動画特徴量からの印象推定に基づく動画 BGM の自動生成

清水柚里奈　（指導教員：伊藤貴之）

1. 概要

近年，デジタルカメラやスマートフォンの普及により，写真や動画を撮影する機会が増え，またその撮影したものを Facebook や Twitter，YouTube などの SNS サイトに投稿することで，多くの人々と共有して楽しむようになった．その際に，ただ撮影したものを投稿するのではなく，撮影映像に BGM を付与するなどの動画編集も行うようになってきた．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．

そこで本報告では，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．また本手法では，ユーザの印象と動画特徴量，音楽特徴量の関係を学習させることで，動画・音楽の印象を推定することから，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り，また膨大な数の動画・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．

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

まず動画から 5 秒ごとに静止画を抽出し，その静止画の各々に対して OpenCV を用いた減色処理を施し，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．現時点で我々はこの 12 色を，色相環や誘目性を考慮に入れ，黒，灰色，白，茶色，赤，オレンジ，黄色，緑，水色，青，ピンク，紫とし，この 12 色で静止画を減色している．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．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，これら計 5 つを動きの特徴量とみなす．

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，それぞれ図 1 に示す音楽特徴量を算出している．

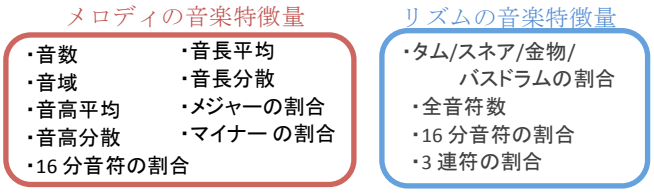

図 1：メロディ・リズムの音楽特徴量

メロディに対しては，文献[3]を，リズムに対しては，文献[4]を参考に定めた．

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．その中で動画の色・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，図 2 のように定めた．

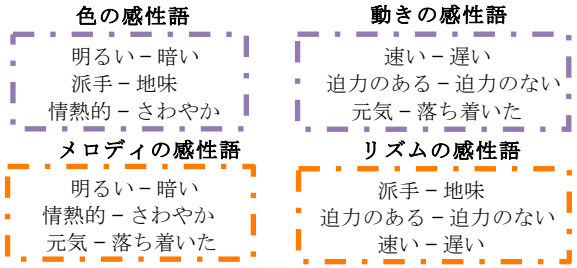

図 2：動画の色・動き,リズム・メロディに関する感性語

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・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．以後，この適合度を印象値と称する．このようにして，各ユーザの印象値を収集する．

3.3.2 色分布からの印象学習

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 3.3.1 項のユーザ印象評価で得られた 6 段階の値を[-1,1]の範囲で 6 等分した値とみなし，選ばれた値に対応する数値を印象値aとする．このようにして得られた各動画の色の印象値をa_i，抽出した 12 色の特徴量を 12 次元ベクトルでv_i，そして求めたい印象値をa_j，求めたい動画の色の特徴量をv_jとし，以下の式で印象値a_jを求める．

$$v = \sum_{i=1}^{N} a_i v_i \qquad \frac{v \cdot v_j}{|v|} = a_j$$

3.3.3 動き分布からの印象学習

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．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．

印象値 $X = a_1$×[速度の平均]＋a_2×[速度の分散]
＋a_3×[角度の分散]
＋a_4×[速度のヒストグラム極大の速度値]
＋a_5×[角度のヒストグラム極大の角度値]

3.3.4 音楽特徴量からの印象学習

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．3.3.3 節と同様，重回帰分析を用いて以下の式の係数を算出する．この式を用いて，ユーザ評価結果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値の違いを考慮した楽曲生成が可能となる．

メロディの印象値 $X = a_1$×[音数]＋a_2×[音域]
＋a_3×[音高平均]＋a_4×[音高分散]
＋a_5×[16 分音符の割合]＋a_6×[音長平均]
＋a_7×[音長分散]＋a_8×[メジャーの割合]
＋a_9×[マイナーの割合]

リズムの印象値 $Y = b_1$×[全音符数]
＋b_2×[16 分音符の割合]＋b_3×[3 連符の割合]
＋b_4×[金物の割合]＋b_5×[バスドラの割合]
＋b_6×[タムの割合]＋b_7×[スネアの割合]

3.4 楽曲生成

次に楽曲の素材となるメロディとリズムを選出し，合成する．3.3.2 項と 3.3.3 項で算出した動画の印象値と，3.3.4 項で算出したメロディ・リズム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メロディ・リズムを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組み合わせて楽曲を生成する．続いて生成した楽曲にコード進行を加える．さらに，動画の再生時間に合うように小節数やテンポを設定する．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

本手法で使用するメロディには自動作曲システム Orpheus[8]を利用して作成した 30 パターンを用意し，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．また動画は 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．

本実験ではユーザ A とユーザ 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生成を行った．以下の 2 種類の動画に対して楽曲を付与した結果を表 1 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 1：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の印象の違いを考慮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．このことから，例えば，ユーザ印象評価の改善や，動画および楽曲の特徴量の見直しなどが必要である．

5. まとめと今後の課題

本報告では動画から一定時間ごとに抽出した動きや色の動画特徴量から動画の印象を推定し，その結果に基づいて楽曲生成を行うことで，動画の印象に合った楽曲を付与する手法を提案した．

今後の課題として，学習段階におけるユーザ印象評価，動画および音楽の特徴量，印象値の推定方法などを再検討することが挙げられる．また現段階では単純な音形で付与しているコードの弾き方を，リズムや曲調に合わせて変えることも検討する．

6. 謝辞

本研究を進めるにあたり，明治大学 嵯峨山先生には楽曲生成にあたり素材を提供していただきました．ここに感謝致します．

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/