視線を用いたターゲット選択に基づく動画内物体の注釈情報提示システム

島里 恵多* 河野 恭之*

概要. 本研究では，動画内の物体を視線で選択し，その注釈情報を提示するシステムを提案する．動画を視聴する際，動画に映る物体やキャラクタについての注釈情報を得ることで動画内容の理解が深まる．しかし近年の動画配信サービスにはこのような動画内のターゲットを選択し，その注釈情報を提示するような機能は少ない．そこで本研究では，動画内の物体を選択しその注釈情報を提示するシステムを提案する．動画内の物体の選択は，マウスやタッチ操作よりも素早いカーソルの移動と選択が可能な視線を用いて行う．まず前処理として，動画内の物体を検出・追跡し，各追跡物体に対して注釈情報を登録する．本処理ではユーザの視線を検出し，ユーザが前処理で注釈情報を登録したターゲットに対して視線を用いて選択した場合その注釈情報を提示する．

1 はじめに

本研究では，視線を用いて動画内に映る物体を選択し，その物体の注釈情報を提示するシステムを開発する．動画を視聴する際，動画に映る物体やキャラクタについての注釈情報を得ることで動画内容の理解が深まる．例えば，近年 COVID-19 の影響で普及されつつあるオンライン水族館や動物園のような生物の動画を視聴する際，その動物の種名や生態を知ることでその動画内容の理解が深まる．しかし，近年の動画配信サービスにはこのような動画内のターゲットを選択し，その注釈情報を提示するような機能は少ない．そこで本研究では，動画内に登場する物体を選択し，その注釈情報を提示するシステムの開発を目指す．

本研究では，動画内に映る物体を選択し，その物体の注釈情報を提示するシステムを提案する．本システムの実装においてユーザが選択するターゲットは 動画内の物体である．動画内には選択したい物体が短いシーンしか映らない場合やターゲットの移動が速い場合がある．そのため本システムではより素早いカーソルの移動やターゲット選択がユーザに求められる．そこで本研究では，マウスやタッチ操作より素早いターゲット選択が可能であると報告[1]のある視線のみを用いて動画内の物体をユーザの視線により選択し，その注釈情報を提示するシステムの開発を目指す．本研究の概要を図 1 に示す．

2 手法

本研究では，動画内の物体を視線で選択し，その注釈情報を提示するシステムを提案する．本システムの視線を用いたターゲット選択には，これまで我々が研究を行ってきた，ユーザのサッケード（跳躍性眼球運動）を検出したとき移動しているターゲットをある間だけ疑似的に静止させ，その間に疑似的に静止しているターゲットを視線のみで選択する手法[2]を用いる．サッケードとは，多くの物から見たい物を見つけるための飛び石を渡るような素早い眼球運動である．サッケードが生じている間，視覚的な認知能力が低下するサッケード抑制という特性がある．この特性からサッケードが起こる際は視覚的な情報の提示は，ほとんど意味がないと考えられる．そこで本研究では，サッケードを検出した際に視線移動に必要な時間だけ動画内のターゲットを疑似的に静止させ，ユーザは疑似的に静止しているターゲットに対してポインティングタスクを行うことでターゲットを選択する．この手法を用いることで，従来の手法[3]よりも軌道の複雑さや速さに関係なく容易にターゲットを選択できる．

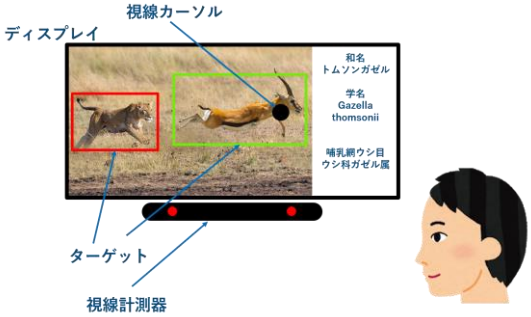

図 1. 研究概要図

2.1 注釈情報の登録

本研究では，前処理として動画ファイルを入力し

* 関西学院大学大学院理工学研究科

物体の検出を行う．本システムは，深層学習に基づく物体検出技術である YOLOv4[4]を用いて物体検出を行った．また本システムにおいてサッケードを検出すると同時にターゲットを疑似的に静止させる際，物体検出の精度が原因でサッケードを検出したときのフレームでターゲットを検出していない可能性がある．そこで本研究では，物体検出した後にカルマンフィルターを用いた SORT[5]アルゴリズムに深層学習を組み込んだ DeepSort[6]というフレームワークを用いてその物体を追跡する．そして各動画フレームに対して追跡した物体バウンディングボックスの座標を記録する．また追跡した物体毎にラベルを割り当て，各ラベルに提示したい注釈情報を登録する．

2.2 ユーザインタフェース

本システムでは，ユーザは前処理で入力された動画を視聴すると同時にディスプレイ上におけるユーザの視線座標を検出する．また動画閲覧の際，選択可能なターゲットにはバウンディングボックスを提示する．そしてサッケードを検出した際にターゲットを疑似的に静止させ，ユーザは疑似的に静止したターゲットに対して視線カーソルを移動させ，ポインティングタスクを行う．本システムでは，サッケードを検出した際にその動画フレームに対応する前処理で記録した検出物体のバウンディングボックスの座標を呼び出し，ターゲットを疑似的に静止させる．ユーザがターゲットを選択した場合，選択した追跡ターゲットのラベルを特定し，このラベルに対応する注釈情報を図 2 のように提示する．

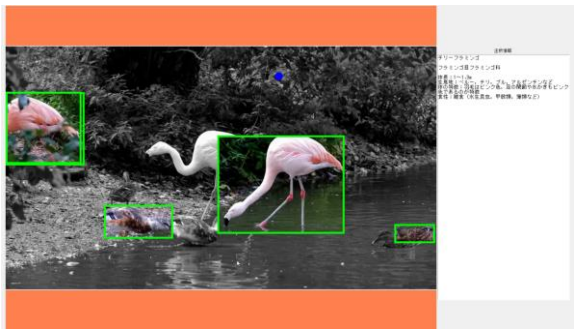

図 2. GUI 操作時の様子

2.3 視線を用いたターゲット選択

本システムでは，サッケードを検出した際に視線移動に必要な時間だけターゲットを疑似的に静止させ，ユーザは疑似的に静止しているターゲットに対してポインティングタスクを行うことでターゲットを選択する．本システムでは Kumar ら[7]が提案した手法を用いて，取得した最も新しい視線座標とその次に取得したいくつかの視線座標とのユークリッド距離がすべてある閾値以上のときサッケードを検出した．そしてサッケードを検出したときのフレームに映る動画内のターゲットを疑似的に静止させる．このとき疑似的に静止したターゲット以外の視覚情報をグレースケール化することでユーザに目立たないように加工して動画を再生する．またユーザの視覚的認知能力に対する影響を最小限にするために，ターゲットを疑似的に静止させる時間はポインティングタスクに必要最低限な時間に設定する．本研究では，ヒトのポインティングタスクのモデルであるフィッツの法則[8]を用いてターゲットを疑似的に静止させる時間を定めた．フィッツの法則は，ポインティングタスクの開始点からターゲットまでの距離とターゲット幅からポインティングタスクにかかる時間を推定できる．本研究では，サッケードを検出したときに用いた視線座標のうち最も新しい視線座標以外のいくつかの視線座標の重心をポインティングタスクの開始点と置く．またサッケードの検出に使用した 2 番目の視線座標をポインティングタスクの開始時間に設定する．そしてユーザは，疑似的に静止しているターゲットを一定時間凝視することで選択を行う．ここでもユーザの視覚的認知能力に対する影響を最小限にするために Isomoto ら[9]の手法を用いてターゲット選択のための凝視時間を低減させる．この手法では，フィッツの法則によって推定したポインティングタスクにかかる時間と実際にかかったポインティングタスクの時間が十分小さくなったときターゲットを選択する．

3 実装

本システムは Python の GUI 構築用ライブラリである Tkinter を用いて実装した．またユーザの視線検出の際には Tobii 社の Tobii Eye Tracker 4C を使用した．本システムの実装環境は Intel(R) Core i7-8750H CPU と NVIDIA GeForce GTX 1060 の GPU を搭載した PC を使用した．

4 おわりに

本研究では，視線を用いて動画内に映る物体を選択し，その物体の注釈情報を提示するシステムを開発した．今後の課題として，物体を検出する際に学習データセットが必要となる点やターゲット選択精度が挙げられる．またユーザが閲覧する動画のストーリー性の有無や動画閲覧時間の長さによるユーザの影響なども考慮して評価していく必要があると考えられる．

参考文献

[1] Linda E. Sibert and Robert J. K. Jacob. “Evaluation of eye gaze interaction”. Proc. CHI'00, pp.281-288, 2000.

[2] Keita Shimasato and Yasuyuki Kono. “Gaze-Based Moving Target Acquisition using Pseudo Stopping for the Time predicted via Fitts' Law”. Proc. AVI’20, Article No. 84, 2020.

[3] Mélodie Vidal, Andreas Bulling, and Hans Gelleren. “Pursuits: Spontaneous Interaction with Displays based on Smooth Pursuit Eye Movement and Moving Targets”. Proc. UbiComp’13, pp. 439-448, 2013.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “YOLOv4: Optimal Speed and Accuracy of Object Detection”. arXiv preprint, arXiv: 2004.10934, 2020.

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, and Ben Upcroft. “Simple Online and Realtime Tracking”. arXiv preprint, arXiv: 1602.00763, 2016.

[6] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple Online and Realtime Tracking with a Deep Association Metric”. arXiv preprint, arXiv: 1703.07402, 2017.

[7] Manu Kumar, Jeff Klingner, Rohan Puranik, Terry Winograd, and Andreas Paepcke. “Improving the Accuracy of Gaze Input for Interaction”. Proc. ETRA’08, pp. 65-68, 2008.

[8] Paul M. Fitts, 1954. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement, In the Journal of Experimental Psychology, Vol.74, pp. 381-391.

[9] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. “Dwell Time Reduction Technique using, Fitts’ Law for Gaze-Based Target Acquisition”. Proc. ETRA’18, Article No 26, 2018.