

附件 一、資料集說明文件格式(範本)

「合成繁中英文數字場景影像資料集」(資料集名稱)

資料集說明文件 v0.2.2

建置團隊資訊

子計畫名稱：__可信賴的生活場景文字識別技術__

子計畫主持人：____ 陳祝嵩____

說明文件版本：__v0.2.2__

日期：____111.08.10____

1 資料集基本說明

資料集名稱		合成繁中英文數字場景影像資料集
資料集摘要 (含目的/用途/價值/標註重點等)		• 含繁中英文數字之合成影像 • 文數字之外框座標標註 • 文數字之答案標註
資料集聯絡人		(姓名/E-mail/聯絡方式) 吳玉辰 winie850609@gmail.com / 陳冠盛 gschen.willy@gmail.com
資料類型		選擇(結構化數據，語音，文字，影像，其他) 影像
資料	資料集類型	影像
	儲存格式	png
	來源機構	(若有多個提供機構請分列) Youtube
	收集期間	110 年 11 月至 111 年 8 月
	案例數	
	筆數/案例	(若每個案例收錄多筆資料請註明) 共 1135834 張訓練資料、1110 張驗證資料、1086 張測試資料
	總共容量	476.7GB
	適法性基礎 (本項暫不對外公開)	☐ A 類(新收案)：000 例 ☐ B 類(曾取得概括同意，已補告知程序)：000 例 ☐ C 類(未曾取得同意，已補告知程序)：000 例 ☒ D 類(資料已去識別，無法進行告知)：000 例 ☐ E 類(資料無個資，不需進行告知)：000 例

	其它說明	
標註資料	標註工具	開發之自動合成方法
	標註方式	☐ 自動標註 ☐ 人工標註 ☒ 自動標註+人工校正 ☐ 其他：_____
	註記方式	☒ 有註記：_有文字的類別、外框外置與文字內容____ ☐ 其他：_____
	標註儲存格式	json
	標註情形	☒ 全部標註完成：000 例 ☐ 僅標註部份項目：000 例 ☐ 僅原始資料，未加標註：000 例
	其它說明	
與其他資料庫的串接		☐ 保留與本團隊其他資料庫串接可能性：未來可透過假 id 串接本團隊其他的資料庫 ☐ 保留與國網中心資料庫串接可能性：與上傳國網中心的資料採同樣的假 ID，以利後續資料更新/維護（但僅限由本團隊主動進行更新與維護） ☒ 無串接規劃 ☐ 其他：_____

2 案例相關資訊欄位說明

案例資訊項目		欄位名稱 (呈現於檔案的名稱)	欄位格式 (數值/字串/文字等)	範例/備註說明
個案編號				
基本 資料	出生日期			
	性別			
	其他			
	...			
資料 取得 設備 資訊	取得日期			
	資料取得設備 型號			
	其他			
	...			
每筆 資料 描述 (選 填)	代碼			
	分類			
	其他			
	...			
延伸 資訊	報告			
	其他			
	...			

3 標註說明範例

使用到下述檔案

標註項目	資料檔	標註檔
文字外框位置、文字內容	1.png	1.json

3.1 資料與標註資料的對應關係

壓縮檔中之 cropped_synth_images 與 cropped_synth_labels 分別存放影像資料與標註檔案，影像資料中的檔名為「數字.png」，對應到標註檔案的「數字.json」。

3.2 標註內容

標註欄位說明

欄位名稱	說明	檔案截圖
labels	文字內容	<pre>{ "version": "0.2.2", "flags": {}, "shapes": [{ "label": "林智雄", "points": [[687, 48], [630, 45], [628, 48], [605, 52]], "group_id": 0, "shape_type": "polygon", "flags": {} }] }</pre>

points	文字外框座標，從左上角的點 [x,y] 順時鐘標註	
group_id	文字的分類，數字代表以下分類： 0: 中文字串 1: 中文字元 2: 英數字串 3: 中英數混合字串 4: 中文單字字串 5: 其他 8: 英數字元	
shape_type	文字外框種類，本標註都是 polygon	

4 工具

可以使用下列工具從儲存的檔案中讀取資料

資料類別	公開/免費軟體名稱/版本	商用軟體名稱/版本
資料	一般的影像瀏覽器 Python 之 opencv 套件	
標註資料	一般的文字編輯器 Python 之 json 套件	

5 去識別化

去識別化方式	☐ ☒ 遵循本專案共同制訂方式
	☐ 本資料集自行定義方式 (勾選本項請逐欄位說明各欄位之處理作法，使用工具，參考資料與連結) 說明： ● 使用 FaceMaskDetection 工具針對場景影像進行人臉偵測，並且將該區域進行模糊化，工具連結： https://github.com/AIZOOTech/FaceMaskDetection ● 使用自行訓練之 yolov5 車牌偵測模型，將影像中的車牌偵測出來，並且進行車牌文字的代換