Vol.2015-MUS-107 No.17
2015/5/23

動画解析・印象推定による動画 BGM の自動生成

清水柚里奈[†1]　菅野沙也[†1]　伊藤貴之[†2]　嵯峨山茂樹[†3]

概要：個人撮影動画を SNS で公開する際に BGM を付与するアプリが増えており，動画共有の楽しみの一環として普及している．本研究では動画特徴量からの印象推定結果に基づいた楽曲生成により，動画の印象に合った楽曲を付与する手法を提案する．まず動画に対して一定時間ごとに色および動きの特徴量を算出し，それに基づいて動画の印象を推定する．また予め用意したメロディとリズム進行についてユーザに印象を回答させ，動画の印象値と類似度が最も高いものを選び出す．これらを合成し，さらにコード進行を付与し，反復回数などを調節することで，動画の長さと印象に合わせた楽曲を生成する．以上の処理により，印象の合った音楽を自分で探すことなく動画に付与できる．

Automatic Background Music Composition Based On Impression Estimation of Input Movies

Yurina Shimizu[†1]　Saya Kanno[†1]
Takayuki Itoh[†2]　Shigeki Sagayama[†3]

Abstract: Recently many people enjoy accompanying background music to the movies in uploading movies in social Web services. Many applications and services to assist the background music editing have been released. This paper presents a technique to automatically generate the background music that matches impression of movies. The technique estimates impression of movies from the feature values of movement and color It then generates the background music by synthesizing melody and rhythm selected based on the impression. The technique learns the relationship between features of movies or music and their impressions answered by the users, so that the music generation process reflects the users' own impression. Users can accompany preferable background music to the movies by this technique, without searching for the tunes by themselves.

1. はじめに

近年，デジタルカメラやスマートフォンの普及により，イベントや旅行先などで気軽に写真や動画を撮影する機会が増えた。そしてその思い出を共有するために，撮影映像に BGM を付与して，Facebook や Twitter，YouTube などの様々な SNS サイトに投稿する人も増えてきた．またそれに伴い，BGM 付与を含む動画編集を支援するアプリも増えてきた．しかしこのような動画編集では一般的に，動画に合った音楽を自分で探したり，動画の長さに合うように音楽を調整したり，といった手間とスキルが必要となる．

そこで本報告では，動画特徴量から動画の印象を推定し，その結果に基づいた楽曲生成を行うことで，印象に合った楽曲を動画に付与する手法を提案する．本手法では，まず動画の印象を推定するために，動画から一定時間ごとに色および動きの特徴量を算出する．また予め用意したメロディとリズム進行についてユーザに印象を回答してもらい，動画の印象値と最も類似度の高いものを選びだし，それらを合成する．さらにコード進行を付与し，反復回数などを調整することで，動画の長さに合った楽曲を生成する．本手法では印象を推定する際に，動画では動画特徴量とそれに対する各ユーザの印象の関係，音楽では音楽特徴量とそれに対する各ユーザの印象の関係を学習させることから，膨大な数の動画・音楽に対して印象を回答してもらうといったユーザの負担を減らすことができ，ユーザ１人１人の動画に対する印象に合った音楽を生成することが可能となる．

2. 関連研究

ビデオに BGM を付与させる研究として，映像の動きと同期する部分を楽曲から抽出し動画へ付与する研究[1]や音楽分析アルゴリズムに基づいてホームビデオの音楽ビデオを自動で生成するシステム[2]などが挙げられる．しかし，[1]では映像の動きだけを考慮して楽曲生成がされており，映像の内容や雰囲気に対しての考慮はされていない．また[2]では動画の内容を予めユーザが指定した上で楽曲生成がなされており，動画解析処理は自動化されていない．

3. 提案手法

本手法は大きく分けて 4 つの処理段階で構成される．具体的には，

(1)動画特徴量：色分布・動き分布の特徴量抽出
(2)音楽特徴量：メロディ・リズムの特徴量抽出
(3)学習：動画，メロディ・リズムの印象の関係性算出_
(4)楽曲生成：ユーザの印象に合った楽曲生成

の 4 段階である．詳細について以下に論述する．

†1 お茶の水女子大学大学院
Ochanomizu University Graduate School
†2 お茶の水女子大学
Ochanomizu University
†3 明治大学
Meiji University

3.1 動画特徴量

現時点での我々の実装では，色分布，動き分布の 2 種類の低レベルな特徴量と印象との関係を学習している．

3.1.1 色分布の特徴量抽出

本処理では動画全体の色分布を数値化する．まず動画から 5 秒ごとに静止画を抽出し，図 1 に示すように，その静止画の各々に対し OpenCV を用いた減色処理を施す．そして各色の画素数を集計することにより，カラーヒストグラムを得る．現時点で我々は色相環や，人の目を惹きやすい誘目性といった色彩心理学の観点から，黒，灰色，白，茶色，赤，オレンジ，黄色，緑，水色，青，ピンク，紫の 12 色に静止画を減色している．得られたそれぞれのヒストグラムの数値から各色の画素数の平均を求め，これを動画全体に対する平均の色の割合とみなし，12 次元の特徴量ベクトルとする．

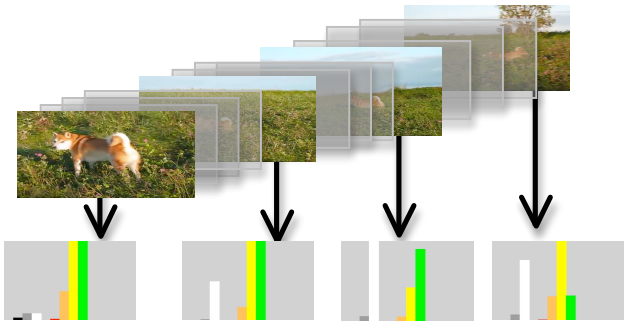

図 1 : 各静止画に対して 12 色のカラーヒストグラム生成

3.1.2 動き分布の特徴量抽出

本処理では動画全体の動き分布を数値化する．まず動画を時間で 4 分割し，各時間帯に対して OpenCV を用いてオプティカルフローを求める．続いて図 2 に示すように，オプティカルフローを構成するベクトル群の速度・角度を集計して各々のヒストグラムを生成し，速度の平均・分散，速度のヒストグラム上で度数が最大となる階級値，角度の分散，角度のヒストグラム上で度数が最大となる階級値を求める．そして各特徴量の全体の平均を求め，これら計 5 つを動きの特徴量とみなす．

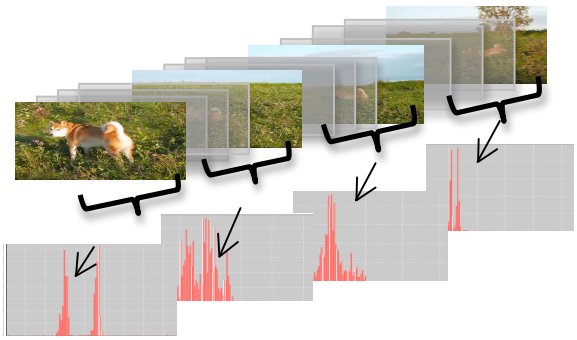

図 2 : 各動画に対して速度ベクトル・角度ベクトルのヒストグラム生成

3.2 音楽特徴量

現時点での我々の実装では，コード進行付きメロディとリズムを別々に素材として用意し，以下の音楽特徴量を算出している．

3.2.1 メロディの特徴量抽出

本手法ではメロディに対して，文献[3]を参考に定めた以下の特徴量を算出する．

メロディの音楽特徴量

・音数	・音長平均
・音域	・音長分散
・音高平均	・メジャーの割合
・音高分散	・マイナーの割合
・16 分音符の割合	

図 3 : メロディの音楽特徴量

次に特徴量の抽出方法について説明する．メロディの音符数をN，n 番目のメロディの構成音の音高を$Note(n)$，n 番目のメロディの構成音の音長を$Ndur(n)$, n 番目の和音中の m 番目の音符の高さを$P_n(m)$とし，以下の計算式で求める．

$$\text{音高平均}：\frac{\sum_{n=1}^{N} Note(n)}{N}$$

$$\text{音長平均}：\frac{\sum_{n=1}^{N} Ndur(n)}{N}$$

$$\text{音高分散}：\frac{\sum_{n=1}^{N}\left(Note(n)-\text{音高平均}\right)^2}{N}$$

$$\text{音長分散}：\frac{\sum_{n=1}^{N}\left(Ndur(n)-\text{音長平均}\right)^2}{N}$$

$$\text{16 分音符の割合}：\frac{\sum_{n=1}^{N}\delta(Ndur(n)-120)}{N} \quad \left(\delta=\begin{cases}1(k=0)\\0(k\neq 1)\end{cases}\right)$$

$$\text{メジャー/マイナーの割合}：\frac{P_i(j)-\min(P_i)}{12}\ \text{の余によって決定}$$

3.2.2 リズムの特徴量抽出

本手法ではリズムに対して，文献[4]を参考に定めた以下の特徴量を算出する．

リズムの音楽特徴量

・タム/スネア/金物/バスドラムの割合
・全音符数
・16 分音符の割合
・3 連符の割合

図 4 : リズムの音楽特徴量

本手法では音符数をN，部位qの音符数を$I(q)$，n 番目の音符の長さを$L(n)$とし，以下の計算式でリズムの各特徴量を求める．

ドラムの各部位を叩いた割合：$\frac{I(p)}{N}$

16 分音符の割合：$\frac{\sum I(p)}{N}$ $(L(n) = 120)$

8 分音符の割合：$\frac{\sum I(p)}{N}$ $(L(n) = 240)$

4 分音符の割合：$\frac{\sum I(p)}{N}$ $(L(n) \leq 480)$

3 連符の割合：$\frac{\sum I(p)}{N}$ $(L(n) = 160)$

3.3 学習

続いて本手法では，動画特徴量とそれに対する各ユーザの印象の関係，またリズム・メロディの音楽特徴量とそれに対する各ユーザの印象の関係を学習する．

3.3.1 ユーザ印象評価

まず予め用意したサンプル動画，サンプルリズム・メロディを評価する際に使用する感性語対を決定する．

本手法では文献[5,6,7]を参考に心理学の観点から，また動画と音楽に共通して適用できそうな感性語対を選んだ．選出した感性語対は以下の通りである．

選出した感性語

明るい － 暗い	速い － 遅い
派手 － 地味	迫力のある － 迫力のない
情熱的 － さわやか	元気 － 落ち着いた

図 5：選出した感性語

この中で動画の色・動きに関して適用する感性語，リズム・メロディに関して適用する感性語を，我々自身の主観に基づいて以下の通りとした．

色の感性語

明るい － 暗い
派手 － 地味
情熱的 － さわやか

動きの感性語

速い － 遅い
迫力のある － 迫力のない
元気 － 落ち着いた

メロディの感性語

明るい － 暗い
情熱的 － さわやか
元気 － 落ち着いた

リズムの感性語

派手 － 地味
迫力のある － 迫力のない
速い － 遅い

図 6：動画の色・動き，リズム・メロディに関する感性語

本手法では各ユーザにサンプル動画を閲覧してもらい，またサンプルメロディ・サンプルリズムを聴取してもらい，上に挙げた感性語への適応度を 6 段階評価で回答してもらう．例えば「明るい－暗い」の場合，1 が最も暗い，6 が最も明るい，という評価とする．以後，この適合度を印象値と称する．このようにして本手法では，ユーザごとの印象値を収集する．

3.3.2 色分布からの印象学習

本手法では以下の処理により 3.1.1 項で示した色分布の特徴量から印象値を推定する．

まず v_{ki} は k 番目の動画における i 番目の色の頻度とする．また 3.3.1 項のユーザ印象評価で得られた 6 段階評価の値を[-1,1]の範囲で 6 等分した値とみなし，j 番目の印象語に対する k 番目の動画の評価に対応する数値を印象値 a_{kj} とする．そして i 番目の特徴量と j 番目の印象語に対する評価の値との関係 c_{ij} を以下の式を用いて求める．

$$c_{ij} = \sum_{k=1} a_{kj} v_{ki}$$

以上の処理によってサンプル動画を用いた学習を終えた後，以下の式を用いて，ユーザ評価結果の与えられていない動画の j 番目の印象語に対する印象値 a_j を算出する．ただし v_i は新しい動画における i 番目の色の頻度とする．

$$a_j = \frac{\sum_{i=1} c_{ij} v_i}{\sqrt{\sum_{i=1} {c_{ij}}^2}}$$

3.3.3 動き分布からの印象学習

本手法では以下の処理により，3.1.2 項で示した動き分布の特徴量から印象値を推定する．まず 3.3.1 項のユーザ印象評価で得られた 6 段階評価の値と，動き分布に関する特徴量から，重回帰分析を用いて以下の式の係数を算出する．

印象値 a ＝ x_1×[速度の平均]＋x_2×[速度の分散]
＋x_3×[角度の分散]
＋x_4×[速度のヒストグラム極大の速度値]
＋x_5×[角度のヒストグラム極大の角度値]
x_1～x_5：標準偏回帰係数

この式を用いて，ユーザ評価結果の与えられていない動画に対して，動き分布の印象値を推定する．

3.3.4 音楽特徴量からの印象学習

本手法では以下の処理により，3.2 節で示したメロディおよびリズムの特徴量から楽曲の印象値を推定する．まず 3.3.1 項のユーザ印象評価で得られた 6 段階評価の値と，メロディおよびリズムの各々に関する特徴量から，重回帰分析を用いて以下の式の係数を算出する．

メロディの印象値 $a = x_1$×[音数]+x_2×[音域]
+x_3×[音高平均]+x_4×[音高分散]
+x_5×[16 分音符の割合]+x_6×[音長平均]
+x_7×[音長分散]+x_8×[メジャーの割合]
+x_9×[マイナーの割合]

リズムの印象値 $b = y_1$×[全音符数]
+y_2×[16 分音符の割合]
+y_3×[3 連符の割合]+y_4×[金物の割合]
+y_5×[バスドラの割合]+y_6×[タムの割合]
+y_7×[スネアの割合]

この式を用いて，ユーザ評価の与えられていないリズムとメロディに対して印象値を推定する．以上の処理により，リズムやメロディに関するユーザごとの印象の違いを考慮した楽曲生成が可能となる．

3.4 楽曲生成

新しい動画が与えられると本手法では，以下の方法によりリズムとメロディを選択し，これを合成することで楽曲を付与する．まず新しい動画に対して，3.3.2, 3.3.3 項の手法により動画の印象値を推定する．また 3.3.4 項の手法により，用意されているリズム・メロディの印象値を推定する．これらの印象値を比較して，ユークリッド空間上で最も距離の近いリズム・メロディを動画の印象に沿った楽曲の素材とする．そしてこの選ばれたリズムとメロディを組み合わせることで楽曲を生成する．

続いて生成した楽曲にコード進行を加える．さらに，動画の長さに合うように楽曲を生成する．この処理では楽曲のテンポ・拍子・小節数を以下の計算式のもと調整することで，動画の再生時間に合うように小節数やテンポを設定する．

以上の処理によって生成された楽曲と動画を合成することで，動画に BGM を付与する．

4. 実行結果

本手法で使用するメロディの作成にあたり，我々は自動作曲システム Orpheus[8]を利用している. Orpheus での作曲に先立ってユーザが与えることができる作曲条件を表 1 に示す．

表 1：Orpheus の作曲条件

歌声指定	リズム形	旋律の型	音域上限	音域下限
音声音量	和声進行	調の設定	拍子設定	速度設定
伴奏楽器	伴奏音形	サブ楽器	サブ音形	ドラムス

このうち歌声指定・伴奏楽器については，現段階の本研究で扱うメロディは歌曲ではないことから，現段階では両方ともエレクトリックグランドピアノを指定している．また拍子は 4/4 拍子のみを対象とする．速度指定ではあとで動画の長さに合うように楽曲の長さを調整することから，現時点では標準的な速さとしてテンポ 112 を指定する．伴奏音形については，今後の課題としてコード進行の弾き方を自動設定できるようにするため，現時点では 2 分音符等の単純な音形とする．以上を前提として，メロディ作成において表 2 中の色のついた作曲条件を自由に選択することで，様々なジャンルの楽曲を生成した．

表 2：自由に与える作曲条件

歌声指定	リズム形	旋律の型	音域上限	音域下限
音声音量	和声進行	調の設定	拍子設定	速度設定
伴奏楽器	伴奏音形	サブ楽器	サブ音形	ドラムス

以上により Orpheus が生成した楽曲は伴奏やリズムが付随された状態であることから，コード進行を付与させたメロディ部分のみを抽出し MIDI データとして保存する．本章で紹介する実験では 30 パターンのメロディを用意した．

またリズムには文献[4]で使われていた 21 パターンを用意した．このうちメロディ 15 種類，リズム 10 種類を学習用のサンプルメロディ・サンプルリズムとした．

本実験では，14 名にユーザ印象評価を依頼し，この結果をもとにしていくつかの異なるジャンルの動画に対して楽曲を生成した．本実験では，人や動物，海などの風景，また運動会などのイベントといったホームビデオとして撮影すると考えられる 1 分以内の 11 種類の動画をサンプルビデオとして用意し, ユーザ印象評価のために閲覧を依頼した．以上の学習過程をもとにして，以下の 2 種類の動画に対して楽曲を付与した．ここでは 14 名のユーザのうち，ある 2 名(ユーザ A とユーザ B)によって生成された楽曲結果を表 3 に示す．

動画 1：人がいない夕暮れの海辺の様子
動画 2：犬が草むらを元気に走っている様子

表 3：動画 1-2 の楽曲生成を行った結果

	ユーザ A	ユーザ B
動画 1	melody22.mid rhythm6.mid	melody29.mid rhythm9.mid
動画 2	melody23.mid rhythm20.mid	melody17.mid rhythm20.mid

ユーザ A とユーザ B では異なる楽曲素材が選ばれており，学習段階の影響によりユーザごとの印象の違いを考量した

楽曲が生成されていることが分かる．しかし動画 2 の明るく元気な動画であるのに対し，ユーザ A とユーザ B でゆったりとした落ち着いた楽曲が生成されてしまった．
また 14 名のユーザに対して生成した楽曲に対する評価は表 4 のようになった。

表 4： 楽曲生成結果における評価

	動画 1	動画 2
4 ： とても合っている	2	1
3 ： 合っている	4	6
2 ： あまり合っていない	8	7
1 ： 合っていない	0	0

これらの評価に対してユーザから頂いたコメントを以下の 3 つに分類することができる．
まずリズムが合っていない，パーカスが合っていないといったコメントからはリズム本来の素材の質に偏りがあると考えられる．ポップやロックといった多様なリズム素材を用いて実験を行う必要がある．曲調が暗い，音の動きが多すぎるといったコメントからは，音数やメジャーマイナーの割合といったメロディの特徴量と曲の印象との間で相関関係が成立しているか確かめる必要がある．また，落ち着いた曲調の方が良い，もっと壮大なイメージが合っているといったコメントからは，曲の音色やコードの弾き方に曲の雰囲気が大きく左右されるものと思われる．これらに関しては今後の課題として扱っていきたい．

5. 考察

4 項で得られた実行結果から例えば，動画および楽曲の特徴量，印象語の見直しや，ユーザ印象評価方法の改善が必要であると考えた．

5.1 特徴量と印象語についての考察

3.3 項の学習によって得られたユーザ A とユーザ B の動画，楽曲に対する印象値をもとに，色分布・動き分布・メロディ・リズムの特徴量と，それぞれに対応する印象語との相関関係について，Hidden というプログラム[9]を用いて可視化した．このプログラムでは高次元データを入力すると図 7 のように，画面右側に次元間距離を表現した散布図が表示され，その中で相関の強い点同士が結ばれる．そしてその結ばれた点の関係性が，PCP (Parallel Coordinate Plot) という可視化手法を適用して画面左側に低次元プロットの集合として表示される．これを用いてユーザ A の色の特徴量と色の印象語について可視化した結果を図 8 に示す．右側の散布図で赤/青の丸で囲まれた部分が，左側の赤/青の四角で囲まれた部分に相関関係として表示されている．

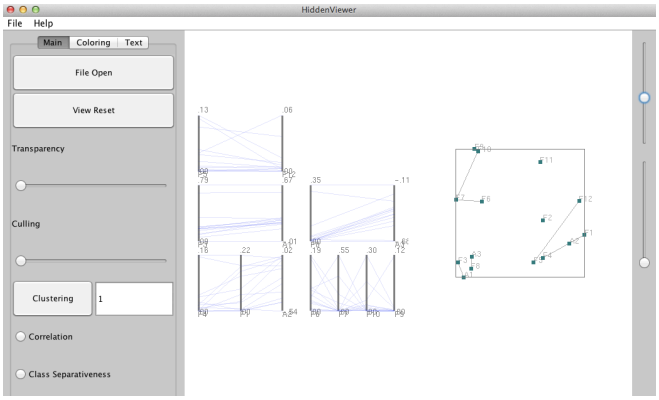

図 7：可視化プログラム Hidden の可視化例

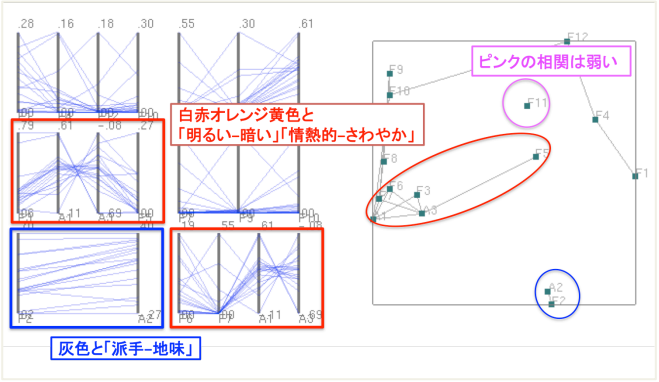

図 8：色の特徴量と印象語の相関可視化

この可視化結果から，灰色と「派手–地味」，白・赤・オレンジ・黄色と「明るい–暗い」「情熱的–さわやか」の相関が強いことが分かった．これはユーザ A の色に対して抱くイメージと近い印象語がそれぞれ選ばれている結果となった．またピンクはどの印象語とも相関が弱く不要なものと思われる．ユーザ B に対しても似たような可視化結果が得られた．

このようにして動き分布・メロディ・リズムでも同様に特徴量と印象語について相関関係を可視化したところ，以下のような結果が得られた．

動き分布に関しては，速度と「迫力のある–ない」，角度と「速い–遅い」の相関が強いことが分かった．しかし，なぜ速度と「速い–遅い」で相関が表れなかったのか，印象語の見直しが必要であると考えられる．

メロディ・リズムに関しては，音高平均，3 連符の割合・金物の割合と印象語の相関が弱かったことから，これらの特徴量を除いて実験を行いたい．

5.2 学習におけるユーザ評価についての考察

3.3.1 項のユーザ印象評価を行った際に，ユーザから動画・楽曲について以下のコメントを頂いた．

動画

- 動画内の動きに差がある場合の評価に迷った
- 色や動きではなく，動画の内容に評価が左右されることも多かった

楽曲

- 評価基準がないと判断が難しい
- 前後に聞いた音楽と比べて評価しがちになる

以上から，動画に関しては，まず動き分布の特徴量抽出方法の見直しが必要である．また一般物体認識を取り入れ，動画から被写体などの高レベルな特徴量を抽出することで，より動画の印象を正確に捉えたい．

また楽曲に関しては，評価方法の改善について，標準的なメロディ・リズムを予め提示しておくことで，それを基準に評価してもらう方法が挙げられる．また 2 つの楽曲を聴き比べていくことで，印象を決定していくようなトーナメント制を導入することも考えられる．評価精度を挙げるためにも，ユーザが評価しやすい評価方法を検討していきたい．

6. まとめと今後の課題

本研究では，動画特徴量，音楽特徴量とその印象の関係をユーザごとに学習させ，その結果として得られる動画および音楽の印象推定結果をもとに，動画の印象に合ったリズムとメロディを選出し，その合成によって動画に楽曲を付与する手法を提案した．

今後の課題として，5 項で挙げた考察をもとに実験・実装を行いたい．また，現時点での本研究では 2 分音符のような単純な音形でコード進行を付与しているが，このコードの弾き方をリズムや曲調に合わせて変えることで，より動画の雰囲気に合った楽曲を生成できると考える．

参 考 文 献

1) 小野佑大, 甲藤二郎, ”音楽のムード分類結果を利用したホームビデオへの BGM 付与支援システム”,情報処理学会音楽情報処理研究会, Vol. 2011-MUS-89, 2011.
2) Jun-Ichi Nakamura, Tetsuya Kaku, ”Automatic Background Music Generation based on Actor’s Mood and Motion”, The Journal of Visualization and Computer Animation, Vol. 5, No. 4, pp. 247-264, 1994.
3) 中山達喜, 吉田真一, ”音楽分類における特徴量の検討”, ファジィシステムシンポジウム講演論文集, Vol. 26, pp. 1256-1261, 2010.
4) 菅野沙也, 伊藤貴之, ”入力文書の印象と感情に基づく楽曲提供の一手法”, 情報処理学会音楽情報科学研究会, Vol. 2014-MUS-103, 2014.
5) 宝珍輝尚, 都司達夫, ”印象に基づくマルチメディアデータの相互アクセス法”, 情報処理学会論文誌, Vol. 43(SIG_2(TOD_13)), pp. 69-79, 2002.
6) 中村均, ”音楽の情動的性格の評定と音楽によって生じる情動の評定の関係”, The Japanese Journal of Psychology, Vol. 54, No. 1, pp. 54-57, 1983.
7) 古賀広昭, 下塩義文, 小山善文, ”画像に合った音楽の選定技術”, 映像情報メディア学会技術報告, Vol. 23, No. 59, pp. 25-32, 1999.
8) 東京大学 大学院情報理工学系研究科 システム情報学専攻, 自動作曲システム Orpheus, http://www.orpheus-music.org/v3/
9) 酒井えりか, 伊藤貴之, 高次元データ可視化のための低次元プロット表示の改良, 第 7 回データ工学と情報マネジメントに関するフォーラム(DEIM), E2-2, 2015.