JPMorgan Chase & Co.

Columbia University Data Science Institute

# Spring 2022 Capstone Project

# Table Extraction via Eye Gaze Tracking

First Progress Report

Shihang Wang (sw3275), Yeqi Zhang (yz3975), Yibai Liu (yl4616), Yijia Jin (yj2682), Yinqiu Feng (yf2579)

March 13, 2022

# INTRODUCTION

## Project Scope and Objective

This Capstone project is a collaboration between Data Science Institute of Columbia University and the AI Research team of JPMorgan & Chase. The objective of this Capstone Project is to utilize Eye Gaze technology and design a system to automatically extract tables from a document in the image file format by leveraging Computer Vision (CV) and Optical Character Recognition (OCR). Automating this process would allow an employee to extract the data in the table(s) of interest from a report by simply looking at the tables, and it would be more efficient compared to screenshot capturing and manual data entry.

*Keywords*: Table Extraction, Computer Vision, OCR

## Progress Overview

The project problem can be divided into two tasks. Task 1 is primarily identifying the Area of Interest (AOI), the regions of a displayed document where the person's eyes fix on, by tracking eye movements and gaze points. Task 2 comprises extracting tables from image documents and recognizing data inside the tables.

So far, we have made solid progress in this project. We paralleled the two tasks since they utilized different data sources, and we carried out experiments and finished data cleaning, wrangling, and exploratory data analysis (EDA) and visualization. Also, we have built a table extraction architecture and trained the baseline model. Next, we will concentrate on the implementation of the model, addressing existing problems and improving model performance.

# METHODS

1. **Eye gaze experiment data collection**

For Task 1, we designed a webcam-based eye tracking experiment in Python using the open-source toolbox PyGaze, a Python wrapper for Gazepoint's OpenGaze API, to establish connection with the Gazepoint equipment. In an experiment trial, 10~15 image-format documents were displayed in front of the participant after calibrations, and the participant was asked to look for table(s) in the image and gaze at one table of interest for 10 seconds per image. The Gazepoint eye tracker recorded the exact position of the pupils, the fixation duration and coordinates, the gaze direction, etc., which provides information about eye movements and areas of interest.

2. **AOI detection based on experiment data**

Since tables in the documents are borderless, generating accurate bounding boxes solely relying on AOIs is not easy to achieve. We have so far followed our initial idea of using the intersection of the AOI bounding box and the box extracted by CV modeling, so if there are multiple tables in a document and the user only wants a specific one, AOI would be able to tell which table to extract.

3. **Table extraction with CV Object Detection**

In the experiments, we utilized IBM's open source dataset FinTabNet which comprises financial table data with PDF and JSON format. The JSON files provide ground truth position of tables and information about rows, columns, and cells. We selected 2,000 documents from the huge dataset and converted PDF documents into images. Then we leveraged computer vision to achieve table extraction with Object Detection architectures.

# RESULTS

## 1. Exploratory Data Analysis (EDA)

To better understand the data, we first carried out exploratory data analysis on the raw experiment data including duplicate value checking, the removal of unnecessary attributes, missing value pattern detecting and correlation analysis. Complete preprocessing process can be found in the code repository shown above.

### 1.1 Missing Values

| | TIME | TIME_TICK | FPOGX | FPOGY | FPOGS | FPOGD | FPOGID | FPOGV | LPOGX | LPOGY | LPOGV | RPOGX | RPOGY | RPOGV | BPOGX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **13** | 3689.03662 | 1203287588860 | 0.0 | 0.0 | 3688.03467 | 0.09863 | 1754 | 0 | -2.11646 | 6.04031 | 0 | 0.64409 | 1.24761 | 0 | -2.11646 |
| **14** | 3689.05322 | 1203287754931 | 0.0 | 0.0 | 3688.03467 | 0.09863 | 1754 | 0 | -2.11646 | 6.04031 | 0 | 0.64409 | 1.24761 | 0 | -2.11646 |
| **15** | 3689.06958 | 1203287918669 | 0.0 | 0.0 | 3688.03467 | 0.09863 | 1754 | 0 | -2.11646 | 6.04031 | 0 | 0.64409 | 1.24761 | 0 | -2.11646 |
| **16** | 3689.08618 | 1203288082609 | 0.0 | 0.0 | 3688.03467 | 0.09863 | 1754 | 0 | -2.11646 | 6.04031 | 0 | 0.64409 | 1.24761 | 0 | -2.11646 |
| **17** | 3689.10254 | 1203288246529 | 0.0 | 0.0 | 3688.03467 | 0.09863 | 1754 | 0 | -2.11646 | 6.04031 | 0 | 0.64409 | 1.24761 | 0 | -2.11646 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **3640** | 3749.62158 | 1203893437831 | 0.0 | 0.0 | 3748.47656 | 0.26294 | 1906 | 0 | -0.13194 | 0.36571 | 0 | 4.02696 | -0.21120 | 0 | 4.02696 |
| **3641** | 3749.63354 | 1203893557871 | 0.0 | 0.0 | 3748.47656 | 0.26294 | 1906 | 0 | -0.13194 | 0.36571 | 0 | 4.02696 | -0.21120 | 0 | 4.02696 |
| **3642** | 3749.64990 | 1203893719918 | 0.0 | 0.0 | 3748.47656 | 0.26294 | 1906 | 0 | -0.13194 | 0.36571 | 0 | 4.02696 | -0.21120 | 0 | 4.02696 |
| **3643** | 3749.66772 | 1203893899312 | 0.0 | 0.0 | 3748.47656 | 0.26294 | 1906 | 0 | -0.13194 | 0.36571 | 0 | 4.02696 | -0.21120 | 0 | 4.02696 |
| **3644** | 3749.70898 | 1203894311222 | 0.0 | 0.0 | 3748.47656 | 0.26294 | 1906 | 0 | -0.13194 | 0.36571 | 0 | 4.02696 | -0.21120 | 0 | 4.02696 |

460 rows × 40 columns

Figure 1: Consistency of Missing Value between FPOGX and FPOGY

As shown above, we filtered out the records with missing fixation coordinates FPOGX and FPOGY (the value 0.0 as missing) and found that there was a consistency between two columns. Since fixation was calculated based on both eyes, the missingness of these columns could be attributed to blinking as the equipment was unable to trace participants' eyes while they were closed. Based on FPOGID (the fixation POG ID number), it is possible to compute the frequency of blinking via regular intervals as follows:

Frequency of Blinking = Total Time / The Count of Blinking

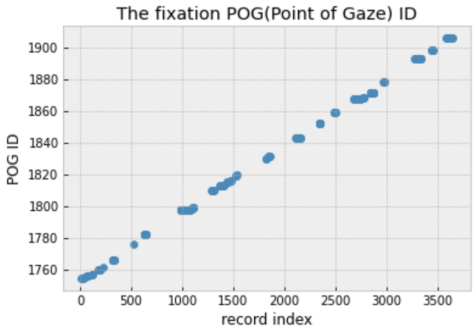



Figure 2: The Fixation POG ID Distribution

By experiment time interval and count of unique FPOGID, the frequency of blinking, as calculated, is 2.02 (experiment participant will blink per 2.02 seconds).

### 1.2 Correlation

With some explorations, we observed tha some features share a similar distribution, which can be considered together. We used heatmap to explore the correlation between each two features.

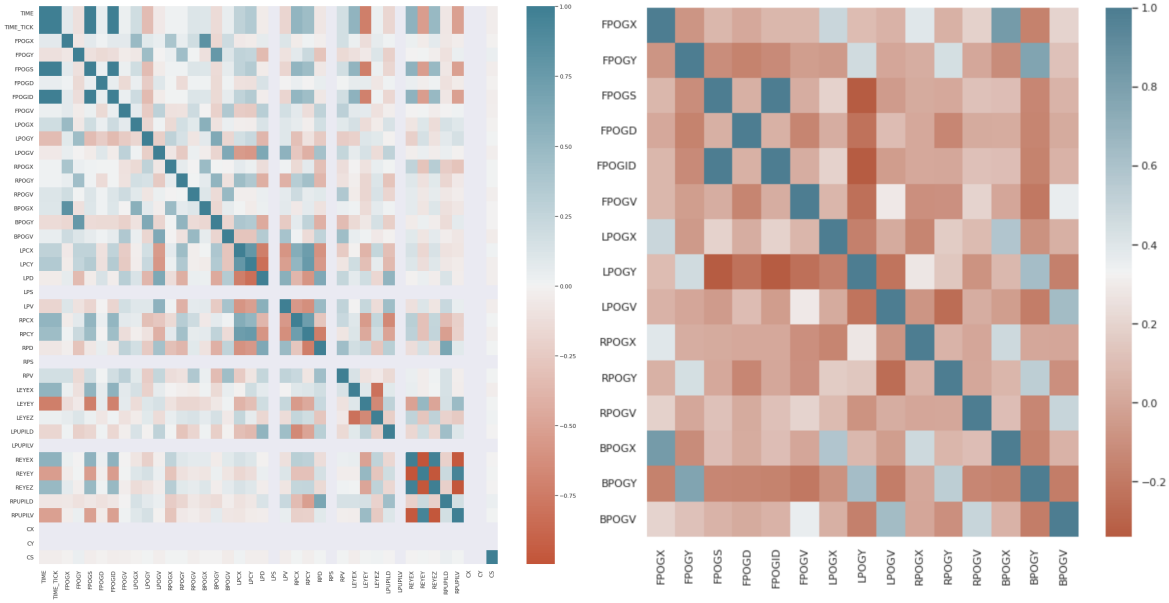



Figure 3: Heatmap of Correlations between features(left) and zoom-in heatmap for features of our interests(right)

With the help of correlation heatmap (Figure 3), we examined correlation between each two features and summarize our observation as following:

### 1.3 Fixation Point and Best POG

We observed that, among all possible pairs, the highest positive correlations are found between FPOGX and BPOGX(fixation of and best of x coordinate value), and FPOGY and BPOGY(fixation of and best of y coordinate value), FPOGV and BPOGV(fixation and best valid flag). This shows that the best POG points are probably filtered from the fixation points. We do not know how “gazepoint” calculate best point, but it may be closely relate to the “best” un-filtered POG (left-right average).

### 1.4 Best (Unfiltered) POG Coordinates

Additionally, we observed that pairs BPOGX and LPOGX/RPOGX, BPOGY and LPOGY/RPOGY are highly correlated. This is explanable because, as indicated in the explanation of best-POG data, the best POG is calculated through averaging unfiltered left and right POG for each coordinate.

### 1.5 Combining X and Y Coordinates Data

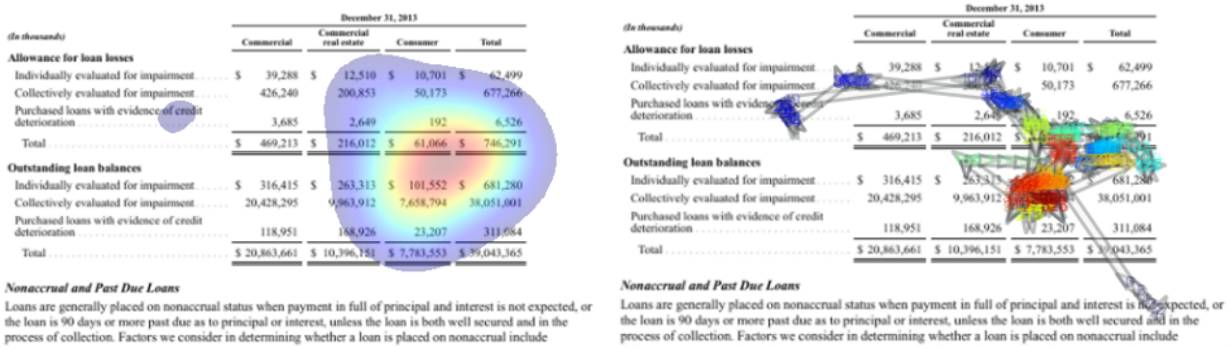
For x-coordinatesm Both BPOGX and FPOGX has a slightly higher correlation with left eye's data(LPOGX) than with right eye's data(RPOGX). This probably indicates that the averaging calculation of both best POG and fixation POG weighs left-eye data heavier than right-eye data. Our guess is that the data reflects the heavier use of the dominant eye over the other(when one use one eye more than the other, has better vision in one eye, or can fixate on something better with one eye).

### 1.6 Eyes 3D Data

At this point of our experiments, we do not find the information in eyes 3d data quite helpful; however, we observe the fixation starting time(FPOGS) and fixation ID(FPOGID) is strongly correlated with x and y coordinates of left eye with 3d space(LEYEX, LEYEY). Such correlation with the coordinates of the right eye(REYEX, REYEY) is also evident, but relatively weaker. Again, this discrepancy is likely due to the heavier use of the participant's dominant eye.

## 2. Data Cleaning and Wrangling

The data cleaning was primarily performed on the experiment data based on exploratory findings. By removing missing values, converting pixels, and trimming data points around the edges, we completed basic denoising tasks and created visualizations like heatmaps, scanpaths, and density plots.



| (In thousands) | December 31, 2013 | | | |
|---|---|---|---|---|
| | Commercial | Commercial real estate | Consumer | Total |
| **Allowance for loan losses** | | | | |
| Individually evaluated for impairment | $ 39,288 | $ 12,510 | $ 10,701 | $ 62,499 |
| Collectively evaluated for impairment | 426,240 | 200,853 | 50,173 | 677,266 |
| Purchased loans with evidence of credit deterioration | 3,685 | 2,649 | 192 | 6,526 |
| Total | $ 469,213 | $ 216,012 | $ 61,066 | $ 746,291 |
| **Outstanding loan balances** | | | | |
| Individually evaluated for impairment | $ 316,415 | $ 263,313 | $ 101,552 | $ 681,280 |
| Collectively evaluated for impairment | 20,428,295 | 9,963,912 | 7,658,794 | 38,051,001 |
| Purchased loans with evidence of credit deterioration | 118,951 | 168,926 | 23,207 | 311,084 |
| Total | $ 20,863,661 | $ 10,396,751 | $ 7,783,553 | $ 39,043,365 |

***Nonaccrual and Past Due Loans***

Loans are generally placed on nonaccrual status when payment in full of principal and interest is not expected, or the loan is 90 days or more past due as to principal or interest, unless the loan is both well secured and in the process of collection. Factors we consider in determining whether a loan is placed on nonaccrual include

Figure 4: Heatmap and Scanpath of fixation points

These plots allow us to quickly evaluate the EyeGaze data quality and also provide insights for eye movements with respect to time.

## 3. Time Series Prediction of Eye Movements

Denote that the movements of eye gazing are time dependent, where we would look next should correspond to what we are currently looking at, especially when the individual are provided with some targets, such as reading the paragraphs or looking at the graphs and tables within the document. Therefore, it's natural for us to hypothesize that when the individuals are looking through tables of a financial document image, the eye gazing movements can be predictable given the history data reflecting where the eyes are gazing at.

To do so, we used data from our previous eye gaze experiments and transformed the X, Y coordinates into [0,1] ratio form, where the original X and Y coordinates were divided by the image width and image height respectively. Then, we conduct time series analysis on this data set, to provide the first image as an example, the eye gazing movements are shown below. It can be drawn that the movements of eyes on X and Y coordinates do correspond to time.

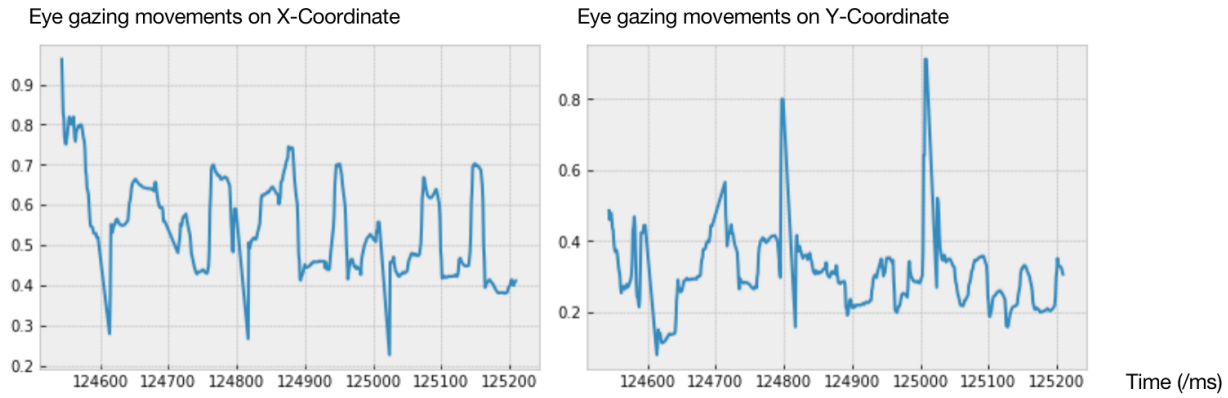



Figure 5: Eye gazing movements on X- and Y-coordinates

Then, we implemented first-order difference on each of the sequences, and found a even stronger correlation between coordinates and time, as shown below:

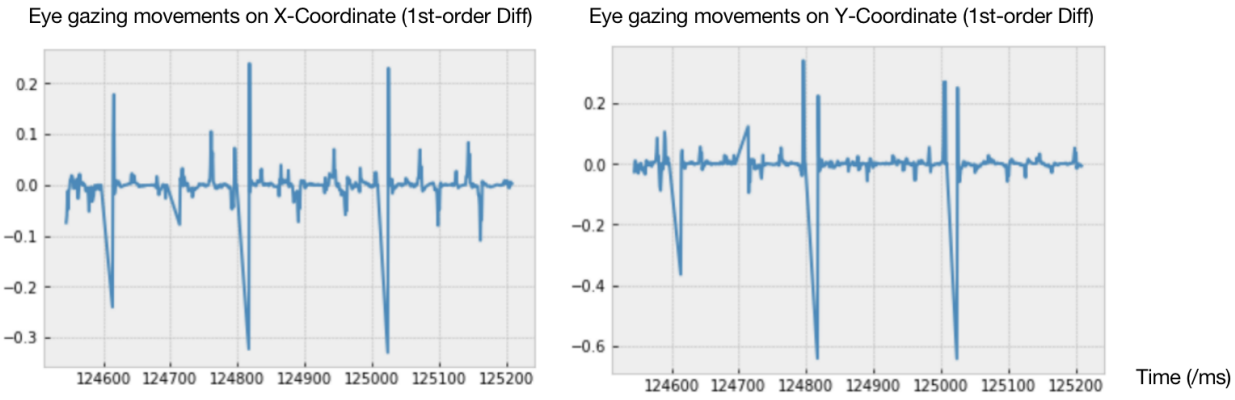



Figure 6: First-order derivatives of eye gazing movements on X- and Y-coordinates

Based on analysis on autocorrelation graph and partial autocorrelation graph for both sequences (the graphs has been omitted due to report length consideration), as well as the ADF test and white noise test, we supposed that the ARIMA(0,1,1) was appropriate for the eye gaze movements prediction on the first graph. We divided both sequences into two parts, the first 80% of the sequence were used as historical data (or training data), and the last 20% were used as testing data.

The ARIMA(0,1,1) model achieved BIC value of -2443.146 and RMSE value of 0.155 for X-coordinates, and BIC value of -1892.588 and RMSE value of 0.120 for Y-coordinates, which

indicates that ARIMA(0,1,1) model the sequences well and the eye gazing movements was actually predictable for our first image.

Next, similar implementations were conducted for each of the images we collected. To obtain more robust results, we also allowed the program to conduct a grid search on the ARIMA(p,1,q) parameters p and q, and the best model was then selected to report the evaluation metrics.

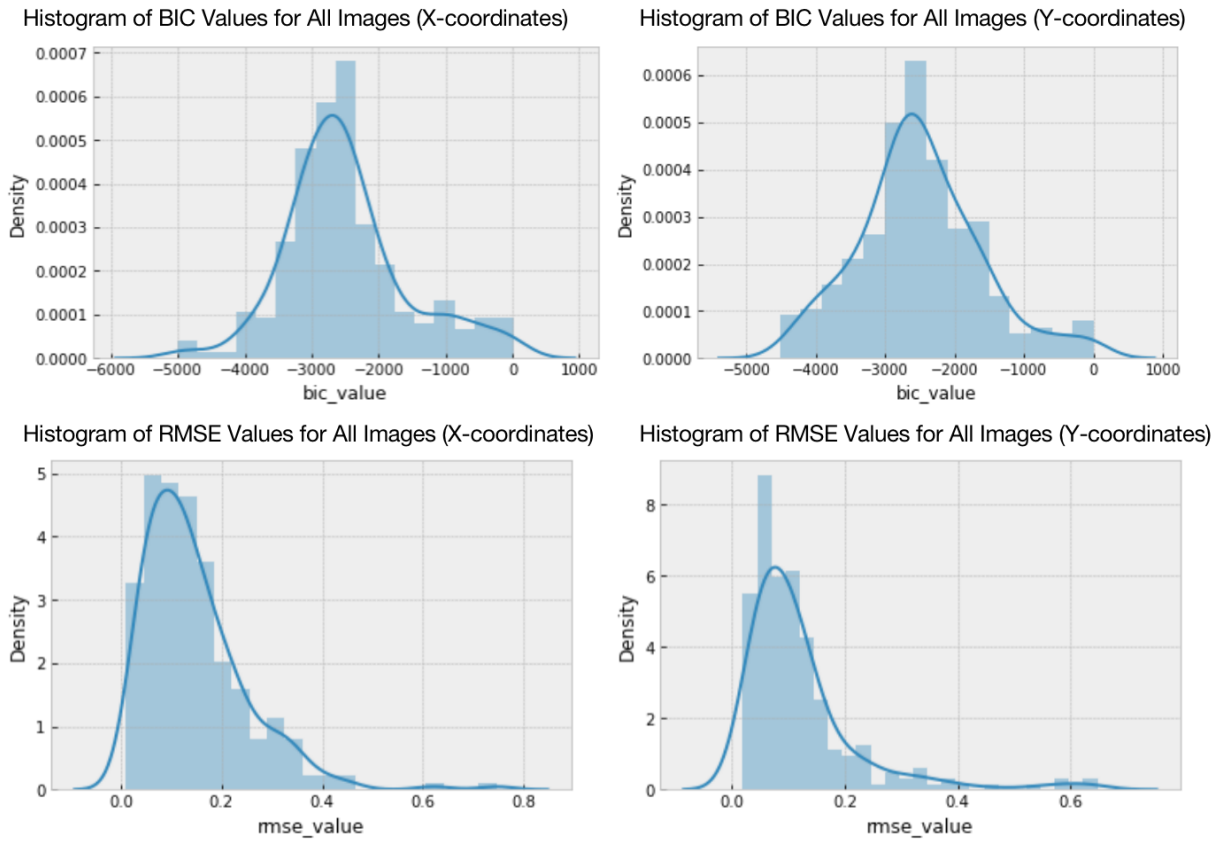



Figure 7: Histogram of BIC and RMSE values on X- and Y-coordinates, respectively

The above results indicate that the eye gazing movements in most of the images are predictable when the individuals are looking through their tables.

## 4. Dimensionality Reduction

To decrease dimensionality of the data, we tried methods including Linear Discriminant Analysis (LDA), Principal Component Analysis(PCA) , and t-SNE. LDA is a supervised technique that also achieves classification of the data simultaneously, whereas PCA is unsupervised and ignores the class label.

### 4.1 PCA RESULT

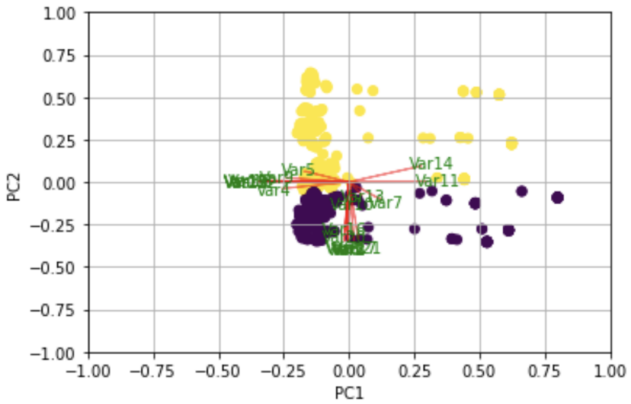



Figure 8: PCA Result Visualization

| | 0 | 1 |
|---|---|---|
| 0 | PC0 | LPOGV |
| 1 | PC1 | FPOGID |
| 2 | PC2 | RPOGX |
| 3 | PC3 | FPOGY |
| 4 | PC4 | LPCX |
| 5 | PC5 | FPOGD |
| 6 | PC6 | FPOGV |
| 7 | PC7 | FPOGY |

Table 1

The first principal component PC1 formed by PCA will account for maximum variation in the data. PC2 does the second best job in capturing maximum variation and so on.

| | |
|---|---|
| LPOGV | The left fixation POG valid flag is 1 for valid and 0 for not valid |
| FPOGID | The fixation POG ID number |
| RPOGX | The X-coordinate of the right eye fixation POG, as a percentage of the screen width |
| RPOGY | The Y-coordinate of the right eye fixation POG, as a percentage of the screen height |
| LPCX | The X-coordinate of the left eye pupil in the camera image, as a percentage of the carama image size |
| FPOGD | the duration of the fixation POG in seconds |
| FPOGV | The FPOG valid flag is 1 for valid and 0 for not valid |
| FPOGY | Y-coordinate of fixation POG, as a percentage of the screen height |

Table 2: Explanation of Important feature terms

From the eye tracking experiments, we discovered that gazing at the center of the table would provide more stable and accurate fixation points.

## 5. Clustering Analysis

In order to extract points or areas of interest, we conduct clustering analysis on the obtained experiment data, in hopes of forming cluster groups of gaze points which can be identified as multiple table entities, or different components and/or the borders of a single table. In order to achieve the latter, we redesign our experiment protocol to have participants staring at each corner of the table on the given page for two seconds. The clustering technique is applied both directly on ordinary experiments and the experiments with the special protocol.

We have tried various popular clustering approaches including DBSCAN, KMeans, and Affinity Propagation. With parameter tunings, the result on our sample data shows that DBSCAN clustering (Figure 5 right) slightly outperforms the others, probably due to the nature of density-based spatial clustering.

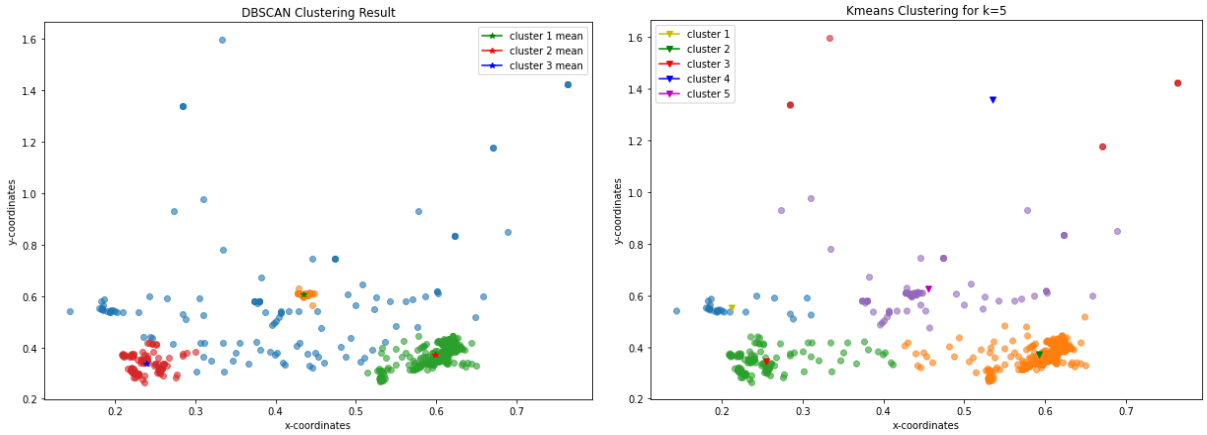



Figure 9: Clustering result for DBSCAN(left) and K-Means(right)

## 6. Table Detection via Tensorflow pretrained model

It is observed that the document pictures in the dataset are mainly black-and-white or single color, and many of them have similar table contents, sizes, and aspect ratios. Therefore, in order to improve the efficiency of training and alleviate over-fitting, we started from annotating the image dataset based on their different sizes, colors, positions and text densities, among which 83 pictures were selected as the final dataset by their diverse forms. Then, we processed the

dataset into a VOC format, and split them into a train set, validation set and test set by the ratio of 0.6 : 0.2 : 0.2.

We choose the YOLO-v5 network to solve the problem as the structure is shown in Figure 9. Based on yolov5x pre-trained model, we finetune the model with the train set for 200 epochs.

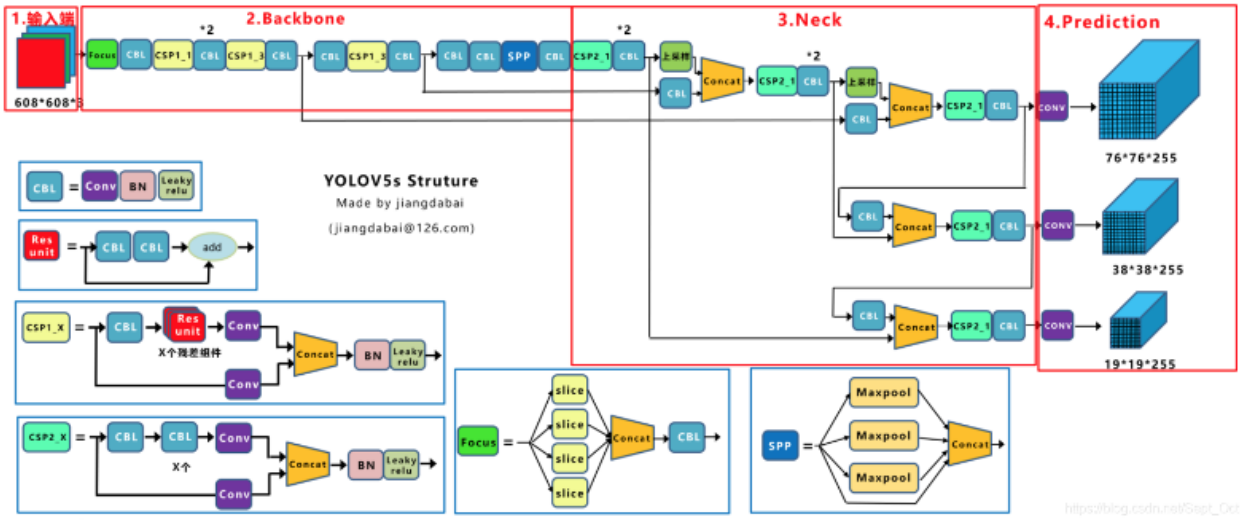



Figure 10: The structure of YOLO.

The MAP of the test set is 0.90. After experimenting with the images, it is observed that the model can bound the table with a high confidence even if the table is very small, thin, or fat. The Figure 10 and Figure 11 shows the bounded table and cropped table content respectively.

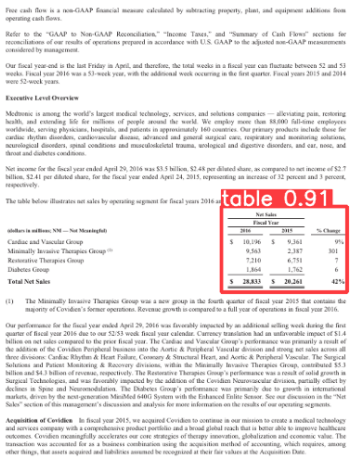



Figure 11: The bounded table

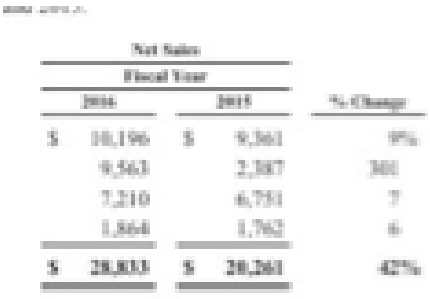


Figure 12: The cropped table

## GOALS AND NEXT STEPS

In the second half of this project, we will concentrate more on the model integration part. Our focus would be on extending the model to address the OCR problem, improving model performance, and integrating AOI detection to the architecture for a systematic solution. After that, we would as well explore possible alternative approaches and make a comprehensive evaluation on model efficiency and accuracy.

## CONTRIBUTION

**Shihang Wang:** Main contributor on public data collection and labeling, experiment design and data collection and times series prediction.

**Yeqi Zhang:** Main contributor on exploratory data analysis, dimensionality reduction, times series prediction and CV table detection algorithm application.

**Yibai Liu:** Team Captain: set up timelines, manage progress and coordinate work; Main contributor on experiment code interface design, experimental data collection and visualization.

**Yijia Jin:** Main contributor on exploratory data analysis, clustering analysis and CV table detection algorithm application.

**Yinqiu Feng:** Main contributor on data preprocessing, feature extraction and dimensionality reduction.

## REFERENCES

Beymer, David & Russell, Daniel. (2005). WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. 1913-1916. 10.1145/1056808.1057055.

Christophe Rigaud, Thanh-Nam Le, J.-C Burie, J.-M Ogier, Shoya Ishimaru, et al.. Semi-automatic Text and Graphics Extraction of Manga Using Eye Tracking Information. 12th IAPR Workshop on Document Analysis Systems (DAS), Apr 2016, Santorini, Greece. pp.120 - 125, ff10.1109/DAS.2016.72ff.ffhal-01336346ff

Farnsworth, B. (2020). *10 Most Used Eye Tracking Metrics and Terms.* Imotions. Retrieved on Feb 4, 2022 from https://imotions.com/blog/10-terms-metrics-eye-tracking/#aoi

Holomb, V. (2021). "Borderless Tables Detection with Deep Learning and Opencv." *Medium*, Towards Data Science, Retrieved on Feb. 15, 2022 from https://towardsdatascience.com/borderless-tables-detection-with-deep-learning-and-opencv-ebf568580fe2.

H. Muñoz, F. Vilariño and D. Karatzas, "Eye-Movements During Information Extraction from Administrative Documents," 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019, pp. 6-9, doi: 10.1109/ICDARW.2019.20045.

S. Karthikeyan, Thuyen Ngo, M. Eckstein and B. S. Manjunath, "Eye tracking assisted extraction of attentionally important objects from videos," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3241-3250, doi: 10.1109/CVPR.2015.7298944.

.