

パーティクルフィルタを用いた ギター演奏の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 尾形 哲也 奥乃 博 (京都大学大学院 情報学研究所)

1. はじめに

本研究では、音楽ロボットのための視聴覚情報の統合による人のギター演奏のビートトラッキングを報告する。ビートトラッキングとは音楽のテンポとビート時刻を推定する手法で、合奏においてタイミングの合った演奏のために不可欠である。本研究では、伴奏楽器としてメジャーで演奏人口も多いギター演奏を扱う。この時、(1) 人の演奏に起因するテンポの揺らぎ、(2) ギター演奏におけるビートパターンの複雑さ、(3) ロボットのファンノイズの3つの問題が存在する。従来の研究でこれら全てに取り組んだものはほとんどなかった。

以上の問題に頑健なビートトラッキングを、オンセット強調によるノイズに頑健な音響情報と演奏タイミングと相関のある手の軌道である画像情報とを、パーティクルフィルタにより統合することで達成する。音響特徴量にはエッジ強調されたスペクトログラムから検出されたオンセットと、その自己相関を用いる。画像特徴量にはオプティカルフローと平均値シフトで求めた手の座標とハフ変換で求めたギターの軸の相対距離の時系列を用いる。

実験により、本手法がテンポ変動やビートパターンの複雑さに頑健であることと、パーティクル数を実時間処理ができる程度に減らしても推定精度 (F 値) の低下は3ポイント程度に抑えられることを示す。

2. 本手法で扱う合奏の仮定と問題

2-1 ギター合奏の設定

合奏を、メロディー担当ロボット1体と伴奏担当の人のギター奏者1人で構成されると設定する。また、楽曲は4/4拍子であるとするが、後述の手の軌道モデルを変更することで、他の拍子の音楽にも対応可能である。

演奏の初めに、共演者とのタイミングやテンポを合わせるために“カウント”を行うものとする。これは主に声やギターの打撃音で行われる。またこのカウントで示されたテンポから、大きく逸脱しないと仮定する。

本手法では、楽譜は用いない。理由は、(1) ギターの楽譜にリズム譜が記載されていない場合が多いこと、(2) 本手法の目標が即興演奏であること、である。ギター奏法は大きくストロークとアルペジオに2分される。本稿では、手の動きとリズムパターンの高い相関を活かすために、ストローク奏法での演奏を仮定とする。ストローク奏法では一般的に、テンポを一定に保つために、空振りをはさみ手の振りを一定に保っている。(例：図1のパターン4)。ここで、小節の初めの手の振りを下向き、手の上下運動の周期を4分音符長と仮定する。

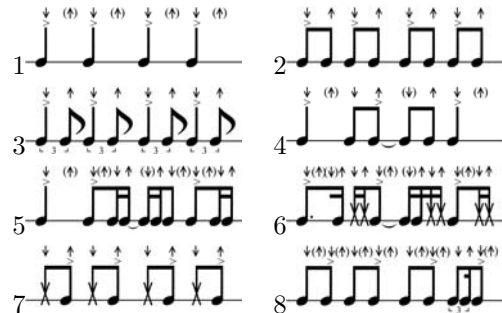


図1 代表的なギターのビートパターン。×は素早く音をミュートすることで打撃音を出す奏法(カッティング)を、>はアクセントを、矢印は手の運動方向を、括弧つきの矢印は空振りを表す。

2-2 本手法のビートトラッキングの入出力と問題

本手法では入力をロボットのマイク、カメラからの音響、画像情報とし、小節内位置とテンポの推定を行う。小節内位置は各小節の今の演奏位置で定義される。最終的に、1小節を4等分したタイミング(以下、“拍子”と書く)でロボットの演奏キューを生成する。

本手法の前述の3つの問題について述べる。

問題(1)は、合奏相手として、プロの演奏家を想定してないために発生する。誰でも参加できる合奏を目指すためには、解決すべき問題である。村田らはこの問題に対し、スペクトログラム上から取得したオンセットを相関をとる Spectro-Temporal Pattern Matching(STPM)[1]を用いて解決した。この手法は、問題(3)のファンノイズなどにも頑健である。

問題(2)はギターでは特に顕著である。この複雑さは裏拍アクセントの多さが原因である。裏拍とは、一小節を複数、特に8以上で等分したときの偶数番目にくる拍をさす。奇数番目は表拍と呼ぶ。図1に、代表的なギターのビートパターンを示す。パターン1,2はパターンの基礎となるもので、3はその3連符版である。これらのアクセントはすべて表拍に置かれる一方、それ以外は裏拍アクセントを含んでいる。パターン7,8のアクセントは裏拍にのみ置かれている。以上より、アクセント位置に対する頑健性が必要である。後藤らの、音楽を音符長毎に切り分ける階層的ビートトラッキング[2]は、ギターに関係なく複雑なビートパターンに頑健である。

問題(3)に関して、ビートトラッキングにおいて、ロボットから発生する音の影響は大きい。従来のビートトラッキング研究では、ロボットのファンノイズなどに言及したものは少なく、従ってノイズに頑健な音響特徴量を使う必要がある。

問題(1),(2)を同時に解決するために、我々は画像情報、特に手のトラッキングを扱うことで音響情報を補完する。手のトラッキングでの問題には、1)他の部分と

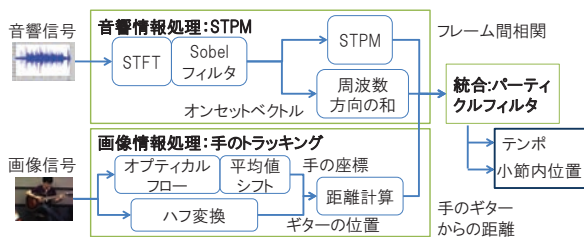


図2 システムの概要図

の混同, 2) 時間解像度の低さの2つが挙げられる. オプティカルフローにより得られる変位ベクトルには, 他の部分も動くのでそれらのノイズが含まれる. また, 従来手法に多い, 色による手のトラッキングは, 周囲環境, 特に照明によりその精度が落ちることが確認されている. 時間解像度は音響情報のそれに比べてその1/4程度しかなく, 両者の単純比較をすることは難しい.

2.3 各問題解決の手法概略

以上であげた問題の解決方法を以下に示す. 音響特徴量にはSTPMで得られたオンセットとその相関を用いる視覚特徴量には手とギターのネック(手で握る部分)との距離を用いる. 手の位置は, オプティカルフローにより手の大まかな位置をとり, 平均値シフト法で詳細な位置を得る. ギターの位置はハーフ変換によりその位置を得る. それらをパーティクルフィルタにより統合し, 解像度差の問題を含めた, 以上の問題に頑健な出力を得る. 図2に処理の概要を示す. 以下の章では, 視聴覚特徴量の検出, パーティクルフィルタの概要について述べるが, 詳しい内容は文献[3]を参照されたい.

3. 視聴覚特徴量の検出

3.1 聴覚特徴量: STPM

本稿では文献[1]で定義されるオンセットベクトルの現フレームと k フレーム前との正規化相互相関関数 $R_t(k)$ 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 を用いる. 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.

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, 正確な手の座標を得られる. カーネルには, 明度変化, つまり影や鏡面反射への頑健性をもつ, D.Miyazakiらの色空間[5]で得た色相ヒストグラムを用いる.

(3) 手の軌跡のモデル化: フレーム 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) はハーフ変換[6]で得たギターの直線パラメータで

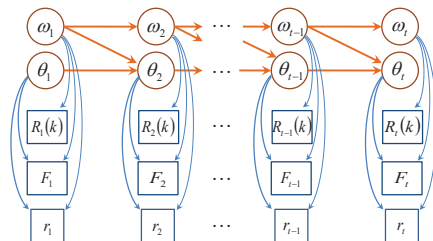


図3 グラフィカルモデル. ○と□はそれぞれ状態変数と観測変数を表す.

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小節を円周でモデル化し, フレーム 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, $0 \leq \theta_t < 2\pi$ 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 は円の角速度で, テンポに反比例する. これらと手の振幅 a_t より, 手の位置 r_t を $-a_t \sin(4\theta_t)$ とモデル化する.

4. パーティクルフィルタによる統合

パーティクルフィルタは観測から, 非線形関数や非ガウス性ノイズにおける隠れ変数の状態空間を推定する手法である. 以下, x を状態変数の集合, z を観測変数の集合とする. このとき, 状態変数の確率密度分布 $p(x_t | z_{1:t})$ は以下のように近似できる.

$$p(x_t | z_{1:t}) \approx \sum_{i=1}^I w_t^{(i)} \delta(x_t - x_t^{(i)}) \quad (1)$$

ここで, I は総パーティクル数で, $\cdot^{(i)}$ は i 個目のパーティクルの変数とする. また, $w_t^{(i)}$ は各パーティクルの重みで, その和は1となる. $\delta(x_t - x_t^{(i)})$ はディラックのデルタ関数である.

問題設定における, 観測を音響, 画像特徴量の時系列, 状態を小節内位置とテンポとしてパーティクルフィルタを適用する. フレーム t における, 状態変数を θ_t (小節内位置), ω_t (ビート間隔)とし, 観測変数を $R_t(k)$ (k フレーム前との相関), F_t (正規化オンセット), r_t (手とギターの距離)とする. また, $\theta_t^{(i)}, \omega_t^{(i)}$ は i 個目のパーティクルのそれぞれの状態変数を表す.

4.1 パーティクルの状態遷移

フレーム t におけるパーティクルの状態変数 $\theta_t^{(i)}, \omega_t^{(i)}$ はフレーム $t-1$ における観測を伴う式(2), (3)でサンプリングされる. グラフィカルモデルを図3に示す.

$$\begin{aligned} \omega_t^{(i)} &\sim q(\omega_t | \omega_{t-1}^{(i)}, R_t(\omega_t), \omega_{init}) \\ &\propto R_t(\omega_t) \times N(\omega_t | \omega_{t-1}^{(i)}, \sigma_{\omega_q}) \times N(\omega_t | \omega_{init}, \sigma_{\omega_{init}}) \end{aligned} \quad (2)$$

$$\begin{aligned} \theta_t^{(i)} &\sim q(\theta_t | r_t, F_t, \omega_{t-1}^{(i)}, \theta_{t-1}^{(i)}) \\ &= M(\theta_t | \hat{\theta}_t^{(i)}, \beta_{\theta_q}, 1) \times \text{penalty}(\theta_t^{(i)} | r_t, F_t). \end{aligned} \quad (3)$$

$N(x | \mu, \sigma)$ は変数 x , 平均 μ , 分散 σ のガウス分布の確率密度関数を表す. ω_{init} はカウントによって設定されるビート間隔を表す. $M(\theta | \mu, \beta, \tau)$ は τ 個のピークを持つように変形されたフォン・ミーゼス分布の確率密度関数である:

$$M(\theta | \mu, \beta, \tau) = \frac{\exp(\beta \cos(\tau(\theta - \mu)))}{2\pi I_0(\beta)}. \quad (4)$$

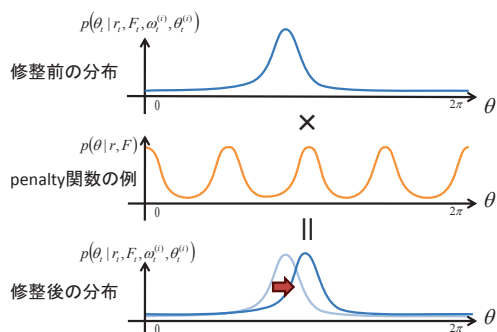


図4 *penalty* 関数による θ の分布の修整例. 例の *penalty* 関数における分布の周期は $\pi/2$ (4分音符長).

ここで, $I_0(\beta)$ は 0 次の第一種変形ベッセル関数, μ はピークの一つを表す. β は集中度を表しており, β が大きくなると, 正規分布に近づく. また, $\hat{\theta}_t^{(i)}$ は, フレーム $t-1$ での観測でのフレーム t における θ の推定であり, 以下のように定義される.

$$\hat{\theta}_t^{(i)} = \theta_{t-1}^{(i)} + \frac{b}{\omega_{t-1}^{(i)}}, \quad (5)$$

ここで, b はビート間隔を小節内位置の角速度 (テンポ) に変換するときの比例定数である.

式 (2), (3) の意味について述べる. 式 (2) はフレーム間相関 $R_t(k)$ に対して, 2 つの窓関数がかかっている. 一つは前フレームのビート間隔, もう一つはカウントで示唆されたビート間隔を中心とするガウス分布である.

式 (3) における *penalty*($\theta|r, F$) は 5 つの窓関数の積で表される. 窓関数は, フレーム t における状態が, それぞれの持つ条件を満たすときはフォン・ミーゼス分布の確率密度関数になり, そうでないときは 1 である関数とする. *penalty* 関数は, θ の分布のピークを自分のピークへと引き込むことで, 仮定やモデルに沿った分布へと修整する働きを持つ. 図 4 にその例を示す.

以下に, それぞれの窓関数が持つ条件と, そのときのそれぞれの確率密度関数を示す.

$$r_{t-1} > 0 \cap r_t < 0 \Rightarrow M(0, 2.0, 4) \quad (6)$$

$$r_{t-1} < 0 \cap r_t > 0 \Rightarrow M\left(\frac{\pi}{4}, 1.9, 4\right) \quad (7)$$

$$r_{t-1} > r_t \Rightarrow M(0, 3.0, 4) \quad (8)$$

$$r_{t-1} < r_t \Rightarrow M\left(\frac{\pi}{4}, 1.5, 4\right) \quad (9)$$

$$F_t > thresh. \Rightarrow M(0, 20.0, 8). \quad (10)$$

それぞれの β は経験的に定めた. *thresh.* は F_t の示す値がオンセットかノイズかを定める閾値である. 式 (6), (7) はストローク方向の仮定, 式 (8), (9) は手の軌道モデル, 式 (10) は手の周期が 4 分音符長, つまり 8 ビートである仮定に対応する.

4.2 重み計算

パーティクルの重みは状態変数の点推定に用いられる. 観測, 状態により逐次的に計算される.

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p(x_t^{(i)}|x_{t-1}^{(i)})p(R_t(\omega_t^{(i)}), F_t, r_t|x_t^{(i)})}{q(x_t|x_{t-1}^{(i)}, R_t(\omega_t^{(i)}), r_t, F_t, \omega_{init})}. \quad (11)$$

ここで関数 q は式 (2), (3) の積, x は状態の集合 ω, θ を表す. 分子は状態遷移関数と観測モデルの積であり, モデルに沿った値を持つパーティクルが大きな重みを持つ

つようになっている. また, 分母は提案分布と呼ばれ, サンプルされにくいパーティクルに比重をかけるようになっている.

状態遷移関数は以下の 2 式から導出される.

$$\theta_t = \hat{\theta}_t + n_\theta \quad (12)$$

$$\omega_t = \omega_{t-1} + n_\omega, \quad (13)$$

ここで, n_ω は正規分布で表現されるビート間隔の, n_θ はフォン・ミーゼス分布で表現される小節内位置のノイズである. よって, 状態遷移関数はこれらの確率密度関数の積で表現できる.

以下で, 観測モデル関数の導出を行う. $R_t(\omega)$ と r_t はそれぞれ平均が $\omega_t^{(i)}$, $-\text{asin}(4\hat{\theta}_t^{(i)})$ であるガウス分布に従う. F_t は経験的に以下のように近似できる.

$$F_t \approx f(\theta_{beat_t}, \sigma_f) \equiv N(\theta_t^{(i)}; \theta_{beat,t}, \sigma_f) * rate. + bias., \quad (14)$$

ここで, $\theta_{beat,t}$ はフレーム t における θ の最近傍拍子位置である. *rate.* は F_t の近似における最大値が 1 に *suru* パラメータで, *bias.* は $[0.35 \ 0.5]$ の一様分布である. 観測モデル関数はこれら 3 つの積で定義できる.

最後に, フレーム t における状態変数の推定のために, 重み付平均を以下のようにとる.

$$\bar{\omega}_t = \sum_{i=1}^I w_t^{(i)} \omega_t^{(i)} \quad (15)$$

$$\bar{\theta}_t = \arctan \left(\frac{\sum_{i=1}^I w_t^{(i)} \sin \theta_t^{(i)}}{\sum_{i=1}^I w_t^{(i)} \cos \theta_t^{(i)}} \right) \quad (16)$$

5. 実験および考察

ヒューマノイドである HRP-2 を用いて実験を行った. まず, 本手法と同様に STPM を用いている村田の手法 [1] との推定精度の比較を行う. また, 画像情報を加える優位性を示すために, 聴覚情報のみを入力とした手法との比較を行う. 次に, パーティクル数を変動させたときの, 計算速度や推定精度の検証を行う. 最後に, 実際のロボットとの合奏実験について述べる.

5.1 実験条件

ギター演奏の録音データは, 被験者 4 名でそれぞれテンポ 3 種 (BPM70, 90, 110), ビートパターンは図 1 に示された 8 種である. 順番は, 数字が小さいほど表拍アクセントが, 大きくなるほど裏拍アクセントが多くなるよう設計した. パーティクル数は特に指定がない場合は 200 個で, カメラの fps は約 19 である. 人とロボットの距離は約 3[m] で, ギター全体が画面に含まれる. また, 推定がビート位置誤差 ± 150 [msec], テンポ誤差 ± 10 [BPM] 以内であるときに推定成功とし, 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. ここで, それらの調和平均である, F 値を導入する:

$$F\text{-measure} = \frac{2}{1/r_{prec} + 1/r_{recall}}. \quad (17)$$

本手法を CPU: Intel Xeon W3565, メモリ: 6GB の計算機上に C++ で実装した. また, 計算速度の向上のために, スレッド分割による並列処理を行った. 以下, 計算速度はパーティクルフィルタの計算時間のみとする.

表 1 各手法ごとの F 値 (%) の比較 (番号は楽譜パターン)

番号	1	2	3	4	5	6	7	8	Ave.
統合	71.9	78.6	79.7	49.7	31.3	20.2	74.8	42.7	56.1
聴覚	42.2	45.2	44.0	28.7	24.0	18.0	42.6	40.1	35.6
村田	84.3	81.7	81.2	43.0	39.9	26.8	26.8	23.7	50.9

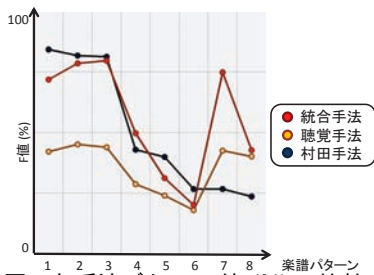


図 5 各手法ごとの F 値 (%) の比較.

表 2 各パーティクル数における計算速度と推定精度

パーティクル数	100	200	300	400
リアルタイムファクタ	0.39	0.70	1.01	1.34
F 値 (%)	54.4	56.1	57.2	57.3

5.2 精度比較

各手法の推定精度の比較結果を表 1 と図 5 に示す。以下、視聴覚統合パーティクルフィルタによるビートトラッキングを統合手法、聴覚のみパーティクルフィルタによるものを聴覚手法、村田らの手法を村田手法と表記する。村田手法はパターン番号が大きくなるにつれて F 値が下がるのに対し、統合手法は比較的高い。以上より、本手法のビートパターンへの頑健性が示された。また、統合手法と聴覚手法を比較すると、ほとんどのパターンで統合手法のほうが F 値が高く、平均では聴覚手法よりも 20.5 ポイント高い。

統合手法がパターン 5,6,8 において他のパターンよりも精度が低い理由は、これらのパターンの手の周期が 8 分音符長であり、2・1 節の仮定 (1) などで示された“手の周期は 4 分音符長である”という仮定に合っていないことが挙げられる。また、手の周期を 8 分音符長と仮定したときの実験においては、パターン 5,6,8 における結果の向上が確認されている。しかし、本手法では最終的に一小節を 4 分音符で分割したときのビート位置を出力するので、8 分音符を一周期とすることによる、出力の 8 分音符ずれが問題となり、これらのエラー対処の必要性が同様に確認されている。

また、統合手法の平均値が 56% に留まっているが、これは (1) 手の軌道モデルと真の軌道との差や、(2) 画像情報の時間解像度が低いことで、penalty 関数が十分に θ に対して機能しないことが原因として挙げられる。

5.3 パーティクル数による結果の変化

パーティクル数を変化させた時のそれぞれの計算速度と推定精度を表 2 に示す。リアルタイムファクタは、その値が 1 より小さい値のとき、システムがリアルタイムで駆動することを示す指標である。ここでは (実行時間)/(演奏時間) で計算される。表より、リアルタイムファクタはパーティクル数に比例して増加しており、300 で 1 を超えている。従って、300 より小さなパーティクル数なら、本システムはリアルタイムで動作する。

また推定精度においては、パーティクル数を上げてあまり精度が向上しない。従って、パーティクル数は



図 6 テルミン演奏ロボットとギター演奏者との合奏

200 でも十分である。

5.4 ロボットを用いたデモ評価

テルミン演奏ロボット [7] との合奏例を図 6 に示す。ロボットには HRP-2 を用いた。入力にはロボットに付属のマイクとカメラで行った。本手法により出力されたテンポとビート位置に従う、ロボットの演奏動作により合奏を実現した。また、ロボットの駆動にはわずかながら遅延が生じるため、テンポとビート位置から動作遅延を考慮したビート予測を内部で行っている。URL:<http://www.youtube.com/watch?v=fuOdhMeF3Y>

6. おわりに

本稿では、音楽ロボットのための人のギター演奏の視聴覚統合ビートトラッキングを報告した。評価実験により、要求条件である (1) テンポ変動、(2) 複雑なビートパターン、(3) ロボットノイズへの頑健さであることを示した。さらに、パーティクル数を上げて精度が大きく上がらず、またパーティクル数が 200 でリアルタイムファクタが 0.88 であるので、本ビートトラッキングの安定した推定精度とリアルタイム性の両方が獲得できる事が示された。

より協調的な合奏の実現のために、ビートトラッキングのさらなる精度向上と、エラー処理の 2 つの問題を解決する必要がある。特に、赤外線センサーを使った手のトラッキングの高性能化は重要な課題である。

謝辞 本研究の一部は科研費 (S)、新学術領域、JST-ANR BINAAHR、GCOE の支援を受けた。また、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] K. Murata *et al.*: “A beat-tracking robot for human-robot interaction and its evaluation.” In *Proc. of Humanoids*, pp. 79–84. IEEE, 2008.
- [2] M. Goto.: “An audio-based real-time beat tracking system for music with or without drum-sounds.” *J. of New Music Research*, pp. 159–171, 2001.
- [3] T. Itoharu *et al.*: “Particle-filter based audio-visual beat-tracking for music robot ensemble with human guitarist.” In *Proc. of IROS*. IEEE, 2011. to appear.
- [4] D. Comaniciu and P. Meer.: “Mean shift analysis and applications.” In *Proc. of Int'l Conf. on Computer Vision*, Vol.2, pp. 1197–1203, 2002.
- [5] D. Miyazaki *et al.*: “Polarization-based inverse rendering from a single view.” In *Proc. of Int'l Conf. on Computer Vision*, pp. 982–987, 2003.
- [6] D. H. Ballard.: “Generalizing the Hough transform to detect arbitrary shapes.” *Pattern recognition*, Vol.13, No.2, pp.111–122, 1981.
- [7] T. Mizumoto *et al.*: “Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model.” In *Proc. of IROS*, pp. 1957–1963. IEEE, 2010.