楽曲から想起される色彩イメージを利用した楽曲推薦システム

菅野桃花
Momoka Kanno
法政大学情報科学部ディジタルメディア学科
momoka.kanno.9h@stu.hosei.ac.jp

Abstract

In this research, we investigate the colors imaged from music and create a system that recommends music from images by using the relationship between the acoustic features of music and color information. By using this system, we can create videos, etc. In this case, it is thought that it is possible to reduce the trouble of searching for a BGM that matches the scene. In the conventional research, the image of the music and the color is once replaced with an impression word, and the impression words are associated with each other. By directly associating music with colors, it is possible to deal with unknown music. In order to directly associate music with colors, multivariate regression of acoustic features and color information is performed. There are 11 types of acoustic features such as Tempo, Roll off, and Zero crossing, and 2 types of color information, RGB and HSV. As a result of performance evaluation of the implemented system, RGB is used for color information and MAP was 0.13. From the results, it is considered possible to recommend songs suitable for the scene by using this system.

1 はじめに

現在, 映画, アニメーション, ミュージカルなど様々なマルチメディアやコンテンツが展開されている. その中で, 音楽は様々なところで利用されており, とりわけ映像作品などの視覚情報に付与されている場合が多い. 音楽と視覚情報の調和は, 情報をより効果的かつ印象的に伝達することにつながるため [1], 音楽が作品全体に与える影響は大きい. また近年では, 動画サイトの普及に伴い, 個人の映像クリエイターなどが増加した. 彼らは著作権フリーの既存の音楽を動画内の BGM として利用することが多いが, 場面に適し, かつ大衆の共感を得られる音楽を探し出すのは手間がかかる. そこで, 手軽に画像に既存の音楽をつけられる検索システムに需要があるのではないかと考えた.

本研究では, 視覚情報の中でも特に色に着目し, 画像にマッチした音楽を既存の楽曲の中から検索できるシステムを作成する. このシステムでは, 入力の画像の色情報を抽出し, その色情報と近い色彩イメージを与える楽曲を検索し出力する. 楽曲と画像を対応付ける従来研究 [2, 3] があるが, この研究では楽曲と画像を主観評価実験により一度印象語に置き換えたうえで両者のベクトルを計算し, 同一平面上にマッピングすることで対応付けている. 印象語に置き換える作業はアンケートによって手作業で行われるため, 印象語に置き換えていない未知の楽曲を推薦システムに対応させるのが難しく, 推薦システムのデータベースを作成するのに手間がかかる. また, 一度言語を介することによって, 直感的なイメージの結びつきが失われてしまう恐れがある. 本研究では, 楽曲の音響特徴量と楽曲から想起される色彩イメージの関係性を主観評価実験によって明らかにし, これを利用した楽曲推薦システムを作成することで, 上記にあげた問題点を解決する.

Supervisor: Prof. Katunobu Itou

2 色彩イメージを利用した楽曲推薦システム

本研究で作成するシステムの概要について説明する. まず, 使用者は画像ファイルをシステムに入力する. 次にシステム内部で画像ファイルの色情報を抽出する. 色情報は HSV(色相・彩度・明度) と RGB のそれぞれ 3 つの要素を想定する. 従来の, 色と音の直接的・直感的な関係性を利用した検索システムや色聴に関する研究の多くで, HSV の各要素と音楽を構成する要素が結びついていることが示唆されている. HSV は回帰処理が難しい点があるため RGB も採用する.

楽曲推薦アルゴリズムのために, 事前に楽曲の楽曲情報や音響特徴量をまとめたデータベースを作成する. アンケート調査により色と対応する楽曲情報や音響特徴量について, またその重みづけについて調査する. 入力された画像の色情報と, 調査結果から導き出した色に対応する楽曲情報・音響特徴量を用い, データベースから適当な楽曲を推薦する. 推薦された楽曲と入力の画像の印象がふさわしいかどうかについて, 性能評価を実施し評価してもらう.

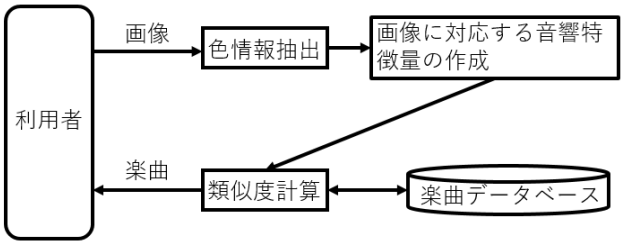


図 1: システムの概要

3 楽曲推薦システム作成のためのアンケート調査

本研究では最初に, 色情報に対応する音響特徴量を作成するための調査を行う. 具体的には, 楽曲から連想する色彩イメージに関する主観評価実験を行い, 得られたデータを画像と楽曲の対応付けに用いる.

本調査では, 一般の人々に対して色彩イメージについての調査を行うが, 共感覚の一つである色聴に関する従来研究 [4, 5, 6, 7] や色調保持者以外も含めて楽曲と色彩 (画像・映像も含む) の関係に着目した従来研究 [3, 8] を参考にしてアンケート内容を作成する. 共感覚とは, ある物理刺激に対してその感覚器官以外に属するはずの感性反応を引き起こす現象である. 色聴とは, 音を聴いて色が見える現象であり, このような現象が起こる人はごく一部に限られている.

従来研究では音と色との対応がおよそ 1 対 1 でとれていることが明らかになっている [4]. 色と音の関係において, 色と音の「明度」の関係は大多数の合意が得られているが, 音程と色彩の関係のように個人の感覚に依るところが大きいものもある. また楽曲と対応の取れる色の数は 20〜30 色であることも明らかになっている [5, 9]. ここでの楽曲と色の対応については, 楽曲を聴取し色を選択する実験に基づくものである. これらのことから, 20〜30 色の対応が確認されることを目指し, 後述する音響特徴量が重複しないよう選曲した. また, それぞれの楽曲で,

いずれかの音響特徴量が特徴的な値を持つものを選曲した．今回のアンケートでは YouTube のオーディオライブラリの中から 30 曲を選び，全ての曲を 30 秒前後に切り取ったものを使用する．30 秒という長さは，被験者の負担を軽減し，かつ各楽曲の音響特徴量が現れる最低限の長さであると考えたためである．被験者は研究室の学生 6 名 (男性 5 名，女性 1 名) を対象に行う．回答は Web 上のアンケートフォームで行う．このアンケートフォームは，3 年次の予備実験で実施した紙のアンケートや色聴について調査しているアンケートを基に設計し，先述した色彩イメージの回答は画面上のカラーピッカーで行ってもらう．また，予備実験を実施した際楽曲から連想できる色が存在しないという回答があったため，楽曲に対して色彩イメージがもてない場合は色彩を感じないという選択肢を選んでもらう．

図 2: アンケート画面の一部

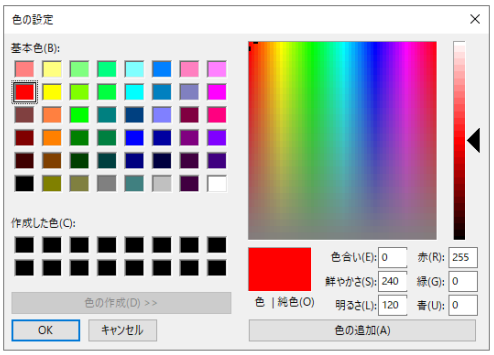


図 3: カラーピッカー

4　楽曲推薦システムの構築

システムの構築は既存の楽曲分類・楽曲検索・楽曲推薦等の研究 [11] を参考に行う．システムは大まかに，画像からの色情報の抽出，抽出した色情報から適切な楽曲を推薦するアルゴリズムの作成の 2 つの段階に分けられる．後者はさらに，楽曲データベースの作成，色情報に対応する音響特徴量の作成，データベース内の検索システムの作成の 3 つに分けられる．

4.1　画像からの色情報の抽出

画像からの色情報の抽出は，画像を 32 色に減色したのちに，使用されたピクセル数が最も多い色を出力するという手順で行う．取得する色情報は HSV と RGB のそれぞれ 3 要素である．

色情報の取得は，画像からカラーコードを取得したのちに RGB に変換し，HSV に計算しなおすという方法をとっている．まず，カラーコードの先頭に含まれる # を除き先頭から 2 文字づつの値がそれぞれ RGB の R, G, B に対応している．これらはそれぞれ 16 進数で表されているため，10 進数に変換する．それぞれの値を R, G, B として HSV に変換した [12]．

4.2　楽曲推薦アルゴリズムの作成

4.2.1　データベースの作成

推薦される楽曲群のデータベースを作成する．楽曲は YouTube ライブラリの中から約 500 曲を使用する．先行研究 [9, 13] から，ボーカルが加わることで，楽曲の色彩イメージに対する歌詞の影響などを考慮する必要があることが判明している．また声の音響特徴量が楽曲からイメージされる色に影響がある関係があるが，これを考慮すると，楽曲からボーカルを抽出し，ボーカルと色の関係性も明らかにしたのちに，伴奏とボーカルそれぞれの色との対応の重みづけを計算する必要があり，非常に複雑になる．そのため本研究ではボーカルのない楽曲や，スキャットなどがわずかに入っていたとしても歌詞が意味を持たない楽曲など，出来る限り歌詞やボーカルそのものの影響を受けない楽曲を中心に使用する．使用する YouTube ライブラリ内の楽曲はどれもおよそ 2〜3 分のものである．楽曲の選択はオーディオライブラリ内に投稿された日付の新しい順に行った．また，データベース化する音響特徴量は先行研究 [10, 11, 14] で用いられ，色と関係がある可能性が示唆されているものから，表のように定めた．例えばテンポの速さは，色の明度の高さと関係があることや，楽曲のキー (調性) が色の色相と関係があるのではないかと考えられている．音響特徴量の分析は MIRtoolbox を用いた．また計算処理を考慮し，Key は以下の表のように数値に変換している．

表 1: 使用する音響特徴量

音響特徴量	説明
Tempo	フレーム内の平均のテンポを示す.
Key	楽曲の調性を示す.
RMS energy	曲の音量の二乗平均平方根を示す.
Low energy	エネルギーが平均エネルギーよりも低いフレームの割合を示す.
Zero crossing	波形がゼロ値を取る回数を示す.
Roll off	低い周波数のエネルギーの合計を計算することにより、総エネルギーの 85 %を占める周波数の割合を示す.
Brightness	1500 Hz 以上の周波数のエネルギーの割合を示す.
Spectral irregularity	音色の変化の大きさを示す.
Inharmonicity	非根音のエネルギーの割合を示す.
Mode	メジャーコードとマイナーコード間のエネルギーの違いを示す.
Key clarity	楽曲の調性の強度を示す.

表 2: Key の変換

Minor	Key	Major
-1	C	1
-2	#C	2
-3	D	3
-4	#D	4
-5	E	5
-6	F	6
-7	#F	7
-8	G	8
-9	#G	9
-10	A	10
-11	#A	11
-12	B	12

4.2.2　色に対応する音響特徴量群の生成

先述のアンケート調査の結果から，多変量回帰を行うことで，ある色情報にふさわしい音響特徴量を算出するための行列を計算する．式は以下のようになる．

$$y_i = X_i \beta + e_i, i = 1, ..., n$$

ここで，X はアンケート 30 問に対する被験者 6 人の回答した色情報を HSV か RGB に変換した 180 × 3 の行列である．y はアンケートに使用した楽曲の音響特徴量の 180 × 11 行列である．X を説明変数，y を説明変数としたときの回帰係数 β を計算し，色情報の 1 × 3 行列と β の積を求める．この行列が，

表 3: 楽曲データベースの一部

Title	RMS energy	Low energy	Tempo	Zero crossing	Roll off	Brightness	Inharmonicity	Mode	Key clarity	Spectral irregularity	Key
FynestLyk	0.172	0.627	116	416	2936	0.273	0.491	0.193	0.782	0.550	A#maj
TrueArtRealAffectionPart4	0.177	0.489	184	312	2285	0.194	0.466	0.037	0.733	0.157	A maj
Find Your Way Beat	0.225	0.550	105	453	6320	0.356	0.474	0.088	0.693	0.329	B maj
Find Me Here	0.226	0.500	130	1095	10577	0.608	0.501	-0.092	0.530	0.352	B min
Elegy	0.124	0.444	148	659	1399	0.130	0.502	-0.037	0.727	0.444	C# min
Eternal Garden	0.156	0.545	141	481	1462	0.146	0.474	0.043	0.811	0.458	C maj
Our Planet	0.296	0.497	120	739	8168	0.456	0.466	0.003	0.486	0.307	D# maj
Beside Me	0.305	0.526	127	1112	9169	0.560	0.524	0.042	0.501	1.095	E maj
No Starlight Dey Beat	0.274	0.488	123	885	11921	0.680	0.464	-0.120	0.766	0.941	F# min
Sweetly My Heart	0.130	0.507	190	766	1364	0.122	0.433	0.303	0.896	0.165	G maj

入力画像の色情報に対応する楽曲の音響特徴量となる．この計算では正規化を行う．アンケートに使用した楽曲 30 曲の音響特徴量に対し平均と分散をあらかじめ求め正規化した値で β を求める．求めた β の要素それぞれに対し，分散をかけ平均値を足す操作を行う．色情報は HSV と RGB の 2 種類の値を用いて行う．

4.2.3 データベース内の検索システムの作成

上で求めた，ある色情報に対応する楽曲の音響特徴量とデータベース内の楽曲の音響特徴量の類似度を計算し，類似度の高い楽曲上位 10 曲のタイトルを検索結果として表示する．今回，類似度の計算はコサイン類似度を用いて行う．コサイン類似度はベクトル空間モデルにおいて文書同士を比較する際に用いられる類似度計算手法であり，−1 以上 1 以下の値をとる．コサイン類似度は以下の式で求められる．

$$cos(A, B) = \frac{A \cdot B}{|A||B|}$$

5 システムの性能評価

5.1 評価実験

システムの性能評価実験を行う．被験者は学内の生徒を対象として行う．図のような色数の少ない画像を 10 枚用意し，システムにそれらの画像を入力し，画像のイメージと推薦された楽曲がマッチしているか評価する．色数が少ない画像を使用する理由として，システムが正常に動いていることを確認し，色以外の画像のイメージに推薦精度が左右されてしまうことをできる限り防ぐためである．評価は，ふさわしい，ある程度ふさわしい，ふさわしくないの 3 段階で行う．評価結果から，「ふさわしい」と「ある程度ふさわしい」と評価された楽曲を True とし，AP(Average Precision) と MAP(Mean Average Precision) を求める．色情報に HSV を用いた時と RGB を用いた場合でどのように差が出るか観察し，有用な方を採用する．

5.2 AP と MAP の算出方法

まず，AP と MAP の算出に使用する Precision と Recall について説明する．今回用いた Precision, Recall は次のような式で表される．

$$Precision = \frac{\text{それまでの } True \text{ の数}}{\text{現在の順位}}$$

$$Recall = \frac{\text{それまでの } True \text{ の数}}{\text{全ての } True \text{ の数}}$$

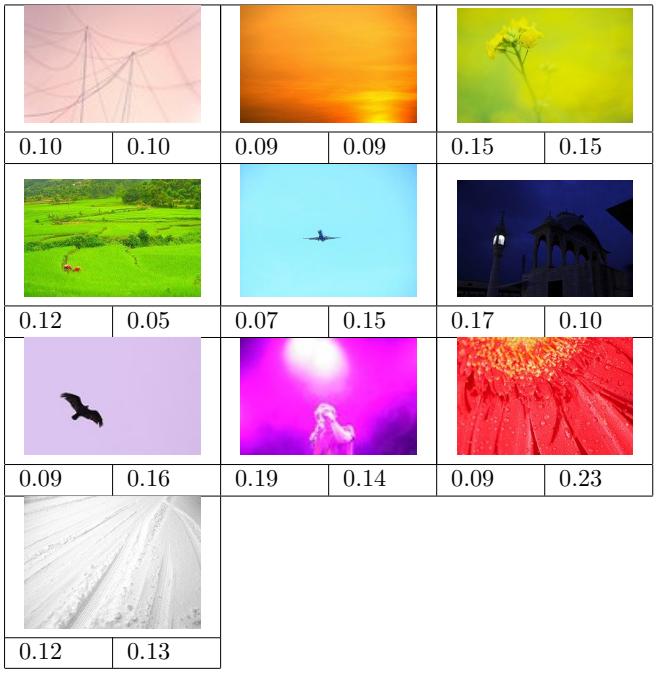
本研究では，データベースにあるすべての曲に対し手動で評価を行うことで全ての True の数を算出している．AP は一般的に次の式で表される．

$$AP = \int_0^1 p(r)dx$$

MAP は AP をそれぞれの事例ごとに計算して平均をとったものであり，式は次のようになる．

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AP(data_i)$$

表 4: 評価実験に用いた画像と各画像の検索結果に対する AP. HSV を使用した場合の AP を画像の左下部，RGB を使用した場合の AP を右下部に記載した．画像は画像番号順に横に並んでいる．



今回は Python の scikit-learn の label_ranking_average_precision_score 関数を使用した．

5.3 実験結果

実験結果は表のようになった．各画像に対しての True は平均 34 個、最少が画像 4 の 16 個で最大が画像 2 の 47 個であった．10 種類の画像で算出した AP の平均である MAP は，HSV を用いた場合が 0.12，RGB を用いた場合が 0.13 となった．複雑な画像を用いた場合の実験結果は表のようになった．各画像に対しての True は平均 17 個，最少が画像 10 の 7 個で，最大が画像 1 と画像 8 の 32 個となった．MAP は RGB を使用し，クラスター数が 3 のときに最も大きくなり，0.048 となった．色数の少ない画像を使用した場合と比べ，AP と MAP は大きく低下した．

表 5: MAP

HSV	RGB
0.12	0.13

5.4 考察

HSV を使用した場合の MAP は 0.12，RGB を使用した場合はの MAP. は 0.13 となったため，HSV よりも RGB を用いたほうがより画像にふさわしい楽曲を推薦することができたと言える．複雑な画像を使用した場合の評価実験でも同じことが言える．また，今回実施した主観評価実験から，楽曲と色の対応

について重みの高い音響特徴量は, RMS energy, Low energy, Inharmonicity, Brightness, Mode, Key clarity の 6 つであることが分かった．従来研究で述べられているいくつかの関係について (楽曲のテンポと明度, 楽曲のキーと色相) 本研究では明らかにすることができなかった．主観評価実験の被験者が従来研究に比べて少なかったため, 十分に傾向を見出すことができなかったからである．被験者を増やす, 被験者の属性を多様にする, 実験に使用する楽曲を多様にするなどして追加の主観評価実験を行うことで改善がみられる可能性がある．

問題点としては次のことが言える．色相は環状になっているため, 回帰行列を求める際に工夫が必要であると考える．現状では色相が 359 から 0 になる, 赤色の画像 9 を入力したときに, HSV を使った場合の MAP が 0.09 となっている．これは同じ画像を入力とし RGB を使った場合の MAP である 0.23 と比べ低い値となっている．また赤色の画像を複数入力してシステムを動作させると, 全く異なった結果が表示されてしまうことがある．本来は似たような色を持つ画像を入力したときには似たような楽曲が推薦される必要があるが, 赤色の画像については色相が 1 周したときの回帰がうまく行えていないためにこのような推薦結果になってしまう．RMS energy, Roll off, Brightness が極端に低いことが多い, ピアノやシンセサイザーを主体としパーカッションがなく静かな楽曲は推薦されにくく, 様々な色情報の画像を入力しても, 常に平均 450 位にある．これらの楽曲は, 聴いた時のイメージとしては似通っているが, 音響特徴量を見ると明らかな共通点は見られない．今回使用した音響特徴量以外に, これらの楽曲に共通する音響特徴量が存在する可能性があると考えられる．またアンケートで回答されにくい．また明度が 100 以下の色情報を持つ画像を入力にしたときに, ふさわしい楽曲が推薦されにくいという問題点がある．明度 100 以下の色は先述のアンケートでの回答率が 5.63% であった．これらの問題は, 上記の楽曲や画像との対応が取れるようなアンケートを重点的に行い, 多変量回帰に用いるサンプル数を増やすことで改善することができる可能性がある [9]. HSV の方が AP が低かった画像として, 画像 9 のほかに, 画像 5 と画像 7 がある．これら 2 つの画像は, 先述の回帰の処理とは関係しないと考えられるため, これらの画像の AP が低い理由を明らかにすることで, HSV を採用した場合の推薦精度を上げられると考えられる．

今後様々な画像に対応できるシステムにするためには以下のことが言える．音楽の構成要素としては [17] で示されているメロディー, コード, リズムなどがあり, 今回用いた音響特徴量以外にも色彩イメージと関係のある特徴量が存在する可能性がある．さらに, このシステムでは画像の色情報が近いものは類似した楽曲が推薦されるが, 色情報が近くても画像のイメージが異なる場合も考えられる．画像の客観的特徴として色彩の空間的配置や画像の中でのオブジェクト情報などがあり [18], 人が画像を見るときに, 何に注目しているのかという知見を利用することで, 推薦結果の精度を上げることができ, 複雑な画像にも対応できると考えられる．

6 おわりに

本研究では, 共感覚の知見を参考に, 楽曲から連想する色彩イメージに関する主観評価実験を行い, 実験結果を基に多変量回帰で求めた行列を利用した楽曲推薦システムの実装を行った．評価実験の結果から, 色情報に RGB を採用した場合の MAP は 0.13 となり, このシステムを利用することで, およそ 1/10 の正解率で画像にふさわしい楽曲を検索することができた．また, HSV よりも RGB を用いたほうが楽曲推薦の性能が良かった．しかし, システムの精度にはまだ改良の余地がある．特に回帰の処理を環状に行うことができれば, HSV を使用した際のシステムの性能が向上する可能性がある．また, より複雑な画像を入力とするためには, 新たな音響特徴量を導入したり, 追加の主観評価実験を行い回帰計算のためのサンプル数を増やしたり, 人間が画像処理を行う上での心理学的な知見などを生かした処理を行ったりする必要がある．さらに, 実用化を考えると, 楽曲データベースの充実や, システムの UI を工夫することが必要である．楽曲データベースの増加は MIRtoolbox を用いて簡単に行うことができるが, 一部の音響特徴量には計算に大幅な時間を取られるものもあった．また, 今回作成したシステムは Python のプログラム内でファイル名を指定することで動作しコンソールに結果が表示されるが, 画像の入力をドラックアンドドロップで行えるようにすることでよりユーザーフレンドリーなシステムとなる．

参考文献

[1] 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, 2000.

[2] 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”, 日本感性工学会論文誌, 2012, Vol.11, No.3, pp.435-442.

[3] 伊藤貴之,“文書や画像の印象にもとづく楽曲生成”,The 31st Annual Conference of the Japanese Societyfor Artici al Intelligence, 2017, No.JSAI2017, pp.2C3OS20a2.

[4] 野村順一,“色の秘密：色彩学入門”, 2015.

[5] 山脇一宏, 椎塚久雄,“音感と色聴感覚”, エンタテインメント感性特集, 2005, Vol.5, No.3, pp.31-37.

[6] 山脇一宏, 椎塚久雄,“カラーイメージスケールを利用した音楽の特徴抽出”, 知能と情報 (日本知能情報ファジイ学会誌), 2005, Vol.17, No.5, pp.615-621.

[7] 岩井大輔, 長田典子, 津田学, 和氣早苗, 井口征士,“音と色のノンバーバルマッピング”, 音楽情報科学, 2002, Vol.47, pp.97-104.

[8] 川野邊誠, 亀田昌志,“楽曲から受ける印象の時系列変化を考慮した楽曲から配色へのメディア変換”, 芸術科学会論文誌, 2006, Vol.5, No.4, pp.95-105.

[9] 仲村哲明, 内海彰, 坂本真樹,“色彩想起と歌詞の関係に基づく楽曲検索”, 人工知能学会論文誌, 2012, Vol.27, No.3, pp163-.175

[10] 斉藤優理, 伊藤貴之,“MusiCube：特徴量空間における対話型進化計算を用いた楽曲提示インターフェース”, 可視化情報学会論文集, 2014, Vol.34, No.9, pp.17-27.

[11] 上原美咲, 伊藤貴之,“楽曲群のコード進行・メタ情報・楽曲特徴量の統合可視化の一手法”, 情報処理学会全国大会講演論文集, 2015, Vol.77, No.2, pp.395-396.

[12] 松橋聡, 藤本研司, 中村納, 南敏,“顔領域抽出に有効な HSV 表色系の提案”, テレビ誌, 1995, Vol.49, No.6, pp.787-797.

[13] 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推定法の設計と評価”, 情報処理学会研究報告, 2011, Vol.2011-MUS-91, No.7, pp1-8.

[14] 草間かおり, 伊藤貴之,“楽曲データの印象表現に基づいた一覧手法の一手法”, 情報処理学会研究報告, 2009 Vol.2009-MUS-81, No.19, pp85-86.

[15] 中野倫靖, 吉井和佳, 後藤真孝,“トピックモデルを用いた歌声特徴量の分析”, 情報処理学会研究報告, 2013, Vol.2013-MUS-100, No.23, pp1-7.

[16] 池田徹志 室田健吾, 石黒浩,“全方位映像から音楽情報へのメディア変換に基づく視覚情報の伝達”, 情報学会論文誌, 2007, Vol.48, No.1, pp.274-283

[17] 秋口俊輔, ”ソフトコンピューティング手法を用いた曲印象からの楽曲自動生成システムの構築”, 知能と情報, 2009, Vol.21, No.5, pp.782-791.

[18] 近藤邦雄, 高橋雅博, 松永政尚, 山嵜秀樹, ”画像データベースのためのイメージカラー検索手法”, 映像情報メディア学会誌, 2000, Vol.54, No.11, pp.1615-1622.