Performance of OCR

Author: Roger Dunham

Date: 8th July 2019

Executive Summary

SolidOCR was compared with Acrobat and ReadIRIS to evaluate the quality and speed of reconstructing PDFs into editable Word Documents.

All three options resulted in generally high quality reconstruction, though some character recognition errors occurred in each case.

In all cases, serif fonts were slower to identify than sans-serif fonts.

SolidOCR offered significantly faster reconstruction per page, and the differential between serif and sans-serif was considerably smaller than for the other OCR options.

Objectives

This document aims to compare the output of three OCR engines when converting the same PDF to Word. The OCR engines are:

- SolidOCR – via JobHandler in the SolidConverter-Jobs sample using Solid Framework 9544
- ReadIris 17
- Acrobat (which uses IRIS 15 – see C:\Program Files (x86)\Adobe\Acrobat DC\Acrobat\plug_ins\PaperCapture\iDRS15)

A comparison will be made of both accuracy of reconstruction, and the time taken for reconstruction.

Source Documents

The source documents are PDFs that were created from scanned images. They do not contain a searchable text layer.

SHA1 of PDF	Number of pages	Serif/ Sans-serif	Type of document	Comments
6938992d5c82606dc43937021f3680ef275d0f90[1]	10	Sans-serif	Report	Contains tables
031cb73c4af71f7ce1272996438f7a8b4333e500	16	Serif	Contract	Sloping text

[1] This file originated with Foxit. See case 21688.

Conversion Environment

Testing was performed on the following machine:

"Leopard" NUC

i7-6770HQ CPU running at 2.6GHz

32 GB RAM

Windows 10 Pro 64 bit

Language specification

Both test documents are English. SolidOCR supports "automatic" language detection. Acrobat and IRIS require the language to be specified.

Method for recording OCR performance

SolidOCR is tested using a JobHandler which reports the total time for reconstruction. This will include the OCR. Conversion is performed using both the 64 bit and 32 bit app. The 32 bit app has been marked as `/LARGEADDRESSAWARE,` allowing it to access 3GB of address space.

Acrobat. Time for OCR is extracted from the log file that is generated.

ReadIRIS. Time, recorded using a stopwatch, taken from a file being dropped into a "Watched" folder to the reconstructed file appearing. ReadIRIS allows the user to specify whether to focus on "Speed" or "Accuracy". Both values were tested. (In this case the quality of the image is quite high, so there is not likely to be much difference).

Snippets of document

6938992d5c82606dc43937021f3680ef275d0f90.pdf

Step 5 & 6: Objectives and Steps

1. Our objective is to provide equal employment opportunities for Black or African American men and women when our organization fills vacancies that become available.

a. The Recruiting Section has established an electronic data base of mid-Atlantic regional college placement officers and faculty advisors to periodically update these criminal justice program and other college/university program contacts on our hiring activities and internship opportunities.

b. The Recruiting Section will maintain its working partnership with the Fairfax County Public School's Criminal Justice instructional program. A mechanism has also been developed to ensure that program participants who show interest in our department at age 18 have a reliable point of contact for potential application after age 20.

c. Developed current social media outreach through our Public Information Office to include Fairfax County Police Department public website with recruitment information; Facebook; Twitter; and other public released information.

d. The Recruiting Section is staffed by at least one black officer to assist in outreach to black male and female potential candidates. Currently, the recruiting staff has a black female who is excellent at recruiting.

e. Recruiting Section will send at least one staff member to the periodic meetings of the Eastern Region Police Recruiters and Applicant Investigators Association. Regular attendance will ensure early identification of competitive agencies hiring activity and evolving personnel hiring practices. This will allow us to learn of the current best practices in recruitment of minorities.

f. Recruiting Section will market Fairfax County Police Department through all their venues as the local law enforcement career opportunity of choice in the Washington Metropolitan Area and the Middle Atlantic Region. Emphasis will be given to the resilience and projected growth of the Northern Virginia economy relative to other Washington Metropolitan Area and Middle Atlantic Region jurisdictions.

g. Recruiting Section and supplemental staff will further develop and maintain contacts with local and regional universities and colleges, as well as with military career transition offices and services, to promote career opportunities with our agency though personal visitation and as enabled by the adoption and utilization of new technologies. Recruiting Section staff have obtained multimedia capable laptops to enhance visitation presentations and to facilitate prospective applicants to apply online.

h. The Recruiting Director and other Command Staff members will continue to emphasize the Department's partnerships with the International Association of Chiefs of Police, the United States Department of State, American University, George Mason University, and the Major Cities Chiefs as indicators of the Police Department's ongoing commitment to improve the standards and delivery of police services nationally and internationally.

031cb73c4af71f7ce1272996438f7a8b4333e500.pdf

Unless the contrary intention appears or the context otherwise requires or admits:

2.2.1 A reference to this Agreement includes the Recitals and Schedules and Annexes.

2.2.2 Headings are for convenience of reference only and shall not affect the construction or interpretation of the provisions of this Agreement.

2.2.3 Any reference to a body corporate shall include the successors and permitted assigns of such body corporate.

2.2.4 Where the context so admits, any reference to the singular includes the plural, any reference to the plural includes the singular, and any reference to one gender includes all genders

2.2.5 Any reference herein to any sum of money or the symbol 'USD' shall be deemed to be a reference to the currency of the United States of America and the obligation to pay such sum shall be deemed to be an obligation to pay such moneys in cash or by bank cheque or by wire transfer in the currency of the United States of America.

2.2.6 As between the Company and each Subscriber, this Agreement shall be construed and interpreted as a separate and independent and severable subscription agreement.

2.3 Entire agreement

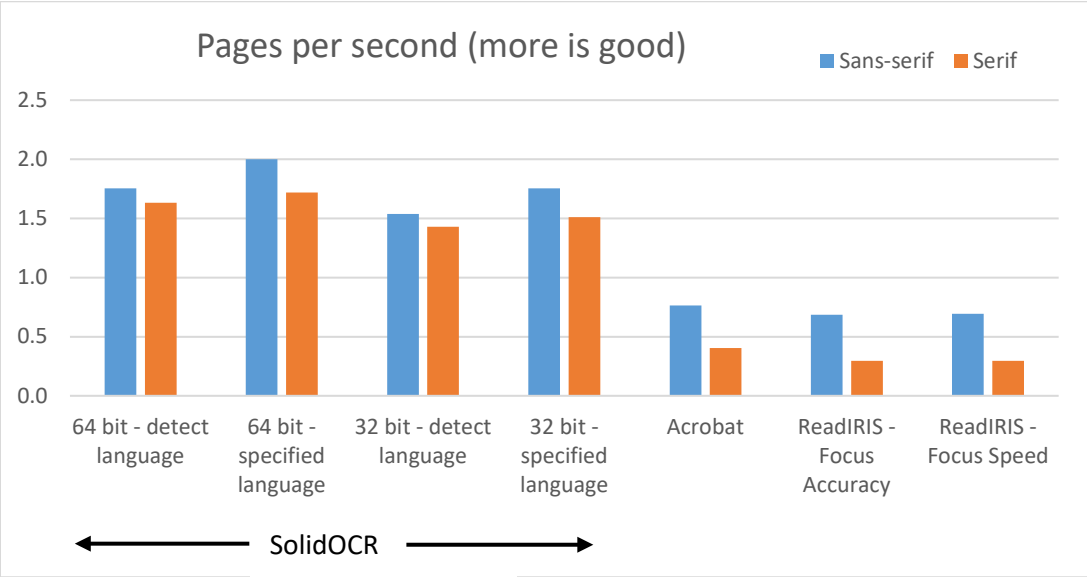
Performance Results

Table 1. Time taken to convert document

File	Solid Framework				Acrobat[2]	ReadIRIS[3]	
	64 bit		32 bit[4]		32 bit	32 bit	
	Language Automatic	Language specified	Language Automatic	Language specified	Language Specified	Focus on Accuracy	Focus on Speed
6938992d5c82606dc43937021f3680ef275d0f90.pdf	5.7	5.0	6.5	5.7	13.1	14.6	14.4
031cb73c4af71f7ce1272996438f7a8b4333e500.pdf	9.8	9.3	11.2	10.6	39.6	54.0	54.0

Table 2. Conversion rate pages per second

File	Solid Framework				Acrobat	ReadIRIS	
	64 bit		32 bit		32 bit	32 bit	
	Language Automatic	Language specified	Language Automatic	Language specified	Language Specified	Focus on Accuracy	Focus on Speed
6938992d5c82606dc43937021f3680ef275d0f90.pdf	1.8	2.0	1.5	1.8	0.8	0.7	0.7
031cb73c4af71f7ce1272996438f7a8b4333e500.pdf	1.6	1.7	1.4	1.5	0.4	0.3	0.3



[2] Based on logfile generated by SolidFramework for time </CPdfDocument::ProcessOCRPipeline ()>.
[3] Calculated by recording time taken for a docx file to appear after dropping a PDF into a watched folder. Recognition set to Accuracy rather than Speed.
[4] Conversion is performed using a JobHandler with Solid Framework 9544.

Quality Results

Sans-serif file

- SolidOCR, Acrobat and ReadIRIS all give extremely good results for text for the *majority* of the document

Table Handling

PDF

	%						
Utilization #/%	-17%	-4%	-6%	-0%	-5%	-0%	-1%
Skilled Craft							
Workforce #/%	1/100%	0/0%	0/0%	0/0%	0/0%	0/0%	0/0%
CLS #/%	15,100/42 %	12,410/35 %	4,005/11 %	80/0%	2,060/6%	0/0%	240/1%
Utilization #/%	58%	-35%	-11%	-0%	-6%	0%	-1%
Service/Maintenance							
Workforce #/%	0/	0/	0/	0/	0/	0/	0/
CLS #/%	14,855/18 %	15,305/18 %	7,365/9%	75/0%	5,390/6%	0/0%	250/0%
Utilization #/%							

Solid OCR

	%						
Utilization #/%	-17%	-4%	-6%	-0%	-5%	-0%	-1%
Skilled Craft							
Workforce #/%	1/100%	0/0%	0/0%	0/0%	0/0%	0/0%	0/0%
CLS #/%	15,100/42 %	12,410/35 %	4,005/11 %	80/0%	2,060/6%	0/0%	240/1%
Utilization #/%	58%	-35%	-11%	-0%	–6%	0%	-1%
Service/Maintenance							
Workforce #/%	0/	0/	0/	0/	0/	0/	0/
CLS #/%	14,855/18 %	15,305/18 %	7,365/9%	75/0%	5,390/6%	0/0%	250/0%
Utilization #/%							

Correct result

Acrobat

	%						
Utilization #f%	-17%	-4%	-6%	-0%	-5%	-0%	-1%
Skilled Craft							
Workforce #/%	1/100%	0/0%	Of0%	Of0%	Of0%	Of0%	Of0%
CLS#f%	15,100f42 %	12,410f35 %	4,005f11 %	80f0%	2,060f6%	Of0%	240f1%
Utilization #f%	58%	-35%	-11%	-0%	-6%	0%	-1%
Service/Maintenance							
Workforce #f%	0f	0f	0f	0f	0f	0f	0f
CLS#f%	14,855f18 %	15,305f18 %	7,365f9%	75f0%	5,390f6%	Of0%	250f0%
Utilization #f%							

“/” incorrectly converted to “f”

ReadIRIS

Utilization #/%	-17%	-4%	-6%	-0%	-5%	-0%	-1%
Skilled Craft							
Workforce #/%	1/100%	0/0%	0/0%	0/0%	0/0%	0/0%	0/0%
CLS#/%	15,100/42 %	12,410/35 %	4,005/11 %	80/0%	2,060/6%	0/0%	240/1%
Utilization #/%	58%	-35%	-11%	-0%	-6%	0%	-1%
Service/Maintenance							
Workforce #/%	0/	0/	0/	0/	0/	0/	0/
CLS#/%	14,855/18 %	15,305/18 %	7,365/9%	75/0%	5,390/6%	0/0%	250/0%
Utilization #/%							

Correct result

Text Where Descenders Intersect Cell Borders

PDF

Captains and 1st Lt.	
Workforce #/%	43/90%
2nd Lt's	
Workforce #/%	83/78%
Colonols, Lt. Col, Majors	
Workforce #/%	8/57%

Solid OCR

Captains and 1st Lt.	
Workforce #/%	43/90%
2nd Lt's	
Workforce #/%	83/78%
Colonols, Lt. Col, Majors	
Workforce #/%	8/57%

Correct result

Acrobat

Captains and 1st Lt.	
Workforce #/%	43/90%
2nd Lt's	
Workforce#/%	83/78%
Colonols, Lt. Col, Majors	
Workforce #/%	8/57%

Correct result

□

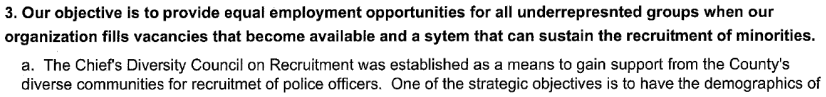
ReadIRIS

Caotains and 1st Lt.	
Workforce #/%	43/90%
2nd Lt's	
Workforce#/%	83/78%
Colonols, Lt. Col, Maiors	
Workforce #/%	8/57%

“Captain” and “Majors” incorrectly converted to “Capitains” and “Maiors”

Handling of Errors in the Original PDF

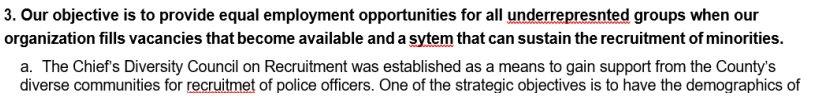
PDF



3. Our objective is to provide equal employment opportunities for all underrepresnted groups when our organization fills vacancies that become available and a sytem that can sustain the recruitment of minorities.

a. The Chief's Diversity Council on Recruitment was established as a means to gain support from the County's diverse communities for recruitmet of police officers. One of the strategic objectives is to have the demographics of

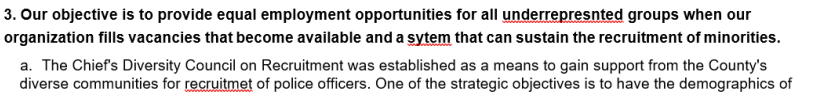
Solid OCR



3. Our objective is to provide equal employment opportunities for all underrepresnted groups when our organization fills vacancies that become available and a sytem that can sustain the recruitment of minorities.

a. The Chief's Diversity Council on Recruitment was established as a means to gain support from the County's diverse communities for recruitmet of police officers. One of the strategic objectives is to have the demographics of

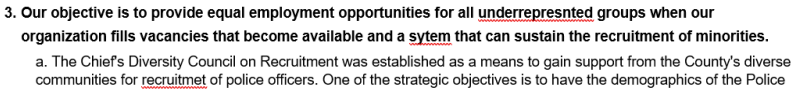
Acrobat



3. Our objective is to provide equal employment opportunities for all underrepresnted groups when our organization fills vacancies that become available and a sytem that can sustain the recruitment of minorities.

a. The Chief's Diversity Council on Recruitment was established as a means to gain support from the County's diverse communities for recruitmet of police officers. One of the strategic objectives is to have the demographics of

ReadIRIS



3. Our objective is to provide equal employment opportunities for all underrepresnted groups when our organization fills vacancies that become available and a sytem that can sustain the recruitment of minorities.

a. The Chief's Diversity Council on Recruitment was established as a means to gain support from the County's diverse communities for recruitmet of police officers. One of the strategic objectives is to have the demographics of the Police

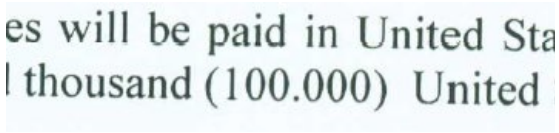
- All three tools correctly reproduce spelling errors in the original file. (This is a good thing, since those spelling mistakes could have been real technical words, and auto-correction would have made them difficult to identify).
- Note that ReadIRIS does not support having the exact text layout retained.

Serif file

- All three tools converted the text quite well for most pages
- All three tools made some errors on pages 14-16 (where text is sparse).

Identification of “1”

PDF



Solid OCR

res will be paid in United Stat
ed thousand (100.000) United

Correct result

Acrobat

res will be paid in United Sta
d thousand (I 00.000) United

“1” is interpreted as “I”

ReadIRIS

the Shares will be paid in Unite
thousand (I 00.000) United Sta

“1” is interpreted as “I”
(and layout is modified)

This problem occurs in multiple places in the document.

Identification of “USD”

PDF

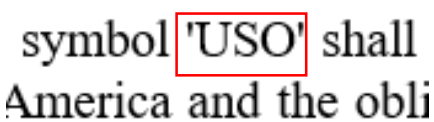
Solid OCR

Correct result

Acrobat

“USD” is not correct

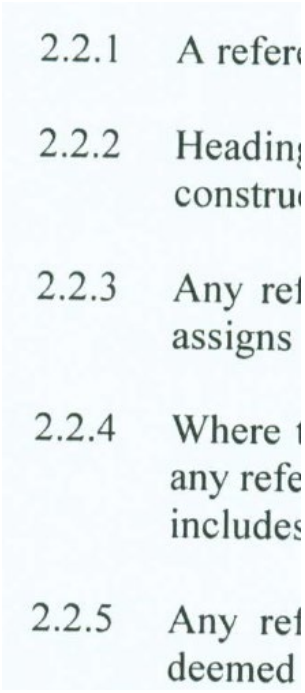
ReadIRIS



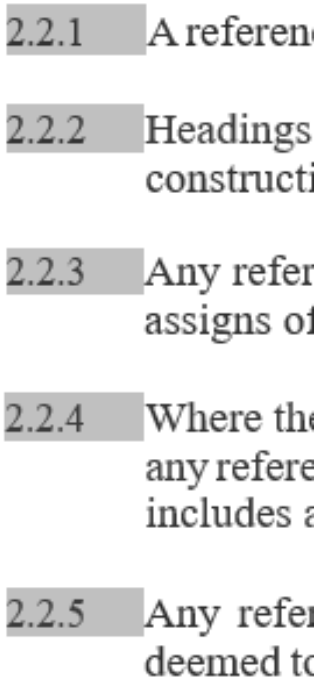
“USD” is not correct

Identification of List Items (Part 1)

PDF

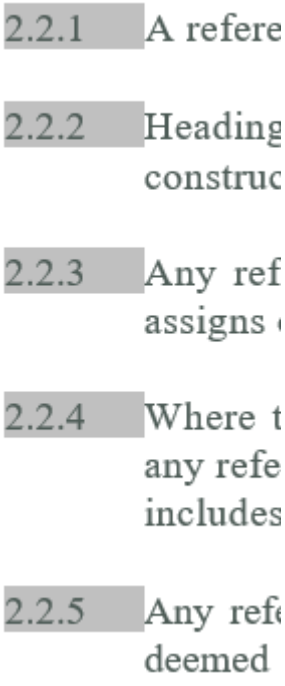


Solid OCR



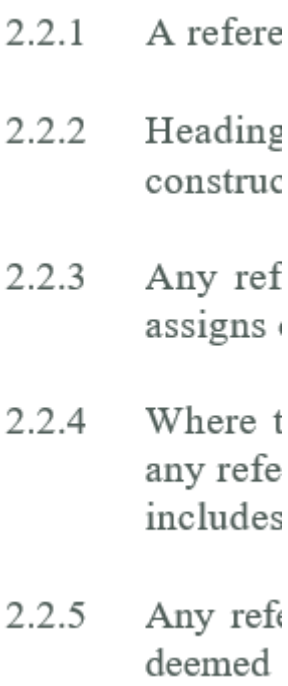
Conversion great!
List items interpreted correctly

Acrobat



Conversion great!
List items interpreted correctly

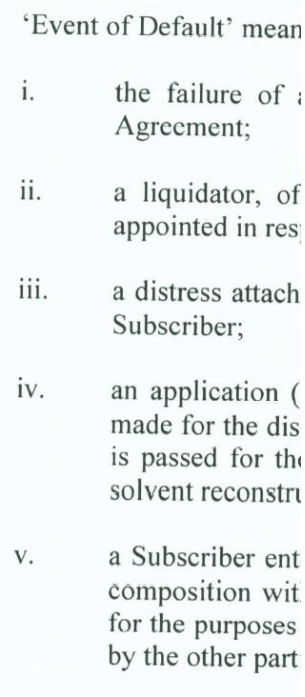
ReadIRIS



List items not detected

Identification of List Items (Part 2)

PDF


‘Event of Default’ mean

i. the failure of
Agreement;

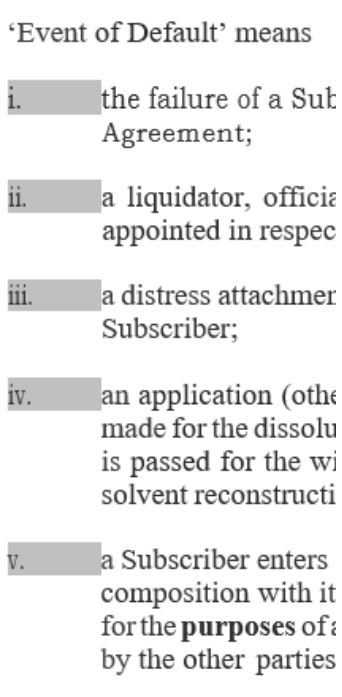
ii. a liquidator, of
appointed in res

iii. a distress attach
Subscriber;

iv. an application (
made for the dis
is passed for th
solvent reconstru

v. a Subscriber ent
composition wit
for the purposes
by the other part

Solid OCR


‘Event of Default’ means

i. the failure of a Sub
Agreement;

ii. a liquidator, officia
appointed in respec

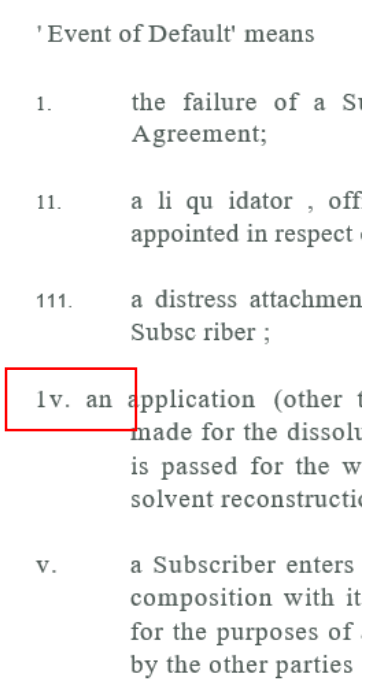
iii. a distress attachmer
Subscriber;

iv. an application (othe
made for the dissolu
is passed for the wi
solvent reconstructi

v. a Subscriber enters
composition with it
for the **purposes** of a
by the other parties

Conversion great!
List items interpreted correctly

Acrobat


' Event of Default' means

1. the failure of a Su
Agreement;

11. a li qu idator , off
appointed in respect

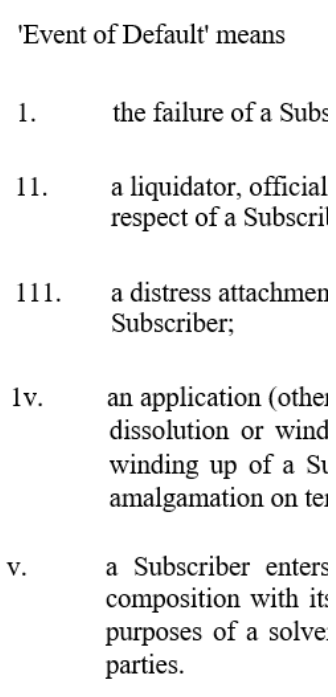
111. a distress attachmen
Subsc riber ;

1v. an application (other t
made for the dissolu
is passed for the w
solvent reconstructi

v. a Subscriber enters
composition with it
for the purposes of
by the other parties

Roman numerals misinterpreted causing list detection to fail

ReadIRIS


'Event of Default' means

1. the failure of a Subs

11. a liquidator, official
respect of a Subscril

111. a distress attachmen
Subscriber;

1v. an application (othe
dissolution or wind
winding up of a Su
amalgamation on te

v. a Subscriber enters
composition with its
purposes of a solve
parties.

Roman numerals misinterpreted