Vol.2021-HCI-193 No.10
Vol.2021-EC-60 No.10
2021/6/1

スマートフォンのみを用いた周囲環境への視線入力インタフェースの検討

永井　崇大[*1]　藤田　和之[*1]　高嶋　和毅[*1]　北村　喜文[*1]

A Gaze Input Interface to Surrounding Environments Using a Single Smartphone

Takahiro Nagai[*1], Kazuyuki Fujita[*1], Kazuki Takashima[*1] and Yoshifumi Kitamura[*1]

Abstract – Gaze is a useful input modality for estimating a user's region of interest and pointing to a distant target, but the challenge is that it usually requires the installation or wearing of additional specialized devices. In this work, we propose a novel user interface that enables gaze input to the user's surrounding environment using only a widely-used smartphone. By simultaneously using both front and rear cameras and a depth sensor on the smartphone, it can track the user's head orientation while recognizing its own 3D position in a known 3D map. This allows the system to estimate the user's 3D head-gaze direction to the surrounding environment. We conducted an early performance test to evaluate the accuracy of head-gaze estimation using our interface. Based on these results, we estimated that the required target size for avoiding erroneous input is $1.64\,m \times 0.94\,m$. Finally we discussed the interactions in which the proposal is effective.

Keywords : gaze-based interaction, mobile interaction, depth sensor

1. はじめに

視線入力は，ユーザの関心領域を自然に示していることが多い点や，離れたターゲットに対しても有効である点などから入力手段としての有用性が示されている．ヒューマン・コンピュータ・インタラクション（HCI）の分野では，視線入力を用いたインタラクション手法を提案する多くの研究が行われてきた．例えば，離れたディスプレイへのポインティングとして視線を用いる手法[5]やユーザの周囲にある IoT 家電を視線により操作対象として選択し，操作する手法[6]，歩行者へのナビゲーション時に視線をユーザのコンテキストとして用いる手法[7]などがある．

しかし，視線入力には導入障壁が高いという課題がある．上記で述べた手法では，環境設置型あるいは頭部着用型のトラッカを使用している．これらのトラッカは視線トラッキングの精度は優れているものの，特定用途でのみ使われる専用のデバイスであり，一般にはあまり普及していない．また，環境設置型のトラッカには限られた方向や範囲内でしか視線入力を利用できないという問題があり，頭部着用型のトラッカには長時間着用する場合にユーザへの負担が大きく，ユーザの視野を妨げうるという問題がある．

これに対し，本研究では，広く普及したデバイスであるスマートフォンを用いて視線入力を実現することを考える．最近では，ほとんどのスマートフォンがその前面・背面にカメラを搭載しているため，ユーザがスマートフォンを把持するだけで，ユーザの頭部や周囲環境の情報を同時に取得することが可能だと考えられる．類似のアイデアを用いた手法として，背面カメラから AR マーカーを読み取り，マーカーの位置を基準とする相対的なユーザの手・頭部・胴体のトラッキングを行った手法[2]やユーザの頭部（視線）方向のトラッキングと背面カメラの映像の物体認識を組み合わせることで注視対象物を推定し，その情報を音声アシスタントへの追加入力とした手法[1]がある．スマートフォン以外では 360 度カメラを用いた手法[10]もある．しかし，このスマートフォンを用いるアイデアで，ユーザの周囲環境への視線入力を検討した研究はまだない．また，最近のスマートフォンには背面に LiDAR スキャナや ToF センサのような高精度な深度センサが搭載されはじめている [1]が，上記の研究では深度センサの活用は検討されていない．カメラと深度センサを組み合わせることで，視線を向ける対象となる周囲環境をより正確かつ頑強に把握できるようになると考えた．

そこで本研究では，ユーザが把持するスマートフォンのみを用い，ユーザの視線トラッキングとユーザの周囲環境の把握を同時に行うことで，周囲環境への視線入力を可能にする手法を提案する．提案手法では，スマートフォンの前面カメラを用いてユーザの頭部（視線）方向を計測し，背面カメラと深度センサを用

*1: 東北大学 電気通信研究所
*1: Research Institute of Electrical Communication, Tohoku University

1: https://www.apple.com/iphone-12-pro/
https://xperia.sony.jp/xperia/xperia1m2/

Vol.2021-HCI-193 No.10
Vol.2021-EC-60 No.10
2021/6/1

いて周囲環境の 3D マップの復元とスマートフォンの自己位置推定を行う．これにより周囲環境の 3D マップ内でのユーザの視線位置が推定できる．この手法が確立すれば，家のリビングでの周囲の IoT 家電に対する操作や，美術館での注目対象に関するスマートフォンへのリアルタイムな情報提示など，幅広い応用が期待できる．本稿では，まず提案手法の実装について述べる．次に提案手法による視線入力の精度評価を目的とした実験を行い，誤入力を避けるために必要なターゲットサイズを考察した．また，それらを踏まえて，提案手法が有効なインタラクションについて議論した．

2. 提案手法

2.1 概要

提案手法の概要を図 1 に示す．提案手法では，スマートフォンのみを用いて周囲環境への視線入力を可能にするために，ユーザの視線トラッキングとユーザの周囲環境の把握を同時に行う．また本手法では，ユーザがスマートフォンを自然に把持するだけで視線入力を可能とすることを目指す．すなわち，視線入力のためにユーザが意識的にスマートフォンの位置や姿勢を変える必要のない手法とする．これを実現するためには，以下の 2 つが必要条件となる．

- 前面カメラがユーザの顔を捉えていること
- 事前に周囲環境の 3D マップを取得済みであること

1 つ目の条件は，スマートフォンの画面が見えるように把持している場合であれば満たされることが報告されている [2]．2 つ目の条件は，背面カメラに映っていない物に対しても位置推定を行い，視線の入力対象とするために必要となる．

これらの条件を満たすことで提案手法は，スマートフォンを把持するだけで，広大な環境であってもユーザはあらゆる位置からあらゆる方向への視線入力が可能となる．また，3D マップは共有することができるため，複数人同時での視線入力も可能となる．

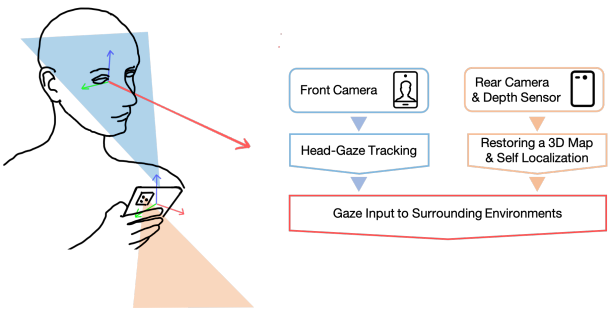

図 1: 提案手法の概観
Fig. 1 Overview of Proposal.

2.2 実装

本手法のプロトタイプに用いるスマートフォンとして，前面・背面カメラの同時使用が可能で，深度センサが搭載されており，頑強な頭部トラッキング API が含まれているフレームワーク ARKit [2] を標準で使用できることから iPhone 12 Pro を採用した．以下では，ユーザの視線トラッキングとユーザの周囲環境の把握のそれぞれの実装について説明する．

2.2.1 ユーザの視線トラッキング

スマートフォンの前面に搭載されたカメラから，顔の 3 次元的な形状を手がかりとしてユーザの頭部の位置･向き（6DoF）をトラッキングする．そして，頭部の正面方向をユーザの視線方向とすることで視線トラッキングを実現する．この頭部トラッキングは ARKit [2] の API を使用して実装した．

なお，頭部方向以外にも，眼球方向を考慮した視線トラッキングも検討したが，我々の予備検討において精度が十分でなかった点と，AR・VR 上でのオブジェクト選択に関する研究のユーザスタディで眼球視線よりも頭部視線の方が好まれた [3], [8] 点から，頭部方向を用いる方法を採用することとした．

2.2.2 ユーザの周囲環境の把握

スマートフォンの背面に搭載されたカメラと深度センサから得られる情報をもとに，取得済みの周囲環境の 3D マップの復元とスマートフォンの自己位置推定を行うことで，ユーザの周囲環境の把握，すなわちユーザと視線入力の対象となる物の位置関係の把握を実現する．3D マップの復元と自己位置推定は，背面カメラから得られる画像の特徴と深度センサから得られる深度マップの特徴を既知の 3D マップの情報と照合することで行う．これらは ARKit [2] の API を使用して実装した．復元する 3D マップは iPhone 12 Pro の深度センサを用いて容易に作成することができる．

3. 性能評価

3.1 実験概要

提案手法による周囲環境への視線入力の精度を評価することを目的とした評価実験を行った．基本的な実験内容と計画は，Mayer らの実施した実験 [1] を踏襲している．実験要因は，壁との距離（3 通り：$1\,m, 2\,m, 4\,m$）とターゲットの水平位置（5 通り），ターゲットの垂直位置（3 通り）の 3 つである．参加者には，壁から 3 通りのうちいずれかの距離だけ離れた位置に立ち，スマートフォンを把持してもらう．そして，壁の上に 5×3 で $0.8\,m$ 間隔のグリッド状に配置された計 15 個のターゲットをランダムな順番で注視してもらい，その際にシステムが推定した注視位置との誤差を計測

2 https://developer.apple.com/augmented-reality/arkit/

Vol.2021-HCI-193 No.10
Vol.2021-EC-60 No.10
2021/6/1

した．

3.2 参加者

参加者は 21 歳から 25 歳の男性 3 名，女性 1 名であった．全員がスマートフォンを日常的に使用しており，矯正視力が 1.0 以上であった．3 名が実験時に眼鏡を着用していた．

3.3 実験環境と手順

実験環境の概観を図 2 に示す．壁にターゲットの目印を付け，ターゲットの注視を妨げない位置にデスクワゴンを設置した．このワゴンを設置した目的は，背面カメラや深度センサへの手がかり（特徴）を増やし，システムのトラッキングをより正確にするためであった．

参加者には初めに，提案手法と実験に関する説明を受けてもらい，その後簡単な練習を行ってもらった．参加者への教示として，スマートフォンの画面が見えるように自然に把持し続けるよう伝えた．練習の後，各距離条件でのターゲットの注視タスクに移った．なお，予備調査において，システムによる推定注視位置には個人差が見られ，同じ水平（垂直）位置のターゲット群では，水平（垂直）方向に同程度の誤差がある傾向を確認した．このため，各距離条件を試行する前に，キャリブレーションフェーズを設けた．具体的には，各参加者にターゲットの水平位置・垂直位置の全通りが含まれるように選ばれた 5 個のターゲットを注視してもらうことで，その水平（垂直）位置における参加者特有の誤差を取得し，これをシステムのキャリブレーションに用いた．毎試行の開始時に，参加者の把持するスマートフォンの画面にターゲット位置を示す図を表示した．参加者は指示されたターゲットに視線を向け，その状態でスマートフォンの画面をタップすることで試行を完了させた．試行中，参加者には視線方向に関するシステムによる視覚フィードバックを与えなかった．指示されるターゲットの順序はランダムであった．各距離条件で 1 つのターゲット位置につき 3 回ずつの試行を実施してもらった．したがって，得られたデータは計 540 試行分であった（4 名 × 15 ターゲット × 3 回 × 3 距離条件）．

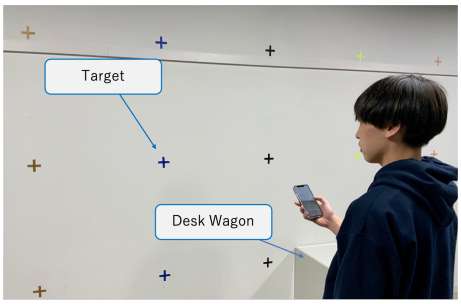

図 2: 実験環境
Fig. 2 Apparatus.

3.4 結果

まず，得られたデータについて外れ値のフィルタリングを行った．フィルタリングの基準は Mayer らの基準 [1] に倣い，絶対誤差の平均値と標準偏差を算出した後，平均値 + 標準偏差 ×3 を超える絶対誤差の試行を除外した．これにより，540 試行のうち 10 試行が除外された．

フィルタリング後の全試行の平均絶対誤差は 0.29 m（$SD = 0.19\,m$）であった．x 方向の平均絶対誤差は 0.22 m（$SD = 0.20\,m$），y 方向の平均絶対誤差は 0.14 m（$SD = 0.12\,m$）であった．図 3 に，(a) 各距離，(b) 各ターゲット水平位置，(c) 各ターゲット垂直位置における平均絶対誤差を示す（エラーバーは標準誤差を表す）．

Shapiro Wilk 検定から絶対誤差には正規性が認められなかったため，Aligned Rank Transform [9] によるデータの変換を行ったうえで，壁との距離，ターゲットの水平位置，ターゲットの垂直位置の 3 要因について多元配置分散分析を行った．その結果，絶対誤差への主効果は壁との距離（$F(2, 485) = 18.15,\ p < .001$）とターゲットの水平位置（$F(4, 485) = 5.57,\ p < .001$）で確認された．1 次交互作用は壁との距離とターゲットの水平位置間（$F(8, 485) = 3.25,\ p < .01$）で確認された．2 次交互作用は確認されなかった．

3.5 考察

本実験は，類似の視線入力手法を実現している Mayer ら [1] の実験計画を踏襲したものであったが，平均絶対誤差は Mayer らの結果の 41% 程度だった．また，Mayer らの結果では，絶対誤差において，ターゲットの垂直位置の主効果が確認されたが，本実験では確認されなかった．同様に，Mayer らの結果で確認されたターゲットの垂直位置と他の 2 要因間での交互作用も本実験では確認されなかった．これは，事前に簡単なキャリブレーションを行ったことと，Mayer らが用いていなかった深度センサの使用により推定精度が向上したことが主な要因だと考えられる．

また，各ターゲット位置（15 通り）で，水平（x）方向と垂直（y）方向のそれぞれの誤差について Shapiro Wilk の正規性検定を行ったところ，x 方向は 12 個，y 方向は 13 個のターゲット位置で正規性が認められた．そこで，Kytö らや Feit らの方法 [3], [4] に倣い，(1) 式から各ターゲット位置でエラー入力を 5%以下に抑えるための，必要ターゲットサイズを算出した．結果を図 3 の (d) に示す．

$$S_{w/h} = 2(O_{x/y} + 2\sigma_{x/y}) \tag{1}$$

なお，この必要ターゲットサイズは本来であれば，各

Vol.2021-HCI-193 No.10
Vol.2021-EC-60 No.10
2021/6/1

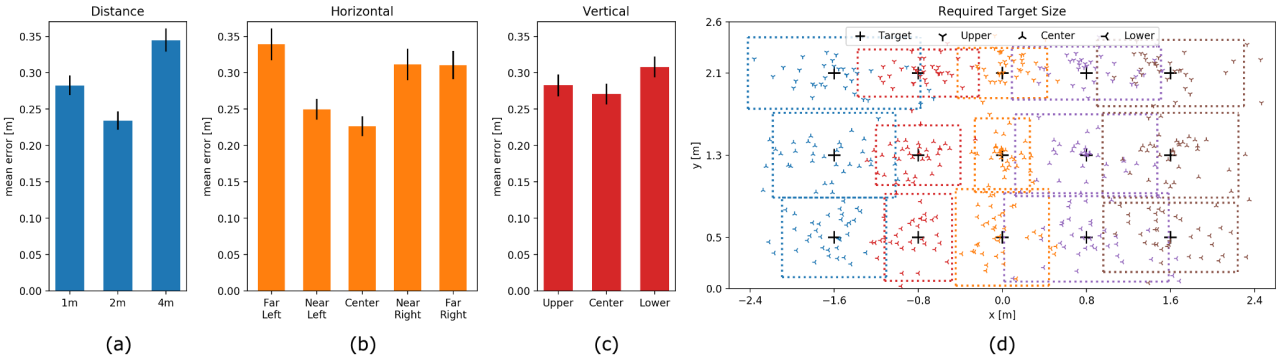

図 3: 実験結果. (a) 各距離, (b) 各ターゲット水平位置, (c) 各ターゲット垂直位置における平均絶対誤差 (d) 各ターゲット位置の必要ターゲットサイズ

Fig. 3 Results; (a) Mean error vs. target distance, (b) horizontal position, and (c) vertical position; (d) Required target size for each target position.

距離条件で独立に算出すべきだが，今回の予備実験では十分な標本サイズが得られなかったため（$n \leqq 12$），全ての距離条件を含めて算出したものであることに注意されたい．算出の結果，必要ターゲットサイズの横幅は上段最左列のターゲットで最長 $1.64\,m$，縦幅は下段中央列のターゲットで最長 $0.94\,m$ であった．

4. 議論

必要ターゲットサイズの横幅と縦幅より，本実験の簡単なキャリブレーション方法と，フィードバックを与えないという条件であれば，視線入力のターゲットは $1.64\,m \times 0.94\,m$ の矩形より大きなサイズに設定すれば十分な精度で入力が可能であると言える．なお，ここでいうターゲットとは視線がその内側に入ることで何らかのインタラクションを発生させる領域を指す．

このことから，本手法を実環境で用いる場合，この必要ターゲットサイズよりも広い間隔で並んだオブジェクトであれば，高い精度で指示可能であると考えられる．すなわち，ユーザの周囲にある IoT 家電や PC のようなデバイスに視線を向けるだけで指示する場面や，美術館のような屋内空間において並べられている展示物を注視することにより指示する場面において十分に利用可能だと考えられる．

また，この必要ターゲットサイズはトラッキング性能の向上や，より高度なキャリブレーションの実装で，さらに小さくできると考えられる．加えて，注視位置に関する視覚や触覚のフィードバックを追加することでも精度の向上が見込めるため，今後は大画面環境に対するポインティング手法としての応用も検討したい．

5. おわりに

本稿では，スマートフォンのみを用いて周囲環境への視線入力を可能にする手法を提案した．性能評価の結果から，提案手法により推定された入力位置の誤差は既存手法よりも 59%小さいことがわかり，$1.64\,m \times 0.94\,m$ 以上のターゲットサイズであれば高い精度での入力が可能であることがわかった．この結果をもとに，今後は提案手法のアプリケーションの実装とその評価を実施する予定である．

参考文献

[1] Mayer, S., Harrison, C., et al.: Enhancing Mobile Voice Assistants with WorldGaze; CHI '20, pp.1-10 (2020).
[2] Babic, T., Haller, M., et al.: Simo: Interactions with distant displays by smartphones with simultaneous face and world tracking; CHI EA '20, pp. 1-12 (2020).
[3] Kytö, M., Billinghurst, M., et al.: Pinpointing: Precise head-and eye-based target selection for augmented reality; CHI '18, pp. 1-14 (2018).
[4] Feit, A. M., Morris, M. R., et al.: Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design; CHI '17, pp. 1118-1130 (2017).
[5] Stellmach, S., Dachselt, R.: Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets; CHI '13, pp. 285-294 (2013).
[6] Sun, Y., Sarma, S., et al: Magichand: Interact with iot devices in augmented reality environment; IEEE VR, pp. 1738-1743 (2019).
[7] Giannopoulos, I., Raubal, M., et al.: GazeNav: gaze-based pedestrian navigation; MobileHCI '15, pp. 337-346 (2015).
[8] Pathmanathan, N., Sedlmair, M., et al.: Eye vs. Head: Comparing Gaze Methods for Interaction in Augmented Reality; ACM ETRA '20 Short Papers, pp. 1-5 (2020).
[9] Wobbrock, J. O., Higgins, J. J., et al.: The aligned rank transform for nonparametric factorial analyses using only anova procedures; CHI' 11, pp. 143-146 (2011).
[10] 木ノ原, 小野, 他: ORANGE: 360 度カメラと画像認識技術によるユニバーサルリモコン; WISS 2020 予稿集, pp. 79-84 (2020).