OCR Engine to Extract Food-Items, Prices, Quantity, Units from Receipt Images, Heuristics Rules Based Approach

Rafi Ullah, Ali Sohani, Athaul Rai, Faraz Ali, Richard Messier
Data Science Department, Cubix Labs Pvt Ltd, Pakistan

Abstract— This paper proposes a some heuristics and intelligent rules for improving OCR results, which is old technique. It is a mixture of some old and few novel techniques to nail down the fundamental problem of Food-Items, Prices, quantities and units recognition, extraction of them from the Grocery Receipts and then correction. There is no specialized OCR system, we found during our literature review, all are generic images to text conversions. We have targeted specialized OCR system, which is actually a wrapper around the basic OCR. This specialized OCR Engine is in the context of Grocery related details like items name, price, quantity and units in Receipt. We have wrapped Tesseract (an open source OCR engine by Google). Our system improve the OCR results by considering some heuristics and intelligent rules. The extracted text passes through some filters (these are advance regular expressions for recognizing item name, quantity, price and units). Context aware spell correction is applied for additional accuracy for items names. OCR systems usually produce garbage results. For ignoring that garbage results we have applied some rules for defining garbage text. We have concluded successfully that our OCR system significantly improve the context based OCR text recognition and having closed matched to reality as compared to general purpose OCR systems (unassisted/ vanilla Tesseract OCR engine). This is the enhance version of our work [1]. We have made grocery receipts images to text converter with proven accuracy as compared to basic OCR systems. The bigger picture will empower Food-Kitchen Assistance Mobile Applications in the market. As when Users won't require to enter what's in their pantry, system can help them to tell what arrived when and what would be required in their next shopping visit.

Index Terms— Generic Receipt parser, Generic OCR System, Rule Based OCR engine, OCR Engine for Receipts to Text

1 Introduction

The main objective of our work is to enhance the efficiency of our proposed generic OCR system's for recognizing items names and prices from grocery receipts [1], accuracy of Tesseract OCR [2] and the limitation of same objective using template matching [18]. Methodology presented in [18] works best but fails in case of new images that are not stored in out template engine. Methodology presented in [1] has a lot of false positive results and weak heuristics. In this paper we will be using some rule and advance heuristics for extracting required and valuable text from the grocery receipts. As we have shown some receipts in [1] that have complex and constantly changeable structure, here out algorithm presents many false positive results and garbage text. To get rid off these false positive and garbage text from receipts, we have design some advance rules and heuristics. These rules and heuristics greatly add efficiency to OCR results. This paper presents the extension of our previous work.

The overall system starts with basic image processing techniques like image binarization, image resizing, non textual area removing etc. These operations are performed as preprocessing for tesseract-ocr, which add value to tesseract-ocr results as shown in figures. These steps are must because image by mobile camera is always noisy. The text (OCR result) is stored in text file. Now from text file, text is read line by line and on each of the line we apply rules and heuristics to extract item names, item quantities, item prices and units used in the receipts. Regular expressions are used to extract names, numbers from the text. Text contains (mainly depend upon the image quality; but receipts images usually of low quality) garbage text, that is not human readable or not valuable. This type of text is filtered using rules, such as item name cannot be greater than 40 letters, cannot contains numeric type etc. The complex part was to identify these parameters from the different receipts as all receipts are of different structure and contents. Receipts usually use short names for items such as **GRLC** for **garlic**. This is handle by dictionary of such short names. At last the item names detected may be mis-spelled, theses mis-guided words are corrected by the context aware dictionary I-e dictionary only contains food item names. For matching we used fuzzy search.

Rest of the paper includes Related work, Tesseract OCR open source API, Image Pre-processing techniques, Proposed methodology, Rules and Heuristic, Regular Expressions, Grocery Dictionary spelling correction, Results, conclusion and future work.

2 Related work

[1] describe related work, about item names and prices retrieved from grocery receipts images. Heuristics and pattern matching used there fails on some of the receipts I-e that results a lot of false positives as given in figure below

Fig. 1. False positive, counted as an item

[18] describe same process by using image template matching. This procedure works if images have constant structure, but we have observed many of the grocery stores having different receipts structures at different occasions. It will not work, when receipt image template is not present in template engine.

[5] ABBY cloud SDK provides paid API. This API provide API plugin in different languages like node.js, python, Java, C++ etc and even for android. Receipts images are noisy due to taken by movable mobile devices. are not always clear. So simple scanning may not give you an accurate results. ABBY SDK uses similar kind of image pre-processing for improving OCR accuracy.

[6] is an R&D about similar purpose. This R&D is basically for receipt parsing. They also took similar steps like image binarization, text finding etc.

OCRDroid framework has been proposed in [8]. This use image processing techniques like deskewing, binarization etc for better results. There is limitation of multiple images OCRing and Text detection from complex backgrounds.

3 TESSERECT-OCR

Tesseract is an open source Optical Character Recognition (OCR) Engine or API, available under the Apache 2.0 license. It can be used directly use or using an API to extract typed text , handwritten text or printed text from images of different formats. It supports a wide variety of languages (we are using python) and almost for all operating systems (we are using Ubuntu 16.01) [2].

For configuring pytesseract in Ubuntu use following command

sudo pip install pytesseract

sudo get-apt install tesseract-ocr

After configuring it, you can select language, configuration according to your need.

We are using 'eng' English as a language, "- psm 6" as a config parameter and Image object as a parameter.

4 IMAGE PREPROCESSING

Tesseract OCR is open source library sponsored by Google, There is accuracy issue. It is generic image to text converter. Basic Image processing steps are same as in [1] and many others papers.

4.1 Image Background Removal

Mobile camera images may have noisy backgrounds. Such images can be scanned using tesseract-OCR, but scanning will take time and may results garbage text. To make this process faster and accurate, we remove background from images as shown in fig 2. We used canny edge detection here.

Fig. 2. Walmart receipt before and after background removal

4.2 Image Binarization

Otsu's Image binarization is used to binarize the image for more accurate results. This image processing operation play very important role in this context based system, because user might play with mobile camera images, which contains shades and noise. We have many options but this is process of converting colored image to black and white image [5]. Dirty, shaded and noisy images are cleaned using this process.

Fig. 3. Image before and after Otsu's binarization

TABLE 2
RESULT BEFORE AND AFTER IMAGE BINARIZATION

gunman 401-: '3	1111311053 JOE'S
'Mmmmmm	I 44 East Ontario Street
Chicago IL 60611	Chicago IL 50511
Store #696 — (312) 951-6369	Store #596 ' (312) 951-5359
OPEN 8:00AM TO 10:00PM DAILY	OPEN 8:00AM 10 10:00PM DAILY
OLIVE OIL POTATO CHIPS.. 1.99	OLIVE OIL POTATO CHIPS.. 1.99
HUMMUS GARLIC ROASTED EC 1.99	HUMMUS GARLIC ROASTED EC
OHEDDAR NEH ZEALAND SHARP 3.71	CHEDDAR NEW ZEALAND SH 3.71
PITA NHOLE NHEAT 5" 1.69	PITA WHOLE WHEAT 5" 1.89
OLIVES MANZANILLA 2.29	OLIVES MANZANILLA 2.29
CREAMY SALTED PEANUT BUTTER 2.49	CREAMY SALTED PEANUT BUT 2.49
SUBTOTAL $14.16	SUBTOTAL $14.15
STATE TAX 1 $0.32	STATE TAX 1 $0.32
ITAL m1m	TOTAL $14.48
ITEMS 6 v, Karl	ITEMS 6 v, Kar1
05-31-2015 03:11PM 0696 06 1173 0559	05'31-2015 03:11PM 0695 06 1173 05
THANK YOU FOR SHOPPING AT	THANK YOU FOR SHOPPING A1
TRADER JOE'S	TRADER JOE'S
www.t,raIJgr'Oe§Im,,,	www.trader'oes.com

4.3 Image De-skewing

Images from mobile camera might be deskew, in that case tesseract-OCR performed poor or even not able to detect text. We apply image deskewing technique [3]. Comparison has been shown in table given below between skewed image and deskewed image.

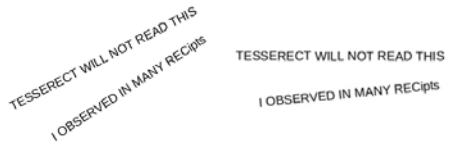

Fig. 4. Image before and after text deskewing

TABLE 2
RESULT BEFORE AND AFTER IMAGE TEXT DESKEWING

<ee~e~eeec< «£va View "As 6 \0%9<N$0 WM" «we	TESSERECT WILL NOT READ THIS ECipIS IOBSERVED IN MANY R

4.4 Image Resizing

There is big issue of processing time and accuracy trade off. Large images (high resolution) will be having accurate OCR results but will talk large time and vice versa. As OCR is game of playing on pixels. We resize images less than height 600 pixels and change the resolution DPI to 300. (300 values has been optimal value in many scenarios). For large images we reduce size to 1/3rd ratios and increase DPI to 300 if less than 300 dpi. This help us to reduce OCR processing time.

5 PROPOSED METHODOLOGY

Algorithm of the proposed system is given below. We have applied some heuristic rules in our algorithm which will be describe in detail after.

5.1 Algorithm

Static:
Symbols = [{, }, [,], /, //, , , ', ", "", ?, &, etc…]
short_names : dictionary contains short form as key and full form as value, I-e "gm" for "gram".
constant_words: dictionary of restaurant names, store names, websites, location, country name etc.
Garbage: Predefined rules for garbage detection
contextAwareDictionary: dictionary contains only food related names I-e food names
Results: [] contain set of item name, quantity, price

1. Image pre-processing
……..1.1. Image cropping (background removal)
……...1.2. Image binarization
……...1.3. Image deskewing
……...1.4. Image Resizing
2. Apply tesseract-OCR on processed image
3. Store OCRed result in text file
4. For every line in text file
…….4.1. Remove symbols from line
…….4.2. For every word in line
…….4.2.1. If word exist in short names
……..4.2.1.1. Replace word by their full form
……..4.3. Extract quantity and unit from line using quantity
……..reg- ex (Regular Expressions)
……..4.3.1. if quantity found
…….4.3.1.2 save it and return remaining line
……..4.4.2. else also check next line
…….4.4.2.1. if quantity found
…….4.4.2.1.1. save previous line as item
……name and next line as quantity
…….4.4.2.1.2. Read Next line from file
…….4.4.2.2. else go to step 4.3.
……...4.4. Extract prices from line using price reg-ex
……..4.4.1. if price found in the line
……4.4.1.1. Save this and go to step 4.4.
……..4.4.2. Else go to step 4.4
……...4.5. Extract word from remaining line
……...4.6. if length of word is less than 3 OR greater than 40
……...words OR word exists in constant_words OR word
……is Garbage then discard every result and go to
…….step 4.
……4.7. else store word as an item name
5. for item in Results
….5.1 find MatchScore between item and words in contextAware
…..Dictionary

…...5.2. if MatchScore < 85%
……...5.2.1. continue
……5.3. else Replace item name by that word

5.2 Short name to full form conversion

It has been observed during our research that most of the grocery reciepts used short names instead of full names. For example reciept will be using **"milk pdr"** instead of **"milk powder".** For mapping we have used a dictionary containing short names as keys and their full form. Every line from OCR result is passed through this filter and replace each short name by their full form found in the dictionary.

Input line: "mlk pdr 12 gm"
Output: "milk powder 12 gram"
Input line: "16 oz onions"
Output line: "16 ounce onions"

5.3 Heuristic

5.3.1 Heuristics 1

We have used some heuristics for OCR results accuracy. In most of the receipt it has been observed that item name, price and quantity is on single line as given in Fig. 3. So we just parsed single line from text file (OCR result), extract price, quantity and item name from that line and ignore rest of the line.

5.3.2 Heuristics 2

Some of the receipts has been observed that they have item name on one line and the item quantity or price on the next line as shown in Fig. 8. This problem cannot be solved by Heuristics 1. If price or quantity is not found in current line, then we read next line using, there may exist price or quantity or both.

5.3.3 Heuristics 3

If numeric type is detected in string/line and the next word is some unit (We have store units) then it is treated as quantity. This can also be detected using regular expression for detecting quantity.

5.3.4 Heuristics 4

we have currencies symbols, if any numeric type is detected, and there is currency symbol before or after the numeric type, then it will be treated as price. This can also be handled by regular expression.

5.4 Rules for garbage text detection

Here are Rules we have define for detecting whether text detected by OCR is garbage or not?

5.4.1 Rule 1

Text detected by OCR will be garbage if length of word is less than 3 letters. For example "F" or "D" or "SC" etc are not valuable for us.

5.4.2 Rule 2

Text detected by OCR will be garbage if length of word is greater than 40 or 50. Words detected such "xcasd.xxxxx.xxxx...xxxx3123.x"
"www.abcassdd.com/newrestatuants" will be simply discarded.

5.4.3 Rule 3

If there 4 consective same letters or digits in a word it will garbage. Raaaafi etc.

5.4.4 Rule 4

Alpha numeric string having length greater than 5 digits will be considered as garbage. For example 4567889966 or 030222343424

5.4.5 Rule 5

If letters to numbers ratio in string /word is greater than 50%, word will treated as a garbage.

5.5 *Regular Expression*

Item name must be string or word detected by following regular expression.

Item name = [a-zA-Z]+

But in grocery receipts may contain number in item names, but we are ignoring them.

Price: Item price can be single decimal number having units like ($, PKR, RS, rupees, etc) or no units.

(?:\$|RS|Rs|rs|Rupees|rupees|pkr)
(\s*?)(\d+(?:\.\d{2}))
or
(\d+(?:\.\d{2})?)(\s*?)((?:rs|RS| pkr|\$|Rupees|rupees))

Quantity: This can be number with units like pound, kg, gram etc, we use Regular Expression

(\d+(?:\.\d+)?)(\s*?)((?:grams| dozens|gm|kg|kilogram|kilo gram| packs| kgram|packets|pair| ounce|spoon| piece))

5.6 Context Aware spell correction

After all the steps defined in algorithm, the last part of our work is to do context aware spell correction. We compare each item name detected by OCR with words in dictionary. If match score of item name with dictionary word is less than 85%, we ignore word of dictionary and considered item name is correct and when the match score is greater than 85%, then item name is replace by dictionary word. The important point here is that we are matching item with a dictionary where all the words are related to grocery. This enhance Chance of correcting item name.

For example word detected by OCR is "**egy**". This word can be replaced by **"ego"** and **"egg"** because both are equal candidates for this. But our system will replace this by **"egg"**, because there is no word like **"egy"** in grocery dictionary.

6 RESULTS

We have tested our proposed algorithm on various receipts from well structured to ill structured receipts images, from clean receipts to mobile camera shaded images, results have been carefully observed and was observed very accurate.

Fig. 5. Wallmart receipt (low quality shaded image)

TABLE 3
OCR RESULT AFTER IMAGE PROCESSING STEPS

m
HIM
mama»:
JIM UILBURN. STORE MHNRGER
UE SELL FOR LESS
HHNHGER UILLIE CHEEKS
(757) 430 - 1836
VIRGINIH BEACH, VA. 23456
ST0 3216 OPfl 00009048 TEG 48 TRfl 08210
PEPSI 001200000129 F 1.08 R
KLG P-TRRTS 003800031120 F 2.18 V
COKE 004900002341 F 2.98 R
COKE 004900002341 F 2.98 R
PEPSI 24 PK 001200000017KF 6.58 R
ARBOR HIST 008210017923 3.37 T
HRBOR HIST 008210017902 3.37 T
HRBOR HIST 008210017901 3.37 T
DISCOUNT GIVEN 1.36
SUBTOTAL 24.55
THX 1 2.500 X 0.25
THX 3 2.500 X 0.61
TOTRL 25.41
DEBIT TEND 25.41
DEBIT CHSH BECK 20.00
TOTRL DEBIT PURCHRSE 45.41
CHRNGE DUE 20.00

TABLE 4
RESULT OF WALLMART RECEIPT IN FIG. 5

Item name	Price	Quantity (unit)
Pepsi	1.08	1 unit
klg p-trrts	2.18	1 unit
coke	2.98	1 unit
coke	2.98	1 unit
pepsi	6.58	24 pack
arbor mist	3.37	1 unit
arbor mist	3.37	1 unit
arbor mist	3.37	1 unit

If price is not detected, by default value will set to "-1" and if quantity is not detected default value is "1 unit".

Fig. 6. Meijer receipt

TABLE 5
RESULT OF MEIJER RECEIPT IN FIG. 6

Item name	Price	Quantity (unit)
frozen onions	-1	1 unit
meijer ketchup	1.59	1 unit
ice cream	5.99	1 unit
tortilla chips	-1	1 unit
cocacola	-1	1 unit
cocacola	-1	1 unit

TABLE 6
RESULT OF TRADER JOE'S RECEIPT IN FIG. 3

Item name	price	quantity
olive oil potato chips	1.99	1 unit
humus garlic roasted economy	1.99	1 unit
cheddar sharp	3.71	1 unit
pita whole wheat	1.69	1 unit
olives hanzanilla	2.29	1 unit
creamy salted pea-nut butter	2.49	1 unit

Fig. 7. Grocery list Receipt (units and price)

TABLE 7
RESULT OF GROCERY LIST RECEIPT IN FIG. 7

Item name	Price	Quantity (Units)
butter	1.99 – $	1 tub
eggs medium	0.76 – $	1 dozen
milk	3 – $	1 gallon
bread sand-wich white	1 – $	1 loaf
coffee	3.50 – $	1 unit
col	0.34 – $	12 pack
cam	0.60 – $	2 can
green beans	0.53 – $	14 ounce
peas dan	-1	1 unit
ground ground	8.95 – $	2.5pound
pork chops	3.99 – $	1 pound
whole dryer	2.07 – $	8 pound
cat food big	3.38 – $	1 unit
cucumber	13.50 – $	1 Ea
lettuce iceberg hezd	0.37 – $	1 Head
tomatoes slic-ing	2.34 – $	1 pound

It work perfect of all types of receipts, here is example of straight forward simple receipt where prices (numbers) are not properly detected.

PRICING COMPARISON	
Product	Amazon
Organic Whole Milk, 1 gallon	$5.99
Half & Half, 1 pint	$2.29
Golden Delicious Apples	$1.83
Tillamook Low Fat Yogurt	$0.79
Dill Pickles, 24 oz.	$3.19
Free Range Chicken,	$7.99
Cage Free Large Brown Eggs	$4.09
Tropicana OJ, Original, No Pulp, 59 oz	$4.83
Tree Top Apple Juice, 46 oz	$3.19
Franz San Juan Island Bread	$4.99
Pirates Booty White Cheddar Popcorn	$3.19
Kellog's Raisin Bran 25 oz	$2.99
TOTAL	$45.36

Fig. 7. Grocery list Receipt (units and price)

TABLE 8
RESULT OF RECEIPT IN FIG 8.

Item name	price	Quantity (unit)
organic whole milk	-1	1 gallon
half half	-1	1 pint
golden delicious apples	-1	1 unit
tillamook low fat yogurt	-1	1 unit
dill pickles	-1	24 ounce
free orange chicken	-1	1 unit
cage free large brown eggs	-1	1 unit
tropicana original pulp	-1	59 ounce
tree top apple juice	-1	46 ounce
franz san juan island bread	-1	1 unit
pirates booty white cheddar popcorn	-1	1 unit
kellogs raisin bran	-1	25 ounce

6 CONCLUSION

We have used state of the art tesseract-ocr open source by Google with novel image processing technique and heuristics rules to get better results of parsing grocery receipt images. Image background is removed, resized, text deskewing and binarization is applied then, Then forwarding to tesseract-ocr. After text extracted from image. Short form words are converted to full form using "short names dictionary". Unwanted text is simply discard using "constant words dictionary". Garbage text removed using heuristics rules (described above). Items names, quantity, units and prices are extracted using regular expressions. And at-last item names are corrected using grocery dictionary by applying fuzzy search.

6 FUTURE WORK

Although this methodology works surprisingly very efficient in most of the cases, but there is still issue of accuracy in some cases like when image is too much dirty and ill structure but this is problem of tessrect-OCR. Numbers detection is still poor, because we have no way of correcting numbers. Future work is to work on more improvement of OCR accuracy through image processing and heuristics and improving number detection. And we will work on retrieving only food items from the receipt. NLP techniques such as Part of speech tagging, entity detection etc. can also be used for the purpose of extracting content of our interest.

ACKNOWLEDGMENT

I would to thanks Cubix Inc Pakistan for providing us a Re-

search and Development platform. I am thankful to all my colleagues at Cubix Inc and especially our mentor Mr ALI SOHANI for his help and guidance.

REFERENCES

[1] Rafi, Ali, Faraz, Athaul, *"OCR Engine to extract Food-items and Prices from Receipts Images via Pattern matching and heuristics approach"*, SMIU, 1st International Conference on computing and related technologies, 2017
[2] https://github.com/tesseract-ocr last visited 6-Oct-2017
[3] http://scikit-image.org/ last visited 6-Oct-2017
[4] http://pyimagesearch.com last visited 9-Oct-2017
[5] Chaki, Nabendu, Soharab Hossain Shaikh, and Khalid Saeed. "A comprehensive survey on image binarization techniques." In *Exploring Image Binarization Techniques*, pp. 5-15. Springer India, 2014.
[6] https://ocrsdk.com/documentation/quick-start/receipt-recognition/
[7] http://rnd.azoft.com/applying-ocr-technology-receipt-recognition/
[8] Church, Kenneth Ward, and Patrick Hanks. "Word association norms, mutual information, and lexicography." *Computational linguistics* 16, no. 1 (1990): 22-29.
[9] Zhang, Mi, Anand Joshi, Ritesh Kadmawala, Karthik Dantu, Sameera Poduri, and Gaurav S. Sukhatme. "OCRdroid: A Framework to Digitize Text Using Mobile Phones." In *MobiCASE*, pp. 273-292. 2009.
[10] GOCR - A Free Optical Character Recognition Program. http://jocr.sourceforge.net/.
[11] OCR resources (OCRopus). http://sites.google.com/site/ocropus/ocr-resources.
[12] OCRAD - The GNU OCR. http://www.gnu.org/software/ocrad/.
[13] OCRdroid - website. http://www-scf.usc.edu/ ananddjo/ocrdroid/index.php.
[14] Simple OCR - Optical Character Recognition. http://www.simpleocr.com/.
[15] Tesseract OCR Engine. http://code.google.com/p/tesseract-ocr/.
[16] http://opencv-python-tutorials last visited 10-Oct-2017
[17] All images from http://google.com
[18] Rafi, Ali, Faraz, Athaul, *"Optical Character Recognition Engine to extract Food-items and Prices from Grocery Receipt Images via Templating and Dictionary-Traversal Technique"*, International conference of computing, 2018 (Accepted)
[19] Modi, Hiral, and M. C. Parikh. "A review on optical character recognition techniques." *Int J Comput Appl* 160, no. 6 (2017): 20-24.
[20] Oudah, Nabeel, Maher Faik Esmaile, and Estabraq Abdulredaa. "Optical Character Recognition Using Active Contour Segmentation." *Journal of Engineering* 24, no. 1 (2018): 146-158.
[21] Zhang, Mi, Anand Joshi, Ritesh Kadmawala, Karthik Dantu, Sameera Poduri, and Gaurav S. Sukhatme. "OCRdroid: A Framework to Digitize Text Using Mobile Phones." In *MobiCASE*, pp. 273-292. 2009.
[22] Kumar, Asit, and Sumit Gupta. "Detection and recognition of text from image using contrast and edge enhanced mser segmentation and ocr." *IJOSCIENCE (INTERNATIONAL JOURNAL ONLINE OF SCIENCE) Impact Factor* 3, no. 3 (2017): 3.
[23] Farahmand, Atena, Hossein Sarrafzadeh, and Jamshid Shanbehzadeh. "Noise removal and binarization of scanned document images using clustering of features." (2017).
[24] Wang, Fu-Bin, Paul Tu, Chen Wu, Lei Chen, and Ding Feng. "Multi-image mosaic with SIFT and vision measurement for microscale structures processed by femtosecond laser." *Optics and Lasers in Engineering* 100 (2018): 124-130.
[25] Zhang, Jing, Guangxue Chen, and Zhaoyang Jia. "An image stitching algorithm based on histogram matching and SIFT algorithm." *International Journal of Pattern Recognition and Artificial Intelligence* 31, no. 04 (2017): 1754006.
[26] Troller, Milan. "Practical OCR system based on state of art neural networks." (2017).
[27] Stadermann, Jan, Denis Jager, and Uri Zernik. "Hierarchical Information Extraction Using Document Segmentation and Optical Character Recognition Correction." U.S. Patent Application 15/620,733, filed September 28, 2017.
[28] Oudah, Nabeel, Maher Faik Esmaile, and Estabraq Abdulredaa. "Optical Character Recognition Using Active Contour Segmentation." *Journal of Engineering* 24, no. 1 (2018): 146-158.
[29] ZHAO, Yan, Yue CHEN, and Shi-gang WANG. "Corrected fast SIFT image stitching method by combining projection error." *Optics and Precision Engineering* 6 (2017): 029.
[30] Sharma, Manoj, Anupama Ray, Santanu Chaudhury, and Brejesh Lall. "A Noise-Resilient Super-Resolution framework to boost OCR performance." In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 466-471. IEEE, 2017.
[31] Brisinello, Matteo, Ratko Grbić, Matija Pul, and Tihomir Anđelić. "Improving Optical Character Recognition Performance for Low Quality Images." In *59th International Symposium ELMAR-2017*. 2017.
[32] Patel, Amit, Burra Sukumar, and Chakravarthy Bhagvati. "SVM with Inverse Fringe as Feature for Improving Accuracy of Telugu OCR Systems." In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pp. 253-263. Springer, Singapore, 2018.

Note: Authors belongs to Cubix Labs, Pakistan
Organization: Cubix Labs, Pakistan
www.cubix.co

Rafi Ullah, Senior Data Scientist at Cubix Labs and Visiting faculty at Pakistan Air Force, Karachi Institute of Economics and Technology. Studying in MS Computer Science specialization in Machine Learning. Working on computer Vision and Artificial Intelligence projects. Graduated from Hamdrad University Karachi, Pakistan. He has 1 publication on video encoding using machine learning, survey of Body Area Network protocols, Book Chapter on Big data Analysis in IoT.

Ali Sohani , Chief Data Scientist and Chief Technical Officer at Cubix Pakistan Ltd. He has more than 16 years of experience in Software Industry working in Design, Research, Development and Management as a founder, co-founder and consultant. He worked in National Technology Group, System Ltd (Visionet Systems Inc), Glotech Inc etc. Developed/ managed and delivered projects for several Fortune 100/ 500, Global 500 and elite-profile organizations. He has publications on recommendation system and OCR Accuracy enhancement and Information Retrieval system's.

Athaul Rai , Junior Data Scientist at Cubix Pakistan. He is Studying MS Computer Science. He has interest in Machine Learning and Natural Language Processing, text mining and computer vision. He is working on Artificial Intelligent system's. He has two papers on specialized OCR system.

Faraz Ali Seelro , Junior Data Scientist at Cubix Pakistan. He is Studying MS Computer Science. He has interest in Machine Learning and Natural Language Processing. He has papers in his field on special purpose OCR and Signature Matching using ANN.

Richard Messier , He is VP United State Region. He has Research Interest in Recommendations system's. He also served as Project manager and Advisor in many Recommendation systems.