E-041

演奏動画の盛り上がり検出に用いる特徴量の検討

A Study on Feature Extraction for Highlights Detection from Musical Performance Videos

小山 健一† 石先 広海‡ 帆足 啓一郎‡ 小野 智弘‡ 甲藤 二郎†
Kenichi Koyama Hiromi Ishizaki Keiichiro Hoashi Chihiro Ono Jiro Katto

1. はじめに

近年，動画共有サイトが普及し，ユーザ自身の楽器演奏を披露する動画(以下，演奏動画と呼ぶ)の投稿が増えている．そして演奏動画を素材として捉え，分割画面に同一曲の演奏動画を合成することであたかも合奏しているかのように見せる動画(以下，合奏動画と呼ぶ)が作られることがある(図 1)．これは，既存のコンテンツからより印象深いコンテンツを再生産する CGM の一例と言える．

図 1　合奏動画の様子

従来の合奏動画は分割画面のままで動画の演出がないものが多い．一方で，一般のライブ映像は，演奏に合わせたズームアップやカメラ転換といった演出が付与されている．そのためコンテンツの完成度が高い．合奏動画にも同様の演出を付与すれば完成度が高められると考えられるが，演出が付与されたものは少ない．その原因の一つに，動画編集経験の少ない一般のユーザにとって演出を付与すべき箇所を決定するのが難しく，時間がかかる点が挙げられる．

そこで本研究では，Web 上の演奏動画から演出が付与された合奏動画を自動合成するシステムの実現を目指す．このシステムにより，ユーザはより印象深いコンテンツを手軽に制作することができるようになる．そのために本稿では，まず演奏動画の盛り上がりを定義する．つぎに，盛り上がりの定義を反映する動画の特徴量を設定する．最後に，盛り上がり検出における特徴量の有効性を実験により検証し，合奏動画自動合成に向けた展望を述べる．

2. 関連研究

演奏動画の盛り上がり箇所は動画の重要箇所と言い換えることができる．重要箇所を抽出するという観点では，本研究の関連研究として動画要約が挙げられる．

料理番組の自動要約を行う研究[1]では，重要箇所として調理中の人間の繰り返し動作を挙げている．そして繰り返し動作をオプティカルフローによって検出し，繋ぎ合わせることで要約を試みている．ニュース番組の自動要約を行う研究[2]では，ニュース番組の一般的な構成を利用し，重要箇所である番組中の各話題の冒頭部を検出することで要約を試みている．ミュージックビデオの自動要約を行う研究[3]では，コーラス箇所，ズームアップ箇所，歌詞の繰り返し箇所が重要箇所であり，これらを検出することで要約を試みている．また，楽曲の重要箇所検出という観点ではサビ検出の従来研究[4]があるが，この研究における重要箇所は楽曲のサビ部分であり，楽曲で何度も繰り返されるフレーズを検出することでサビ検出を実現している．

以上より，重要箇所の定義がコンテンツの種類によって異なることが分かる．そのため従来手法は演奏動画に対してそのまま適用することができない．また，演奏動画における重要箇所，すなわち盛り上がり箇所を特定する上で着目すべき要素は従来研究で明らかになっていない．したがって，演奏動画における盛り上がり箇所の定義と，その箇所の検出に適した特徴量を検討することが必要である．

3. 演奏動画の盛り上がり

以上をふまえて，3.1 項では演奏動画における盛り上がりを定義し，3.2 項では定義に基づいて必要と思われる特徴量を挙げる．

3.1 盛り上がりの定義

定義にあたって，あらかじめ学生 11 名に対してアンケートを取った．回答者は同一の演奏動画を視聴した後に「演奏動画における盛り上がりとはどんな箇所か」という設問に自由記述形式で答えた．このアンケートを参考にしつつ，以下に挙げる 3 つの箇所を盛り上がりと定義した．

・難しい演奏をしている箇所

一般のライブ映像において，ソロなどの難しい演奏をしている箇所は観客が歓声を上げることが多く，視聴者が注目する箇所といえる．またアンケートで 7 名がこのような箇所を盛り上がりだと感じると答えた．よって，難しい演奏をしている箇所を盛り上がりと定義する．

・奏者の身体の動きが激しい箇所

奏者の身体は曲のリズムに同期して動くことが多く，この動きにより音楽的な盛り上がりが表現されることがある．また，アンケートで 7 名が奏者の身体の動きが激しい箇所を盛り上がりだと感じると答えた．よって，奏者の身体の動きが激しい箇所を盛り上がりと定義する．

・音が大きくなった箇所

一般的に，クレシェンドする(音量を徐々に大きくしていく)箇所は演奏の最も盛り上がる箇所の直前に置かれ，盛り上がりの高まりを期待させることが多い．よって，音が大きくなった箇所を盛り上がりと定義する．

3.2 利用する特徴量

3.1 の定義に基づき，以下の 3 つの特徴量を盛り上がり箇所の抽出のために用いる．なお，特徴量はすべて 0～1

† 早稲田大学基幹理工学研究科　Waseda University
‡ 株式会社　KDDI 研究所　KDDI R&D Laboratories, Inc.

の値を取るように正規化した後，細かい変化を除去するため 2 項フィルタによる平滑化を行う．

・オンセット数

一般に，音符の密度が高い演奏は手を細かく動かす必要があるため，難しい演奏といえる．そこで音符数と比例関係にあるオンセット数を測定する．オンセットは演奏音の音響信号の立ち上がりを検出することで得られ，1 秒間あたりのオンセット数を特徴量として利用する．

・動き

ブロックマッチング法により動画像の 30 フレーム間のオプティカルフローを取り，動きベクトルの大きさの総和を求める．これを 1 秒間の動きの変化量とすることで奏者の身体の動きが激しい箇所を検出する．

・音量 RMS

演奏音の音響信号の 1 秒間の RMS(二乗平均平方根)を取ることで 1 秒間の音量の変化を測定し，音が大きくなった箇所を検出する．

4. 予備実験

4.1 実験内容

各特徴量が盛り上がりの定義を反映しているか検証するために，人が感じる演奏動画中の盛り上がり箇所を調査する予備実験を行った．実験では，被験者は演奏動画を視聴しながら，演奏の盛り上がりに対する気分の高まりを表す値として「高揚度」を記録する．そして高揚度と各特徴量の変化の様子を比較し，考察を行う．なお，被験者は学生 5 名であり，1 人につき 4 つの演奏動画を視聴し，高揚度を記録する．

4.2 実験手順

あらかじめ，実験用に動画視聴アプリを自作した．このアプリは左側に動画を表示し，右側には上下に動くスライドバーを備えている．被験者はこのアプリを用いて演奏動画を視聴する．被験者は視聴している動画が盛り上がってきたと感じたら右のバーを上に，落ち着いてきたと感じたら下に移動させて，高揚度を指定する．この作業を 4 つの演奏動画 A～D について行う．各被験者の高揚度は 0～1 の値を取るように正規化される．そして被験者の高揚度の平均値を 2 項フィルタにより平滑化したものを特徴量との比較に用いる．なお，演奏動画 A～D は，国内の動画共有サイトであるニコニコ動画[5]で収集した動画である． A から順にベース，ドラム，ギター，キーボードの演奏動画であり，全て同一の楽曲を演奏している．

4.3 実験結果

4.1 項で述べた高揚度と，動画 A～D から抽出した各特徴量の変化を比較する．具体的には，高揚度と各特徴量間の相関係数を求めた．一例として，動画 A について，相関係数を表 1 に，高揚度と各特徴量の関係を図 2 に示す．

音量 RMS は表 1 より高揚度と中程度の相関があること，また図 2 より音量 RMS は高揚度の変化に追随して同様の箇所で変化していることが多い．そのため，音量 RMS は盛り上がりを表現するために有効な特徴量と言える．またオンセット数についても同様に中程度の相関が見られ，図 2 より 20 秒,40 秒,230 秒付近で高揚度の大きな変化に追随しているので有効な特徴量と言える．

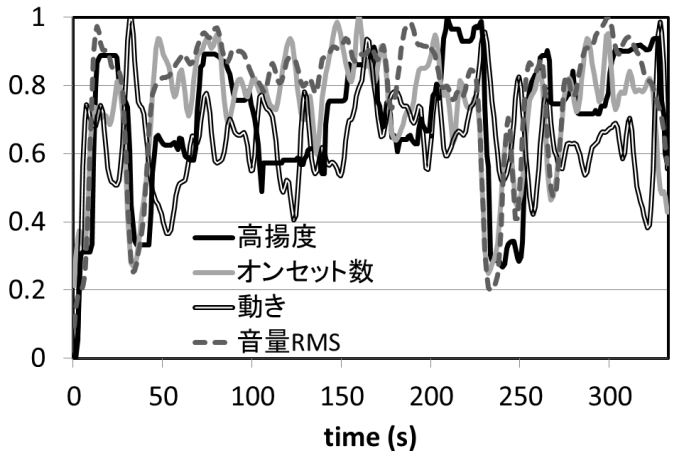

図 2　高揚度と各特徴量の関係

表 1　高揚度と各特徴量間の相関係数

	音量 RMS	動き	オンセット数
高揚度	0.643	0.186	0.564

一方で動きについては，表 1 より，高揚度とほとんど相関がないと言える．原因としては，奏者の身体の動きが少なくても，手元は細かく動かして演奏する場合があることが考えられる．また，実際に動画 A を視聴すると，動きのピークがある 30 秒付近では奏者が演奏しておらず，椅子に座り直す動作が入っていた．この動作が動きのピークになったと考えられる．よって動きと高揚度の相関性を高めるには，盛り上がりと関連した動きのみを抽出する必要があると言える．また，音量 RMS とオンセット数について，60 秒，120 秒付近などで高揚度の変化と対応していない値の動きが見られる．よって，この 2 つの特徴量だけでは高揚度を説明する変数としては不十分であると考えられる．したがって，盛り上がりを反映する別の特徴量の導入が必要である．

5. まとめ

本稿では，演奏動画における盛り上がり箇所を定義した．そして動画の特徴量と実際の盛り上がりの関係を検証した結果，音量 RMS とオンセット数が演奏動画の盛り上がり検出に有効な特徴量であることが示唆された．

今後は合奏動画自動合成に向けて，特徴量の扱い方を改善するとともに，定義を反映する新たな特徴量の導入を検討することで，盛り上がり箇所の検出精度の向上を目指す．

参考文献

[1] 三浦 宏一 , 浜田 玲子 , 井手 一郎 , 坂井 修一 , 田中 英彦, “動きに基づく料理映像の自動要約”, 情報処理学会論文誌 CVIM_7, pp.21-29, 2003.

[2] 工藤 大樹, 西川 博文, 加藤 嘉明, “ニュース番組の要約作成に関する検討”, 電子情報通信学会ソサイエティ大会講演論文集 2006 年_基礎・境界, p.71, 2006.

[3] Changsheng Xu *et al.*, “Automatic music video summarization based on audio-visual-text analysis and alignment”, Proc. of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.361-368, 2005.

[4] 後藤 真孝, “リアルタイム音楽情景記述システム : サビ区間検出手法”, 情報処理学会研究報告. [音楽情報科学] 2002(100), pp.27-34, 2002

[5] ニコニコ動画, http://www.nicovideo.jp/