IDENTIFYING ENGLISH IDIOMS USING OPTICAL CHARACTER RECOGNITION ON MOBILE APPLICATION

Robby Kurniawan Budhi[1], Dwi Taufik Hidayat[2] and Jonathan Putra Pangapul[3]

[1] Informatics Engineering, Faculty of Engineering, Widya Kartika University
Surabaya, Indonesia
robby@widyakartika.ac.id

[2] Informatics Engineering, Faculty of Engineering, Widya Kartika University
Surabaya, Indonesia
dwitaufikhidayat@widyakartika.ac.id

[3] Informatics Engineering, Faculty of Engineering, Widya Kartika University
Surabaya, Indonesia
jonathanpangapul@gmail.com

Abstract

The use of English as a communication tool has become a common thing. However, it is sometimes difficult to solve compilations of phrases that sometimes cannot be interpreted literally. This phrase is commonly referred to as an idiom. This article commented on research consisting of introductory phrases in English using a mobile application. Identification is carried out using the Optical Character Recognition (OCR) facility in the Android application. Identification is done by analyzing images that contain certain text with variations in size, font type, writing color, and background color. The observations show a low success rate for text with Times New Roman font type, 6pt size with various writing colors. While for other font types, identification is also less successful for small font sizes and green or blue text colors.

Keywords: ***Optical Character Recognition (OCR), image processing, English idioms, mobile application***

Introduction

Idiom is a group of words established by usage as having a meaning not deducible from those of the individual words. To understand the meaning of idioms, vocabulary knowledge is needed. Idiom could not be translated word by word. Common translator application such as Google Translator also could not recognize the meaning of idioms.

An idiom “I put my foot in my mouth” has the meaning of “I said something that I shouldn’t have”, not literally by putting my foot into my mouth. People usually using translator to translate the meaning of words, but idioms could not be translated well. Therefore, English idioms translator is needed.

There are some English idiom dictionary that can be used to know the meaning of idioms in bahasa. But, the text should be entered using textpad or keyboard. For some people who had not familiar with English words, this input method could make some invalid results because the words typed can be wrong. To solve the problem, Optical Character Recognition (OCR) can be a good solution.

OCR is a tool that can be used to identifying text from a picture. By using OCR, some texts can be recognized from a picture. The picture could be taken from files or camera instead. Using the result, searching can be more easily done by matching some words from a database.

Research by Mollah (2018) and Mithe (2013) explain that OCR can be used to identify writings on the extension of the image. This method is in fact already a library to be used globally. Therefore, this study was using OCR to identify the image writing idiom. However, the use of OCR method on mobile applications require testing in identifying a particular piece of writing.

Optical Character Recognition (OCR) is an algorithm used to convert character shape on an image into a form of its original character. OCR has several processes, including: receiving, preprocessing, segementation, normalization, and the identification. However, OCR could not identify properly if there is disruption on the writing.

The working procedure of the OCR application are as follows:

1. Take the object of text using a camera so obtained a file image format (.jpg), or load some files that contained the text needed.

2. The image files were processed using the applications to recognize the text, where this device was doing the process of identifying the characters that exist in the image file.
3. The output of the software application was a text file that contains characters that have been identified and were ready to be processed further.

The success rate of text identifying application is rely on some factors:

1. Image quality text documents read as well as the level of its complexity (size, format, text, colour, background).
2. The quality of the software.
3. The quality of optical devices used (camera).

This study was made to evaluate the optimal text setting which could be recognized using OCR. In advance, the result of the process were used as input data to develop English-Indonesia idiom dictionary application.

Research methods

The data for this research was collected from literatures, and also from questionaires to evaluate and to know the respon of the user.

A. Dataset

The dataset used in this study were derived from English-Indonesia idioms dictionary by as much as 30 printed idioms with different font, size, text color, and backgound colors, and 30 image files made digitally using Image Processing Software, also with different conditions.

B. OCR Process

1. Data input: data collected using camera or loaded from files.
2. Preliminary process:
    a. Greyscalling and thresholding: converting image into greyscale; apllying threshold to convert image into binary image.
    b. Smoothing: image noise reduction process.
    c. Scaling: change of image size to minimize the load of process.
    d. Stroke thinning: mark the character lines.
3. Segmentation: to limit the image area.
4. Normalization: to change the thickness of detected characters.
5. Feature Extraction:
6. Character Recognition: to compare between detected characters and stored characters in database.
7. Finalization: spelling correction of detected words.

C. System Development Method

1. Requirement analysis: to analyze collected data.
2. System design: to design the system interface, system flow, and databases.
3. System implementation: to develop the system using Android Studio.
4. Testing and evaluation: to test the system flow, to identify bugs, and to know the user's respond for the application.

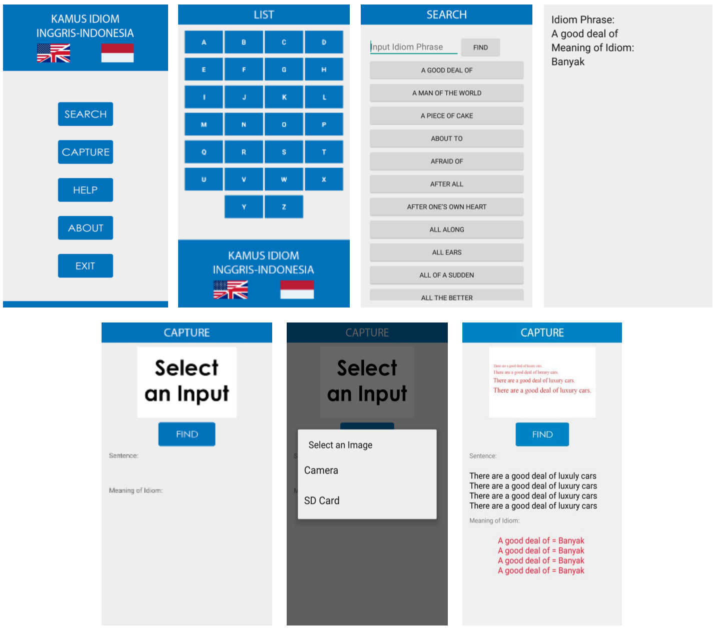
Research Results and Discussion

This experiment was conducted to test the OCR methods in identifying idiom phrases using mobile applications. Testing was done through two ways: First, OCR tested to identify words in the existing image in the memory of the smart phone. Second, the OCR tested to identify words in images obtained from the smart phone's camera. This was done to find out the capabilities of the OCR analysis in some cases.

Data on the experiment was images taken from memory card and also by taking the image by using default smart phone camera. So, the images have different conditions. With the difference of these conditions will strengthen the role of OCR in identifying the pictures. In addition to the results of the identification, condition of the image and typeface or font were also important to consider.

A. System Implementation Result

Picture 1 shows the user interface and screenshots of the application built as English-Indonesia Idioms Dictionary Application.



Idiom Phrase:
A good deal of
Meaning of Idiom:
Banyak

There are a good deal of luxuly cars
There are a good deal of luxury cars
There are a good deal of luxury cars
There are a good deal of luxury cars

A good deal of = Banyak
A good deal of = Banyak
A good deal of = Banyak
A good deal of = Banyak

Picture 1. Application Screenshots

There are some ways to search the meaning of idioms. First, the user can type the idioms using keypad. Second, the user can load images from files. Third, the user can capture the image by using camera. OCR proceeds the second and the third way of searching.

B. Testing Result

To test the result of OCR, four commonly used fonts were used as samples. There were Times New Roman, Arial, Calibri, and Comic Sans. The size of the fonts were 6, 8, 10, and 12. For the digital image files, font color was black, and the backgrounds were white, red, green, and blue. For the printed text, font colors were black, red, green, and blue.

Table 1. Identification result from digital image files

Font Type	Size (pt)	Text Color	Background Color			
			White	Red	Green	Blue
Times New Roman	6	Black	√	√	√	√
	8		√	√	√	√
	10		√	√	√	√
	12		√	√	√	√
Arial	6		√	√	√	√
	8		√	√	√	√
	10		√	√	√	√
	12		√	√	√	√
Calibri	6		√	√	√	√
	8		√	√	√	√
	10		√	√	√	√
	12		√	√	√	√
Comic Sans	6		√	√	√	√
	8		√	√	√	√

	10		√	√	√	√
	12		√	√	√	√

As the result of text identification, all of the images could be identified well on any background colors as shown in Table 1. The sentence “there are a good deal of luxury cars” which contains idioms “a good deal of” was identified perfectly. This sentence was also used in the printed text as experimental data for OCR using camera.

Table 2. Identification result from image taken by camera

Font Type	Size (pt)	Background Color	Text Color			
			Black	Red	Green	Blue
Times New Roman	6	White	X	X	X	X
	8		√	√	X	√
	10		√	√	√	√
	12		√	√	√	√
	14		√	√	√	√
Arial	6		√	√	X	√
	8		√	√	√	√
	10		√	√	√	√
	12		√	√	√	√
	14		√	√	√	√
Calibri	6		√	√	√	√
	8		√	√	√	√
	10		√	√	√	√
	12		√	√	√	√
	14		√	√	√	√
Comic Sans	6		√	√	√	X
	8		√	√	√	X
	10		√	√	√	√
	12		√	√	√	√
	14		√	√	√	√

Table 2 shows the result for the image taken using camera. Some images could not be identified correctly. There were some reasons to explain those conditions.

1. Serif type font, like Times New Roman, has extended line on every corner of the letter. In a small size, the line could make the identification invalid.
2. Greyscalling process for green and blue font color were also suspected as the reason of the fail.
3. The quality of the camera was also affecting the result.
4. The lighting condition when taking the picture could also be the reason of the fail.

When the experiment was done using the digital image files, OCR could identify all of the text correctly. Image noise was none, compared with the result of images taken by camera. In addition, digital image files usually improved by the software so the color composition is fixed.

An experiment was also done to measure the time for searching some idioms using the application compared with manual search in dictionary book. As the result, the using of English-Indonesia Idiom Dictionary Application was faster then the manual way.

Conclusions and recommendations

As the result of the experiments, OCR can identify most of the text with various font type, size, colors, and background colors with some limitation. The problem was in identifying small Serif type font like Times New

Roman, and also for green and blue text when using the camera. As the recomendation for the further research, OCR can be used to identify text in other application by considering those conditions.

Acknowledgment

This research was funded by Research Technology and Higher Education Ministry in 2018 Novice Lecturer Research Grant.

References

Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

Makkai, A. (2013). *Idiom structure in English* (Vol. 48). Walter de Gruyter.

Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical character recognition. *International journal of recent technology and engineering (IJRTE)*, *2*(1), 72-75.

Moeliono, A. M., & Ruddyanto, C. (1989). *Kembara bahasa: kumpulan karangan tersebar*. Gramedia.

Mollah, A. F., Majumder, N., Basu, S., & Nasipuri, M. (2011). Design of an optical character recognition system for camera-based handheld devices. *arXiv preprint arXiv:1109.3317*.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491-538.

Pangestu P. Penerapan Histogram Equalization pada Optical Character Recognition Preprocessing. ULTIMATICS. 2015 Jun 1;7(1).