5S-05

ダンス共演ロボットのためのマルチモーダルビートトラッキング

大喜多 美里[1]　坂東 宜昭[2]　池宮 由楽[2]　糸山 克寿[2]　吉井 和佳[2]

[1]京都大学 工学部 情報学科　[2]京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

ダンス共演ロボットは，人と共に踊ることを目的としたロボットである．社交ダンスの動作をインタラクティブに生成したり [1]，高度な制御でヒューマノイドロボットによる自然なダンスを実現する技術 [2] が開発されている．多くのダンスには音楽が付随しており，特に人間はダンスを音楽のテンポに同調させることが知られている [3]．そのため，ロボットがより人間と協調してダンスを踊るためには，リアルタイムで正確に音楽のビートトラッキングを行う技術が不可欠である．ここで，ビートとは楽曲のテンポと表拍時刻を指す．

音響信号のみを用いたビートトラッキング手法として，マルチエージェントに基づくもの [4] や音の立ち上がり時刻の自己相関をとる STPM (Spectro-Temporal Pattern Matching) [5] などが提案されている．しかし本目的において，前者はテンポ変動追従性，後者は音符長追従性が十分ではなかった．糸原ら [6] はギター演奏に特化し，手の動きと音響信号を用いたマルチモーダルビートトラッキングを提案した．本稿では共演者（人間）のダンスを表す骨格情報を用いることで，ダンス共演ロボットの音楽理解能力の向上のためのマルチモーダルビートトラッキングを提案する．音響信号だけでなく, 人間のダンスから得た骨格情報を用いることで性能向上を行う．

2. 視聴覚統合ビートトラッキング

ダンス共演では音楽と共演者のダンスという 2 つの情報が存在するため，両者を統合することで精度向上を図る．本稿では，音楽はマイクにより取得した音響信号，ダンスは Kinect やモーションキャプチャで取得した骨格時系列情報で表現する．音楽とダンスに含まれる情報の統合という課題に対しては，複数モーダルによる観測を用いて柔軟にモデルを組むことができる状態空間モデルを用いる．従って，本稿で扱う問題を以下のように定める．

入力 音響信号：y_t
骨格情報 (各関節の 3 次元座標)：$\boldsymbol{b}_{1,t}, \cdots, \boldsymbol{b}_{M,t}$
出力 テンポ BPM(Beats per Minute)：f_t
表拍時刻：θ_t

M は関節数，$\boldsymbol{b}_{m,t} \in \mathbb{R}^3 (m = 1, \cdots, M)$ は時刻 t での首や腰などの関節 m の 3 次元座標である．

手法概要を図 1 に示す．音響信号と骨格情報からそれぞれ特徴量抽出を行い, 得られたビート特徴量から f_t, θ_t の確率密度を状態空間モデルを用いて推定する．音響信号からの特徴量抽出にはテンポ変化追従性と対ノイズ性に優れている STPM を用いる．骨格情報からの特徴量抽出には Chu ら [7] の Rhythm of Motion (RoM) 推定法を応用する．

2.1 音響信号からの特徴量抽出：STPM

STPM は音響信号 y_t から各時間フレームごとのオンセットベクトルを求め，自己相関によりビート間隔信頼度を求める．オンセットベクトル $d(t, f)$ は，音響信号 y_t から周波数解析で得た，人間の音高知覚特性を反映したメル尺度のスペクトログラム上でパワーが上昇している時刻をソーベルフィルターを用いて検出する．ここで, f はメルフィルタバンクの次元を表す．

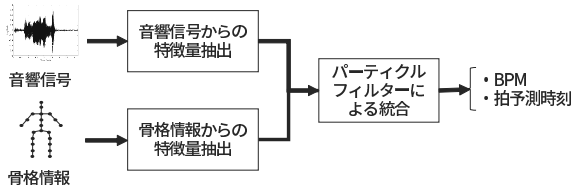


図 1: マルチモーダルビートトラッキングシステム

次に以下で定義される正規化相互相関マッチングを用いることでビート間隔信頼度 $R(t, k)$ を求める．

$$R(t,k) = \frac{\sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-i,j) d(t-k-i,j)}{\sqrt{\sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-i,j)^2 \sum_{j=1}^{F_\omega} \sum_{i=0}^{P_\omega - 1} d(t-k-i,j)^2}}$$

P_ω はパターンマッチングの窓幅で k はシフトパラメータである．$R(t, k)$ のローカルピークを表拍時刻候補として BPM と表拍時刻を計算する．これにより一般的な自己相関関数を用いるよりも短い窓幅でテンポを抽出する．

2.2 骨格情報からの特徴量抽出：RoM 推定

本手法による RoM 推定は次の 3 つの手順で行う (図 2)． 1) 各関節の骨格情報 $\boldsymbol{b}_{m,t-N:t}$ から関節の動作が停止・回転する時点を抽出し, 2) ガウス関数を用いて波形を作成, 3) 周波数領域に変換した後全関節で足しあわせ, パワーが最大になる周波数を BPM q_t とする．

1) 停止・回転する時点の抽出　関節の移動距離が極小となる時刻を停止点とし集合 $\mathcal{I}_m^{\mathrm{st}} \subset \mathbb{R}$ で表す．移動距離を $g_{m,i} = \|\boldsymbol{b}_{m,i+1} - \boldsymbol{b}_{m,i}\|$ と定め，$t - N \leq i < t$ で $g_{m,i}$ が極小となる時刻 i の集合が $\mathcal{I}_m^{\mathrm{st}}$ となる．また，関節の内積が極大となる時刻を回転点とし $\mathcal{I}_m^{\mathrm{tr}} \subset \mathbb{R}$ で表す．内積とは，$\boldsymbol{o}_{m,i} = (\boldsymbol{b}_{m,i+1} - \boldsymbol{b}_{m,i})/g_{m,i}$ としたときの $h_{m,i} = \boldsymbol{o}_{m,i} \cdot \boldsymbol{o}_{m,i+1}$ を指し，$t - N \leq i < t$ で $h_{m,i}$ が極大となる時刻 i の集合が $\mathcal{I}_m^{\mathrm{tr}}$ となる

2) 波形の作成　$\mathcal{I}_m^{\mathrm{st}}, \mathcal{I}_m^{\mathrm{tr}}$ をもとにガウス関数を用いて波形 $y_m^{\mathrm{st}}(t), y_m^{\mathrm{tr}}(t)$ を作成する．$\mathcal{N}(x|\mu, \sigma)$ は変数を x とする平均 μ，分散 σ^2 の正規分布の確率密度関数である．

$$y_m^{\mathrm{st}}(t) = \sum_{i \in \mathcal{I}_m^{\mathrm{st}}} \mathcal{N}(t|i, \sigma^2), \quad y_m^{\mathrm{tr}}(t) = \sum_{i \in \mathcal{I}_m^{\mathrm{tr}}} \mathcal{N}(t|i, \sigma^2)$$

3) 周波数領域での足しあわせ　$y_m^{\mathrm{st}}(t), y_m^{\mathrm{tr}}(t)$ をフーリエ変換により周波数領域に変換したものを $\hat{y}_m^{\mathrm{st}}(f), \hat{y}_m^{\mathrm{tr}}(f)$ とする. それらを全関節で足し合わせパワーが最大となる周波数が時刻 t における BPM q_t である．

$$S_t(f) = \sum_{m=1}^{M} (|\hat{y}_m^{\mathrm{st}}(f)| + |\hat{y}_m^{\mathrm{tr}}(f)|)$$

$$q_t = b \times \operatorname*{argmax}_f S_t(f)$$

ただし, b は周波数を BPM に変換する定数である．

Multi-modal Beat Tracking for a Dancing Robot Playing with Humans: *Misato Ohkita, Yoshiaki Bando, Yukara Ikemiya, Katsutoshi Itoyama, Kazuyoshi Yoshii(Kyoto Univ.)*

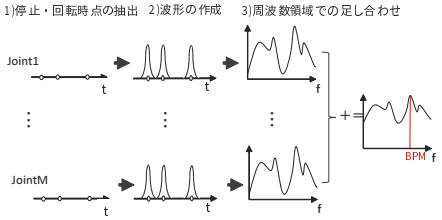
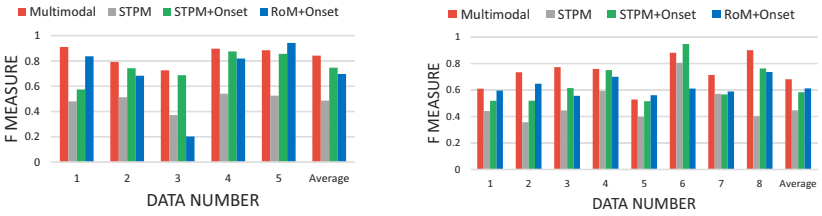


図 2: RoM 推定提案手法概要図



1. MotionCapture data　　2. Kinect data

図 3: 評価結果. 横軸はデータ番号を表す.

2.3　統合：パーティクルフィルタ

音響特徴量と視覚特徴量の統合は，状態空間モデルを用いて行う．現在のテンポを表す状態変数 $\boldsymbol{x}_k$ は表拍時刻 θ_k と BPMf_k で表現する．

$$\boldsymbol{x}_k = [f_k, \theta_k]^{\mathrm{T}}$$

観測モデル　観測変数 $\boldsymbol{z}_k$ は STPM の出力である BPM M_k，オンセットベクトル $d(\theta_k, f)$ の周波数方向の和を正規化したオンセット和 $F_k(\theta_k) = \sum_f d(\theta_k, f)$ ，RoM 推定で得られるパワースペクトル $S_k(f_k)$ である．

$$\boldsymbol{z}_k = [M_k, S_k(f_k), F_k(\theta_k)]^{\mathrm{T}}$$

観測変数は全て独立とみなし，観測モデルを以下とする．

$$p(\boldsymbol{z}_k|\boldsymbol{x}_k) = p(M_k|\boldsymbol{x}_k)p(S_k(f_k)|\boldsymbol{x}_k)p(F_k(\theta_k)|\boldsymbol{x}_k)$$
$$p(M_k|\boldsymbol{x}_k) = \mathcal{N}(M_k|f_k, \sigma_M)$$
$$p(S_k(f_k)|\boldsymbol{x}_k) \propto S_k(f_k)$$
$$p(F_k(\theta_k)|\boldsymbol{x}_k) \propto F_k(\theta_k)$$

状態遷移モデル　状態遷移は random walk で表現する．

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}_k|[f_{k-1}, \theta_{k-1} + 60/f_{k-1}]^{\mathrm{T}}, \mathbf{Q})$$

$\mathbf{Q} \in \mathbb{R}^{2\times 2}$ はモデル誤差を表す共分散行列である．

推論アルゴリズム　状態空間モデルが非ガウスモデルであるため，本状態空間の推定にはパーティクルフィルタを用いる．また，計算効率化のために SIR (Sequential Importance Resamling) パーティクルフィルタ [8] を用いて行う．提案分布は状態遷移モデルと等しい．

3.　実験

本手法の有効性を確認するため，次の手法と比較する：STPM 単体，観測変数 $\boldsymbol{z}_k$ から RoM 特徴量 S_k を除いた場合 (STPM+Onset)，観測変数 $\boldsymbol{z}_k$ から STPM 特徴量の BPMM_k を除いた場合 (RoM+Onset)．実験には the Dance Motion Capture Database of the University of Cyprus[§]のモーションキャプチャデータ 5 曲と，ダンス経験者によるポップスの曲に合わせたミックスダンスを Kinect で取得したデータ 8 曲を使用した．後者では，残響下における音楽音響信号をマイクで録音した．また，データ 1,3,4,6,8 には音声や手拍子が含まれている．

STPM の計算にはロボット聴覚ソフトウェア HARK[¶]を使用し，BPM は 91〜180bpm に制限している．表拍時刻の推定結果と正解結果の差が ±100ms 以内ときを正解とし，各データについて適合率 (=推定成功拍数/検出拍総数)・再現率 (=推定成功拍数/正解拍総数) から F 値を求めた．これは音の立ち上がりのタイミングが ±100ms 以上の場合は人間には音がずれて感じられることに基づいている [9]．各データに対してパーティクルの初期値を変更して 5 回推定を行い F 値の平均により評価した．

[§]http://dancedb.cs.ucy.ac.cy/　[¶]http://www.hark.jp/

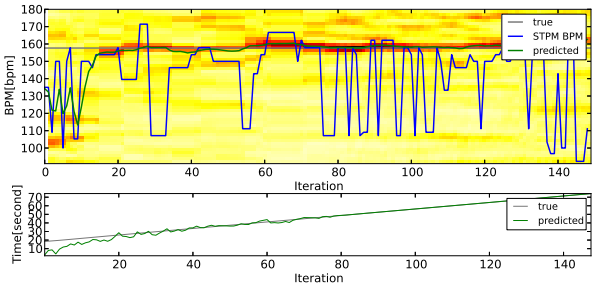


図 4: 推定結果 (緑) と正解 (灰) の一例．横軸は推定回数．上図は BPM で M_k(青) と S_k(濃淡)．下図は表拍時刻．

図 3 に評価結果を示す．図 4 は推定結果の一例であり，BPM・表拍時刻推定結果が共に正解に収束している様子が分かる．評価結果から，本手法により全データにおいて STPM より精度が向上し，特に両データセットについて本手法が平均で最も高い精度を実現した．一方，モーションキャプチャデータのデータ 3 において RoM+Onset の推定結果が 0.2 程度となっているのは，足が細かく動くダンスで RoM 推定が失敗したためであると考えられる．Kinect データにおいて推定結果平均がモーションキャプチャデータより低いのは，モーションキャプチャで取得している関節数が 54 であるのに対し Kinect では 16 であること，さらに Kinect で生じるオクルージョンが原因であると考えられる．

4.　おわりに

本稿では音響信号と骨格情報の特徴量をパーティクルフィルタにより統合したマルチモーダルビートトラッキングを開発した．本手法により，従来の各特徴量を個別に用いた手法と比較し，ビートトラッキングの精度が向上することを確認した．今後は Kinect データでの精度向上，実ロボットを用いた評価などを行う．

謝辞　本研究の一部は、科研費 24220006，26700020，24700168 および OngaCREST プロジェクトの支援を受けた．

参考文献

[1] K. Kosuge *et al.* Partner Ballroom Dance Robot-PBDR. *SICE*, 2011.
[2] K. Kaneko *et al.* Cybernetic Human HRP-4C. *Humanoids*, 2009.
[3] T. Shiratori *et al.* Detecting Dance Motion Structure through Music Analysis. *FGR*, 2004.
[4] M. Goto. An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. *J. New Music Research*, 2001.
[5] K. Murata *et al.* A Beat-Tracking Robot for Human-Robot Interaction and Its Evaluation. *Humanoids*, 2008.
[6] T. Itohara *et al.* Particle-filter Based Audio-visual Beat-tracking for Music Robot Ensemble with Human Guitarist. *IROS*, 2011.
[7] C. Wei-Ta *et al.* Rhythm of Motion Extraction and Rhythm-Based Cross-Media Alignment for Dance Videos. *ACM Multimedia*, 2012.
[8] M. Sanjeev *et al.* A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *Signal Processing*, 2002.
[9] R. A. Rasch. Synchronization in Performed Ensemble Music. *J. Acta Acustica united with Acustica*, 1979.