第 25 回日本バーチャルリアリティ学会大会論文集（2020 年 9 月）

VR ゴーグルに内蔵されたアイトラッカを用いた視線ジェスチャの識別

張翔 [1)]，坂本大介 [2)]，杉浦裕太 [1)]

1) 慶應義塾大学 理工学部 情報工学科（〒 223-8522 神奈川県横浜市港北区日吉 3-14-1）

2) 北海道大学 大学院情報科学研究院（〒 060-0814 札幌市北区北 14 条西 9 丁目）

概要: 本研究では VR ゴーグルに内蔵されたアイトラッカを用いて視線ジェスチャを識別する手法について提案をする．ユーザがトリガなしでジェスチャを入力できるように，ユーザがジェスチャをしているか，どのジェスチャを行っているか，という 2 つの識別にそれぞれ機械学習で学習モデルを作成し，それらを組み合わせてジェスチャ識別するようにした．実際に VR コンテンツをプレイしながらジェスチャをして，識別精度検証を行った．

キーワード： ジェスチャ認識，アイトラッキング，インタラクション，ユーザインターフェース

1. はじめに

近年，アイトラッカが内蔵された VR ゴーグルが発売されている．VR ゴーグルにアイトラッカを内蔵することで，目の動きや瞬きを VR 空間のアバタに反映することや，ユーザの視線に基づいて解像度を最適化してレンダリングにかかる負担を軽減できるほか，VR コンテンツ内で視線による入力や選択ができる．現在，VR 環境における入力方法としては手にコントローラを持つのが一般的であるが，内蔵アイトラッカによって視線入力が行えるようになると，VR ゴーグルを装着した上でさらにコントローラを手に持つという手間がなくなったり，コントローラを手に保つ場合でもさらに入力方法のバリエーションを増やすことができる．

アイトラッカによる視線入力では，アイトラッカからの視線データそのものでは，ユーザが入力を行おうとしているのかどうかが分からない Midas Touch 問題と呼ばれる問題が存在する [1]．Midas Touch 問題を解決するために，画面上に表示されたメニュー内の項目を一定時間見つめることで選択を行う方法 [1] や，うなずきや頭の回転をトリガにする方法 [2] が考えられている．一方で，これらの方法では入力が完了するまでに時間がかかったり，視線入力以外の方法と組み合わせて入力を行うのでユーザにとって負担となったりするという課題がある．

そこで，本研究では視線入力の方法の 1 つである視線ジェスチャについて，ユーザによるトリガを必要とせずに入力を行う方法を提案する．VR ゴーグル内蔵のアイトラッカから取得した両眼を瞳孔の位置および瞼の開き具合のデータを使用して，畳み込みニューラルネットワーク (CNN) を用いてジェスチャを意図しているかどうか及びどのジェスチャを行っているか，についてそれぞれ学習モデルを作成し，2 つの学習モデルを組み合わせて識別を行う．実際に VR コンテンツをプレイしながらジェスチャを行った結果 70 %の精度でジェスチャを識別することができた．

2. 関連研究

アイトラッカを用いたコンピュータへの入力では Jacob ら [1] や，Vertegaal ら [3] が一定時間画面上の選択したいもの見つめることで選択をおこなう方法について検討を行っている．また，一定時間見つめて選択を行う方法以外での視線による入力方法としては，Bee ら [4] が視界をいくつかの空間に分け，それぞれの空間にアルファベットを割り当て，視線をその空間の間で移動させることによって文字入力を行う方法を提案しているほか，Esteves ら [5] がスマートウォッチ上で円運動をする点に視線を追従させることで入力を行う方法を提案している．

視線ジェスチャによる入力ついては Vaitukaitis ら [6] が，モバイル端末において一定時間見つめることで選択する方法とジェスチャによる入力方法の比較をしている．また，Drewes ら [7] はスマートフォンのカメラを用いて 4 つの視線ジェスチャを推定する方法について提案している．

また，VR ゴーグルを用いた視線入力に関しては Piumsomboon ら [2] が VR 空間内で，オブジェクトを視線で選択をする方法について，一定時間見つめる，うなずきや頭の回転をトリガにする，オブジェクトを少しずつ移動させ視線の追従を計測する，という 3 つの方法の比較を行っている．

本研究では、VR ゴーグルとその内蔵アイトラッカを用いて視線ジェスチャによって入力を行う方法を提案する．視線ジェスチャによる方法では音量の操作といった VR 空間内のオブジェクトとは紐付かないような入力もおこなうことができるメリットがある．

3. 実装

本研究では，図 1 に示す 4 種類のジェスチャについて，ジェスチャをしているかどうか及びジェスチャの種類をそれぞれ識別するモデルを機械学習で作成し，2 つのモデルを組み合わせてジェスチャを識別するプログラムを作成した．

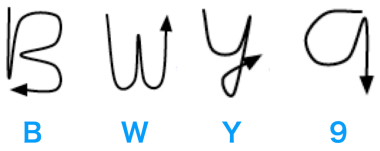

図 1: 識別する 4 種類のジェスチャ

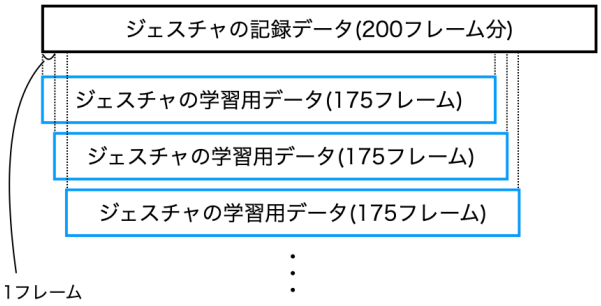

図 2: ジェスチャの記録データと学習用データの関係

この 4 種類のジェスチャについては PDA 端末の Palm で用いられた文字入力用のジェスチャセットから利用した．また，VR ゴーグル及びアイトラッカは HTC VIVE Pro Eye を使用した．

3.1 学習データの用意

CNN で学習をする際に使用するデータとして，4 種類のジェスチャを行った際の視線データ及び VR コンテンツをプレイした際の視線データを取得した．視線データを収集するプログラムは Unity を用いて作成され，両眼の瞳孔の位置の 2 次元データ及び瞼の開き具合を 0.02 秒間隔 (50Hz) で記録するようにした．

4 種類のジェスチャの視線データの記録は VR コントローラのボタンを押してから 4 秒間 (200 フレーム分) 記録するようにし，それぞれのジェスチャを 20 回分記録した．

VR コンテンツをプレイした際の視線データの記録では，Valve 社が提供する VR ゲーム「The Lab」に含まれる 8 つのミニゲームをそれぞれ 6 分以上プレイして記録した．

3.2 ジェスチャをしているかどうかの学習

ジェスチャをしているかどうかを識別するモデルは 4 種類のジェスチャの視線データ及び，VR コンテンツをプレイした際の自然な視線データをそれぞれ 1 つのクラスした合計 2 クラスでの CNN による機械学習で作成した．

ジェスチャの学習データの取得：用意した 4 種類× 20 回=80 個のジェスチャの視線データの 75%(60 個) を学習データ，25%(20 個) をテストデータにランダムに分けた．さらに図 2 のように 200 フレームで記録した視線データ 175 フレーム分のデータを 1 フレームずつずらしながらスライスして，1 個の記録したジェスチャデータから 25 個のスライスしたデータを作成した．この方法でスライスされた学習データは 60 × 25=1500 個，テストデータは 20 × 25=500 個となった．

自然な視線移動の学習データの取得：記録した視線データを 200 フレームごとに範囲が被らないように分割し，ランダムに 60 個を学習データ，20 個をテストデータとした．さらに，ジェスチャ中の学習データ取得と同様にそれぞれのデータを 1 フレームずつずらしながら 175 フレーム分のデータにスライスした．

このように作成されたデータを用いて CNN で学習を行ったところ，テストデータの識別精度は 95%となった．

3.3 ジェスチャの種類の学習

ジェスチャの種類の学習では 4 つのジェスチャの記録した視線データについて，それぞれのジェスチャを 1 つのクラスとして CNN で機械学習をした．3.2 節と同様に 4 種類 × 20 回=80 個のジェスチャの視線データの 75%(60 個) を学習データ，25%(20 個) をテストデータにジェスチャの種類が偏らないようにランダムに分けた後，それぞれのデータを 1 フレームずつずらしながら 175 フレーム分のデータにスライスした．

このように作成されたデータを用いて CNN で学習を行ったところ，学習後のモデルによるテストデータの識別精度は 96%であった．

3.4 ジェスチャ識別プログラムの作成

3.2 節及び 3.3 節で作成したモデルを組み合わせて VR コンテンツをプレイしながら視線ジェスチャを行った際の視線データからジェスチャを行っている部分とそのジェスチャの種類を識別するプログラムを作成した．プログラムでは学習する際と同様に視線データを 1 フレームずつずらしながら 175 フレーム分スライスし，そのすべてのスライスしたデータをジェスチャをしているかどうかを識別する学習モデルで識別を行う．識別を行った結果，ジェスチャをしていると識別されたデータのみジェスチャの種類を識別する学習モデルで識別を行う．

さらにこのようにして得られた時系列の識別結果に対して，1 回のジェスチャが 2 回以上のジェスチャと識別されることを防ぐために，直前 175 個のスライスされたデータの識別結果のうち 45 個以上のデータがジェスチャであると識別された場合はその中で最も識別頻度が高かったジェスチャを，それ以外の場合はジェスチャをしていないとするフィルタをかけ，1 つのジェスチャの区間では識別結果が 1 つだけになるようにした．

4. 実験

3.4 節で作成した識別プログラムの精度を検証する実験を行った．

4.1 概要

実験では，VR ゲーム「The Lab」に含まれるミニゲームの 1 つの「Postcards」をプレイしながら視線ジェスチャを行い，視線データを記録した．視線ジェスチャは 45 秒ごとに音が鳴るタイマーを用意し，音が鳴るごとに 1 回ジェスチャを行うようにした．4 種類のジェスチャをそれぞれ 5 回ずつ，合計 20 回ジェスチャを行った．

ジェスチャの真値を記録するために，ジェスチャをする際はキーボードの特定のキーを押し続けるようにし，その

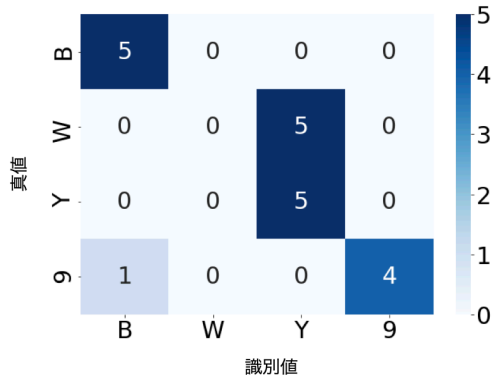

図 3: 実験結果

キーが押されているかを視線データとともに記録した．さらに，記録終了後にジェスチャを行っていると記録された部分には手動でジェスチャの種類を表す値を付与した．

このようにして得られた視線データを識別プログラムで識別を行い，真値と比較することで精度を検証した．

なお，新型コロナウイルスの影響で実験の被験者は著者のみである．

4.2 結果

ジェスチャを行っているどうかの識別に関しては，真値がジェスチャであるすべての区間内にそれぞれ 1 つだけジェスチャを行っていると識別された区間が含まれ，逆に真値がジェスチャをではない区間でジェスチャを行っていると識別された区間はなく，識別精度は 100%となった．

ジェスチャを行っていると識別された部分のみで行われたジェスチャの種類の識別に関しては全体の識別精度の平均は 70%であった．それぞれのジェスチャの識別精度の混合行列は図 3 の通りである．特に真値が W のジェスチャはすべて Y のジェスチャであると識別されるという結果となった．

なお，識別ではいずれのジェスチャでも一定区間連続してジェスチャをしていると識別され，3.4 節で述べた通りに識別結果を 1 つにするためにフィルタをかける処理が行われた．

5. 議論

5.1 考察

実験では W のジェスチャが Y のジェスチャと識別されたのが，全体の識別精度を大きく押し下げる要因となった．どちらのジェスチャも最初の部分では上から下へ視線を移動させた後再び上に視線を移動させるという流れとなっていて，ジェスチャの動きが似ているためであると考えられる．そのため，動きが似ているジェスチャがないようにジェスチャの組み合わせを変えることで，さらに精度を上げることができるのではないかと考えられる．

5.2 制約，今後の課題

本研究では 4 種類の視線ジェスチャの識別を行ったが，VR コントローラを視線ジェスチャで置き換えることを考えた場合はより多くのジェスチャが必要になる．そのためには多くのジェスチャのサンプルデータを用意して，その中でジェスチャ同士の動きがかぶっておらず，識別精度が良好な組み合わせを見つける必要があると考えられる．

また，視線のジェスチャは 4 秒間 (200 フレーム) で記録を行い，識別は 3.5 秒間 (175 フレーム) 分のデータを用いた．さらに，連続してジェスチャであると識別される区間ではその区間でデータを 1 つにまとめて識別結果を出力する処理を行う必要があるため，リアルタイム識別を行った場合，ジェスチャを始めてから結果が出力されるまで 3.5 秒以上かかってしまう．識別までの時間を短くするためには，短い時間でジェスチャを完了できるようにジェスチャの動きをよりシンプルにする必要があるが，よりシンプルなジェスチャではジェスチャかどうかの識別やジェスチャの種類の識別の精度が低下する可能性がある．

今回の研究ではジェスチャを予め定義したが，それらのジェスチャがユーザにとって行いやすいのかどうかについては検証を行っていない．日常の生活において視線ジェスチャを行う場面はなく，ユーザは視線ジェスチャに慣れていないため，ユーザにとって行いやすいジェスチャを設計する必要がある．また，ジェスチャの設計だけではなく，どのジェスチャをどの操作に割り当てるのが直感的であるかを検証することも今後の課題の 1 つである．

6. 結論

本研究では VR ゴーグル内蔵のアイトラッカを用いて視線ジェスチャを識別する方法を提案した．ジェスチャをしているかどうか，ジェスチャの種類をそれぞれ CNN で学習をして識別を行い，精度を検証したところ平均 70%でジェスチャ識別ができた．

謝辞 本研究は，JST AIP-PRISM 課題番号 JPMJCR18Y2 の支援を受けたものです．

参考文献

[1] Robert J. K. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. ACM Transactions on Information Systems (TOIS), Vol. 9, No. 2, 152-160, 1991.

[2] Thammathip Piumsomboon, Gun Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. 2017 IEEE Symposium on 3D User Interfaces (3DUI), 36-39, 2017.

[3] Roel Vertegaal. A Fitts Law comparison of eye tracking and manual input in the selection of visual targets. Proceedings of the 10th international conference on Multimodal interfaces (IMCI '08), ACM, 241-248, 2008.

[4] Nikolaus Bee, and Elisabeth André. Writing with Your Eye: A Dwell Time Free Writing System

Adapted to the Nature of Human Eye Gaze. Perception in Multimodal Dialogue Systems, Vol. 5078, 111-122, 2008. 457-466, 2015.

[5] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements. Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15), 457-466, 2015.

[6] Vytautas Vaitukaitis, and Andreas Bulling. Eye gesture recognition on portable devices. Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12), 711-714, 2012.

[7] Heiko Drewes, Alexander De Luca, and Albrecht Schmidt. Eye-Gaze Interaction for Mobile Phones. Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology (Mobility '07), ACM, 364-371, 2007.