



Using Computer Vision to Identify and Extract Tables from Various File Formats

www.GenInvo.com

North America | UK | Europe | Asia



Kanishk Mair
Shashank Pagidimarri

Overview

1. Why Table Extraction?
2. Present State of Affairs
3. Approach Using Computer Vision
 - i. “File-to-Image” Conversion
 - ii. Pre-process Image for Extraction
 - iii. Computer Vision on Table Image
4. Challenges
5. Conclusions



www.GENINVO.com
North America | UK | Europe | India

Why Table Extraction?

- “Just for fun” ;-)
- Pulling data directly from the Protocol rather than re-entering, e.g.,
 - Protocol -> Trial Domain Datasets
 - Protocol -> CRF/DB specifications
- Porting tables from TFLs into CSR templates (embedded tables)
- Regenerating datasets from CSR Appendixes (Data Listings)
- However, extracting tables can be challenging considering various types of file formats

Present State of Affairs

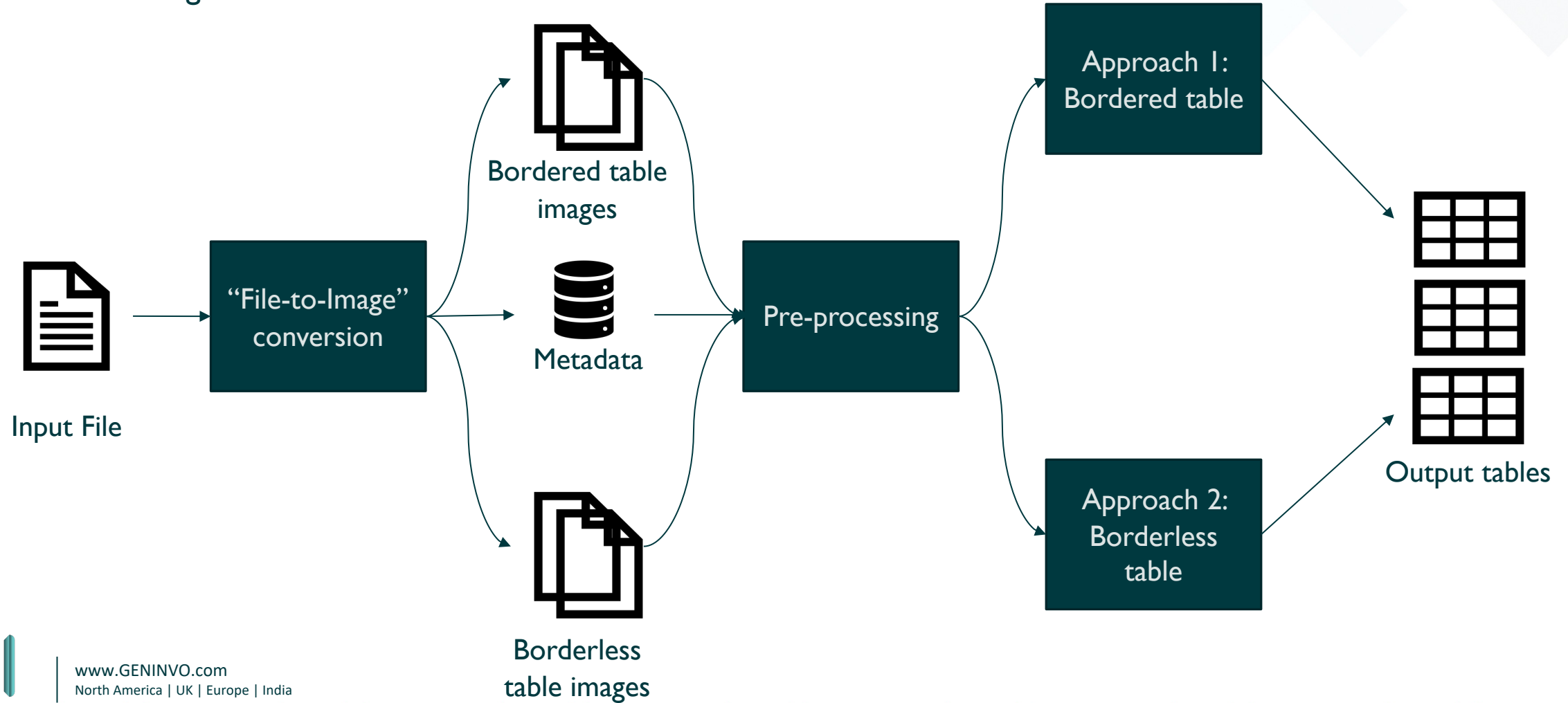
- Historically, study teams have manually reviewed the protocol and entered key information for their deliverables into specification sheets
- Programmers have manually reviewed the protocol to pull data needed for generation of Trial Domain datasets
- Medical Writers review “source tables” and sometimes copy/paste or regenerate those tables in the Clinical Study Report as embedded tables
- Thus, using automation for table extraction can:
 1. Prevent transcription errors
 2. Introduce time-savings



www.GENINVO.com
North America | UK | Europe | India

Approach Using Computer Vision

Process Flow Diagram of the Table Extraction Process...



“File-to-Image” Conversion

- For any given file, we first convert it to PDF to have a standard format.
- Next, all the pages containing a table are detected and saved as an image.
- The quality of the images saved is to be kept >200 dpi* for more accurate/consistent results by the Computer Vision (CV) process downstream.
- Each table’s structure is determined using a deep learning model called “CascadeTabNet”.
- This model identifies the coordinates of each table and whether the table is bordered or not.

* - Recommendation by PDF2IMG Package



www.GENINVO.com
North America | UK | Europe | India



“File-to-Image” Conversion (cont’d)

- Processed page from the PDF analyzed via the Cascade TabNet.
- The table metadata predicted is saved for subsequent processing.

Supplemental Investigations

If CR determination depends on a residual lesion that decreased in size but did not disappear completely, it is recommended the residual lesion be investigated with biopsy or fine needle aspirate. If no disease is identified, objective status is CR.

If progression determination depends on a lesion with an increase possibly due to necrosis, the lesion may be investigated with biopsy or fine needle aspirate to clarify status.

Objective/Subjective Progression

Patients requiring discontinuation of treatment without objective evidence of disease progression should not be reported as PD on tumor assessment CRFs. This should be indicated on the end of treatment CRF as off treatment due to Global Deterioration of Health Status. Every effort should be made to document objective progression even after discontinuation of treatment.

Table 11. Objective Response Status at each Evaluation

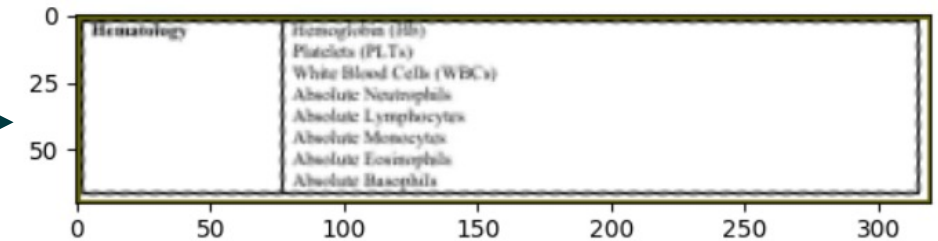
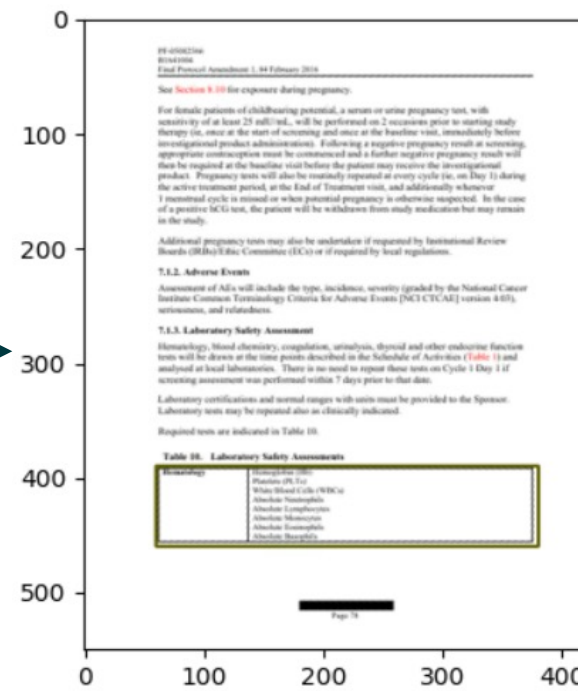
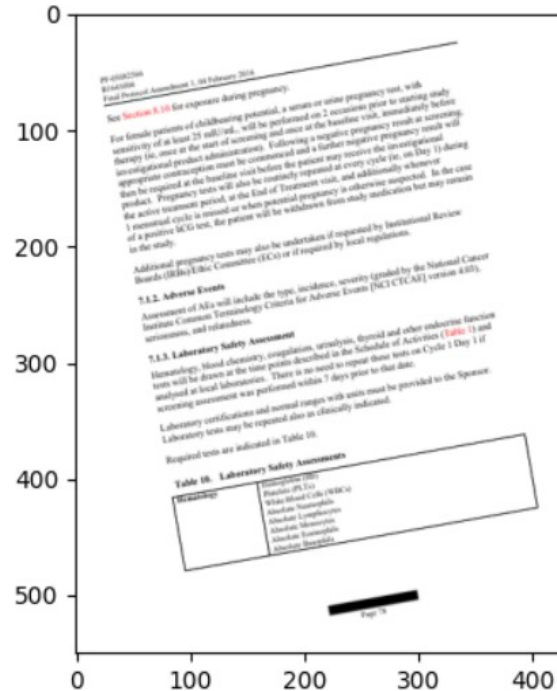
Bordered 1.00

Target Lesions	Non-target Disease	New Lesions	Objective status
CR	CR	No	CR
CR	Non-CR/Non-PD	No	PR
CR	Indeterminate or Missing	No	PR
PR	Non-CR/Non-PD, Indeterminate, or Missing	No	PR
SD	Non-CR/Non-PD, Indeterminate, or Missing	No	Stable
Indeterminate or Missing	Non-PD	No	Indeterminate
PD	Any	Yes or No	PD
Any	PD	Yes or No	PD
Any	Any	Yes	PD

Pre-process Image for Extraction

- This step precedes image analysis for main detection of tables, which can be bordered or borderless.
- For each image, the image orientation is identified, and its orientation corrected.
 - This is done to help the CV process easily group the relevant horizontal and vertical lines in a bordered table.
 - This step is necessary only if the image is converted from image inputs (e.g., scanned pages) and can be skipped if the file contained standard tabular formats.
- Finally, we crop the image using metadata.

Pre-process Image for Extraction (cont'd)



Pre-processing of the page for final conversion

Computer Vision on Table Image

Bordered Tables

Step 1: Image Kernels

- Bordered tables make it simpler to identify the table structure as we can target the horizontal and vertical lines.
- For identifying rows and columns, we need to determine the horizontal and vertical lines, for which we use horizontal and vertical kernels.



Horizontal Kernel



Vertical Kernel

- These kernels help to identify the table cells and make it convenient for us to iterate over each cell.

Computer Vision on Table Image (cont'd)



0
250
500
750
1000
1250
1500
1750
2000

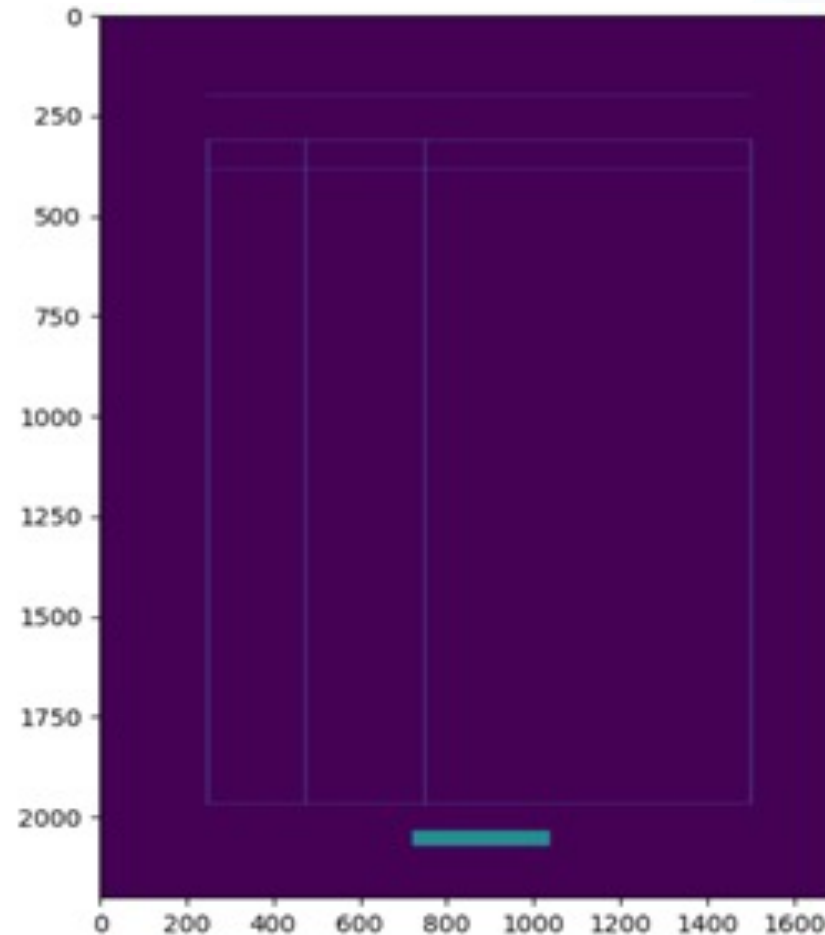
0 200 400 600 800 1000 1200 1400 1600

01-0000000
00000000
Final Protocol Amendment 1, 04 February 2016

Document History

Document	Version Date	Summary of Changes and Rationale
Amendment 1	04 February 2016	Protocol Summary: Secondary Endpoints - Endpoint 4 changed: V_{12} to V_{14} and added AUC_{0-12} and AUC_{0-14}. Protocol Summary: Study Design: Dose Finding Portion: - Changed the maximum tested dose from 10 mg/kg to 5 mg/kg. - Added the following in bold: Dose finding may stop if V DLT evaluable patients have been treated at the estimated MTD (if no DLT observed at any dose). - Added the 100 mg flat dose cohort to the dose finding portion of the trial which will only enroll SCCEN and/or squamous NSCLC patients. - Added that the 7.5 mg/kg cohort (if safe and tolerable) will enroll at least 6 squamous NSCLC patients. Protocol Summary: Study Design: Dose Expansion Portion: - Added that the expansion cohorts may include CRC, bladder, ovarian, squamous esophageal, SCCEN and squamous NSCLC. Schedule of Activities: - Revised the Baseline Signs and Symptoms CRF. - Updated, changed and clarified the timepoints for the PK, ADA and Cytokine blood draws. - Footnote 119: Changed time period for reporting of pregnancy/breast feeding from 28 days after last dose of PP-47082566 to 60 days.

Page 1



Computer Vision on Table Image (cont'd)

Bordered Tables

Step 2: Contour Filtering and Clustering

- After producing an image containing horizontal and vertical lines, we use contour filtering to distinguish those within the table from other lines and potentially redactions.
- Being an image, points on the same line might differ by a few pixels. Hence, we apply the clustering algorithm to determine where the lines along both X and Y axes align.
- We use a density-based DBSCAN algorithm for clustering. When writing to the excel table, this input allows us to determine if the cells are to be merged.

Computer Vision on Table Image (cont'd)

Bordered Tables

Step 3: Text Extraction

- The final step of this ETL process is to determine the text from the image. Using the clusters, we iterate over the cell positions and apply Optical Character Recognition (OCR) to the text within the cell boundaries.
- Tesseract was chosen as an OCR tool to extract text because of the ease of integration with the rest of the code. Also, this is a free open-source tool with continuous and frequent improvements being released.

Computer Vision on Table Image (cont'd)

Bordered Table - Results

Document	Version Date	Summary of Changes and Rationale
Amendment 1	04 February 2016	Protocol Summary: Secondary Endpoints - Endpoint 4 changed V_d to V_{ss} and added AUC_{inf} and deleted AUC_{∞}. Protocol Summary: Study Design: Dose Finding Portion: - Changed the maximum tested dose from 10 mg/kg to 5 mg/kg. - Added the following in bold: Dose finding may stop if 9 DLT evaluable patients have been treated at the estimated MTD (6 if no DLT observed at any dose). - Added the 100 mg flat dose cohort to the dose finding portion of the trial which will only enroll SCCHN and/or squamous NSCLC patients. - Added that the 5.0 mg/kg cohort (if safe and tolerable) will enroll at least 6 squamous NSCLC patients. Protocol Summary: Study Design: Dose Expansion Portion: - Added that the expansion cohorts may include CRC, bladder, ovarian, squamous esophageal, SCCHN and squamous NSCLC. Schedule of Activities: - Removed the Baseline Signs and Symptoms CRF. - Updated, changed and clarified the timepoints for the PK, ADA and Cytokine blood draws. - Footnote #19: Changed time period for reporting of pregnancy/breast feeding from 28 days after last dose of PF-05082566 to 60 days.

Document	Version Date	Summary of Changes and Rationale
Amendment 1	04 February 2016	o Endpoint 4 changed V_d to V_{ss} and added AUC_{du} and deleted AUC_{w}. Protocol Summary: Study Design: Dose Finding Portion: o Changed the maximum tested dose from 10 mg/kg to 5 mg/kg. o Added the following in bold: Dose finding may stop if 9 DLT evaluable patients have been treated at the estimated MTD (6 if no DLT observed at any dose). o Added the 100 mg flat dose cohort to the dose finding portion of the trial which will only enroll SCCHN and/or squamous NSCLC patients. o Added that the 5.0 mg/kg cohort (if safe and tolerable) will enroll at least 6 squamous NSCLC patients. Protocol Summary: Study Design: Dose Expansion Portion: o Added that the expansion cohorts may include CRC, bladder, ovarian, squamous esophageal, SCCHN and squamous NSCLC. Schedule of Activities: o Removed the Baseline Signs and Symptoms CRF. o Updated, changed and clarified the timepoints for the PK, ADA and Cytokine blood draws. o Footnote #19: Changed time period for reporting of pregnancy/breast feeding from 28 days after last dose of PF-05082566 to 60 days.

Computer Vision on Table Image (cont'd)

Bordered Table - with Cell Merges and Redactions

Table 2. Schedule of Activities: Pharmacokinetic, Pharmacodynamic, and Pharmacogenomic Assessments (Dose-finding and Expansion Cohorts)

Protocol Activities	Screening ^[1]	On Treatment (1 Cycle=28 days)									Post Treatment	
		Cycle 1				Cycles 2-4		Cycle 5		Cycles 8, 12, 16, 20, 24	EOT ^[9]	Follow Up ^[32]
Visit Window Cycle 1(+/-) 2 days Cycles 2-5: (+/-) 2 days Cycles >5: (+/-) 7 days	(<28 days of registration)	Day 1	Day 8	Day 15	Day 22	Day 1	Day 1	Day 8	Day 15	Day 1		
Blood for PF-05082566 PK ^[21]		X				X	X	X	X	X	X	
Blood for Mogamulizumab PK ^[22]		X	X	X	X	X	X	X	X	X	X	
CCI												
Banked Biospecimen (DNA) ^[23]		X										
Blood for Pharmacogenomics (RNA) ^[24]		X				X (C 2 & 3 only)	X				X	
Blood for PF-05082566 Immunogenicity testing ^[27]		X				X (C3 only)	X			X	X	X
Blood for Mogamulizumab Immunogenicity testing ^[28]		X				X (C3 only)	X			X	X	X
CCI												

Protocol Activities	Screening	On Treatment (1 CychZS days)										Post Treatment	
		Cycle				Cycles 2– Cycle				Cycles 8, 12, 16, 20, 24	[OT [31]	Follow U [3.	
Visit \Window Cycle I(+2 days Cycles 2-5: (1:) 2 days Cycles >5: (i) 7 days	(528 days of registration	Day	Day 8	Day 15	Day 22	Day	Day	Day	Day 15	Day			
Blood for PF-05082566 PK	:I		:I	S	E							E	
Blood for Mogamulizumab PKm]	S											S	
												J	
												H	
Banked Biospecimen (DNA)ood for Pharmacogenomic A)[261	:		:	:	:	:	:	:	:	I:	E	E	
Blood for PF-05082566 Immunogenicity testingm]	S		E	E	E	(C 2 & 3 only)		E	S				
Blood for Mogamulizumab lggggunogenicifl testingm	S		E	E	E	(C3 only)		E	S				

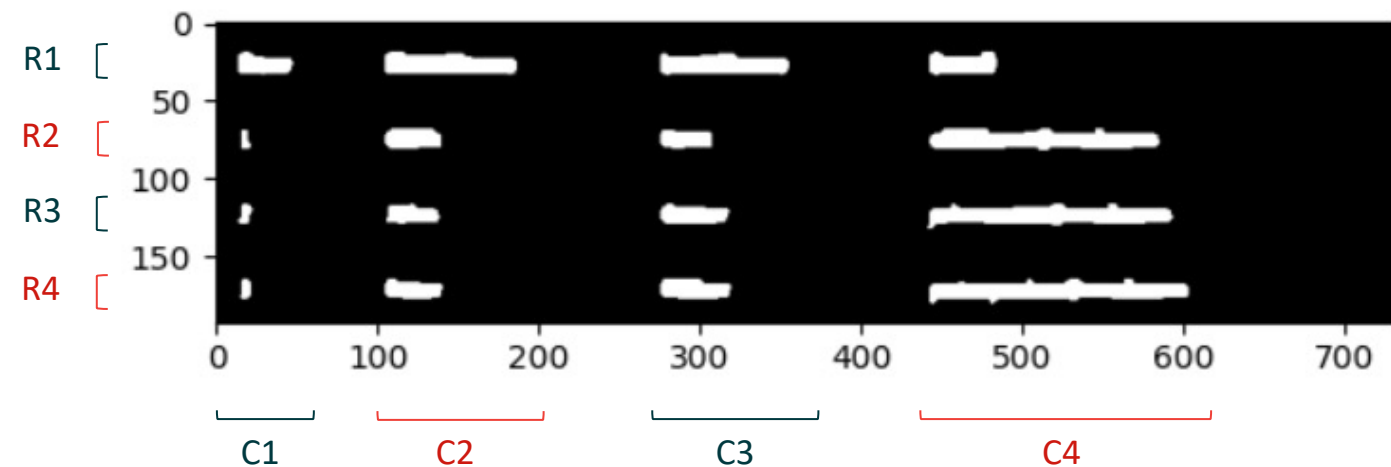
Computer Vision on Table Image (cont'd)

Borderless Tables

Step 1: Morphological Operations

- We have to group text which is part of the same cell but not from other cells.
- To accomplish this, morphological operation of image dilation followed by erosion is carried out.
- A simple rectangular kernel is used for this.

Row	First Name	Last Name	Email
1	Clark	Kent	clarkkent@mail.com
2	John	Carter	johncarter@mail.com
3	Peter	Parker	peterparker@mail.com



Computer Vision on Table Image (cont'd)

Borderless Tables

Step 2: Contour Filtering

- The contours of the image are found.
- The bounding boxes of these cells can be directly used for text extraction, but close proximity of border and text can reduce OCR accuracy.
- So, the bounding boxes are then padded with a few pixels as a pre-processing step to OCR.

Step 3: Text Extraction

- This step follows the same strategy used for bordered tables.

Computer Vision on Table Image (cont'd)

Borderless Tables - Results



Name	Age	Location
John	28	New York
Kelly	29	London
Chung Fu	21	Hong Kong

	A	B	C
1	Name	Age	Location
2	John	28	New York
3	Kelly	29	London
4	Chung Fu	21	Hong Kong
5			

Challenges

1. Using Computer Vision to identify table structure of borderless tables is extremely difficult. Without cell uniformity (e.g., merged cells, inconsistencies in orientation/alignment), the resulting table structure can be incorrect.
2. For our test/pilot, we had not further trained the OCR to understand special characters or to better decipher text in proximity to borders. This resulted in:
 - a. Inaccurate predictions for text in close proximity to cell borders.
 - b. Special characters were predicted as one of the alphanumeric characters.
 - c. In some cases, similar characters (ex: 0 and o) were getting inappropriately cased (possibly thrown off by bullets).
3. Performance:
 - a. Image quality has a very high impact on OCR performance.
 - b. Additional considerations for CPU, memory, real-time vs. batch processes, “watch” directories, etc., should be made to determine what works best for your organization. COTS offerings offer additional capabilities.



Recommendations

Despite the challenges, you may be able to directly leverage tabular data in PDFs by adhering to the following recommendations:

#	Challenge	Recommendation
1	Table/Cell Borders – for bordered tables, text in close proximity to borders reduces accuracy	Cell Margins should be set to a minimum of .02 in on all sides.
2	Table Text – low quality, nonoptimal font settings reduce accuracy	Images: >=300 DPI (the smaller the font size, the greater the DPI needed for best results) Font Size: 10-14 Font: OCR-A, OCR-B (designed for OCR), Microsoft Sans Serif or any font in the Sans Serif family
3	Use of Non-ASCII Characters	Use of ASCII characters is recommended as other characters are predicted to ASCII characters.
4	Borderless Tables – variable cell formats/structure reduces accuracy	The structure of these tables must not be too complex, cell merges should be avoided.
5	Redactions – proximity to remaining text reduces accuracy	Cell redactions must be carefully done to avoid overlap with remaining text.



Conclusion

- The pilot project involved batch processing of a standard file input to return all the tables within a document. The process can be revised to run for only specific page ranges or specific tables.
- The pilot confirmed for us the potential value in enabling study team members the ability to leverage data contained in PDFs.
- With document preparation based on recommendations/standards and additional/ongoing model training, we expect to achieve 99.8% results. This will allow us to save time in generating first draft deliverables such as specifications, trial domain datasets, protocol deviations/violations datasets, etc.
- There is extensive research on PDF and OCR development, and with improvements in this field, quality and efficiencies will continue to advance.
- We will be continuing our testing and development and are planning for a sequel presentation at the EU connect. In the meantime, please drop us your business card and we will share updates of our progress and evaluations of our opensource pilot versus COTS offerings.



www.GENINVO.com
North America | UK | Europe | India



Ideas Delivered Intuitively

Thank You!

(309) 807-5006 | info@GenInvo.com | www.GenInvo.com

