Comparative Analysis of Efficient Image Segmentation Technique for Text Recognition and Human Skin Recognition

Septian Cahyadi¹, Febri Damatraseta², Victor Ilyas Sugara¹

¹Faculty of Mathematic and Natural Science Pakuan University of Bogor, Indonesia
²Department of Computer Science, Faculty of Master, Budi Luhur Universit, Jakarta, Indonesia.

Corresponding author: septian.cahyadi@unpak.ac.id

Abstract. Computer Vision and Pattern Recognition is one of the most interesting research subject on computer science, especially in case of reading or recognition of objects in real-time from the camera device. Object detection has wide range of segments, in this study we will try to find where the better methodologies for detecting a text and human skin. This study aims to develop a computer vision technology that will be used to help people with disabilities, especially illiterate (tuna aksara) and deaf (penyandang tuli) to recognize and learn the letters of the alphabet (A-Z). Based on our research, it is found that the best method and technique used for text recognition is Convolutional Neural Network with achievement accuracy reaches 93%, the next best achievement obtained OCR method, which reached 98% on the reading plate number. And also OCR method are 88% with stable image reading and good lighting conditions as well as the standard font type of a book. Mean while, best method and technique to detect human skin is by using Skin Color Segmentation: CIELab color space with accuracy of 96.87%. While the algorithm for classification using Convolutional Neural Network (CNN), the accuracy rate of 98%.

1. Introduction

Object detection is a topic of Computer Vision and one of the most interesting research subject on computer science, especially the real-time reading of the camera device. Object detection itself is the first step for the development of advanced research such as the introduction of text and skin for various purposes. The first segmentation is to select the part to be extracted, by detecting the exact object to be processed. In 2015 International Conference of Document Analysis and Recognition (ICDAR) report, on Text on Video challenge, the highest result of text recognition on video has not reached 50% or more precisely 45.18% for Text Localization and 41.84% for End-to-end Recognition. In its development there has been the application of the latest method to answer the challenge of Text Recognition in the video until reaching presentation level of 60%, although there has been a significant increase but seen from the value of the percentage still allows the application of other methods better. Unlike the results of the challenge to read text on still-image both focused text and incidental text, the percentage obtained has reached more than 90%. This is because the scene text is very difficult to read and recognize, because the text that appears on the video is affected by uneven lighting, textured background image or the slope of the position when the video is taken. Because of this, reading texts in the video is a challenge for this researchers.

For object detection to recognize patterns on human skin have different levels of difficulty, human skin has a special color and can be translated into a form of human skin color space. In general, human skin color can be seen on the face, neck, hands and feet. Some studies have successfully analyzed or created a pixel-based skin space in both image and video, including RGB color space, HIS / HSV, TSL, YebCr, etc. Several studies have conducted experiments on this technique such as, detect the gesture on a face...
in a video with different light conditions against different backgrounds, which will later be stored into memory [5]. The skin color space can also be used to detect gestures from the hand by extracting features on the hand to reduce noise on complex backgrounds with different light conditions, the way used by selecting vectors from pixels in blocks around the skin tone to obtain skin pixels the new one.

Technology of Computer Vision can also be used for people with disabilities such as helping illiterates (buta huruf) and deaf people (tuli) to know and learn about the letters of the alphabet. In this study we are trying to find better methods and techniques to implement this technology. We are trying to find a good method and technique to make an introduction to a text and detect the presence of human skin on a picture or video in real-time. In this study, we set up a framework for finding good methods and techniques to use.

1) Obtain video or image data from persons with disabilities (inputs).
2) Classify each frame on a video or image to find the detected object.
3) Display the classification results that are most likely or closer than the classification score (output).

The following is an identification of the problems that need to be addressed in the search for good methods and techniques:

1) The use of different light intensity and background on each image or video.
2) The position of the camera to capture an object on a picture or video is very influential.
3) Co-articulation: (when a sign is affected by the preceding or succeeding sign).

2. Method

In this study we will try to compare the segmentation used to recognize an object taken from image or video, that object is a text and human skin. Various methods will be presented in this research ranging from segmentation how to detect a text and detect human contours.

To detect an object to recognize text there are some of the most popular methods, based on the journal literature we refer to. The first most popular method is Neural Network, with variations, such as Artificial Neural Network, Convolutional Neural Network, and Recurrent Neural Network. This method generally performs the initial detection by reading the frame, one of which uses a MESR based detector to take the right frame for use and the easiest to read. Then we use the Neural Network method to classify and classify the text for segmentation. After segmentation done then can be used to do word recognition.

While to detect the presence of human skin contours stored in images or video by converting RGB image into the color space model. In addition to using segmentation of skin color can use certain techniques such as calculating the depth of an image and so forth. The use of skin color segmentation is a simple pixel based feature but its capability is very powerful.

3. Result and Discussion

First of all we look for research what has been done related to image processing for text recognition. This search leads us to a paper entitled "Estimating hidden parameters for text localization and recognition" written by Neumann L and Matas J from the university of Prague. In 2016 [34] also re-proposed this technique with improvements by adding synthetic fonts that have been trained for the classification process, the result is a significant improvement. Tests performed using the ICDAR dataset 2013 achieved an accuracy of 76.3% for text localization and for case-sensitive end-to-end text recognition achieved 45.2%. In the ICDAR dataset 2015 the results achieved for end-to-end text recognition of 41.8% and also in SVT achieved 68.1% result for end-to-end recognition.

Dibayan Chakraborty et al (2013) discusses the technique of taking the character or process of scanning a book more quickly using a digital video camera. This method is assumed to be faster process than scanning a book using a scanner engine. The steps in this method are (1) running a feature learning
algorithm on a set of image filters captured before training data to study a set of image features, (2) evaluating the convolutional features of the training image. Reduces the number of features that use spatial pooling, (3) trains a linear classification for either text detection or character recognition. Using the ICDAR dataset 2003 this method achieved an average yield of 82.66% accuracy.

In the experiment (Neumann L, 2013) at the ICDAR conference, a new method was introduced: Oriented Stroke Detection was used to segment the text. The end result of this approach shows a significant result compared to the previous study, tested on the ICDAR dataset 2011, the accuracy value obtained for text localization is 72%, but for end-to-end text recognition only get 45.2% percentage.

In a study conducted by (Baoguang Shi et al., 2016) with a new approach, RARE (Robust Scene Text Recognition with Automatic Refraction) with the same journal title received excellent results. Applied to two datasets, SVT and CUTE80, each received an accuracy of 77% and 59.

Result of research related to best text recognition so far obtained by using method of convolutional neural network (CNN). Introduction is achieved by performing a multi-directional classification across the dictionary to obtain potential words. As a result, of the many datasets tested, from Synthetic Fonts, ICDAR03, SVT, ICDAR13, and IIT5k, an average of 93% was obtained for accurate cropped word recognition. But for end-to-end text spotting achieved an average yield of 75%.

Prajit Ramachandran (2016) discusses the implementation of Faster R-CNN algorithm to detect objects on Video. The results of this implementation are presented at ICCV 2015 ImageNet and MS COCO Visual Recognition Challenges Joint Workshop. This study uses a video clip from the ImageNet 2015 Object Detection from Video Challenge. In this journal it was submitted that the author uses VGGNet as the Convolutional backend of Faster-RCNN. VGGNet provides substantial improvements beyond the actors described by Zeiler and Fergus. From the result of set validation, the writer get the accuracy of mAP 53%.

U. Bhattacharya et al (2016) discusses the technique of detecting Scene Text without limitation using a method that is easy to implement. Based on efficient stroke detection of pixels and comparing with surrounding pixels. Stroke-specific keypoints detect efficiently and text fragments sequentially in extracts using local threshold following keypoint properties. The classification method effectively calculates features to eliminate non-text areas. This method supports a wide variety of scripts (Latin, Hebrew, Chinese, etc). The results of stroke detector testing with FASText on ICDAR dataset 2013 showed an accuracy of 82% with an increase in process speed 4 times higher than MSER.

Annmaria Cherian (2016) discusses the application of the Hough Transform method to improve the orientation of the scene image and use efficient character detection and localization techniques. The method used to filter non-text components is SVM. After the filtering process, character recognition is used to recognize the text accurately. The dataset used in this study is the Text-Perpective Scene dataset which gives the image oaoan ads or billboards located on the edge of the highway with a perplexal image is skewed. And the results of the test show that the method used is quite simple and exceeds the existing method.

Lukas Neumann et al (2016) tried to approach the Extremal Regions used in this journal to read Scene Text localization and recognition or referred to as Photo OCR. The highest probability of characters is then selected in the final stages when the context of each character is known. This method is tested on three public datasets. First use ICDAR 2013 dataset and achieve outstanding results in text localization. Both use SVT datasets and the results significantly outperform other methods. The third dataset uses ICDAR 2015, and achieves exceptionally competitive accuracy.

Raja Vikramdeep Singh et al (2016) discusses the implementation of the OCR methodology to help collect incoming vehicle data through toll roads and traffic monitoring in India. Stages in the process include four phases, the first phase is image acquisition, second extraction, third pixelation and segmentation, and fourth is Recognition. The result obtained an accuracy of 98%.
Gokhan Yildirim et al (2016) describes text recognition techniques on natural images using multiclass hough forest methods. This journal offers a technique for detecting and recognizing text in various conditions by searching for words directly without reducing the image to a region or individual character. They dedicate three contributions. And by using the ICDAR 2003 dataset we get outstanding results in word recognition in terms of performance, which is 85.7%.

In the research (Hartanto and Kartikasari, 2016) conducted HGR research using SIBI as the object of his research. The approach used is non-contact approach by utilizing webcam camera that has been embedded, the method used is trying to catch or detect the movement of the hand by using Skin Color Segmentation model YCbCr color space and Back Propagation Neural Network classification to look for patterns on hand movements that are detected at Webcam in real-time. In their experiments, the first step used to perform the HGR are catching the HGR on the frame. The last stage is the process of classification with Back Propagation Neural Network to find the pattern of input and match it with the SIBI pattern alphabet dataset and display the classification results into a text. From the experiments that have been tested, the accuracy is 62.6% of 26 letters of the SIBI alphabet.

A further reference to the research belongs (Garcia and Viesca, 2016), they experimented on HGR to translate American Sign Language (ASL) in real-time. This research is similar to previous research using Skin Color Segmentation to know the existence of HGR. The HSV color space model and the Convolutional Neural Networks (CNN) classification were used in ASL testing. However they use the ASS pattern dataset ILSVRC2012 from letter a to k (without j) that have been issued by Surrey University and Massey University. From the results of the tests they have done, CNN is used to determine the percentage of accuracy of the object under study. Obtained a percentage of 98% for ASL letters from a to e. While from the letters ASL a through k (without the letter j) obtained by 70%.

Translating sign languages was re-examined by (Shinde and Kagalkar, 2015). Marathi's sign language became the object of their research. The use of Skin Color Segmentation with HSV color space model is used to detect human skin pixel contours in the image. While the classification to understand the input image using Euclidean Distance. Accuracy value in detecting 6 letters Marathi Sign Language can reach 30%-100%. However, the time of classification process used is quite time consuming.

Thai Sign Language (TSL) becomes the object of research (Pariwat and Seresangtakul, 2017) to translate sign language into text. The technique used is using SVM technique to get accuracy value in classifying TSL applying Linear Kernel, Polynomial Kernel, and RBF Kernel methods. The accuracy value in detecting 15 TSL letters was 91.20%.

The Arabic Sign Language (ArSL) translator authentication is investigated by (Saleh et al, 2016) using SOFTKINECT sensors to detect the movement of the human hand, while for the classification of the results taken by sensors using PCANet. The accuracy value for the introduction of 28 ArSL reaches 96%-99.5%.

The research belongs (Shweta et al, 2016) trying to translate ASL sign language into a voice or text to speech. Techniques used in detecting HGR simply change the image into binary form by using Geometric operations, Neighborhood and block operations. For classification use Mel Frequency Cepstrum Coefficients provided by Matrix Laboratory. The results achieved in detecting 9 ASL letters obtained an accuracy of 91%.

The Alphabet Bangla Sign Language is researched by (Azher and Chwdhury, 2016) to translate it into text using Skin Color Segementation model HSV color space techniques to detect the presence of human skin contours and classification for recognition using SVM. From 6 letters Bangla Sign Language got an accuracy score of 97.7%.

(George et al, 2016) conducted a study on how to detect an autonium disease in human skin by using the object of research from the image of psoriasis. To know Skin Map (SM) in image use two approaches that is Otsu's Thresholding and Decision Tree (DT) by using Morphological operation. And to know skin on image using Skin Color Segmentation with model YCbCr, HSV, RGB, XYZ and CIELab color
space. Histogram-based Bayesian classification is used to conduct trials of 100 psoriasis images. Where this classification is so effective to test the methods presented because Skin Probability Maps (SPMs) generated from the color space presented can produce the best SM for psoriasis images. From the test results obtained for color space YCbCr with Otsu's thresh percentage highest in TNR is 98.81% while for DT the largest percentage value is in TNR of 97.61%. The HSV color space of Otsu's thresh percentage is TNN 98.81% and DT is 97.91%. And the CIELab color space of the largest percentage value is in the TPR of 92.49% with Otsu's Thresh and DT of 94.60% in TPR.

The research (Ting and Lin, 2016) attempted to test the CIELab color space technique on 20 images with minimal contrast. They want to prove that using CIELab and Fuzzy Automatic Contrast Enhancement (FACE) techniques can increase the value of the contrast better so as to get clearer and better viewed images. From the results of experiments they have done obtained a contrast increase value of 20 images obtained RMSC value of 12.6 to 243.7.

4. Conclusions

From the elaboration of the use of methods and techniques that have been done by some previous researchers, it can be concluded that in detecting an object, the most preferred is the speed in process to produce better and more accurate information.

4.1. Table comparison of Methodology Text Detection

Based on table 1, many discuss about text recognition, the most widely used and accurate method is Convolutional Neural Network with achievement accuracy reaches 93%. Although researchers generally modify the CNN method, resulting in a variety of results, it can generally be argued that the use of CNN is the best method for the initial phase of text recognition and segmentation. While the next best achievement obtained OCR method, which reached 98% on the reading plate number that has the same font on each vehicle, seen from the value it does look very promising, but the results can not be achieved for reading the text in general with various font types. The best results obtained by the OCR method are 88% with stable image reading and good lighting conditions as well as the standard font type of a book, rather than real-time text readings encountered on the road with very diverse variations.

| No | Title                                                                 | Author                                                  | Year Publish | Methodology                          | Result  |
|----|----------------------------------------------------------------------|----------------------------------------------------------|--------------|--------------------------------------|---------|
| 1  | Text recognition in natural images using multiclass hough forests    | Gokhan Yildirim, Radhakrishna Achanta, Sabine Susstrunk | 2013         | Hough Forest                         | 85.7%   |
| 3  | SceneTextReg: A Real-Time Video OCR System                          | Haojin Yang, ChengWang, Christian Bartz, Christoph Meinel | 2016         | Convolutional Neural Network (CNN)    | 82%     |
| 4  | An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition | Baoguang Shi, Xiang Bai                                  | 2016         | Convolutional Recurrent Neural Network (CRNN) | 83.4%   |
5\(^b\) Robust Scene Text Recognition with Automatic Rectification

Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai

2016 Robust text recognizer with Automatic Rectification (RARE) 59%

6\(^c\) Automobile Number Plate Recognition And Extraction Using Optical Character Recognition

Raja Vikramdeep Singh, Navneet Randhawa

2016 Optical Character Recognition Methodology 98%

7\(^c\) Automatic Localization and Recognition of Perspectively Distorted Text in Natural Scene Images

Annmaria Cherian

2016 Hough Transform & SVM 72%

8\(^c\) Efficient Scene text localization and recognition with local character refinement

Lukas Neumann, Jiri Matas

2015 MSER with Character Refinement 77.1%

9\(^c\) Reading Text in the Wild with Convolutional Neural Networks

Jaderberg et al

2016 CNN AVG: 93%

\(^a\) Research base on Text Recognition by Scene Text Video Footage (dataset)

\(^b\) Research base on Text Recognition by Video Real-Time (Streaming)

\(^c\) Research base on Text Recognition by Image

4.2. Table comparision of methodology Human Skin Detection

From table 2 it is found that a good skin color segmentation model is used, ie CIELab color space. It is better than other color models that with the value of accuracy can reach 96.87%. And for the classification method obtained Convolutional Neural Network has an accuracy of 98%.

Table 2. A table with result of comparison methodology of Human Skin Detection.

| No | Title                                                                 | Author                                      | Publish | Methodology                      | Result                  |
|----|-----------------------------------------------------------------------|---------------------------------------------|---------|---------------------------------|-------------------------|
| 1  | Implementation of skin color selection prior to Gabor filter and neural network to reduce execution time of face detection | Mejda Chihaoui, Akram Elkefi, Wajdi Bellil dan Chokri Ben Amar | 2015    | Skin Color Segmentation: YCbCr color space dan Sd-Gabor-NN | Execution time (sec): 12.46500 – 77.77300 |
| 2  | A Sign Language Recognition Approach For Human-Robot Symbiosis        | Afzal Hossian, Shahrin                      | 2014    | Skin Color Segmentation: YIQ    | Accuracy: 80%-92%       |
| Year | Title                                                                 | Authors                                      | Model/Technique                                                                 | Accuracy |
|------|----------------------------------------------------------------------|----------------------------------------------|--------------------------------------------------------------------------------|----------|
| 2016 | Android Based Real-Time Static Indonesian Sign Language Recognition System Prototype | Chowdhury, and Asma-ull-Hosna | Back Propagation Neural Network | 62.6% |
| 2016 | Real-time American Sign Language Recognition with Convolutional Neural Networks | Rudy Hartanto, Annisa Kartikasari | Skin Color Segmentation : YCbCr color space; Back Propagation Neural Network | 70%-98% |
| 2017 | Thai Finger-Spelling Sign Language Recognition Using Global and Local Features with SVM | Thongpan Pariwat dan Pusadee Seresangtakul | SVM Teknik : Linear Kernel Accuracy, Polynomial Kernel, RBF Kernel, Sigmoid Kernel | 91.20% |
| 2016 | Arabic Sign Language Fingerspelling Recognition from Depth and Intensity Images | Saleh Alyl, Basma Osman, Walaa Aly dan Mahmoud Saber | Sensor SOFTKINECT and PCANet | 96%-99.5% |
| 2016 | Real Time Two Way Communication Approach for Hearing Impaired and Dumb Person Based on Image Processing | Shweta S. Shinde, Rajesh M. Autee dan Vitthal K. Bhosale | Mel Frequency Cepstrum Coefficients and Geometric operations, Neighborhood and block operations. | 91% |
| 2016 | Hand Sign Language Recognition for Bangla Alphabet using Support Vector Machine | Md Azher Uddin dan Shayhan Ameen Chowdhury | Skin Color Segmentation : HSV color space dan SVM | 97.7% |
| 2016 | Pixel-based Skin Segmentation in Psoriasis Images | Y. George, M. Aldeen, and R. Garnavi | Skin Color Segmentation : CIELabs color space and Bayesian Classifier | 90%-97% |
5. References

[1] Shinde, Amitkumar dan Kagalkar, Ramesh. *Advanced Marathi Sign Language Recognition using Computer Vision*, International Journal of Computer Applications, Volume 118 – No. 13, pp.1-7. 2015.

[2] Hartanto, Rudy dan Kartikasari, Annisa. *Android Based Real-Time Static Indonesian Sign Language Recognition System Prototype*, International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-6. Oct. 2016.

[3] Garcia, Brandon., dan Viesca, Sigberto Alarcon. *Real-time American Sign Language Recognition with Convolutional Neural Networks*, Stanford University., Stanford, CA, 2016.

[4] Pariwat, Thongpan. dan Seresangtakul, Pusadee. *Thai Finger-Spelling Sign Language Recognition Using Global and Local Features with SVM*, 9th International Conference on Knowledge and Smart Technology (KST), pp. 116-120. March 2017.

[5] Saleh Alyl, Basma Osman , Walaa Aly dan Mahmoud Saber. *Arabic Sign Language Fingerspelling Recognition from Depth and Intensity Images*, International Computer Engineering Conference (ICENCO), pp. 99-104. 2016

[6] Shweta S. Shinde, Rajesh M. Autee dan Vitthal K. Bhosale., *Real Time Two Way Communication Approach for Hearing Impaired and Dumb Person Based on Image Processing*, International Conference on Computational Intelligence and Computing Research (ICCIC), pp.1-5. 2016.

[7] Md Azher Uddin dan Shayhan Ameer Chowdhury., *Hand Sign Language Recognition for Bangla Alphabet using Support Vector Machine*, International Conference on Innovations in Science, Engineering and Technology (ICISET). Pp 1-4, 2016.

[8] Y. George, M. Aldeen, and R. Garnavi., *Pixel-based Skin Segmentation in Psoriasis Images*, International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Pp 1352- 1356, 2016.

[9] Po Ting Lin dan Boting Rex Lin., *Fuzzy Automatic Contrast Enhancement Based on Fuzzy C-Means Clustering in CIELAB Color Space*, International Conference on Mechatronic and Embedded Systems and Applications (MESA). Pp 1-10, 2016.

[10] Y. George, M. Aldeen, and R. Garnavi, *Pixel-based Skin Segmentation in Psoriasis Images*, 2016 IEEE 38th Annual International Conference, Engineering in Medicine and Biology Society (EMBC), 2016.

[11] Lukas Neumann, Jiri Matas, *Real-Time Lexicon-Free Scene Text Localization and Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016

[12] Haojin Yang, ChengWang, Christian Bartz, Christoph Meinel, *SceneTextReg: A Real-Time Video OCR System Haojin*, Hasso Plattner Institute (HPI), University of Potsdam, Germany. 2016

[13] Baoguang Shi, Xiang Bai. *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. IEEE Transactions On Pattern Analysis And Machine Intelligence. 2016

[14] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai. *Robust Scene Text Recognition with Automatic Rectification*. IEEE Conference on Computer Vision and Pattern Recognition Robust. 2016