International Journal of Engineering & Technology, 8 (1.9) (2019) 199-204

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Review on Text Extraction in Complex Image using Five Different Web OCR

Sari Dewi Budiwati[1], Dahliar Ananda[2], Siska Komala Sari[3]

[123] *Information System Study Program, Telkom Applied Science School, Telkom University*
**Corresponding author E-mail: saridewi@tass.telkomuniversity.ac.id*

Abstract

Price product comparison is needed for the consumer in order to find the cheapest price. In this research, we build the system to compare the price between several modern markets. The input is product catalogs since consumer often receives it from the modern market. We upload 75 grayscale 75 color input images into five web OCR. We compare the results based on characteristic and segmentation parameter. We define 0.5 and 1 point if web OCR recognizes the product price or/and names from product catalogs. Characteristic parameter is a parameter which identified product price and name using box line or empty line. Meanwhile, for segmentation parameter, we use image properties such as image dimension, dots per inch (dpi), and bit depth. From both parameters, OCR4 gives a better result as it is can recognize 50.67% grayscale image and 62.67% color images. Although it is recognized by OCR4, some of its results shown with data noise. In order to remove the noise, we proposed to used Sparse Binary Polynomial Hashing (SBPH) algorithm with 5-8 letters combination. As the result, some of the text was able to recognize in order to compare its price, while the others with much data noise were not.

Keywords: *Online Character Recognition (OCR), web OCR, text extraction, modern market catalogs, SBPH algorithm*

1. Introduction

Information extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources (Sarawagi, 2007). The example of unstructured sources are document images which contain text with its characteristic: grayscale or color images, compressed or uncompressed image, and the text may move or may not (Giri, 2013). Product catalogs from the modern market (supermarket or hypermarket) were classified as document images which contain text, graphic and acquired by scanning (Sumanthi, Santhanam, & Gayathri Devi, 2012). Product catalogs provide crucial information about product name and price (Figure 1). It can be used by the consumer to find the cheapest price or discount. The modern market often promotes their goods by offline product catalogs and online product catalogs. They put offline product catalogs in front of the entrance store, while online products catalogs were put on the websites. In previous research, online product catalogs have poor result in order to recognize the text since most of its images have lower bit depth and dimension (Budiwati, Ananda, & Komala Sari, 2015).

Currently, there are some web OCR which is freely available to recognize the text in images by its free service (API, engine or web service). In this research, we are using five web OCR to found which one has able to recognize the text from product catalogs. Most of these web OCR can recognize JPG, GIF, TIFF, BMP, PDF, complex scanned documents, ancient scripts, rare languages, and non-standard fonts. The five web OCR are CustomOCR, Free OCR, i2OCR, Online OCR and Google Drive.

CustomOCR and Free OCR are using tesseract engine. Online OCR has services which can be used freely, implemented using SOAP and REST interfaces and can be accessed through API by HTTP or HTTPS requests. Google Drive is a cloud service which provided by Google which one of its features is to recognize text in the image. The discussion and result of the recognition will be placed in section Result and Discussion.

The result of text recognition from 5 webs OCR has diverse noise (Figure 6)). In order to clean it, we proposed using Sparse Binary Polynomial Hashing (SBPH) algorithm. SBPH is a way to create a lot of distinctive features from the incoming text (Yerazunis, 2003). SBPH often used in email spamming recognition (Yerazunis W. S., 2004), (Thomason, 2007). We used 8 letters combination to find the text which contains the words "Rp" and product name. The discussion and the result of using SBPH algorithm will be explained in section Result and Discussion. The clear data of text recognition will be stored in the database. From here we compare the price with a query in the web user interface. The discussion and the result of the query will be placed in Result and Discussion section. In the last section, we describe future research from our research.

2. Research Method

We modify the methodology from (Aggarwal & Ghosal, 2011) into several phases:

Copyright © 2019 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Scanned and Segmentation Products Catalogs

In this phase, we scanned and segmented the offline product catalogs as the input images. As the result, we use 75 grayscale images and 75 color images (Figure 4). The input images are limited to 9 (nine) staple according to Indonesia Ministry of Trade. The staples are 1) Rice, flour, and corn; 2) Sugar; 3) Vegetables & fruits; 4) Chicken and meat; 5) Cooking oil and margarine; 6) Milk; 7) Eggs; 8) Gas; 9) Salt.

As we collect the input images, we identify that images product price and product name in input images have two characteristics: 1) bounded with the line (Figure 2); 2) empty line (Figure 3).



Fig 1: Supermarket catalogs



Fig 2: Box line



Fig 3: Empty Line



Fig 4: Segmentation in grayscale and color image

2. Images upload

The input images then upload to 5 web OCR. The five web OCR are:

a) CustomOCR, we identified as OCR1 (CustomOCR, 2015)

b) Free OCR, we identified as OCR2 (FreeOCR, 2015)

c) i2OCR, we identified as OCR3 (i2OCR, 2015)

d) Online OCR, we identified as OCR4 (OCRwebservice, 2016)

e) Google Drive, we identified as OCR5 (Google Drive, 2015)

OCR1 and OCR2 are using tesseract engine. Tesseract is an open-source OCR engine that began as a Ph.D. research project and developed at HP between 1984 and 1994 (Smith, 2007). Now, it is developed and maintained by Google (Google, 2016). The tesseract architecture consists of 4 phase (Patel, Patel, & Patel, 2012):

a) Adaptive thresholding, which converts the image into binary images;

b) Connected component analysis, which is used to extract character outlines;

c) Blob for finding lines and regions;

d) Recognizing text through each word and it is passed to an adaptive classifier.

3. Noise cleaning

As the result from web OCR, the input images often has noise in the text (Figure 6). We identified the noise consist of 1) Long unidentified characters; 2) Error; 3) Empty result and 4) Mixed text with some unidentified characters. We processed the text with categorized in the mixed text with some unidentified characters using SBPH algorithm, we ignore the result from another result since there is no information using the keyword "Rp" and "product name".

Although SBPH algorithm often used in spam filtering, we used this as it can be used to find the keyword, with the step are:

a) Use the mixed text with some unidentified characters category

b) Construct a sub-text consist of 8 letters.

c) Generate features from each sub-text, starting from the first sub-text until the last sub-text.

d) Store the generates features in database

4.Displaying product information

The generate features from the previous step were shown by web user interface. The result is a comparison of product price and name.

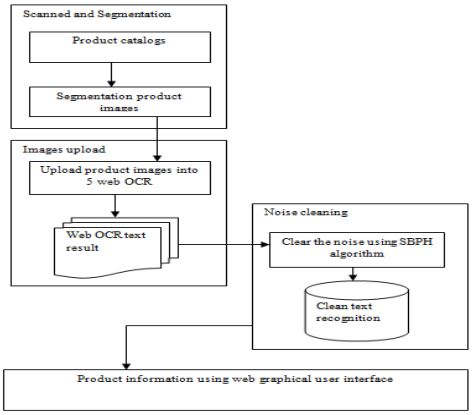


Fig 5: Research Methodology

3. Result and Discussion

From the scanned and segmentation, we identify input images from product catalogs example by its staples category, consist of milk, fruits & vegetables, fish & meats, and diapers. Our first hypotheses stated that the text in input images will be more recognize if using box line, rather that empty line. This is because web OCR will be easy to find the line and then the text. But, after the segmentation, we identify that more half of input images using the empty line.

Table 1: Staples and segmentation input images

Category	Box line	Empty line
Milk	6	46
Fruits & Vegeta-ble	30	16
Fish & Meats	18	16
Diapers	0	18
Total	54	96

The input images were upload to five web OCR. An example of the result can be seen in (Figure 6) and (Figure 7).



Fig 6: Text recognition result from color image

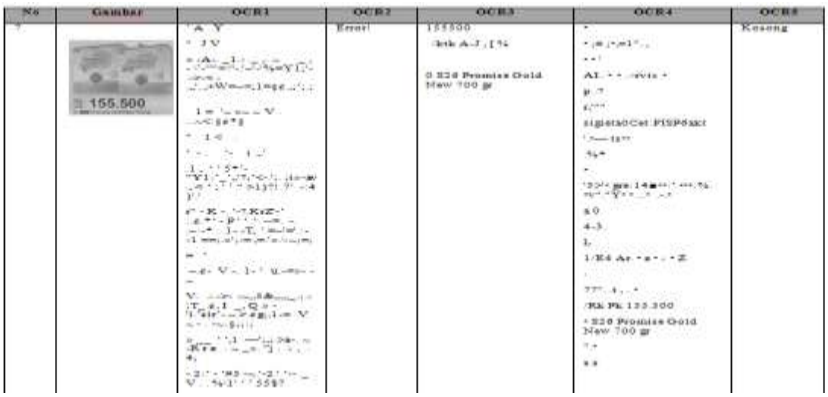


Fig 7: Text recognition result from grayscale image

Although we are using same input images, but when its differentiate with its color (grayscale and color) it will have a different result. The result from uploading input images into web OCR are varied but with example result for the input image number 7 are:

a) OCR1 will give different long unidentified characters.

b) OCR2 have the same result that is "error"

c) OCR3 & OCR4 give different mixed text with some unidentified characters

d) OCR5 have the same result that is "empty"

We ignore the result with "error" and "empty" since our objectives were to find the keyword with "Rp" and "product name". We identify that the keyword is often found in the mixed text with some unidentified characters. Further, we use the text from this category.

From the category, we evaluate the extraction result with a point:

1. 0.5 point if products name OR price was recognized by web OCR;

2. 1 point if products name AND price was recognized by web OCR.



Fig 8: Input images with ID 7

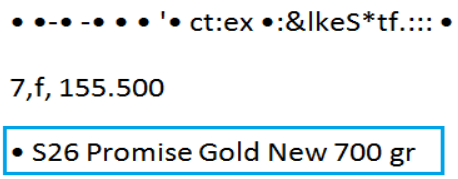


Fig 4: Text recognition for 0.5 point



Fig 10: Input images with ID 31

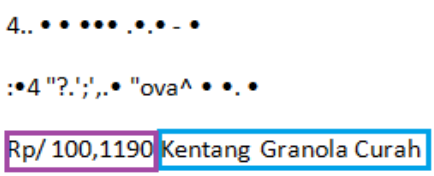


Fig 11: Text recognition for 1 point

Afterward, we categorize the result based on characteristic and segmentation parameter. Characteristic parameter refers to how much web OCR recognize the text whether appeared using box line or empty line. Meanwhile, segmentation parameter refers to how much web OCR recognize the text based on its property: image dimension, dpi, and bit depth.

a) Characteristic parameter

In (Table 2), we summarize the recognizable input images. Based on the result, OCR4 has better recognition among other web OCR to recognize the product name and price, whether it is grayscale or color. It is shown with the total of recognition has more than other web OCR. The result also shows that when we use the empty line in the input image, it gives more recognition result compared to box line. This result breaks our first hypothesis which stated that the text in input images will be more known if using box line.

Table 2: Image evaluation based on text recognition

Images type	Point	OCR1	OCR2	OCR3	OCR4	OCR5
Grayscale with box line (27 input images)	0.5	6	1	3	7	0
	1	1	0	4	2	0
Grayscale with empty line (48 input images)	0.5	5	0	16	20	5
	1	0	0	2	9	2
Color with box line (27 input images)	0.5	3	0	1	4	3
	1	1	0	6	9	1
Color with empty line (48 input images)	0.5	6	1	15	18	6
	1	1	0	7	16	0

Table 3: Percentage result based on text recognition

Images type	OCR1	OCR2	OCR3	OCR4	OCR5
Grayscale	16.00%	1.33%	33.33%	50.67%	9.33%
Color	14.67%	1.33%	38.67%	62.67%	13.33%

b) Segmentation parameter

Input images have properties such as dimension range, DPI, and bit depth. We use this property as an evaluation. We use image dimension from 532X788 to 2230X2851, 96 and 600 DPI, and 24 and 8-bit depth. Based on Table 4, OCR4 also has better recognition than others. Grayscale images give better recognition with DPI 96, bit depth 24 and image dimension more than 1000.

Table 4: Grayscale image recognition based on segmentation parameter

Images type	Point	Segmentation parameter	OCR 1	OCR2	OCR3	OCR4	OCR5
Grayscale	0.5	DPI : 96	9	1	9	14	0
		DPI: 600	2	0	10	13	5
		Bit depth: 24	9	1	9	14	0

		Bit depth: 8	2	0	10	13	5
		Dimension: >1000	6	0	14	17	3
		Dimension: <1000	5	1	5	10	2
	1	DPI : 96	1	0	5	4	1
		DPI: 600	0	0	1	7	1
		Bit depth: 24	1	0	5	4	1
		Bit depth: 8	0	0	1	7	1
		Dimension: >1000	1	0	3	8	2
		Dimension: <1000	0	0	3	3	0

Table 5: Color image recognition based on segmentation parameter

Images type	Point	Segmentation parameter	OCR 1	OCR2	OCR3	OCR4	OCR5
Color	0.5	DPI : 96	7	0	8	14	3
		DPI: 600	2	1	8	8	6
		Bit depth: 24	9	1	16	22	9
		Bit depth: 8	0	0	0	0	0
		Dimension: >1000	7	1	10	15	4
		Dimension: <1000	2	0	6	7	5
	1	DPI : 96	1	0	6	11	1
		DPI: 600	1	0	7	14	0
		Bit depth: 24	2	0	13	25	1
		Bit depth: 8	0	0	0	0	0
		Dimension: >1000	1	0	7	15	1
		Dimension: <1000	1	0	6	10	0

As shown in (Fig 4) and (**Error! Reference source not found.**), the text extraction result is not always the same as it is written in the input image. The result give a noise with 4 category: 1) Long unidentified characters; 2) Error; 3) Empty result and 4) Mixed text with some unidentified characters. Before we input the text recognition result into our system, we clear the text using SBPH algorithm with the steps below

a. Use the mixed text with some unidentified characters category

"/Rk Pk 155.500 .S26 Promise Gold New 700 gr"

b. Construct a sub-text consist of 8 letters. Below is some example of subtext from the text above:

i. /Rk Pk 15

ii. Pk 155.5

iii. 155.500

in the example above, the space letter is included, except in the 2nd sub-text, the front space is removed.

c. Generate features from each sub-text, starting from the first sub-text until the last sub-text. Because of the massive number of features by using 8 letters, we use only the combination of 5-8 letters. From each mixed text, it generates 64 features characters combination.

d. The next step is to check for each feature into databases. The checking is started from features with the most letters.

i. If a match with one of database items, the items found and the process stop.

ii. If not match, then continue to the next features

iii. If no features match, continue to the next sub-text.

As OCR4 gives a better result more than other web OCR, we use all text result from OCR4 to implemented it with SBPH algorithm and stored in database. The product searching result can be seen through web user interface using reference table. The references table consist of 1) Category table; 2) Reference Product Table. Category table is a category based on 9 (nine) staples as we mention in the introduction . The reference table is a detail table consist of many product and size. The tables are in **Error! Reference source not found.**and **Error! Reference source not found.**



Fig 12: Staples category table

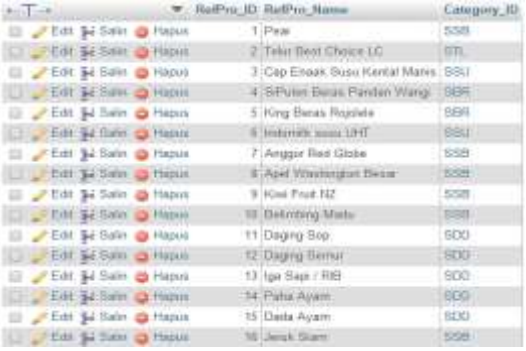


Fig 13: Reference table



Fig 14: Product comparison result

4. Discussion and conclusion

Modern market catalogs have important information regarding consumer needs. Information can be used for comparing the product price between another supermarket. Product catalogs are considered as a complex image because it has a different font, writing size, color, and writing position. To be recognized, product catalogs need an OCR to extract the text inside images. In this research, we upload 150 input image into five web OCR, considering the web has API and web service so it can be used in our system. Before upload, the image was segmented manually. This has consumed time and effort. For further research, the image can be segmented using an algorithm so it can be done effectively.

OCR4 has a better result in recognizing the text. This is shown with the testing result using 2 parameters: characteristic parameter and segmentation parameter. In each parameter, OCR4 able to recognize more input image among other web OCR. We use its web service partially with our system.

Although SBPH algorithm often used in spamming, in this research we used SBPH algorithm to generate the feature of product catalogs: product price and product name. We used combination of 5-8 letters and compare it to the reference product table from our database. Although the generate text consist of noise, but it has shown some feature that has the same product name in our table. For further research, the generate features needs more filtering in order decreasing the process for comparing with the table.

Acknowledgements

This work is based on research project held by Telkom University in Internal Research schema in 2015. Authors thank the University for supporting this research.

References

[1] Budiwati SD, Ananda D, Komala Sari S. Data Extraction on Hypermarket Catalogs using Sparse Binary Polynomial Hashing. In Konferensi Nasional Sistem Informasi; 2015.
[2] CustomOCR. [Online].; 2015. Available from: HYPERLINK "http://www.customocr.com/index.php?r=site/page&view=demos.tesseract_ocr" http://www.customocr.com/index.php?r=site/page&view=demos.tesseract_ocr .
[3] FreeOCR. [Online].; 2015. Available from: HYPERLINK "http://www.free-ocr.com/" http://www.free-ocr.com/ .
i2OCR. [Online].; 2015. Available from: HYPERLINK "http://www.i2ocr.com/" http://www.i2ocr.com/ .
OCRwebservice. [Online].; 2016. Available from: HYPERLINK "http://www.ocrwebservice.com/" http://www.ocrwebservice.com/ .
Google Drive. [Online].; 2015. Available from: HYPERLINK "https://drive.google.com/drive/" https://drive.google.com/drive/# .
Smith R. An Overview of the Tesseract OCR Engine. In Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR), IEEE Computer Society ; 2007.
[4] Google. Github. [Online].; 2016. Available from: HYPERLINK "https://github.com/tesseract-ocr" https://github.com/tesseract-ocr .
Patel C, Patel A, Patel D. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. International Journal of Computer Applications. 2012 October; 55.
[5] Sarawagi S. Information Extraction,. Foundations and Trends in Databases. 2007; 1: p. 261-377.
[6] Giri P. Text Information Extraction and Analysis From Images Using Digital Image Processing Techniques. Special Issue of International Journal on Advanced Computer Theory and Engineering (IJACTE). 2013; 2(1): p. 2319-2526.
[7] Yerazunis WS. The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How To Get Past It. In MIT Spam Conference; 2004.
[8] Yerazunis WS. Sparse Binary Polynomial Hashing and the CRM114 Discriminator. In Cambridge Spam Conference Proceedings Vol 1; 2003.
[9] Thomason A. Blog Spam: A Review. In CEAS 2007 - The Fourth Conference on Email and Anti-Spam; 2007; Mountain View, California.

[10] Aggarwal G, Ghosal S, inventors; United States patent US6728706 B2. 2011.