ヒューマノイドロボットのための顔位置予測を用いた人の顔追従制御

Face-to-Face Contact Method for Humanoid Robots using Face Position Prediction

岡藤　勇希[1*]　馬場　惇[2]　中西　惇也[3]　倉本　到[3]
Yuki Okafuji[1]　Jun Baba[2]　Junya Nakanishi[3]　Itaru Kuramoto[3]

[1] 立命館大学
[1] Ritsumeikan University
[2] 株式会社サイバーエージェント
[2] CyberAgent, Inc.
[3] 大阪大学
[3] Osaka University

Abstract: When humanoid robots communicate with human beings, face-to-face contact behavior plays an important role. However, while the target subject who is talked by the robot is moving, smooth communication is interfered due to a mechanical delay of the robot. This is because the simple method that the robot gazes at the observed point of the subject is used. Therefore, we proposed a face-to-face contact method using the predicted face position to reduce the robot delay. We compared two face-to-face contact methods that the robot gazes at the current and predicted face point. In the situations with various human motions, our proposed method generated the robot motion with less delay.

1　はじめに

アイコンタクトは，人間同士の会話において重要な役割を担う．人の視線によって，口語的なコミュニケーションの意味合いを強化することや [1]，人の注意を惹くことも可能なことから [2] [3]，非言語的なコミュニケーションの中でも特に重要な機能であるとされている．同様に，アイコンタクトを用いたコミュニケーション方法は，ヒューマノイドロボットにおいても重要な機能となると考えられ，様々なロボットに応用されている [4]．例えば，視線によってロボットの意図を伝えることができることから [5]，人通りが多い環境下で目標人物に話しかける際には，その人を見ながら話しかけることで，ロボットが対象としている人物を明確にする効果が期待できる．しかしながら，ロボットの視線を人の顔 (もしくは目) に向けて動かす時に，目的とする行動を取れないことが多くある．これは，話しかける対象人物が動いている時，対象者の顔の観測位置に視線を移動させるという単純な手法では，ロボットの機械的な遅れ等によって正確な視線移動を生成出来ないためである．これにより，対象者からの注意を惹きにくく，円滑な対話に支障をきたすことがある．

本研究では，人の動作を予測し，顔の予測地点に視線を向けるロボットの制御手法を提案する．対象者の現在の観測位置ではなく，予測位置に視線を向けることで，機械的な遅れ等があったとしても，人から見たロボットの遅れが軽減されることが期待される．提案手法と観測地点に視線を向ける単純な手法の比較を行い，様々な動きをする人物の顔を見る動作において，ロボットの遅延が減少することを確認する．

*連絡先：立命館大学
滋賀県草津市野路東 1-1-1
E-mail:okafuji@gst.ritsumei.ac.jp

2　提案手法

本節では，ロボットが人の顔に視線を向けるための，ロバストな顔認識・予測手法について述べる．本研究では検出の容易さを考慮して，ロボットは人間の目の位置ではなく，顔の中心に視線を向けることとした．

2.1　顔認識手法

近年，急激に発展してきた機械学習の手法を用いることで，人や顔の検出は誰しもが容易に扱えるように

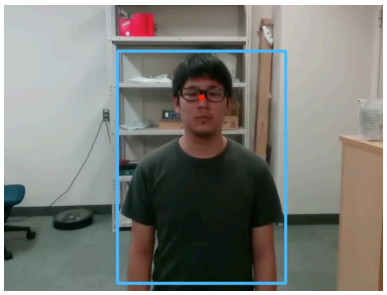

図 1: YOLOv2 の人間認識結果．顔の位置を示す赤い点は，Bounding Box の上位 20%としている．

なってきた．一例として，OpenPose [6] を用いることで，人の骨格推定を行うことができ，目や顔の中心と言った点は容易に取得することが可能である．しかしこのような手法では，高い計算コスト求められることから，安価にロボットを運用したい状況では，使用するのが困難である．また高速な機械学習の手法として，YOLOv2 [7] などが挙げられる．しかし，OpenPose などに比べて精度は劣っており，確実に人や顔を検出することはまだ困難であると言える．そこで本研究では，高精度ではないが高速な機械学習の手法と，従来からの画像特徴量を用いた手法を組み合わせた，ロバストな人の顔検出手法を提案する．また，オプティカルフローを用いて人間の行動予測をすることで，ロボット固有の機械的な遅れを減少させる．

人の認識器では，高速な機械学習の手法として YOLOv2 (図 1) を用いた人検出と顔検出，画像特徴量の一つである Haar-Like 特徴量を用いた顔検出の 3 つを用いた．また人の行動を予測する手法として，Farneback アルゴリズムを用いたオプティカルフローを用いた．しかしこれだけでは，1. 光などの影響によって人や顔の認識を上手く行えない，2. 人がその場で回転している，などの状況には対応することが出来ない．そのため，ロバストな認識を行うために，全ての手法からそれぞれ速度・顔位置を取得することとした．YOLOv2 の人認識結果から，顔の位置は Bounding Box の上位 20%の位置とした．また，YOLOv2 と Haar-like 特徴量の顔認識結果からは，後進差分を用いて画像上の速度を求めた．オプティカルフローから顔の位置を求める手法として，画像上のオプティカルフローを二次元ガウス分布に当てはめて，分布平均を顔位置とした．以上によって，4 つの顔位置 (X, Y) と画像上の速度 (u, v) を取得可能であるが，これらの観測値から真値を推定する必要がある．そこで，パーティクルフィルタを用いて，真値を推定する手法を用いた．

2.2 パーティクルフィルタ

以下に，パーティクルフィルタの概要を簡略に示す．まず，以下の状態方程式を仮定する．

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{v} \tag{1}$$

ただし，$\boldsymbol{v} = \mathcal{N}(0, \alpha^2)$ はガウス分布に従ったシステムノイズである．

この状態方程式に対して，1. 予測ステップ，2. フィルタリングステップ (尤度計算)，3. リサンプリング，の大きく 3 つのステップを踏みながら状態量を推定していく手法である．

[予測ステップ]

前時刻における状態量 x_{t-1} と，観測値 y_{t-1} が分かっていたとして，状態量を示す粒子を N 個用意する．各粒子 i が重みを有しており，この重みによって予測分布を表現する．現在の時刻における状態量の予測分布を以下のように設定する．

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{t-1}) = \sum_{i=1}^{N} \delta\left(\boldsymbol{x}_t - \boldsymbol{x}_{t|t-1}^{(i)}\right) \tag{2}$$

$$\boldsymbol{x}_{t|t-1}^{(i)} = f(\boldsymbol{x}_{t-1}^{(i)}, \boldsymbol{u}) + \boldsymbol{v}^{(i)} \tag{3}$$

[フィルタリングステップ]

次に，各粒子の尤度を計算する．全体の尤度の総和が 1 になるように，正規化してある．

$$\tilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{i=i}^{N} \omega_t^{(i)}} \tag{4}$$

$$\omega_t^{(i)} = p\left(\boldsymbol{y}_t|\boldsymbol{x}_{t|t-1}^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \sqrt{-\frac{\left(\boldsymbol{y}_t - \boldsymbol{x}_{t|t-1}^{(i)}\right)}{2\sigma^2}} \tag{5}$$

式 (5) では，一つの観測値に対してガウス分布を仮定して尤度を計算しているが，本研究では，式 (5) を複数重ねて混合ガウス分布を形成する．また，各計測値の信頼度を考慮して，尤度に対して更に重み w を考慮する．つまり，式 (5) は以下で表される．

$$\omega_t^{(i)} = \sum_{j=1}^{M} \boldsymbol{w}_j \boldsymbol{p}_j \left(\boldsymbol{y}_t|\boldsymbol{x}_{t|t-1}^{(i)}\right) \tag{6}$$

[リサンプリング]

各粒子が持つ尤度を比較して，尤度の高い粒子を抽出する．リサンプリングにはいくつかの手法があるが，本研究ではシステマティックサンプリングと呼ばれる手法を用いた．リサンプリング後に，以下で現在時刻の状態量を推定する．

$$\boldsymbol{x}_t = \tilde{\omega}_t p(\boldsymbol{x}_t|\boldsymbol{y}_{t-1}) \tag{7}$$

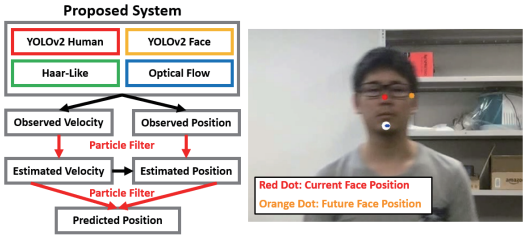

図 2: 提案システムの外観

図 3: 人間の顔に視線を向けるロボットシステム

[N ステップ先予測]

パーティクルフィルタによって求められた顔位置・速度を用いて，N ステップ先の人間の顔位置の予測を行う．N ステップ先の速度予測には，以下を用いた．

$$u_{t+1} = u_t + \alpha(u_t - u_{t-1}) \tag{8}$$

$$u_{t+N} = u_t \sum_{i=0}^{N} \alpha^i + u_{t-1} \sum_{j=1}^{N} \alpha^j \tag{9}$$

α は加速度調整のための定数値である．

次に，上記を用いて顔位置予測を以下のようにした．

$$X_{t+N} = X_t + \sum_{i=1}^{N} u_{t+i} \tag{10}$$

この N ステップ先の点を見るようにロボットに指令することによって，人間が感じる機械的な遅れを減少させることを目的とする．図 2 に，提案手法の外観と，顔位置予測の結果を示す．

3　提案手法の精度評価

顔の現在の観測位置と，提案手法によって推定した予測位置を用いて，ヒューマノイドロボットの視線をそれらの位置に向けるシステムを構築した (図 3)．このシステムにおいて，予測精度の評価と遅れ軽減の評価を行った．

3.1　予測の精度評価

カメラの前で人が様々な動きを行う動画 (30 FPS) を用意して，オフラインで顔の現在値と予測値の推定を行った．図 4 に予測の精度結果を示す．青で示しているのが観測した顔位置で，オレンジで示してるのが，提案手法を用いて予測された 10 ステップ後の顔の位置である．結果から，提案手法が将来の顔位置を推定できていることが確認できる．

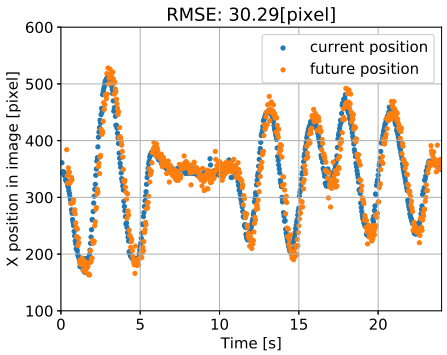

図 4: 10 ステップ後の予測精度

表 1: 動画内の人間の動き

Situation 0	上下・左右・前後の複合運動
Situation 1	左右の運動 (顔検出可能)
Situation 2	左右の運動 (顔検出不可)
Situation 3	上下・左右の複合運動
Situation 4	上下の運動 (ジャンプ)
Situation 5	その場で回転運動

3.2　遅れ軽減の評価

カメラの前で人が様々な動きを行う，6 種類の RGB-D 動画 (30 FPS) を用意した (表 1)．オフラインで現在と予測の顔位置に視線を向ける指令値を生成し，ロボットから得られる出力値を得る．提案手法では，顔位置の予測値を指令値として送り，現在の顔位置と出力値の差を評価した．遅れの評価を行う際には，入力値と出力値が合致するように，時間軸をずらすことで評価を行った．例として，図 5 は入力値と出力値の時間経過を示している．ここで図 6 に示すように，出力値のプロットを 0.47 秒早めることで，入力値と出力値の差が一番小さくなる．この時の 0.47 秒を，本研究ではロボットの遅れと定義する．

図 7 にロボットの遅延時間の結果を示す．結果より，

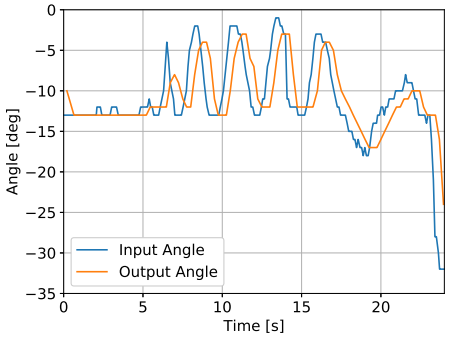

図 5: ロボットの機械的な遅れ

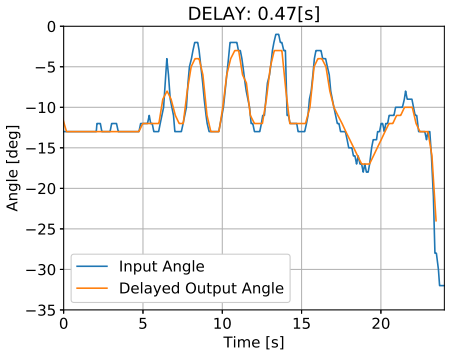

図 6: ロボットの機械的な遅れの評価

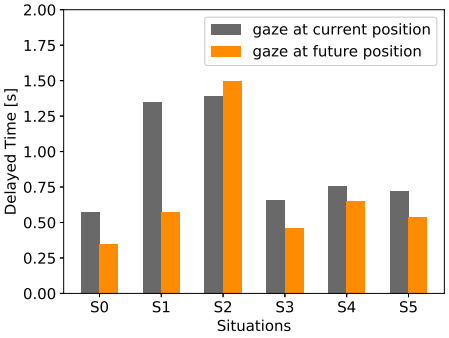

図 7: 6 種類の動画を用いたロボットの遅延時間

Situation 2 を除いて，現在の顔位置を向くときよりも，予測地点を向くほうが遅れを軽減させることが可能となった．Situation 2 は，人の顔が見えない状況で横方向に移動する動きである．そのため，顔位置の認識精度が低かったため，結果として予測の精度も低くなったことが原因として考えられる．

4 おわりに

本研究では，Deep Learning ベースの手法と画像特徴量を用いた，ハイブリッドな人間の動作予測手法を提案した．結果として，人間の行動予測は良い精度を得られ，ロボット自身の遅れも軽減することができた．しかし，全ての状況において遅延を軽減できなかったこともあり，特に顔認識手法の精度向上が望まれる結果となった．手法の改善としては，OpenPose に更に，畳み込みニューラルネットワークを導入することで，人間の行動を予測する手法もある [8]．様々な手法を検討しながら，より自然なコミュニケーションを取れるロボットの動作生成手法の構築を目指す．

参考文献

[1] S. Goldin-Meadow, “The role of gesture in communication and thinking”, *Trends in Cognitive Sciences*, vol. 3, no. 11, p. 419-429, 1999

[2] M. Argyle, M. Henderson, M. Bond, & Y. Iizuka, “Cross-cultural variations in relationship rules”, *International Journal of Psychology*, vol. 21, p. 287-315 1986

[3] C. Sherrard, “Six principles for developmental communications: Silent-film montage and adult-infant interaction”, *Language and Communication*, vol. 13, p. 163-168, 1993

[4] H. Admoni, & B. Scassellati, “Social Eye Gaze in Human-Robot Interaction: A Review”, *Journal of Human-Robot Interaction*, vol. 6, no. 1, p. 25-63, 2017

[5] T. Ono, M. Imai, & R. Nakatsu, “Reading a robot’s mind, amodel of utterance understanding based on the theory of mind mechanism”, *Advanced Robotics*, vol. 14, no. 4, p. 311-326, 2000

[6] Z. Cao, T. Simon, S. E. Wei, & Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields”, *Proceedings of CVPR*, 2017

[7] J. Redmon, & A. Farhadi, “YOLO9000: Better, Faster, Stronger”, *arXiv preprint arXiv:1612.08242*, 2016

[8] Y. Horiuchi, Y. Makino, & H. Shinoda, “Computational Foresight: Forecasting Human Body Motion in Real-time for Reducing Delays in Interactive System”, *Proceedings of ACM International Conference on Interactive Surface and Spaces*, 2017