第 27 回 NICOGRAPH 論文コンテスト (NICOGRAPH2011) 論文集, I-2, pp.1-8, September 2011.

DanceReProducer: Web 上の音楽動画を再利用して新たな音楽動画を自動生成する N 次創作支援システム

中野 倫靖 [†1]　　室伏 空 [‡3]　　後藤 真孝 [†2]　　森島 繁生 [‡3]

[†] 産業技術総合研究所　　[‡] 早稲田大学

[1]t.nakano[at]aist.go.jp　　[2]m.goto[at]aist.go.jp　　[3]shigeo[at]waseda.jp

アブストラクト　本研究では、既存の音楽動画の映像部分を、新たな別の音楽に合うように切り貼りして映像を自動生成し、かつインタラクティブに編集できるシステム DanceReProducer を実現した。これによって音楽鑑賞の楽しみ方を拡張し、誰でも自分の好きな音楽に映像を付与して、視覚的にも楽しむことを可能にした。音楽動画は、楽曲（音響信号）と映像（画像の時系列）で構成され、楽曲の内容（リズムやテンポ、印象など）に、映像が時間的に同期していることが多い。従来、楽曲に映像を自動付与する研究事例はあったが、既存の音楽動画を再利用した動画生成はできなかった。そこで我々は、Web 上で公開されている二次創作のダンス動画を多数収集し、音楽と映像の多様な対応関係を機械学習によってモデル化した。次に、ユーザが与える入力音楽を分析してテンポや楽曲構造等を自動推定し、そのテンポに同期し内容が合うように、収集した動画の映像を切り出して伸縮・連結することでダンス動画を自動生成した。さらに、その生成結果にユーザが「ダメ出し」するだけで、誰でも手軽に自分の好みを反映した編集ができる新たなインタラクション技術を実現した。

1　はじめに

近年、既存の音楽・画像・動画等のメディアコンテンツを再利用して、新たなコンテンツを作成して楽しむコンテンツ鑑賞の新しい文化が形成されつつある。例えば、ヒップホップ DJ (Disc Jockey) や VJ (Video Jockey) は、新たなライブ演奏行為として過去 20 年間で既に定着し、近年では、「MAD 動画」や「マッシュアップ動画」と呼ばれる新たなコンテンツが多く制作され、動画共有サイト等で人気を集めている。マッシュアップ動画や MAD 動画では、既存の映像コンテンツを断片的に切り貼りし、別の音楽のテンポに合わせて伸縮しながら連結することで新たな音楽動画を作成する。

音楽や動画などの既存のメディアコンテンツを一次創作とすると、それらを素材として新たに作成された MAD 動画は、二次創作に位置付けられる。近年、動画共有サイトの普及によって興味深いのは、この二次創作すら素材としてさらに三次創作、四次創作がなされるような大規模な協調的創造活動 [1] が起きていることであり、これは「*N* 次創作」と命名されて [2]、学術的観点からも分析・考察がなされてきた。図 **1** に、MAD 動画制作における *N* 次創作の例を示す。

このような MAD 動画、*N* 次創作コンテンツの制作には、高度な専門知識と手間のかかる作業が必要とされ、スキルのない人々にとっては、好みの楽曲に映像を自在に付けて楽しむことはできなかった。「コンテンツ立国」という言葉に代表されるように、第 3 期科学技術基本計画 分野別推進戦略や経済産業省「技術戦略マップ 2010」等においても、コンテンツは日本の戦略産業

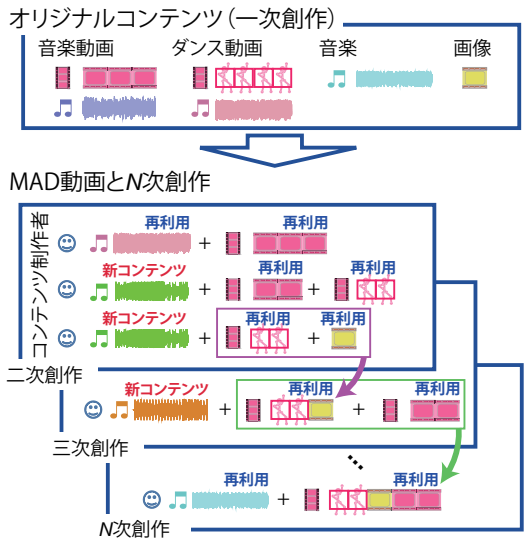

図 1: MAD 動画制作における *N* 次創作の例

分野に位置付けられており、前者ではクリエータ育成問題も指摘されている。そこで、「一億総クリエータ時代」に向け、様々な人が手軽に *N* 次創作コンテンツを制作できれば、鑑賞の楽しみが拡張された上で、さらに創作の楽しみも味わうことができる。しかし、MAD 動画の制作を支援するためには、音楽と映像のリズム・印象・文脈的な意味等を考慮した上で、それらを適切に対応付ける必要があるが、技術的に困難なために従来実現されていなかった。

そこで本研究では、図 2 のように、Web 上に多数公開されている既存の音楽動画中の映像を組み合わせて再利用して、誰でも手軽に新たな音楽動画を制作できるシステム DanceReProducer を実現した [3–5]。本シ

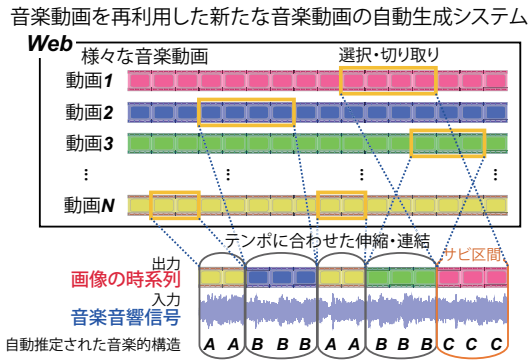

図 2: 音楽動画コンテンツを再利用して新たな音楽動画を自動生成するシステム

ステムは、実写表現も含め、内容を問わず任意の音楽動画を再利用可能だが、我々の実証実験では、音楽動画として人気の高いコンピュータグラフィックス表現を中心としたダンス動画を対象とした。そして、Web 上のダンス動画（一次創作だけでなく二次、三次創作等も含む）を多数収集して、それらの N 次創作動画の制作を支援した。

本システムの技術的特長は、「音楽にどのような映像を付与するのが適切か」を、人手でアドホックに設計するのではなく、大量の音楽動画から音楽と映像の対応付けを機械学習できる点にある。そのために、音楽と映像の特徴量の対応関係の多様さを扱える新手法を提案した。さらに、音楽と映像を同期させる上で本質的な楽曲の拍や小節、テンポ、楽曲構造も、人手で付与するのではなく、音響信号処理に基づく自動音楽理解技術によって、任意の楽曲を与えるだけで全自動で推定可能とした。これにより、音楽の拍のタイミングでダンスが変化し、映像が切り替わるだけでなく、サビの盛り上がりを反映した映像表現等も自動生成可能になった。ただし、「一億総クリエータ時代」に向けてユーザのスキルアップを促すには、全自動とは違う「自分らしさ」を反映したコンテンツ創作をいかに可能にするか、も大切である。そこで、自動生成結果の気に入らない箇所にユーザが「ダメ出し」すると他候補が表示され、選択操作だけで自分の好みを反映した編集ができる新たなインタラクション技術も開発した。これにより、ユーザはプロデューサ感覚でコンテンツ創作を楽しみつつ、音楽と映像の組み合わせの変化が、どのような印象の変化を生むのかを知ることもできる。

2 先行研究と本研究の位置づけ

MAD 動画を制作するためには、音楽のリズムや印象に合った既存動画を探し出して、それを切り貼りして音楽のリズム・テンポに合うように伸縮するといった、高度で手間隙のかかる作業が必要となる。さらに、音楽の文脈に沿って動画を構成するためには、音楽の構造や文脈を耳で聴いて判断するしかなく、それが動画制作の作業効率を下げていた。

従来、音楽に合わせた映像付与に関する先行事例がいくつか存在する。例えば近年の音楽再生ソフトウェアの多くには、音楽の周波数成分や強弱に反応した視覚効果を描画する機能がある。また様々な色や形（視覚効果）を描画する研究 [6] や、楽曲のムードの可視化 [7]、CG ダンサーなどのキャラクタモデルを音楽に同期させて動かす研究 [8,9] などがあった。また、ホームビデオを自動生成する目的で、撮影した画像を音楽に合わせて切り貼りする研究 [10,11] や撮影した映像を切り貼りする研究 [12,13] があった。しかし、これらの方法では、既存の音楽動画を再利用して音楽に合わせた動画を生成することを支援できなかった。

本論文で述べる DanceReProducer では、ユーザは好みのダンス動画をデータベースとして与え、それらを再利用してダンス動画を制作できる。ここで、データベースには一次創作動画だけでなく、N 次創作（MAD）動画も含めて、音楽に合った動画を自動生成する。さらに、自動生成された結果が好みに合わない場合には、インタフェースを通じて「ダメ出し」を行うことで容易に編集できる、**ダメ出しインタラクション**を導入する。

3 DanceReProducer の設計

本章では、システムの設計方針と、それに基づいたインタフェースの機能について述べる。なお本論文では以降、説明の便宜上、音楽付きのダンス動画コンテンツは単に**動画**と呼び、楽曲付きかそうでないかを区別するために、動画の映像部分を**映像**と呼ぶ。

3.1 音楽に合った映像についての考察

システムの設計にあたり、音楽に映像が合っていると感じられる要素を、以下に示す「局所的な対応関係」と「文脈的な対応関係」の二つの観点から考察する。動画の自動生成に関する従来研究 [12,13] や、Web 上等で公開されている二次創作の制作過程を参考にした。

局所的な対応関係 音楽と映像の印象があっていると感じる要素

- リズム ダンス動作、カメラワークや画面の切り替えによる映像のリズムが、音楽のリズムやアクセントと同期している。
- 印　象 ダンス動作、映像全体の色彩や明るさ、ライティング等の映像効果の印象が音楽の印象と合っている。

文脈的な対応関係 音楽と映像の文脈的な流れが合って

いると感じる要素

音楽的構造　音楽的構造（A/B メロ、サビ等）に合わせて映像の印象が変化する。音楽の盛り上がりに合わせて映像が盛り上がる。

時間的連続性　音楽的まとまりの境界では映像シーンが切り替わり、それ以外の箇所では映像の印象が連続している。

以上は常に満たされるわけではなく、分類が相互に独立してはいないが、これらを考慮することで音楽に合った映像が作成できると考えられる。

3.2　映像の作成方法

3.1 節で述べた対応関係を満たすためには、音楽のリズムや印象に合った映像を既存の動画群から探し出し、音楽的な文脈に沿って映像を構成する必要がある。しかし、既存の動画群から音楽に合った映像を探し出すことは、仮に動画数が数十程度であっても、その一部の断片が切り貼り映像候補となるために膨大な候補数が存在し、人手で絞り込むのは容易でなかった。また、音楽の印象に合わせて映像を切り出して、伸縮してリズムを合わせる必要があった。さらに、候補映像を繋ぎ合わせて映像を構成するためには、音楽を何度も聴きながら、その構造や文脈を把握する必要があった。

そこで本システムでは、そのような作業を効率化するために、まず入力音楽にあった映像を自動生成する。しかし、自動選出された候補は必ずしもユーザにとって好みの映像とは限らず、制作者のオリジナリティを反映させることも出来ない。そこで気に入らなかった場合には、ユーザが膨大な候補から選ぶことなく、単にシステムに「ダメ出し」をするだけで、容易に他の映像候補を選択できるようなインタフェースを提供する。

3.2.1　映像の自動生成

映像を自動生成するために、まずデータベース中の動画を音楽の小節単位で切り出して、局所的な対応関係へ対処する。これ以後、小節単位で分割された動画を**素片**と呼ぶ。音楽とそれに合わせたダンスは、曲のフレーズやダンス動作が楽曲のリズムに同期しているため、映像の切り貼りの最小単位として小節を用いた。

次に、入力音楽の各小節において、印象の合った動画素片の映像を選択する。それを入力音楽のテンポに合わせて伸縮させながら、繋ぎ合わせることで映像を自動生成する。この際、文脈的な対応関係を考慮するために、音楽的構造に基づいて、同一構造内では映像の印象が連続し（時間的連続性を保ち）、構造の切り替わり箇所では印象を変えるように別の映像を選択する。

以上の処理を 3.1 節に対応させて次に示す.

リズムの同期　入力音楽の小節に合わせて動画素片の映像を時間方向に伸縮させることで、入力音楽とダンス映像のリズムの同期を実現する。

印象の近い映像選択　動画素片の音楽と映像それぞれの印象に関する特徴量を自動抽出し、音楽と映像の対応関係をモデル化することで、入力音楽の印象にあった映像を自動選択する。

音楽的構造　音楽的構造 (自動推定したサビ区間や繰り返し区間) の境界で、映像の印象を変えて、音楽的な文脈の流れに合わせた映像を作成する。

時間的連続性　素片選択時に時間的な前後関係を考慮し、不連続な接続にペナルティを与えて、音楽的構造毎に映像の印象が連続するようにする。

以上のようにして、音楽と映像が局所的・文脈的に対応付けられた動画を生成する。

3.2.2　インタフェースによる視聴及び映像編集

図 3 に、本システムのインタフェース画面を示す。ユーザはシステムが自動生成した動画の視聴だけでなく、映像が気に入らなかった場合には、「ダメ出し」機能によって容易に映像を訂正・編集できる。

視聴のための基本機能として、現在時刻の映像の描画（図 3, ①）、音楽の読み込みや作成動画の保存（図 3, ②）、動画の再生・停止機能（図 3, ③）、再生時刻を表すスライダバーと音楽的構造の推定結果（図 3, ④）がある。ここで構造の各区間はサビが緑、それ以外の区間が青で着色して表示される。また楽曲全体を通じて、時間的に等間隔な 15 箇所のサムネイル画像が描画される（図 3, ⑤）。

また、好みに合わない映像に「ダメ出し」するために、インタフェースに以下の機能を与えた。

ダメ出し機能　NG ボタン（図 3, ⑥）をクリックすることで、映像の描画領域を四分割して候補映像を確認できる（図 4, ⑧）。これら候補映像を同時再生して見比べ、よりイメージに合った映像を選択する。この「ダメ出し」を、音楽的構造の構造区間毎に行うことで、音楽的な構造や文脈を考慮した映像を手軽に作成できる。好みの映像と出会う可能性を広げるため、印象の異なる候補が提示されるようにした。

構造の頭出し機能　見通しよく編集するため、頭出しボタン（図 3、⑦）により音楽的構造の先頭へと移動ができる。また、構造区間（図 3、④）を直接クリックしても移動できる。

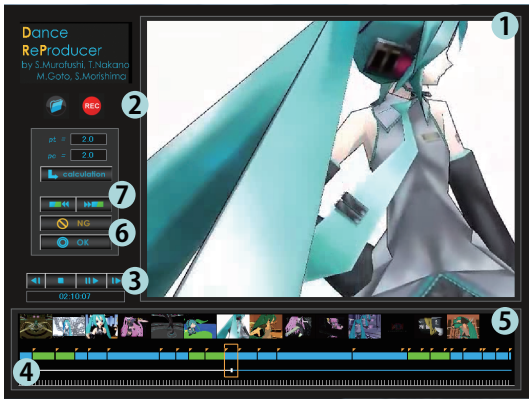

図 3: DanceReProducer のインタフェース画面

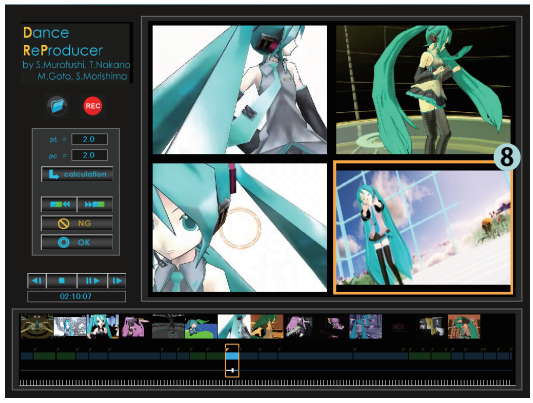

図 4: 「ダメ出し」インタラクションによって表示された四つの映像候補

4 DanceReProducer の実現方法

本システムは、音楽と映像の対応関係をモデル化し、自動推定した音楽的構造を考慮した映像の自動選出によって実現する。一般に音楽と映像との自動対応付けは困難な課題だが、本研究では既存の二次創作動画群から学習することで実現する。二次創作動画には、同一の動画素材を再利用した様々な楽曲の動画や、逆に同一の楽曲に全く別の映像が対応付けられた動画が存在する。すなわち、人手で音楽と映像が対応付けられた大量の事例を入手して利用でき、音楽と映像に関する対応付けの多様性をモデル化できる可能性がある。このように音楽と映像との対応関係の多様な解釈をモデル化することは、学術的にも意義が深い。

システムは、既存の動画群から音楽と映像の特徴量を自動抽出して保存するデータベース構築フェーズと、局所的・文脈的な対応関係を考慮しながら映像選択を行う動画生成フェーズの二段階で構成される（図 5）。本章では、上記の観点を踏まえて、図 5 を参照しながらシステムの実装について述べる。

4.1 データベース構築フェーズ

データベース構築フェーズでは、Web 上から収集した動画コンテンツ群と、そこから抽出した音楽・映像特徴量について以下のような処理を行う。

Step 1) Web から動画を収集し、映像のフレームレート (fps) を 30fps、音楽のサンプリング周波数を 44.1kHz にリサンプリングする (図 5, Ⓐ)。

Step 2) ビートトラッキングによって、音楽動画のテンポと小節線の位置を自動推定する (Ⓑ)。

Step 3) 音楽と映像の対応関係をモデル化するために、1 フレーム（約 33 ms）毎に音楽と映像特徴量を抽出し（図 5, Ⓐ）、これをフレーム特徴量と呼んで用いる。次に、推定された楽曲のテンポと小節線の位置を用いて、フレーム特徴量を小節単位にまとめて小節特徴量を得る (Ⓑ–Ⓒ)。フレーム特徴量の抽出の際、分析窓のシフト幅は映像のフレームレートに合わせて 1470 点（= 44100 Hz ／ 30 fps）とした。

以下、処理の詳細を述べる。

4.1.1 ビートトラッキングによる楽曲のテンポ及び小節線の推定

楽曲のテンポ及び小節線を推定するビートトラッキングには様々な先行研究 [8] がある。将来的にはそうした成果を利用することも検討しているが、現段階では、予備実験において比較的良い結果が得られた、音響信号のパワーに基づく簡易的な方法で計算を行った。

まず、入力音響信号のパワーの自己相関関数のピーク時刻を求める。これはパワーの周期性を表すため、これを一拍の時間長としてテンポ推定する。ただし、倍テンポ誤りや半テンポ誤りを回避するため、テンポに 60〜120bpm（一拍が 0.5〜1.0 秒）の制限を設けて推定した。続いて、推定されたテンポから一拍毎の時刻にピークを持つパルス列を生成し、それと入力音響信号のパワーの相互相関関数を計算してピーク時刻を求める。これは楽曲中の一拍目の時刻を表すため、本論文では非常に単純な手法として、一拍目を小節線の開始位置とみなし、また 4/4 拍子を仮定して、機械的に小節線の位置を決定した。

4.1.2 音楽のフレーム特徴量抽出

音楽特徴量は、音楽と映像の対応付けに関する先行研究 [14,15] と、楽曲ジャンル分類に関する先行研究 [16] を参考に、アクセントおよび印象に関する特徴量を決定して抽出した。

アクセントに関する特徴量としては、主に楽曲のパ

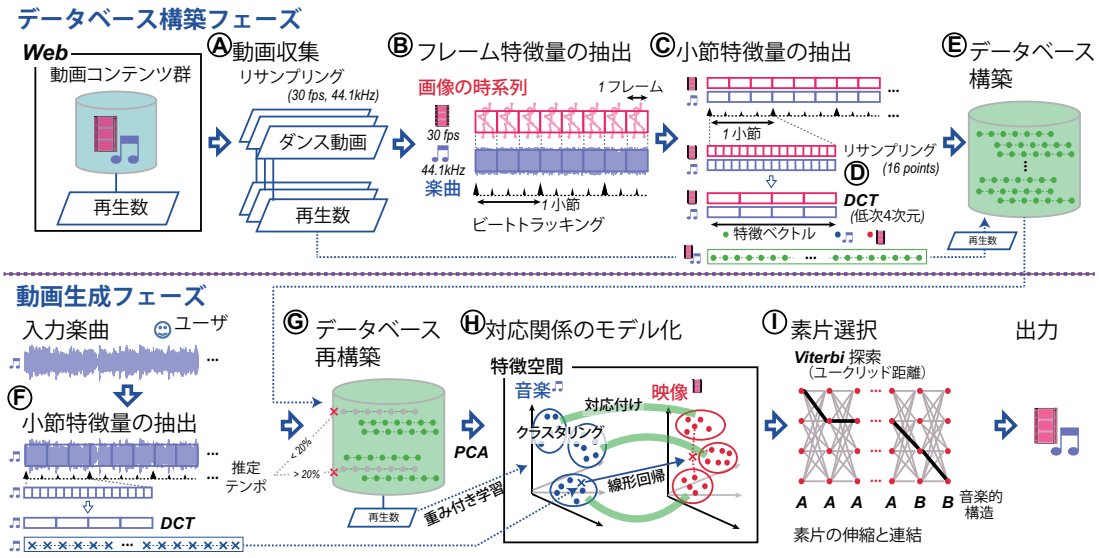

図 5: DanceReProducer の処理の流れ

ワーとその時間変化を表現するために、フィルタバンク毎のパワー（4 次元: フィルタバンク数が 4）と Spectral Flux（1 次元）を用いた。

印象に関する特徴量としては、楽曲の音色に関連した Zero-crossing rate（1 次元）と MFCC（Mel-Frequency Cepstral Coefficients）の直流成分と低次 12 項を用いた（13 次元）。

4.1.3　映像のフレーム特徴量抽出

映像特徴量には、音楽と映像の対応付けに関する先行研究 [14, 15] を参考に、アクセントおよび印象に関する特徴量を決定して抽出した。特徴抽出は、画面サイズを 128×96 にリサンプリングして行った。

アクセントに関する特徴量としては、画面の動きやダンス動作とそれらの時間変化や画面の切り替わりを表現するために、オプティカルフローと輝度値の時間微分の平均値を用いた（各 1 次元）。オプティカルフローはブロックマッチング法を用い、ブロック数 64×48、シフト幅 1、最大シフト幅を 4 として計算した。

印象に関する特徴量としては、映像の雰囲気を表現するために、全画素における色相・彩度・明度のそれぞれの値の平均と標準偏差を用いた（全 6 次元）。また映像全体の印象を表現するため、二次元の離散コサイン変換（DCT: Discrete Cosine Transform）の係数 12 次元（＝横の低次 4 項×縦の低次 3 項）を用いた。

4.1.4　小節特徴量の抽出

上述した音楽と映像のフレーム特徴量を小節特徴量としてまとめ、これを音楽と映像を対応付けるための特徴量として利用する。従来、楽曲のジャンル識別に関する研究等では、フレーム毎に抽出した特徴ベクトルを、各次元の平均と分散をとって楽曲の特徴量として用いることが多かった [16] が、そのような方法では音楽と映像の時間方向の特徴が失われてしまう。

本論文では、時間的な変化を反映するために DCT を用いて小節特徴量にまとめる（図 5, Ⓓ）。ある小節のフレーム特徴量を、時間方向に 16 点にリサンプリングして各次元毎に DCT し、DCT 係数の低次 0 ～ 3 項の計 4 次元を特徴量とした。すなわち、小節特徴量の次元数はフレーム特徴量の次元数の 4 倍となる。

4.2　動画生成フェーズ

動画生成フェーズでは以下の処理を行い、映像の選択候補を算出していく。

Step 1) データベース作成フェーズと同様に入力音楽からテンポ、小節線の推定、小節特徴量の抽出を行う (図 5, Ⓕ)。

Step 2) 入力音楽のテンポに応じてデータベースから使用する動画を選別する (Ⓖ)。これは、映像生成の際に不自然に速い（遅い）動作のダンス映像が生成されることを防ぐために、テンポが入力音楽と ±20% 以上異なる動画素片を選択候補から除外した。

Step 3) 選別後の素片群に対し、それぞれの特徴量を主成分分析 (PCA) で次元削減して利用する（累積寄与率 95%）。予備実験では、**4.1.4** 項で述べた小節特徴量が、音楽特徴で 76 次元、映像特徴で 68 次元であったものが、それぞれ 62 次元、68 次元へと削減された。ただし、入力音楽のテンポに応じて素片を選別するため、次元削減数は若干変動する.

Step 4) 次元削減後の素片群から、音楽と映像の小節特徴量から線形回帰によって対応モデルを求める (Ⓗ)。

Step 5) 得られた対応モデルを用いて、入力音楽の全小節特徴量から映像の小節特徴量を推定する。最後に、局所的な対応関係として、推定された映像特徴量とデータベース中の全映像特徴量との距離を算出する。また文脈的な対応関係として、音楽的構造毎で連続する映像となるように時間方向にペナルティを設け、これらの最適解として Viterbi アルゴリズムによる最小パスを求めて候補を決定する（Ⓘ)。

以下、処理の詳細を述べる。

4.2.1 複数の線形回帰モデルを用いた音楽–映像間の対応モデルの取得

本論文では、説明変数に音楽の小節特徴量、目的変数に映像の小節特徴量を与え、線形回帰によって対応付けを行うことを考える。しかし、前述したように二次創作動画には、同一の音楽に異なる映像、異なる音楽に同一の映像、といった多様な対応付けが存在する。しかし、単一の線形回帰のみ用いたのでは、そうした解釈の多様性を適切にモデル化できない。

そこで解決法として、複数の線形回帰モデルによって対応付けを行う。そのために、音楽・映像の小節特徴量を用いてクラスタリングし、各クラスタ毎に線形回帰を学習した。ここで、クラスタリングには k-means 法を用いた。

4.2.2 局所的な対応関係と文脈的な対応関係を考慮した映像選択

局所的及び文脈的な対応関係を考慮して映像選択するために、小節毎にコストを求めて、それが入力楽曲全体で最小となるように映像を選択する。

局所的な対応関係は、小節毎に入力音楽から推定した映像特徴量と、データベース中の全映像の映像特徴量との距離をコストとして考慮する。入力音楽に対応する映像特徴量は、**4.2.1** 項で得た回帰モデルを用いて、入力音楽の小節特徴量を映像の小節特徴空間へ写像して推定した。ここで回帰モデルには、入力音楽の小節特徴量に距離が最も近い重心を持つクラスタの線形回帰モデルを用いた。

また、文脈的な対応関係を考慮するために、生成される動画の時間的連続性や音楽的構造もコストとして考慮することで、動画像の生成をビタビ探索によるコスト最小化問題として解いた。ここで、音楽的構造とサビ区間は RefraiD [17] で求め、繰り返し区間の始端と終端を音楽的構造が切り替わる時刻として用いた。また、推定された繰り返し区間のうち、長さが四小節に満たないものは利用しなかった。

小節数が N の入力音楽について、小節番号を $n(n = 1, 2, ..., N)$、データベース中の楽曲集合 M 中の k 番目の小節を $k_m(k = 1, 2, ..., K_m, m \in M)$ で表した。小節毎のコストをユークリッド距離 $d(n, k_m)$ として、選択されるローカルコスト $c_l(n, k_m)$ と累積コスト $c_a(n, k_m)$ を次式で定義した。

$$c_l(n, k_m) = \begin{cases} d(n, k_m) & \text{if } ch(n) = 1 \\ & \wedge ch(k_m) = 1\ , \\ p_c \times d(n, k_m) & \text{otherwise} \end{cases} \tag{1}$$

$$c_a(n, k_m) = \min_{\tau,\mu} \begin{cases} c_l(n, k_m) & \text{if } (\mu = m \wedge \kappa = k - 1) \\ \quad +c_a(n-1, \kappa_\mu) & \vee st(n) \neq st(n-1) \\ p_t \times c_l(n, k_m) & \\ \quad +c_a(n-1, \kappa_\mu) & \text{otherwise} \end{cases} \tag{2}$$

ここで $ch(n)$ は小節 n がサビの場合に 1 を返し、$st(n)$ は音楽的構造の番号を返す関数である。また p_c が高いほど、データベース中の動画でサビに使われた映像が、入力音楽のサビで選択され易くなり、p_t が低いほど選出される映像が頻繁に切り替わる。

累積コストを最小化する映像系列は、最終小節 N において最も累積コストが小さい映像 $d_{\min}$ を次式で求めたのち、バックトレースによって得る。

$$d_{min} = \underset{k,m}{\operatorname{argmin}} \quad c_a(N, k_m). \tag{3}$$

3.2.2 項で述べた「ダメ出し」機能では、選択された構造区間の最終小節において、累積コストが異なる四つの候補からバックトレースして映像を生成して表示する。生成される映像の幅を広げるために、最も累積コストが小さい候補、全候補数の 1/3 番目と 2/3 番目に累積コストが小さい候補、最も累積コストが大きい候補、の四つを選んで映像を生成した。これによって、多様な印象を持つ映像が生成できる。

4.3 再生数に応じた回帰モデルの学習

本研究では、Web 上に公開された二次創作動画を再利用しているが、これは次の問題を含むと考えられる。動画制作者は音楽と映像の対応関係を各々で解釈して動画を作成するため、音楽と映像の対応付けの信頼度、すなわち「映像が音楽と合っていると感じられる度合い」には、ばらつきがあると考えられる。したがって、全ての動画を均等に用いて対応付けを行うと、適切に学習できない可能性がある。

この問題を解決するために、動画共有サイトにおける再生数が、楽曲と映像の対応付けの信頼度を間接的に反映していると仮定し、対応関係の学習で利用する。具体的には、再生数に応じた重み付けのモデル学習を行った。再生数 V_c の動画は、モデル学習時に以下の式によって求まる重み w を用いる。

$$w = \alpha \times \lfloor \log_{10}(V_c) + 0.5 \rfloor + \beta. \quad (4)$$

ここで $\lfloor \cdot \rfloor$ は切捨を表し、0.5 を足して四捨五入する。

なお本研究では、$\alpha = 2$、$\beta = -7$ とした。すなわち 1 万回再生された動画は $w = 1$、10 万回再生された動画は $w = 3$ となるよう重み付けを行った。具体的には、モデル学習における特徴ベクトルの数を、w の値に応じて増やすことで重み付き学習を実現した。

5 システムの運用結果

本章では DanceReProducer の実装に利用した動画コンテンツと、運用の結果について述べる。

5.1 収集した動画コンテンツについて

本研究では、既存のダンス動画から音楽に合ったダンス動画を切り貼りして生成し、また、音楽からの動画生成に関する多様な対応付けをモデル化するために、システムが扱う動画コンテンツは以下に示す四つの条件を満たす必要がある。

条件 1 内容がダンスを中心に構成されていること

条件 2 動画を切り貼りして生成された動画であり、その素材が統制されていること

条件 3 再生数を対応付けの学習等に用いて動画生成を行うことを考慮し、個々の動画が再生数を持つこと

条件 4 上記二つの条件を満たすコンテンツが大量に存在し、かつ、入手が容易であること

このような条件を全て満たすコンテンツとして、バンダイナムコゲームスから販売されているアイドル育成シミュレーションゲーム「THE IDOLM@STER」とそのライブシミュレーションゲーム「アイドルマスター Live for You！」[1] を素材として二次創作された Web 上の動画を対象とした。それに加え、3DCG によるダンスモーション制作ツールである MikuMikuDance (MMD)[2] によって制作された音楽動画も対象とした。これらを動画コミュニケーションサービス『ニコニコ動画』[3] から、再生数が 1 万回以上のものに限定して、それぞれ 100 件収集した。

[1] http://wwwbandainamcogames.co.jp/cs/list/idolmaster/
[2] http://www.geocities.jp/higuchuu4/
[3] http://www.nicovideo.jp/

5.2 システムの使用結果

本システムにより作成された動画は、リズムの同期や印象の近い映像が作成されていた。したがって、既存の動画群における音楽と映像の対応付けを適切にモデル化することができたといえる。

インタフェースを用いた実際の運用では、音楽と合っていない動画が自動生成される場合があっても、「ダメ出し」機能を使用するだけで映像が訂正できるため、好みの動画を簡単に作成できた。一方システムの改善点としては、「ダメ出し」による訂正の際に、映像の印象がほとんど類似した候補ばかりが提示されてしまって、適切に訂正できない場合もあった。

また、二次創作経験の無いユーザからの意見としては、映像候補数をさらに多くすることで、イメージが浮かびやすくなるという意見を得た（現状では四候補）。二次創作に精通したユーザからは、音楽の小節線や音楽的構造の推定結果を訂正することができれば、より良い動画を作れそうだという意見を得た。

6 おわりに

本論文では、既存のダンス動画を再利用して音楽に合った動画を作成できる DanceReProducer について述べた[4]。本システムは、音楽に合った動画が手軽に自動作成できるだけでなく、「ダメ出し」機能で制作者のオリジナリティを反映させながら、新たな音楽動画を制作することを可能とした。技術的な成果としては、多数の二次創作動画における音楽と映像の多様な対応関係を機械学習してモデル化し、アドホックでない映像付けを実現した点にある。ダンスに特化した特徴量の使用を意図的に避けたことで、学習対象の音楽動画を変更するだけで、任意の表現の音楽動画の創作を支援できる。さらに、その機械学習において、動画共有サイト上での再生数を活用するという世界初の試みをし、再生数の多い人気の高い音楽動画の特性をより重視した（より学習結果に反映されやすい）学習を実現した。

最後に、DanceReProducer が切り拓く未来のビジョンについて議論する。

6.1 DanceReProducer の発展

現在の実装では、上述のように敢えてダンスに特化した処理はしていなかったが、今後、ダンス動作に関する特徴抽出[5]による動作の考慮や、顔認識技術の導入によるダンサーの一貫性の考慮などにより、自動生成の精度をより上げることができる。

[4] http://staff.aist.go.jp/t.nakano/DanceReProducer/index-j.html でデモ動画を確認できる。
[5] EyesWeb: http://www.infomus.org/EywMain.html など。

我々は混合音（多重奏）中の歌詞（音素）認識技術 [18] を有しているが、映像の口形状の認識技術と組み合わせることで、例えば歌っている映像を切り貼りする際に、リップシンクも考慮してより精度を上げる拡張が考えられる。

また、拍や小節線、音楽構造の自動推定結果をユーザが手軽に訂正できるインタラクション技術の導入により、さらに高度な動画制作も今後可能になっていく。

6.2 社会とのインタラクション技術

DanceReProducer では、システムによる自動生成の枠組みを超えて、「ダメ出し」という**ユーザとのインタラクション**技術により、N 次創作支援におけるオリジナリティの反映を実現した。さらに、「動画共有サイトの再生数」を考慮する**社会とのインタラクション**技術により、音楽と映像の対応関係を学習する新しい手法を実現した。つまり、社会的にその音楽動画がどのように評価されているか、も反映可能とした。

本研究はこのように、音楽と映像の対応関係のような人間の心や社会的な文化に依存する問題を扱う場合に、人間とのインタラクションと社会とのインタラクションという二つの観点を適切に導入する方法を示した点でも貢献がある。これは今後、同様の性質を持つ他のシステム構築の際にも、重要となる考え方である。

6.3 豊かな社会の実現に向けて

DanceReProducer は、誰もが手軽にコンテンツを N 次創作して楽しむ未来社会を目指した研究である。コンテンツの制作は「自己表現」につながるだけでなく、自分で制作を体験してみることは、コンテンツ理解能力の向上につながり [19–21]、コンテンツをより楽しむことができるようになる。これらは QoL (Quality of Life) 向上の観点から重要であり、精神的に豊かな社会の実現に資する。

謝辞

長谷川 裕記 氏、平井 辰典 氏（早稲田大学）に感謝する。

参 考 文 献

[1] 濱崎雅弘，武田英明，西村拓一: 動画共有サイトにおける大規模な協調的創造活動の創発のネットワーク分析–ニコニコ動画における初音ミク動画コミュニティを対象として–，人工知能学会論文誌， Vol. 25, No. 1, pp. 157–167 (2010).

[2] 濱野 智史: インターネット関連産業，デジタルコンテンツ白書 2009，pp. 118–124 (2009).

[3] 室伏　空，中野倫靖，後藤真孝，森島繁生: ダンス動画コンテンツを再利用して音楽に合わせた動画を自動生成するシステム，情報処理学会 音楽情報科学研究会 研究報告，Vol. 2009-MUS-81, No. 7, pp. 1–5 (2009).

[4] 室伏　空，中野倫靖，後藤真孝，森島繁生: DanceReProducer：既存のダンス動画の再利用により音楽に合った動画を作成できるシステム，WISS 2009 講演論文集, pp. 63–68 (2009).

[5] Nakano, T., Murofushi, S., Goto, M. and Morishima, S.: DanceReProducer: An Automatic Mashup Music Video Generation System by Reusing Dance Video Clips on the Web, *Proc. of SMC 2011* (2011).

[6] 藤澤隆史，谷　光彬，長田典子，片寄晴弘: 和音性の定量的評価モデルに基づいた楽曲ムードの色彩表現インタフェース，情報処理学会論文誌， Vol. 50, No. 3, pp. 1133–1138 (2009).

[7] Laurier, C. and Herrera, P.: Mood Cloud : A realtime music mood visualization tool, *Proc. of CMMR 2008*, pp. 163–167 (2008).

[8] Goto, M.: An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds, *Journal of New Music Research*, Vol. 30, No. 2, pp. 159–171 (2001).

[9] 白鳥貴亮，中澤篤志，池内克史: 音楽特徴を考慮した舞踊動作の自動生成，電子情報通信学会論文誌 D， Vol. J90-D, No. 8, pp. 2242–2252 (2007).

[10] Cai, R., Zhang, L., Jing, F., Lai, W. and Ma, W.-Y.: Automated Music Video Generation using WEB Image Resource, *Proc. of ICASSP 2007*, pp. II–737–II740 (2007).

[11] Hua, X.-S., Lu, L. and Zhang, H.-J.: Automatically Converting Photographic Series into Video, *Proc. of the 12th annual ACM international conference on Multimedia*, pp. 708–715 (2004).

[12] Foote, J., Cooperand, M. and Girgensohn, A.: Creating music videos using automatic media analysis, *Proc. of the tenth ACM international conference on Multimedia*, pp. 553–560 (2002).

[13] Hua, X.-S., Lu, L. and Zhang, H.-J.: Automatic music video generation based on temporal pattern analysis, *Proc. of the 12th annual ACM international conference on Multimedia*, pp. 472–475 (2004).

[14] 西山正紘，北原鉄朗，駒谷和範，尾形哲也，奥乃　博: マルチメディアコンテンツにおける音楽と映像の調和度計算モデル，情報処理学会音楽情報科学 研究報告 2007-MUS-069，Vol. 2007, No. 15, pp. 111–118 (2007).

[15] Gillet, O., Essid, S. and Richard, G.: On the Correlation of Audio and Visual Segmentations of Music Videos, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 17, No. 2, pp. 347–355 (2007).

[16] Tzanetakis, G. and Cook, P.: Musical Genre Classification of Audio Signals, *IEEE Transactions on Speech and Audio Processing*, Vol. 17, No. 2, pp. 293–302 (2002).

[17] Goto, M.: A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music, *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1784–1794 (2006).

[18] 藤原弘将，後藤真孝，奥乃　博: 多重奏中の歌声の基本周波数と有声音素の同時推定手法，情報処理学会論文誌, Vol. 51, No. 10, pp. 1995–2006 (2010).

[19] 後藤真孝: 音楽音響信号理解に基づく能動的音楽鑑賞インタフェース，情報処理学会音楽情報科学研究会 研究報告，Vol. 2007, No. 37, pp. 59–66 (2007).

[20] Goto, M.: Augmented Music-Understanding Interfaces, *SMC 2009: Inspirational Session* (2009).

[21] Goto, M.: Music Listening in the Future: Augmented Music-Understanding Interfaces and Crowd Music Listening, *Proc. of the AES 42nd International Conference on Semantic Audio*, pp.21–30 (2011).