中央大學MIAT實驗室

鄧祺文

IMG & PDF OCR WEBUI

技術手冊

**系統設計架構**

**辨識模型**

* 論文主題：

SVTR: Scene Text Recognition with a Single Visual Model

* 論文地址：

<https://arxiv.org/abs/2205.00159>

* 模型代碼開源地址：

<https://github.com/PaddlePaddle/PaddleOCR>

* 技術簡介：

場景文字識別可以看作是一個從圖像映射到序列的任務。大多數的識別算法通常由兩個模塊構成，即用於特征提取的視覺模塊以及用於文本輸出的序列模塊。比如早期基於CNN-RNN的CRNN[1]，和現在一些基於注意力機制，進行自回歸式解碼的算法。但是這樣的雙階段算法的推理速度往往較慢，難以滿足工業應用的需求。因此本文從推理速度和模型性能的雙重角度出發，提出了只由Transformer構成的純視覺模塊網絡SVTR，在NVIDIA 1080Ti GPU上達到了 4.5 ms的推理速度，並且參數量僅有6.03M。

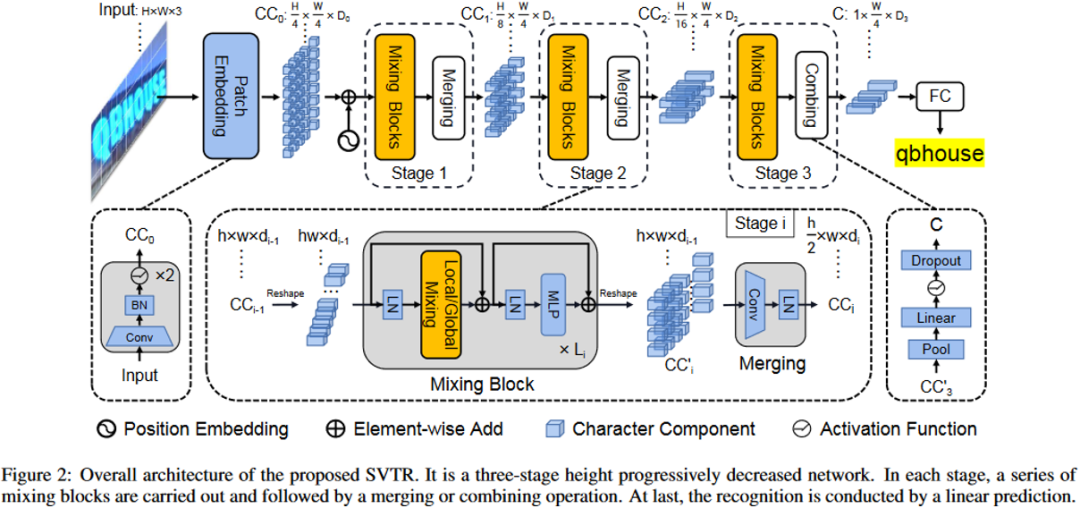
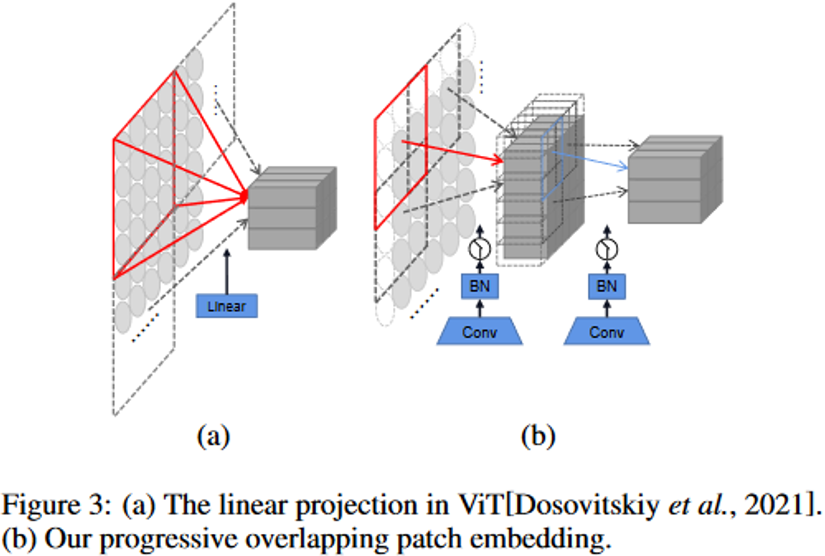
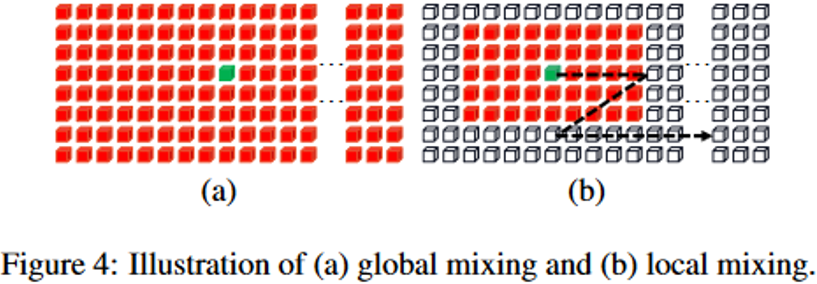


圖2是這篇文章提出的SVTR的整體結構，采用類似於SwinTransformer[2]的視覺模型和一個全連接層以及CTC解碼器進行文本序列預測。

首先和ViT[3]類似，將輸入尺寸為圖像按照Patch進行劃分, 得到  Embeddings。本文采用的Patch Embedding操作和ViT中的有些許差異，其由兩層步距為2，卷積核大小為卷積層3X3，以及BN層構成。這樣不同的Patch之間是存在著重疊的，如圖3所示。經過Patch Embedding後的序列將經過一系列的Stage，每一個Stage都由一系列的Mixing Block和Merging Layer構成。



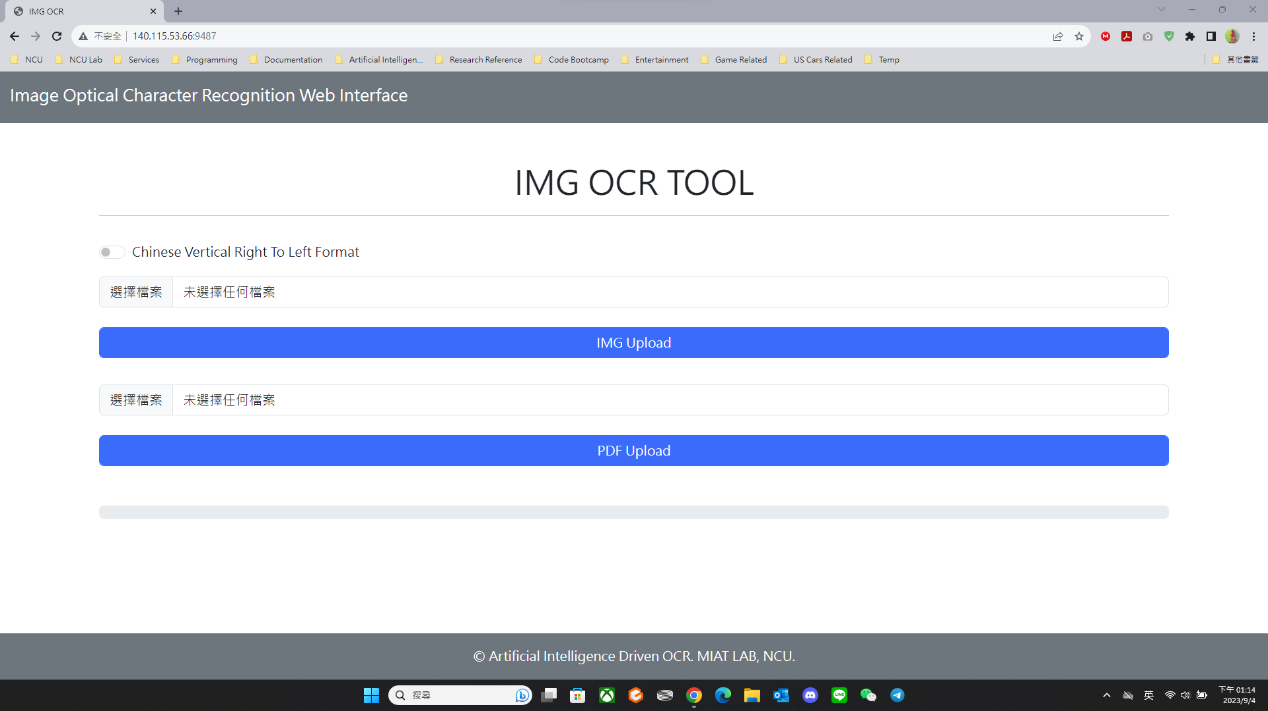
作者認為文本識別需要兩種特征。第一種是局部特征，如筆畫特征。它編碼了字符的不同部分之間的形態特征和相關性。第二種是字符間的依賴性，如不同字符之間或文字與非文字成分之間的相關性。因此，作者設計了兩個混合模塊，即 Global Mixing 和 Local Mixing， 通過使用不同大小感受野的自注意層來實現。如圖4 所示。Global Mixing層本質上就是一個Transformer block，由一個多頭自注意層，一個Layer Norm 層，以及一個MLP層構成。通過自注意力機制的全局建模特性來進行全局字符建模。Local Mixing則是采用了帶窗的自注意層，窗大小設置為了



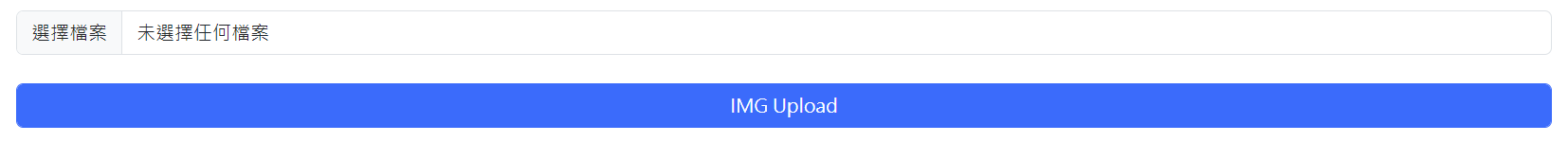
Merging層扮演著將輸入序列進行下采樣的角色。其由高度方向步距為2，寬度方向步距為1，卷積核大小為3X3的卷積層構成。將輸入序列的尺寸由縮小為。同時每經過一次Merging層，序列的Channel維度也會增大，從而彌補在高度上的信息損失。

**網頁功能入口**

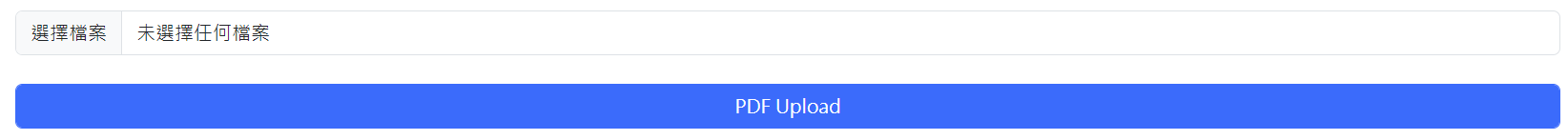
* 功能入口



* 上半部選擇框：支持JPG、JPEG、PNG



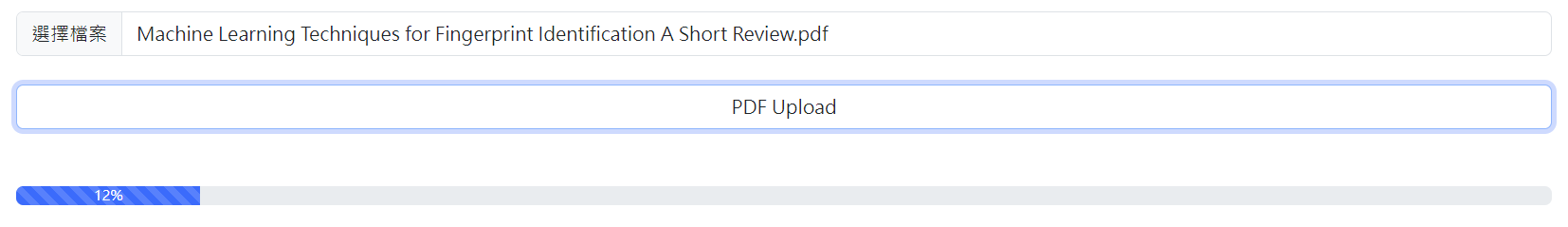
* 下半部選擇框：支持PDF



* 圖像特殊排序：中文直式由右至左（預設標準為由上至下）

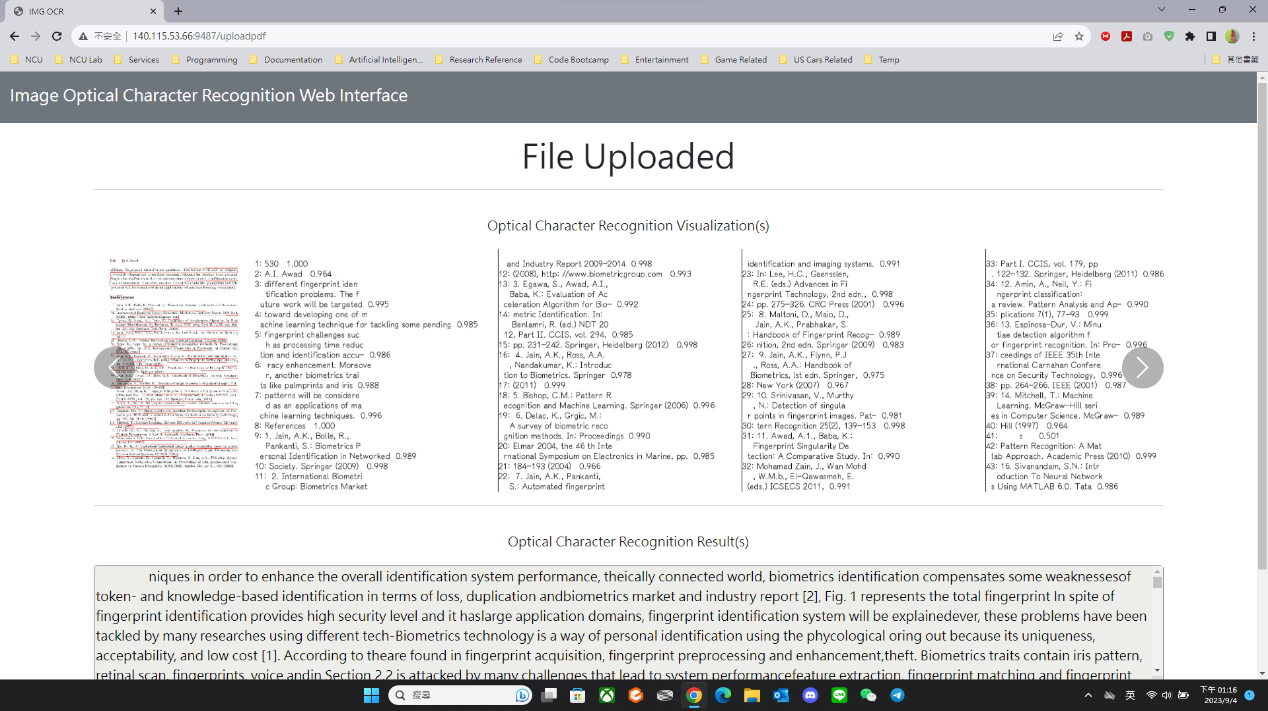


* 成功讀取資料後，可於下方看見處理進度條

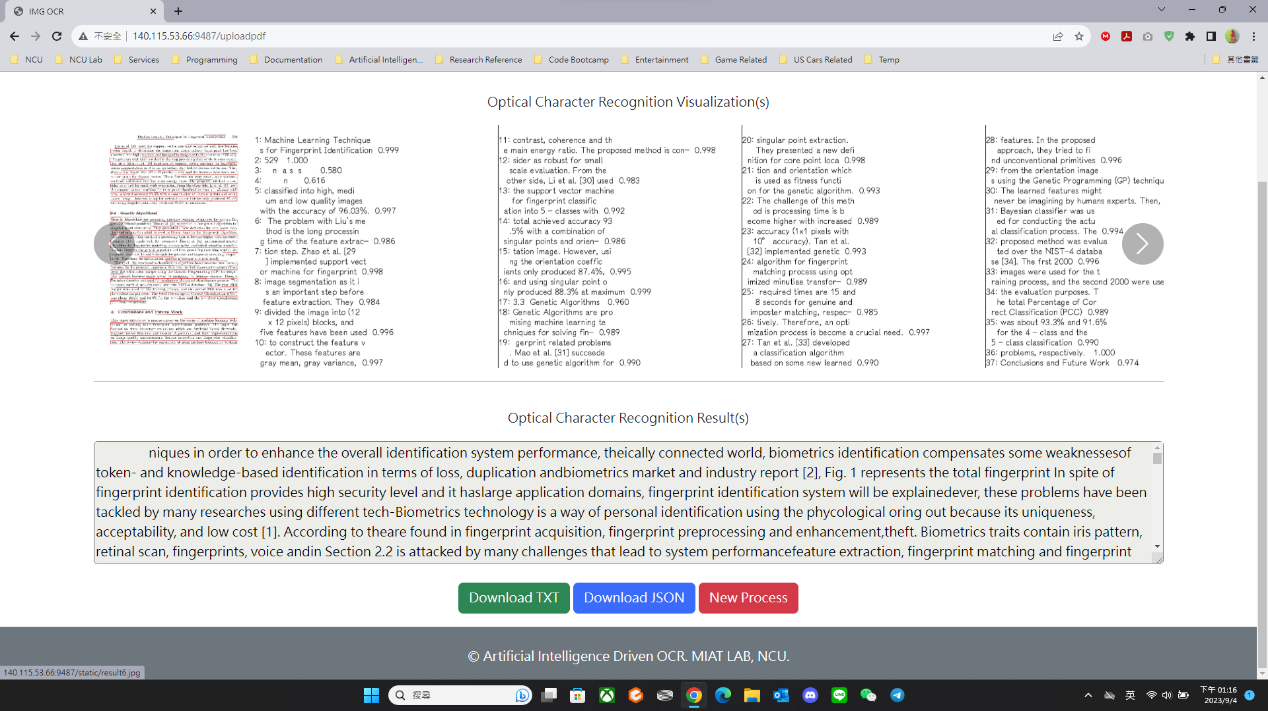


**網頁結果輸出**

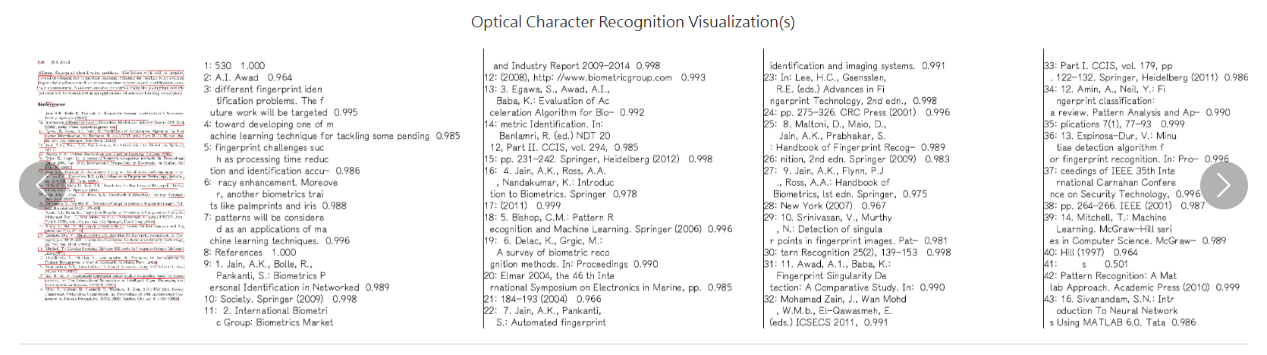
* 結果輸出頁面：上半部



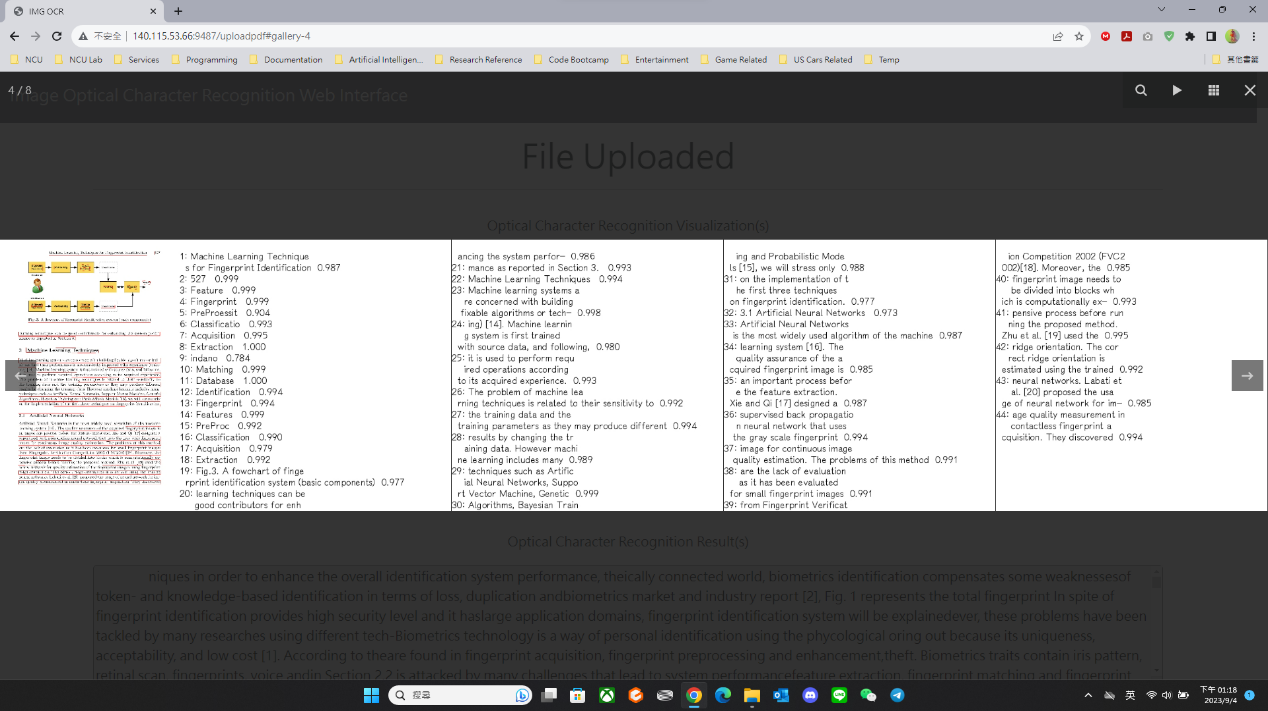
* 結果輸出頁面：下半部

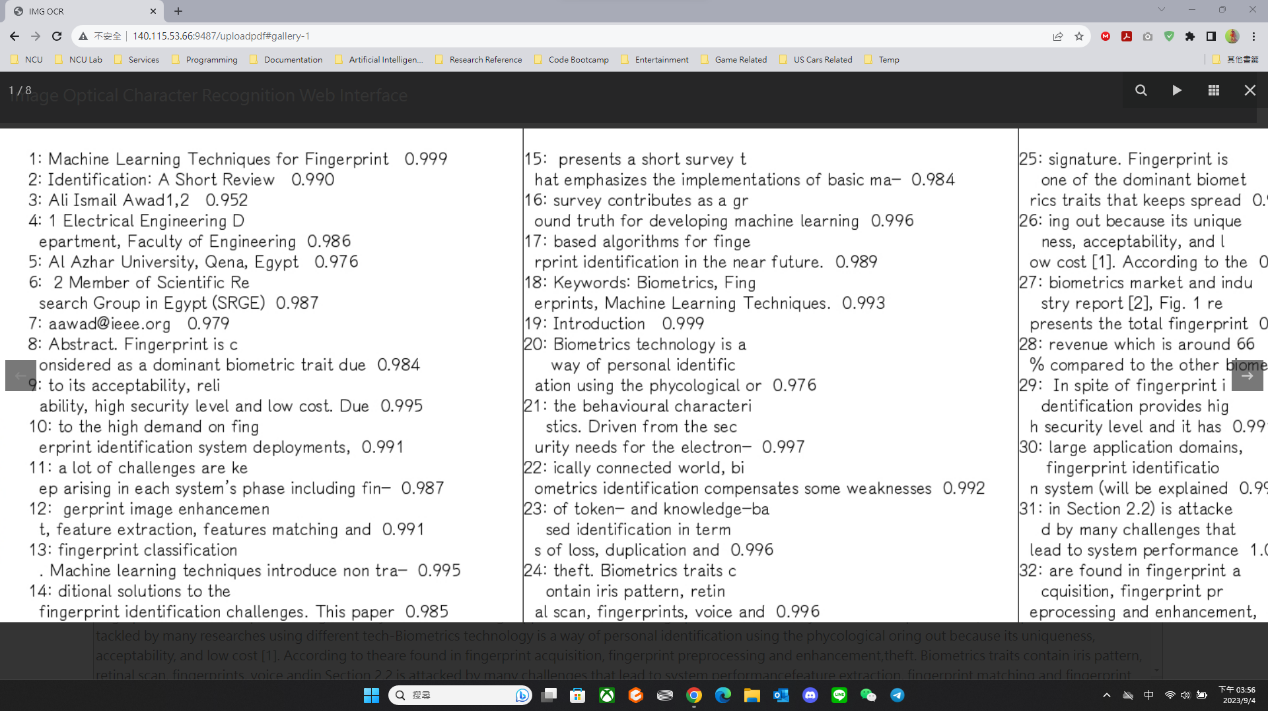


* 輸出視覺化：包含辨識框及輸出結果展示，左右按鍵可變換當前展示頁面

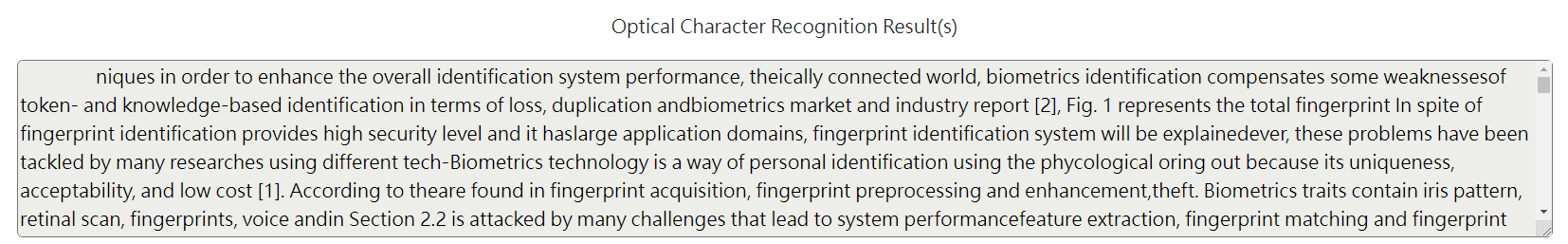


* 直接點擊圖片進入相簿模式，可以對圖片進行放大





* 完整文字輸出：包含所有頁面辨識出之完整文章



* 文章打包功能：包含TXT文字檔下載、API串接用JSON檔案下載，以及重新開始新一輪識別

