MIST：音楽に印象の合うアイコンを自動選択する一手法

小田瑞穂 [1)]　伊藤貴之 [2)]
1)　お茶の水女子大学大学院　人間文化研究科
2)　お茶の水女子大学　理学部情報科学科
E-mail: {miz-oda, itot} @itolab.is.ocha.ac.jp

概要

近年ではマルチメディア技術の発達により、計算機上で音楽を鑑賞する機会が増えている。計算機の記憶容量の増加に伴い、個人の１台の計算機に大量の音楽が取り込まれるようになった。しかし曲数が増えると、聴きたい音楽を探し出すのに苦労する場合も多い、また、同じフォルダの中に同じようなタイトルの曲ばかり入っていると、一曲ずつ聴いてみるまで曲の区別がつかないこともある。

そこで本研究では、任意の音楽に印象の合うアイコンを自動選択することを目的として、印象の合う画像と音楽を自動的に組み合わせる手法を提案する。本研究では画像や音楽の特徴を自動抽出して数値化し、その特徴から画像や音楽の印象を推測する。この推測のために本研究では、人間による印象も数値化し、特徴から印象を求める数式を用いる。この数式の決定には応答曲面法を用いる。

なお本研究は、アンケート形式により大衆の印象に合致する組み合わせというより、個人の、一人一人の印象に合致する組み合わせを提示することを目的としている。本報告ではその研究成果として、いくつかの画像と音楽の組み合わせを示す。

1. はじめに

マルチメディア技術は計算機の普及と発達に大きく寄与してきた。特に近年では、計算機上で画像や音楽を楽しむ技術が発達している。しかし計算機の研究開発の経緯から、画像技術と音声技術は個別に発達をとげたものであり、その統合技術に関する研究課題は現在でも残っている。

聴覚と視覚の両方に同時に刺激を与えるようなメディア鑑賞技術、例えば画像に合う BGM を提供する技術や、音楽に合う背景画像を提供する技術があれば、ユーザは今まで以上に、計算機上で画像や音楽を楽しむことができると考えられる。現在でも例えば、Windows Media Player に代表されるように、音楽に同期した映像を表示する技術を搭載したメディア観賞用ソフトウェアは存在する。しかしながら、視覚と聴覚の双方の印象があまりにも違いすぎると、メディアの印象が却って不鮮明になると考えられる。よって、このようなメディア観賞用ソフトウェアにおいては、視覚と聴覚の印象を合わせる、という点が重要になると考えられる。

また計算機では、視覚メディアを使いやすくするために聴覚的にアシストする、逆に聴覚メディアを使いやすくするために視覚的にアシストする、という方法はよく使われている。例えばウィンドウシステム上では、マウスなどを使った視覚操作にブザー音などを連動することで、ユーザはウィンドウシステムを直感的に利用できる。また例えば、音楽ファイルはウィンドウシステム上では特定のアイコンで表示されるために、ユーザはそれが音楽ファイルであることを認識できる。ここでもし、音楽ファイルのアイコンが画一ではなく、音楽の印象に合わせた画像が使われていたとしたら、その曲ごとのアイコンに設定することで、多くの曲を一覧表示することが可能であると考えられる。アイコンと曲の印象が近ければ、一目見ただけで、このアイコンの曲がどのような曲なのか、すぐに推測できる。

本報告では、任意の音楽に印象の合う画像を、計算機で自動選出する手法を提案する。本手法では、画像や音楽の特徴を自動抽出し、「特徴値」という形で数値化する。また同時に、人間による画像や音楽の印象を、「評価値」という形で数値化する。本手法の準備段階では、ユーザにサンプル画像やサンプル音楽を提示し、その印象を評価値とする。続いて応答曲面法を用いて、特徴値から評価値を算出する式を導出する。本手法の実行段階では、任意の画像や音楽を提示すると、計算機はその特徴値を自動算出し、続いてその評価値を自動算出することにより、画像や音楽の印象を推定する。ここで評価値の距離の近い画像と音楽を組み合わせることにより、印象の近い画像と音楽を自動選出できると考えられる。

筆者らは本手法を、音楽に印象の合うアイコンを自動選出する目的で実用できると考えている。そこで本報告では本手法を、MIST (Music Icon Selector Technique) と称する。ただし、本報告で提案する手法の原理は、音楽から画像を選ぶ、画像から音楽を選ぶ、という両方の目的に適用が可能である。

2. 関連研究

楽曲に関する関連研究として、大山らは画像に印象の合うように音楽を自動アレンジする手法を提案している[1]。また印象に基づいて楽曲を検索するシステムについての研究がある[2]。しかし、この研究では画像と音楽の組み合せは実現していない。また、類似度の高い曲に類似度の高い色を割り当てるインタフェースの研究[3]も存在しているが、色だけでなく別の要素も組み込むことを可能にしたい。

画像に関する関連研究としては、画像メディアデータから、メタデータを自動抽出する方式を実現させ意味的画像検索への応用を可能にした[4]。メタデータの抽出には人間の感性や感覚を解釈させるため意味的連想検索方式を入れることで、より直感に合致する検索を可能にした。しかし、この研究では音楽と画像を組み合わせるという目的ではない。

3. 提案内容

3.1 概要

本報告の提案手法は、図 1 に示すように、2 つの準備段階と 1 つの組み合わせ作成段階の 3 段階から構成される。ここで本手法では、「特徴値」という数値をもって、画像および音楽の特徴を抽出する。また「評価値」という数値をもって、画像および音楽のユーザによる印象を表現する。ここで特徴値および評価値の定義は以下のとおりである。

[特徴値:] 画像および音楽の特徴を表す値で、コンピュータによる自動抽出が可能なもの。

[評価値:] 画像または音楽に対して、人間による主観的な印象を表す値。本研究では、画像や音楽の印象を表す数種類の形容詞を用意し、その形容詞の各々に対して適用度を主観評価させた値を評価値として用いる。例えば「軽い」という形容詞に対し、与えられた画像や音楽の適用度を 10 段階のいずれかの値で主観評価した値を用いる。よって本研究では、用意した形容詞の数だけの評価値が用いられる。なお本研究では、画像と音楽の両者に対して同じ形容詞を用いるものとする。

本手法の準備段階は、どの特徴値や評価値を使用するかを決定する相関図作成段階と、特徴値から印象を自動推測するための関係式作成段階の 2 つに分けられる。そして、その後の組み合わせ作成段階にて、初めて画像と音楽の組み合わせ結果を得る。以下に各々の処理について解説する。

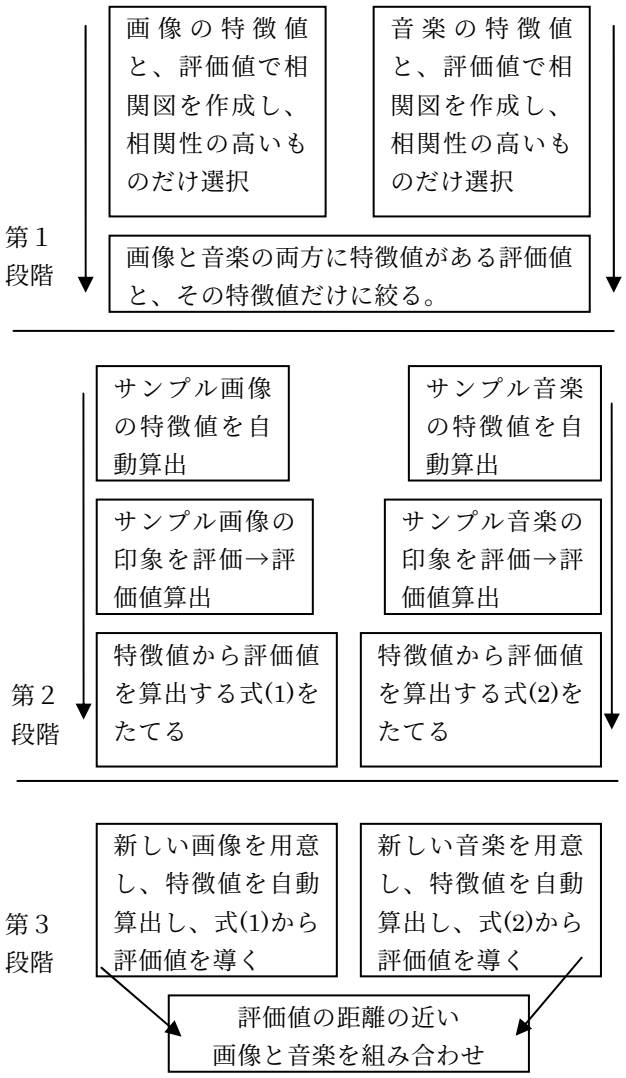

図 1. 提案手法の構成

[第 1 段階(相関図作成):] 画像や音楽のさまざまな特徴が、どのような印象に大きく影響を与えているか、その相関性を調べる。相関性の高さは、特徴値を横軸、評価値を縦軸とするグラフを作成して判断する。本手法ではさまざまな評価値の中から、画像と音楽の両方の特徴値に対して相関性が高い評価値だけを選択的に利用する。

[第 2 段階(関係式作成):] サンプル画像とサンプル音楽をユーザに提示し、その印象を評価値として回答させる。また、サンプル画像とサンプル音楽の特徴値をあらかじめ算出する。特徴値、評価値、第 1 段階での結果を利用して、応答曲面法を用いて、特徴値から評価値を推測する式を立てる。

[第 3 段階(組み合わせ作成):] ユーザが任意の画像と音楽を用意すると、計算機はその画像と音楽の特徴値を算出する。続いて第 2 段階の応答曲面法によって導き出された式から、その画像と音楽の評価値を算出する

ことで、それらの印象を推定する。この評価値の距離が近い画像と音楽を組み合わせることで、計算機は印象の近い画像と音楽の組み合わせを自動提示する。

3.2 特徴値と評価値の選定

画像と音楽にはさまざまな特徴があり、さまざまな印象を与えているが、そのすべてを実装に用いるのは難しい。特徴値および評価値を多く採用することは、第 2 段階で応答曲面法を解くために必要な応答式の数が増大する。つまりサンプル画像とサンプル音楽を多く用意する必要が生じ、ユーザによる回答の手間が大きくなる。よって使用する特徴値および評価値の種類は、ユーザによる回答の手間を増大しすぎない程度の種類にとどめる必要がある。そこでユーザに回答させる評価値は、特徴値と相関性の高いもののみに絞ることが望ましい。

本手法の第 1 段階では、ひとつの評価値に影響を与えている特徴値を探すために、特徴値を横軸に、評価値を縦軸にとり、散布図を作成する。評価値ごとに相関性の高い特徴値を決定する。これは検定による自動選択も可能だが、本研究では視覚的に判断をした。具体手金は、散布図の点が点在するような場合には相関性が低く、ひとつの関数で近似できるような場合を相関性が高いとし、それぞれの評価値ごとに相関性の高い特徴値を選定する。

画像と音楽の評価値は共通のものを使う。よって第 1 段階では、画像と音楽の両方の特徴値に対して相関性の高い評価値だけを選択することで、少ない回答で適切な組み合わせを得る。

3.3 特徴値から評価値への定式化

本報告では、上述した特徴値と評価値を、以下のように表現する。

- 画像の特徴値： $P(p_1, p_2, ..., p_l)$
- 音楽の特徴値： $M(m_1, m_2, ..., m_m)$
- 両者共通の評価値： $A(a_1, a_2, ..., a_n)$

（ただし l,m,n は正整数）

ここで上述の通り、提案手法の本処理では、特徴値から評価値を推定する数式が必要である。本報告では、この数式を以下のように表現する。

- P から a_k を求める数式： $a_k = f_k(P)$ …(1)
- M から a_k を求める数式： $a_k = g_k(M)$ …(2)

提案手法の第 2 段階は、特徴値と評価値の相関性を表現する式(1)(2)の関数 f,g を求めることに相当する。著者らの現時点での実装では、関数 f および g の導出に応答曲面法を用いている。応答曲面法とは、実験から得られる数値の相関性を数式化するために、実験数値に対して最小二乗曲面を生成する手法である。ここで応答曲面法では、応答関数に n 次多項式を用いると、応答曲面法を立てるのに、n!個の応答式が必要となる。ここで仮に n=4 とすると、必要な応答式が 120 個と多大になってしまい、第 2 段階の負担が大きくなってしまう。そこで本報告では、n=2 として実験を行う。

本報告では式(1)(2)における関数 f,g を、応答曲面法による以下の数式により表現する。

$$a_k = \beta_0 + \sum \beta_i x_i + \sum \beta_{ii} x_i^2 + \sum \beta_{ij} x_i x_j \quad (3)$$

a_k：ユーザが回答した、画像や音楽の k 番目の評価値
x_i：計算機が算出する画像(P)や音楽(M)の特徴

提案手法の第 2 段階では、ユーザの回答結果である a_k と、計算機が算出する x_i を与え、その点群の近傍を通るような曲面を生成できるように係数 β_0、β_i、β_{ij} を決定する。

本手法では最初に準備段階として、s 個のサンプル画像と t 個のサンプル音楽を用意し、ユーザの回答結果として評価値 a_1～a_k を決定する。つまりここには(s+t)個の評価値が存在することになる。

それと同時に本手法では、s 枚のサンプル画像における特徴値 $P(p_1 \sim p_l)$を算出する。これを(1)に代入して、 f_1～f_k の各々に対して応答曲面法を適用する。同様に本手法では、t 曲のサンプル音楽における特徴値 $M(m_1 \sim m_m)$ を算出する。これを(2)に代入して、g_1～g_k の各々に対して応答曲面法を適用する。

以上により a_1～a_k を求める数式が完成する。これ以降、ユーザが任意の新しい画像や音楽を追加しても、a_1～a_k を求める数式(1)と(2)を用いて評価値を算出することで、計算機は画像や音楽の印象を推測できる。

3.4 特徴値と評価値の具体例

本研究で用いた特徴値と評価値の一覧を、表 1 に示す。著者らの現時点での実装では、画像や音楽のデータ形式には、特徴値を数値として抽出しやすいものを選んでいる。そこで画像データ形式は SVG を採用し、画像データ中に数値として直接記述されている形状情報や色彩情報をもとにして特徴値を抽出した。また音楽データ形式には MIDI を採用し、音楽データ中に数値として直接記述されている音符情報やメタ情報をもとにして特徴値を抽出した。また評価値を選定するために著者らは、SD 法（Semantic differential method）[5]が提示する形容詞群の中からいくつかを抜粋した。サンプル画像およびサンプル音楽をユーザに提示する際には、これらの形容詞群がどのくらい適合するかをユーザに回答させ、それを 10 段階に数値化したものを評価値とした。

表 1. 特徴値と評価値の一覧

画像 P(p_1,…)	評価値 A(a_1,…)	音楽 M(m_1,…)
色相	明暗	休符含有率
彩度	静動	平均音程
明度	重軽	テンポ
面積	攻防	単位時間当音数
要素数	華素	
	幸悲	
	硬柔	
	複単	

3.5 評価値の近い画像と音楽の組み合わせ

提案手法の第 3 段階では、評価値をベクトルであるとみなし、その距離が小さい画像と音楽を「印象が近い画像と音楽」と判断する。例えば A の項目が 3 個のとき、画像の評価値を $A_1(a_{11},a_{12},a_{13})$とし、音楽の評価値を $A_2(a_{21},a_{22},a_{23})$と仮定する。3 次元座標系で A_1 と A_2 を座標にプロットし、画像に対して印象の近い音楽を選ぶには 2 点の距離が最小である点を探し出せばよいことになる。2 点間距離は$(a_{11}-a_{21})^2+(a_{12}-a_{22})^2+(a_{13}-a_{23})^2$ によって求められるので、距離が最小のものを選び出せばよい。A の項目を 3 項目から n 項目に拡張したときには、

$$2\text{点間距離} = \sqrt{\sum_{k=1}^{n} (a_{1k} - a_{2k})^2}$$

の値が最小になる画像と音楽の組み合わせを選べばよい。

4. 実行結果

4.1 第 1 段階(相関図作成)

表 1 に挙げた特徴値と評価値にたいして、相関図を作成した。この相関図は被験者ごとに作った。被験者 A の相関性の高いものだけを選定した結果を表 2 に示す。

表 2. 被験者 A の第 1 段階の選出結果

画像の特徴値	評価値	音楽の特徴値
明度,彩度,面積	静的な⇔動的な	音程,音数
明度,彩度,数	重い ⇔軽い	テンポ,音程
明度,彩度,面積	攻撃的⇔保守的	音程,音数
明度,彩度,面積	華やか⇔素朴な	テンポ,音程,音数
彩度	硬い ⇔柔らか	休符含有率
彩度,明度	明るい⇔暗い	テンポ

さらに、応答曲面法を利用するにあたって、画像と音楽の特徴値を各 2 個以上必要とするため、表 2 の下 2 項目は候補からはずした。したがって、被験者 A は 4 つの評価値によって本研究を実行した。

画像と音楽の特徴の中にも、表 1 に上げた特徴値と評価値には入れなかった項目も存在する。さらに、今回使用しなかった特徴値や特徴値がある。これらは更なる組み合わせの精度向上につながると思われるが、現在では保留としている。

4.2 第 2 段階(関係式作成)

いくつかのサンプル画像（図 2）とサンプル音楽（図 3）を、数名の被験者に評価をしてもらう。第 1 段階で選定された評価値と特徴値の関係に沿って、計算機が自動で算出する特徴値 P と M と、被験者ごとの評価値 A を応答曲面法に代入することで、関数 f と g を求めることができる。被験者ごとに、評価値の項目ごとに応答曲面法の式を立てることになるので、被験者 A の場合は、関数 f が 4 項目、g が 4 項目の合計 8 個の式を持つことになる。

ここで、被験者ごとに違う数式を立てているのは、被験者の好みを反映させやすくするためである。

図 2. サンプル画像

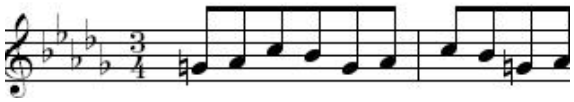

図 3. サンプル音楽

4.3 第 3 段階(組み合わせ作成)

サンプル画像やサンプル音楽とは違う、画像と音楽をそれぞれ数枚ずつ用意し、被験者ごとに立てた式を利用して、印象値を算出し、組み合わせをする。被験者 A の場合は、第 1 段階の結果から 4 つの印象値を利用したので、4 つ印象値の軸として、すなわち 4 次元空間に、画像と音楽をプロットし、距離の近いものを組み合わせる。完成した組み合わせに対して、満足できるかどうか、再び同じ被験者に判定をしてもらう。

実験は、音楽に対して画像が近いものを組み合わせる方法と、画像に対して音楽が近いものを組み合わせる方法の 2 種類がある。

ところで、組み合わせるために利用する距離とは、距離を測定したい、1 枚の画像と一曲の音楽の両者に対し、4 項目の評価値 a_1～a_4 をベクトルとみなし、その 2 点間距離を求める。4 つの評価値は、それぞれ 0～10 ま

での値をとるので、最短で距離 0 を取り、最長で距離 sqrt(10×10×4)=20 を取る。

(1)音楽に対して近い画像を組み合わせる
印象値を軸とした４次元空間にまず画像を配置する。次に音楽を一曲持ってきて、配置し、その曲とすべての画像との距離を最小 2 乗法により求める。求めた距離の中で一番小さい距離をもつ画像が、その音楽と組み合わさるべき画像。

(2)画像に対して近い音楽を組み合わせる
(1)とほぼ同様で、印象値を軸とした 4 次元空間にまず音楽を配置する。次に画像を一枚配置し、その画像とすべての音楽との距離を求める。求めた距離の仲で一番小さい距離をもつ音楽が、その画像と組み合わさるべき音楽。

(1)の方法で組み合わせを行ったものが、音楽に対して印象の近い画像をアイコンとすることができる、本研究の MIST という仕組みにあたる。

また、サンプル画像やサンプル音楽ではなく、さらに 4.3 で用意した画像や音楽でもない、さらに新しい画像や音楽を容易に追加することができる。

4.4 実行結果の考察

本研究の実験結果では、被験者ごとにさまざまな結果を得たが、その一因としてサンプル画像・サンプル音楽の評価値に個人差が大きいことがあげられる。被験者 A は評価値のばらつきが大きいため、画像と音楽の距離が大きくなる傾向にあった。一方、被験者 B は評価値のばらつきが小さいので、画像と音楽の距離が小さくなる傾向にあった。従って画像と音楽の距離を絶対値として扱うことが難しかった。本手法を「個人の嗜好に合致する手法」ではなく「アンケート形式で大衆の嗜好に合致する手法」にするためには、この問題を克服する必要があると思われる。

また、サンプル画像やサンプル音楽の印象に偏りがあり、特定の画像ばかりが選択されてしまう結果があった。これは、サンプル画像やサンプル音楽の特徴値の分布が極度にばらついている場合や、分布の密度が極端にばらついている場合に起こりえる。この問題を解決するためには、サンプル画像やサンプル音楽の選び方にも検討が必要であると考えられる。さらに、サンプル音楽に短調の曲ばかりを使用したため、長調の曲の組み合わせ結果ばかりがよい結果を得られない、サンプル画像の色相と違う色相を新画像として持ち込むとよい組み合わせ結果を得られない、いう被験者もいたので、引き続きサンプル音楽と画像については十分な検討をしていきたい。

また、実験の精度を上げるためには、サンプル数を増やさなければならないと考えるが、ただサンプル数を増やすと被験者一人一人の負担が増える。また、実際に使用する時点でも、準備段階からユーザが評価をしなければいけないので、手軽さという面を考慮すると、サンプル数は少ないままで、精度の高い評価値を被験者から得られるような工夫が必要となる。応答曲面法の利用のしやすさのために、ユーザの評価を１１段階にしているが、その点についても解答しやすさを考慮したい。

5. まとめ

本研究では、音楽の印象に近い画像をアイコンとして組み合わせる手法を提案し、いくつかの組み合わせ結果について検討した。

今後の課題として、結果の精度を高めるために、表２の下２項目の評価値についても応答曲面法を適用できるようにしたい。また、音楽については、数値以外の特徴（例えば音色や調など）や、1 曲の中での曲想の変化などを考慮できるようにしたい。そのためには、応答曲面法ではない方法を用いて、式を立てる必要があるので、その方法も検討していきたい。さらに、画像については、今回使用しなかった特徴値や数値以外の特徴値(例えば色相など)はもちろん、静止画だけでなく CG アニメーションなども対象にしたい。

さらに、本研究は、ユーザの印象に合うような結果を出すので、具体的な結果をここに載せても、ユーザ以外の人にとっては、満足のいく組み合わせでないことが多々ある。したがって、具体的な組み合わせ結果の満足具合を提示する方法についても検討していきたい。

また、アイコンとなる画像を選ぶのではなく、本研究の成果を利用して、作り出すことを今後の目標としていきたい。画像のすべての特徴値を一つずつずらして、何千枚もの画像をあらかじめ用意しておくことが非合理的であると考えられるからだ。画像の選択をやめることは、形や色など、ユーザの要望を取り入れやすくすることも可能である。

謝辞

応答曲面法は京都大学小山田耕二教授からご提供いただきました。この場をお借りして御礼申し上げます。

参考文献

[1] 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, 第 68 回情報処理学会全国大会、2006.

[2] 熊本, 太田, 印象に基づく楽曲検索システムにおけ

る程度語の理解, The 18th Annual Conference of the Japanese Society for Artificial Intelligence, 2004.
[3] 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, WISS2004, pp. 53-58, 2004.
[4] 中西, 北川, 清木, 画像メディアデータを対象としたメタデータ自動抽出方式の実現とその意味的画像検索への応用, DEWS2002.
[5] Osdoog, C. E., Suci, G. J., Tannenbaum, P., The measurement of meaning, Univ. Illinois Press, 1957.