニコ動画における初音ミク動画コミュニティを対象として，人工知能学会論文誌，Vol.25, No.1, pp.157–167 (2010).

[7] 濱野智史：アーキテクチャの生態系—情報環境はいかに設計されてきたか，NTT 出版 (2008).

[8] 剣持秀紀：歌声合成の過去・現在・未来：「使える」歌声合成のためには，情報処理（情報処理学会誌），Vol.53, No.5, pp.472–476 (2012).

[9] Miyamori, H., Nakamura, S. and Tanaka, K: Generation of Views of TV Content Using TV Viewers' Perspectives Expressed in Live Chats on the Web, *Proc. ACM Multimedia2005*, pp.853–861 (2005).

[10] 森山　剛，坂内正夫：ドラマ映像の心理的内容に基づいた要約映像の生成，電子情報通信学会論文誌 D，Vol.J84-D2, No.6, pp.1122–1131 (2001).

[11] Nakamura, S. and Tanaka, K.: Video Search by Impression Extracted from Social Annotation, *Proc. WISE2009*, pp.401–414 (2009).

[12] 中村聡史，小松孝徳：スポーツの勝敗にまつわるネタバレ防止手法の検討，情報処理学会インタラクション 2012, pp.9–16 (2012).

[13] rankingloid，日刊 VOCALOID ランキング：ランキング動画全自動生成システムとその運用，情報処理学会研究報告音楽情報科学（SIGMUS），2012-MUS-94, Vol.5, pp.1–6 (2012).

[14] 佃　洸摂，中村聡史，山本岳洋，田中克己：ソーシャルアノテーションに基づく動画の登場人物の重要度の推定，*Proc. WebDB Forum 2010* (2010).

[15] Uehara, H. and Yoshida, K.: Annotating TV drama based on viewer dialogue-analysis of viewers' attention generated on an Internet bulletin board, *Proc. SAINT2005*, pp.334–340 (2005).

[16] 吉高淳夫，松井亮治，平嶋　宗：カメラワークを利用した感性情報の抽出，情報処理学会論文誌，Vol.47, No.6, pp.1696–1707 (2006).

[17] Logan, B. and Chu, S.: Music Summarization Using Key Phrases, *Proc. ICASSP 2000*, pp.749–752 (2000).

[18] Bartsch, M.A. and Wakefield, G.H.: To Catch A Chorus: Using Chroma-based Representations for Audio Thumbnailing, *Proc. WASPAA '01*, pp.15–18 (2001).

[19] Cooper, M. and Foote, J.: Automatic Music Summarization via Similarity Analysis, *Proc. ISMIR 2002*, pp.81–85 (2002).

[20] 後藤真孝：楽曲再生方法及び装置，特許第 4273202 号 (2003).

中村 聡史　（正会員）

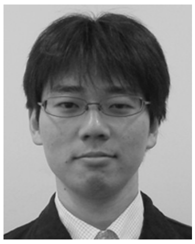
1976 年生．2004 年大阪大学大学院工学研究科博士後期課程修了．同年独立行政法人情報通信研究機構専攻研究員．2006 年京都大学大学院情報学研究科特任助手，2009 年同特定准教授，2013 年明治大学総合数理学部准教授，現在に至る．サーチとインタラクションや，情報曖昧化技術，ソーシャルアノテーション分析等の研究活動に従事．ヒューマンインタフェース学会等各会員．博士（工学）．

山本 岳洋　（正会員）

1984 年生．2011 年京都大学大学院情報学研究科博士後期課程修了．同年日本学術振興会特別研究員（PD），2012 年京都大学大学院情報学研究科特定研究員，現在に至る．博士（情報学）．情報検索，特に情報検索におけるユーザインタラクションに関する研究に従事．日本データベース学会会員．

後藤 真孝　（正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．同年電子技術総合研究所に入所し，2001 年に改組された産業技術総合研究所において，現在，情報技術研究部門首席研究員兼メディアインタラクション研究グループ長．統計数理研究所客員教授，筑波大学大学院准教授（連携大学院），IPA 未踏 IT 人材発掘・育成事業プロジェクトマネージャーを兼任．ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，科学技術分野の文部科学大臣表彰若手科学者賞，情報処理学会長尾真記念特別賞等，31 件受賞．



濱崎 雅弘　（正会員）

2000 年同志社大学工学部知識工学科卒業．2002 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2005 年総合研究大学院大学数物科学研究科博士後期課程修了．博士（情報学）．同年より，産業技術総合研究所情報技術研究部門勤務．オンラインコミュニティや知識共有に関する研究に従事．人のネットワークを活用した情報システムに興味がある．人工知能学会，ACM 各会員．

（担当編集委員　上田 真由美）