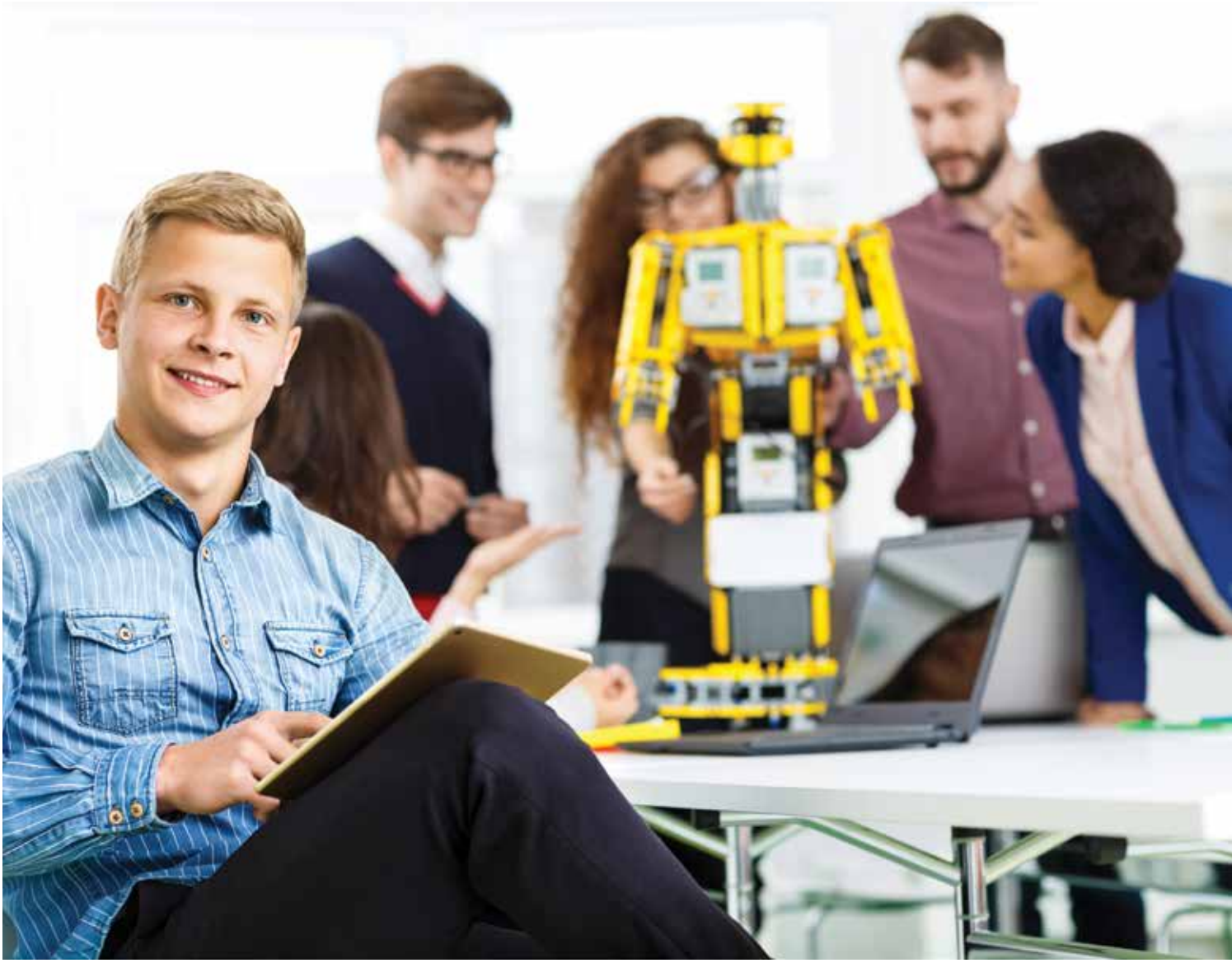
Contents

State of automation in modern enterprises p3/Overview of OCR p5/Need for intelligent OCR p7/
OCR complexities faced by RPA developers p8/UiPath 2017 vs UiPath 2018 comparison p10

Robotic process automation and intelligent character recognition: Smart data capture



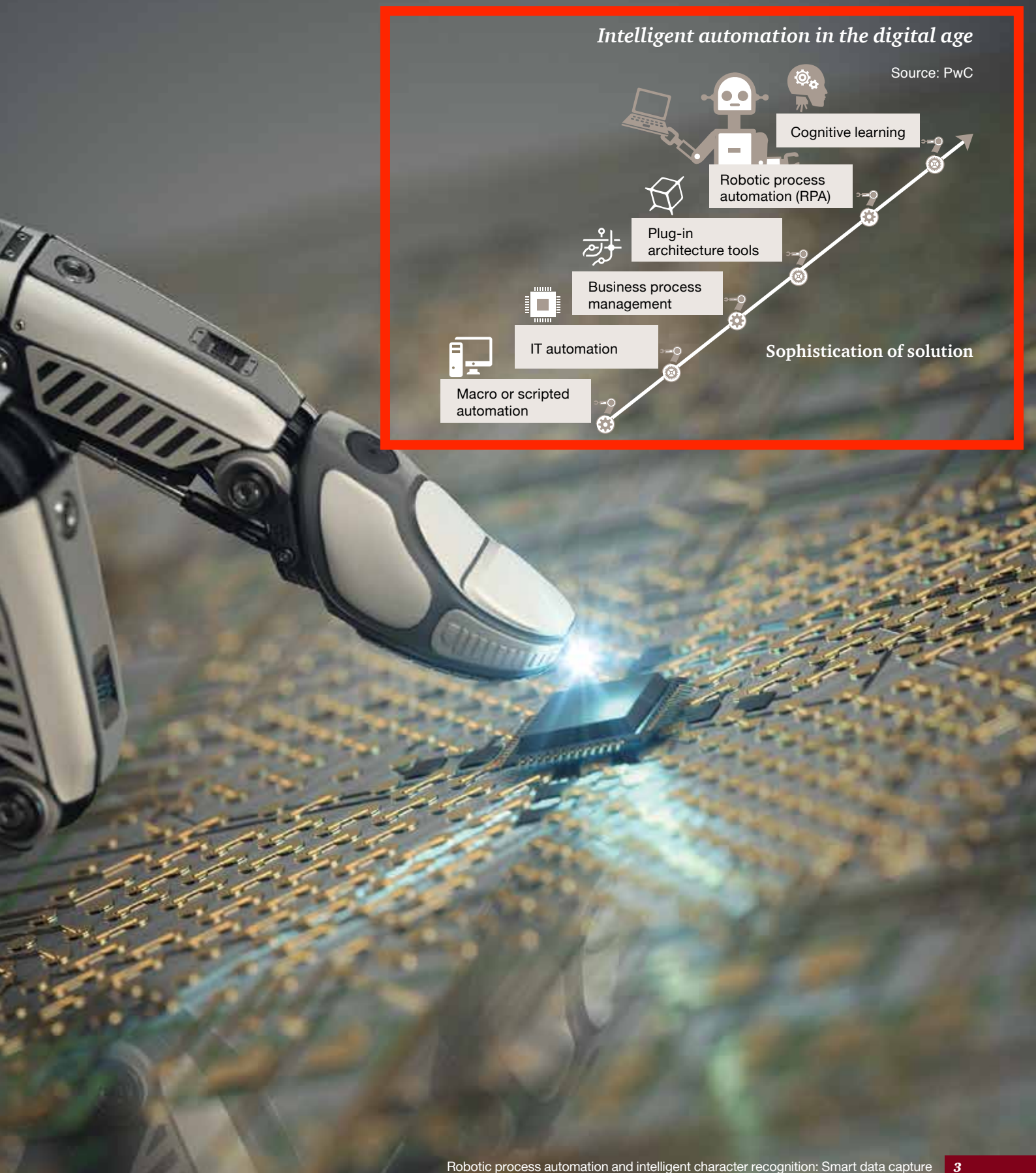
pwc

www.pwc.in

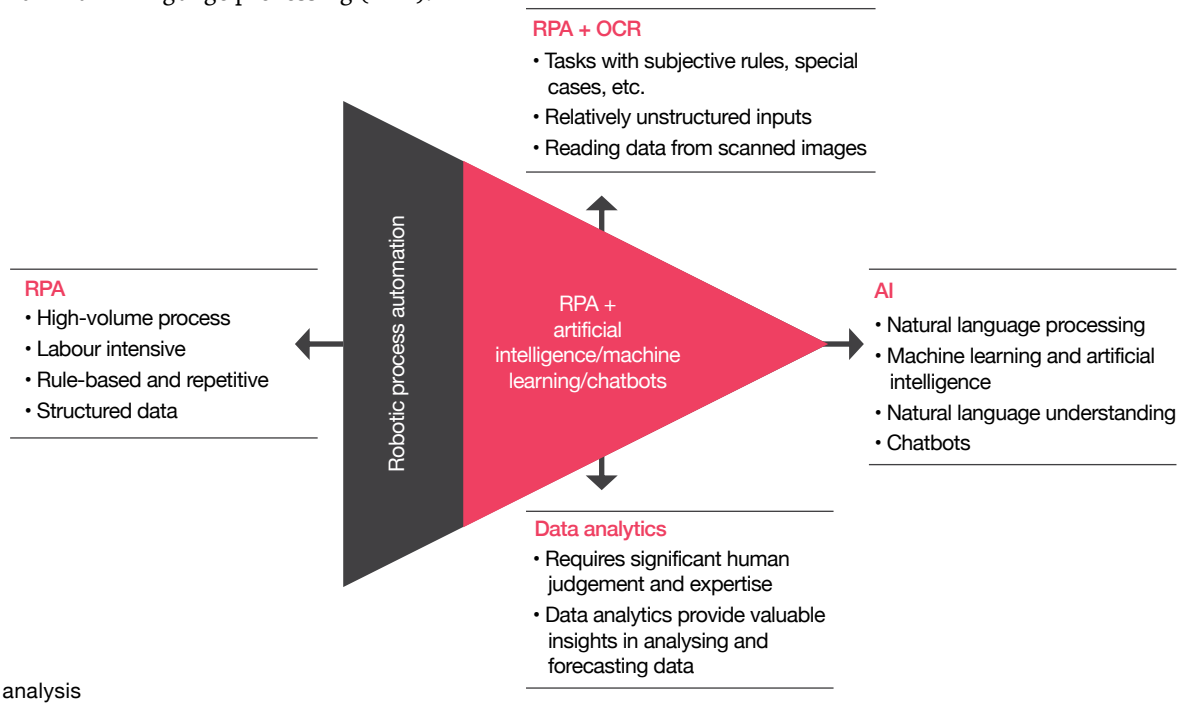
State of automation in modern enterprises

In this era of technology disruption, enterprises are under immense pressure to digitise operations, and they are gearing up for a future where human work can be augmented by software robots. Digitisation and automation continue to be the key business drivers across various sectors and industries globally, including government organisations, which have now jumped onto the automation bandwagon.

Enterprises are looking to build a digital workforce as part of their automation strategy by combining elements of robotic process automation (RPA), artificial intelligence (AI), optical/intelligent character recognition (OCR/ICR) and analytics to automate their business processes. While RPA technologies are capable of taking on low-value activities in a quick and efficient manner, the next phase is leveraging such automation technologies to deliver intelligent process automation (IPA).



With software robots becoming more advanced in recent years and undertaking more than just the automation of mundane rule-based processes, organisations are expanding the scope of process automation end to end to include sections that were initially deemed non-automatable as inputs were in the form of unstructured data, documents/scanned images, texts/human judgement and natural language processing (NLP).



Source: PwC analysis

Customers are the new market makers, reshaping the automation requirements and largely influencing product vendors to constantly upgrade and meet those requirements. Success depends on how well and fast an organisation responds. The RPA landscape is now maturing to a state where product vendors have started integrating their product offerings with other tools to distinguish themselves and stay ahead in the market. The impact of such collaboration helps them achieve seamless automation in some areas coupled with workflow tools while also improving the automation percentage that could be achieved in processes that require image/character recognition.

In our experience of having automated processes across enterprise organisations, we have found that lack of extensive OCR capabilities within RPA tools resulted in failed attempts at automating some complex OCR processes. For the purpose of this paper, we will focus on how RPA tools facilitate process automation that requires OCR, and discuss challenges around such automation.

While multiple RPA solutions are available in the market, this paper presents a case study on UiPath (a leading RPA product vendor[1(1)]), which has integrated ABBYY Flexicapture (an intelligent OCR platform) in its latest offering, UiPath v.8 (Firefly) to tackle advanced OCR automation requirements.

'In conversations with various clients on key challenges they faced in their RPA journey, extracting information from unstructured data formats and inputs emerged as the most common issue. They also pointed out the limited options available to address advanced intelligent optical character recognition capabilities in RPA tools, which limited their automation scope.'

Sumit Srivastav,
Intelligent Process Automation Leader, PwC India

1 Le Clair, C. (26 June 2018). The Forrester Wave™: Robotic Process Automation, Q2 2018. Retrieved from https://www.uipath.com/hubfs/The_Forrester_Wave_RPA_2018_UiPath_RPA_Leader.pdf (last accessed on 16 July 2018)

What is OCR and how does it work?

OCR is a technology that primarily aims to analyse an image, detect based on patterns if the image contains text, and extract that text into a machine readable format. This helps convert scanned documents into a digitally editable format while comparing the images of the available characters to the ones stored on its database for traditional OCR engines. The newer versions of OCR use machine learning (ML) techniques to recreate the characters and render the best possible match to the user.

Based on the image type and type of data that needs to be extracted, these character recognition engines use the options below to recognise text.

ICR/OCR:

ICR helps in converting handwritten text characters into a machine-readable format. The core difference between OCR and ICR is that in the case of OCR, its capability is restricted to printed data that looks the same given the standardisation of multiple fonts. However, in the case of ICR, it is intelligent enough to decipher data from non-standard documents, which contain handwritten texts with varied formats.

Optical mark recognition (OMR):

This technology helps in recognising tick/check marks and also free-form check marks like underlined text and shaded circles.

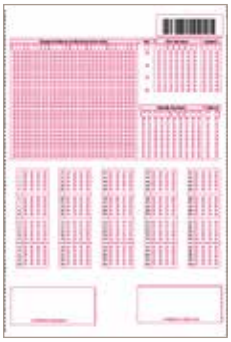
Optical barcode reader (OBR):

An OBR helps in reading barcoded data from a document.

What are the different data types for which the above engines can be used to extract data?

Structured documents:

These document types are standard in format and templatised with a fixed location for specific data sets. This makes it easier for the OCR engines within the RPA tools to search for data as they can be trained to look for a particular data set on the document at a specified place. Nothing usually moves or changes around these forms of data; therefore, it becomes easier for the bot to look in the same place every time for the information to be extracted. Examples of structured data documents are banking forms, surveys, exam papers, etc.

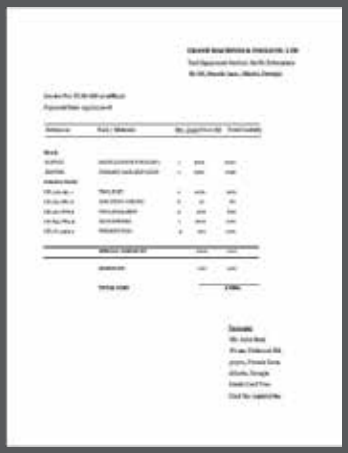


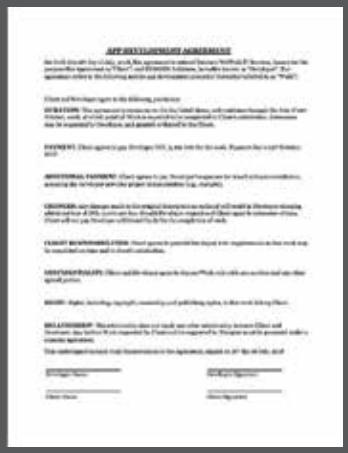
Semi-structured documents:

Semi-structured documents do not have a formal structure in place for information. The document is usually the same, but design and layout may differ. The information will be tagged in the document, but the placement of the information may vary from document to document. Common examples of semi-structured documents are invoices and purchase orders.

Unstructured documents:

In this case, documents have no standard structure. The data is usually free-flowing and lacks consistency. Due to complexities in the way the data is presented in these documents, it becomes challenging to come up with a solution for data extraction and companies usually have to appoint staff who extract key information and feed the same into the internal business systems. This is a time-consuming and costly task, and prone to manual errors. Examples include contracts, agreements and letters.





Where do enterprises need intelligent OCR?

Intelligent OCR is needed to simplify paper-driven processes where inputs are received in varied multiple formats such as PDF, scanned, fax and handwritten documents. Examples of such processes where OCR can be implemented are:

Financial services

- Confirmations and pre-/post matching
- Customer onboarding
- Account opening
- Loan applications
- Compliance-related processes
- Receipt processing
- Vendor onboarding

Insurance

- Claims handling
- Mortgage processing

Manufacturing

- Sales order processing
- Accounts payable/ receivable
- Parts requests from customers
- Remittance processing

HR

- Employee onboarding
- Extracting key data from candidate CVs
- HR records processing

Supply chain management

- Order scheduling and tracking of shipments
- Bill of lading
- Transport notes

Healthcare

- Billings and claims management
- Insurance processing

Government sector

- Immigration applications
- Education system applications
- Passport management applications

What are the complexities faced by RPA developers?

Scaling the image to the correct size

Most of the market OCR engines require a minimum image quality size. This usually ranges from 200–300 dots per inch (DPI). Anything less than the minimum requirement will result in unclear and inaccurate results. On the flip side, having an exceptionally good DPI quality (e.g. 500 DPI) will not help in increasing the quality of the output. Rather, only the image size will increase, which will result in increased storage. If the quality of the image is not good, the OCR engine can get confused. Instead of reading an S, the output can be provided as 5, the number '0' or the letter 'O'. Hence, the better the quality of the image, the better is the output provided by the OCR engine.

Handwritten text/ink stamps over printed text

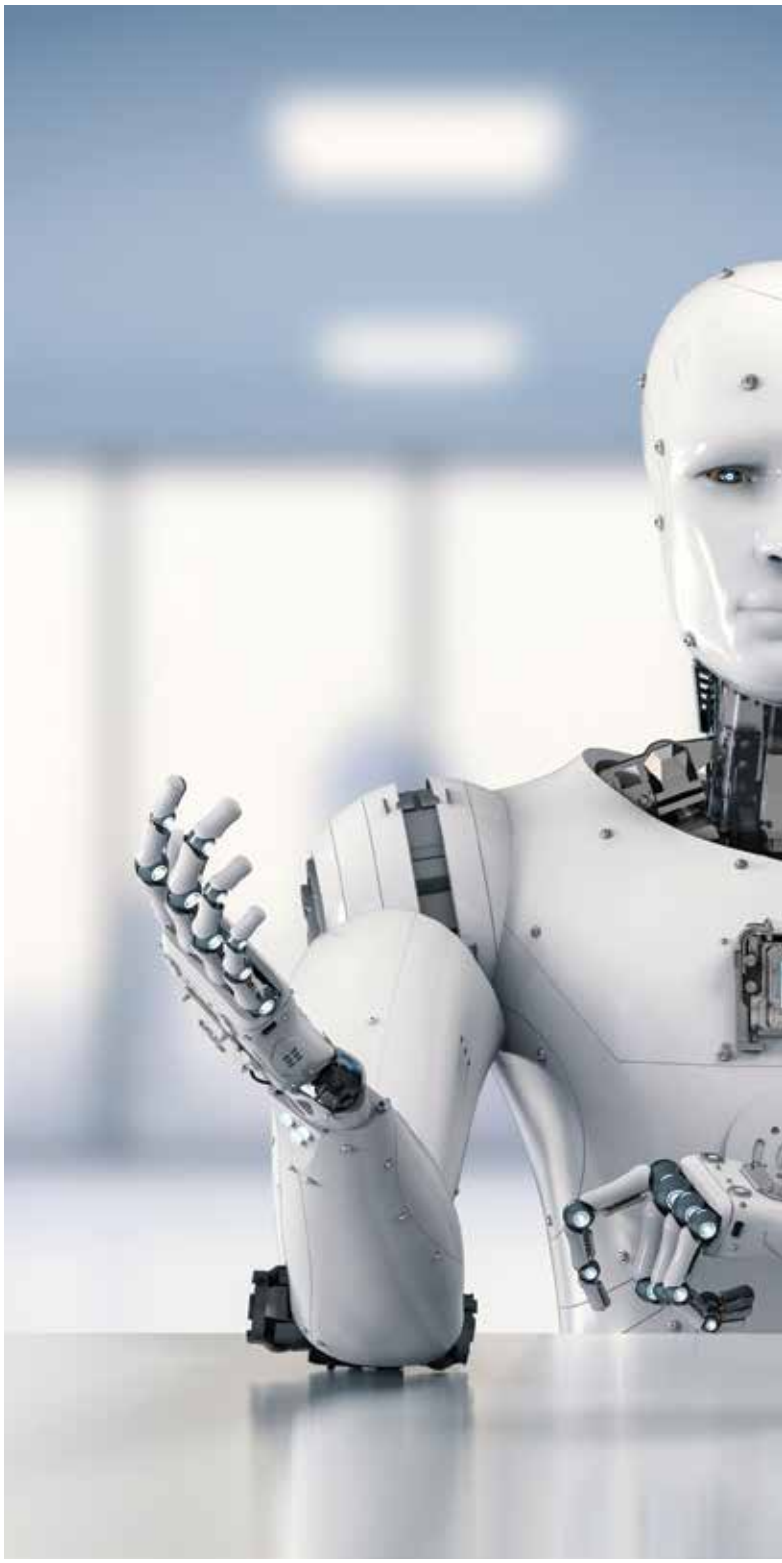
As part of internal procedures (Maker/Checker, audit etc.), people tend to write critical information or use stamps over documents which are then scanned. Such handwritten text usually interferes with the printed text and makes it difficult for the OCR engine to capture the text from the document. Moreover, this reduces the quality of the document.

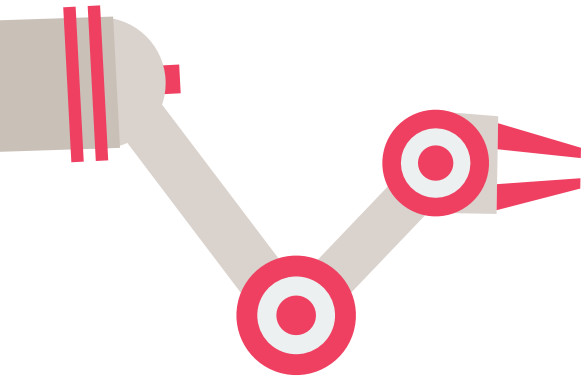
Noise/distortion on the scanned image due to bad scanner quality

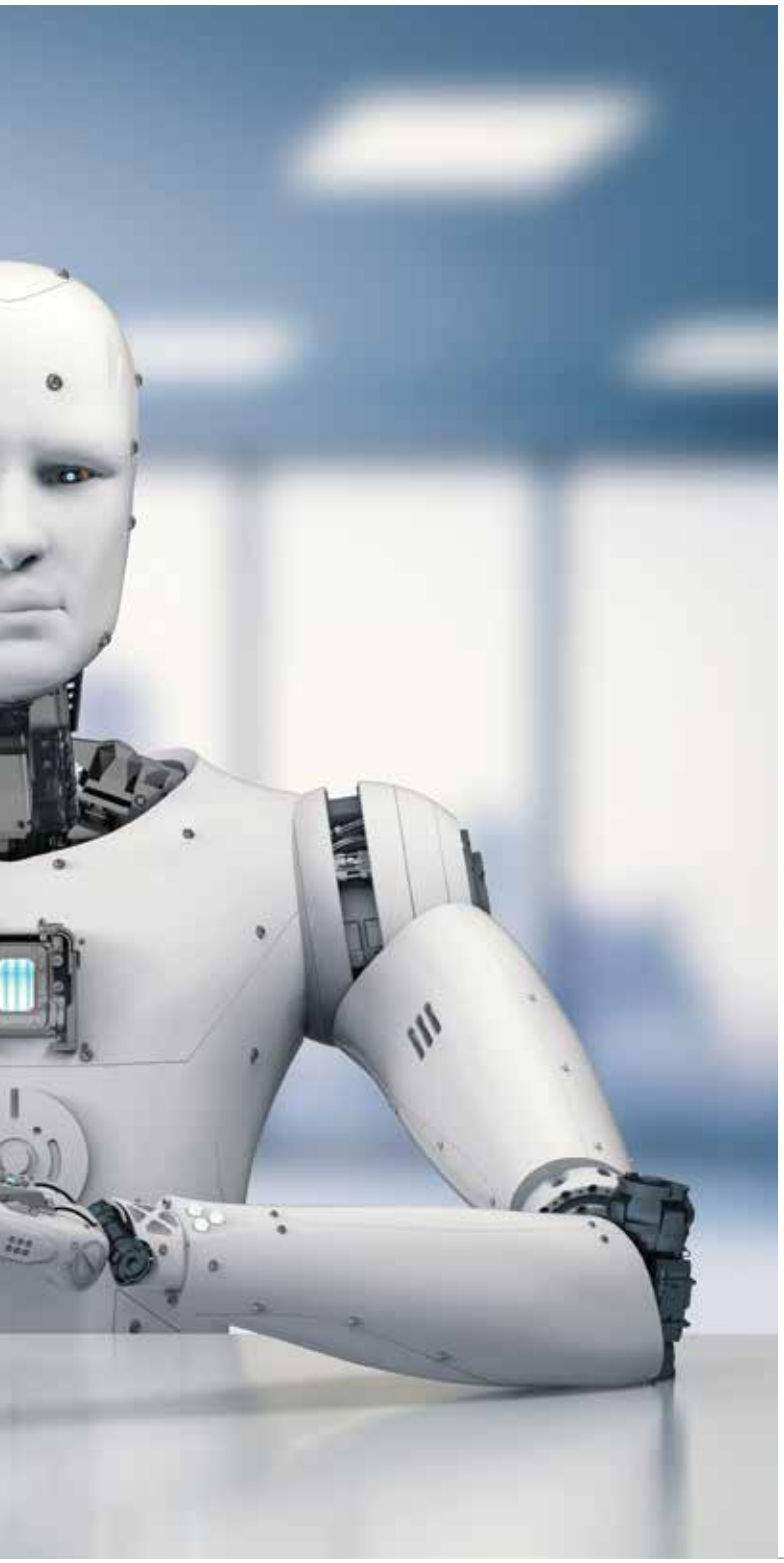
The presence of noise (distortion) on the scanned document can significantly reduce the output from the OCR engine. Noise usually appears on a document due to improper scanning or a bad scanner. Examples of noise are spots in the background of the document, uneven contrast, etc.

Higher number of sample documents required for training

In our previous implementation experience, we have observed that all types of documents are not made available during implementation, which makes it challenging as the OCR engine has to be trained on the major types of documents. The higher the number and variety of samples, the more efficiently the OCR engine can be trained to handle exceptions and errors.







Scanning an already scanned image

Many a times, a hard copy document is printed which is a scanned image. Scanning a printed copy of an already scanned image would definitely impair the quality of the document, thereby influencing the accuracy levels of the extracted data.

No labels on tables

Many invoices or purchase orders have tables where the particulars are mentioned but they do not contain headers or labels like amount, description and quantity. This makes it challenging for the OCR engine to search for the appropriate data.

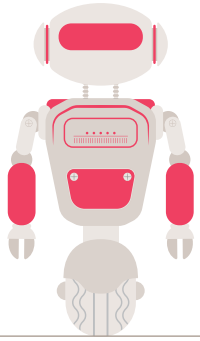
Background images and colour

Often, business documents include design elements such as textures and background images. The presence of these has an adverse effect on the quality of text recognition from the scanned image.

Multiple formats of inputs

A document can be received for further processing in various formats. The multiple formats increase the complexity of implementation. Examples of such formats are TXT, EML, XLSX, VSD, HTML, DOCX, XLS, VSDX, DOC, PPTX, HTM, PPT, RTF, BMP, PCX, DCX, JPEG, TIFF, GIF, PNG, PDF.

There are various options in the market when it comes to OCR engines. While many of the popular OCR engines do a good job, each comes with its own strengths and weaknesses. Choosing the correct engine depends on various important factors like accuracy required, budget, type of use case chosen and ease of integration with the current technology landscape.



The current RPA workarounds to potential OCR automation roadblocks may not be perfect and a foolproof solution may not exist. UiPath has been working behind the scenes to tackle the roadblocks related to OCR automations and has integrated the ABBYY OCR engine with its current OCR toolset to bring about a revolution in OCR automation. The 2018 version of UiPath has been codenamed Firefly.

PwC was given a preview of Firefly and performed a comparative study of some of the key OCR functions/commands used in Firefly and UiPath's previous version, the 2017 enterprise RPA platform codenamed Moonlight. The table below compares and assess the OCR capabilities of the two versions and scores them on three parameters.

Data extraction, data validation and data classification

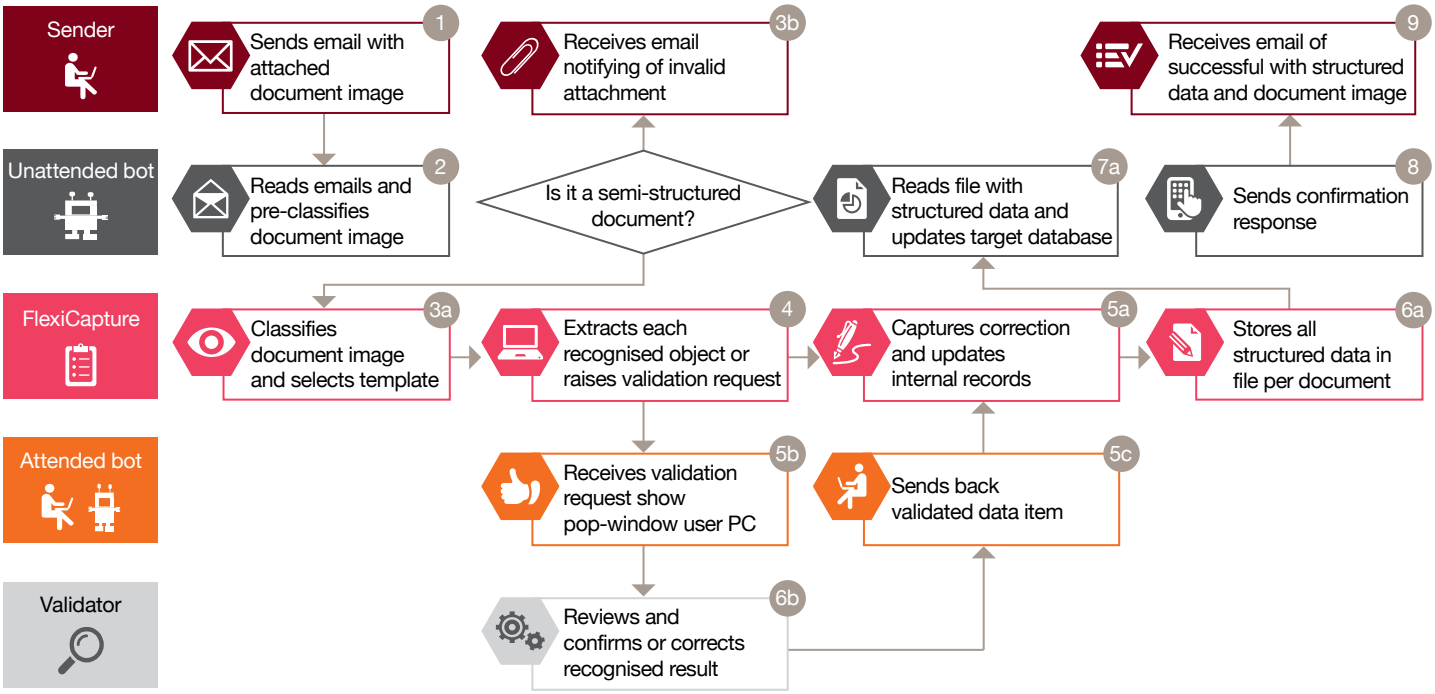
OCR evolution		UiPath 2017 (Moonlight)			UiPath 2018 (Firefly)		
		Google Tesseract 3.0	Microsoft Modi	Abbyy Fine reader 11	Google Tesseract 4.0	Microsoft Modi	Abbyy Flexicapture 12
Data extraction							
	Screen scrapping - desktop/web	●	●	●	●	●	○
	Screen scrapping - documents	◕	◕	●	●	◕	○
	Structured data	●	●	●	●	●	●
	Semi-structured data	◗	◗	◕	◕	◗	●
	Unstructured data	◗	◗	◕	◗	◗	●
	Multilingual support (same document - multiple languages)	○	○	●	○	○	●
	Multilingual support (different documents - multiple languages)	◗	◕	●	◕	◕	●
	Barcode extraction	○	○	●	○	○	●
	Signature extraction (validation not included)	○	○	◗	○	○	◗
	Table extraction (output available in data table format)	○	○	◔	○	○	●
	Handwritten information (ICR)	○	○	○	○	○	◗
Data validation							
	Manual validation (verification panel for data validation, post extraction)	○	○	○	○	○	●
	Confidence score (data extraction quality score)	○	○	◔	○	○	●
	Data comparison across multiple documents	○	○	○	○	○	●
Data classification							
	Classification - different data types (checkboxes/barcodes, etc.)	○	○	○	○	○	●
	Classification - document (based on document templates)	○	○	○	○	○	●

● Best in class ◕ High ◗ Medium ◔ Low ○ NA

UiPath 2018, Firefly, can help enterprises achieve intelligent OCR automation with ease using RPA. The integration of the ABBYY OCR engine not only enhances automation for rules-driven processes, but also adds the flavour of NLP and widens the scope of automation.

Firefly is an intelligent and enhanced version of its predecessor, Moonlight. Firefly brings to the table RPA coupled with cognitive abilities, which help enterprises overcome the burden of comprehending unstructured data using the cognitive capabilities of ABBYY FlexiCapture, amongst other ML/AI components.

Illustrative example



Firefly: A preview

Some of the key highlights of UiPath v.8 are listed below:

Segment	UiPath 2017 (Moonlight)	Evaluation	UiPath 2018 (Firefly)	Evaluation
OCR capabilities - Google	Tesseract 3.0		Tesseract 4.0 • Higher scraping accuracy across all languages • New OCR based on long short-term memory neural networks	
OCR capabilities - ABBYY	FineReader • Simple documents • Same formats • Multiple languages		FlexiCapture • Multiple documents • Multiple formats • Complex documents • Many languages • Human validation • Advance reporting	
Cognitive and natural language processing	Out-of-the-box activities for integration with third-party cognitive platforms (separate licences) • Google Text Analysis • Google Text Translate • IBM Watson Text Analysis and • Microsoft Text Analysis		Out-of-the-box activities to utilise Stanford Natural Language Processing libraries (free) • Text analysis • Entity extraction • Capturing intent from unstructured content in emails and documents	
Machine learning and AI	• Machine learning and AI activities were not included.		• Python activities integrated to support executing and embed Python code machine learning models • Automated alerting mechanisms using machine learning models in Elasticsearch (X-Pack)	
Scalability	• Multiple terminal sessions and • Invoke codes (custom activity creation)		• Hyperscalability (simultaneously host and manage up to 10K robots in Orchestrator • Out-of-the-box REFramework offering an automation template for large-scale deployments • RPA adapter for Oracle • Integration with Newgen Soft • Centralised Runtime (Robot) configuration settings through Orchestrator	
Licensing	• Node locked and authorised user licences • Centralised licensing, automatic studio activation		• Concurrent licences (licence consumption is not based on machine but on actual users • Licensing robots using Orchestrator • Regutil.exe to activate (online/offline), deactivate or export licence information to a file	
Security (authentication)	• CyberArk integration • Password complexity configuration • Multi-tenancy with Orchestrator host admin implemented		• Secure deployment added through secure NuGet feed • Foolproof packages • Organisation units to allow separation of orchestration resources • Azure AD SSO implemented and • Entry into Veracode verified directory	
Online academy	**Free online courses** • Foundation Course • Orchestrator • SAP Automation Training		**Free online courses** • 360° training • Single topic tutorials covering RPA Center of Excellence roles: • Business analyst • Implementation methodologist • Solution architect • Infrastructure engineer • RPA awareness	
Analytics	• Integration with Elasticsearch and Kibana for data visualisation		• Integration with Tableau, • Machine learning extensions: Elasticsearch (X-Pack) build machine learning for anomaly detection • Enhanced robot logging capabilities with queues reporting, review and audit • Improved monitoring by providing transperancy (new dashboards and visual reports in Kibana and Tableau to monitor data processing in APIs and robot-to-robot process automation)	

Best in class High Medium Low NA

Organisations globally have accepted the reality that striking gold with processes that are high on volume will become harder to find in the days to come. RPA vendors must get smarter to escape the tag of structured rule-based automation. While AI will not replace RPA, RPA tools that use AI components will replace those that do not disrupt the modest roots of RPA. UiPath is headed in this direction by collaborating with partners providing automation essentials such as data analytics, NLP, ML and intelligent OCR engines.

As other RPA vendors in the market work on future versions of their tools and similar enhancements, we will bring out a series of thought papers that cover other key players and product enhancements in the RPA world:

Issue 1 – RPA in a virtual environment (May 2018)

Issue 2 – RPA and intelligent optical character recognition (July 2018)

Notes

Notes

About PwC

At PwC, our purpose is to build trust in society and solve important problems. We're a network of firms in 158 countries with more than 2,36,000 people who are committed to delivering quality in assurance, advisory and tax services. Find out more and tell us what matters to you by visiting us at www.pwc.com

In India, PwC has offices in these cities: Ahmedabad, Bengaluru, Chennai, Delhi NCR, Hyderabad, Kolkata, Mumbai and Pune. For more information about PwC India's service offerings, visit www.pwc.com/in

PwC refers to the PwC International network and/or one or more of its member firms, each of which is a separate, independent and distinct legal entity. Please see www.pwc.com/structure for further details.

© 2018 PwC. All rights reserved

Contacts

Sumit Srivastav

Partner and Intelligent Process Automation Leader
PwC India
sumit.srivastav@pwc.com

Authors

Nitin Kamra
Hariprasad Gajapathy

pwc.in

Data Classification: DC0

This document does not constitute professional advice. The information in this document has been obtained or derived from sources believed by PricewaterhouseCoopers Private Limited (PwCPL) to be reliable but PwCPL does not represent that this information is accurate or complete. Any opinions or estimates contained in this document represent the judgment of PwCPL at this time and are subject to change without notice. Readers of this publication are advised to seek their own professional advice before taking any course of action or decision, for which they are entirely responsible, based on the contents of this publication. PwCPL neither accepts or assumes any responsibility or liability to any reader of this publication in respect of the information contained within it or for any decisions readers may take or decide not to or fail to take.

© 2018 PricewaterhouseCoopers Private Limited. All rights reserved. In this document, "PwC" refers to PricewaterhouseCoopers Private Limited (a limited liability company in India having Corporate Identity Number or CIN : U74140WB1983PTC036093), which is a member firm of PricewaterhouseCoopers International Limited (PwCIL), each member firm of which is a separate legal entity.

AW/July 2018-13763