Vol.2009-MUS-81 No.21
2009/7/30

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

室伏　空[†1]　中野倫靖[†2]
後藤真孝[†2]　森島繁生[†1]

本研究では、既存のダンス動画コンテンツの複数の動画像を分割して連結（切り貼り）することで、音楽に合ったダンス動画を自動生成するシステムを提案する。従来、切り貼りに基づいた動画の自動生成に関する研究はあったが、音楽–映像間の多様な関係性を対応付ける研究はなかった。本システムでは、そうした多様な関係性をモデル化するために、Web 上で公開されている二次創作された大量のコンテンツを利用し、クラスタリングと複数の線形回帰モデルを用いることで音楽に合う映像の素片を選択する。その際、音楽–映像間の関係だけでなく、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。

An Automatic Music Video Creation System by Reusing Dance Video Content

Sora Murofushi,[†1] Tomoyasu Nakano,[†2]
Masataka Goto[†2] and Shigeo Morishima[†1]

This paper presents a system that automatically generates a dance video clip appropriate to music by segmenting and concatenating existing dance video clips. Although there were previous works on automatic music video creation, they did not support various associations between music and video. To model such various associations, our system uses a large amount of fan-fiction content on the web, and selects video segments appropriate to music by using linear regression models for multiple clusters. By introducing costs representing temporal continuity and music structure of the generated video clip as well as associations between music and video, this video creation problem is solved by minimizing the costs by Viterbi search.

1. はじめに

本研究では、音楽を耳で聴いて楽しむだけでなく、目で観て視覚的に楽しむこともできる技術を実現することで、音楽鑑賞の可能性を広げ、人々の音楽生活をより豊かにすることを目指す。音楽に合った映像を、その内容を考慮しながら自動生成できるシステムが実現できれば、音楽の新しい鑑賞手段をそのユーザに提供できるだけなく、音楽と映像の関係性を明らかにできる可能性がある点で学術的にも意義が深い。

音楽に合わせた映像の生成方法として、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ビートに同期した視覚効果を描画する機能がある。また従来、様々な色や形（視覚効果）を新規に描画する研究[1)] や、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[2),3)]、音楽に合わせてホームビデオを切り貼りしながら動画を生成する研究[4)–6)] があった。また、映像の生成以外では、音楽と映像の調和度をモデル化した研究[7)] や、音楽と映像の相関を定義して動画検索や動画のジャンル分類へ応用した研究[8)] があった。しかし、音楽と映像が対応付けられた大量の動画を用いて、その対応関係を学習し、新規音楽への映像付与を対象とした研究はなかった。

本研究では、音楽に合った動画を自動生成するシステム構築の第一段階として、楽曲を入力として与えると、既存のダンス動画コンテンツからその映像を分割・伸縮させながら連結（切り貼り）することで、音楽に合ったダンス動画を自動生成するシステムを提案する（**図 1**）。音楽に合ったダンス動画を自動生成できれば、誰でも手軽に音楽をダンス付きで鑑賞できるだけでなく、ダンスの印象などを利用して楽曲を分類するといった応用も考えられる。本システムは、動画共有サイト上に存在する大量のダンス動画コンテンツを再利用して動画を生成する。近年、音楽をリミックスするように、既存の動画コンテンツを断片的に再利用しながら、別の音楽へ映像を付与して楽しむユーザが増えている。そのような二次創作物を利用することで、同じ音楽に対して付与された異なる映像、同じ映像に対して付与された異なる音楽といった、複数の対応関係を学習できる。ここで、動画共有サイトにおける再生数は、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、対応関係の学習時に利用できる。また、動画の生成では、楽曲の音楽的構造を考慮しながら行う。

†1 早稲田大学
Waseda University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

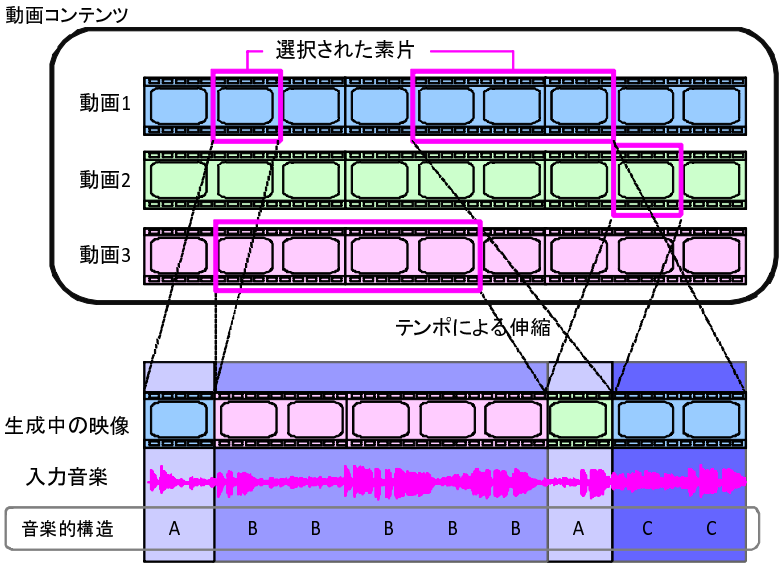

図 1 ダンス動画コンテンツから音楽に合った動画を自動生成するシステムのイメージ

2. ダンス動画コンテンツから音楽に合った動画を自動生成するシステム

本システムは、既存の動画コンテンツを入力としてそれをデータベース化するデータベース構築フェーズと、楽曲を入力としてそれに合った動画を生成する動画生成フェーズの二種類で構成される（**図 2**）。本章では、従来研究や人間の制作過程に関する調査に基づいたシステムの設計方針と、本システムを構成する二つのフェーズの処理概要を述べる。これ以降、説明の簡単化のために音楽付き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。また、既存の動画を再利用しながら、入力音楽に映像を付与して生成された動画は**再構成動画**と呼ぶ。

2.1 システム設計: 既存のダンス動画から音楽に合った動画を生成する方法

音楽に合った動画の自動生成システムを実現するために、動画の自動生成に関する従来研究[4)–6)] や、二次創作コンテンツ制作者が Web 上などで公開している制作過程を参考にした。その結果、動画の再構成に必要と考えられる条件を、内容に応じて分類しながらまとめた結果を以下に示す。

印象　楽曲の音楽的印象とダンスの動作・衣装に関する印象に相関がある

リズム　音楽のアクセントとダンス動作が同期している

音楽的構造　音楽的構造を反映した動画が生成される

- 音楽的構造の境界において何らかのアクションがある（ダンスの切り替わり、等）
- 音楽的構造を考慮したストーリーが構成される

時間的連続性　映像はある一定の時間長の間切り替わらない

- 印象やリズム、音楽的構造を反映しながら切り替わる

歌詞　歌詞とダンスが同期している

- ダンサーの口の動きが歌詞の音素と同期している（リップシンク）
- 歌詞の意味的内容をダンスが反映している

映像のエフェクト　映像エフェクトを新規に付与することで上記が実現される場合がある

- ライティングや画面切り替えによって映像のリズムを楽曲のリズムに合わせる

これらは、厳密な定義に基づいて得られた結果ではなく、また分類が相互に独立してはいないが、音楽–映像間の対応付けを考える上で考慮すべき条件であるといえる。

本稿では以上の条件を参考に、その中でも特に重要と考えられる「リズム」「印象」「音楽的構造」「時間的連続性」に関して、それを考慮した再構成動画の生成システムを設計する。まずリズムの同期は、視聴者が音楽との同期を最も感じる要素と考えられる。また、印象の同期を実現することは実用的なだけでなく、音楽とダンスの内容を考慮した対応付けを実現する上で重要である。さらに動画として成立するためにはある程度の時間長が必要であり、音楽を処理する上では音楽的構造の考慮が重要である。

そこで本システムでは、データベース構築（**2.2**）と動画生成（**2.3**）における時間の最小単位を小節にすることで対処する。ここで、小節単位で分割された動画を**素片**と呼ぶ。小節毎に特徴量を抽出することで、「リズム」に関する特徴量を抽出でき、また動画の時間的連続性をより保ちやすいと考えた。また動画生成において、データベース中の素片を入力楽曲のテンポに応じて伸縮させながら連結し、その際に素片の選ばれやすさを「印象」「音楽的構造」「時間的連続性」の三つに基づいて決定することで、それらの要素を考慮した再構成動画を生成する。

2.2 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 上に存在する動画コンテンツ群から、それぞれのコンテンツの素片毎に動画の特徴量（音楽特徴量と映像特徴量）を自動抽出し、また楽曲の音楽的構造を自動推定して、データベースを構築する（図 2）。さらに、コンテンツに対す

Vol.2009-MUS-81 No.21
2009/7/30

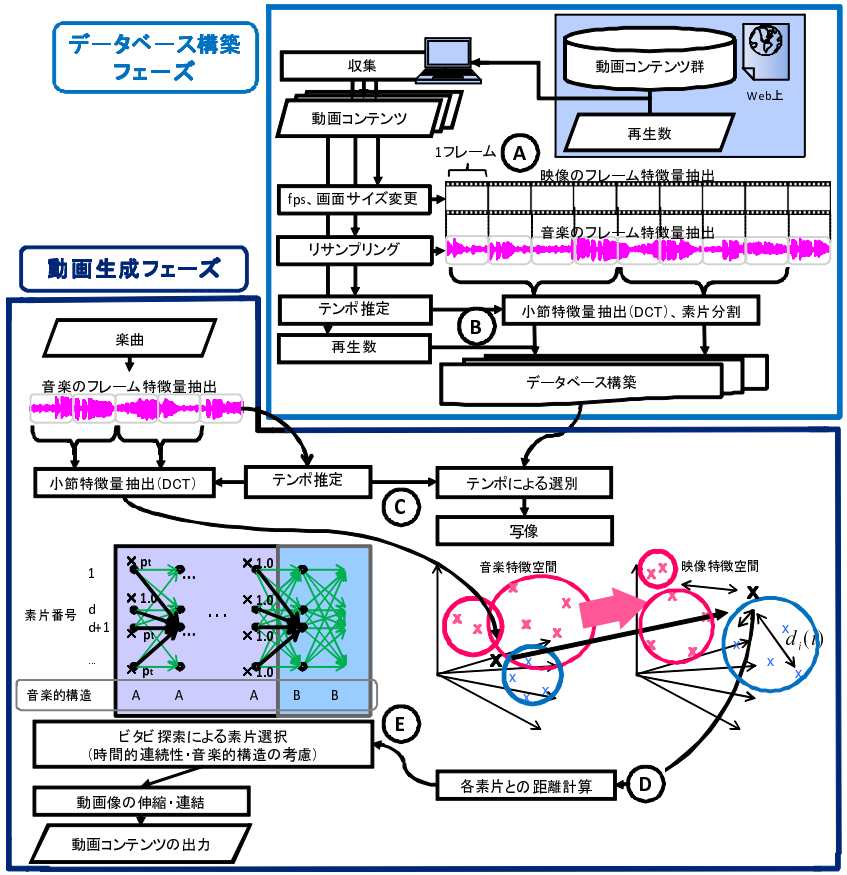

図 2 ダンス動画コンテンツから音楽に合った動画を自動生成するシステムの全体像

る再生数も同時に取得して追加・更新する⋆1。

処理の流れは、まず動画のフレームレート（以降 fps と略記⋆2）を 30 fps にリサンプリングしてフレーム毎の特徴量を抽出した後（図 2，Ⓐ）、それを小節毎の特徴量へ変換する（図 2，Ⓑ）。ここで、前者の特徴量をフレーム特徴量と呼び、これは fps を処理の時間単位として 1 秒間に 30 回抽出した音楽・映像特徴ベクトルである。また後者を小節特徴量と呼び、これは楽曲のテンポ及び小節線位置の推定結果に基づいて、フレーム特徴量を小節単位にまとめた特徴ベクトルである。

以上のようにして、動画コンテンツ群から、動画とそのテンポ・小節線位置・小節特徴量・音楽的構造、及び再生数を保持したデータベースが構築される。

2.3 動画生成フェーズ

動画生成フェーズでは、動画生成に先立ち、まず、入力楽曲のテンポに応じてデータベースを動的に再構築する。これは入力楽曲と大きくテンポが異なるダンス映像を使用してしまうと、素片の時間的な伸縮を行った際に動作として不自然に速い（遅い）ダンスが生成されてしまうが、それを防ぐために行う。またこの再構築においては、視聴者による再生数の利用など、必要に応じた様々な尺度を適用可能である⋆3。再構築されたデータベース中の全素片について、それぞれで音楽と映像の対応付けを求める（図 2，Ⓒ）。

続いて、入力楽曲から音楽に関する小節特徴量を求め再構成動画を生成する。これまでで得られた音楽–映像間の対応付けを用いることで、入力楽曲の全小節に対し、データベース中の全素片との近さを算出できる（図 2，Ⓓ）。そして、再構成動画の全体的な印象が楽曲に合うように、音楽的構造と時間的連続性を考慮しながら素片選択して動画を生成する。

以上のようにして得られた動画は、リズム、印象、音楽的構造、時間的連続性について再構成動画の生成に関する必要な条件を満たすと考えられる。

3. システムの実現方法

2 章で述べた本システムについて、本論文における具体的な実現方法について述べる。

3.1 動画コンテンツの取得

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三つの条件を満たす必要がある。

（条件 1）コンテンツの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

⋆1 時代の流れやその時々の流行の違いによって、視聴者やコミュニティ及びその評価が変わる可能性があるが、このような動的な構成を取ることで、その変化に柔軟に対処することができる。
⋆2 fps は frame per second の略であり、1 秒あたりのフレーム数（映像を構成する静止画の数）を表す。
⋆3 ユーザ毎にとって好みの動画を生成しやすいシステムとなるよう、楽曲の音楽ジャンルや映像の雰囲気等に基づいた自動分類法を利用できることを想定している。

Vol.2009-MUS-81 No.21
2009/7/30

（条件 3）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！」[9] を素材として二次創作された Web 上の動画を対象とした。ここで、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、再生数がカウントされている必要がある。そこで、動画共有サイト「ニコニコ動画」[10] から、再生数が 10,000 回以上の動画を 100 件収集した。

3.2 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の推定には様々な先行研究があり、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システムを実現した[⋆1]。音響信号のパワーの自己相関関数、及びそのパワーと推定されたテンポで生成されたパルス列との相互相関関数を計算し、それぞれピークピッキングによってテンポと小節線の位置を推定した。

サンプリング周波数 44.1kHz のモノラル音響信号（楽曲冒頭の 60 秒）を、その絶対値を取ってから、エイリアシング除去フィルタを通して、1 kHz の信号にダウンサンプリングすると、音響信号のパワーに相当する関数 $E(t)$ が求まる[⋆2]。

長さ T のパワー $E(t)$ の自己相関関数 $R_a(\tau)$ は次式で計算される。

$$R_a(\tau) = \frac{1}{T}\sum_{t=1}^{T}(E(t)\cdot E(t+\tau)). \quad (1)$$

$R_a(\tau)$ のピーク時刻は $E(t)$ の周期性を示しており、これを一拍の時間長としてテンポを推定する。ただし、倍テンポ誤り、半テンポ誤りを回避するため、推定されるテンポの範囲を 60 〜 120 bpm （$0.5 \leq \tau \leq 1.0$ 秒）として制限した。このようにして推定されたテンポを用いて、一拍毎にピークを持つパルス列 $P(t+\tau)$ と $E(t)$ との相互相関関数 $R_c(\tau)$ に基づいて、小節線の位置を推定する。

$P(t+\tau)$ と $E(t)$ との相互相関関数 $R_c(\tau)$ は次式で計算される。

⋆1 ダンス楽曲 16 曲を用いた予備実験では、サブバンド毎の出力や Spectral Flux によるテンポ推定と比べて性能が良かったため採用した。

⋆2 実際には、メモリの関係上、一度 16 kHz にダウンサンプリングしてその絶対値を取り、その信号をさらに 1 kHz にダウンサンプリングして実装している。

表 1 音楽と映像のフレーム特徴量

特徴量の意味	音楽特徴量	映像特徴量
アクセント特徴量	1 – 4 サブバンド毎のパワー	1 オプティカルフローの時間微分
	5 Spectral Flux	2 輝度値ヒストグラムの時間微分
印象特徴量	6 Zero-crossing rate	3 – 4 色相値の平均と標準偏差
	7 – 19 MFCC (0 次項 + 低次 11 次元)	5 – 6 彩度値の平均と標準偏差
		7 – 8 明度値の平均と標準偏差

$$R_c(\tau) = \frac{1}{T}\sum_{t=1}^{T}(E(t)\cdot P(t+\tau)). \quad (2)$$

$R_c(\tau)$ のピーク時刻は、楽曲中の一拍目の時刻を表わしている。本論文では非常に単純な手法として、仮にこの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。

3.3 特 徴 抽 出

特徴量抽出では、動画のフレームレートを 30 fps に合わせた後、音楽と映像のフレーム特徴量を抽出し、その後、それを小節単位にまとめる。本節では、2.1 節で述べた印象とリズムに着目して特徴抽出を行う。ここで “リズム” 特徴量とは、楽曲のパワーや画面の切り替わりといった “アクセント” に関する特徴量や、音色や雰囲気といった “印象” に関する特徴量の、時間的な変化を抽出することで表現する。したがって、フレーム特徴量としてアクセントと印象を抽出し、それらを小節単位にまとめてリズム特徴量とする。

本節では、それぞれの特徴量について、その意味と具体的な抽出方法について述べる。提案する特徴量は、$\boxed{n}$のように特徴量の番号を四角で囲んで示す。また、フレーム特徴量抽出における処理（フレームシフト）の時間単位は 30 fps（約 33 ms）である。

3.3.1 音楽に関するフレーム特徴量の抽出

音楽特徴量には、音楽と映像の対応付けに関する先行研究[7),8)] だけでなく、楽曲ジャンル分類に関する先行研究[11)] 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（**表 1** “音楽特徴量” 列）。特徴抽出は、サンプリング周波数が 44.1 kHz のモノラル音響信号を、音響信号の振幅が 0.9 となるように正規化してから行った[⋆3] 。分析窓のシフト

⋆3 収集した動画の中には、サンプリング周波数が 48 kHz のコンテンツが存在したが、それは 44.1 kHz にリサンプリングして扱った。

幅は 1470 点（30 fps）とし、窓幅はシフト幅に合わせて 1470 点とした。

アクセントの特徴量としては、主に楽曲のパワーとその短時間での変化を表現するために、フィルタバンク出力（1 – 4）と Spectral Flux（5）を用いた。ここで、フィルタバンク出力はフィルタバンク数を 4 として算出した。また Spectral Flux（S_t）は、楽曲の短時間フーリエ変換（STFT: Short-term Fourier Transform）によって得られる時刻 t、周波数 f の振幅スペクトルを $S(t,f)$ とすると、

$$A_s(t) = \int_0^N (S(t,f) - S(t-1,f))^2 df, \quad (3)$$

として計算される。ここで N はナイキスト周波数である。

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（6）と MFCC（Mel-Frequency Cepstral Coefficients）の直流成分と低次 11 項を用いた（7 – 20）。

以上のように、音楽のアクセントと印象に関する計 19 次元のフレーム特徴量を抽出した。

3.3.2 映像に関するフレーム特徴量の抽出

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[7),8)] 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（表 1 “映像特徴量” 列）。特徴抽出は、映像のフレームレートを 30 fps、画面サイズを 128×96 にリサンプリングして行った。

アクセントの特徴量としては、画面の動きやダンサーの動きとそれらの時間変化、また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に関する特徴量を用いた（それぞれ 1 と 2）。具体的には、それぞれを $A_o(t)$、$A_b(t)$ とすると、

$$A_o(t) = \frac{1}{P} \sum_{b_k=0}^{P} \sqrt{(O_x(t,b_k) - O_x(t-1,b_k))^2 + (O_y(t,b_k) - O_y(t-1,b_k))^2}, \quad (4)$$

$$A_b(t) = \frac{1}{Q} \sum_{b_n=0}^{Q} |B(t,b_n) - B(t-1,b_n)|, \quad (5)$$

として計算される。ここで、$B(t,b_n)$ は時刻 t における輝度値のヒストグラムであり、b_n がその bin 番号、Q が bin 総数（$Q = 128$）である。また、$O_x(t,b_k)$ は、時刻 t、ブロック番号 b_k における画面横軸方向のオプティカルフローの大きさ、$O_y(t,b_k)$ は画面縦軸方向のオプティカルフローの大きさ、P はブロック番号の総数である。オプティカルフローは、ブロック数 64×48、シフト幅 1、最大シフト幅を 4 としたブロックマッチング法によって求めた。

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、HSV (HSB) 色空間における色相（**H**ue）、彩度（**S**aturation）、明度（**V**alue もしくは **B**rightness）のそれぞれの値について、全画素の平均と標準偏差を用いた（3 – 8）。

以上のように、映像のアクセントと印象に関する計 8 次元のフレーム特徴量を抽出した。

3.3.3 小節特徴量の抽出

これまで述べてきた音楽と映像のフレーム特徴量を、それぞれについて、時間的な変化を良く反映するような小節特徴量としてまとめる。従来、楽曲のジャンル識別に関する研究では、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あった[11)] が、そのような方法では時間方向の特徴が失われ、リズムを表現するために不十分と考えられる。しかし、単純に各次元毎に全フレームの特徴量を用いたのでは、小節特徴量が非常に大きな次元数となってしまう。

そこで、次元数を抑えながら、かつ時間的な変化を反映した小節特徴量を、離散コサイン変換（DCT: Discrete Cosign Transform）を用いてフレーム特徴量から抽出する。まず素片毎のフレーム特徴量（ベクトル）の各次元毎に、16 点にリサンプリングして DCT を行う。その後、DCT 係数の低次 0 〜 3 項の計 4 次元を用いた。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。

以上のようにして、小節単位の素片について印象とリズム（アクセント）に関する特徴量を抽出した後、次元削減のために主成分分析を行った（累積寄与率 95 %）。主成分分析によって削減される次元数は、データベースの再構築を行うために一定ではないが、およそ音楽では 76 次元から 62 次元、映像では 32 次元から 26 次元へ削減された。

3.4 素片選択及びその連結に基づく音楽に合った動画生成

データベース中の全動画及び入力楽曲について小節特徴量を求めた後、入力楽曲に合った動画を生成する。ここでは入力楽曲の全小節について、それぞれに印象やリズムに関する特徴量が最も近い素片を選択する必要がある。本研究で用いる動画における音楽と映像の関係には、同一の音楽に異なる映像、同一の映像に異なる音楽といった多様な（矛盾した）対応付けがなされる可能性があるからである。特徴量の観点から考えるとそれはさらに顕著で、音楽特徴空間上では距離が近いにもかかわらず、映像特徴空間上では距離が遠いといった非線形な対応付けが起こると考えられる。また、このような問題は、データベースのサイズが大きくなればなるほど頻出する可能性がある。しかも、適切な対応付けが行えたとしてもそれだけでは不十分で、生成される動画は「時間的連続性」と「音楽的構造」が考慮されている必要がある（2.1 節 参照）。

3.4.1 複数の線形回帰モデルを用いた音楽–映像間対応付け

本論文では、音楽と映像を回帰モデルによって対応付けを行うが、その際には音楽と映像の多様な関係性に対処するために、音楽と映像の小節特徴量を結合したベクトルをクラスタリングし、それぞれのクラスタ毎に線形回帰を学習する。このような方法を取ることで、音楽と映像が共に類似している素片同士がまとめられるため、より矛盾の少ない音楽–映像間の対応付けが可能となる。

この複数の線形回帰モデルを学習するために、データベース中の全素片に対して、音楽と映像の小節特徴量を結合したベクトルを k-means 法によってクラスタリングする。入力楽曲の小節特徴量に近い素片選択では、まず特徴空間上の距離が最も近いクラスタの重心（平均ベクトル）を選択する。次に、そのクラスタで学習した回帰モデルを用いて、音楽の小節特徴量を映像の小節特徴空間へ写像する。最後に、写像によって得られた映像の小節特徴量と特徴空間上での距離を素片毎に計算する。ただし距離計算の際には、同じクラスタに属する素片だけでなく、データベース中の全ての素片を対象として距離を計算する。なぜなら、異なるクラスタに割り当てられてはいても、映像特徴空間上で近い素片は、潜在的には入力楽曲に合っている可能性があると考えたためである。このような方針を取ることで、クラスタリングによるデータベースの縮小というデメリットなしに、多様な関係性を活用して動画生成できる。

さらに回帰モデルの学習では動画の再生数などの利用も考える。これによって、再生数を重視した重み付けのモデル学習が行える。重み付け学習では、再生数 V_c から以下の式によって重み w を求める。

$$w = \alpha \times \lfloor \log_{10}(V_c) + 0.5 \rfloor + \beta. \tag{6}$$

ここで、$\lfloor \cdot \rfloor$ は切り捨てであり、0.5 を足して四捨五入する。$\alpha = 2$、$\beta = -7$ とすれば、10,000 回再生された動画は $w = 1$、10,0000 回再生された動画は $w = 3$ として重み付け学習が行える。

3.4.2 コスト最小化による素片選択

「時間的連続性」と「音楽的構造」を考慮して、かつ音楽に印象やリズムが合った再構成動画を生成するために、フレーム毎にコストを求めて、それが入力楽曲全体として最小となるように素片を選択する。すなわち、動画の生成をビタビ探索によるコスト最小化問題として解く。ビタビ探索アルゴリズムで、各ノードから次のノードへ遷移する場合に、時間的連続性と音楽的構造を考慮する。本論文では、音楽的構造とサビ区間を RefraiD[12] を用いて求めた。RefraiD では楽曲中の繰り返し区間の推定を行うが、得られた繰り返し区間の始端と終端を音楽的構造が切り替わる位置として用いた。また、推定された複数の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。

まず、小節数が N の入力楽曲において、n 番目の小節 $i_n (n = 1, 2, ..., N)$ でデータベース D 中の楽曲集合 M 中の t 番目の素片 $d_{(t,m)} (t = 1, 2, ..., T_m, m \in M)$ が選択されるローカルコスト $d(i_n, d_{(t,m)})$、及び累積コスト $c(i_n, d_{(t,m)})$ を次式によって定義した。

$$d\left(i_n, d_{(t,m)}\right) = \begin{cases} \sqrt{\sum_f \left(v(d_{(t,m)}, f) - v_n(i_n, f)\right)^2} & \text{if } p_{ch}(v(d_{(t,m)}, f)) = 1 \\ & \wedge ch(n) = 1 \\ p_c \times \sqrt{\sum_f \left(v(d_{(t,m)}, f) - v_n(i_n, f)\right)^2} & \text{otherwise} \end{cases}, \tag{7}$$

$$c\left(i_n, d_{(t,m)}\right) = \min_{\tau,\mu} \begin{cases} d\left(i_n, d_{(t,m)}\right) + c(i_{n-1}, d_{(\tau,\mu)}) & \text{if } \mu = m, \tau = (t-1) \\ & \vee s(n) \neq s(n-1) \\ p_t \times d\left(i_n, d_{(t,m)}\right) + c(i_{n-1}, d_{(\tau,\mu)}) & \text{otherwise} \end{cases} \tag{8}$$

ここで、$v(d_{(t,m)}, f)$ はデータベース中の映像小節特徴ベクトル、$v_n(i_n, f)$ は入力楽曲の n 番目の小節 i_n から推定された映像小節特徴ベクトルである。また p_c は、選択しようとした素片がサビで使われたことがない（$p_{ch}(v(d_{(t,m)}, f)) = 0$）、もしくは現在の小節 n がサビでない（$ch(n) = 0$）場合にかかるペナルティである。$ch(n)$ は小節 n がサビである場合に 1 を返す。また p_t は、選択しようとした素片が元の動画中で連続しておらず、楽曲構造も切り替わらない $s(n) \neq s(n-1)$ 場合にかかるペナルティである。すなわち、パラメータ p_c と p_t を操作することで、元の動画で連続していた素片は連続しやすく、音楽的構造が切り替わる箇所では素片が切り替わりやすく、サビに使われた素片をサビに使われやすくすることができる。

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

$$d_{\min} = \underset{t,m}{\operatorname{argmin}} \quad c\left(i_N, d_{(t,m)}\right) \tag{9}$$

3.4.3 システムの出力結果例

本システムの有効性を検証するために、時間的連続性と音楽的構造に関するパラメータを $p_t = 5.0, p_c = 1.0$ として、k-means 法によるクラスタリングでは $k = 10$ として実装した。また、既に人間が楽曲に合った映像を切り貼りして生成した動画から、その楽曲を抽出して、システムに入力として与えた出力結果の一部を**図 3** に示す。図 3 では、本システムによって生成された動画の時間的連続性、推定した音楽的構造、人間による同一楽曲の再構

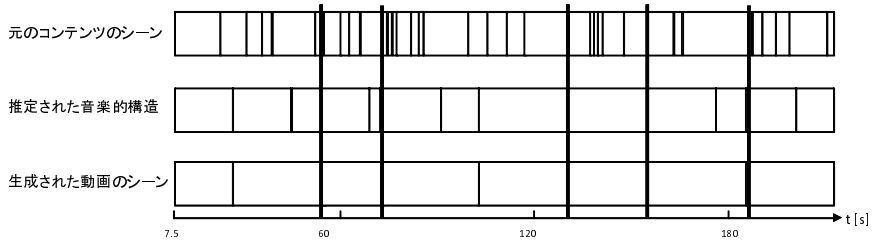

図 3 本システムによって生成された動画の時間的連続性（下）、RefraiD[12] によって推定された音楽的構造（中）、人間による同一楽曲の再構成動画の時間的連続性（上）を示す。図中、五箇所（太線のひかれた箇所）において、それぞれ三つの結果が揃っていることを示す。

成動画の時間的連続性が示されている。図中、五箇所（太線の箇所）は、それぞれ三つの結果において同期が取れていることを示す。

本結果から、本システムで生成される動画は、時間的連続性を持っているといえる。また音楽的構造を一部考慮できていた。ただし、人間による動画生成結果はさらに複雑であるため、このモデル化が今後の課題である。

4. おわりに

本論文では、既存の動画コンテンツを再利用し、音楽に合わせた動画を生成するシステムを提案した。まず、人間によって二次創作された大量のコンテンツを利用した。その際、複数の回帰モデルを用いることで、音楽と映像の多様な対応付けに対処した。また、コンテンツの再生数に着目し、コンテンツの選別や、音楽とダンス動画の対応関係を学習する際の重みとして利用した。最後に、動画の生成を、時間的連続性と音楽的構造を考慮しながら、ビタビ探索によるコスト最小化問題として解いた。

本システムの定量的な評価は今後の課題であるが、自動生成された動画は、楽曲の印象やリズムを反映していると感じられた。また、音楽的構造や時間的連続性を考慮した動画を生成することができていた。今後は、人間の動画生成過程をより良くモデル化するために、特徴量の検討や推定精度の向上、サビ以外の音楽的構造への対応にも取り組んでいきたい。また、この技術の応用として人間のダンスにより着目した研究を展開し、楽曲に対するダンスの振り付け例の提示なども考えたい。

参考文献

[1] 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).

[2] 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).

[3] 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. 90-D, No. 8, pp. 2242–2252 (2007).

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

[5] Hua, X.-S., Lu, L. and Zhang, H.-J.: AVE: automated home video editing, *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 490–497 (2003).

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

[7] 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，情報処理学会研究報告，2007-MUS-069, pp. 111–118 (2007).

[8] Gillet, O. and Richard, G.: Comparing Audio and Video Segmentations for Music Videos Indexing, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, pp. V–21–V–24 (2006).

[9] バンダイナムコゲームス：THE IDOLM@STER OFFICIAL WEB, http://www.bandainamcogames.co.jp/cs/list/idolmaster/.

[10] ニワンゴ：ニコニコ動画, http://www.nicovideo.jp/.

[11] Tzanetakis, G. and Cook, P.: Musical Genre Classification of Audio Signals, *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, pp. 293–302 (2002).

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