REXTRACT

AUTOMATED KEY-VALUE PAIR INFORMATION EXTRACTION SYSTEM FOR INTELLIGENT DOCUMENT PROCESSING

TeamPREV

Pradeep Jankiraman – A0140188H

Roy, Chiu Man Shan – A0249252A

Ethan, Kuch Swee Cheng – A0249264X

Vikram Sankireddypally – A0249306A

March 25, 2023

TABLE OF CONTENTS

TABLE OF CONTENTS

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

PROJECT OVERVIEW

The purpose of this project is to develop a computer vision solution for scanning images containing key-value pairs and extracting the information to a JSON format. This solution is designed to improve the efficiency of logistics and manufacturing processes by automating the extraction of key information from shipping labels, packaging information and component labels.

PROJECT BACKGROUND

In the logistics and manufacturing sectors, there is a need for efficient and accurate processing of large amounts of data. Some of the most time-consuming tasks is the manual extraction of key information from documents such as invoices, material picklist, packaging information, components labels and shipping labels. This process is prone to errors and can significantly impact the efficiency of these sectors.

To address this challenge, computer vision technology can be used to automate the extraction of key information from these documents. This technology can accurately recognize text and identify key-value pairs, making it an ideal solution for optimizing the logistics and manufacturing processes.

EXECUTIVE SUMMARY

PROJECT DESCRIPTION

The project leverage on multiple computer vision algorithm that can accurately identify key-value pair on images and extract them to JSON format. The algorithm will use deep learning techniques such as Text Detection, Text Recognition and Handwritten Expression Recognition alongside OCR (Optical Character Recognition) to recognition text and identify key-value pairs from the image.

The key-value pair will be defined by the user and can be customized according to the specific needs of the logistics or manufacturing sector processes.

- For manufacturing sector, a key-value pair for a component label may include date code, manufacture part number, expiration date and country of origin.
- For logistic sector, a key-value pair for an invoice may invoice the vendor's name, invoice number and invoice date.

The algorithm will be trained on a sector specific dataset of images containing key-value pairs and will be optimized for accuracy and speed. The user will be able to upload image and a csv file to the system through their cell phone. And the algorithm will automatically extract the key-value pairs and output them in a JSON format. The extracted key-value pair could also be compared with the fields value in the csv file uploaded by the users, for data verification and validation.

BUSINESS OBJECTIVE

BUSINESS OBJECTIVE

BUSINESS USE CASE

ReXtract has the potential to be used in a variety of manufacturing and logistics application such as:

- **Label Verification**: The solution can be used to verify the accuracy of component/packaging labels, such as barcodes, lot numbers, and expiration dates. This can ensure compliance with regulations and prevent costly mistakes.
- **Assembly Verification:** The solution can be used to verify the accuracy of product assemblies, such as the correct placement of components during the material kiting process. This can ensure that products are assembled correctly and reduce downtime during the assembly production process.
- **Inventory Management:** The solution can be used to scan and track inventory levels and locations in a warehouse or production facility. This can help streamline inventory management and reduce the risk of stockouts or overstocking.
- **Data Entry:** The solution can be used to extract and enter key information from documents, such as invoices, work orders or production reports, into SAP or other enterprise level software. This can save time and reduce errors in data entry processes.
- **Packaging Label Processing:** The solution can be used to extract key information from packaging labels, such as date code, manufacture part number, expiration date and country of origin, which can be used to automate label verification process.
- **Shipping Label Processing:** The solution can be used to extract key information from shipping labels, such as tracking number, carrier name, and delivery date, which can be used to automate the shipping and receiving process.

BUSINESS OBJECTIVE

BUSINESS MODEL

ReXtract will be sold as a software-as-a-service (SaaS) solution. Customers will be charged a monthly or annual fee based on the number of users and the amount of data processed.

MARKET RESEARCH

Global intelligent document processing market is expected to witness significant growth over the forecast period from 2022 to 2030, driven by factors such as increasing investments in digital transformation and the need for cost-effective and efficient document processing solutions. The COVID-19 pandemic has also had a positive impact on the market, with businesses across industries increasing investments in automation and adopting technologies such as Artificial Intelligence (AI) in response to the pandemic.

PRODUCT CONCEPT AND DEVELOPMENT

Based on technology, the market is segmented into machine learning, Natural Language Processing (NLP), and computer vision, with the machine learning segment accounting for the largest revenue share in 2021. By deployment, the cloud segment accounted for the largest revenue share in 2021, while the on-premises segment is likely to capture significant growth over the forecast period.

Finally, the BFSI segment currently accounted for the largest revenue share in 2021, with intelligent document processing solutions serving many use cases in this sector, such as KYC, mortgage applications, and bank statements. However, there is an opportunity to trickle this intelligent document processing technology to the next market segment such as manufacturing and logistic.

GO-TO MARKET STRATEGY

GO-TO MARKET STRATEGY

TARGET MARKET

ReXtract solution can be customized to meet the specific needs of each company, making it a valuable tool for a wide range of applications across different industries.

MARKETING STRATEGY

ReXtract will focus on creating awareness of the solution through targeted advertising on social media platforms, industry publications, and trade shows. The company will also leverage partnerships with software providers and consultants to reach a wider audience and market groups.

PRICING STRATEGY

ReXtract will be sold as a software-as-a-service (SaaS) solution. Customers will be charged a monthly or annual fee based on the number of users and the amount of data processed. The pricing strategy will be competitive with other solutions on the market, while still providing value to the customer.

SALES STRATEGY

The sales team would be focus on building relationships with key decision-makers in logistics and manufacturing companies as the next step of growth, while engaging other industry needs. Sales representatives will target these decision-makers through phone calls, email outreach, and in-person meetings. The company will also offer a free trial period to allow potential customers to test the solution before making a purchase.

GO-TO MARKET STRATEGY

LAUNCH PLAN

Targeted advertising campaign on social media platforms and industry publications. The company will also participate in trade shows to create awareness of the solution. Sales representatives will reach out to key decision-makers in logistics and manufacturing companies to offer a free trial period and demonstrate the value of the solution.

GROWTH PLAN

Our business focus would be on expanding the customer base and increasing revenue. The company will continue to build relationships with key decision-makers in logistics and manufacturing companies, while also targeting new industries and markets. The solution will be further developed to include other types of documents and data and integration with other software and systems used in the logistics and manufacturing sectors.

SYSTEM DESIGN

SYSTEM DESIGN

REXTRACT SYSTEM DESIGN

ReXtract will consist of the following component to achieve key-value pair information extraction from the image:

- **Image Acquisition**: The system will acquire images containing key-value pairs from a variety of sources, including scanned documents, digital images, and photographs acquired using mobile phones.
- **Image Preprocessing**: The acquired images will undergo preprocessing to enhance their quality and improve the accuracy of the OCR process. This may include noise reduction, image normalization, and contrast enhancement.
- **OCR engine**: The engine will use deep learning techniques such as Text Detection, Text Recognition and Handwritten Expression Recognition alongside OCR (Optical Character Recognition) to recognition, identify and extract key-value pairs from the image. The OCR engine will be trained on a dataset of images containing key-value pairs to optimize its accuracy.
- **Key-Value Pair Extraction**: The key-value pairs will be extracted from the OCR output using pattern matching and rule-based techniques. The extraction process will be customized according to the specific needs of the logistics or manufacturing process.
- **Data Output**: The extracted key-value pair will be output in a JSON format for use in downstream applications, such as ERP or MRP systems. The data would be stored in a centralized repository for ease of data management and analytics. The data could also be used for field verification against a user uploaded csv file format contain the key (e.g., Manufacture Part No. or Invoice No.)
- **Data Verification**: The extracted data could also be used for field verification/comparison against a user uploaded csv file format contain the key (e.g., Manufacture Part No. or Invoice No.), removing the need for manual eyeballing of data from packaging labels.

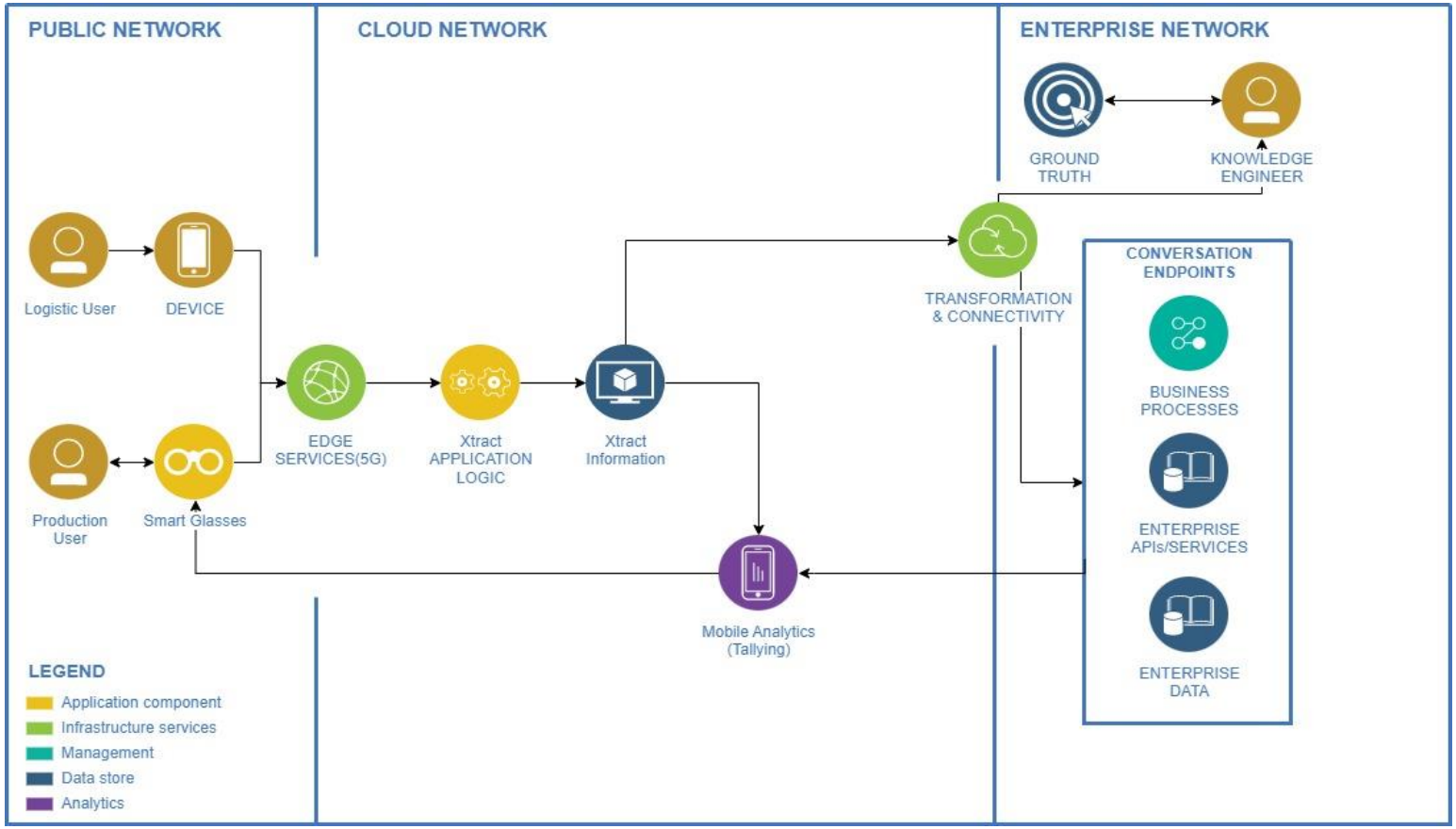
SYSTEM DESIGN

SYSTEM MODEL

The system model for ReXtract will consist of the following components:

- **Image Acquisition Module**: This module will acquire images containing key-value pairs from a variety of sources, including scanned documents, digital images, and photographs.
- **Preprocessing Module**: This module will preprocess the acquired images to enhance their quality and improve the accuracy of the OCR process.
- **OCR Engine**: This module will use deep learning techniques to recognize text in the image and identify key-value pairs.
- **Key-Value Pair Extraction Module**: This module will extract the key-value pairs from the OCR output using pattern matching and rule-based techniques.
- **Data Output Module**: This module will output the key-value pairs in a JSON format for use in downstream applications, such as ERP or MRP systems.
- **User Interface Module**: This module will provide a user interface for interacting with the system, including uploading images, customizing key-value pairs, and viewing the extracted data.
- **Database Module**: This module will store the extracted data for future analysis and use in downstream applications. Allowing users to update and retrieve from the enterprise datab

SYSTEM DESIGN



SYSTEM DESIGN

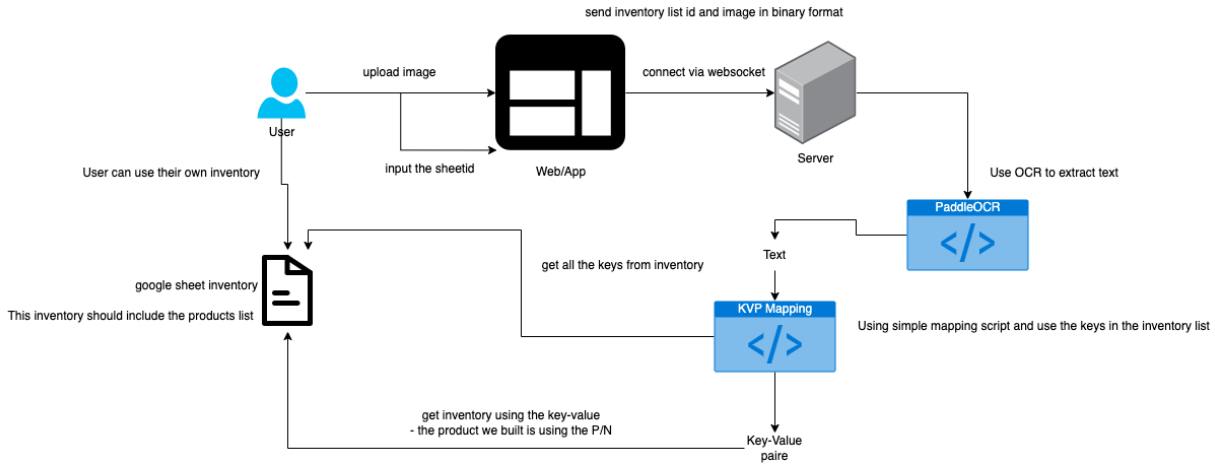
SYSTEM DESIGN

Insert system design information such as process map.(Ethan)

Other System design information here!

SYSTEM DESIGN

TECHNICAL IMPLEMENTATION



The proposed system aims to provide a portal that can be accessed via mobile or laptop devices, connected to a WebSocket that allows for rapid response times. Users will be able to upload images to the portal, which will be processed by a server-side application utilizing the PaddleOCR tool to extract text and its corresponding coordinates. The application will also access an inventory database containing predefined keys and products, such as "P/N" and "Product Name". The extracted key-value pairs will be matched using the closest coordinate pairs, and further validated using the inventory data to ensure correct formatting. Finally, the key will be used to search the inventory database, returning product details to the user.

To achieve this, a user interface will need to be designed to allow for image uploading, while the server-side application will need to be developed to receive and process the images. An inventory database will also need to be created, and the keys and products defined. The application will need to be able to use the coordinate data to accurately match key-value pairs, and further validate the extracted data against the inventory data.

CONCLUSION

This project will require expertise in user interface design, server-side programming, and database management, but the resulting system will be a powerful tool for extracting information from images and searching an inventory database.

CONCLUSION

PROJECT CONCLUSION

Rextract is a tool that is developed to intelligently extract key words in typical documents that are prevalent in logistic applications. A typical document would have some key information, for example, invoice numbers and serial numbers which need to be extracted and checked against existing databases. Rextract simplifies this process by automating the process of manually checking documents and cross checking the database.

The solution implemented is simple and effective which allows it to be showcased as a Proof-of-Concept and is potentially scalable in a feasible way in large organizations.

PROJECT FINDING

Two main concerns were identified in this project:

1. Image Quality: The application is sensitive to the quality of the image. Image enhancements techniques can be used upstream to address this issue.
2. Labelling Standards: Different documents have different standards for the same labels, which makes it difficult to use simple methods to extract key label and value pairs. Potentially, other machine learning methods can be augmented to the current implementation to address non-standardization of labels.

ANNEX

ENVIRONMENT AND DEPENDENCIES

Environment

Python3.9

Dependencies / Packages

ANNEX

paddlepaddle

websockets

numpy

gspread

INSTALLATION GUIDE

1. Install the dependecies `pip install -r requirement.txt`

2. Need to get a service account from google ([Doc](https://cloud.google.com/iam/docs/service-accounts-create "Doc"))

3. Put the exported json file under the Systemcode directory

4. Update the `start.sh` first two lines to be your `HOST` and `PORT`. Default host is `127.0.0.1`, port is `8081`

5. Run `sh start.sh`

6. A page will be opened up and wait for the light change to green and you can start using it

ANNEX

APPLICATION GUIDE

This is the portal page. You can change the inventory list by changing the `Google Sheet Inventory:`. For this guide, i am using [this public sheet](https://docs.google.com/spreadsheets/d/1BiNltufbq1F4iW5SKrtP_QaQ3wxPYPhsfGPBlyvBR7g/edit?usp=sharing "this public sheet")

To start, you can click `choose file` and upload a label. For this sample, i am using `web/images/test2.jpg`. There would be a preview for the uploaded image.

Click `Recognize Text` to start the process. After processing, the system would help you to find the item in the inventory and also display all the texts detected from the image.

Note: If you need to test it in mobile, please use a valid `HOST` instead of localhost as well as hosting the webpage (`web`) in a server.

PROJECT MEMBER INFORMATION

PROJECT MEMBER INFORMATION

- **Pradeep Jankiraman**
  Chief Executive Officer

- **Roy, Chiu Man Shan**
  Chief Technology Officer

- **Ethan, Kuch Swee Cheng**
  Enterprise Solutions Architect

- **Vikram Sankireddypally**
  International Business Director

PROJECT MEMBER INFORMATION