IMAGE RECOGNITION

AI Machine Learning Project

1 CONTENTS

2 Basic understanding of the project 2

3 Tools and Technologies 2

4 About Team 2

5 About the Consultant 3

6 Project Current Stage 4

6.1 Block identification from Image 4

6.2 Text extraction 8

6.3 Product Identification 15

6.4 Table data extraction 15

7 Clarifications 16

IMAGE RECOGNITION SYSTEM

2 BASIC UNDERSTANDING OF THE PROJECT

Image files will be processed and the text and corresponding images will be extracted with its positions (x & y co-ordinates). Manual intervention is needed to confirm and correct if required. The changes will be stored into the database.

3 TOOLS AND TECHNOLOGIES

- Python 3.6
- TensorFlow 2.0
- Keras
- Python-tesseract
- Flask / DJango
- OpenCV 2

4 ABOUT TEAM

AI ML Project Team		
Project Administration		
Project Leader	Suraj P	
AI ML Developer Details		
#	1	2
Developer	Amal Sebastian	Berlin Mary Joseph
Qualification	B.Tech Information Technology	B.Tech Electronics and Communications
Experience	2 years	2 years
Designation	Software Engineer	Software Engineer
Basic Skills	**Java Developer** Core Java, SpringBoot, Swing, JSP,Servlet, HTML, CSS, JavaScript, AngularJS, MySql	**Java Developer** Core Java, SpringBoot, Swing, JSP,Servlet, HTML, CSS, JavaScript, AngularJS, MySql

Specialized Skills	**Artificial Intelligence and Machine Learning** Computer Vision and Image Processing Deep Learning, Neural network, OCR Python, Open CV, Anaconda, Pandas, SkLearn, Tensorflow	**Artificial Intelligence and Machine Learning** Computer Vision and Image Processing Deep Learning, Neural network, OCR Python, Open CV, Anaconda, Pandas, SkLearn, Tensorflow

5 ABOUT THE CONSULTANT

AI Consultant	
Consultant Name	**Shailesh Sivan**
Designation	AI Architect
Qualification	Ph.D in COMPUTER SCIENCE (AI - Machine Learning)
	Master's Degree in MATHEMATICS
Experience	12 years
Technical Skills	
Programming Skills	**Python**, R, C, C++, Visual C++(Win32,MFC), Java, C#, Android , JavaScript,JSP, PHP, Android, HTML, CSS, JQuery, AngularJS, Bootstrap
Frameworks and Technologies	**Open CV, Pandas, Tensorflow**, Tableau, SkLearn, Open GL, Anaconda, Hadoop, OmNet++, NS2, Latex, Doxygen. Microsoft.NET
AI Skills	1. Design of Computational Products & Services. Intelligent Automation. 2. Digital Image Processing and Computer Vision for Image and Video Analysis. 3. Machine Learning, Deep Learning. 4. Probabilistic Programming, Predictive Analysis & Decision Making. 5. Artificial Intelligence / Cognitive Computing. 6. Architect for different academic and industry AI based projects. 7. Solution architect for IoT and BigData projects 8. Design of Computational Models & Architectures. 9. Distributed Computation 10. NLP / Deep NLP & Text Mining
Research & Development Skills	1. Image and Video Processing, Computer Vision. 2. Deep Learning, Machine Learning, Neural Network and Artificial Intelligence 3. Big Data Processing with Hadoop 4. Data mining and Information Retrieval 5. Information Security and Cryptography 6. Computational Linguistics and Natural Language Processing

6 PROJECT CURRENT STAGE

We have taken the image samples from the shared catalog image URL provided and done some work for block identification, getting the position and size, block extraction, table details identification from block, text extraction from image using OCR library etc.

6.1 BLOCK IDENTIFICATION FROM IMAGE

Method:

For identify the blocks in an image, we have used the following process

Step 1: Input Image-> Edge Detection ->Convert to Grayscale -> Convert Binary Image

Step 2: Contour Extraction -> Split Page Image ->

Step 3: Locate Block Position and Size -> Extract and Save Blocks

Example:

The sample page processed from the catalog images is given below

https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/23

Figure: Sample Page 23 Image

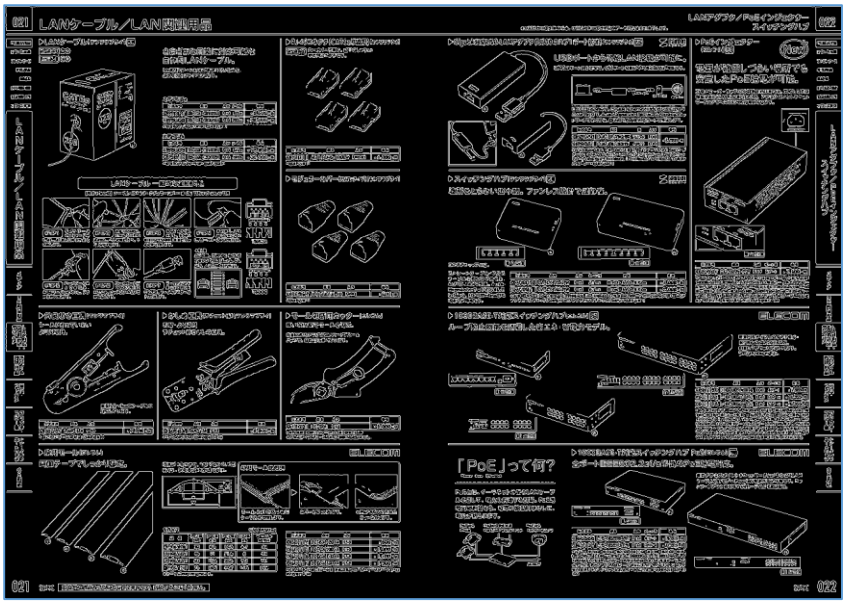


Figure: The image after completing the Step 1:
Input Image-> Edge Detection ->Convert to Grayscale -> Convert Binary Image

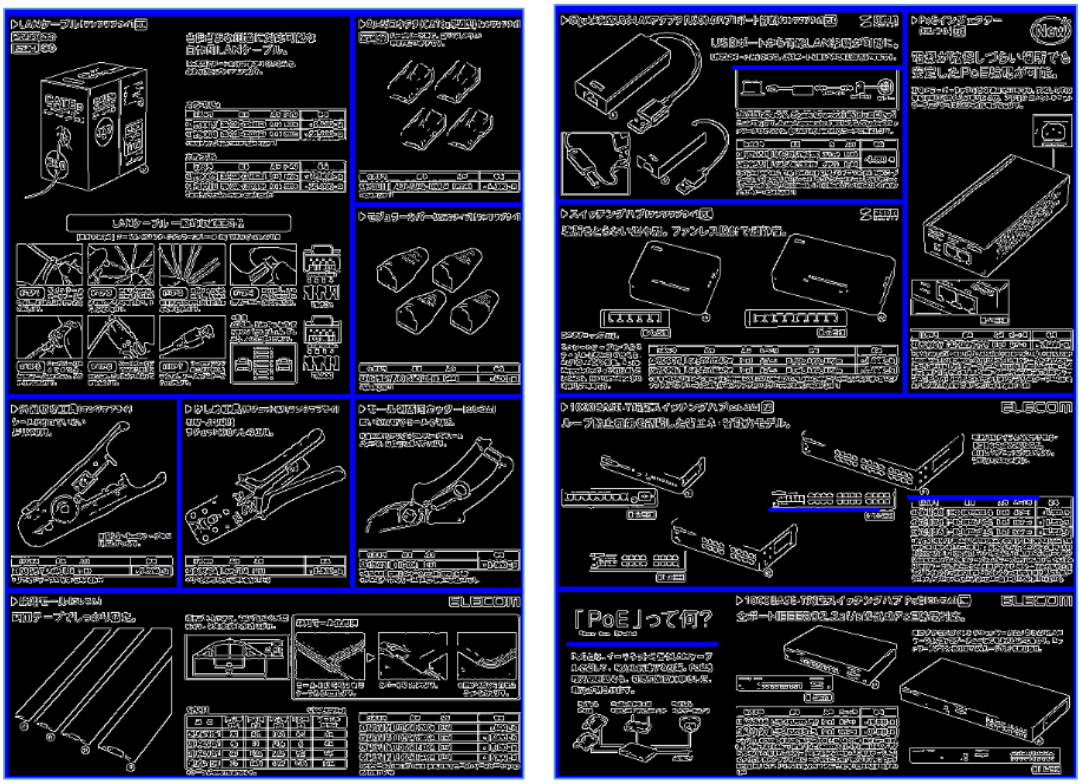


Figure: Left Page Image and Right Page image extracted separately
Step 2: Contour Extraction -> Split Page Image

```
        if w>100 and h>80:
            ROI = dilated_value[y:y+h, x:x+w]
            cv2.imwrite('BLOCK_{}_{}.png'.format(ROI_number,page_name), ROI)
            datab=[[x],[y],[w],[h],['BLOCK_{}_{}.png'.format(ROI_number,page_name)]]
            dataB1.append(datab)
            dilated_value=cv2.rectangle(dilated_value,(x,y),(x+w,y+h),(0,0,255),1)
            ROI_number += 1


    plt.subplot(n),plt.imshow(dilated_value,cmap = 'gray')
    plt.title(pagename), plt.xticks([]), plt.yticks([])
    n=122
    pageno=pageno+1
    ROI_number =ROI_number
    page_name='right'
df1=pd.DataFrame(dataB1,columns=['TopLeft-X','TopLeft-Y','Width','Height','Image'])
print(df1)

cv2.waitKey()
```

```
    TopLeft-X TopLeft-Y  Width Height                Image
0         [3]     [496]  [435]  [155]    [BLOCK_0_left.png]
1       [298]     [332]  [140]  [159]    [BLOCK_1_left.png]
2       [150]     [332]  [143]  [159]    [BLOCK_2_left.png]
3         [3]     [332]  [143]  [159]    [BLOCK_3_left.png]
4       [298]     [169]  [140]  [158]    [BLOCK_4_left.png]
5       [298]       [7]  [140]  [157]    [BLOCK_5_left.png]
6         [3]       [7]  [290]  [320]    [BLOCK_6_left.png]
7         [3]     [496]  [435]  [155]   [BLOCK_7_right.png]
8         [3]     [332]  [435]  [159]   [BLOCK_8_right.png]
9         [3]     [169]  [290]  [158]   [BLOCK_9_right.png]
10      [298]       [7]  [140]  [320]  [BLOCK_10_right.png]
11        [3]       [7]  [290]  [157]  [BLOCK_11_right.png]
```

Block names for LEFT page and TopLeft Position (X,Y) Size - Width & Height

Block names for RIGHT page TopLeft Position (X,Y) Size - Width & Height

Figure: Python code execution output displays the top left position **(X, Y)**, **Width** and **Height** along with the extracted block image file names

Extracted blocks from LEFT page are saved as image files

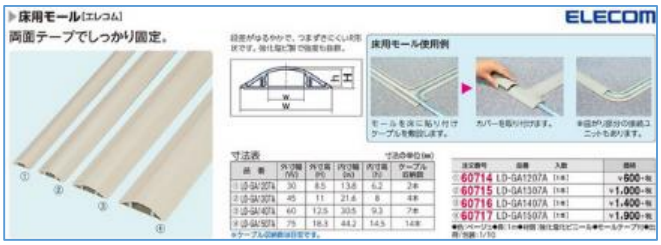


Fig: Block_0_left.png

Fig: Block_3_left.png

Fig: Block_6_left.png

Extracted blocks from RIGHT page are saved as image files

Fig: Block_7_right.png

Fig: Block_8_right.png

Fig: Block_9_right.png

Fig: Block_11_right.png

Current Problem and Solution:

1. Primary challenge in identifying the bounding box representing each product seems it is uneven, i.e. there is no general method to extract the blocks in a page.
2. The current method will not work for all other pages having different bounding block formats.

Another possible solution is to dedicatedly build an intelligent model that automatically identify the bounding box using deep learning – RCNN and object localization.

6.2 TEXT EXTRACTION

Python-tesseract, optical character recognition (OCR) tool for python for recognizing and read the text from images.

Steps involved in reading text from images are

1. Apply denoising on image
2. Convert to Grayscale
3. Dilation and Erosion
4. Enlarge the image size to five times larger image
5. Extract the text from image

Example:

The sample pages processed for text extraction is given below:

https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/110
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/131
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/168
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/172
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/189
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/205
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/209
https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/247

Block from Page and its Extracted text (Errors are highlighted in Yellow)

ラベルプリンタタ専用ソフトマックス]
自由にレイアウトして、お好みのラベルを作成。

楽らくラベル Pro Windows
注文番号。 品番 入数
171834 SLP-10 (11 ¥40,000+%

@対応 OS:VWindows(7 以降)@出荷/包装: 1

Block and its Extracted text (Errors are highlighted in Yellow)

GBC ババウチフィルム [アコ・ブランズ]
大判 A2 に対応。

一一

注文番号 品番 入数 規格
D39517 LP100A2 [OO 枚] A2 判 \24.000+税

@汰寸:綱 604X 横 430mm@フィルム厚:100(0.1 環)@出荷/包装:1/5

Block and its Extracted text(Errors are highlighted in Yellow)

4 穴式 リングファイル [テージー]

紙をしっかりじられる
4 穴タイプ。

4 穴式は 1 つの穴へかかる負担が
少なくなるため、 紙が破れにくく
なります。

A4 判タテ型 (育幅 35mm)
注文番号 品番　　色 入数

| 価格
35038 FR-414 田ブルー 穫]

@條寸:縦 314x 横 265mW@痛幅:35mm@適正収容:120 枚リング内径:
ーーて Tr1 4 穴@穴間隔:80 呆@材質 :表紙=ニ PP,綴じ具 ABS@出荷/

Block and its Extracted text(Errors are highlighted in Yellow)

板目表紙[ワラウン] 謙及台き GPN 提載

書類綴じ・工作など活用法いろいろ。
重物などの保護材にも便利。

※綴りひもは別売です。
※綴じひも用の穴は別途あける必要があります。
注文番号　品番　入数 規格　外二
115560 CR-JH45A4-W [100 初 A4 判 縦 306x 横 215mm　y2,560+税
? 20460 CR-JH45-W [100 枚] 美濃判 縦 394x 横 273mm　y3.650+税

@多厚:0.59mme@坪量: 485g/n?450g/me@材質:古紙 70%以上使用@出荷/包装:1/4

▶ニューホルダー[テージー]

タテ入れ

B5～A2（新聞1ページ）サイズまでラインアップ。

綴じ具とカバーの間に余裕があるため、枚数の多い資料なども収納できます。

中紙あり

（20ポケット）

規格	注文番号	品番	色	入数	外寸	価格
B5判タテ型	①00532	H-30-00	■黒	[1冊]	縦275×横212mm	¥1,400+税
A4判タテ型	②00533	H-40-00	■黒	[1冊]	縦315×横242mm	¥1,600+税
B4判タテ型	③00534	H-50-00	■黒	[1冊]	縦381×横288mm	¥2,600+税
A3判タテ型	④00535	H-60-00	■黒	[1冊]	縦437×横330mm	¥3,400+税
B3判タテ型	⑤00536	H-70-00	■黒	[1冊]	縦540×横417mm	¥5,800+税
A2判タテ型	⑥00537	H-80-00	■黒	[1冊]	縦620×横475mm	¥7,400+税

●ポケット数：20ポケット●背幅：①～③12mm④～⑥15mm●ポケット厚：①～③0.055mm④0.06mm⑤⑥0.065mm●材質：表紙=ビニール,ポケット=PPフィルム●中紙付●出荷/包装：1/5

Block and its Extracted text(Errors are highlighted in Yellow)

ニューホホルダー[テージー]
一生一一 5 ンムの 2 (新聞 1 ページ) 綴じ貞とカバパーの間に余裕があるため、
嗣 サイズまでラインアップ。 枚数の多い資料なども収納できます。

中移あり

(20 ポケット)

" 現格　注文番号。 尽重。 色 入数　外十
B5 判タテ型 や 00532 H-30-00 司黒 1 衝] 縦 275x 模 212mm
A4 判タテ型 00533 日-40-00 革昧 氏] 縦 315x 横 242mm　　ヤ 1i500+科
B4 判タテ型 $00534 H-50-00 騙昧 [(加] 縦 381x 模 288m
A3 判タテ型 で 00535 昌-60-00 玉黒 [1] 終 437 メ模 330 画
B3 刊タテ型 @00536 H-70-00 屋昧 [1 弄】 縦 540x 模 417m
A2 判タテ型 00537 日-80-00 厚時 [1f] 縦 G20x 模 475mn

あポケット麗:20 ポケット机棚:やの一一る 12 画の一一の 15 画ゆポケット厚や一
⑬0.055mn4D.06 王(5@⑤0.065
画借材弓:表抵二ビニール,ポケット一PP フィルム@中抵付$出荷/包装: 1/5

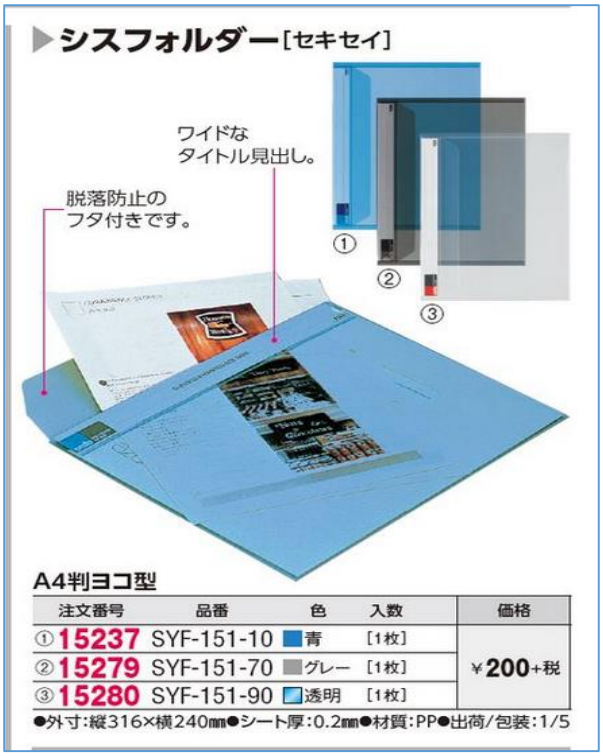
Block and its Extracted text (Errors are highlighted in Yellow)

クリアホルダー [ナカバパヤシ]
経 旨削減にお勧めの 100 枚入り。

A4 判

注文番呈 品番 色 ス数

①07570 CH1036C クリア [00 松] y2.400+税
@外寸:綱 310X 横 220mm@シート了ぽ:0.2xm@材質:PP@出荷/包装:1/6



Block and its Extracted text (Errors are highlighted in Yellow)

シスフォルダー[セキセイ]

ワイドな
タイトル見出し。

脱落防止の
フタ付きです。

4 判ヨコ型

し 昌| ee |
①15237 SYF-151-10 較埋 n 枚]

る 15279 SYF-151-70 回グレー [(払 \200+税

$⑧15280 SYE-151-90 可透明 i 抽]
@寸:拉 316X 損 240 破シート屋:0.2m@桂距:PP@出座/包装:1/5

Current Problem and Solution:

Some text are not identifying correctly. When we try to enlarge the images we took from the screen shot, it is getting blurred. There is an issue regarding the clarity of the image we used to

process. Currently we are taking the screen shot of the catalog image, in which the clarity is getting lost.
We have to get good quality images for getting text extracted correctly.

6.3 PRODUCT IDENTIFICATION

Use CNN for train the dataset and object identification. We need the image of products to train the dataset.

6.4 TABLE DATA EXTRACTION

Initially we have used Edge detection and Contour Extraction for identifying the blocks.
We have tried used CNN model for train sample tables and then predict the tables.
Here we faced the issue of over fitting, i.e. only trained tables are getting identified, since the data varies from table to table.

Current Problem and Solution:

All the tables have only three sides, the fourth side is not closed. Tables are of not uniform structure, and non-white background colored table are not identified in this method.

We are trying other alternate methods to identify the table and by ignoring the inside contents and identify the table structure.

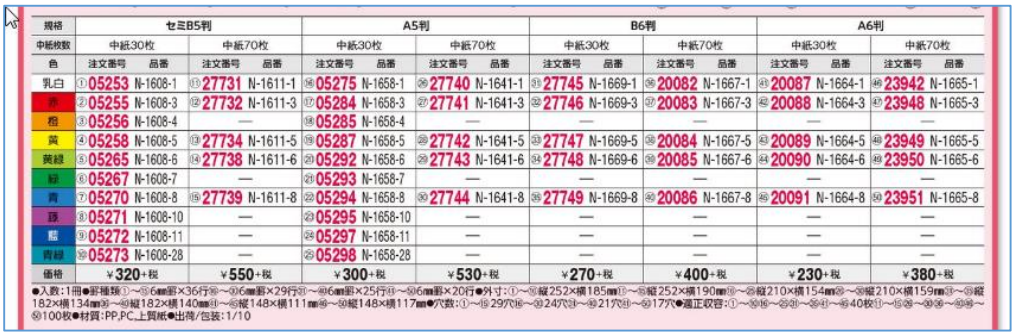
Example:
Table structure (commonly found)

In the above image, only Top, Right and Bottom side is having a boundary line. The LEFT side of the table is open, not closed.

Table with non-white background color (marked in RED color rectangle)

Below table is having a different format than the previous tables (Page369)

https://www.crowngroup.co.jp/us/gra-o/crown/ebook/group_2019/HTML5/pc.html#/page/369



規格	セミB5判				A5判				B6判				A6判			
中紙枚数	中紙30枚		中紙70枚		中紙30枚		中紙70枚		中紙30枚		中紙70枚		中紙30枚		中紙70枚	
色	注文番号	品番	注文番号	品番	注文番号	品番	注文番号	品番	注文番号	品番	注文番号	品番	注文番号	品番	注文番号	品番
乳白	① 05253	N-1608-1	⑬ 27731	N-1611-1	⑯ 05275	N-1658-1	㉖ 27740	N-1641-1	㉛ 27745	N-1669-1	㊱ 20082	N-1667-1	㊶ 20087	N-1664-1	㊻ 23942	N-1665-1
赤	② 05255	N-1608-3	⑫ 27732	N-1611-3	⑰ 05284	N-1658-3	㉗ 27741	N-1641-3	㉜ 27746	N-1669-3	㊲ 20083	N-1667-3	㊷ 20088	N-1664-3	㊼ 23948	N-1665-3
橙	③ 05256	N-1608-4	—		⑱ 05285	N-1658-4	—		—		—		—		—	
黄	④ 05258	N-1608-5	⑭ 27734	N-1611-5	⑲ 05287	N-1658-5	㉘ 27742	N-1641-5	㉝ 27747	N-1669-5	㊳ 20084	N-1667-5	㊸ 20089	N-1664-5	㊽ 23949	N-1665-5
黄緑	⑤ 05265	N-1608-6	⑭ 27738	N-1611-6	⑳ 05292	N-1658-6	㉙ 27743	N-1641-6	㉞ 27748	N-1669-6	㊴ 20085	N-1667-6	㊹ 20090	N-1664-6	㊾ 23950	N-1665-6
緑	⑥ 05267	N-1608-7	—		㉑ 05293	N-1658-7	—		—		—		—		—	
青	⑦ 05270	N-1608-8	⑮ 27739	N-1611-8	㉒ 05294	N-1658-8	㉚ 27744	N-1641-8	㉟ 27749	N-1669-8	㊵ 20086	N-1667-8	㊺ 20091	N-1664-8	㊿ 23951	N-1665-8
藤	⑧ 05271	N-1608-10	—		㉓ 05295	N-1658-10	—		—		—		—		—	
藍	⑨ 05272	N-1608-11	—		㉔ 05297	N-1658-11	—		—		—		—		—	
青緑	⑩ 05273	N-1608-28	—		㉕ 05298	N-1658-28	—		—		—		—		—	
価格	¥320+税		¥550+税		¥300+税		¥530+税		¥270+税		¥400+税		¥230+税		¥380+税	

●入数：1冊●郵種別①～⑤6mm罫×36行⑥～⑩6mm罫×29行⑪～㉕6mm罫×25行㉖～㉙6mm罫×20行●外寸：①～⑤縦252×横185mm⑥～⑩縦252×横190mm⑪～⑮縦210×横154mm⑯～⑳縦210×横159mm㉑～㉕縦182×横134mm㉖～㉚縦182×横140mm㉛～㉟縦148×横111mm㊱～㊵縦148×横117mm●穴数：①～⑨29穴⑩～⑳24穴㉑～㉚21穴㉛～㊵17穴●適正収容：①～⑨10～25枚⑩～⑳4～40枚㉑～㉕26～30枚㉖～㉚100枚●材質：PP,PC,上質紙●出荷/包装：1/10

7 CLARIFICATIONS

Question: Is there any way to provide an option to download the image in its original clarity?

Question: Whether we get product images separately for training the dataset and object identification?