基於螢幕的視線追踪模型
Screen-Based Gaze Tracking Model

組員：潘品齊、陳俊騰　　指導教授：林惠勇 教授

摘要

人眼視線追蹤一直是電腦視覺領域最具挑戰性的問題之一，該研究已被廣泛的使用在各個領域中，例如駕駛人輔助、虛擬實境和人機互動等。其中駕駛人輔助能夠藉由視線追蹤模型預測車輛駕駛人視線範圍並評估其狀態，在駕駛安全中起著重要作用。

傳統視線追蹤模型的視線預測方法主要是由眼睛狀態來預測其視線範圍，該方法在穿戴式視線追蹤設備上表現優異，不過其缺點是穿戴式設備成本過高，若改為使用非穿戴式設備進行視線追蹤，便能夠有效降低成本，但也會因為頭部姿勢的運動而影響視線預測效果。

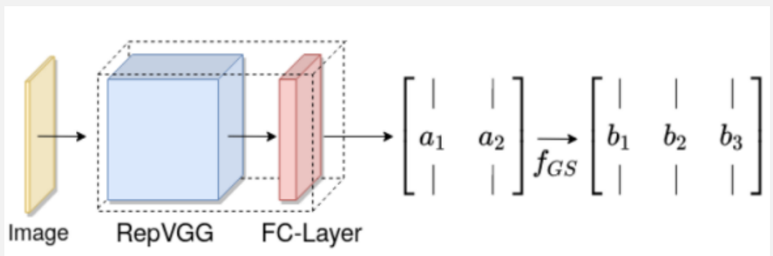
本研究提出了一種基於外觀的頭部姿勢估計並結合視線追蹤的預測方法，除此之外我們還使用機器學習中的決策樹演算法來預測駕駛人正在觀看的視線區域。

研究方法

在頭部姿勢預測方面，我們選擇使用 6DRepNet 作為主要網路架構，並將 Vision Transformer 以及 Swin Transformer 模型替換其中的 RepVGG 進行比較與優化。在損失函數上，我們選擇使用 Geodesic Loss 進行訓練並與我們常使用的損失函數 MSE Loss 及 L1 Loss 進行比較。

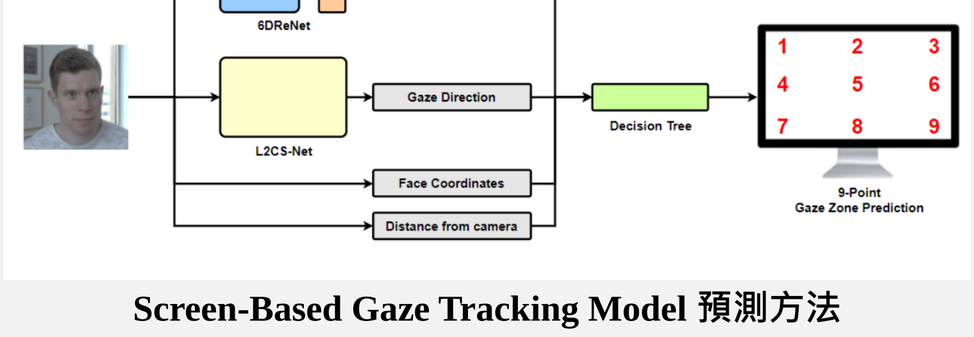
$$L_g = cos^{-1}\left(\frac{tr(R_p R_{gt}^T) - 1}{2}\right)$$

Geodesic Loss



6DRepNet 頭部姿勢預測方法

在人眼視線預測方面，我們選擇使用 L2CS-Net 作為我們的主要網路架構，並同樣將 Vision Transformer 以及 Swin Transformer 模型替換其中的 ResNet50 進行比較與優化。



L2CS-Net 人眼視線預測方法

為了預測使用者正在觀看的螢幕區域，我們將臉部姿勢預測出的 Yaw、Pitch、Roll 值，以及人眼視線預測出的 Yaw、Pitch 值，加入人臉 x、y 軸座標以及與鏡頭的距離作為輸入來訓練我們的決策樹模型，為了避免決策樹的深度太深導致過擬合的問題，我們將決策樹的最大深度設置為 11。

Screen-Based Gaze Tracking Model 預測方法

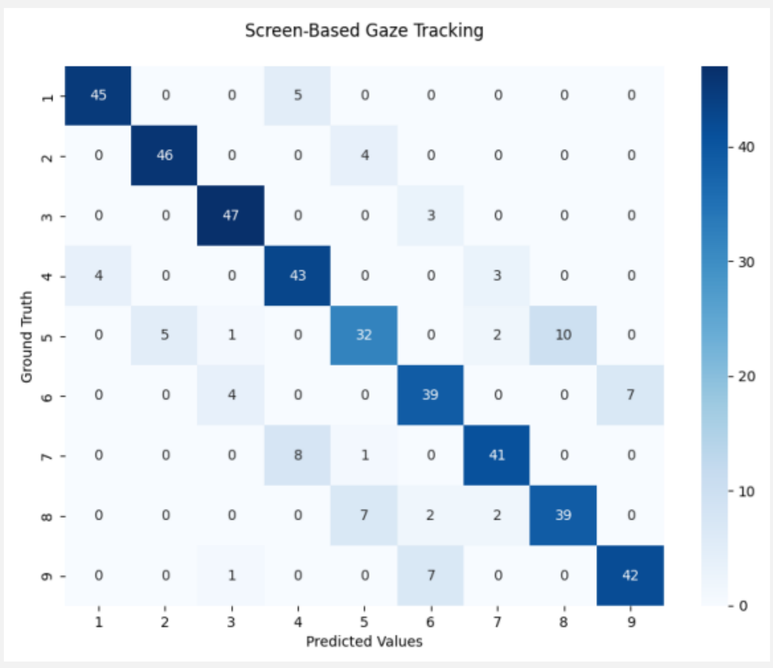
實驗結果與討論

本專題將臉部姿勢以及人眼視線預測模型透過決策樹模型成功建構出基於螢幕的視線追蹤模型。我們也嘗試使用不同數量的特徵當作決策樹的訓練資料進行測試。根據我們的研究表明當特徵數量越多時，決策樹模型的準確率也相對提升。若不加入臉部深度，我們的決策樹模型表現會下降約一個百分點，而當不加入臉部座標或角度進行訓練時，決策樹模型的表現會下降更多，這也證明了我們最初的想法，即頭部姿勢的運動會影響視線追蹤模型的性能，所以在訓練視線追蹤模型時，應該要始終加入頭部姿勢預測結果以提升模型效能。

# Features	Gaze Pitch	Gaze Yaw	Face Pitch	Face Yaw	Face Roll	Face X	Face Y	Face Depth	Accuracy
5	✓	✓	-	-	-	✓	✓	✓	81.78%
6	✓	✓	✓	✓	✓	-	-	✓	81.78%
7	✓	✓	✓	✓	✓	✓	✓	-	82.00%
8	✓	✓	✓	✓	✓	✓	✓	✓	**83.11%**

使用不同特徵數量訓練決策樹之結果

本專題最終將螢幕分為九個區域，其結果以混淆矩陣表示。我們的模型在每個區域的預測上，最可能會誤判的類別為其上方與下方的區域。根據我們的研究發現越靠近鏡頭的區域，其準確率比其他區域還高出 10%。在本研究中，我們的基於螢幕之視線追蹤模型經過訓練後可達到 83.11% 的準確率。



9-Point 螢幕視線區域預測結果

結論

在本研究中我們提出了一種基於螢幕的視線追踪模型(Screen-Based Gaze Tracking Model)，改善了非穿戴式設備在進行視線預測時會因為頭部姿勢而導致視線預測產生誤差的問題，並且我們也分析了在不同架構與損失函數的訓練下，頭部姿勢與人眼視線預測模型的性能表現。

根據我們的研究表明在頭部姿勢預測方面基於 Transformer 的網路模型，比起基於 CNN 的網路模型表現更好。而在人眼視線預測方面，基於 CNN 的網路模型表現較為優異。在螢幕視線區域預測方面，除了頭部姿勢與人眼視線預測外，透過加入人臉座標以及人臉與鏡頭的距離當作決策樹的訓練參數，能使模型在預測使用者所觀看的螢幕區域時能更準確。