第二章 文獻探討

自動化文件處理的研究領域，往往根據應用需求的不同或文件類型的差異，採用截然不同的處理方法。一般而言，自動化文件處理會先經前處理、文件分析及文件了解等前三步驟，並依不同的應用需求，將文件切割成表格、欄位、圖形或照片、直線或標題、子標題及本文等不同形式之區塊。而不同類型之區塊，再依照其實際需求做資料抽取、資料壓縮或光學字元辨識等後續處理工作。

前處理的工作包含了文件影像二元化、雜訊去除及傾斜角度校正等。而文件分析及文件了解皆為擷取文件結構資訊的技術。其中文件分析的工作在擷取文件的幾何結構，而文件了解的工作則是將幾何結構與邏輯結構作對應轉換。

表單文件的處理程序與一般類型的文件處理相同，第一步先分析並抽取出表單的結構，加以判別表單類別，並儲存為樣本表單供比對之用。第二步則是輸入未知表單，與樣本表單進行比對，鑑別其所屬表單類別，若為已知表單類別，則利用已儲存之相關資料欄位特徵資訊，擷取輸入表單中的重要資料；反之為一新表單類別，即分析未知表單之結構及欄位特徵資訊，儲存於樣本表單資料庫中。

本章節分別針對表單文件處理與手寫資料處理等研究領域，探討現有的處理技術與限制。

2.1 表單文件處理

表單文件處理的研究，可區分出各獨立的研究領域，針對各領域之相關文獻，分述如下。

2.1.1 結構切割

首先針對表單的幾何結構與邏輯結構作簡單的定義。所謂幾何結構是利用表單文件外觀與設計格式為基礎之結構分割結果。而邏輯結構則是以類似於人類認知觀點，對表格文件所進行之分析結果，常依照應用需求的不同，而有不同的呈現方式。

在所有表單文件中，表格為表單處理研究中最被廣泛採用的研究對象[2]。由於一般表格本身的結構較明確，且格式較固定，結構複雜度亦不高，因此通常表格中的框線，已經足以表示表格本身的幾何結構。針對這個特質，已有不少研究提出不同的框線抽取方法，再根據所抽取之框線進行表格結構的切割。

Zheng 等人[3]以 Run-based 演算法[1]為基礎，提出根據 Directional single-connected chain (DSCC)的特性，分析出表格中的框線所在。而為了進一步改進 DSCC 方法所造成的線段破損問題，Zheng 等人又提出利用 Hidden Markov Model 的訓練模組[4]，根據線段在水平及垂直投影的端點位置[4,5]，分析表單框線所在。Xi 等人[6,7]則是針對灰階背景及複雜背景之影像，分析小波轉換所得子影像間之關聯性，進行框線的抽取。

Lu 等人[8]提出兩種方法，一是將輸入影像細線化之後，利用特徵點群聚法

求出框線所在，另一方法則是追蹤影像中連通成分的輪廓，求出框線之連通成分。Lam 等人[9]則提出利用鏈結碼追蹤連通成分的外形，以取得表格框線。

傳統的文件切割方法可分為 Top-down 與 Bottom-up 兩種模式。Top-down 的切割方式，針對固定格式之文件，效果與速度較佳。因此針對一般格式的表格文件，多數研究都利用表格中水平分界線與垂直分界線間之相互關聯性，進行表格結構的切割處理，以定義表格的幾何結構。例如：以水平與垂直分界線間所圍成之區域，作為表格中區塊切割的依據[10,11,12,16,18,19,20,24]。Diligenti 等人[17]則是以表格中的水平及垂直的空白區域，作為分隔文件中區塊的依據。Watanabe 等人[14,15]提出以水平及垂直方向切割點為切割依據的方法，分別針對整體結構特性、區域性結構特性、及區塊間邏輯結構關係來分析表格結構。

然而僅以表格中線段關係或區塊分佈來表示的結構特性，無法滿足於表單文件處理中，填寫欄位及非填寫欄位擷取時，特殊結構物件的分析。因此不同於 Top-down 的切割方式，Bottom-up 的方式則是先擷取影像中最基本的單位像素元素，將其作適當的合併，以便處理格式較複雜的表單影像。張貴雲[24]提出以 Run-based 演算法為基礎，藉由擷取表單影像中的 runs，並分析 runs 所構成的資料物件結構，以進行表單中填寫欄位的分類與擷取。

2.1.2 結構表示方式

表格結構的表示方式，大致上可分為矩陣形式及階層式樹狀結構[14,15,16,17]來表示。Watanabe 等人[14,15]分別建立三種二元樹，用以表示表格整體結構特性、區域性結構特性及區塊間邏輯結構的關係。Diligenti 等人[17]提出一種新的 Hidden Tree Markov Model 結構表示方式，將 XY-Tree 與 HMM 做結合，用以表示各區塊之間的邏輯結構關係。Amano 等人[18,19,20]則以 XML 語法為基礎，提出一種表格結構表示語法(TFML)，將切割出之區塊按照其類別作標記，並將區塊之間的邏輯結構關係以文件結構文法(Document Structure Grammar)[17]或階層式樹狀結構[19,20]表示之。而不同的結構表式方式，對於後續表單比對的效果，亦具有不同的影響，結構所能包含的資訊越多，比對之效果越佳。

2.1.3 區塊物件識別

經觀察現有的表單文件可以發現，表單文件已不再只是簡單的表格格式，多數表單文件會因其應用需要，設計成包含多種不同型態之資料物件的複合式表單。例如：方形、圓形核對框、虛線或齒狀間格等。因此 Busch 等人[13]及張貴雲[24]的研究，皆特別針對不同型態的欄位物件，分析每種物件的特徵，逐一設定判斷條件，如物件長寬、大小、形狀、封閉區域數等，將已擷取出之區塊作進一步的區別。而 Tabbone 等人[25]則是分析每個單一物件，在不同角度剖面上的投影量，並利用曲線圖表示之。進一步利用特定型態之物件會產生其特定曲線圖的特性，來區辨不同型態之物件。

2.1.4 相似度比對

在比對表格文件的相似度研究中，Duygulu 等人[10]提出以樹狀結構表示表單的幾何結構，以二維矩陣記錄樹中每一層之節點所含之子節點數，並根據不同表單間矩陣之差異，判斷表單之相似程度。Fan 等人[21]以水平及垂直線段所切割出之表格結構，分別以三個二維矩陣，記錄表單中的水平與垂直線的交叉關係、水平線間之距離及垂直線間之距離特徵，分別利用這些特徵差異值，進行相似度比對。Peng 等人[22]則以連通區塊為表單結構切割依據，並利用一維陣列形式之 Component Block List (CBL)，依區塊的位置排序後，串連儲存起來。並對每張表單影像建立 Component Block Tree(CBT)，以比較 CBT 的結構差異進行相似度的比較。Shimotsuji 等人[23]提出以二維的雜湊表(Hash Table)來記錄區塊的中心位置，將兩表單影像之雜湊表做比較，差異越小即相似度越高。

2.2 手寫資料處理

2.2.1 框線去除及破碎字修補

由於在擷取表單文件中的手寫資料時，常會因填寫人的書寫習慣，發生手寫筆劃與表單中的印刷線條重疊的情況。而目前市面上現有的部分字元辨識系統，針對手寫字的辨識效果仍很有限，若將重疊的線段和手寫字資料一併進行辨識，則容易因多餘的線段而降低其辨識率。因此，框線去除[27]及破碎字修補仍有其研究的必要性。

此研究領域的相關文獻中，Ye 等人[26]利用重疊部分的線段形狀，以不同大小的遮罩進行修補範圍的測量，並針對筆劃的一致性、平滑性、鄰近性等三種特性進行破碎筆劃的修補工作。Yu 等人[27]以 Run-based 演算法為基礎，判斷出表單影像中屬於重疊筆劃的框線範圍，將範圍內之 runs 去除，並將內插法應用於破碎區域的修補工作上，Chen 與 Tseng[28]則是針對去除的框線尋找切割點，依據切割點所在之 runs，進行鄰近 run 數的分析，根據不同狀況進行修補的工作。

2.2.2 光學字元辨識

現有的光學字元辨識研究，多數都僅針對單一種類文字進行研究，不同語言種類的文字，所採用之辨識方法亦不同[29,30,31,32,33,34]。由於實際文件中的文字資料，通常皆包含兩種以上的字體或語言文字，如印刷文字與使用者填入之手寫字體。然而受限於現有的技術，為了能進行多種類型文字的辨識工作，亦開始有研究針對文字類型的分類進行研究[29,35,36]，再將分類出之文字，交由適當的文字辨識系統進行辨識工作。

多數表單文件處理之相關研究，僅針對上述之部分領域，進行個別的研究，並無對各領域間的相關性，加以整合探討。因此本研究即期望整合上述表單之研究領域，針對現有之表單文件，提出對於一般簡單表單文件格式與複雜表單文件皆適合之填寫欄位擷取及手寫資料萃取的研究方法，將文件自動化中欄位擷取與資料萃取兩部分領域進行整合，以達到文件自動化處理之完整性。