**實 踐 大 學**

**高雄校區**

資 訊 科 技 與 通 訊 學 系

專 題 論 文 報 告



字體大小22點標楷體

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

( 論 文 題 目 )

學 生：

指導教授：

中 華 民 國 年 月 日

# 摘要

車輛行駛中駕駛的分心行為可能造成嚴重的交通事故，識別出車內駕駛的行為是近年來重要的應用。此研究使用的資料集以車內的三處攝影機針對駕駛進行攝影，分心行為共 17 種，目標是在影片中識別出駕駛分心行為，屬於時序動作檢測 (Temporal Action Localization) 任務。本研究使用 X3D 網路搭配時序定位演算法進行駕駛分心動作的識別，研究重點包括三個視角模型的不同組合對於評估分數的影響，並嘗試調整視角模型合併的權重；另外將分心動作做不同分組重新訓練，觀察對於時序定位與動作偵測的影響；基於文獻[[1]](#_bookmark82)的時序定位演算法進行改良實驗，包括放寬不易偵測的動作類別之閾值以及動作重疊後處理兩部分，並分析改良後的成效。

關鍵字：動作辨識、駕駛分心行為、時序動作檢測、深度學習、時序定位(參考範例)

# 目錄

[誌謝 i](#_bookmark0)

[摘要 ii](#_bookmark1)

[Abstract iii](#_bookmark2)

[目錄 iv](#_bookmark3)

[圖目錄 vi](#_bookmark4)

[表目錄 vii](#_bookmark5)

[第一章 緒論 1](#_bookmark6)

* 1. [研究背景 1](#_bookmark7)
  2. [研究動機 2](#_bookmark9)
  3. [論文架構 3](#_bookmark11)

[第二章 文獻探討 4](#_bookmark12)

* 1. [動作辨識 4](#_bookmark13)
  2. [多攝影機下的動作辨識 5](#_bookmark15)
  3. [時序動作檢測 5](#_bookmark16)
  4. [駕駛行為辨識 6](#_bookmark17)

[第三章 研究方法 8](#_bookmark18)

* 1. [資料集 8](#_bookmark19)
  2. [Backbone 模型：X3D 9](#_bookmark21)
  3. [視角模型訓練 10](#_bookmark25)
     1. [比較三個視角模型評估之成效 12](#_bookmark28)
  4. [將動作做不同分組進行訓練 13](#_bookmark29)
  5. [時序定位改良 14](#_bookmark31)
     1. [不易偵測的動作類別之閾值設定 15](#_bookmark34)
     2. [動作重疊問題與後處理 16](#_bookmark36)
  6. [訓練參數 19](#_bookmark40)

[第四章 實驗結果 20](#_bookmark41)

* 1. [實驗環境 20](#_bookmark42)
  2. [評估方法 20](#_bookmark43)
  3. [視角模型訓練 21](#_bookmark45)
     1. [不同訓練方式之比較 21](#_bookmark46)
     2. [抽換 X3D 模型實驗 24](#_bookmark51)
     3. [不同視角下的模型評估比較 25](#_bookmark54)
     4. [調整視角模型合併之權重 26](#_bookmark57)
  4. [將動作做不同分組進行訓練 27](#_bookmark59)
  5. [時序定位改良實驗 31](#_bookmark65)
     1. [不易偵測動作類別之閾值設定 31](#_bookmark66)
     2. [動作重疊問題之後處理 33](#_bookmark71)
     3. [時序定位改良的搭配組合 37](#_bookmark77)

[第五章 結論與未來展望 38](#_bookmark80)

[參考文獻 39](#_bookmark81)

# 圖目錄

[圖 1 三個視角下的 REACHING BEHIND 動作 1](#_bookmark8)

[圖 2 DASHBOARD 視角下的TEXT (LEFT) 動作 2](#_bookmark10)

[圖 3 SLOWFAST 網路架構[3] 4](#_bookmark14)

[圖 4 X3D 網路示意圖[4] 9](#_bookmark22)

[圖 5 文獻[1] MODEL ENSEMBLE 示意圖 11](#_bookmark26)

[圖 6 使用 3 個視角模型進行預測之示意圖 12](#_bookmark27)

[圖 7 文獻[1]之時序定位演算法 14](#_bookmark32)

[圖 8 改良後的時序定位流程圖 15](#_bookmark33)

[圖 9 文獻[1]於 A2 資料集-VIDEO 9 的 INFERENCE 結果 17](#_bookmark37)

[圖 10 動作重疊範圍示意圖 17](#_bookmark38)

# 表目錄

[表 1 AIC22 TRACK3 資料集 18 種動作類別 8](#_bookmark20)

[表 2 X3D 網路架構[4] 9](#_bookmark23)

[表 3 不同大小 X3D 模型的規格[4] 10](#_bookmark24)

[表 4 分九類實驗之分組 13](#_bookmark30)

[表 5 各個分心動作在 A2 資料集的 DETECTION RATE (#ID: VTCC-UTVM) 16](#_bookmark35)

[表 6 文獻[1]於 A2 資料集-VIDEO 9 的 INFERENCE TOP-3 SCORES (節錄 146 ~ 171 秒) 19](#_bookmark39)

* 1. 研究背景(參考範例)

# 第一章 緒論

在車輛行駛中，駕駛之分心行為可能造成嚴重的交通意外事故，而識別出這些駕駛異常行為是降低這類型車禍的重要手段。使用深度學習與電腦視覺等技術分析車內駕駛之行為是當今熱門的研究主題之一，每年也有舉辦許多競賽如 AI City Challenge，試圖激盪出更好的解決方案。本研究以 2022 AI City Challenge Track3 [[2]](#_bookmark83)為主題，使用其資料集進行研究，透過車內之三部攝影鏡頭（儀表板處、後照鏡處、右側副駕駛處）同時錄影紀錄駕駛在行駛中的行為，目標是在這三處攝影機視角之影片中識別出駕駛分心行為

（包含講電話、撿拾物品等 17 種動作）。以動作類別 8: Reaching Behind 為例，如[圖 1](#_bookmark8)

所示。

本研究基於 2022 AI City Chal，以期取得更佳的駕駛分心行為識別效果。



圖 1 三個視角下的 Reaching Behind 動作

## 1.2.研究動機

關於駕駛行為識別任務，有以下困難點：

* + 1. 車內駕駛動作辨識上的困難：對於車內駕駛所進行之分心動作，部分相異的動作之間相似性高而難以辨別。如[圖 2](#_bookmark10) 所示。
    2. 模型訓練上的挑戰：因車內的駕駛行為辨識領域缺少公開 benchmark 資料集， 關於駕駛行為識別相關的預訓練模型較為缺乏，故對於 transfer learning 形成挑戰。

* 1. 動作辨識(參考範例)

# 第二章 文獻探討

動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[[3]](#_bookmark84)、X3D[[4]](#_bookmark85)等。動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等

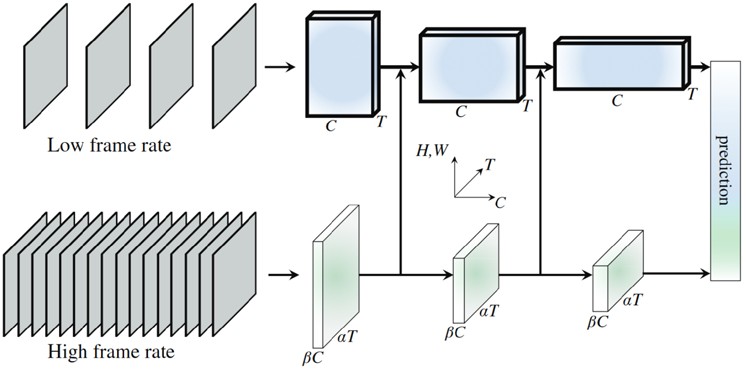


圖 3 SlowFast 網路架構[[3]](#_bookmark84)

## 時序動作檢測(參考範例)

時序動作檢測 (Temporal Action Localization/Detection, TAL) 任務是在一段較長、未經裁切的影片中識別出動作之類別，以及動作的起始與結束時間，大致上可分成Two-staged 與 Single-staged 兩種方式。Two-staged 方式的第一階段會先產生可能為動作的 proposals，並在第二階段對這些 proposals 做動作分類或優化時序的邊界，經典的方法例如 R-C3D[[11]](#_bookmark92)。Single-staged 方式則整合生成動作 proposals 與動作分類在一個end-to-end 架構，例如 Single Shot Action Detector (SSAD)[[12]](#_bookmark93)。

* 1. 資料集

# 第三章 研究方法

本研究使用 2022 AI City Challenge Track3 提供之資料集，此資料集中包含訓練集A1、驗證集 A2。

此外，僅訓練集 A1 提供分心動作之類別及起始/結束時間 ground truth 標記，A2 資料集僅供驗證使用動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等動作辨識 (Action Recognition) 任務是電腦視覺領域中影片理解 (video understanding) 的基礎，其做法可分為 CNN-based 及 Transformer-based 兩大類。CNN- based 方法為 2D 影像分類 CNN 的延伸，從空間域拓展至空間及時間域，常見的方式包含 SlowFast[3]、X3D[4]等

## 將動作做不同分組進行訓練

由於 17 種分心動作中包含較難辨別的動作，例如動作類別 16: Singing with music 與類別 17: Shaking or dancing with music；另外還有一些動作的左右邊方向也需區分，例如動作類別 2: Phone call (Right) 與類別 3: Phone call (Left)。故我們好奇若將模型在較為寬鬆的

第二種分組方式為自行重新分組，共區分為九類，。

表 4 分九類實驗之分組

|  |  |  |  |
| --- | --- | --- | --- |
| **New Class ID** | **Action (Original Class ID)** | **New Class ID** | **Action (Original Class ID)** |
| **0** | Normal Driving (0) | **5** | Reaching behind (8) |
| **1** | Drinking (1)  Eating (4) | **6** | Adjust control panel (9) |
| **2** | Phone Call right (2)  Phone Call left (3) | **7** | Pick up from floor (Driver) (10)  Pick up from floor (Pass.) (11) |
| **3** | Text right (5)  Text left (6) | **8** | Talk to pass. at the right (12) Talk to pass. at backseat (13) Yawning (14)  Singing with music (16)  Shaking or dancing with music (17) |
| **4** | Hair/makeup (7) Hand on head (15) |

* 1. 實驗環境

# 第四章 實驗結果

實驗使用之硬體設備為 Intel® Core™ i7-8700 CPU、32GB RAM、NVIDIA GeForce® GTX 1080 Ti GPU，作業系統為 Ubuntu 64-bit。程式相依性為 python 3.8、torch 1.8.1+cu101、torchvision 0.9.1+cu101、cudatoolkit 10.1。

## 評估方法

評估採用 2022 AI city challenge Track3 指定的計算方式，說明如下。評估指標包含Precision、Recall、F1-score，一個動作區間包含動作的類別以及起始/結束時間，以動作區間作為單位進行以下判定。一個動作被成功識別成 True Positive (TP) 之條件為：與ground truth 動作類別相符，且動作之起始/結束時間與 ground truth 標記相差在一秒內。非 TP 者，包含時間區間誤差大於一秒或動作類別分類錯誤，為 False Positive (FP)。若動作在 ground truth 中有標記但未被系統成功識別出，為 False Negative (FN)。最終以Precision 及 Recall 的調和平均 F1-score，做為綜合的評估分數。

由於實驗會在 A2 資料集進行評估而 A2 資料集並未給予 ground truth，且 2022 AI City Challenge Track3 線上評估系統有上傳次數限制並已下線(2023/01)，為了方便評估分數，因此我們對 A2 進行人工標記。為了確認人工標記的正確性，我們將其上傳至評估系統的 General Leaderboard 進行檢查，其中人工標記的 F1-score 分數達 0.5455、Precision 0.5549、Recall 0.5363，在 General Leaderboard 排名第二，顯示我們的人工標記具備一定的標記品質。

# 第五章 結論與未來展望

本研究使用影片識別 X3D 模型搭配時序定位演算法，來進行駕駛分心行為的偵測。在模型訓練上，嘗試了兩種視角模型訓練整合方式；另外將三個視角（儀表板 處、後照鏡處、右側副駕駛處）模型做不同的搭配組合進行評估，並比較各個組合評估分數的差異；我們也嘗試調整視角合併時的權重，觀察對於最終合併後的評估分數

之影響。另一部分，將 17 種分心動作以不同的分組重新訓練，觀察對於動作分類以及時序定位的影響。最後，我們嘗試在原先的時序定位演算法上找尋可改進的部分，包括對不易偵測的動作之閾值進行放寬，以及使用兩種後處理方式來解決原先演算法的動作重疊問題，最後將這些改進方法做搭配組合，比較對於最終評估分數的改善。

未來展望部分，包括更佳的視角資訊整合方式的研究、最佳化各個動作類別的閾值設定與更佳的時序定位演算法的設計等等，以及未來在有更充足的自然駕駛行為資料集的前提下進行訓練與評估，來驗證這套方法在駕駛分心行為偵測上的可靠性與準確性。

# 參考文獻

1. Tran, Manh Tung, et al. "An effective temporal localization method with multi-view 3D action recognition for untrimmed naturalistic driving videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
2. https:[//w](http://www.aicitychallenge.org/)ww[.aicitychallenge.org/](http://www.aicitychallenge.org/)
3. Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
4. Feichtenhofer, Christoph. "X3d: Expanding architectures for efficient video recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
5. Fan, Haoqi, et al. "Multiscale vision transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
6. Liu, Ze, et al. "Video swin transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
7. Kay, Will, et al. "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).
8. Gao, Zan, et al. "Adaptive fusion and category-level dictionary learning model for multiview human action recognition." IEEE Internet of Things Journal 6.6 (2019): 9280- 9293.
9. Nguyen, Chuong, et al. "Learning generalized feature for temporal action detection: Application for natural driving action recognition challenge." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
10. Zhang, Chenlin, Jianxin Wu, and Yin Li. "Actionformer: Localizing moments of actions with transformers." arXiv preprint arXiv:2202.07925 (2022).
11. Xu, Huijuan, Abir Das, and Kate Saenko. "R-c3d: Region convolutional 3d network for temporal activity detection." Proceedings of the IEEE international conference on computer vision. 2017.
12. Lin, Tianwei, Xu Zhao, and Zheng Shou. "Single shot temporal action detection." Proceedings of the 25th ACM international conference on Multimedia. 2017.
13. Choi, In-Ho, and Yong-Guk Kim. "Head pose and gaze direction tracking for detecting a drowsy driver." 2014 international conference on big data and smart computing (BIGCOMP). IEEE, 2014.
14. Abtahi, Shabnam, Behnoosh Hariri, and Shervin Shirmohammadi. "Driver drowsin