NICOGRAPH 2016, pp. 177-184

動画特徴量からの印象推定に基づく動画 BGM の自動素材選出

清水柚里奈 [1)]　菅野沙也 [1)]　伊藤貴之 [1)]　嵯峨山茂樹 [2)]　高塚正浩 [3)]

1)お茶の水女子大学大学院理学専攻　2) 明治大学総合数理学部　3) シドニー大学工学部

Automatic Background Music Composition Based on Impression Estimation of Input Movies

Yurina Shimizu[1)]　**Saya Kanno**[1)]　**Takayuki Itoh**[1)]

Shigeki Sagayama[2)]　**Masahiro Takatsuka**[3)]

1) Ochanomizu University　2) Meiji University　3) The University of Sydney

{yurina, saya, itot} (at) itolab.is.ocha.ac.jp　sagayama (at) meiji.ac.jp
masa.takatsuka (at) syadney.edu.au

アブストラクト

撮影した動画に BGM を付与することで，昔の思い出を振り返る，SNS に公開する，といった形で動画を楽しむ新しいスタイルが生まれ，またそれを支援するアプリも増えてきた．しかし動画に印象の合った音楽を選ぶ作業は必ずしも簡単ではない．本報告では，動きや色，被写体のキーワードといった動画特徴量から印象を推定し，その結果に基づいて選出されたメロディとリズム進行をマッシュアップする形で楽曲を生成することで，動画に印象に合った楽曲を付与する手法を提案する．まず動画から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．また事前に用意したメロディとリズム進行についても同様に印象を推定する．そして動画の印象値と最も類似度の高いと推定されるメロディとリズムを選出し，それらを合成する．さらに音色，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．以上の処理により，印象に合った音楽を自ら探すことなく動画に付与することができる．

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. However, it is sometime bothering to manually select background tunes those impressions are close to ones of the given movies. 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 and 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

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．これらの作業が自動化されることで，BGM を付与した動画の投稿がより手軽になると考えられる．

動画に楽曲を付与する従来の研究には，ユーザまたはシステムがあらかじめ用意した楽曲の中から動画の印象にあったものを選曲する手法が多い．また動画の印象には色・動きといった低水準な特徴に左右される場合もあれば物体やストーリーといった高水準な情報に左右される場合もあるが，その両者を考慮して動画に楽曲を付与する手法はほとんど見当たらない．また従来の研究には，どのユーザに対しても 1 つの動画に対して固定的に同じ楽曲を付与するものが多い．逆に言えば，各ユーザの嗜好を反映させて各ユーザ向けの楽曲を付与する手法は少ない．これらの観点から本報告では，動画の印象に合った楽曲を付与する一手法を提案するものである．

本報告では，動画の印象に合った楽曲を自動付与することを目標として，動画特徴量からの印象推定結果に基づいて楽曲を生成する手法を提案する．本手法では動画と楽曲の印象を表現する印象語対を数組用意し，各々の印象語対への適合度（以下「印象値」と称する）という実数値ベクトルを用いて動画と楽曲の印象を表現する．そして動画や楽曲の印象値を算出する回帰式を各ユーザに対して導出することで，動画と楽曲の印象を推定する．そして，入力動画に印象が近いと推定されるメロディおよびリズムを合成することで，入力動画のための楽曲を生成する．

2. 関連研究

静止画や動画に印象の合う音楽を提供する研究は旧来から多く発表されている．静止画や動画，アニメーショングラフィックスを含む映像作品に BGM を付与する研究では，予め用意された楽曲の中から，映像に合った楽曲を推薦する手法[2][3][4][5]や，MIDI ファイルの音符情報を自動的に編集することで楽曲を生成する手法がある．楽曲推薦において，Dunket らは写真のスライドショーに合わせた楽曲を推薦する手法[6]を提案している．この研究では，画像の顔認識を適用することで，人の年齢や表情といった高レベルな特徴量を抽出している．Feng らはホームビデオの色や動きといった低水準な特徴量をもとに，複数の楽曲を選出し，それらを統合することで楽曲を生成する手法[7]を提案している．しかし，このような楽曲推薦で対象としている楽曲は音響データを使用しているため，自由に音楽を編集しなおせるとは限らない．

MIDI 等の楽譜情報を前提として楽曲を生成する手法の例として，画像の色分布や対象物からその印象を推測し，それに合わせて楽曲をアレンジする手法[8]がある．また，色や動きといった低水準な動画特徴量から音楽特徴量を計算することで楽曲を自動生成する手法[9]がある．これらと違って本研究では，ユーザの感性を学習し，ユーザの好みを反映させて楽曲を生成する．

3. 提案手法の処理手順

提案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，高水準情報の抽出

(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．処理手順の全体像を図 1 に示す．詳細について以下に論述する．

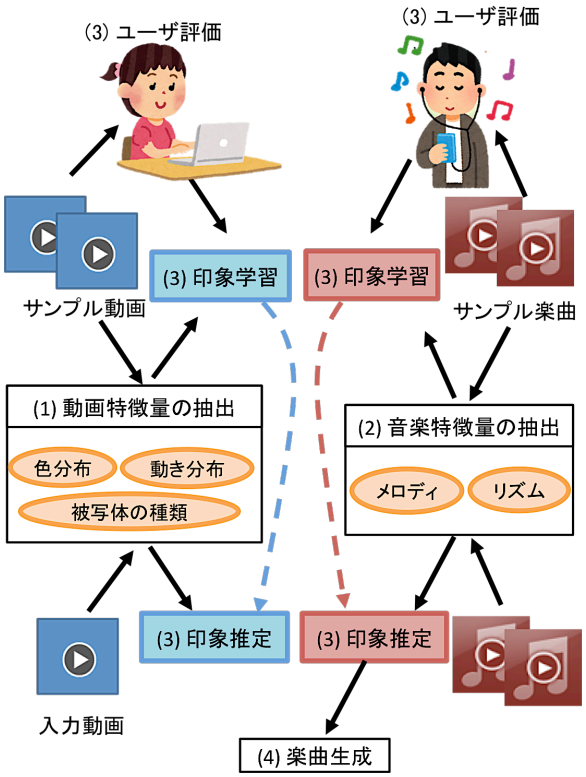

図 1：処理手順の全体像

3.1 動画特徴量

画像から受ける印象には個人差がある．文献[8]の実験結果からは，色などの低水準な特徴から印象を受ける人と，被写体などの高水準な特徴から印象を受ける人に二分されることが示唆されている．そこで，現時点での我々の実装では，色分布，動き分布の 2 種類の低水準特徴量と，被写体の種類にもとづく高水準情報を抽出し，これらの特徴量と印象との関係を学習している．

色分布，動き分布の特徴量抽出に関して，文献[1]時点での実装では，色分布や動き分布を一定時間ごとに抽出していた．この抽出を改良するために現在の実装では，動画変換ライブラリ FFmpeg (http://www.ffmpeg.org/) を用いて入力動画像を「I フレーム」と呼ばれるキーフレームごとに分割し，分割された各動画に対して画素値およびオプティカルフローを求めることにし

た．映像中の重要なキーフレームを単位として特徴量を抽出することで，より動画の特徴を捉えることができる．特徴量抽出のためのそれ以降の処理に関して，以下に示す．

3.1.1 色分布の特徴量抽出

まず動画から I フレームごとに静止画を抽出し，その静止画の各々に対して OpenCV を用いて 12 色（黒，灰色，白，茶色，赤，オレンジ，黄色，緑，水色，青，ピンク，紫）への減色処理を施し，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．得られたそ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．

3.1.2 動き分布の特徴量抽出

まず動画を I フレームごとに分割し，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．次にその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，各々のヒストグラムを生成する．そして速度の平均・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．各特徴量の全体の平均を求め，これら合計 5 つを動きの特徴量とみなし，5 次元の特徴ベクトルとする．

3.1.3 高水準情報の抽出

現時点では，動画から認識されると思われる被写体のうち，最も視聴者の印象に残ると思われる 1 つの被写体を主観的に選び，そのキーワードを手動で付与している．

今後はこの処理を自動化するために，動画からの一般物体認識技術を適用して，被写体のキーワードを動画に自動付与したい．一般物体認識技術の例として，DeepBeliefSDK（https://github.com/jetpacapp/DeepBeliefSDK）などのディープラーニング手法を適用したい．また，一般物体認識結果から，視覚刺激のサリエンシーに基づく「印象に残る被写体」を特定することで，キーワードを効果的に選別したい．

3.2 音楽特徴量

現時点での我々の実装では，メロディとリズムを別々の素材として用意し，それを合成することで楽曲を提供する．現時点で我々は，メロディとリズムに対してそれぞれ，文献[10,11]を参考にして図 2 に示す音楽特徴量を算出している．以上により，メロディについて 9 次元の特徴ベクトル，リズムについて 7 次元の特徴ベクトルを用いる．

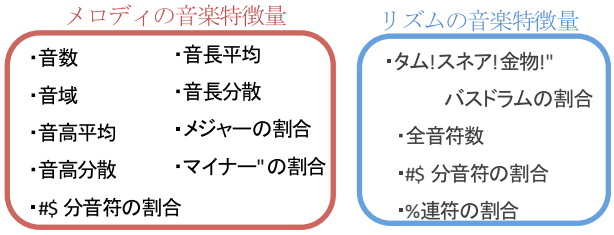

図 2：メロディ・リズムの音楽特徴量

3.3 学習

続いて本手法では，低水準の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．高水準の動画特徴量に関しては，word2vec を用いて単語を定量化することにより，印象値を求める．

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[12,13,14]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．

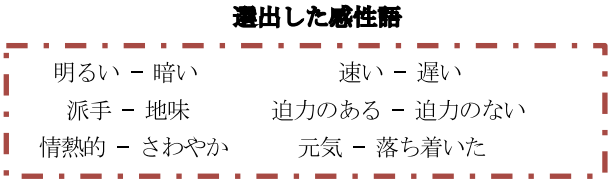

図 3：選出した感性語

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．

図 4：動画の色・動き，リズム・メロディに関する感性語

以前の実装[1]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，上にあげた感性語への適応度を 6 段階評価で回答してもらった．例えば「明るい－暗い」の場合，1 が最も暗い，6 が最も明るい，と評価することで，ユーザごとの印象値を収集した．しかし，ユーザに回答してもらった際に評価システムに関して以下のコメントを頂いた．

- 評価基準がないと判断が難しい
- 直前・直後に鑑賞したサンプルと比較してしまう

このような問題点を解決するために，評価対象となる動画，楽曲を順に並べてもらうことで相対評価を可能とする評価システ

ム（図 5 参照）を開発した．

図 5：ユーザ評価システム

このシステムの使い方は以下の通りである．ユーザに鑑賞してもらったサンプル動画・メロディ・リズムを，図 3 に記載されている感性語への適応度順に，スライダーを用いて並べる．そして，そのスライダーの位置から各動画・メロディ・リズムの印象値を算出する．このようにして，各ユーザの印象値を収集する．

3.3.2 SOM を用いた印象学習

提案手法では動画および音楽に対して，特徴ベクトルを入力情報として印象を推定する．これを以下のように定式化する．

$$a_{ci} = f_{ci}(\boldsymbol{x}_c)$$
$$a_{vi} = f_{vi}(\boldsymbol{x}_v)$$
$$a_{mi} = f_{mi}(\boldsymbol{x}_m)$$
$$a_{ri} = f_{ri}(\boldsymbol{x}_r)$$

ここで$\boldsymbol{x}_c, \boldsymbol{x}_v, \boldsymbol{x}_m, \boldsymbol{x}_r$はそれぞれ，色，動き，メロディ，リズムに関する特徴ベクトルである．$a_{ci}, a_{vi}, a_{mi}, a_{ri}$はそれぞれ，色，動き，メロディ，リズムに関する i 番目の印象値を示す．よってこの処理は，特徴ベクトルから印象値を算出する関数群を構築する問題に帰着される．

初期の実装[1]では付録に示すように，特徴ベクトルから印象値を算出するために線形重回帰分析を適用していた．しかし，線形重回帰分析では教師信号が少ないと誤差が大きくなる，また非線形の特性を有する応答値を適切に推定できないという制約がある．本手法ではあらかじめユーザにサンプル動画・リズム・メロディを評価してもらい，それをもとに算出した印象値を学習データとして扱うため，ユーザの負担の観点から多くの学習データを集めることは期待できない．また我々は既に，色分布の特徴量と感性語の間で線形の相関関係が見られず，重回帰分析を適用できなかったことを指摘している．

これらの問題を解決するために現在の実装では，教師なしのニューラルネットワークアルゴリズムである SOM (Self Organizing Map) [15]を適用している．この実装では，まずユーザが印象値を回答したサンプル動画・サンプルメロディ・サンプルリズムを，平均が 0，分散が 1 となるように基準化を行った特徴量と印象値から構成される多次元ベクトルで表現する．基準化する理由として，各特徴量では，値の範囲や分散の度合いが異なるために，値が小さい場合は，相対的に SOM による影響度が小さくなってしまうことがあげられる．そのため，このように基準化した値で構成された多次元ベクトルを SOM に入力することでマップを生成する．続いてこれから印象値を推定したい動画・メロディ・リズムについて，学習データによって得られた各特徴量の平均値，分散値を用いて基準化した特徴量ベクトルからマップ上の位置を特定し，その近傍にあるサンプル動画・サンプルメロディ・サンプルリズムの印象値を補間することで，印象値を推定する．

SOM を利用するにあたって，これまで使用していた重回帰分析と SOM を比較して精度検証を行った．

[実験] ある被験者 1 名のメロディ 15 曲分の学習データ（文献[1]の 3.3.1 項参照）を 10 曲の訓練データと 5 曲のテストデータに無作為分類する．そして訓練データを重回帰分析および SOM に適用し，その結果を用いてテストデータを構成する各曲の印象値を計算する．そして，重回帰分析，SOM によって計算された印象値と実際の回答値の乖離度を以下の式を用いて求める．

$$\text{乖離度} = \frac{\sum_{i=1}^{5} \left| \text{実際の回答値} X_i - \text{SOM / 重回帰分析による計算値} Y_i \right|}{5}$$

この **[実験]**を 6 回繰り返し，実際の回答値と重回帰分析および SOM で計算した値との乖離度の平均を計算した．結果の例を表 1 に示す．

表 1：実際の回答値との乖離度平均

	明るい-暗い	さわやか-激しい	落ち着いた-元気な
SOM	0.3818	0.3996	0.4937
重回帰	7.6375	6.3932	16.5378

この実験から，重回帰分析より SOM の方が印象推定結果の誤差が小さいことがわかった．この結果を受けて以降の説明では，被験者評価結果の与えられていない動画，楽曲に対して，色分布，動き分布，メロディ，リズムの印象値を SOM により推定するものとする．

3.3.3 被写体の印象推定

高水準情報である被写体のキーワードから印象値を計算するにあたって，我々は word2vec （https://code.google.com/archive/p/word2vec/）を使用した．word2vec は単語の意味ベクトルを大規模なコーパスから学習することで, 単語を定量化する．この word2vec を用いて, 被写体のキーワードと本手法で使用している感性語間の cos 類似度を計算する．現時点では学習データとして, 日本語版 Wikipedia を構成する文書群を用いている．

単語間の類似度を計算するにあたって, まず 3.3.1 項で選出した対となっている感性語（「明るい－暗い」であれば「明るい」と「暗い」）それぞれと単語間との cos 類似度 X_j，Y_jを計算する．さらに，求めた値に対して以下の計算を行うことで，j 番目の感性語対に対する印象値 a_{pj}を求める．ただし，現段階で使用している学習データには 「迫力のある」,「迫力のない」の感性語が存在しないため，この感性語対のみ「迫力」と単語間の cos 類似度を計算した値を印象値としている．

$$a_{pj} = \cos\left(\frac{\cos^{-1} X_j}{\cos^{-1} X_j + \cos^{-1} Y_j}\pi\right)$$

3.3.4 動画の印象値計算

3.3.2 項で求めた色分布，動き分布の印象値と，3.3.3 項で求めた被写体の印象値から，動画の印象値を求める．3.1 節の冒頭でも述べたように，画像閲覧者には低水準な特徴から動画の印象を受ける人と高水準な情報から印象を受ける人に二分されることが示唆されている．このことから，ユーザにあらかじめ，どちらの印象に左右されるかを自己回答してもらった上で，以下の計算式を用いて BGM を付与したい動画の i 番目の印象値 a_iを求める．ただし，a_{ci}, a_{vi}は低水準特徴量から得られた印象値とし，a_{pi}は高水準情報から得られた印象値である．

$$a_{\mathrm{i}} = s(a_{ci} + a_{vi}) + ta_{pi}$$

現時点では，低水準な特徴に印象が左右されると回答した場合は, s=0.8, t=0.2 とし，高水準な情報に印象が左右されると回答した場合は, s=0.2, t=0.8 として重みづけ計算を行うことで，ユーザの特性を印象値に反映させている．今後の課題として，どちらの印象に左右されるかを自動判定することを考えたい．

3.4 楽曲生成

提案手法では多数のメロディとリズムが既に用意されていることを前提として，その中から動画の印象値に近いと思われる 1 個ずつを選択して合成することにより，楽曲を生成する．

3.4.1 メロディ・リズムの合成

楽曲の素材となるメロディとリズムを選出する．ここで，用意されたメロディおよびリズムの印象値は，SOM を適用する（3.3.2 項参照）ことであらかじめ推定されているものとする．

新しい動画が与えられると提案手法は, まず 3.3.4 項で説明した方法でその動画の印象値を推定する．続いて，ユークリッド空間上で動画の印象値と各メロディ・リズムの印象値の距離を算出し，距離が最も近いメロディおよびリズムを選出し，動画の印象に沿った楽曲の素材とする．そしてこの選出したメロディとリズムを合成することで楽曲を生成する．続いて，生成した楽曲に，あらかじめメロディに付与されていたコード進行を加える．さらに，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．

3.4.2 音色の選択

続いて生成した楽曲のメロディの音色を自動選択する．現時点での実装では文献[16]をもとに，24 種類の楽器の各々について音色から連想される色分布を算出した．楽曲を付与したい動画の色分布と，各楽器の音色の色分布との類似度を cos 類似度推定法により求め，最も色分布の類似度は大きいとされる音色をメロディの音色として指定する．

以上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. 実行結果と考察

本章では提案手法の実行例を示し，その結果について考察する．我々は本実験のために，自動作曲システム Orpheus[17]を利用して作成した 30 パターンのメロディを用意し, リズムには文献[1]で使用した 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. また 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た.

本実験では，普段動画を閲覧する際に色や動きといった低水準な特徴に印象が左右されると回答した被験者 A と，被写体のような高水準な特徴に印象が左右されると回答した被験者 B の各々に対して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．以下の 2 種類の動画に対してメロディおよびリズムを選出した結果を表 2 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 2：動画 1,2 の楽曲生成を行った結果

	被験者 A	被験者 B	音色
動画 1	Melody23.mid Rhythm4.mid	Melody3.mid Rhythm9.mid	ハーモニカ
動画 2	Melody11.mid Rhythm21.mid	Melody27.mid Rhythm21.mid	トランペット

被写体のキーワードを動画 1 では海，動画 2 では犬として楽曲を生成した．表 2 から，被験者 A と被験者 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被験者の印象の違いを考慮した楽曲が生成されていることがわかる．この結果を被験者 A，被験者 B に評価してもらった．被験者 B からは，動画の雰囲気に合っている楽曲が生成されたという意見があった．一方で被験者 A からは，動画 1 の穏やかな海の情景であるのに対し，スネアを多用した激しいリズム素材が選ばれている，という意見があった．著者らはこの原因として，リズム素材が少

ないために SOM によって最適な素材を導くことが難しかったか，あるいはリズムの特徴量を再考する必要があるのではないかと考えている．

さらに，この被験者 A，被験者 B に対し，

(1) 本手法で生成した楽曲
(2) 従来の低水準特徴量のみを反映させた楽曲
(3) 重みづけの s 値と t 値を逆にして生成した楽曲

の 3 種類の楽曲を動画 1 に対し生成した．選出されたメロディおよびリズムを表 3 に示す．

表 3：(1)(2)(3)の楽曲生成結果

	(1)	(2)	(3)
被験者 A	Melody23.mid Rhythm4.mid	Melody25.mid Rhythm4.mid	Melody23.mid Rhythm9.mid
被験者 B	Melody3.mid Rhythm9.mid	Melody23.mid Rhythm10.mid	Melody3.mid Rhythm10.mid

この生成結果に対し，被験者 A からは(2)の楽曲より(1)(3)の楽曲の方が動画の印象にあっているが，(1)と(3)ではリズムが落ち着いて響いている(3)の方が好ましいという意見があった．また被験者 B からは, (1)も(3)も曲の雰囲気が似ていて評価しにくいという意見があった．表 3 を見ると，被験者 A によって選出された(1)のリズムと(2)のリズム, また被験者 B によって選出された(3)のリズムと(2)のリズムがそれぞれ同じである．これは, 低水準特徴量のみを反映させた場合と，低水準特徴量から得られた印象値に重みを付与して計算した場合では，リズム選択に違いが生じていないと考えられる．一方で (1)と(3)でメロディの選択結果は同一である．以上の結果から，動画の低水準特徴量と高水準情報の扱いにはまだ調整の余地があると考えられる．

そこで動画の低水準特徴量に関して，3.3.2 項で示した SOM を用いて印象値を求めた際に，この動画 1，動画 2 に対してどのサンプル動画の印象値を補間することで印象値を推定したか検証した．動画 2 に関して，被験者 A では犬が部屋で遊んでいる動画と，多くの人が流れるプールで遊ぶ様子が映った動画の印象値を補間していた．また被験者 B では犬が部屋で遊んでいる動画とお花畑の動画の印象値を補間していた．これらの動画に対して両被験者から，色彩や動きに関して似た印象を抱くとの回答を得た．一方，動画 1 では，被験者 A，B ともに多数の人が歩いているスクランブル交差点の動画の印象値を補間していた．この結果について両被験者から，色彩に関してはサンプル動画に似た印象を抱くが,「速い－遅い」,「元気－落ち着いた」といった動きに関する印象語についてはサンプル動画とは印象に差異がある，との回答を得た．このことから，SOM による印象推定精度の向上，動き分布に関する特徴量の再検討と印象語との相関関係についての検討が必要であると考える．

さらに 3.3.3 項で示した被写体の印象推定手法に関する検証, 3.3.4 項で示した式の見直し，あるいは重みづけに用いた定数値 s,t の設定方法などについて議論が必要であると考える．

この結果を踏まえて我々は，3.3.3 項で論じた被写体の印象推定結果を検証するために，いくつかのキーワードについて印象値を確認した．その例として表 4 に，動画 1 のキーワードである海，動画 2 のキーワードである犬に対する印象値を示す．

表 4：キーワードに対する印象値

感性語	海	犬
明るい-暗い	-0.064931023	0.020301235
派手-地味	-0.015323806	-0.054356361
情熱的-さわやか	0.047744734	-0.041416347
速い-遅い	0.045046726	0.009789739
迫力のある-迫力のない	0.221131	0.162828
元気-落ち着いた	0.050239425	0.070866794

著者は犬に対して主観的に「明るい」「元気」という印象を有しているが，表 4 からもそれに近い数値が算出されていることがわかる．一方で著者は海に対して主観的に「明るい」「さわやか」という印象を有するが，表 4 に示された数値は著者の印象とは近くない．このことから，word2vec に用いる学習データの検討，またユーザの主観を反映する仕組みが必要であると考えられる．また本手法では印象値の範囲を[-1,1]としているが，それに対して印象値の算出結果が極端にゼロに近い傾向にある．このことから, 感性語対の印象値算出式の再検討が必要である．

5. 現在の取り組み

現在取り組み中の課題として，動画特徴量と音楽特徴量，感性語対の見直しがあげられる．これまでの実装では先行研究で用いられてきた動画・音楽特徴量と感性語対をそのまま用いてきたが，本研究においてもこれらが有効であるか検証する必要がある．そこで我々は現在，感性語対を用いた印象評価結果と各特徴量との相関を求め，その結果から動画・音楽特徴量と感性語対を再選定しようとしている．具体的な選定方法は以下の通りである．

まず感性語対と動画・音楽特徴量の候補を選出した．表 5 に選出された感性語対と動画・音楽特徴量を示す．動画特徴量に関しては 3.1 項で列挙したものを採用した．音楽特徴量と感性語に関しては再度文献[10,11,12,13,14]を参考に選定した．この感性語対を用いて，図 5 にある評価システムを使用しユーザアンケートを実施した．アンケート回答者は情報科学を専攻する大学生・大学院生 17 名であった．評価システムに用いるサンプル動画・楽曲の選定には，まず動画・音楽特徴量にもとづいて K-means 法により動画・楽曲をクラスタリングし，各クラスタから代表動画・楽曲を選出することでサンプル動画・楽曲を選定した．これにより 50 個の動画から 10 個，40 曲のメロディから 15 曲，21 曲のリズムから 10 曲を選び出しサンプルとした．

そしてアンケート回答をもとに，感性語対に関するサンプル動画･楽曲の適合度とそれらの特徴量との相関係数を算出した．その結果，表 6 にあげた感性語対，特徴量が相関の高いものとして選出された．動画に関しては動きの特徴量と感性語対との相関が見られなかったことから，動きの特徴量として別の特徴量を検討する必要がある．以上を踏まえ，再度動画の低水準特徴量を検討したい．

表5：動画・楽曲の感性語と特徴量の候補

動画の感性語候補	動画の特徴量候補
・ 楽しい―切ない ・ 情熱的―さわやか ・ 勇ましい―かわいらしい ・ 激しい―穏やか ・ 元気―落ち着いた ・ 明るい―暗い ・ 静か―うるさい ・ 軽い―重い ・ 派手―地味 ・ 速い―遅い	・ 黒/灰色/白/茶色/赤/オレンジ/黄色/緑/水色/青/ピンク/紫 ・ 速度の平均 ・ 速度の分散 ・ 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
メロディの感性語候補	メロディの特徴量候補
・ 楽しい―切ない ・ 情熱的―さわやか ・ 勇ましい―かわいらしい ・ 激しい―穏やか ・ 元気―落ち着いた ・ 明るい―暗い	・ 音数 ・ 音域 ・ 音高平均 ・ 音高分散 ・ 音長平均 ・ 音長分散 ・ 16 分音符の割合 ・ メジャーの割合 ・ マイナーの割合
リズムの感性語候補	リズムの特徴量候補
・ 静か―うるさい ・ 軽い―重い ・ 激しい―穏やか ・ 派手―地味 ・ 速い―遅い	・ 音符数 ・ 8 分音符の割合 ・ 16 分音符の割合 ・ 3 連符の割合 ・ シンバルの割合 ・ ハイハットの割合 ・ バスドラの割合 ・ タムの割合 ・ スネアの割合

表 6：選出された動画・楽曲の感性語と特徴量

動画の感性語候補	動画の特徴量候補
・ 楽しい―切ない ・ 激しい―穏やか ・ 元気―落ち着いた ・ 明るい―暗い ・ 軽い―重い ・ 派手―地味 ・ 速い―遅い	・ 黒 ・ 白 ・ 緑 ・ 水色
メロディの感性語候補	メロディの特徴量候補
・ 楽しい―切ない ・ 激しい―穏やか ・ 元気―落ち着いた ・ 明るい―暗い	・ 音数 ・ 音域 ・ 音高分散 ・ 音長平均 ・ 16 分音符の割合 ・ メジャーの割合 ・ マイナーの割合
リズムの感性語候補	リズムの特徴量候補
・ 軽い―重い ・ 激しい―穏やか ・ 派手―地味 ・ 速い―遅い	・ 音符数 ・ 8 分音符の割合 ・ 16 分音符の割合 ・ シンバルの割合 ・ バスドラの割合

6. まとめと今後の課題

本報告では，動画からＩフレームごとに抽出した動きや色の低水準特徴量と，被写体のキーワードである高水準情報から動画の印象を推定し，その印象評価結果に近いとされるメロディとリズムを合成することにより，動画の印象に合った楽曲を付与する手法を提案した．

5 章であげた課題に加え，4 章での考察をもとに，以下を今後の課題としてあげたい．

高水準情報の自動付与．現在では動画の内容や物体を手動でタグづけしているが，一般物体認識手法の導入によりこの作業の自動化したい．その上で，一般物体認識結果にもとづく高水準情報を採用しても動画の印象を適切に推定できるかを検証したい．

低水準特徴と高水準情報の重みづけ．現状では 3.3.4 項の式における係数 s,t に 2 種類の定数値を用いており，しかも被験者の自己申告にもとづいて定数値のいずれかを採用している．被験者の入力に応じてこの値を動的に設定する方法を検討したい．

印象推定方法の精度向上．現時点での我々の実装では，限られた数のサンプル動画・サンプル楽曲への回答にもとづくユーザごとの印象推定を採用しているが，回答数が少なすぎるために SOM による印象推定がうまく機能していない状況が起こりえる．この問題点を解決する一手段として，サンプル動画・サンプル楽曲への回答にもとづいてユーザをクラスタリングし，クラスタごとに SOM を適用する，という方法を試みたい．

これらの改良の後に被験者実験を再度実施し，有効性を検証したい．

参考文献

[1] 清水柚里奈，菅野沙也，伊藤貴之，嵯峨山茂樹，“動画解析･印象推定による動画 BGM の自動生成”，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), F2-3, 2015.

[2] C. C. S. Liem, A. Bazzica, A. Hanjalic, MuseSync: Standing on the Shoulders of Hollywood, ACM International Conference on Multimedia, pp. 1383-1384, 2012.

[3] A. Stupal, S. Michel, Picasso-to-Sing, You Must Close Your Eyes and Draw, ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 715-724, 2011.

[4] A. Stupar, S. Michel, Benchmarking Soundtrack Recommendation Systems with SRBench, ACM International Conference on Information and Knowledge Management, pp. 2285-2290, 2013.

[5] J. Wang, E. Chng, C. Xu, H. Lu, Q. Tian, Generation of Personalized Music Sports Video Using Multimodal Cues, IEEE Transactions on Multimedia, Vol. 9, No. 3, pp. 576-588, 2007.

[6] P. Dunker, P. Popp, R. Cook, Content-Aware Auto-Soundtracks for Personal Photo Music Slideshows, IEEE International Conference on Multimedia and Expo, pp. 1-5, 2011.

[7] J. Feng, B. Ni, S. Yan, Auto-Generation of Professional Background Music for Home-Made Videos, International Conference on Internet Multimedia Computing and Service, pp. 15-18, 2010.

[8] 大山喜冴, 伊藤貴之, ”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] M. Nayak, S. H. Srinivasan, M. S. Kankanhalli, Music Synthesis for Home Videos: an Analogy Based Approach, Information, Communications and Signal Processing, pp. 15-18, 2003.

[10] 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.

[11] 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.

[12] 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 02 (TOD 13)), pp. 69-79, 2002.

[13] 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.

[14] 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.

[15] 高塚正浩, Ying Xin WU, ”球面 SOM のデータ構造と量子化誤差の考察およびインタラクティビティの向上”, 日本知能情報ファジィ学会誌, Vol. 19, No. 6, pp. 611-617, 2007.

[16] 赤井良行, 李昇姫, ”音色からイメージされる色彩の寒暖と音色構造の関係”, 日本感性工学会論文誌, Vol. 13, No. 1, pp. 221-228, 2014.

[17] 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/

付録：印象学習に関する以前の実装

3.1.1 項，3.2 項で示した低水準の動画特徴量・音楽特徴量から印象値を推定するために，以前の実装[1]では以下の処理を適用していた.

色分布からの印象学習

以前の実装では，以下の処理により，色分布の特徴量から印象値を推定した.

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

動き分布からの印象学習

以前の実装では，以下の処理により，動き分布の特徴量から印象値を推定した．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.

音楽特徴量からの印象学習

以前の実装では，以下の処理により，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した. 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_{mel} ＝ 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 a_{rhy} ＝ 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.