第五章 實驗結果與討論

本研究針對表單文件提出表單文件手寫欄位擷取與手寫資料擷取之方法，本章將利用不同格式之表單文件，透過實驗來驗證其可行性。本章共分三節討論，5.1 節說明實驗資料來源；5.2 節將分別針對簡單格式表單文件與複雜格式表單文件，對所提出之方法進行實驗及實驗結果與討論；5.3 節為實驗結果之總結。

5.1 實驗資料來源

本研究所使用之文件影像來源分為兩種類別，依照表單文件格式分成簡單格式與複雜格式之表單文件，並設計兩組實驗，分別以不同格式類型之表單文件為實驗對象，每組實驗皆分成手寫欄位分類、擷取與手寫資料擷取兩階段。

第一組實驗對象為簡單格式類別之問卷式表單文件。第一階段採用臺灣師範大學 95 學年度教學評鑑問卷表為空白表單樣本，進行手寫欄位擷取試驗。第二階段採用該學年度 231 門課程，共 9730 張回填問卷，為已填寫表單樣品，供手寫資料擷取實驗之用。所有表單文件皆為 A4 大小，掃描解析度為 100dpi 之 JPEG 格式影像。

第二組實驗對象則為格式複雜的各式表單文件。第一階段是以 18 張類型、用途不同之空白表單影像為樣本，表單內文包含了中、英、日三種文字(詳見附錄)。而第二階段則選擇表單編號 3,4,5,6,8,9 共 6 張表單文件，各填寫 3 份，共計 18 張 A4 大小，掃描解析度為 100dpi 之 JPEG 格式影像為已填寫之輸入樣本。表單編號與名稱詳列於表 5.1。

表 5.1　實驗之表單文件列表

表單文件編號	表單名稱
1	中華民國簽證申請表
2	中華民國普通護照申請書 (1)
3	中華民國普通護照申請書 (2)
4	中文姓名申請書
5	Application for a Visitor’s Visa To New Zealand Health Details
6	Application Form for Sponsorship Letter
7	Application for Residence in New Zealand (1)
8	Application for Residence in New Zealand (2)
9	Application for Residence in New Zealand (3)
10	Application for Residence in New Zealand (4)
11	Application Form for a Student's Pass (1)
12	Application Form for a Student's Pass (2)
13	Application Form for 2008 Panasonic Scholarship (1)
14	Application Form for 2008 Panasonic Scholarship (2)
15	Application Form for 2008 Panasonic Scholarship (3)
16	Application Form for 2008 Panasonic Scholarship (4)
17	Application Form for 2008 Panasonic Scholarship (5)
18	Application Form for 2008 Panasonic Scholarship (6)

5.2 實驗驗證

根據表單格式的複雜度，採用不同類型之表單文件為影像樣本，分成兩組實驗進行手寫欄位擷取與手寫資料萃取之探討，並採用 5.1 式 Precision 值及 5.2 式 Recall 值來評估資料物件擷取之辨識率與精確率。其中 $N_{correct}$ 為分類正確之物件數，N_{false} 為分類錯誤之物件數，N_{miss} 為分類錯誤中未辨識出之物件數，$N_{redundant}$ 為分類錯誤中辨識出之非填寫欄位物件數。

$$precision = \frac{N_{correct}}{N_{correct} + N_{false}} \tag{5.1}$$

$$recall = \frac{N_{correct}}{N_{correct} - N_{miss} + N_{redundant}} \tag{5.2}$$

本實驗亦針對手寫欄位擷取中不同類型欄位物件，所發生之擷取錯誤情況，定義如表 5.2 所列之欄位擷取錯誤類型，以供對照與討論之用。後續將分別對兩組實驗進行說明及分析，透過所求得之正確率來驗證研究方法的可行性。

表 5.2　手寫欄位擷取錯誤類型表

填寫欄位類型 / 錯誤類型	橫直線 (A)	虛線 (B)	核對框 (C)	方格 (D)	表格 (E)
視填寫欄位物件為非填寫欄位物件	A1	B1	C1	D1	E1
視非填寫欄位物件為填寫欄位物件	A2	B2	C2	D2	E2
因連通物件斷線或相連造成之誤判	A3	B3	C3	D3	E3
因掃描導致表單影像部分錯位造成之誤判	A4	B4	C4	D4	E4

5.2.1 實驗一、簡單結構之問卷式表單文件處理

實驗目的：

驗證本研究所提出之手寫欄位分類擷取與手寫資料擷取方法，對於結構簡單之表單文件，能有效且正確地擷取出手寫欄位區域以及填入之手寫資料。

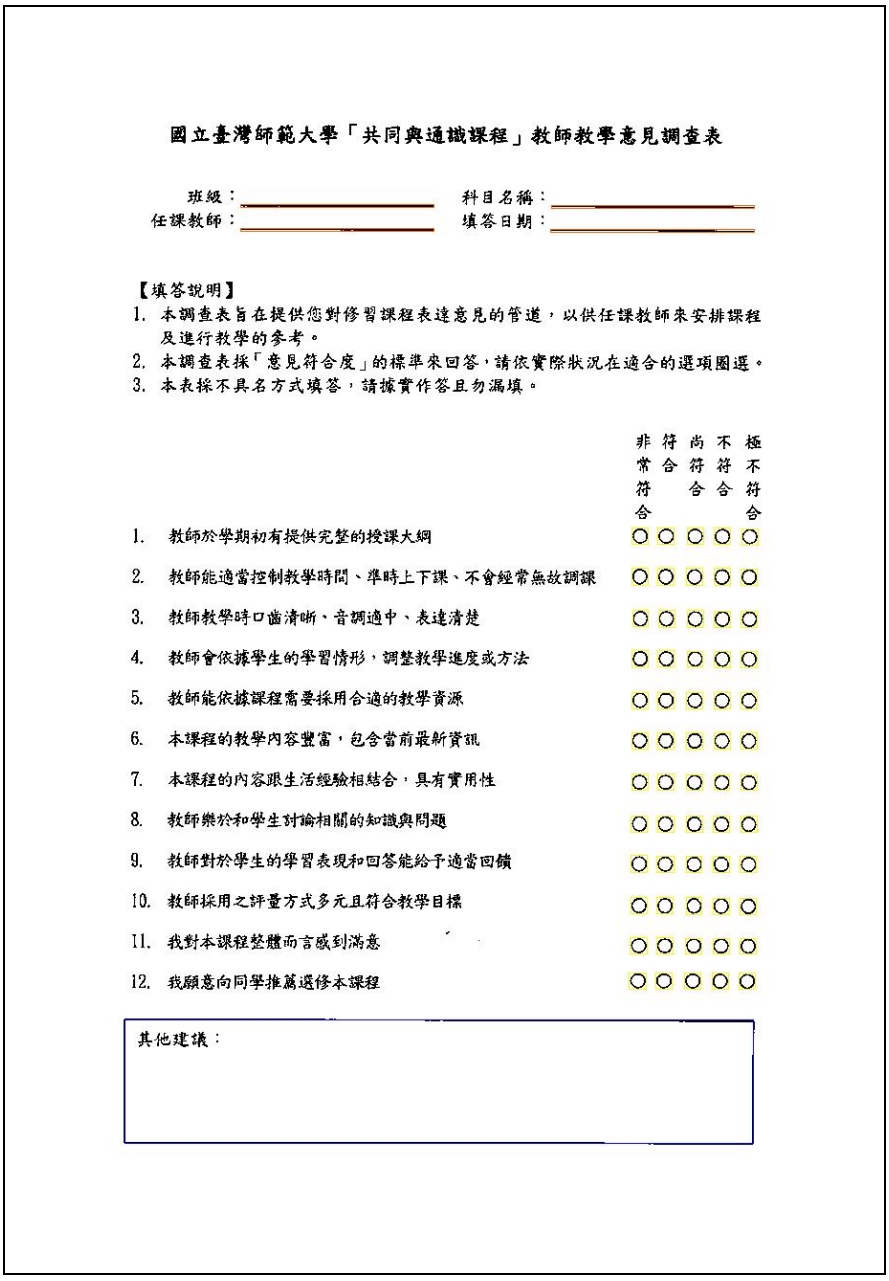
實驗結果：

本實驗主要分成兩個階段：手寫欄位分類擷取與手寫資料擷取。在手寫欄位擷取階段中，本實驗採用臺灣師範大學 95 學年度教學評鑑問卷表為空白表單樣本。此表單格式為五選一式問卷表單，共有 675 個物件，當中包含了三類手寫欄位物件：4 個橫直線物件、60 個核對框物件、及 1 個方格物件共 65 個手寫欄位物件。其中並無結構複雜的表格欄位物件，屬於結構較為單純之表單文件。表 5.3 為第一階段手寫欄位分類、擷取之結果，針對表單文件中所包含的三類欄位物件，皆可達到 100%的辨識率。如圖 5.1 所示，紅色框區域表示擷取出之橫直線欄位區域，黃色框為核對框欄位區域，而藍色框即方格欄位區域。

在第二階段手寫資料擷取實驗中，分成已填寫表單之手寫欄位擷取與手寫資料之框線去除及破碎字修補兩部分。本實驗以 95 學年度 231 門課程共 9730 張回填評鑑表，為已填寫表單之影像樣品。表 5.4 為透過表單比對後，已填寫表單中手寫欄位擷取之結果。

表 5.3　簡單結構之空白表單文件手寫欄位擷取結果

手寫欄位物件類型	手寫欄位物件數量	Precision (%)	Recall (%)	擷取時間 (s/張)
橫直線	4	100	100	
核對框	60	100	100	
方格	1	100	100	
	65	100	100	63.1



國立臺灣師範大學「共同與通識課程」教師教學意見調查表

班級：＿＿＿＿＿＿＿＿　科目名稱：＿＿＿＿＿＿＿＿
任課教師：＿＿＿＿＿＿＿＿　填答日期：＿＿＿＿＿＿＿＿

【填答說明】
1. 本調查表旨在提供您對修習課程表達意見的管道，以供任課教師來安排課程及進行教學的參考。
2. 本調查表採「意見符合度」的標準來回答，請依實際狀況在適合的選項圈選。
3. 本表採不具名方式填答，請據實作答且勿漏填。

	非常符合	符合	尚符合	不符合	極不符合
1. 教師於學期初有提供完整的授課大綱	○	○	○	○	○
2. 教師能適當控制教學時間、準時上下課、不會經常無故調課	○	○	○	○	○
3. 教師教學時口齒清晰、音調適中、表達清楚	○	○	○	○	○
4. 教師會依據學生的學習情形，調整教學進度或方法	○	○	○	○	○
5. 教師能依據課程需要採用合適的教學資源	○	○	○	○	○
6. 本課程的教學內容豐富，包含當前最新資訊	○	○	○	○	○
7. 本課程的內容跟生活經驗相結合，具有實用性	○	○	○	○	○
8. 教師樂於和學生討論相關的知識與問題	○	○	○	○	○
9. 教師對於學生的學習表現和回答能給予適當回饋	○	○	○	○	○
10. 教師採用之評量方式多元且符合教學目標	○	○	○	○	○
11. 我對本課程整體而言感到滿意	○	○	○	○	○
12. 我願意向同學推薦選修本課程	○	○	○	○	○

其他建議：

圖 5.1 實驗一之空白表單文件手寫欄位擷取結果

紅色表橫直線物件，黃色框表核對框物件，藍色框為方格物件。

表 5.4　回填教學評鑑表單之手寫欄位擷取結果

手寫欄位物件類型	手寫欄位物件總數	Precision (%)	Recall (%)	錯誤類型 (%)	擷取時間 (s/張)
橫直線	37720	100	100		
核對框	565800	99.71	100	C4	
方格	9430	99.32	100	D4	
	612950	99.67	100		1.712

在手寫資料框線去除及破碎筆劃修補部分，本實驗自 231 門課程中選擇 5 門課共 95 張回填表單為實驗對象。透過表單比對擷取出已填寫表單中之手寫欄位後，針對橫直線類之手寫欄位區域，共計擷取出 311 筆手寫資料，自其中取出 423 筆與框線相交之手寫資料，進行框線去除及破碎筆劃修補。共計 1544 個手寫字元，錯誤修補字元有 135 個，修補正確率達 91.26% (正確修補字元/總字元數)。所謂正確修補乃指可將框線去除後所造成之斷裂筆劃連接，且不誤連原本即不相連之筆劃。以圖 5.2 為例，(a)為填寫表單影像中，框線未去除前所擷取出之 3 筆手寫資料，而(b)經過框線去除及筆劃修補後之手寫資料影像。

(a)

(b)

圖 5.2 手寫資料之框線去除與破碎字修補結果

(a)框線去除前　(b)去除框線並修補破碎字

實驗結果討論：

由實驗一之手寫欄位擷取階段之實驗結果可知，本研究所提出之表單手寫欄位分類與擷取方法，對於格式簡單之空白表單文件，能相當正確地擷取出表單文件中的手寫欄位物件，以供後續手寫資料擷取之用。

在第二階段手寫資料擷取實驗中，根據表 5.4 可知，經過與空白表單比對後，填寫表單中之橫直線類欄位物件之所在，皆可相當正確地擷取出來。針對其他兩類欄位物件，亦可達到 99%以上之擷取正確率。由於本階段實驗之 9430 張輸入表單影像中，因掃描時表單放置不當，導致其中 23 張輸入表單影像中，文件下半部影像發生錯位及傾斜不均的情況，2 張輸入表單部分影像遺失，以及 3 張表單影像傾斜角度過大(>10°)，共計 28 張錯誤輸入表單影像，以致在表單比對時，無法正確判定出填寫表單中手寫欄位之位置，而造成位於表單偏下半部分之核對框欄位，及方格欄位擷取之錯誤情況，即錯誤類型 C4 與 D4，進而影響到擷取之正確率。

在框線去除與破碎字修補時，造成的修補錯誤的主要原因，如圖 5.2 (b)之第三圖中，由於字元「6」其手寫筆劃與框線重疊處幾乎重合，導致框線下方無明顯筆劃像素可供連結修復，若保留此種情況之框線處，則必使得所有非相交區段皆被保留，而造成更大的錯誤。

5.2.2 實驗二、複雜結構之複合式表單文件處理

實驗目的：

驗證本研究所提出之手寫欄位分類擷取與手寫資料擷取方法，對於結構複雜，不同類型與用途之複合式表單文件，亦能正確地擷取出手寫欄位區域以及填入之手寫資料。

第一階段實驗結果：

本實驗同樣分成手寫欄位分類擷取與手寫資料擷取兩階段。在手寫欄位擷取階段中，採用 18 張不同類型及用途之複合式表單文件，作為空白表單影像樣本。在所採用之空白表單樣本中，含括了本研究中定義之五大類常見手寫欄位物件：橫直線(A)、橫虛線(B)、核對框(C)、方格(D)及表格(E)物件等。而本階段將針對每一張空白表單文件，分析其手寫欄位物件擷取之正確率，以驗證本研究所提出之手寫欄位分類、擷取方法，對於複合式結構之表單文件，亦能達到相當之成效。表 5.5 中分別列出每張表單文件中所有連通成份物件數、手寫欄位物件數與類型，及手寫欄位擷取之結果，其中錯誤類型代碼則對應至表 5.2 所定義之錯誤類型。

表 5.5　複雜結構之空白表單文件手寫欄位擷取結果

表單編號	表單物件總數	手寫欄位物件數	手寫欄位物件類型	Precision (%)	Recall (%)	錯誤類型
1	2291	30	A,B,C	98.82	98.69	C2,C3
2	1066	115	C,D,E	98.97	99.62	C3,E1,E2
3	1690	27	A,C,E	99.23	99.35	A2,C3,E2
4	1557	24	A,B	99.42	99.48	A2,B1,C2
5	1550	31	A,C,D,E	99.87	99.94	C2,C3
6	2105	18	A	99.76	99.53	C2
7	1926	40	C,D,E	99.79	99.89	A2,E1
8	1879	45	C,D,E	99.95	100	C3
9	1341	118	E	100	100	
10	1610	73	C,D,E	100	100	
11	1202	66	A,B,C,D,E	97.00	94.34	B1,C2,C3
12	1485	17	B,C,D,E	97.64	96.28	C2,C3,E1
13	925	33	A,C,D,E	98.16	96.80	E2,*
14	724	75	A,D,E	97.79	97.25	E2
15	1054	24	A,C,D,E	98.39	97.01	A2,C2,E1,E2
16	373	7	A,D,E	99.46	99.46	A3,E2
17	400	8	D,E	99.25	99.50	E2,*
18	259	11	D,E	99.61	99.23	E2
	23437	762		99.06	98.69	

*其他錯誤類型　表示未擷取出欄位物件外之印刷文字間所包含之手寫區域

第一階段實驗結果討論：

由表 5.5 可知，針對結構複雜之各種複合式表單文件，手寫欄位擷取之平均辨識率為 99.06%，平均精確率亦可達 98.96 %，表示大部分表單物件皆能被正確辨識與分類，進而擷取出手寫欄位物件，驗證本研究所提出之表單手寫欄位分類與擷取方法，對於複合式表單文件之可行性。針對表 5.5 中造成欄位擷取錯誤之錯誤類型，下面將分別探討不同欄位物件類型之錯誤發生原因。

一、橫直線欄位物件

根據表 5.5 中歸納得知，主要造成橫直線欄位擷取錯誤之錯誤類型有兩類：說明文字底線誤判為手寫欄位(A2)，以及因與說明文字筆劃相連而未辨識出之錯誤類型(A3)。造成 A2 錯誤類型的原因，乃因說明文字底線上之說明文字分佈，不符合大多為置中分佈的假設，而導致底線上方包含了較大範圍的空白區域，而誤判為橫直線手寫欄位，如圖 5.3(a)所示，紅色框區域即誤判為橫直線之說明底線。A3 錯誤類型發生的原因，則因底線與上方說明文字之部分筆劃相連，導致於表單物件分類時，誤判為其他類型之表單物件，如圖 5.3(b)，字母 g 底部與橫直線相連。

(a)

(b)

圖 5.3　橫直線欄位物件擷取錯誤類型

(a)A2 錯誤類型，紅色框表示誤判為橫直線之底線。

(b)A3 錯誤類型

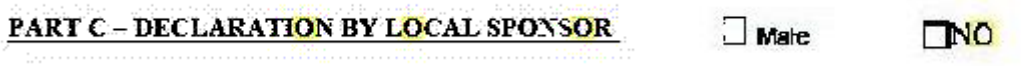
二、橫虛線欄位物件

由表 5.5 可知，橫虛線欄位擷取錯誤類型共計一類，即視橫虛線欄位物件為非填寫欄位物件(B1)。造成 B1 錯誤類型之原因，乃因表單文件中能構成橫虛線之點物件，所呈現之物件結構特徵與直線物件之結構特徵相似，不符合物件結構特徵所定義之條件，導致表單物件分類時，無法正確分類至點物件類，進而影響到後續虛線重組之正確性，以致未擷取出橫虛線手寫欄位所在。

三、核對框欄位物件

核對框欄位擷取錯誤類型可分為兩類。第一類即誤將印刷字判定為核對框欄位(C2)，造成此類錯誤發生之原因，乃因印刷文字中英文字母 D、O 與中文字中「口」等印刷文字物件，所呈現之物件結構特徵與封閉區域特徵皆與核對框物件相同，導致誤判，如圖 5.4(a)，黃色框區域即誤判為核對框之物件。

第二類錯誤類型則分為兩種情況。第一種情況即因表單文件經掃描與二元化處理後，造成印刷筆劃斷裂，導致核對框欄位物件結構不完整。第二種情況則因核對框欄位邊緣與印刷字部分筆劃相連。上述兩種情況皆會造成無法於表單物件分類階段，正確分析核對框之結構特徵與封閉區域數，以致誤判為非填寫欄位(C3)。圖 5.4(b)為框線斷裂之情況，(c)則為與印刷字相連之情況。



(a) (b) (c)

圖 5.4 核對框欄位物件擷取錯誤類型 (a)C2 錯誤類型，黃色框為誤判為。(b)(c)C3 錯誤類型

四、方格欄位物件

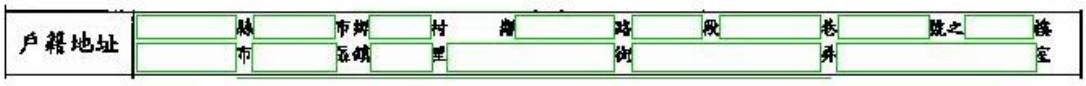
於本實驗中，無論是空白方格填寫欄位或是內含印刷文字之外框方格物件，本實驗皆能正確地進行分類與擷取，因此針對此類欄位物件並無造成任何分類及擷取上之錯誤。

五、表格欄位物件

表格欄位物件擷取錯誤類型，主要分為未擷取出之填寫區域(E1)，與視說明文字列空白區域為填寫區域(E2)兩類型。由於本研究為能擷取出所有的空白填寫

區域，將表格物件所細分出之非空填寫欄位區域，根據其中所包含之印刷文字列數，進一步縮小填寫欄位之判別範圍，針對每一列印刷文字與空白區域範圍，擷取其中所含之填寫區域。

因此，造成 E1 錯誤類型之原因，在於當印刷文字列中所包含的可填寫之空白區域大小，小於設定之門檻值時，則會被誤判為非填寫欄位，如圖 5.5(a)，「村」與「鄰」間之空白間距即誤判為非填寫欄位之區域。反之，當印刷文字分佈，不符合說明文字置中且平均分佈之假設，且所包含之空白區域範圍亦大過門檻值時，則會將非填寫空白區域誤判為填寫欄位，即 E2 錯誤類型，如圖 5.5(b)，上半部表格欄位內之綠色框即誤判為填寫欄位之非填寫區域，而下半部表格欄位即辨識正確之填寫欄位區域。



(a)

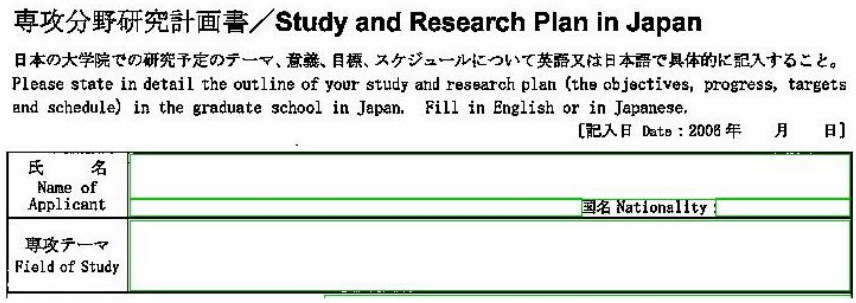
(b)

圖 5.5 表格欄位物件擷取錯誤類型

(a) E1 錯誤類型　(b)E2 錯誤類型，綠色框為辨識出之填寫欄位。

六、其他錯誤類型

由於本實驗所採用之編號 15 與編號 19 之表單文件中，在非欄位物件之印刷文字間，亦包含了手寫欄位區域。如圖 5.6 所示，在表格上方之印刷文字間，提供了可填寫年、月、日等日期資訊之手寫區域。由於本研究並未針對此類型填寫區域進行分類與擷取，固造成手寫區域擷取錯誤。


専攻分野研究計画書／Study and Research Plan in Japan

日本の大学院での研究予定のテーマ、意義、目標、スケジュールについて英語又は日本語で具体的に記入すること。
Please state in detail the outline of your study and research plan (the objectives, progress, targets and schedule) in the graduate school in Japan. Fill in English or in Japanese.

〔記入日 Date：2006年　月　日〕

氏　名 Name of Applicant	国名 Nationality：
専攻テーマ Field of Study	

圖 5.6 其他錯誤類型

第二階段實驗結果：

本實驗於手寫資料擷取階段中，採用空白表單其中 6 張表單文件，各填寫 3 份共 18 張已填寫之表單文件為輸入影像樣本。經過與空白表單文件比對後，擷取出 336 筆手寫資料，共計 2409 個字元。其中 512 個字元發生與框線相交之情況。本實驗在進行框線去除與破碎筆劃修補後，得 113 個錯誤修補字元，筆劃修補正確率為 95.31%。

5.3 總結

一般而言，表單中的手寫欄位是以橫直線、橫虛線、核對框(包含圓圈與方格)、方格及表格等形式呈現。在 5.1 節的實驗中，針對上述五種手寫欄位物件之分類與擷取，以及手寫欄位資料擷取與修補之方法，分別設計了兩組實驗探討手寫欄位擷取與手寫資料擷取之準確性。根據實驗結果，我們利用表單物件本身的結構特徵進行欄位物件分類與擷取，能相當穩定將物件分類並擷取出，並不會受到表單文件本身格式或內文的差異，而對結果造成影響。在實驗二，即使於表格物件中因擷取出多餘之說明欄位填寫區域，而降低整體辨識率，卻也因此降低了錯誤類型 E1，即填寫欄位區域未辨識出之錯誤機率，且由所得到之精確率值，表示欄位擷取之結果，確實有相當不錯之正確性。

在手寫資料擷取實驗階段，除實驗一中因掃描時的人為因素導致輸入表單影像內容變異的問題，造成手寫欄位經比對擷取後，欄位位置擷取錯誤之情況外，大部分已填寫之輸入表單與空白表單比對後，皆能正確取得欄位物件中之手寫資料。而根據兩組實驗第二階段所得之結果，利用內插法來還原框線去除後之相交處筆劃，對於所擷取出之手寫資料，皆能夠準確地修復大部分常見之手寫筆劃與框線相交之情況。

本實驗之輸出資料，包括每一張表單文件中，所擷取出之手寫資料及手寫欄位資訊兩部分，並以影像方式儲存手寫資料於資料庫，以供未來進行表單比對，光學字元辨識及手寫資料管理之用。