# 第一章 緒論

## 1.1 研究動機與目的

隨著科技的進步，日常生活之各方面應用，皆朝向數位化及自動化的趨勢發展，以便利人們的生活。長久以來，大多數人皆習慣使用紙類文件記載資料，而在各公家機關及民間機構中，對於各種重要資訊的收集、記錄、搜尋及保存，又以表單文件為最普遍的文件格式。然而紙張來源取得固然容易，長久保存卻相當地困難，紙類文件容易因遺失或損壞，造成重要資料的流失。此外，現今的人力及空間，之於與日俱增的文件資料數量，實為有限。因此近年來，為了避免重要文件資料的流失，完善保存及管理數量龐大的文件資料，同時亦能提高各機構辦公室的工作效率，節省人力及空間資源，因此推廣文件數位化與自動化處理，已成為各機構提昇工作效率及改善文件保存方式之趨勢。

由於各行業在業務性質上的差異，使得對於自動化文件處理的需求也各有不同。例如郵政單位為提高信件遞送過程的效率，需要透過辨識郵遞區號或信件上之地址以加速信件的分類；而學校機關在學生事務上，為有效建立學生個人資料，須事先對各種學務表格欄位進行資料擷取及辨識，並存入資料庫中，以供日後修正、查詢或學生事務上的管理之用；針對一般公家單位的公文自動化處理，除了要擷取出公文的內容外，仍須將公文上之簽名及職章等資訊，完善地保存於電子文件中，以確保日後公文的效力及追朔未來行政責任之用。

因此為了滿足如此多變的應用需求，自動化文件處理流程即須整合包括文件數位化、自動化文件辨識、自動化文件資料擷取、光學字元辨識( OCR )、文件影像壓縮、資料庫儲存等技術。而其中以自動化文件辨識、自動化文件資料擷取及光學字元辨識等步驟最為重要。

另一方面，為因應各公家機關及民間機構的不同應用需求，文件類型亦是琳瑯滿目，如表單文件、報章雜誌、統計圖表及工程圖表等。對於各種重要資訊的收集、記錄、搜尋及保存，因表單文件具有完整的資料整合、歸納及分類等特性，故表單文件即成為各機構單位最為廣泛使用之文件格式。由於表單文件的重要性，正確儲存表單文件中所記載之資料，其重要性亦相對地提高，因此在表單文件自動化的處理過程中，表單資料欄位辨識、表單資料萃取及辨識等技術，在目前的文件自動化處理研究中，有著舉足輕重的地位。

目前多數機構所使用的文件處理系統，皆會設置文件資料庫，然而在儲存文件中的資料時，仍需人工自行輸入資料至資料庫中，以建立檔案。因此在整個文件處理的過程中，加強已知與未知文件的自動比對技術，以避免耗時的人工輸入過程，亦為相當重要的一環。

此外，在文件自動化處理中，最重要的目的在於擷取出文件中的資料。由於文件資料中包含各式應用所需的全部資訊，因此文件資料辨識的準確度，會直接的影響其使用及管理資料的應用。而表單文件中的資料，多是以填寫者手寫填入的形式，因此針對資料萃取後，可能降低資料辨識之正確率等因素，如表單框線

與手寫筆劃的重疊，或去除框線後造成的斷裂筆劃等，亦為重要之研究議題。

本研究期望能夠廣泛地針對不同類型之表單文件，整合表單手寫資料欄位擷取及表單手寫資料萃取等兩方面，提出準確有效的解決方法。

## 1.2 研究範圍與限制

表單資料擷取之流程可分為兩個部分。第一：分析未填寫之空白表單格式，判斷需使用者填寫之欄位，並記錄填寫欄及表單架構的相關資訊。第二：根據所記錄的相關資訊，與使用者已填寫好之表單進行比對處理，進一步取出表單中的手寫資料，以圖像方式儲存，可供未來字元辨識之用。

其中，正確地找出表單中的填寫欄位及取得完整的填寫資料，是自動數位化表單資料工作中十分重要的一環，對於未來手寫資料辨識的正確性，也有密不可分的影響。

本研究的目的在整合上述兩項工作，針對表單文件的自動化處理，尋求更正確且有效的解決方法。為求研究的順利進行，設定了以下的限制：

一、本研究所討論之表單影像的傾斜角度範圍，介於 0 至 5 度之間。

二、根據表單功能的不同，表單中之資料物件亦有不同之呈現方式。本研究為了能適應各種常用的表單格式，根據所處理之表單，歸納出表格、實橫線段、虛橫線段、圓圈、方格等五種常用資料物件形式，在本研究中則針對上述資料物件進行處理。

三、為顧及系統執行效率，且不影響表單品質的情況下，本研究所使用之輸入影像品質，為掃描解析度 100dpi 之 JEPG 格式表單影像。

四、為考量手寫資料擷取之精確性，本研究採用背景簡單之表單文件，避免因顏色較深之背景經二元化後，與手寫資料產生混淆。

## 1.3 論文結構

本論文共分為六章。第二章中探討目前在文件影像分析領域與手寫資料辨識領域中，相關之研究文獻。第三章為本研究之系統簡介。第四章中將探討本研究所使用之方法與步驟。第五章分析實驗結果與討論。最後第六章為結論與未來研究探討。