Shufti Pro

Identity Verification

Fully Embracing Digital Transformation with Automated AI OCR Engine

White Paper

Shufti Pro

> Digitalization is becoming the key business driver and making it quick, secure, and convenient for organizations is the primary focus for Shufti Pro. OCR for businesses is designed to help businesses in automating manual data extraction and making data management efficient. Our team worked hard to deliver a technically advanced AI-powered OCR engine that is designed to eliminate manual data extraction and help businesses in achieving seamless digitalization.

Shahid Hanif

CTO and Co-founder, Shufti Pro

Outline

State of Automation in this Technological Era

In this era of technological disruption, businesses are under immense pressure to digitize operations and they are looking for a future where manual tasks can be augmented by using software robots. Digitalization and automation have been the key business drivers for organizations in different sectors globally. According to Mickensey, 64% of the businesses could save 30% of their time with workflow automation.

Enterprises around the world are looking for workflow automation by combining Robotic Process Automation (RPA), Artificial Intelligence (AI), big data analytics, and Optical Character Recognition (OCR). While RPA applications can automate low-value activities in a quick and efficient manner, companies mostly lag in the automation of high-value tasks that require time, energy, and money. One such task is data extraction from important business documents.

With technology becoming more advanced in recent years and undertaking more than just automation, organizations are expanding the scope of process automation to include tasks that were previously seen as non-automatable, because inputs were in the form of unstructured data, scanned images/documents or handwritten texts.

Technology isn’t only transforming business operations but also the way businesses interact with customers. The word digital transformation is on everyone’s lips from logistics to utilities and smart factories. Customers are becoming well versed with digital means and expect a seamless digital experience when interacting with any business.

However, the challenge for businesses is to adopt digital transformation by leaving the existing systems. The most efficient way for organizations to embrace digital transformation is by relying on technology that could easily fit into the existing systems i.e. transform analog/physical data around them into digital data.

Optical Character Recognition(OCR) is one of the similar technologies that could help businesses in digitizing the data. Adopting this technology, businesses can simplify everyday tasks such as onboarding a new customer, digitalizing the scanned invoices, managing accounts sheets, to name just a few.

This whitepaper will focus on how the AI-based OCR engine could help businesses to embrace digitalization completely.


Businesses spend an estimated **3%** of their revenue on paper consumption, **50%** of their waste is composed of paper. (Gartner, Inc. 2018)
28% of businesses gained a full return on investment (ROI) in the first six months of going paperless, **59%** achieved a full ROI in under 12 months, and **84%** within an **18-month** period. (Business Partner magazine, Nov 2019)
Businesses can save **$2 trillion** in total annual wages by automating **45%** of current paid activities. (McKinsey report)
Office workers spend around **69 days** per year on administration tasks, resulting in an annual productivity loss of **$5 trillion.** (Unit 4 study)
Automation in the healthcare and pharmaceutical industry is estimated to generate a revenue of more than **US$ 900 million.** (Fact.MR)
Small and medium enterprises will create a market for the opportunity of more than **US$ 1,600 million** between **2017-2026** by adopting business workflow automation. (Fact.MR)
Marketing automation drives a **14.5%** increase in sales productivity and marketing overhead is reduced by **12.2%.** (Sessioncam)
Businesses using cloud automation witness year-over-year **revenue growth of 15%.** (Instapage)
80% of companies using marketing automation saw an increase in their number of leads, and **77%** saw the number of conversions increase. (VB Insight, 2015)

AI-Based OCR Engine

Data is a high-octane fuel that helps run your business operations. When the data isn't available in the form your business system consumes, it chokes off the fuel supply and leads to stammering efforts to improve productivity (to gain new insights for your business). If you want to transform your business into a digital platform, your business needs to extract the unstructured documented papered documented data or crudely embedded pixels in the form of images.

Today you probably have a team of employees manually managing the data entry task or you may have been using free OCR engines available online to augment efforts of your team by automating some of the data extraction tasks. Even these free OCR engines have their limits. You can only extract structured data that too with a lot of errors that require manual correction.

But with AI-based OCR engines, things work differently. Let's look at What is AI OCR and How it works?

What is AI OCR?

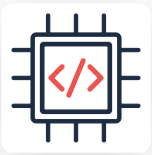
AI-based OCR engines combine both artificial intelligence and OCR technology to transform a document into a machine-readable and editable digital format. Traditional OCR engines aim to analyze a document in the form of an image by detecting based on patterns if the image contains text and then extracts the 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 AI-based OCR engines use different machine learning, computer vision, and natural language processing (NLP) algorithms to render the text in images and deliver more accurate results to the users. By looking at and understanding the language, document type, context, and other specific details of the document, AI OCR engines build a comprehensive understanding of the document and the data within. Delivering accuracy of up to 99.9%, AI OCR engines eliminate the need for a human resource to make corrections.

How does the AI OCR Engine Work?

AI-based OCR works in 3 main steps:

1. Pre Processing
2. Data Extraction i.e OCR
3. Post Processing



Pre Processing

The preprocessing step is further divided into 4 distinctive steps

Step 1: Cleaning

In this process, the brightness and contrast of the image are adjusted and the noise, distortion from the data is cleaned by detecting borders and adjusting the detection threshold.

Step 2: Deskewing

On a regular page, the characters are in a straight line. However, this may not be true in a less-than-perfect page scan. The book's binding could raise pages above the scanner's glass, to create a skew or curved image of the text. The deskewing process helps straighten such images.

Step 3: Shake Reduction and Sharpening the Image:

The input may be captured on a smartphone rather than a flat-bed scanner, where the shaking of images is a distinct possibility.

Step 4: Normalization

The last step in the pre-processing is to normalize the image that could be under changing conditions of light. The image is then smoothed to normalize the noise created during the afore-mentioned processes.

Data Extraction

Data extraction in AI OCR engines is also done in two distinct steps:

Step 1: Segmentation

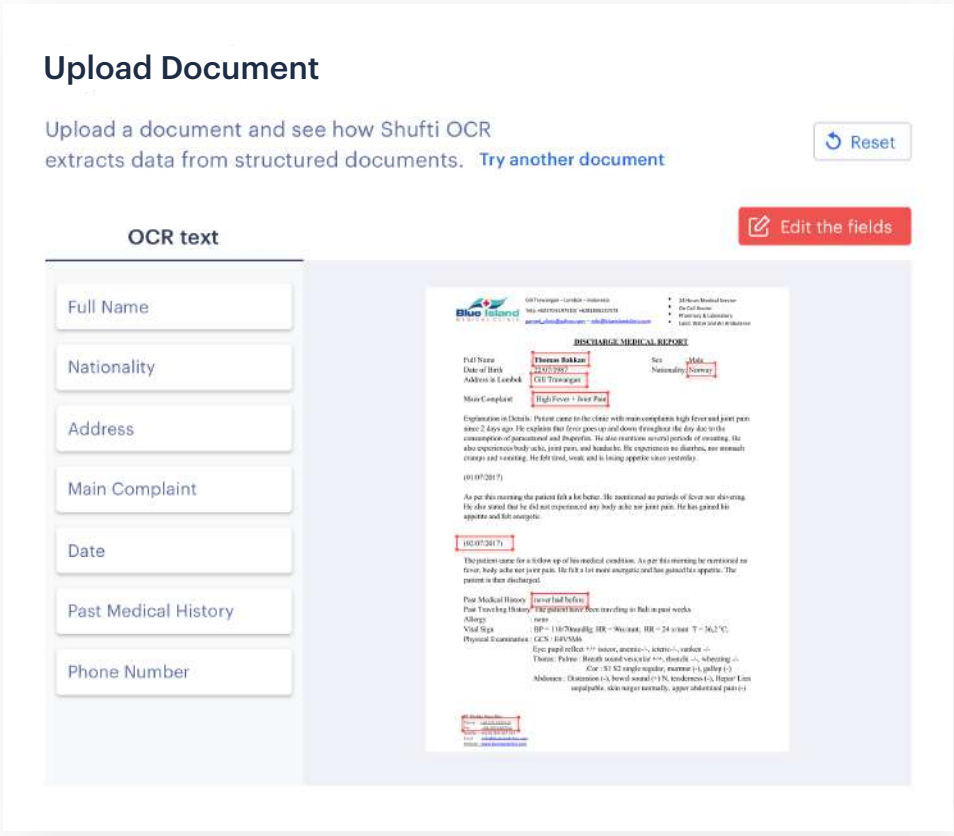
Sometimes the text may be in stylized bitmaps, making it more difficult to detect. Deep learning neural networks are used to detect such texts. The software attempts to find text-block structures, separation of paragraphs, creation of lines, and finally, recognition of characters. This operation is important as character recognition works one line at a time.

Step 2: Feature Extraction

The OCR tries to detect geometric features like curves, straight lines, holes, different loops, and similar. More modern OCR systems replace feature extraction with deep learning, in which neural networks can recognize features on their own, without human inputs.

Post Processing

This step actually serves as a learning curve for machine learning and deep neural network models used for extracting the text. The data extracted helps the machine learning algorithms to learn by extracting text in different font styles, sizes, and different types of documents.



Different Data Formats for Which AI-Based OCR can be used to Extract Data

Structured Documents

The documents with a predetermined format and composition are structured documents. OCR can extract information from structured documents with the help of artificial intelligence and machine learning. Image to text structured OCR just requires an image of the document and extracts the required information such as name, nationality, date of birth, issue date, MRZ code, picture, etc.

Some Structured Documents:

- ID cards
- Driving licenses
- Passports
- Credit/debit cards
- Utility bills

Do you Know?

Shufti Pro's OCR allows extended data extraction from structured documents to help you gather detailed insights about your customers.

Extended information may include:

- ✓ MRZ Codes
- ✓ Age
- ✓ Nationality
- ✓ Alias Name
- ✓ Authority
- ✓ Authority signature
- ✓ Authorized patronage
- ✓ Category
- ✓ Country
- ✓ Country Code
- ✓ Father Name
- ✓ Gender
- ✓ Mother Name
- ✓ City
- ✓ Code
- ✓ License class
- ✓ Logo
- ✓ Marriage immigration
- ✓ Municipality
- ✓ Office
- ✓ Particularities
- ✓ Permit type
- ✓ Previous type
- ✓ Processing center
- ✓ Race
- ✓ Emergency contact
- ✓ Employer id
- ✓ Occupation
- ✓ Personal Number
- ✓ Place of Birth
- ✓ Place of issue
- ✓ Birth certificate number
- ✓ Blood group
- ✓ Booklet Number
- ✓ Citizenship
- ✓ Middle Name
- ✓ Personal Number
- ✓ Cast
- ✓ Restrictions
- ✓ Certificate date
- ✓ Collectorate
- ✓ Conditions
- ✓ Country of stay
- ✓ Country state
- ✓ Document Type
- ✓ Date of payment
- ✓ Degree of Comanche
- ✓ Disability
- ✓ District
- ✓ Division
- ✓ Domicile
- ✓ Donor
- ✓ Donor status
- ✓ Endorsement
- ✓ Finger print
- ✓ Relationship
- ✓ Family name
- ✓ Guardian name
- ✓ Replaced date
- ✓ Residence country
- ✓ Resident date
- ✓ Resident since
- ✓ Restriction
- ✓ Revised card date
- ✓ Section
- ✓ Signature
- ✓ Social security number
- ✓ Sponsor
- ✓ Sponsor rank
- ✓ Sponsor service
- ✓ State
- ✓ Status
- ✓ Sub location
- ✓ Identification mark
- ✓ Telephone number
- ✓ Vehicle classification
- ✓ Vehicle restriction number
- ✓ Invoice
- ✓ voting number
- ✓ weight

Semi-Structured Documents

Semi-structured documents are those in which the location of the information and of the data fields can vary from document to document. They can be termed as a bridge between structured and unstructured documents. These document types do not follow a defined format the way structured forms do, and they are not bound to specified data fields either. They do, however, follow a common format which makes it easier to automate them as compared to unstructured documents.

Some Semi-Structured Documents:

- Invoices
- Purchase orders
- Emails
- Sales orders

OCR technology is vital for the automatic processing of such documents. It uses document automation technology to identify text within the document images enabling businesses to fully automate recurring data entry and management operations.

Unstructured Documents

There always exists a confusion when it comes to differentiating between semi-structured and unstructured documents. This happens mostly because the difficulty level in extracting data from both of the documents is similar. The main difference between both is the level of standardization and density of the information. An unstructured document does not follow any defined format and the placement of information in such documents is not specified as well. For example, a legal agreement, which is a type of unstructured document, there are always dates, terms, definitions, and parties involved, but there is always a variation in these things depending upon the negotiation process and the needs of the people involved. The location of the information in such documents is typically not standard.
AI-based OCR analyzes the data in these documents and after developing a context, extracts the required information with more than **90% accuracy.**

Some Unstructured Documents:

- Photos
- Social media contents
- PDFs
- Hand-written documents

Multilingual Documents

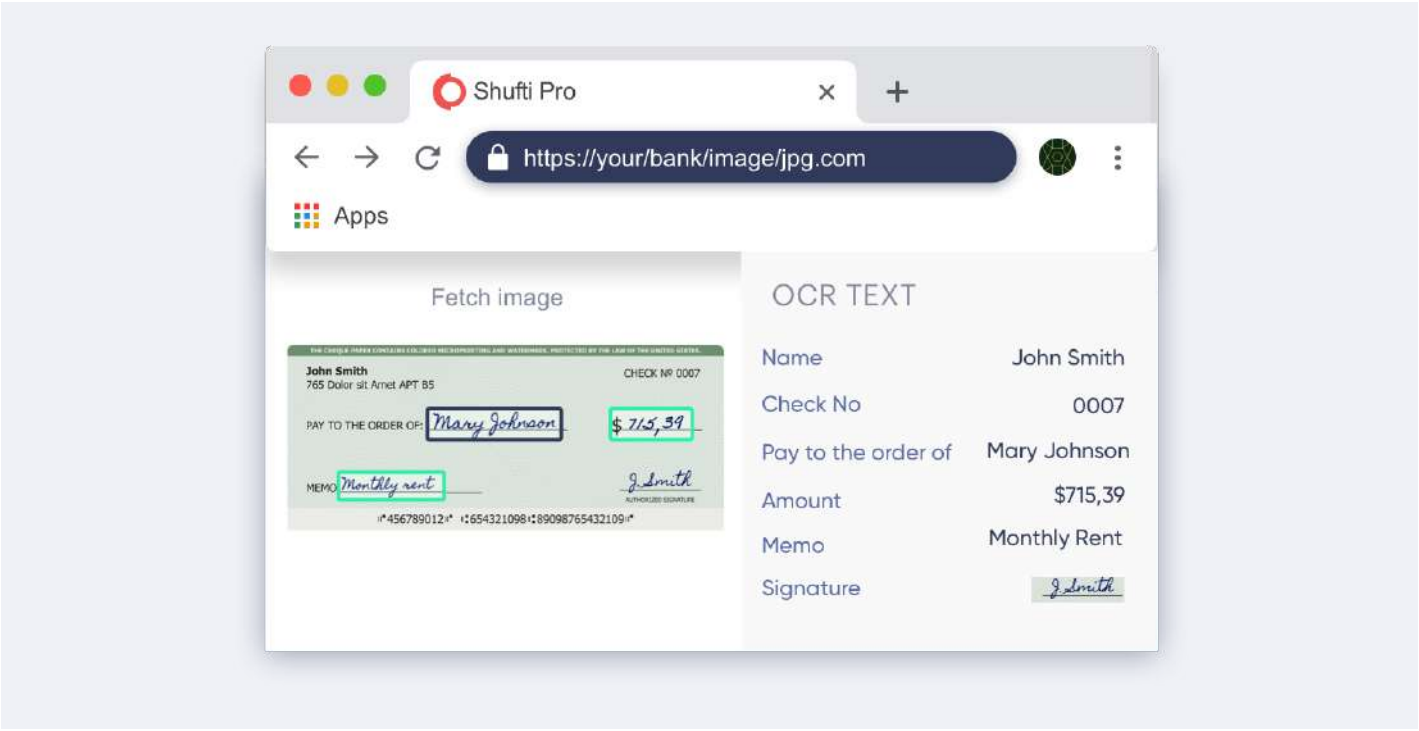
Aside from the difference in the formats or data placement, different documents may contain text written in more than one language. For example, a publication that contains the same content in two or more language types, such as airline magazines, or multilingual identity documents, etc. Doing a manual pre-sorting of information is probably not an option when dealing with multilingual documents.

AI OCR technology uses language data and dictionaries to achieve high recognition quality during the process of data extraction. OCR identifies the language of the document uploaded and extracts the data with perfect accuracy.

The multilingual documents or data may exist in the form of images, scanned files, pdf documents, web pages, etc. To perform Optical Character Recognition (OCR) on such multilingual documents, it is essential to identify different languages of the input document.

Image URL

Usually, we have to provide a scanned copy of a document or uploaded images for OCR to analyze and extract data from them. But advanced OCR engines, just like Shufti OCR also have the ability to extract data from simple URLs as well. For that users have to provide the URL (Google Drive, Dropbox, etc.) of an image they want to extract data from, and the software will fetch the image from the URL and after that extracts data from that image.

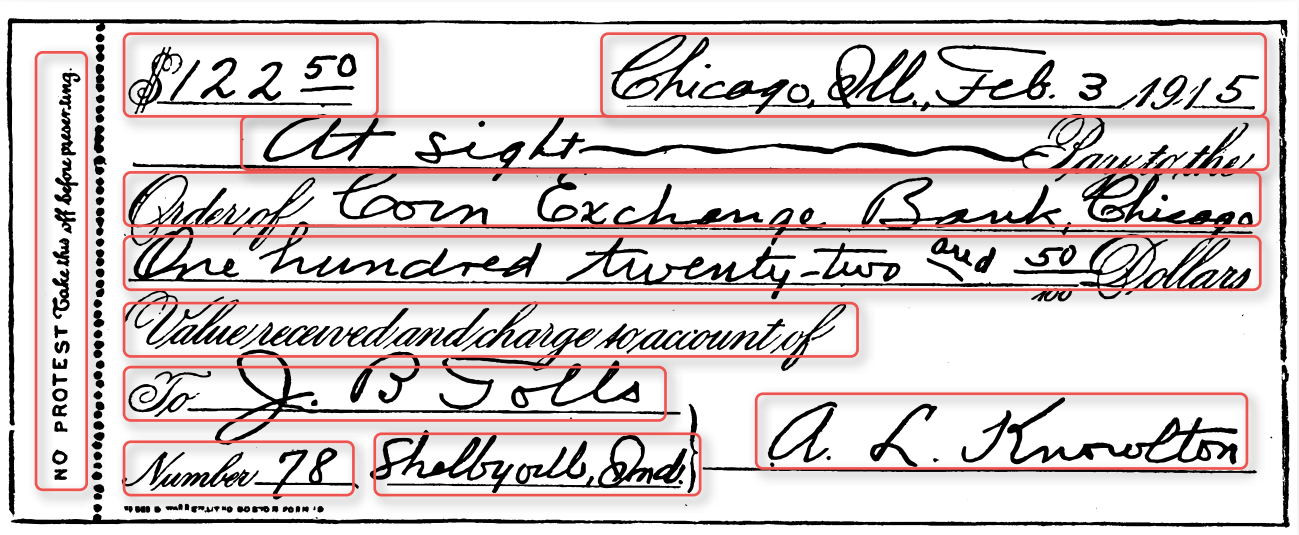


Annotated Documents

Annotations can be notes, comments, explanations, or different types of external remarks that can be attached to a selected part of a document or web document, to signify the need for a correction or just a suggestion.

Such documents can be termed as annotated documents. As the annotations are external, it is possible to add them to any web document independently, without having the need to edit the document or data itself. In technical form, annotations are usually seen as metadata, which gives additional information about an already existing piece of data.

Annotated documents can be found in the form of images, scanned copies, or web documents. To extract data from such documents, you just have to upload the file for the OCR engine to analyze and fetch your specified information for you. It can help in extracting selective data e.g circled text, highlighted content, MRZ codes, or just the handwritten content.



AI-Powered OCR Can Automate the Workflow in Numerous Industries

OCR is one of the technologies that has its application throughout the entire industrial spectrum. With OCR, a huge number of paper-based documents either scanned, printed, or hand-written can be transformed into machine-readable text making data storage and accessibility simple and efficient.

Here are a few industries that could make use of OCR engine;

Financial Services

(Banking, Accounting Firms, Credit and Insurance Companies)

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

- Claims handling
- Bookkeeping
- Mortgage processing
- Account management

Health Care

- Billing and claims processing
- Insurance processing
- Maintaining patient records

Manufacturing

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

Retail/e-Commerce

- Account opening

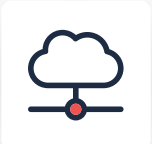
- Invoice and bill processing
- Customer’s document verification

Government Sector

- Immigration applications
- Education system applications
- Passport management applications

Human Resource

- Employee onboarding
- Extracting key data from candidate CVs
- HR records processing
- Storing legal documents (employment agreements, HR policies, meeting notes, etc.)



SaaS

- Expanding online storage solutions
- Document conversion in cloud storage
- Document verification and data extraction

How Businesses Could Benefit from AI OCR Technology?

Automated OCR technology has eased business workflows and operations. OCR based data extraction requires lesser resources, time, and effort. Businesses can reduce the resources required for manual data processing as well as save time in data processing and data management. Having higher accuracy AI OCR engines can help firms in extracting data from multiple documents either in electronic or paper formats. With the world moving towards digitalization, businesses around the globe need to adopt an intelligent solution to automate their workflows, reduce human efforts, speed up the processes, and cater to the customers in the most efficient way possible.

These are some benefits that AI OCR engine brings to businesses;

- Reduce Cost:

  Automating the document extraction process reduces the extra cost required to manually maintain all the record-keeping process. According to Import.io, automating the manual data entry could reduce cost upto 66% for small and medium-sized businesses.

- Reduce Manual Identification:

  Manually maintaining the data entry is a long and tiring process and is also prone to human errors.

Replacing manual data extraction with an automated AI-powered OCR engine reduces the probability of human error to a minimal level.

- Save Human Resources:

  The number of employees to manually enter the customer details in sheets can be reduced by integrating OCR services to their system. Reduction in human resources helps cut operational costs.

- Improved Business Productivity:

  Robust data collection process eventually improves business productivity. The resources that previously used to perform operation manually can be utilized in other useful work.

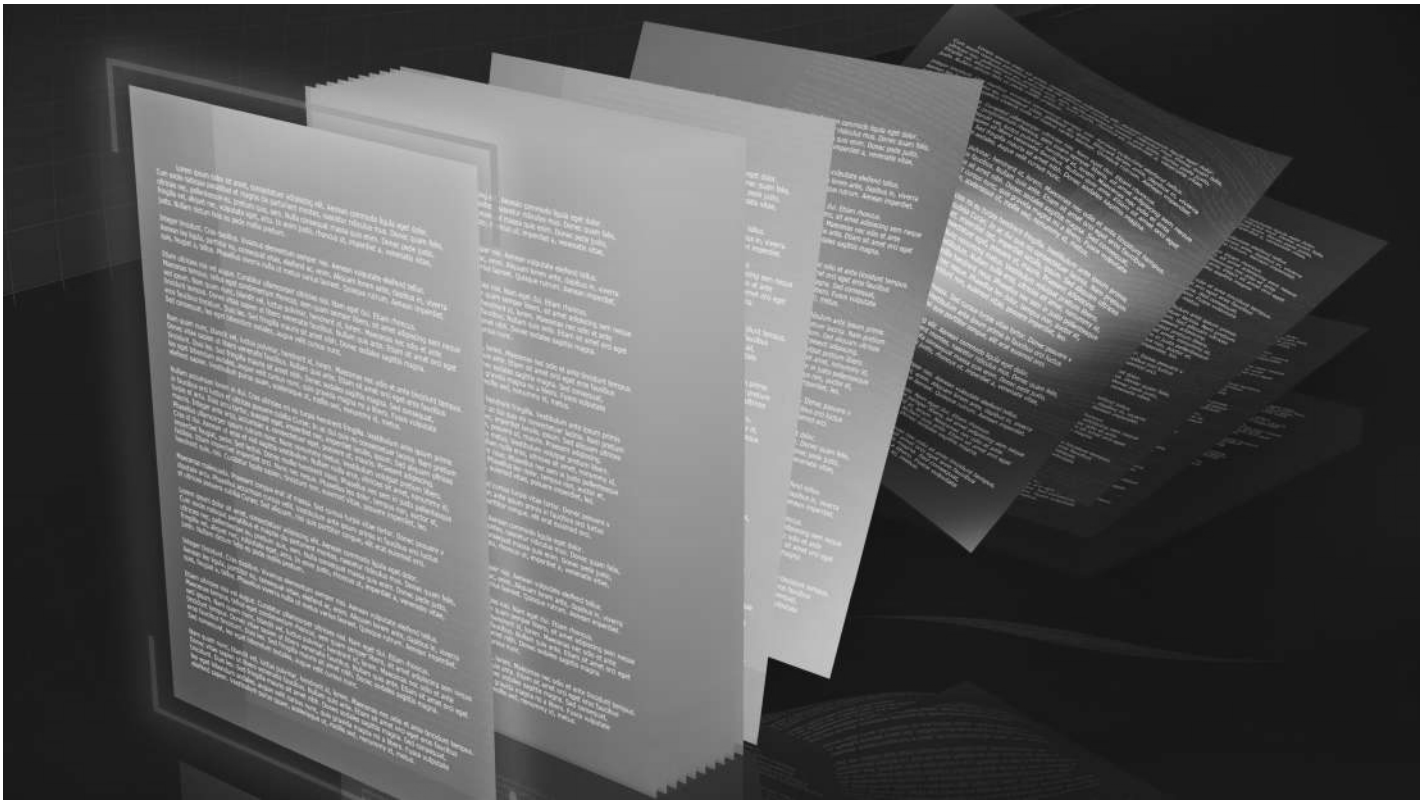
- Automated Content Processing:

  AI OCR engine extracts the information from the documents and fills the data in the form accordingly. For instance: during the identity verification process the customer's data is extracted from the identity document and is automatically filled in the form.

AI OCR engine employs algorithms and techniques that help businesses utilize their data as a competitive advantage by effectively embracing digitization. AI OCR engines employ algorithms and techniques that help businesses to effectively embrace digitalization with all their data available to them in digital formats to compete in the digital world.

For instance, in the banking industry, a lot of work is paper-based or requires manual data entry. From customer onboarding to keeping policies intact, the manual data entry process is required, given the sensitivity of the information, human errors could result in huge financial and reputational losses to the bank. By embracing AI OCR engine banks could simply automate the workflow and reduce the chase of error.

Consumers are demanding digital solutions so online banking emerged with a variety of online services. In almost every industry, data extraction and management have become a critical task and require a technology that could be cost-effective, efficient, and error-free at the same time. AI-based OCR is an all-in-one solution that could streamline the data extraction process with ease.



Automate Your Business Operations with Shufti Pro's AI-Powered OCR

Shufti Pro is keen to enhance the digitization process for companies that want to go digital. With its remarkable accuracy of above 90% and secure data management, Shufti Pro's AI-powered OCR engine can help businesses to automate their data extraction process. In mere seconds, the banking industry, e-commerce, digital payment services, and many more can extract out the user information from any type of document by taking advantage of OCR technology. Just upload a document and the AI OCR engine will extract all the relevant data for you. You can easily access, sort, download, or delete the data in your back office anytime.

Here are some key benefits of adopting Shufti Pro's AI OCR;

- Banking grade data Security
- High Accuracy
- Cloud Storage
- Optimized data extraction
- Global Support
- Multi-document support

Shufti Pro's AI OCR engine has the ability to extract data from multiple languages including some of the most difficult languages (Arabic, Chinese, Urdu,etc.) and multiple types of documents both structured and unstructured. With unprecedented global coverage in 230+ countries and supporting 150+ languages, businesses could benefit greatly Shufti Pro's AI-based data extraction engine.

Do you want to learn more about how Shufti Pro can help your company enhance its document processing capabilities?

Discuss it with our experts

www.shuftipro.com

sales@shuftipro.com

Resources

1 https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/four-fundamentals-of-workplace-automation

2 https://mydatascope.com/blog/en/2018/03/08/how-much-paper-waste-is-costing-your-business/

3 https://businesspartnermagazine.com/should-your-business-go-paperless/

4 https://www.globenewswire.com/news-release/2018/01/18/1296235/0/en/6-Key-Insights-on-Business-Workflow-Automation-Market-for-Forecast-Period-2017-2026.html

5 https://itchronicles.com/technology/repetitive-tasks-cost-5-trillion-annually/

6 https://link.springer.com/content/pdf/10.1007%2F3-540-45869-7_49.pdf

Shufti Pro True Identity Builds Trust

Expanding services to 230+ countries and territories in a short period of time, Shufti Pro envisioned playing a pivotal role in creating cyberspace where every transaction is verifiable and secure. With enough experience in technologies like machine learning (ML), OCR, artificial intelligence, and Natural Language Processing (NLP), Shufti Pro strives to provide the best identity verification services to verify customers and businesses online.

Shufti Pro's cost-effective solutions help businesses to prevent fraud and illicit crimes that can ruin the integrity and brand reputation of your business. Our perfect solution suite consisting of KYC verification, AML screening, ID verification, Facial Recognition, Biometric Authentication, Video KYC, OCR, and KYB helps to improve your company's fraud prevention, Know your Customer (KYC) and Anti Money Laundering (AML) regulatory efforts by automating the workflow. With single API integration, Shufti Pro empowers you to verify customers with document checks from 3000+ ID templates and business entities from 200 million companies data.