音楽アイコン自動選択手法 MIST への音響データファイル適用の試み

小田瑞穂† 伊藤貴之†

我々は，楽曲と画像をそれらの印象に基づいて自動的に組み合わせる，という考え方に基づいて楽曲のアイコン画像を自動選択する手法 MIST を提案している．MIST を利用することで，楽曲ファイルのブラウザには多種多様なアイコン画像が表示され，ユーザは好みや気分に応じて容易に選曲できるようになる，と考えられる．本報告では MIST の拡張手法として，楽曲ファイルに WAVE 形式や MP3 形式などの音響データファイルを用いる実装手法を提案し，そのユーザテスト結果を示す。

1. はじめに

マルチメディア技術の発展のおかげで，コンピュータを使った音楽の録音・再生が可能になった．さらに今日では，インターネットが主たる音楽購買メディアになりつつある．このような技術の発展に伴い，我々はインターネットからダウンロードした音楽を，コンピュータやポータブルプレイヤなどで聴く機会が増えている．また，ハードディスクドライブの容量増加などに伴い，個人のパソコンに大量の楽曲をダウンロードすることも容易になった．

コンピュータ上で楽曲を再生するためのソフトウェアには，楽曲選択を支援するユーザインタフェースが搭載されていることが多い．例えば iTunes などに見られるプレイリストのように，いくつかのソフトウェアでは楽曲の雰囲気や印象で楽曲を選択するユーザインタフェースを提供している．しかしながら，これらのユーザインタフェースの多くは，タイトルや演奏者名をはじめとする文字情報だけを表示している．これだけでは，ユーザの好みや気分にあった雰囲気をもつ楽曲を，大量の楽曲の中から直感的に選ぶのは難しい場合があると考えられる．

我々は，楽曲と画像をそれらの印象に基づいて自動的に組み合わせる，という考え方に基づいて楽曲のアイコン画像を自動選択する手法 MIST (Music Icon Selector Technique)[1),2)] を提案している．一般的なファイルシステムでは，同一のファイル形式で保存された楽曲には同一のアイコン画像（多くの場合，その再生用アプリケーションを特定するアイコン画像）を適用するのに対して，MIST は図 1 に示すように，同一のファイル形式で保存された楽曲に対して，多様なアイコン画像を自動選択する．これによって，楽曲ファイルのブラウザには多種多様なアイコン画像が表示され，ユーザは好みや気分に応じて容易に選曲できるようになる，と考えられる．

また我々は，これらのアイコン選択結果を一覧表示する一手法を示している。我々の実装では，楽曲はフォルダによって分類されていると仮定する．そしてフォルダを枝ノード，各楽曲を葉ノードとする木構造を構築し，フォルダを長方形の枠で，楽曲をアイコンで表示する．図 1 において，灰色の長方形の枠の中に属するアイコンは，同じフォルダに保存されている楽曲ファイル，と解釈できる．このように MIST は，多数のフォルダにまたがって保存されている多数の楽曲を，一度に，かつ直感的に眺められるシステムとして機能できると考えられる．この表示スタイルは，我々自身によって発表されている階層型データ可視化手法「平安京ビュー」[3)] を応用したものである．

MIST に関するこれまでの報告では，音楽ファイルに MIDI 形式ファイルを利用していた．本報告では新たに，音響ファイル（WAVE ファイル，MP3 ファイルなど）を適用する実装について論じるものである。また本報告は，これまでの報告よりも多様なユーザテストを実行した結果を示すものである．

† お茶の水女子大学大学院

図 1　MIST による楽曲表示例.

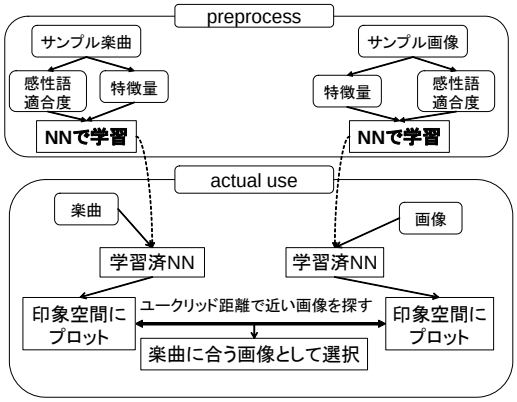

図 2　MIST の処理手順.

2. MIST の概要

MIST の処理手順を図 2 ☆に示す．MIST は事前学習段階と実用段階の 2 段階で構成される．

事前学習段階では，いくつかのサンプル画像とサンプル楽曲を用意し，それらの特徴量をそれぞれ抽出する．それと同時に，ユーザにサンプル画像とサンプル楽曲を提示し，これらに対して感性語の適合度を回答させる．ここで用いる感性語を表 1 に示す．そして

☆ 図 2 では，ニューラルネットワークを NN と略している．また図 2 では，感性度適合度によって構成される多次元空間を，印象空間と称している．

表 1　感性語一覧.

明るい 暗い	重い 軽い	硬い 柔らかい	安定 不安定
澄んだ 濁った	滑らか 歯切れ良い	激しい 穏やか	厚い 薄い

MIST は，これらの特徴量と感性語適合度の関連性を，ニューラルネットワークを用いて学習する．

実用段階では，ユーザが蓄積しているアイコン画像と楽曲に対して特徴量を抽出し，ニューラルネットワークを用いて感性語の適合度を推定する．続いて，感性語適合度によって構成される多次元空間において，各々の楽曲に対して全てのアイコン画像までのユークリッド距離を計算し，楽曲からの距離が最短である画像を，その楽曲のアイコンとして選択する．

以下の各節では，MIST の各々の処理工程について詳細を述べる．

2.1 画像の特徴量

MIST ではアイコン画像として，デジタルカメラで撮影した写真や，コンピュータ上で描いた絵を，縦横 48 画素の正方領域に切り取った BMP 画像を用いる．

画像からの特徴量には，色分布，周波数分布，境界線形状，などを用いることが多いが，現時点での我々の実装では色分布のみを用いている．一般的に，特徴量の次元が高くなればなるほど，その学習には多くの教師信号を必要とする．MIST ではユーザビリティの観点から，教師信号となるサンプル画像をあまり増やしたくない．そのため我々は，特徴量の次元数を低く抑えることを優先的に考え，特徴量は色分布に基づくもののみを採用し，周波数分布や境界線形状に基づく特徴量を採用していない．

我々はよく知られた数種類のアイコン画像を題材として，被験者調査を行った．具体的には，アイコン画像から数種類の代表的な色を抽出し，どの色が印象に残っているか，という質問をし，被験者の回答を集計した．その結果から我々は，アイコン画像のような小さな画像において印象に残りやすい色は，

- 面積の大きい色領域（例えば背景色領域）
- 原色に近い鮮やかな色領域

の 2 種類が多いと結論付けた．そこで我々の実装では，この仮説に該当する 2 色を画像から抽出し，その 2 色の画素値を YCbCr 表色系に変換した合計 6 値を，特徴量に用いている．

2.2 特徴量に基づく感性語適合度の推測

MIST の事前学習段階では，サンプル画像とサンプル楽曲をユーザに提示し，その感性語適合度をユーザ

に回答してもらう．現時点での我々の実装では，表 1 に示す 8 種類の感性語をユーザに提示し，各々の画像および楽曲に対して，各々の感性語がどの程度適合するかを，1 から 7 までの 7 段階の整数で回答してもらっている．なお MIST では，画像と楽曲の両方に対して，同じ感性語群を用いる必要がある．

続いて MIST では，画像から得られる n 次元の特徴量，および楽曲から得られる m 次元の特徴量に対して，感性語適合度との関連性を学習する．ここで，

- N 枚の画像の集合を $P = \{p_1, p_2, ..., p_N\}$
- M 曲の楽曲の集合を $Q = \{q_1, q_2, ..., q_M\}$
- i 番目の画像の特徴量を $C_i = \{c_{i1}, c_{i2}, ..., c_{in}\}$
- j 番目の楽曲の特徴量を $D_j = \{d_{j1}, d_{j2}, ..., d_{jm}\}$
- i 番目の画像の感性語適合度を $S_i = \{s_{i1}, s_{i2}, ..., s_{i8}\}$
- j 番目の楽曲の感性語適合度を $T_j = \{t_{j1}, t_{j2}, ..., t_{j8}\}$

とすると，MIST が必要とする学習は，

- $S_i = f(C_i)$ を満たす関数 f を導く
- $T_i = g(D_i)$ を満たす関数 g を導く

という独立した 2 種類の学習に相当する．MIST では，サンプル画像およびサンプル楽曲の特徴量と，これらに対してユーザが回答した感性語適合度をもとに，関数 f および g を導くような学習を実現する．

MIST では上述の学習に，教師付き学習手法であるニューラルネットワークを用いる．現時点の我々の実装では，回路網に階層型神経回路網を採用し，その学習方法に誤差逆伝播法を採用している．

MIST は以上の処理による学習結果を保存することで，事前学習段階を完了する．実用段階においては，まず関数 f および g を用いて，新しく蓄積されたアイコン画像および楽曲の感性語適合度を推測する．

2.3 楽曲と画像の距離計算

MIST では，前節までに述べた手法で導出した感性語適合度を利用して，各々の楽曲に対して，最も距離の近い画像を選択する．我々の実装では，8 項目の感性語を採用しているので，この適合度を 8 次元ベクトルとみなし，8 次元空間におけるユークリッド距離として楽曲と画像の距離を計算する．つまり，i 番目の画像と j 番目の画像の距離を，以下の式で計算する．

$$distance = \sqrt{\sum_{k=1}^{8}(s_{ik} - t_{jk})^2} \qquad (1)$$

なお我々は，MIST において，最大マッチングや 1 対 1 マッチングの適用は必要ないと考えている．むしろ我々は，同じような印象の楽曲には同じ画像を選択するべきだと考える．したがって MIST では，楽曲に対する画像の自動選択は，単純に感性語適合度の距離が最小なものを選ぶにとどめている．

2.4 アイコン選択結果の一覧表示

図 1 に示したとおり，我々は「平安京ビュー」[3] を用いて，アイコン選択結果を一覧表示するユーザインタフェースを開発している。我々の実装では，楽曲群を蓄積しているファイルシステムのフォルダの階層構造を，入れ子の長方形領域で表現し，その中にアイコンを隙間なく画面に敷き詰めるようにして表示する．また，このユーザインタフェースでは，個々のアイコンは楽曲ファイルに紐つけられており，アイコンをダブルクリックすると楽曲の再生を開始するようになっている．このような表示スタイルは，多数のフォルダにまたがる大量楽曲の一覧表示，またそれを利用した楽曲ファイルの整理に有効である．

3. 音響データファイルの適用

我々は過去の発表[1),2)] において，楽曲ファイルに MIDI 形式ファイルを採用していた．MIDI 形式ファイルに代表される楽譜情報は，近年では携帯電話の「着メロ」などの形で普及してきた．しかし，我々が楽曲ファイルとして日常生活で扱うものは，MP3 ファイルに代表される音響情報のほうが圧倒的に多い．そこで我々は，MIST に音響データファイルを適用する試みを行った．現時点の我々の実装では，楽曲の特徴量算出において，MP3 ファイルを WAVE ファイルに変換したものを用いている．

MIST では楽曲の特徴量についても，画像の特徴量と同様の考え方に基づき，サンプル楽曲を増やさないことを優先して，その次元数を低く抑えている．現時点での我々の実装では，以下の 2 種類の数値を特徴量として試験的に用いている．

音場感を表す特徴量．具体的には，以下の変数を用いる．

- 時間別の全ての音量の合計を時間で割った平均音量
- 時刻別の音量の分散
- 平均音量よりも大きい音量を出した回数を総時間で割った単位時間当たりの数
- 音量のピークが最も頻繁に現れる周波数の対数

楽典的な意味を表す特徴量．具体的には，音楽の 3 要素（メロディ・ハーモニー・リズム）に関係ある以下の変数を用いる．

- メロディ音域となることが多い音域（我々の実装では 200〜600Hz）の音量比
- 長調を構成する倍音と，短調を構成する倍音の音量比

- リズム演奏によって生じる音量ピークの出現頻度

4. 実　　験

4.1 実験方法

本章では，提案手法を実装して実験した結果を示す．我々は，画像の特徴量算出モジュール，およびニューラルネットワークによる学習モジュールを GNU gcc 3.4 上で実装し，楽曲の特徴量算出モジュールを MATLAB および MATLAB signal processing unit 上で実装し，ユーザインタフェースを Java SDK 1.5 で実装した．これらを HP Compaq dc5700 (CPU 1.6Hz, RAM 1.0GB) および Windows XP 上で実行した．

本実験にて我々は，100 曲の MP3 形式楽曲ファイルと，9 種類の BMP 形式アイコン画像ファイルを用意した．そして被験者に対して，100 曲の MP3 形式楽曲ファイルのうち 11 曲を，サンプル楽曲ファイルとして試聴させ，その感性語適合度を回答させた．同様に，9 種類の BMP 形式アイコン画像ファイルを閲覧させ，その感性語適合度を回答させた．続いて，これらの楽曲および画像の特徴量と感性語適合度を学習させ，その学習結果から残りの楽曲ファイルに対してアイコン画像を決定した．以上の結果から得られる合計 100 曲の楽曲ファイルのアイコンを一覧表示し，被験者に閲覧させた．そして被験者に，100 曲のアイコンの中からランダムに 10 曲をダブルクリックさせ，再生される楽曲とアイコンの印象について，

(1)　ぴったり合っている
(2)　概ね合っている
(3)　まあまあ合ってないこともない
(4)　あまり合っていない
(5)　合っていない

の 5 段階の言葉，あるいはそれに準じる言葉で回答させ，備考意見がある場合にはそれも併せて記述させた．以下の集計結果では，上記（1）〜（3）を「満足」とみなして集計し，それを試聴曲数（本実験の場合 10 曲）で割ることで満足度とした．

4.2 実験結果

以下に，本実験による集計結果と，アイコン表示結果の例を示す．なお本実験では，20 代の女子大生 7 名を被験者とした．

まず我々は，3 節で示した 2 種類の特徴量の有効性を比較した．具体的には，ランダムにアイコンを選択した場合の満足度に対する，MIST を使ってアイコンを選択した場合の満足度の比を算出した．その結果，「音場感を表す特徴量」を採用した場合には，満足度の比が 0.89 と，ランダムにアイコンを選択した場合よりも悪い結果が出てしまった．それに対して，「楽典的な意味を表す特徴量」を採用した場合には，満足度の比が 1.51 という良好な結果が得られた．よって我々の実験環境においては，楽典的な意味を表す特徴量を採用するほうが望ましい，という結論が得られた．

図 3　2 名の被験者におけるアイコン表示結果．

表 2　被験者の満足度．

被験者	A	B	C	D	E	F	G
満足度 (%)	30	60	50	70	60	80	60

図 3 は，ある 2 名の被験者におけるアイコン表示結果である．100 曲の楽曲は全く同一で，楽曲の画面配置も全く同一であるが，選択されたアイコン画像が大きく異なるのがわかる．この結果から MIST が，事前学習段階における個人の回答結果にあわせて，多彩なアイコン選択をしていることがわかる．

表 2 は，7 名の被験者（A〜G と表記）の満足度と，その平均値を示したものである．この結果から，被験者によって満足度に大きな隔たりがあるのが観察される．この要因について，次節にて議論する．

4.3 フィードバックと考察

本実験では，各楽曲に対するアイコン選出結果を回答させるだけでなく，実験全体についてフリーコメントを求めた．以下，このフリーコメント内容から，今後の課題について考察する．

まず MIST に好意的な意見として，以下のコメントが得られた．

- 自分の中で印象と画像の結びつきの法則ができあ

がった．(被験者 B)

このコメントは，MIST を使い続けるうちに，徐々に「このアイコンなら，このような印象の音楽が再生されるだろう」という慣れが生じることで，MIST がユーザにとって徐々に使いやすくなる，という可能性を示唆していると考えられる．

続いて本実験に対する問題点として，以下のコメントが得られた．

- 自分だったらアイコンに使いたくないと思う画像が多かった．(被験者 A)
- 格好いい楽曲が多いのに，格好いいと思える画像がなかった．(被験者 C)
- 画像の枚数が少ない．

これらのコメントを残した被験者のうち A,C の 2 名は，他の被験者よりも満足度の低い結果を出している．これらのコメントから，本実験で満足度が伸び悩む要因があるとすれば，アイコン画像の準備が一因だったことが考えられる．今後の課題として，我々が用意したアイコン画像を使うかわりに，被験者自身に選ばせたアイコン画像を用いた実験に着手したい．

また楽曲に対しても，以下のコメントが得られた．

- 普段聞かないジャンルの楽曲は，アイコン画像が合ってないような印象を持った．それに対して，普段聞くジャンルの楽曲は，アイコン画像が合っているような印象を持った．(被験者 G)

このコメントは，事前学習段階におけるサンプル楽曲の試聴においても，普段聞くジャンルの楽曲を用いることで，満足度が向上する可能性があることを示唆するものである．今後の課題として，我々が用意した楽曲を用いるかわりに，被験者自身が既に日常的に聴いている楽曲を用いた実験に着手したい．

アイコン画像の選択結果について，以下のようなコメントが得られた．

- 9 種類の画像がまんべんなく選択されるような結果が欲しかった．(被験者 D)

我々は画像選択の過程において，単純なユークリッド距離を採用している．しかし，もし被験者 D のような感想をもつ被験者が多いとしたら，ユークリッド距離に基づく画像選択を再考し，まんべんなくアイコン画像が選択されるようなマッチング手法を適用したほうが望ましいかもしれない．

平安京ビューによる可視化結果については，以下のようなコメントがあった．

- 同じアイコンは画面上の近い位置に，まとまって表示されるほうがよい．(被験者 F)

これについては意見の分かれるところであろうと考えられる．見かけだけで判断するのであれば，多くのユーザは被験者 F と同様に，同じアイコンはまとまって表示されるほうがよいと感じるであろう．しかし，アイコンの並び順に楽曲上の意味（例えばアルバムのトラック順）があるのであれば，それを尊重してアイコンを配置すべき場合もあると考えられる．

5. 関 連 研 究

アイコンの自動生成技術

アイコン画像の自動合成に関する有名な手法として，Semanticons[4] という手法が報告されているが，これは音楽ファイルに焦点を絞った研究ではない．音楽ファイルに焦点を絞ってアイコンを自動選択する手法としては，Kolhoff らの研究[5] があげられる．しかしこの論文では，あらかじめ規定された幾何学模様状のアイコンの中からの選択，という限定的なアイコン表現を示している．MIST では，写真画像などをユーザが自由にアイコンとして採用することを前提としており，Kolhoff らよりも柔軟なアイコン表現を目指している．

音楽を色で表現するシステム

プレイリストに類似した使い方を想定した楽曲推薦システムの例として，後藤らによる Musicream[6] があげられる．Musicream は楽曲の雰囲気を色で表し，意味や印象の近い楽曲を画面上の近い位置に再配置する．また，音楽に色を適合する研究として，川野辺らの研究[7] では，楽曲を任意の 3 色で表現する手法を提案している．MIST はアイコン画像の特徴量に色を用いているという点で，これらの手法と類似している．一方で MIST は，ファイルシステムのブラウザ上でのアイコン表現という，我々が日常的に用いるユーザインタフェース上にそのまま導入することを意識した手法，という点に特徴があると考えられる．

音楽と画像の組み合わせを求める技術

大山らの提案する DIVA[8] では，画像の印象に基づいて楽曲をアレンジする．これは画像の特徴量とキーワードを基にアレンジと画像の距離を計算するという点で，MIST と関連があると考えられる．しかしながら，DIVA はアレンジ過程において楽曲の特徴量を参照していないため，原曲に合わないアレンジを提供してしまうことがある，という問題点が残っている．

音楽や画像を感性語で表現する技術

池添らは，感性語を利用して音楽の印象を表現する手法[9] を提案している．この手法では，SD 法 (Semantic Differential Method) の中から，音楽の印象として影響を強く与える感性語を決定し，楽曲と感性語適合度の関係を導出している．我々の研究において

も，これと同様な感性語を用いている．

楽曲の特徴量

楽曲から抽出できる特徴量として有効に使えるものは，本報告で MIST が採用した特徴量以外にも多々考えられる．例として，テンポの速さと揺らぎ，アクセントの強さ[10]，スペクトル変化度，周波数強度の分散[11]，周波数空間の部分領域における音量，周波数分析結果の重心や幅など[12]，コードの変化度，ダイナミクス[13] などがあげられる．

可視化システム「平安京ビュー」の応用システム

我々は既に「平安京ビュー」を，静止画像や CG コンテンツなどの一覧表示に適用している[14),15)]．これらの適用事例では，静止画像や CG 画像を，「平安京ビュー」の葉ノードである黒いアイコンに貼り付けることで，静止画像や CG 画像の一覧表示を実現している．しかし，音楽コンテンツの一覧表示への適用は，MIST が初めてである．

6. ま と め

本報告では，音楽アイコン選択手法 MIST に音響データファイルを適用した試み，またその結果の一覧表示のために可視化手法「平安京ビュー」を適用した試みを論じた．さらに，被験者実験による満足度の集計結果を示し，被験者によるフリーコメントから提案手法や被験者実験の問題点について考察した．

今後の課題として，以下のような点があげられる．

- 画像や音楽の特徴量算出手法の妥当性の検証．
- サンプル楽曲を増やすことで学習効果が高まるか，という点の検証．
- 被験者自身が持参した楽曲および画像を用いた被験者実験．
- 楽曲一覧 GUI としてのユーザビリティの向上．

謝　　辞

ニューラルネットワークに関する情報提供をしてくださったお茶の水女子大学小林一郎准教授，実験に参加いただいた被験者の皆様に，感謝の意を表します．また本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．

参 考 文 献

1) 小田, 伊藤, MIST: 音楽に印象の合うアイコンを自動選択する一手法, 第 22 回 NICOGRAPH 論文コンテスト (2006).
2) M. Oda and T. Itoh: ”MIST: A Music Icon Selection Technique Using Neural Network”, NICOGRAPH International 2007 (2007).
3) T. Itoh, H. Takakura, A. Sawada, K. Koyamada, ”Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space”, IEEE Computer Graphics and Applications, Vol. 26, No. 2, pp. 40-47, 2006.
4) V. Setlur, C. Albrecht-Buehler, A. A. Gooch, S. Rossoff, and B. Gooch: ”Semanticons: Visual Metaphors as File Icons”, Computer Graphics Forum (Eurographics 2005), Vol. 24, No. 3, pp. 647-656, 2005.
5) P. Kolhoff, J. Preub and J. Loviscach: ”Music Icons: Procedural Glyphs for Audio Files”, IEEE SIBGRAPI’06, pp. 289-296, 2006.
6) 後藤, 後藤, Musicream: 楽曲を流してくっつけて並べることのできる新たな音楽再生インタフェース, 日本ソフトウェア科学会 第 12 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2004), pp. 53-58, 2004.
7) 川野, 亀田, 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, Vol. 5, No. 4, pp. 95-105, 2005.
8) 大山, 伊藤, DIVA：画像の印象に合わせた音楽自動アレンジの一手法の提案, 芸術科学会論文誌, Vol. 6, No. 3, pp. 126-135, 2007.
9) 池添, 梶川, 野村, 音楽感性空間を用いた感性語による音楽データベース検索システム, 情報処理学会論文誌, Vol. 42, No. 12, pp. 3201-3202, 2001.
10) ミュージックソムリエ, http://www.watch.impress.co.jp/av/docs/20031222/dal127.htm
11) 大塚, 梶川, 野村, PCM データに対応した感性語による音楽データベース検索システムに関する研究, 第 14 回電子情報通信学会データ工学ワークショップ (DEWS), 2003.
12) D. Liu, L. Lu, H.-J. Zhang, Automatic Mood Detection from Acoustic Music Data, International Symposium on Music Information Retrieval (ISMIR), 2003.
13) 中澤, 白鳥, 池内, 観察に基づく音楽およびモーションキャプチャデータからの舞踏動作生成手法, 画像の認識・理解シンポジウム (MIRU2005), pp. 1137-1144, 2005.
14) 五味, 宮崎, 伊藤, Li, CAT:大量画像の一覧可視化と詳細度制御のための GUI, 画像電子学会誌, Vol. 38, No. 4, pp. 436-443, 2008.
15) 建部, 伊藤, 3 次元 CG アニメーションデータの分類結果の可視化の一手法, 可視化情報学会第 35 回可視化情報シンポジウム, 2007.