5ZD-1

Kinectによる楽器マスキングを用いた視聴覚統合ビートトラッキング

糸原 達彦 † 大塚 琢馬 † 水本 武志 † 中臺 一博 ‡ 尾形 哲也 † 奥乃 博 †

†京都大学 大学院情報学研究科 知能情報学専攻 ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン (HRI-JP)

1. はじめに

楽器を演奏するロボット (音楽ロボット) と人が同期して演奏できれば, 人とロボットの共生を促進すると期待できる. 音楽ロボットが人と合わせて演奏を行うためには, 人の演奏から小節内位置とテンポを推定するビートトラッキングが必須である. 我々はこれまでに音響情報と相関の強い弾き手の軌道を用いた視聴覚統合ビートトラッキングを開発した [1]. しかし, ギターの動きや手との色が類似していることが原因で, 十分な手の軌道追従及びビートトラッキング性能ではなかった.

本稿では, 視聴覚センサに加えて深度センサを持つ Kinect を用いて, 深度情報によるギターのマスキングを行い, 手の領域を抽出することで, 上記の問題を解決する. Kinect によるギターのマスキングには 3 次元ハフ変換を用いる. 本手法により, 従来の RGB のみの手法に比べ最小二乗誤差距離が 13.3[pixel] 減少した. 一方, ビートトラッキングの精度の向上は得られなかった. これに関して, Kinect の解像度の高さと使用した手の軌道モデルについての議論を行う.

2. 従来の視聴覚統合ビートトラッキング

ビートトラッキングのまとめ

入力 Kinect で取得した音声, 画像 (RGB, 深度)
出力 (1) テンポ (BPM)
(2) 小節内位置
課題 ギターのビートトラッキング:
(1) 周囲雑音への対処
(2) 人の演奏揺らぎへの追従
(3) 複雑なギタービートパターンへの追従
手のトラッキング:
(1) 手とギターの色の近さへの頑健性
(2) 演奏者の揺らぎへの頑健性

2.1 合奏タスクと課題

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. 演奏の最初に, 共演者とタイミングやテンポを合わせるために “カウント” を行う. これは主に声やギターの打撃音で行われる. 演奏テンポはカウントで示されたテンポから大きく逸脱しないものとする.

ロボットのセンサには Kinect を用いる. Kinect は視覚情報として, RGB 情報及び赤外線センサを用いた距離の情報 (深度情報) を取得できる. また, マイクは 4 つあるうちの 1 つを用いた. ビートトラッキングは各時刻のテンポと小節内位置を出力する. 小節内位置は各小節を基準とした現在の演奏位置 ([rad]) とした.

音響ビートトラッキングの取り組むべき課題は, (1) 周囲雑音への対処, (2) 人の演奏揺らぎへの追従, (3) 複雑なビートパターンへの追従である. 本稿では 1 小節内のアクセント (音量の大きい拍) 位置が, 表拍で少なく, 裏拍で多い場合, そのビートパターンは複雑であるとする. こ

Audio-visual beat-tracking with hand-tracking of a guitarist using instrumental masking with Kinect: *Tatsuhiko Itohara*†, *Takuma Otsuka*†, *Takeshi Mizumoto*†, *Kazuhiro Nakadai*‡, *Tetsuya Ogata*†, *and Hiroshi G. Okuno*† (†*Kyoto Univ.* ‡*HRI-JP*)

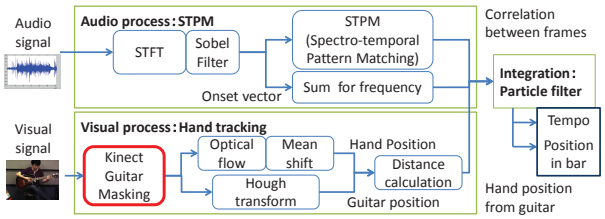


図 1: 楽器マスキングを用いた視聴覚ビートトラッキング.

こで, 表拍, 裏拍はそれぞれ 1 小節を偶数個に分けたときの奇数番目の拍, 偶数番目の拍とする. 複雑なビートパターンは他楽曲にも多く含まれるが, ギターの単独演奏は音がまばらで, 人も検出誤りを起こす困難な問題である. この問題を解決するために, これまで, 我々は手のトラッキングと統合し, まばらな音声情報を補完する, 視聴覚統合ビートトラッキングを報告した (2.2 節参照)[1].

通常, 物体のトラッキングに用いる特徴量は, Condensation[2, 3] などの色情報, もしくはオプティカルフローに代表される移動情報である. しかし, 手とギターの色が似ているため, 前者の色情報のみを用いた手のトラッキングは困難である. またギター演奏では, 追従したい手以外の物体, 例えば頭, 足, ギター本体も動くため, 後者の移動情報のみではトラッキングが難しい.

2.2 従来の視聴覚統合ビートトラッキング

2.1 節で述べた視聴覚統合ビートトラッキングは, 音響情報に加え, 手のストローク動作の画像情報を用いたビートトラッキングである. これにより, ギター演奏のような音がまばらで拍検出が難しい状況下でのビートトラッキングが可能になった. 音響特徴量には, STPM(Spectro-temporal Pattern Matching) から得られる STPM ベクトルとオンセットベクトルを用いた. STPM は従来のビートトラッキング [4] でも使用されており, 得られる特徴量は定常雑音や楽曲のテンポ変化に頑健である. 画像特徴量には手とギターの距離を用いた. 手の位置はオプティカルフローと平均値シフトを, ギターの位置はハフ変換を用いて位置検出を行う. これらの特徴量をパーティクルフィルタで統合する. 時刻 t の手とギターの距離 r_t のモデルには sin 関数を利用した (式 (1)).

$$r_t = -a\sin(4\theta_t), \tag{1}$$

a は手の振幅 (定数), θ_t は小節内位置である. これにより手の移動と音響情報の変化を紐づけし人の揺らぎ追従と複雑なビートパターンの両方に頑健な処理を実現した.

2.3 従来法の問題点

従来法では手のトラッキング精度に問題があった. 手のトラッキングでは, (1)“画像内で手が最も大きく動く” という仮定の下, オプティカルフローによりフレーム間変位を求め, (2) その中央値の座標を初期値とした色相カーネルの平均値シフトにより詳細な位置を求める. (2) で中央値をとることで, (1) のノイズ動作の除去が可能であり, またその後, 色情報を用いた平均値シフトを行うことで, 探索範囲を手を中心とした範囲に絞ることができる, ギターの色情報に対してある程度頑健である. しかし, その

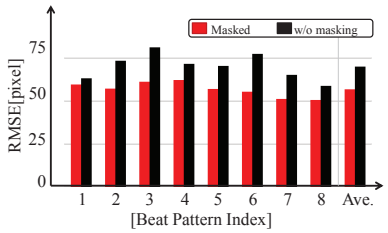


図 2: 結果:手のトラッキング

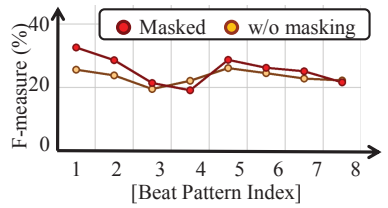


図 3: 結果:ビートトラッキング

ノイズ除去性能や範囲の絞り方は十分ではなく, ギターの部分を誤って追従する問題があった.

本稿では, Kinect の深度センサから得られる深度情報を用いたギターのマスキングを行い, 手のトラッキングの精度の向上する手法を提案する (図 1 参照).

3. 深度情報によるギターマスキング

Kinect の入力は, サイズ 640×480[pixel] の RGB と深度画像である. 深度画像は各ピクセル座標における x, y, z 方向の値 (単位:[m]) である. x, y, z 正の方向はそれぞれ Kinect から見て左, 鉛直下向き, カメラ方向である.

以下にギターのマスキングの過程を示す.

1. 背景閾値以上の奥行き (z) を持つ座標を, RGB 画像, 深度画像においてマスクする
2. 深度画像を縮小. 非マスク部を “特徴点” とする.
3. 特徴点からギターの平面を導出
4. 深度画像の各座標と 3 の平面の距離を計算し, 閾値以下なら対応する RGB 画像上の点をマスクする

本稿では, 背景閾値を 3[m], 画像の量子化は 16×12[pixel], 4. の平面との距離閾値は 5[cm] とした.

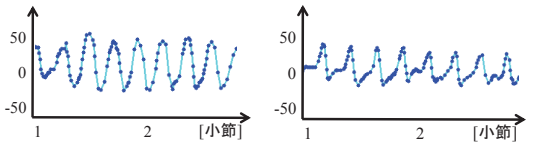
3. におけるギターの表板の検出は Hough 変換で行う. Hough 変換では, 画像中の各特徴点を通る全ての平面のパラメータを算出, パラメータ空間に対して投票を行い, 最大票を獲得したものを推定平面のパラメータとする. 平面を表すパラメータに, 球座標における原点 (カメラ) を始点とする平面の法線ベクトル (ρ, θ, ϕ) を使用した. ρ, θ, ϕ はそれぞれ原点と平面の距離, 偏角, 方位角を表す.

4. 実験及び考察

ビートトラッキングを行うデータとして, 1 名の演奏者にそれぞれ, テンポ 3 種類, ビートパターン 8 種類, 計 24 の演奏パターンを用いた演奏データを用いた. 使用したビートパターンの順番は, 数字が大きくなるに連れて表拍アクセントが減り, 裏拍アクセントが増えるように設計した. 聴覚情報には環境雑音としてロボットのファンノイズが混入している. カメラのフレームレートは約 20[fps] である. 演奏者と Kinect との距離は 2[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. 以下で, ギターマスキングの有り無しの場合での, 手のトラッキング, ビートトラッキングの評価を行う.

4.1 手のトラッキング

手のトラッキングの最小二乗距離による比較を図 2 に示す. 図の横軸は演奏パターンである. 全パターンにおいて, マスク後のトラッキングの方が勝っており, 平均で



(a) パターン 1 (b) パターン 3

図 4: 手の軌道の正解データの例

13.3[pixel] 減少した. 特に, 手の移動が一定間隔でないかつ短い時間に大きく移動するパターン 3, 6 では大きな改善がみられる. 逆に比較的等間隔で動きも緩やかなパターン 1 では, 有意な差は認められなかった.

4.2 ビートトラッキング

ビートトラッキングの評価として, 視聴覚統合のギターマスキングの有, 無の 2 手法を比較する. 評価基準には, 推定結果と正解結果の差が ±150[ms] 以内かつ推定テンポが ±10[bpm] 以内のときを推定成功とし, それらの適合率, 再現率をそれぞれ $(r_{prec} = N_e/N_d)$, $(r_{recall} = N_e/N_c)$ で定義する. ただし, N_e, N_d, N_c はそれぞれ推定拍数, 推定成功拍数, 正解拍数を表す. 比較の指標として, それらの調和平均である F 値 $(2/(1/r_{prec} + 1/r_{recall}))$ を用いる.

実験結果の演奏パターンごとの総計を図 3 に示す. ギターマスキング有りと無しでは有意な差は認められなかった. 絶対値としても決して高いとは言えず, これは Kinect を使う以前 [1] と比べても全体で 39.0%の低下となっている. この理由の検証のために, 手のトラッキングの正解データを入力としてビートトラッキングを行った. 結果は平均で 37.7%であり, 同様に低下が認められた.

以上の精度低下の原因は, sin 曲線のような単純な手の軌道モデルを用いたためと考えられる. 図 4 に手の軌道の正解データの例を示す. 正解データは人が各フレームを見て手の位置を記録して作成した. 図 4(a) は形が sin 曲線に似ており, ビートトラッキングの結果も比較的よいが, 図 4(b) は sin 曲線とは程遠く, ビートトラッキングの結果も悪い. 従来用いていたカメラは, フレームレートが 15[fps] と低く, sin 関数のような単純な近似でモデル化できた. しかし, Kinect のフレームレートは 30[fps] と比較的高く, より精密な手の軌道が取得できる. 今後, さらにフレームレートが高いデバイスの開発が期待でき, 手のモデルの改良の必要性が本実験を通じて示唆された.

5. おわりに

本稿では Kinect の深度情報を用いたギターのマスキングを行うことで, ギター演奏の手のトラッキング精度向上を報告した. また視聴覚統合ビートトラッキングにおける手の軌道モデルについての議論を行った. 今後, 更なるモデル検証及び, モデルの改良が必要である. また残る課題として, 音高情報を用いての精度向上が挙げられる. ギター演奏では, メロディーと和音情報のみの楽譜が多く用いられており, これとのスコア追従との組み合わせによる精度向上が期待できる.

謝辞 本研究の一部は科研費 (S), GCOE の支援を受けた. また, HRI-JP の中村圭佑氏の協力に感謝します.

参考文献

[1] T. Itohara *et al.* Particle-filter based audio-visual beat-tracking for music robot ensemble with human guitarist. In *Proc. of IROS*, pages 118–124. IEEE, 2011.

[2] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *Int'l J. of computer vision*, 29(1):5–28, 1998.

[3] K. Petersen *et al.* Development of a real-time instrument tracking system for enabling the musical interaction with the Waseda Flutist Robot. In *Proc. of IROS*, pages 313–318. IEEE, 2008.

[4] K. Murata *et al.* A beat-tracking robot for human-robot interaction and its evaluation. In *Proc. of Humanoids*, pages 79–84. IEEE, 2008.