G-19

深層学習を用いた非接触式入力システムの検討

Investigation of Contactless input System using Deep Learning

○郭知洋[1]，泉隆[2]，滕琳[2]，香取照臣[2]
*Guo Zhiyang[1], Izumi Takashi[2], Teng Lin[2], Katori Teruomi[2]

Abstract: In recent years, research and development on non-contact input interfaces are prosperous. Non-contact inputs include voice, gaze, and motion, and this research focuses on gaze and motion input. Existing systems include a gaze input system used to support people with disabilities, but while it is capable of high-precision detection, it requires expensive sensors and external devices such as infrared cameras, which are not easily available. Therefore, in this research, we are using a general WEB camera to construct an inexpensive non-contact input system with deep learning model.

１．まえがき

近年、新型コロナウイルスの流行により、衛生面でのニーズが高まり、手に触れるモノに対する警戒感が顕在化している。そのため、手を触れずに身振りだけで操作できる非接触式入力インターフェース技術への関心が一気に高まっている。非接触式入力には、音声、視線、動作などによる入力方式があり、本研究では視線と動作入力に着目している。既存の視線入力システム[1]には障害者支援に用いるシステムが挙げられるが、高精度な検出ができるに対して赤外線カメラなど高価なセンサ・外部装置が必要のため、簡単に入手できない。

そこで、本研究では一般的な WEB カメラを用いる安価な入力システムの構築を検討する。我々は深層学習を用いた画像から、ユーザの視線推定、ジェスチャ検出により操作意思を推定する手法を検討している。

２．システム構成

本研究における非接触式入力システムは Figure1 に示すように、主の処理は視線によるポインティング操作と手ジェスチャによる意思判定の 2 部分により構成されている。

まず、Web カメラで取得した画像からユーザに当たる対象を検出し、顔画像と手部分画像を抽出する。顔画像から目の位置と瞳孔の中心を検出し、瞳孔の中心座標により視線の注視点（方向）を推定する。リアルタイムで視線を追跡し、注視区域の判定によりポインティング操作を行う。また、手部分画像からジェスチャの認識による意思判定を同時に行うことで、非接触式の入力を実現する。

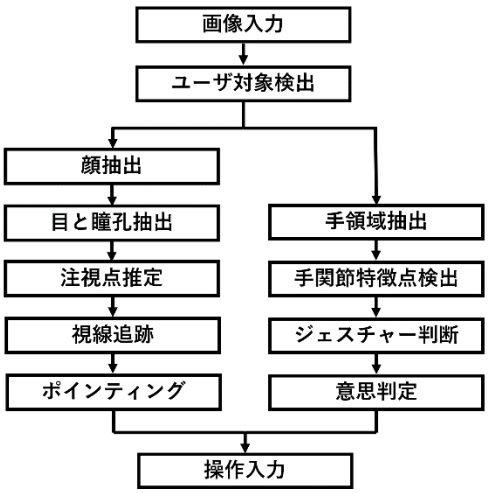

Figure 1. System flow chart.

３．視線推定

＜3・1＞顔と目の抽出　　本研究では Python 環境でオープンソースのライブラリ OpenCV と dlib[2]を利用して顔や目の検出を行う。入力画像から、顔の中の各パーツの位置関係、寸法を識別可能な Facial Landmark 情報を検出する。そして検出結果から両目位置 Landmark の座標情報を抽出する。実際に抽出した結果例を Figure 2 に示す。

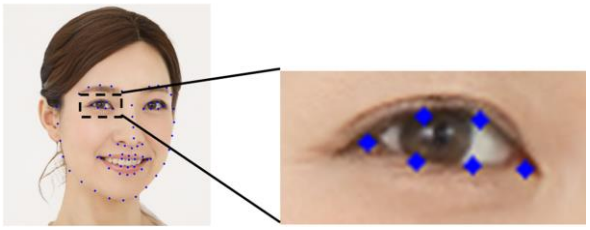

Figure 2. Detection example from Facial Landmark.

１：日大理工・院（前）・情報　２：日大理工・教員・情報

＜3・2＞瞳孔の検出　　眼球の構造により、瞳孔の部分は周りより輝度値が低いため、画像をグレースケール化してから、二値化により黒部分のみを抽出する。面積が小さいノイズを除去し、残る部分は瞳孔の候補領域とする。瞳孔候補領域に対し、カーネル平均による平滑化を施し、抽出された輪郭を Hough 変換で円形に表し、この円心座標を瞳孔の中心座標とする。その一連の検出結果の例を Figure 3 にて示す。

(a) 2 値化結果　　(b) 瞳孔候補抽出　(c)中心座標特定

Figure 3. Pupil detection process.

＜3･3＞Appearance-based 視線推定　　本研究では顔画像から直接視線を推定する Appearance-based モデル[3]を導入し、顔特徴と視線方向の関連付けを示す CNN モデル[4]を用いた。顔特徴入力のほか、瞳孔位置を追加し、出力が視線方向となる。このモデルは MPIIGaze[3]データセットを用いて、15 人の約 21 万セットの学習データにより高精度な視線推定ができ、低画質画像にも対応できる。さらに、複数ユーザの同時検出も可能である。実装結果例を Figure 4 に示す。

単独対象の検出例　　複数対象の検出例

Figure 4. Appearance-based detection example.

推定した視線方向と距離情報を用いて、画面上の注視エリア（座標）を推定する。検証実験は被験者 1 人に対し、ノード PC 画面全体を 12 個の注視エリアに分け、テスト画像に注視点を提示する。推定された注視座標にポインタを移動させ、注視点と一致するかを記録する。なお、注視点は 10 秒間隔でランダムに変更し、100 回の計測を行った。各プロセスの平均実行時間と検出精度を Table 1 に示す。

Table 1.　Experimental result.

実行項目	実行時間(ms)	精度
瞳孔位置特定	8.3	93%
注視エリア特定	27.7	87%

４．ジェスチャ検出

本研究ではユーザの視線推定とともに、手のジェスチャを検出してユーザの意思判定を行う。手のジェスチャ検出は BlazePalm[5]モデルを用いた。手を検出した後、手の Landmark である 21 個のキーポイントを検出する。キーポイントには腕から指までの全関節の座標が示されたので、手と指の状態は関節の累積角度によって計算される。本研究では、あらかじめ指の状態集合を定義し、検知された手のジェスチャによる意思判定を行う。実際に検出した結果例を Figure 5 に示す。

Figure 5. Appearance-based detection example.

５．まとめ

本研究では一般の Web カメラを用いて安価な非接触式入力システムの構築を検討した。

視線推定には CNN ベースの Appearance-based 視線検出モデルを導入し、リアルタイムの視線推定を実現し、検証実験では 8 割以上の精度を実現した。

また、ユーザ意思判定について、手のジェスチャ検出には BlazePalm モデルを導入し、事前定義操作をジェスチャにより実現できた。

今後は視線推定の精度向上及びジェスチャ集合にパターンの追加などを検討していきたい。

６．参考文献

[1] 竹原 一行：「小型高精細度カメラを用いた重度障がい者用視線入力システム」，FIT2017 ， K-011(2017)

[2] Jiangjing Lv,Xiaohu Shao,Junliang Xing,Cheng Cheng,Xi Zhou：“A Deep Regression Architecture With Two-Stage Re-Initialization for High Performance Facial Landmark Detection”，IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3317-3326

[3] ARK,Seonwook, et al：“Few-shot adaptive gaze estimation”，Proceedings of the IEEE International Conference on Computer Vision. 2019. p. 9368-9377.

[4] ZHANG, Xucong, et al：“It’s written all over yourface: Full-face appearance-based gaze estimation”，Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017. p. 51-60.

[5] Fan Zhang, Valentin Bazarevsky, et al：“MediaPipe Hands: On-device Real-time Hand Tracking”,CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, USA, 2020