1ZB-2

視聴覚統合ビートトラッキングを用いた音楽ロボットとギターとの合奏システム

糸原 達彦 † 大塚 琢馬 ‡ 水本 武志 ‡ 高橋 徹 ‡ 尾形 哲也 ‡ 奥乃 博 ‡

† 京都大学 工学部情報学科 ‡ 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

人とロボットとのインタラクションが重要視されている中で，音楽ロボットに関する研究が盛んに行われている [1]．中でも、ビートトラッキングは入力された音楽の小節内位置及び楽曲のテンポを推定する音楽ロボット合奏の基礎技術である．しかし，人のギター伴奏のような演奏テンポの揺らぎや裏拍ビートを含む多様なリズムに頑健なビートトラッキングのためには，マルチエージェント [2] ではテンポ変動追従性が，STPM(Spectro-Temporal Pattern Matching)[3] では音符長追従性が不足していた．これらはトレードオフの関係にあり，聴覚情報を扱うだけでは解決が困難である．

そこで、本稿では視聴覚の両情報を用いたビートトラッキングを報告する．聴覚情報は STPM で，視覚情報は手のトラッキングで取得し，両者をパーティクルフィルタを用いて情報統合し，多様なリズムに対し頑健な拍時刻とテンポ推定を達成する．

2. 問題設定と課題

入力 (1) ロボットのマイクで取得したノイズ込のギター音，(2) ロボットのカメラからの映像

出力 (1) テンポ (BPM)，(2)1 小節内のどの部分であるか (以下，小節内位置)

課題 (1) ロボットノイズへの対処
(2) 人の演奏揺らぎへの追従性
(3) 手のトラッキング: 1) 手の色の人や場所による多様性，2) カメラや演奏者の揺らぎへの頑健性，3) 時間解像度の低さ

本研究で扱う合奏形態は“ロボット (主旋律) と人のギター (伴奏) ”である．ギター伴奏はその演奏動作とビート時刻・間隔の相関が強いので，ビートトラッキングに有利である．一方，伴奏の 1 小節内のビートのパターン (以下，ビート構造) が，1 小節を偶数個に分けたときの奇数番目の拍 (表拍) だけでなく偶数番目の拍 (裏拍) にもアクセントが置かれ易いという課題がある．本稿では，ギター演奏と手の運動に関して，手を振る演奏動作 (バッキング)，手の振りが下から上ならば裏拍，演奏前カウントからテンポは大きく逸脱しないを仮定する．

3. 視聴覚統合ビートトラッキング

視聴覚統合ビートトラッキング問題を，視聴覚情報からなる観測 $z_{1:t}$ が与えられたとのき，テンポ・小節内位置・手の振幅 x_t の確率密度 $p(x_t|z_1:t)$ の逐次推定問題と定式化する．本逐次推定をパーティクルフィルタを用いて解く．処理の概要を図 1 に示す．

3.1 聴覚特徴量：STPM

本稿では文献 [3] で定義されるビート間信頼度 $R_t(i)$ と，オンセットベクトルの周波数方向の和を正規化した F_t をパーティクルフィルタの入力に用いる．ビート間信頼度は i フレーム前とのオンセットベクトルを用いた正規化相互相関関数で，オンセットベクトルの各成分は周波数ビンごとの音の立ち上がり度合で表される．この手法の利点は，定常雑音の白色化による高い定常雑音頑健性，マッチングの窓幅が小ささによるテンポ変化への高適応性である．また最大 1 秒とレイテンシが比較的低いので実時間処理に適している．

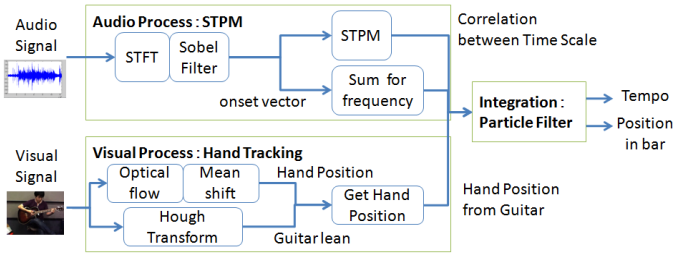


図 1: システムの概要図

3.2 視覚特徴量：手のトラッキング

手のトラッキングを以下の 3 つの手順で行う．

(1) オプティカルフローによる手の存在範囲の推定：画像の 2 フレーム間での差分をオプティカルフローで求める．カメラや人の揺らぎを考慮するため，ここでは手の存在範囲のみを取得する．

(2) 平均値シフト法の適用による手の座標 $(h_{x,t}, h_{y,t})$ の推定：平均値シフト法 [4] は与えられたデータセット内での極大点を見つける手法で，さらに注目点以外のデータの異常値に頑健であるので，比較的安定な手の座標を得られる．カーネルに D.Miyazaki らの色空間 [5] で得た色相ヒストグラムを用いる．この色空間は明るさの変化に頑健であり，影や鏡面反射に影響されにくいので，他の色空間より人の肌の検出精度が高い．

(3) 手の軌跡のモデル化：時刻 t での手の位置 r_t を $r_t = \rho_t - h_{x,t}\cos\theta_t + h_{y,t}\sin\theta_t$ と定義する．ただし，(ρ_t, θ_t) は Hough 変換で得たギターの直線パラメータである．r_t の正負がそれぞれギターの上，下に手があることを表す．ここで，1 小節を一周の円でモデル化し，時刻 t のビート間隔とビート時刻をそれぞれ ω_t，θ_t とする．ただし，$0 \le \theta_t < 2\pi$ で，$\theta_t = \pi/2 * n$ は表拍を，$\theta_t = \pi/2 * n + \pi/4$ は裏拍を表すとする $(n = 0,1,2,3)$．また ω_t はテンポ (円の角速度) の逆数に比例する．これらと手の振幅 a_t を用いて，手の位置 r_t は $-a_t \sin(4\theta_t)$ と表せる．

3.3 視聴覚統合：パーティクルフィルタ

時刻 t の状態変数を，θ_t:小節内位置，ω_t:ビート間隔，a_t:手の振幅と定義する．観測変数は，i フレーム前との相関関数 $R_t(i)$，正規化オンセット和 F_t，ギターと手の距離 r_t を用いる．多数のパーティクルを用いて，状態変数の確率密度 $p(\theta_t, \omega_t, a_t | R_t(i), F_t, r_t)$ を次式で推定する．

$$p(\theta_t, \omega_t, a_t | R_t(i), F_t, r_t) = \sum_{i=1}^{I} w_n^i [\theta_t^i, \omega_t^i, a_t^i] \quad (1)$$

ただし，i はパーティクルの添え字，I はパーティクル総数であり，$w_t^i \theta_t^i, \omega_t^i, a_t^i$ はそれぞれ i 番目のパーティクルの重み，小節内位置，ビート間隔，手の振幅である．

Music Robot Ensemble System with Human Guitarist using Audio-Visual Beat Tracking: *Tatsuhiko Itohara, Takuma Otsuka, Takeshi Mizumoto, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

パーティクルフィルタで行う処理を示す: (1) 提案分布からのサンプリング，(2) パーティクルの重み計算，(3) 状態の推定・リサンプリング．

(1) 提案分布からのサンプリング：時刻 $t-1$ における観測から，時刻 t の状態 $[\theta_t^i, \omega_t^i, a_t^i]$ を次式を用いてサンプルする．

$$\begin{aligned}\omega_t^i &\sim q(\omega_t|\omega_{t-1}^i, R_t(i), \omega_{init}) \\ &\propto R_t(\omega_t) \times \mathrm{G}(\omega_t|\omega_{t-1}^i, \sigma_{\omega_q}) \times \mathrm{G}(\omega_t|\omega_{init}, \sigma_{\omega_q}) \quad (2)\\ \theta_t^i &\sim q(\theta_t|r_t, F_t, \omega_{t-1}^i, \theta_{t-1}^i) \\ &= \mathrm{G}(\theta_t|\theta_{t-1}^i + \omega_{t-1}^i, \sigma_{\theta_q}) \times \mathrm{penalty}(\theta_t^i|r_t, F_t) \quad (3)\\ a_t^i &\sim q(a_t|r_t, \omega_{t-1}^i, \theta_{t-1}^i) \\ &= \mathrm{G}(a_t|r_t / \sin(4(\theta_{t-1}^i + \omega_{t-1}^i \Delta T)), \sigma_{a_q}) \quad (4)\end{aligned}$$

ただし，$\mathrm{G}(x|\mu,\sigma)$ は平均 μ，分散 σ の変数 x のガウス関数，$\sigma_{\omega_q}, \sigma_{\theta_q}, \sigma_{a_q}$ はそれぞれビート間隔，小節内位置，手の振幅のサンプリング分散，ω_{init} は初期カウントで得たビート間隔，ΔT は 1 フレーム間の時間を表す．また，$\mathrm{penalty}(\theta|r,F)$ は手の位置の遷移の仮定，また音のピークの有無に従う窓関数である．

(2) パーティクルの重み計算：次式で更新する．

$$w_n^i = \frac{p(\omega_t^i, \theta_t^i, a_t^i|\omega_{t-1}^i, \theta_{t-1}^i, a_{t-1}^i) p(R_t(\omega), F_t, r_t|\omega_t^i, \theta_t^i, a_t^i)}{q(\omega_t^i, \theta_t^i, a_t^i|R_t(i), F_t, r_t, \omega_{t-1}, \theta_{t-1}, a_{t-1})} \quad (5)$$

式 (5) の分子の項はそれぞれ状態遷移モデル，観測モデルと呼ぶ．分母は状態変数のサンプルされた提案分布である．パーティクルの持つ値が各モデルに合致するほど重み w_n^i は大きくなり，提案分布からサンプルされやすい値ほど重みは小さくなる．

状態遷移モデルを導出する状態方程式を以下に示す．

$$\theta_t = \theta_{t-1} + \omega_{t-1}\Delta T + n_\theta \quad (6)$$
$$\omega_t = \omega_{t-1} + n_\omega \quad (7)$$
$$a_t = a_{t-1} + n_a \quad (8)$$

ここで n_ω, n_θ, n_a はそれぞれビート間隔，小節内位置，振幅に関するガウス分布に従うノイズである．それぞれの平均を 0，分散を $\sigma_\omega, \sigma_\theta, \sigma_a$ とすると，状態遷移モデルは以下のようになる．

$$p(\omega_t|\omega_{t-1}) = \mathrm{G}(\omega_t; \omega_{t-1}, \sigma_\omega) \quad (9)$$
$$p(\theta_t|\theta_{t-1}, \omega_{t-1}) = \mathrm{G}(\theta_t; \theta_t + \omega_{t-1}\Delta T, \sigma_\theta) \quad (10)$$
$$p(a_t|a_{t-1}) = \mathrm{G}(a_t; a_{t-1}, \sigma_a) \quad (11)$$

次に観測モデルを導く． $R_t(\omega)$, r_t の観測はそれぞれ $\omega_t^i, -a_t^i \sin(4(\theta_t + \omega_t^i \Delta T))$ を平均とするガウス分布に従うとする．F_t は $\mathrm{G}(\theta_{beat_t}, \sigma_f) + bias.$ という形で近似できる．ここで θ_{beat_t} はフレーム t に対して最も近い表拍か裏拍の小節内位置，$bias.$ は $[0.35, 0.5]$ の一様分布に従う．従って観測確率は以下のように表せる．

$$p(R_t(\omega_t)|\omega_t^i) = \mathrm{G}(\omega_t; \omega_t^i, \sigma_\omega) \quad (12)$$
$$p(F_t|\omega_t^i, \theta_t^i) \propto \mathrm{G}(F_t; \theta_{beat_t}, \sigma_f) + \mathrm{G}(F_t; 0.4, 0.05) \quad (13)$$
$$p(r_t|\omega_t^i, \theta_t^i, a_t^i) = \mathrm{G}(r_t; -a_t^i \sin(4(\theta_t + \omega_t^i \Delta T)), \sigma_a) \quad (14)$$

(3) 状態の推定・リサンプリング：最終的な推定は重み付きの平均で求める．リサンプリングは各パーティクル重みに深刻な縮退が生じたときに行う操作で，各パーティクルの重みに比例した確率で，I 個のパーティクルを独立にサンプルすることで行う．

4. 実験及び考察

本手法の評価実験を 2 名の演奏者にそれぞれ，テンポ 3 種類，ビート構造 8 種類，計 24 の演奏パターンを用いて行った．聴覚情報にはロボットのファンノイズが混入

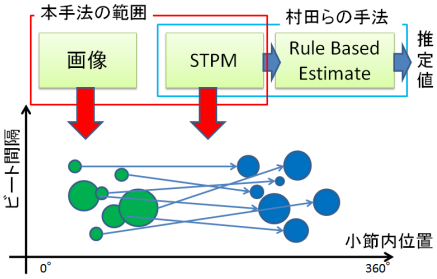
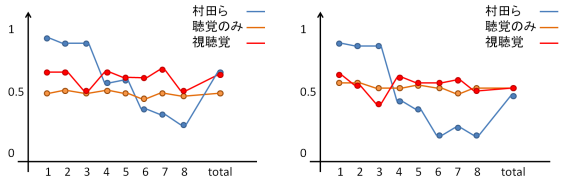


図 2: 従来手法との比較



(a) 適合率　(b) 再現率

図 3: 評価結果

している．カメラのフレームレートは約 19[fps] である．演奏者とロボットとの距離は 3.5[m] で，ギター全体がカメラフレーム内に収まるよう調整した．以上の条件で，本手法 (視聴覚統合，聴覚のみ) と村田らの手法 [3] の 3 手法を比較した．図 2 に本手法と村田らの手法の比較イメージを示す．評価基準には，推定結果と正解結果の差が ±200[ms] 以内かつ推定テンポが ±10[bpm] 以内のときを正解とし，それぞれの演奏パターンについて適合率 ($r_{prec} = N_e/N_d$)，再現率 ($r_{recall} = N_e/N_c$) を用いた．ただし，N_e は推定成功拍数，N_d は検出拍総数，N_c は正解拍総数である．実験結果の演奏パターンごとの総計を図 3 に示す．図の左 3 つは全ての表拍に音が存在するパターンであり，右に行くにつれて表拍の音数が少なく，裏拍の音数が多くなる．グラフからもわかるように村田らの手法は演奏パターンへの依存性が非常に強いが，本手法は演奏パターンによらない結果を出している．また，視聴覚統合の結果で他よりも極端に成績が悪いパターンは，裏拍での仮定合わない，3 連符のようなものを用いたことによるものである．また，本手法の結果が 0.5 程度になっている原因は視覚情報 r の fps の低さと誤推定の多さにより，ビート誤推定の修正分布である $\mathrm{penalry}(\theta|r,F)$ の影響が小さくなることだと考えられる．

5. おわりに

本稿では視聴覚の特徴量を用いたパーティクルフィルタによる視聴覚統合ビートトラッキングを開発した．残る課題として，精度向上，ロボットでの実装などがあげられる．また，赤外線センサを持つデバイスを使うことで手のトラッキング部の性能向上が期待できる．

謝辞 本研究の一部は科研費 (S)，特定領域，新学術領域，GCOE の支援を受けた．STPM の使用許可をいただいた HRI-JP に感謝します．

参考文献

[1] Takeshi Mizumoto *et. al*: “Integration of flutist gesture recognition and beat tracking for human-robot ensemble”, *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, 2010.

[2] Masataka Goto: “An Audio-based Real-time Beat Tracking System for Music WithorWithout Drum-sounds”, *J. of New Music Research*, pp.159-171, 2001.

[3] Kazumasa Murata *et. al*: “A Robot Singer with Music Recognition Based on Real-Time Beat Tracking,” *ISMIR*, pp.199-204, 2008.

[4] D. Comaniciu and P. Meer: “Mean Shift Analysis and Applications,” *in Proc. of Int. Conf. on Computer Vision*, pp. 1197-1203, 1999.

[5] D. Miyazaki *et. al*: “Polarization-based Inverse Rendering from a Single View,” *in Proc. of Int. Conf. on Computer Vision*, pp.982-987, 2003.