観察に基づく音楽およびモーションキャプチャデータからの舞踊動作生成手法

中澤篤志† 白鳥貴亮†† 池内克史††

† 大阪大学サイバーメディアセンター 560-0043 豊中市待兼山町 1-32
†† 東京大学生産技術研究所 153-8504 目黒区駒場 4-6-1
E-mail: †nakazawa@cmc.osaka-u.ac.jp, ††{siratori,ki}@cvl.iis.u-tokyo.ac.jp

あらまし 本論文は，音楽を入力すると，それに合ったキャラクターアニメーションを生成する手法を提案している．本手法はモーションキャプチャデータの観察に基づく動作解析，音楽解析，および解析に基づく動作生成手法によって構成される．動作解析部分では，モーションキャプチャーデータベースからフレーム間相関に基づいたグラフをつくる．同時に動きにおける特徴的なフレーム（キーフレーム）や動きのダイナミクス等の特徴量が抽出される．音楽からは，ビートおよび調の変化，音の盛り上がり等の特徴量が検知される．これら動き特徴量および音楽特徴量をそれぞれ，動き特徴ベクタ，音楽特徴ベクタとして表す．動作生成は，グラフ探索においてこれらの特徴ベクタの相関が高い枝を次々と選択することで行われる．生成された動きは，音楽のリズムおよび曲調に応じたキャラクターアニメーションとなって出力される．本手法はオンラインでも生成ができるよう拡張することも可能である．実験結果から，本手法によってあたかも「音楽を聴いている」ような動作の合成が可能であることが示された．
キーワード コンピュータアニメーション，モーションキャプチャ，舞踊，動作解析，音楽解析

Bust A Move Mr. Avatar!: Character Animation Dance to Music

Atsushi NAKAZAWA†, Takaaki SHIRATORI††, and Katsushi IKEUCHI††

† Cybermedia Center, Osaka University 1-32 Machikaneyama, Toyonaka, Osaka, Japan.
†† Institute of Industrial Science, University of Tokyo 153-8504 4-6-1 Meguro, Komaba, Tokyo, Japan.
E-mail: †nakazawa@cmc.osaka-u.ac.jp, ††{siratori,ki}@cvl.iis.u-tokyo.ac.jp

Abstract This paper will present our method for the synthesization of dancing avatar animation to music. Our method consists of a motion analysis, a music analysis, and a motion synthesis based on results of the analyses. In the analyses step, a motion and music feature vector for each frame is acquired. These vectors are derived from motion keyframes, motion intensity, musical beats, chord changes, and music intensity. The motion is then synthesized by motion graph tracing. For this step, the most correlated trajectory between music and motion feature vectors is selected and the resulting motion is generated. We enhanced our method not only for off-line synthesis but on-line synthesis too by considering the dependencies of each step. Our experimental results indicate that our proposed method actually creates dance as the system 'hears' the music.
Key words Computer Graphics, Three-Dimensional Graphics and Realism, Animation, Arts and Humanities, Performing Arts, Music

1. はじめに

リアリスティックな人体アニメーションを自動生成する試みは近年のコンピューターアニメーションのトピックであり，様々な取り組みがなされている．これらは，ダイナミクスシミュレーションによって動作を生成するものと，様々なモーションキャプチャデータを連結し動作を生成するものに分類でき，特に後者はモーションキャプチャアシスティッドアニメーション (motion capture assisted animation) として知られている．

これらの動作生成技術を実際のコンピューターアニメーション製作に応用するためには，デザイナによる意図による生成が可能でなければならない．過去の技術においては，Gait Motion において歩行軌跡をデザインする手法などがある [16].

一方で「音楽にあわせて体を動かす」能力は，誰にでも備わっ

た能力である．音楽を聴いていると，人は自然に音楽に合わせて体を動かすのは誰もが経験していることである．舞踊者たちは，流れている音楽に合わせて踊りをリアルタイムに構成することができる．この能力はどのように構築されているかを考えると，舞踊者は動きをまったくゼロの状態から作りだしているのではなく，既存の動きを音楽に合わせて適宜組み合わせ，新しい動作を生成していると考えられる．

我々は，人や舞踊者に備わったこのような能力を理解し，無形文化財のデジタルアーカイブに適用する研究を行っている [1]．その一環として，コンピュータアニメーションやロボット動作に適用可能な動作生成手法の開発が挙げられる．本論文では，音楽を入力とし，あらかじめ取得されたモーションキャプチャデータベースから，その楽曲に合った適切な動作を検索・合成し，新規動作列を生成する手法を提案する．これは人に備わる「音楽に合わせて既存の動きを組み合わせる」能力の実現であると言える．本技術により，音楽からの自動的なキャラクタアニメーション生成ができるだけでなく，ロボットの人工知能としての「音を聞きながらそれにあわせて踊る」能力を実現する手段として応用できる．本手法は，音楽情報とモーションキャプチャデータの解析による観察から，人間の動きと音楽の関係を明らかにし，それに基づき新たな動作を生成するイメージメディアとして，広用途に応用が可能な技術であるといえる．

またロボット等への応用を考えると，音楽解析および動作合成プロセスをオフラインで実行するだけでなく，オンラインでも行えることが必要である．この要求に対し我々は，本手法に２つのモードを持たせ，オフライン推定モードでは音楽の盛り上がりも含めた完全な解析に基づいた生成を行い，オンライン推定モードでは音楽のビートと調の変化のみを用いて動作合成を行うこととした．

本論文の構成は以下のようになる．まず次章では，本目的に必要な音楽解析技術および動作合成技術に関する関連研究を述べ，その後３章にて提案アルゴリズムの概要を述べる．４章，５章でそれぞれ動きの解析，生成技術と音楽解析技術について述べ，６章では最終的な動作合成手法について説明する．７章では本システムのオンライン化への処理について説明し，８章，９章で実験結果とまとめを行う．

2. 関 連 研 究

2.1 キャラクターアニメーションの合成手法

コンピュータグラフィックスの分野では，モーションキャプチャに基づいた動作合成手法が数多くなされている．フィルターバンクを用いた信号処理的なアプローチによるもの [7]，動きの補間（モーションワーピング）によるもの [34]，単なる関節角度の再マッピングではなく，体のサイズ等の違いを考慮した動作変形手法 [12] [18] などが研究されている．また，より自然な動作を合成するための，物理シミュレーションによる動作生成手法も提案されている [26] [21]．

近年ではモーションキャプチャの再使用を目的とした手法も数多くなされている．データベースに格納された複数のモーションキャプチャクリップより相関あるフレームを見つけ，その間を補間することで生成されるモーショングラフ (Motion Graph) アルゴリズムは，その代表的なものである．ユーザーは任意の分岐点，あるいは与えられた評価値によってグラフ分岐を指示し，システムはそのグラフを探索することで新しい動作が生成される [16] [25] [2] [19] [20] [3]．また近年では，動作データベースからの比較や学習により，共通的な動作やスタイル部分を抽出する研究もなされている [4] [5]．これにより，動作シミュレーションにおける制約も容易に解けるという利点がある [35] [14]．

音から動作を合成した研究も，若干ながらなされている．Stone らの手法 [32] では，人のスピーチにおける身振り動作の合成を目標としている．ここでは，音声情報を用いることが前提となっており，舞踊等の音楽に応じた動作合成を目的とはしていない．Kim らは，動作の周波数と音楽の周波数の解析から，踊り等を対象とした動作合成手法を提案している [15]．しかし本手法は，入力音楽のテンポ（周波数）が一定であることを仮定しており，音楽情報は MIDI データによって入力される．しかし，音楽においては一般にテンポが変化する場合が普通であり，この手法は限定された状況にしか応用することができない．本手法では，これらの問題点を解決する方法を提案している．

2.2 音楽からの特徴解析手法

音楽情景描写の解析は，人間がどのように音楽の要素を認識しているかを解明する上で重要であり，音楽情報処理の分野では Computational Auditory Scene Analysis と呼ばれている [6] [9]．特に舞踊の解析や生成においては，我々は音楽の拍節，および盛り上がり度（ダイナミクス）が非常に重要であると考えており，これらの音楽的特徴の認識に関する関連研究を述べる．

MIDI 信号からビート成分を抽出する研究では，発音成分，ピッチ (音高)，各音素のボリューム等の音楽的要素が容易に得られ，これらを用いた手法が提案されている [10] [27] [28] [11] [17]．一方音響信号を対象とする場合，これらの要素を正確に抽出することは困難であり，発音成分やそれに関連したパワーの変化等を用いた手法が多く提案されてきた [33] [23]．後藤はポピュラー音楽を対象として，発音成分だけでなく，コード (調) の変化や打楽器の発音時刻などの音楽的要素を考慮することによって，ビート間隔やビート構造等の把握が可能な手法を提案している [13]．Scheirer は *accel.*(リズムの加速) や *rit.*(リズムの減速) などが含まれるようなリズムが一定でない曲に対しても適用可能なビートトラッキング手法を提案している [29]．その他にもカルマンフィルターを用いた手法 [8] やベイジアンネットワークを用いて次のリズム時刻を推定する手法 [30] も提案されている．また，音楽のダイナミクス認識手法としては，ケプストラム領域の特徴量の一つである MFCC(Mel-Frequency Cepstrum Coefficients) を用いる手法 [22] や，振幅の大きさの変動を用いる手法 [23]，変形離散 cos 変換を用いる手法 [31] などが提案されている．

3. アルゴリズムの概要

提案アルゴリズムの概要を図 1 に示す．システムは動作解析

フェーズと音楽解析フェーズ，動作合成フェーズから構成される．動作の生成手法は，動作の相関を評価しそれを連結する一連の研究 ([16] [25] [19]) を改良し使用する．そのため，動作解析フェーズにおいてあらかじめモーションキャプチャーデータの全フレーム間の相関を評価し，連結可能なフレームセットをあらかじめ検出しておく．これを連結し，モーショングラフを生成する．同時に各動作の「動き特徴ベクトル」を検出しておく．音楽解析フェーズでは，入力された音楽を周波数解析しビートフレームおよび小節の開始ビートを検出，音楽特徴ベクトルを求めておく．動作生成フェーズでは，動作解析フェーズで構築されたモーショングラフを探索し新たな動作を生成するが，ここでは動き特徴ベクトルと音楽特徴ベクトルの相関性を評価することで最適なパスを計算することになる．

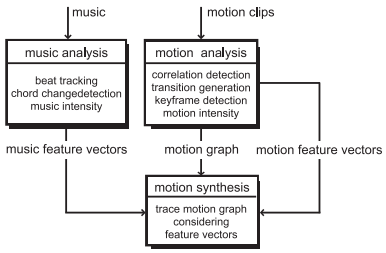

図 1　提案アルゴリズムの流れ．本手法は動作解析，音楽解析および動作生成部分より構成される．

4. モーションキャプチャデータからの動作解析と連結手法

本章ではモーションキャプチャデータからの舞踊動作解析・合成手法について述べる．我々の手法は，動作の類似度評価に基づいているが，従来手法と異なり以下のような特徴をもっている．

- 体全体の姿勢と動きを用いた新たな類似度距離
- 類似度距離に基づく連結フレーム数 (時間) の制御

またデータの各フレームにおける動き特徴ベクトルを抽出する手法についても解説する．

4.1 動作の類似度評価と連結可能性

本手法で用いる人体モデルを図 2 のように表す．フレーム f における人体姿勢は，体全体の位置姿勢を表す行列とベクトル $\mathbf{R}, \mathbf{t}$，および体の形を表現する 17 のベクトル $\mathbf{v_n}$，およびリンクの長さを表す 9 つのスカラー値 l_n によって表現できる．

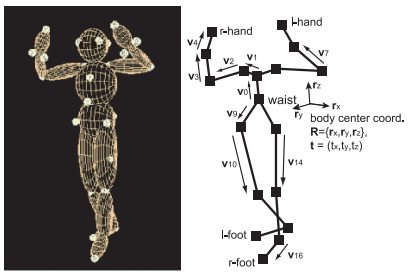

図 2　使用する人体モデル．体中心座標系 $\{\mathbf{R}, \mathbf{t}\}$ およびリンクを表す 17 のベクトル $\mathbf{v_n}$ によって表現される．

$$S(f) = \{\mathbf{R}, \mathbf{t}, \mathbf{v_0}, \mathbf{v_1}, ..., \mathbf{v_{16}}, l_0, l_1, ..., l_8\}.$$

$$\mathbf{R} = \{\mathbf{r_x}, \mathbf{r_y}, \mathbf{r_z}\}.$$

$$\mathbf{t} = \{t_x, t_y, t_z\}.$$

正規化ベクトル $\mathbf{v_n}$ は体中心座標系 $\{\mathbf{R}, \mathbf{t}\}$ における体リンクの方向を表している．ここで，体中心座標系 $\{\mathbf{R}, \mathbf{t}\}$ としては，データの各フレームにおける腰を基準にした位置および正面方向を採用している．Z 軸が腰から腹部方向を表しており，Y 軸が正面方向を表す．この座標系はモーションキャプチャデータから容易に求めることができ，逆にこの表現からモーションキャプチャのマーカー点に戻すことも可能である．

この記述を用いると，データベース中の動作のフレーム $S^A(f_A), S^B(f_B)$ 間の相関は以下で計算する．

$$Dist(S^A(f_A), S^B(f_B)) = \sum_i \{\alpha_i \cdot \mathbf{v_i^A}(f_A) \cdot \mathbf{v_i^B}(f_B) + \beta_i \cdot \dot{\mathbf{v}}_\mathbf{i}^\mathbf{A}(f_A) \cdot \dot{\mathbf{v}}_\mathbf{i}^\mathbf{B}(f_B)\}$$

第 1 項は姿勢のスタティックな相関（ポーズの相関）を表し，第 2 項は体中心座標系でのエンドエフェクタの動き相関を表している．この評価式を用いることで，人の動作の相関をより正しく求めることができることが分かっている [36]．図 3 に，これで求められた 2 動作の相関マトリックスを示す．

この相関マトリックスから，連結可能性のあるフレームを抜き出すことが次の課題である．固定しきい値法によって相関の高い blob を抽出し，その極大点を接続点とする手法がもっとも単純である [16]．しかしこの著者らも指摘しているように，しきい値の設定は対象の動作群の類似度に依存する．すなわち，2 種類の Gait Motion のように互いが類似した動作の場合その値は高くするべきであり，逆に Gait Motion と Dance Motion の比較のように類似したフレームがほとんど期待できない場合は，しきい値を低くしなければ接続フレームを取り出すことはできない．そこで我々は，比較する動作の種類に関係なく一定数の接続フレームを確保するため，可変しきい値による抽出手法を用いる．

相関マトリックスを $\mathbf{M} = M(j,k)$ とし，n_0 はユーザーが指定した連結フレーム数であるとする．ここで類似度に対するしきい値の範囲 $thresh_{max}$ および $thresh_{min}$ を設定する．マトリックスに対する 2 値化およびラベリング処理をこのしきい値範囲で順次行い，ラベル数が n_0 にもっとも近くなる値を採用する．連結フレームは，このラベリングされた領域の極大点として求められる．

これにより，すべての動作の組み合わせにおいて一定数の接続フレームを得ることができるが，接続フレーム間の類似度は比較する動作の組み合わせによって異なる．一方で，これらのフレームを補完する時間（フレーム数）を一定とすると，補間中の体部位の速度が前後の動作にくらべ大幅に速くなったり遅くなったりし，不自然な動作が生成される原因となる．そこで我々は，接続フレーム間の各リンクの角度差を保存しておき，適応的な接続時間の設定に用いる．

4.2 連結動作の生成

接続フレーム間のトランジションを生成する．体姿勢の合成

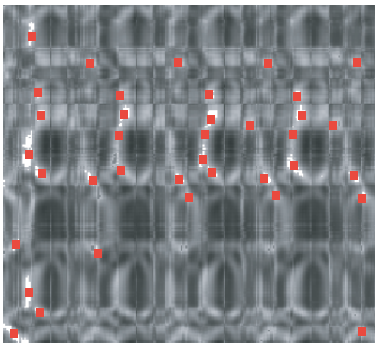

図 3　日本舞踊（ジョンガラ節と会津磐梯山）における相関マトリックス．赤い点は連結可能フレームを示している．

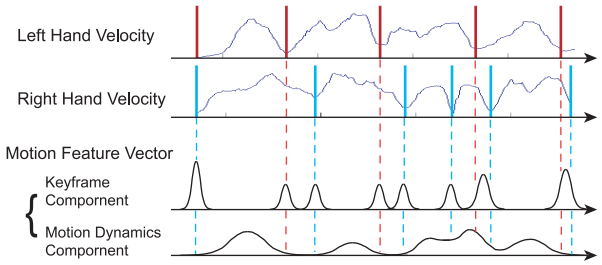

図 4　両手のエンドエフェクタの速度グラフと動き特徴ベクトルの関係．動き特徴ベクトルはキーフレーム成分およびダイナミクス成分から構成される．各要素の時間的広がりはガウス分布で規定される．

は基本的に，接続されるフレーム間の各体リンクを三次補完したものとなる．すなわち，連結するフレームを S^A および S^B とし，連結時間を T (frames) としたとき，連結動作ベクトル $\{\mathbf{v_i}(f)|0 \leqq f \leqq T, 0 \leqq i \leqq 16\}$ は以下で求められる．

$$
\begin{aligned}
\mathbf{v_i}(f) &= f^3 \cdot \mathbf{a_i} + f^2 \cdot \mathbf{b_i} + f \cdot \mathbf{c_i} + \mathbf{d_i} \\
\mathbf{a_i} &= \{T \cdot (\dot{\mathbf{v}}_\mathbf{i}^\mathbf{B} - \dot{\mathbf{v}}_\mathbf{i}^\mathbf{A} - 2 \cdot (\mathbf{v_i^B} - \mathbf{v_i^A} - \dot{v_i^A}T)\}/T^3 \\
\mathbf{b_i} &= \{-T \cdot (\dot{\mathbf{v}}_\mathbf{i}^\mathbf{B} - \dot{\mathbf{v}}_\mathbf{i}^\mathbf{A}) + 3 \cdot (\mathbf{v_i^B} - \mathbf{v_i^A} - \dot{\mathbf{v}}_\mathbf{i}^\mathbf{A}T)\}/T^2 \\
\mathbf{c_i} &= \dot{\mathbf{v}}_\mathbf{i}^\mathbf{A} \\
\mathbf{d_i} &= \mathbf{v_i^A}
\end{aligned}
$$

また体全体の動きに対しても，基本的には $\{\mathbf{t^A}, \dot{\mathbf{t}}^\mathbf{A}, \mathbf{R^A}, \dot{\mathbf{R}}^\mathbf{A}\}$ and $\{\mathbf{t^B}, \dot{\mathbf{t}}^\mathbf{B}, \mathbf{R^B}, \dot{\mathbf{R}}^\mathbf{B}\}$ に対する三次補完関数を用いる．しかし，補間前後での対大地への姿勢および足の接触状態を保存するため，体の鉛直方向 t_z の角度 $\theta_z = arccos(\mathbf{r_z} \cdot [0\ 1\ 0]^T)$ が保存されるように調整を行う．このため，まず前出の方法において体の正面方向を決定し，次に θ_z が保存されるように補間フレーム全体に対して調整を加える．

一方，ここで問題となるのが補間にかかる時間である．前後の動作に比べ不自然に早くならないよう，先に述べた関節角度差とトランジション前の動作の最大速度から，この値を求める．前出した動作解析ステップにおいて，連結前後のクリップにおける各関節における最大回転速が求められている．三次補完による連結動作フレームにおいて，この最大速を超えない最小の連結時間を選択する．この処理によって，連結時間中の不自然な手足速度の変化を防ぐことができる．

4.3　動作特徴量の抽出

データベース中のすべてのクリップに対して，あらかじめフレーム毎の動作特徴ベクトルを計算し保存しておく．動作特徴ベクトルは，キーフレーム成分 $x_{keyframe}$ と動きダイナミクス成分 $x_{intensity}$ から構成される．ここで，キーフレーム成分は我々が従来提案した手法により抽出され [36]，動きのダイナミクスは手先の最大速度から求められる．

まず，エンドエフェクタの体中心座標系における速度を求め，ここから速度が極小となるフレームをキーフレームとして登録する．また，キーフレーム間での最大速度はダイナミクス成分として用いられる．すなわち，フレーム f における動作特徴ベクトルは以下で表される．

$$
\begin{aligned}
\mathbf{MotionFeature}(f) &= \begin{bmatrix} x_{keyframe}(f) \\ x_{intensity}(f) \end{bmatrix} \\
x_{keyframe}(f) &= \sum_{i=1}^{N^L} exp\{\frac{-(f - f_i^L)^2}{\sigma(K)^2}\} \\
&\quad + \sum_{j=1}^{N^R} exp\{\frac{-(f - f_j^R)^2}{\sigma(K)^2}\} \\
x_{intensity}(f) &= \sum_{i=1}^{N^L-1} v_{max}^L(i) \cdot exp\{\frac{-(f - \frac{f_i^L + f_{i+1}^L}{2})^2}{\sigma(f_{i+1}^L - f_i^L)^2}\} \\
&\quad + \sum_{j=1}^{N^R-1} v_{max}^R(j) \cdot exp\{\frac{-(f - \frac{f_j^R + f_{j+1}^R}{2})^2}{\sigma(f_{j+1}^R - f_j^R)^2}\} \\
\sigma(x) &= -2.0 \cdot x/ln(0.01) \\
v_{max}^{\{R,L\}}(n) &= maximum\ velocity\ between\ f_n^{\{R,L\}} \\
&\quad and\ f_{n+1}^{\{R,L\}}
\end{aligned}
$$

ここで，N^L, N^R および f_i^L,f_i^R は，左手，右手におけるキーフレーム数である．図 4 に速度グラフと，動き特徴ベクトル成分の例を示す．キーフレーム成分は，キーフレームに対して時間的に近いほど大きく，また両手のキーフレームが同時に生じるほど大きな値となる．また隣接するキーフレーム間における手先の速度が大きいほどダイナミクスの値は大きくなる．ここで，各成分に対して時間的幅を与えているのは，人間は必ずしも音楽ビートに完全にキーフレームを一致させているのではなく，ある程度のズレが生じる可能性があるからである．逆にこの幅を規定する値 σ が，キャラクターの個性を表現しているといえる．

5.　音楽的要素の抽出

舞踊者が音楽に合わせて動作を生成する場合，音楽的特徴としては以下の要素が重要であると考えられる．

ビート　人は音楽のリズム（ビート）合わせ体を動かすことができる．舞踊動作においても同様のアルゴリズムが存在すると考えられる．ここでは，三拍子や四拍子と言ったビートの構造も重要である．

音楽ダイナミクス　ゆったりとした音楽や激しい音楽など，音楽の印象の違いは舞踊動作にも影響を与える．この音楽におけ

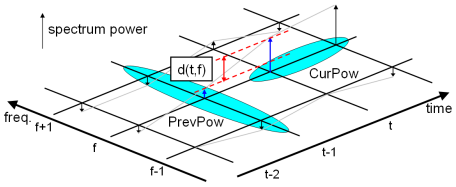

図 5　オンセット成分抽出手法．まず $p(t-1,f)$，$p(t-1,f\pm1)$ 間の最大値が PrevPow として表現され，次に $p(t,f)$，$p(t,f+1)$ 間の最小値が CurPow で抽出される．CurPow と PrevPow の差が計算されオンセット成分が求められる．

る「盛り上がり」を音楽のダイナミクスとして求め，動作生成に利用する．

ビート構造解析としては，ビート間隔とコード変化度を用いる．ダイナミクスの計算はこれらの指標に加え周波数的な変化度も影響するものと考えられる．これら 3 つの音楽的要素から以下の music feature vector を算出する．

$$\mathbf{MusicFeature}(f) = \begin{bmatrix} x_{beat}(f) \\ x_{music\ intensity}(f) \end{bmatrix}$$

$$x_{beat}(f) = beat\_existence \cdot degree\_of\_chord\_change$$

以下ではこれら 3 つの音楽的要素の導出について述べる．

5.1　ビート時刻の推定

ビート時刻を推定する上では，以下の仮定を用いる．

仮定 1　音の発音時刻とビート時刻は一致する可能性が高い

仮定 2　発音成分のピークの間隔はビート間隔とほぼ一致する

まず仮定 1 を用い，時刻 t，周波数 f におけるスペクトルパワーの増分 $d(t,f)$ を下式を元に算出する (figure 5) [13].

$$d(t,f) = \begin{cases} \max(p(t,f),p(t+1,f)) - \mathrm{PrevPow} \\ \quad (\min(p(t,f),p(t+1,f)) \geqq \mathrm{PrevPow}), \\ 0 \quad (\mathrm{otherwise}) \end{cases} \quad (1)$$

$$\mathrm{PrevPow} = \max(p(t-1,f),p(t-1,f\pm1)), \quad (2)$$

ここで $p(t,f)$ は時刻 t，周波数 f におけるスペクトルパワーを表す．次に全周波数にわたる合計 $D(t)=\sum_f d(t,f)$ を求める．ここから，時刻 t においてどのくらいの強さの音が発音されたか（発音成分）を算出できる．

次に仮定 2 に基き，発音成分 $D(t)$ の自己相関関数を求めることによって，平均ビート間隔を算出する．ここでは，パルス間隔が平均ビート間隔となっているパルス列と発音成分 $D(t)$ との相互相関関数を求めることによって，ビートの開始時刻およびビート間隔が求められる．しかし実際の楽曲ではビート間隔が常に一定ではなく，わずかながら変化しており，この誤差が累積するとリズムの追従が不可能になる．そこで，再び仮定 1 を用い，求まったリズム時刻の周辺で発音成分 $D(t)$ のローカルピークを抽出することによって，より正確なリズム時刻が推定できる．

5.2　コード変化度

大まかなビート構造を把握するために以下の仮定を用いる．

仮定 3　コード (調) の変化が起こる時刻に小節線 (拍節の区切れ) が現れる可能性が高い

コード変化を認識することによってよりハイレベルなビート構造を把握することが可能となり，舞踊動作の生成の際にコード変化を考慮することによって拍節等にあった舞踊動作を生成することが可能となる．

ある音が発せられると，その音の整数倍の周波数を持つ音（倍音）も同時に発生する．コードは，その楽曲の中で最も低い音とその音の倍音によって構成される．コードが変わる時刻では低音とその倍音の周波数成分が著しく変わり，またコードが変わらない場合は低音部分の周波数成分をほとんど変化しないことが知られている．本手法では，周波数成分のヒストグラムを求め，そのヒストグラムの変化の度合いをコード変化度として算出した．

具体的には，最初に 5.1 で推定されたビート時刻を元に，一拍間での周波数成分のヒストグラムを算出し，そのヒストグラムのピークを算出する．次に 0〜1000Hz におけるヒストグラムのピークの合計値を求め，連続するヒストグラム間で合計値の比を求めることによってコード変化度を算出した．この手法では，コードそのものの抽出ではなく，コードの変化の割合を求めるアプローチを取っているため，比較的容易にパラメータの算出，大まかなリズム構造の把握が可能となった．

5.3　音楽のダイナミクス

音楽のダイナミクスを算出する上では，以下に示す音楽的知識を用いた．

仮定 4　音楽が盛り上がるにつれて，楽曲のメロディラインのスペクトルパワーが増加する

このダイナミクスの算出には，5.2 で求めた一拍間での周波数成分ヒストグラムのピークを用いて行なう．一般的に可聴領域のメロディラインは 400-600Hz であるため，その倍音を考慮して 400-1500Hz における周波数成分ヒストグラムのピークの合計を求め，連続する拍間で増減比を求めることによってダイナミクスを算出する．

6.　動作の生成

動き特徴ベクトルおよび音楽特徴ベクトルを用いて，音楽に合った動作生成を行う．ここで我々は，音楽と体の動きの間には以下のような関係があると考え，アルゴリズムの設計を行っている．

- 動きのキーフレームは音楽のビートフレームに生じる．
- 音楽のダイナミクスと動きのダイナミクスは一致する．

動きのキーフレームの仮定は人の動きに関する仮定であるが，多くのコンピュータアニメーションの研究が，人の動き速度の極小値をキーフレームとしていることからも自然であり，生物学的な知見からも支持されると思われる．また，動きのキーフレームと音楽の関係，および，動きのダイナミクスと音楽のダイナミクスの関係は，我々が日常的に体験していることであり，また音楽に合わせた舞踊者の演技を観察することでも理解できる．音楽のビートがキーフレームと完全に一致することはありえないが，キーフレームが生じるフレームが必ずビートフレー

ムであるという仮定は正しいと考えられる．

図 6 に，動作生成アルゴリズムの流れを示す．動作生成は，動作解析部分で得られたモーションキャプチャデータの連結関係に基づき，モーションクリップおよび連結動作を順次たどることによって生成できる．どの連結関係を選択し，次の動作に遷移するかの判定は，遷移先の動き特徴ベクトルとそれに対応する音楽特徴ベクトル間を，以下の計算によって相関評価することで求められる．

$$\mathbf{MatchingEval} = \begin{bmatrix} Beat\ Eval. \\ Intensity\ Eval. \end{bmatrix}$$

$$= \begin{bmatrix} \sum_i x_{beat}(f+i) \cdot x_{keyframe}(f'+i) \\ \sum_i x_{music\ intensity}(f+i) \cdot x_{motion\ intensity}(f'+i) \end{bmatrix}$$

ここで，パラメータ f は将来的に遷移が起こる音楽フレーム番号を示しており，f' は，遷移が起こった先の動作クリップのフレーム番号を示している．図 6 を例に示す．ここでは motion 1 のフレーム 0 から，動作合成が開始されることを想定している．まず，motion 1 の最初の遷移（クリップ A からクリップ C）が上式によって評価される．このとき用いられるパラメータは $f = T_A + T_I$, $f' = T_B$ である．ここで，評価に用いるフレーム数 i も指定する必要がある．理論的にはこのパラメータ i は遷移先のモーションクリップの長さと同じにすべきだが，遷移先で再遷移が起こる可能性や，計算コストを考慮すると，これを有限の値として計算してもよいと考えられる．我々は i として 3 秒 (120Hz の場合 360 フレーム) の値を用いている．

システムは遷移を起こす前に遷移先の評価を行い，評価が高い枝を選択することで動作生成が行われる．ここでの評価手法としては，まずキーフレームと音楽ビートの相関を評価し，その後動きのダイナミクスの相関を評価する 2 ステップをとる．これは，キーフレームと音楽ビートの対応が，動きと音楽間のマッチングの印象により大きく働くと考えているためである．すなわち，まず将来的に遷移する可能性のあるフレームの中から，キーフレームと音楽ビートの相関の高い候補を有限個選び出し，次にダイナミクスの相関を用いて遷移を選択する．2 段階目の最終的な決定方法として，以下の手法を用意している．

第 1 は，動作クリップ全体を通して，ダイナミクスの相関値の和がもっとも高いパスを選択する最適戦略である．しかし，この探索はＮＰ完全問題であり，すべての可能性を試行することは不可能である．そのため準最適なパスを選択するヒューリスティックな計算方法を導入する．各探索過程において，高い評価値を示す 5 つの遷移フレームを，3 階層にわたって計算し，その中でもっとも高い評価パスを選択する．これを順次計算することで，動作生成全体においてもっとも高い評価値を持つものが選ばれると期待できる．

第 2 はランダムに選択する方法である．この場合は，キーフレームと音楽ビートの相関で選ばれた候補の中から，ランダムな 1 つを選択することを繰り返す．この場合，生成される動きは条件が同一であった場合でも，異なる動作が生成される．キーフレームと音楽ビートの相関は評価されているため，この場合でも音楽に合った動きの生成が可能である．

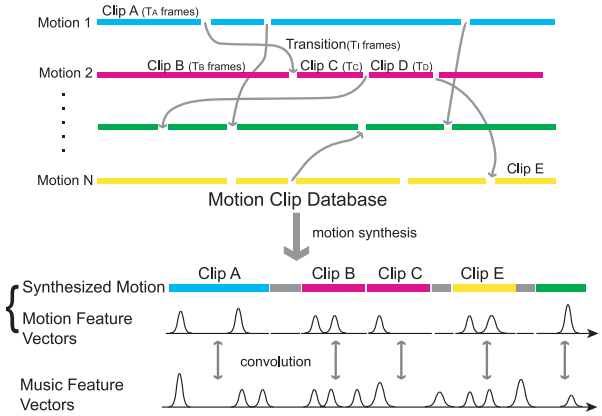

図 6　動作生成アルゴリズムの流れ．新規動作はモーショングラフをたどることにより生成される．各枝の評価値は，各フレームに与えられる動き特徴ベクタと音楽特徴ベクタの相関を評価することで求められる．システムは，生成動作全体でもっとも評価値の高い枝の列を探索することにより，もっとも音楽に合う動作を生成する．

7.　動作のオンライン生成

上記手法は音楽をオフライン解析し，全フレームについての特徴列が得られていることを前提としているが，音楽解析および動作生成を一部工夫すれば，オンラインでの動作生成が可能である．

7.1　音楽特徴量のオンライン抽出

オンライン生成時には，ビート成分のみを特徴量として用いビート構造は用いない．リアルタイムで取得された音声情報からビートを抽出するため，以下の手法を用いる．まず，ビート間隔推定における過去の自己相関関数の計算結果より，現在のビート間隔を推定する．このビート間隔を持ったパルス波形を生成し，新たに得られた音楽波形と組み合わせて再び自己相関関数の計算を行う．これにより，次のビート時刻の推定を行うことが可能となる．

7.2　オンライン動作合成

オンライン動作合成では，遷移先の選択として最適選択ではなく best-first サーチあるいはランダム選択を用いる．best-first サーチでは，システムは現在の選択肢の評価値を保持しており，その評価値よりも高い評価値を持つ遷移先が見つかった場合にはそれに遷移する．いずれの選択手法においても，探索速度の向上のため，評価値計算は遷移フレーム毎に別のスレッドで並列計算を行う．これらのスレッドは，主となる探索が別のクリップに遷移した場合，計算を終え消滅する．

8.　実　　験

本手法の有効性を確認するため 20 クリップからなるモーションキャプチャデータベースに対して適用した．このうち，10 クリップは光学式モーションキャプチャシステム (VICON8) で得

られたものであり，4 クリップは磁気式 (Ascention Tec.) で得られたものである．残る 6 クリップは CMU motion capture database [37] を使用した．光学式は，32 点のマーカーから構成されており 120Hz で記録されている．磁気式は 12〜15 点のマーカー点をもち，フレームレートは 60Hz である．処理においては，すべてのデータが図 2 で表される標準形に 120Hz で変換される．

データベースは 5 種の歩行動作 (normal walk, sad walk, active walk, kick walk, run)，エクササイズ動作，8 種の日本舞踊の動作と 6 種のダンスの基本動作からなる．クリップの長さは 231 フレーム (1.9 秒) から 2400 フレーム (20 秒) であり，平均では 844 フレーム (7.0 秒) である．

このデータベースに対してモーショングラフを生成した．約 190 組のクリップの相関を評価するが，要した時間は PENTIUM4 1.7GHz PC で約 3 時間であった．各クリップにおける連結可能フレーム数は，元のクリップの長さにより 10〜30 フレームを指定した．動作特徴量の抽出は同様の環境で 2 分以内で完了した．

オフライン生成モードでは，入力として wave ファイルを使用した．60 秒のサウンドデータに対し，特徴量抽出に要した時間は 2 分以下である．13 サンプルの音楽データを準備したが，そのうち 10 サンプルにおいてビートトラッキングが成功していることが確認できた．動作生成実験では，これら成功したサンプルを使用して行った．動作生成には，前述した最適探索とランダム選択の両者を用いた．最適探索戦略では，3 レベルの展開で平均 83 分ほど要したが，1 レベルの展開の場合は 6 分ほどで探索が完了した．図 7 に，”La Bamba”に対する動作生成結果を示す．動作生成結果より，本手法が与えられた音楽に対してよくマッチする動作を生成していることが確認できた．このことは，音楽のビート特徴とキャプチャデータのキーフレーム特徴の相関を示した図 8 からも確認できる．両者のピークがよく相関しており，提案手法が意図通り働いていることがわかる．図 9 に，別の音楽 (”Mamma-mia”) に対して，最適探索戦略とランダム選択によって生成された違いを示している．初期条件やその他のパラメータは同一であっても，両者の結果がまったく異なるものになることが確認できる．しかし，いずれの場合においても，ビートとキーフレームの相関がとられており，正しくマッチングした動作が生成されていることが確認できた．

9. 考　　察

本アルゴリズムは音楽のビート特徴と動作のキーフレーム特徴，および動きと音楽のダイナミクスの類似性を利用し，動作を合成する．既存の研究 [15] では，音楽のビート間隔が一定であることが仮定されているが，本手法では音楽の特徴を直接波形データより取り出している．事実，実験においても，音楽のビートは一定間隔ではなく，揺らぎのあることがわかっており，本手法がどのような音楽に対しても有効に働くことが示されたと考える．

動作合成においては，これらの 2 要素以外の特徴量を導入することも可能である．具体的には，音楽の調と動きのムードの関係，音楽のカテゴリーと動きのカテゴリー等の相関が考えられる．この場合，特徴ベクトルの要素を増やし，最適性評価の手法を変更するだけで対応が可能となる．

オンライン生成においては，ビート特徴のみを動作の合成に用い，ダイナミクス特徴は用いなかった．この理由は，システムは未知の音楽に対してそのダイナミクスを予想できないからであるが，これは舞踊の初心者が初めて聞く音楽に対しては踊ることができないことに類似している．一方，熟練舞踊者の場合，未知の音楽であってもある程度の動作は可能である．これは彼らが過去の経験則等により，類似の音楽や動作を想像しているためだと考えられる．現在のところ本システムでは，このような「熟練者」のテクニックをシステムに導入していないが，将来的には予測システム等を導入することでこの能力に対しても模倣可能になると考えている．

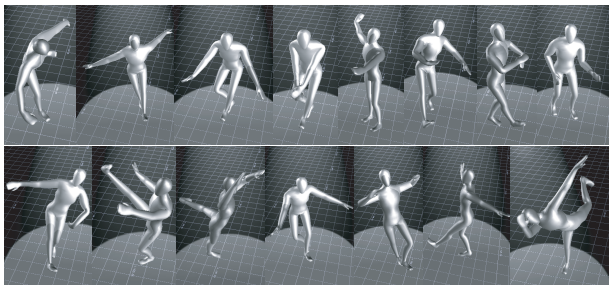

図 7　”La Bamba”への動作生成結果．上：ビートおよび調の変化が用いられた場合，下：ビートおよびオンセット成分が用いられた場合

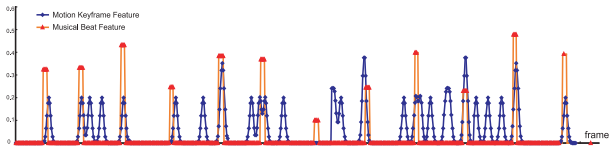

図 8　”Do Me”におけるビート成分（赤）とキーフレーム成分（青）のマッチング結果．ほとんどのビート成分が動きのキーフレーム成分と一致していることが確認できる

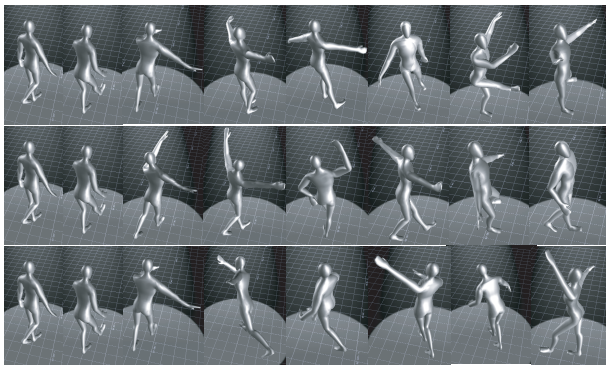

図 9　”Mamma Mia!”に対する動作生成結果．上：最適探索，中，下：ランダム選択．初期条件が同一であっても，探索手法の違いによりまったく異なる動作を生成することができる．しかし，いずれの動作も音楽に合った動作として確認できる．

10. 結　　論

本論文では，音楽を入力すると，それに合った新規舞踊動作を生成する手法を提案した．本手法は，音楽のテンポ（ビート）およびダイナミクスが，動きのキーフレームおよびダイナミクスとの相関があるという仮定に基づいている．これは，実際の舞踊者の「音楽を聴きながら踊る」能力を実現したものであるといえる．システムは音声の信号ファイルから自動的に特徴を抽出し，その特徴に良くあったモーションキャプチャデータのクリップを連結することで，音楽に合った動作を生成する．本システムはオンライン化可能であり，音楽入力からリアルタイムに動作を合成することが可能である．実験結果から，本手法が有効に働くことが確認された．これにより，音楽からのアニメーション合成やインタラクティブロボットの知能としての活用が期待できる．将来的には，本グループが提案している舞踊ロボットに組み込むことで，音楽データに対してロボット自ら動作計画を立て，舞踊を演じることが実現可能になると考えている．

文　　献

[1] 池内，中澤，小川原，高松，工藤，中岡，白鳥：民俗芸能のデジタルアーカイブとロボットによる動作提示，**日本バーチャルリアリティ学会誌**, Vol. 9, No. 2, 2004.

[2] Arikan, O., and Forsyth, D. A. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics 21*, 3, 483–490.

[3] Arikan, O., Forsyth, D. A., and O'Brien, J. F. 2003. Motion synthesis from annotations. *ACM Transactions on Graphics 22*, 3, 402–408.

[4] Brand, M. E., and Hertzmann, A. 2000. Style machines. *ACM Transactions on Graphics 22*, 3, 402–408.

[5] Blanked 2000. Blanked *Blanked*

[6] Bregman, A. S. 1990. *Auditory Scene Analysis: The Perceptual Organization of sound.* The MIT Press.

[7] Bruderlin, A., and Williams, L. 1995. Motion signal processing. In *Proceedings of ACM SIGGRAPH 95*, 97–104.

[8] Cemgil, A. T., Kappen, B., Desian, P., and Honing, H. 2001. On tempo tracking: Tempogram representation and kalman filtering. *Jounal of New Music Research 29*, 4, 259–273.

[9] Cooke, M., and Brown, G. 1993. Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication 13*, 391–399.

[10] Desain, P., and Honing, H. 1989. The quantization of musical time: A connectionist approach. *Computer Music Journal 13*, 3, 56–66.

[11] Desain, P., and Honing, H. 1994. Advanced issues in beat induction modeling: Synchopation, tempo and timing. In *Proceedings of Internationl Computer Music Conference*, 92–94.

[12] Gleicher, M. 1998. Retargetting moiton to new characters. In *Proceedings of ACM SIGGRAPH 98*, 33–42.

[13] Goto, M. 2001. An audio-based real-time beat tracking system for music with or without drum-sounds. *Jounal of New Music Research 30*, 2, 159–171.

[14] Grochow, K., Martin, S. L., Hertzmann, A., and Popović, Z. 2004. Style-based inverse kinematics. *ACM Transactions on Graphics 23*, 3, 522–531.

[15] Kim, T., Park, S. I., and Shin, S. Y. 2003. Rhythmic-motion synthesis based on motion-beat analysis. *ACM Transactions on Graphics 22*, 3, 392–401.

[16] Kovar, L., Gleicher, M., and Pighin, F. 2002. Motion graphs. *ACM Transactions on Graphics 21*, 3, 473–482.

[17] Large, E. W., and Kolen, J. F. 1994. Resonance and the perception of musical meter. *Connection Science 6*, 1, 64–76.

[18] Lee, J., and Shin, S. Y. 1999. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of ACM SIGGRAPH 99*, 39–48.

[19] Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., and Pollard, N. S. 2002. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics 21*, 3, 491–500.

[20] Li, Y., Wang, T., and Shum, H. Y. 2002. Motion texture: a two-level statistical model for character motion synthesis. *ACM Transactions on Graphics 21*, 3, 465–472.

[21] Liu, C. K., and Popović, Z. 2002. Synthesis of complex dynamics character motion from simple animations. *ACM Transactions on Graphics 21*, 3, 408–416.

[22] Logan, B., and Chu, S. 2000. Music summarization using key phrases. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing.*

[23] Lu, L., and Zhang, H.-J. 2003. Automated extraction of music snippets. In *Proceedings of ACM Multimedia*, 140–147.

[24] Nava, G. P., and Tanaka, H. 2004. Finding music beats and tempo by using an image processing technique. In *Proceedings of International Conference on Information Technology for Application.*

[25] Pullen, K., and Bregler, C. 2002. Motion capture assisted animation: Texturing and synthesis. *ACM Transactions on Graphics 21*, 3, 501–508.

[26] Rose, C., Guenter, B., Bodenheimer, B., and Cohen, M. F. 1996. Efficient motion generation of motion transition using spacetime constraints. In *Proceedings of ACM SIGGRAPH 96*, 147–154.

[27] Rosenthal, D. 1992. Emulation of human rhythm perception. *Computer Music Journal 16*, 1, 64–76.

[28] Rosenthal, D. 1992. *Machine Rhythm: Computer Emulation of Human Rhythm Perception.* PhD thesis, Massachusetts Institute of Technology.

[29] Scherier, E. D. 1998. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustic Society of America 103*, 1, 588–601.

[30] Sethares, W. A., Morris, R. D., and Sethares, J. C. 2005. Beat tracking of musical performance using low-level audio features. *IEEE Transactions on Speech and Audio Processing 13*, 2.

[31] Shao, X., Xu, C., Wang, Y., and SKankanhalli, M. 2004. Automatic music summarization in compressed domain. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing.*

[32] Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., and Bregler, C. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics 23*, 3, 506–513.

[33] Todd, N. P. M. 1994. The auditory primal sketch: A multiscale model of rhythmic group. *Jounal of New Music Research 23*, 1, 25–70.

[34] Witkin, A., and Popović, Z. 1995. Motion warping. In *Proceedings of ACM SIGGRAPH 95*, 105–108.

[35] Yamane, K., Kuffner, J., and Hodgins, J. K. 2004. Synthesizing animations of human manipulation tasks. *ACM Transactions on Graphics 23*, 3, 532–539.

[36] 中澤，中岡，竹村，池内：スケーラブル DP を用いた人間動作のマッチングと生成，**画像の認識理解シンポジウム** *(MIRU2004)*., 2004.

[37] CMU Graphics Lab. 2003. CMU Graphics Lab. Motion Capture Database *http://mocacp.cs.cmu.edu/*