THE DEVELOPMENT OF INDOOR OBJECT RECOGNITION TOOL FOR PEOPLE WITH LOW VISION AND BLINDNESS

Rhio Sutoyo[1]; Andry Chowanda[2]

[1,2]Computer Science Department, School of Computer Science, Bina Nusantara University
Jln. K.H. Syahdan No 9, DKI Jakarta 11480, Indonesia
[1]rsutoyo@binus.edu; [2]achowanda@binus.edu

Received: 8th February 2017/ **Revised:** 7th March 2017/ **Accepted:** 13th March 2017

Abstract **-** The purpose of this research was to develop methods and algorithms that could be applied as the underlying base for developing an object recognition tools. The method implemented in this research was initial problem identification, methods and algorithms testing and development, image database modeling, system development, and training and testing. As a result, the system can perform with 93,46% of accuracy for indoor object recognition. Even though the system achieves relatively high accuracy in recognizing objects, it is still limited to a single object detection and not able to recognize the object in real time.

Keywords: object recognition, computer vision, tool, blindness, low vision

I. INTRODUCTION

According to Indonesia Blind Union (PERTUNI - Persatuan Tunanetra Indonesia) in 2009, there were more than 1,5 million blind people and more than 1,2 million people who had low vision (weak sight) in Indonesia. Then, the number is increasing over time. In addition, at the end of 2012, Indonesia was rated as the second highest rate of blindness in the world (Jakarta Globe, 2012). Across the globe, the number of visually impared people or blind people is predicted to increase by 20% within the next 50 years (Bujacz & Strumillo, 2006). The reason is because the aging demographic around the world. With a sight problem, daily activities such as walking, reading, recognizing object and people, and visiting places can be quite troublesome for those people.

There have been many researches to help visually impaired people, escpecially in the field of computer vision. For example, computer vision could be used to create an "artificial eye" for blind or visually impared people. An artifical eye would allows sightless-human to move to some destinations or places, read books (with the help of hearing or Braille), recognize objects and faces, and many more. According to a survey from previous works by Manduchi and Coughlan (2012), the most useful tools were Braille note takers, text magnifiers and screen readers, and document scanners with Optical Character Recognition (OCR). With the help of technology, Braille notetaker such as BrailleNote Touch can help blind or visually impared people to take notes. Furthermore, text magnifiers and screen readers have become the standard feature in modern laptops and smartphones nowadays (i.e. VoiceOver feature in Apple's iOS, TalkBack feature in Google's Android). Meanwhile, the OCR technology helps character recognition in images. After the scan and recognition processes have completed, the recognized text can be readed back to the users. In short, Braille and OCR technology help blind and visually impared people to read without using their sight. Besides character recognition, the other problem of people with bad eyesight is object recognition. Bujacz and Strumillo (2006) utilized stereophonic to detect whether there was an object nearby. The research used high sound to represent objects that were close to the users, low sound to show objects that were far from the user, and no noise for no objects near the user. The representation of the environment was generated by several scanning process of 5°x5° sections of the environment view. The type of the scans included basic (front) scan, wide (vertical) side-scan, and horizontal side-scan. The test were conducted with 10 volunteers participated in three types of trials which were mobility, orientation, and a combination of both. The results indicated that the users were able to imagine indoor structure based on the casted sound from the indoor-objects. However, this research had weaknesses in recognizing objects and difficulty in implementation in the room that had a complicated structure.

Then, Ran *et al.* (2004) developed a tool to help blind people to walk and know the surrounding objects by using ultrasound sensors, wireless, and Global Positioning System (GPS). This tool could be used in both indoors and outdoors. In indoor, Ran *et al.* (2004) used an ultrasound sensor that was attached to the house and also to the user. The tests were conducted by using 4 HE900M pilots, 2 HE900T beacons, and a RS485/RS232 converter. The pilots were placed on the four corners of the house. With the choice of the placement, the pilots provided complete map of the house. Moreover, the beacons were attached to the user to receive the ultrasound signals from the pilots. The genereated outputs were position (i.e. the distance to the refrigerator was 2 meters) and orientation (i.e. turn right 45 degrees and walk ahead). Thus, users could know the location of the object. Meanwhile, in outdoor, they used GPS and wireless sensors. The tests were conducted by using a Trimble PROXRS, 12 channel integrated GPS/Beacon/Satellite receiver with multi-path rejection technology. Users were able to use voice guidance to determine whether they were out of the lane or not. Even though these approaches were useful, it required expensive setup and could only be used in the rooms that had sensors installed. Moreover, more sensors were needed to achieve a better precision result or the room should be bigger.

On the contrary, Anjum (2012) used a map of existing objects combined with the constructed dataset. This research helped blind and visually impared people to reach his/her destination by providing direction. In order to get the destionation route, users needed to send an image of their current locations to the trained system for localization.

After the localization process had finished, the system generated the path to the destination. This research used 128-dimensional Scale-Invariant Feature Transform (SIFT) method for classifying objects in the room. This method was invariant to rotation, luminosity, and scale. Thus, it could determine the location of the room. While the system showed promising results for recognizing space and navigation, it was only limited to rooms that already had well-built maps. In this research, computer vision technique was utilized to achieve affordable setup cost with the precise result.

According to Pinto *et al.* (2008), the fundamental problem for the recognition of the object was the unlimited number as input (e.g. position, scale, pose, illumination, background images, and others) from a 2D image captured by the retina or camera, and position changes of the image captured by the camera. To this day, it is still a major problem to recognize real-world objects. Moreover, several researchers have been successfully addressing the problem of object recognition. For example, Gu *et al.* (2009) used region extractions to overcome the problem of estimation combination toward the bag of regions derived from a region tree. Regions were described by using a rich set of cues (i.e. shape, color, and texture) inside them. Furthermore, a discriminative max-margin framework were used to measure the region weights. Then, Hough voting scheme was used to generate object locations, scales, and support scenario possibilities. The 2D image was processed by using a region tree and bag of regions. Thus, detection and image segmentation could be done. The evaluation were done by using ETH Zurich (ETHZ) shape and the Caltech 101 databases. They claimed that the algorithms were able to reduce the total number of estimation combination. In addition, Pinto *et al.* (2011) used several measurement such as Pyramid Histogram of Gradients (PHOG), Pyramid Histogram of Visual Words (PHOW), Geometric Blur, Sparse Localized Features (SLF), Scale-Invariant Feature Transform (SIFT), Pixels, V1-like for object recognition. Caltech-101 image categorization task were used as a point of reference. Based on the conducted experiment, V1-like measurement had the highest performance among the others.

Similarly, Farhadi *et al.* (2009) used attributes that were specific to an object to improve the accuracy of object recognition. Their research utilized common features to classify the object and compared it with the specific features resulting in the accuracy in object recognition. Hence, it could not only recognize and name familiar objects, but also report unusual aspects of an object (i.e. "spotty cat" instead of "cat"), describe about unfamiliar objects ("skinny and four-legged" instead of "unknown"), and learn about new objects even without visual examples. Unfortunately, their datasets were limited only to few objects. Moreover, this research was not perfect as there were multiple class objects and efficiency issues in the process. The disadvantages of this research could be answered by Hochbaum and Singh (2009). They adopted co-segmentation by using Markov Random Field (MRF). With this technique, the segmentation process for similar object's regions could be done on similar (or same) object that appears in two different images. It increased efficiency and accuracy of the object recognition. Although the algorithm run similary on normal images, it gives time-saving process for larger images.

In addition, Li, Crandall, and Huttenlocher (2009) constructed datasets of images obtained from the internet such as Facebook and Flikr using geo-tagging. The datasets were used to detect place name by using Scale-Invariant Feature Transform-based bag-of-word features. This research showed a good performance when it processed approximately 30 million existing images. Meanwhile, Li, Socher, and Fei-Fei (2009) utilized classification, annotation, and segmentation by finding the relationship between objects so that it could be concluded that there was a likelihood scene. Unfortunately, these methods required complete datasets. Next, Carreira *et al.* (2012) used a ranking system towards object segmentation results. Their research used figure-ground segments generated by bottom-up computational processes. The method was conducted by using three steps. Those were producing a set of figure-ground segmentation hypotheses for each image using the combinatorial CPMC segmentation algorithm, scoring each object category to assess the likelihood that a segment hypothesis belongs to that class, and sorting the segment hypotheses by their scores and also consecutively making detection and segmentation decisions depending on a weighted combination of responses collected at top-level segments. It showed more accurate results of object recognition. Nevertheless, this research still indicated an error in some cases such as boundary objects that were not clear and objects with a similar class. Furthermore, Divvala *et al.* (2009) analyzed the object detection methods based on the existing variables. It was the changes in confusion matrices and accuracies with respect to size and occlusion, and analysis of sources and uses of context. This research explored the importance of contextual reasoning for object recognition. The contextual reasoning reduced the detection of errors and also made the remaining errors more reasonable. This research was used as reference for object detection, segmentation, and scene recognition.

This research presents a model for categorization and recognition limited to indoor objects as the first step to develop a tool to help blind and low vision people. Then, the aims are to take the advantages of the computer vision technique to build a tool that can help those people. With the help of this "artifical eye" tool, visually troubled people are able to recognize everyday object easier. With computer vision technique, the researchers can build tool that can recognize people (Chowanda *et al*., 2014, 2016) and their expressions (Sutoyo, Harefa, & Chowanda, 2016), object (Farhadi, Endres, Hoiem, & Forsyth, 2009), places (Li, Crandall, & Huttenlocher, 2009), presentation process (Sutoyo *et al.*, 2015, 2017), and others.

II. METHODS

Figure 1 illustrates the overall system architecture. The camera captures all the information from the world in 2D format. The captured image will be preprocessed (e.g. image enhancement) before the researchers extract the features of the image. Then, the system classifies and recognizes the object in the processed image based on the trained models by using existing datasets. Finally, the system will provide a sound feedback using Text-to-Speech to the user.

The first stage is image processing and enhancement. In this stage, images captured by the camera are processed by the feature extraction stage. The images will be converted into gray-scale images with the help of OpenCV with the command:

```
IAT_Rgb2Gray * R2G = new

IAT_Rgb2Gray();
```

Then, they will be processed by using canny edge detector with the command:

```
cvCanny (img1, img1, 10, 50, 3);
```

Lastly, those images will be converted into HSV (Hue, Saturation, Value) mode with the help of OpenCV with the command:

```
cvCvtColor(img, Img1, CV_RGB2HSV);
```

After the image-processing stage has been done, the processed images will be sent to the feature extraction step.

The second stage is feature extraction. In this step, image separation will be done before the feature extraction. This process takes the work of Kim *et al.* (2010) where they separate the image into five regions or area block. The idea of this approach is because the object of interest in photos is mostly located in the center of the images (e.g. people's faces). Thus, for indoor or outdoor, the border image classification has more information than the center. The example for image separation can be seen in Figure 2.

Image feature is calculated for each block by setting Region of Interests (ROI) with a particular block. The next process is edge features extraction which used equation (1):

$$F_{i,n}^{EOH} = \frac{E_{i,n}}{\sqrt{\sum_{j=1}^{k}(E_{i,j})^2+\varepsilon}} \quad (1)$$

Ei, n is the value of an angle n, and block i. ε is a small integer. Moreover, the feature is calculated for eight quantised corners of each block that provides 40-D feature vector. After edge feature extraction has been done, the next process is color feature extraction which is implemented by using equation (2).

$$F_{i,n}^{COH} = \frac{E_{i,n}}{\sqrt{\sum_{j=1}^{k}(E_{i,j})^2+\varepsilon}} \quad (2)$$

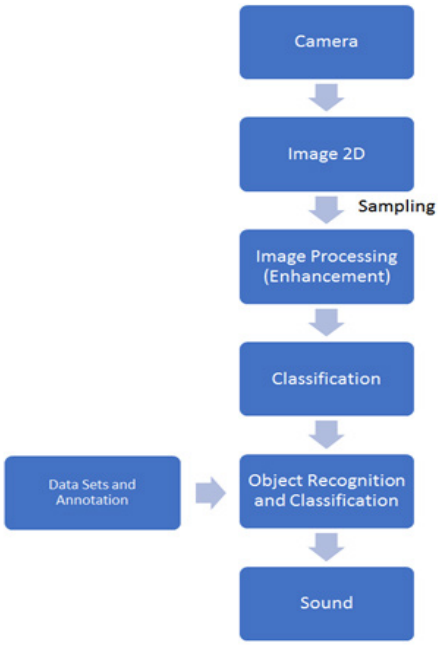

Figure 1 System Architecture

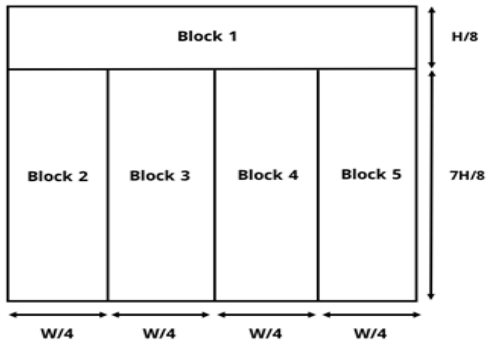

Figure 2 Image Division in Five Areas or Regions
Source: Kim et al. (2010)

This process is similar to edge feature extraction process by using Edge Oriented Histogram (EOH). The vector of features obtained from the previous process (i.e. edge extraction features, color feature extraction) is initially stored in an Extensible Markup Language (XML) file. Then, these features are combined and stored in a Comma-Separated Values (CSV) file so the data can be read easily for Support Vector Machine (SVM) classifier training. The process to create Edge and Color Orientation Histogram (ECOH) can be seen in Figure 3. After all the process have been done, the feature extraction phase is complete. Furthermore, the data obtained from this stage will be used at SVM image classification step.

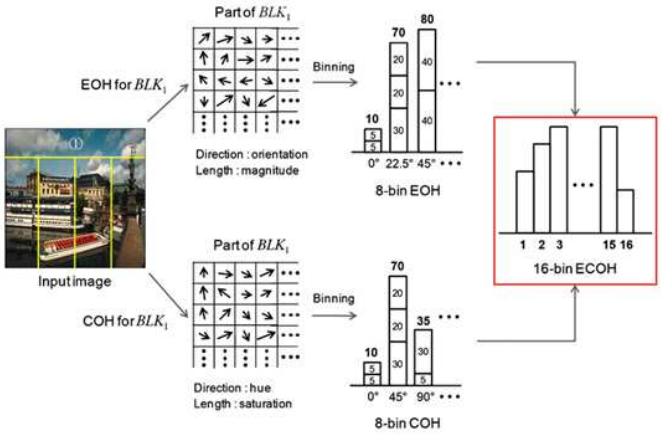

Figure 3 Feature Extraction Process
(Source: Kim et al., 2010)

The third stage is SVM Classification. To train SVM image classification (i.e. data training), all features in the CSV file from the previous step are converted to the form of a matrix. At the end of each attribute of the vector, the feature is placed to determine an image category. Once the matrix has a parameter, data practice must be organized to train the SVM classifier. SVM image classification will be implemented by using the LIBSVM. LIBSVM is a software developed by Chih Chung Chang and Chih Jen Lin that has General Public License (GPL) (Chang, 2011). Next, the classifier will be stored in XML file format. The XML file created is used to predict the class of new images. New images must be classified. In addition, all features needed for ECOH are calculated and predicted from the image. Then, the result will be stored in a matrix. Next, the matrix will be given to the SVM classifier to determine the image class. The result depends on the feature score which given it

as an output. These results are used for the final step, which is object detection and recognition.

The final stage is object detection and recognition. This stage requires the sampling process. Sample creation can be done with the help of OpenCV library which generates a vector file (VEC) of a positive description file. The sample is created by using sub-image marked from the real image. Width and height of this image are 24 x 24 samples (pixels). However, that size can be changed by the input parameter. To run it, the sample must be placed in folder bin in OpenCV where the application is located in the file directory. If it is success, this process produces the amount of sample which is identified successfully and unsuccessfully by inserting the existing dataset.

Figure 4 Successful Detected Objects

The classification in this research can identify several objects such as chairs, clocks, and other indoor objects. The example can be seen in Figure 4. The image is collected from a database of images on the website of the California Institute of Technology (CALTECH, 2017). The dataset used in this research can be seen in Table 1. Moreover, the researchers model an image database for training purposes by combining three existing datasets. Then, by using the image database, the researchers train the SVM classifier for object recognitions using algorithm described (SVM Classification).

Table 1 Datasets

Name	Number of Images	Providers
House Dataset	1000 jpg	California Institute of Technology (Philip & Updike, 2001)
LHI 8 Image	1200 jpg	Yao, Yang, and Zhu (2007)
MIT Image	4289 jpg	MIT Media Lab (Oliva & Torralba, 2001)

In Figure 5, the example of an image as the identification results of the system prototype is shown. The current system still uses 2D images and are no able to do the real-time identification. The image is taken from a database of images on the website of the California Institute of Technology. The monitor has 78% level of confidence, the glass has 83% level of confidence, and the watch has 78% level of confidence.

Figure 5 Identification Results of the System Prototype

III. RESULTS AND DISCUSSIONS

The evaluation is done by using the image database from California Institute of Technology website and MIT website. About 107 images are used in the testing with a wide variety of objects from the provided dataset. The result of object recognition achieves a high percentage of detection such as 93,46%. The accuracy or confidence level represents the number of images that are correctly detected by the system during the evaluation (i.e. number of true positive divided by all images used in the evaluation). Hence, it is:

$$100/107 \times 100\% = 93{,}46\%$$

The high percentage of detection in this system is affected by a combination of the three datasets used for the process of object recognition. There is some limitation to the system. This system can only detect one object (single-detection). Moreover, objects with the same shape or similar features can produce several estimations. The examples can be seen in Figure 6 with different percentage levels of confidence (i.e. bed: 52% confidence, table: 32% confidence). The current system cannot be used for real-time conditions and is still context-based. Hence, if there are a form and features that are not common to certain objects, the system will be difficult to recognize those objects.

Table 2 represents the results of the evaluation by involving 107 images for object detection. The number of images represents the number of sampling images in each object. Correct is the number of times that the object is successfully and properly identified. Max is the percentage of the highest confidence ever achieved by an object. Meanwhile, min is the percentage with the lowest confidence ever achieved by an object. The mean is the average confidence for the introduction of an object. Then, detect is the success percentage for an object which is successfully detected. It can be seen from Table 2 that five objects manage to get 100% detection by the system. Moreover, six objects achieve relatively high confidence level, which is more than 70%. Last, only three objects have low confidence level (glass, watch, and bed).

Figure 6 Objects with the Same Shape or Similar Features

Table 2 The Result Summary of Image Detection

	Images	Correct	Incorrect	Max	Min	Mean	Detect
Table	15	15	0	89	30	75	100%
Chair	16	16	0	88	35	78	100%
Monitor	14	14	0	91	32	77	100%
PC	16	16	0	91	42	85	100%
Lamp	4	2	2	85	52	83	50%
Bottle	16	16	0	89	62	77	100%
Glass	14	12	2	96	41	64	85,7143%
Watch	5	3	2	70	20	58	60%
Bed	7	6	1	72	54	68	85,7143%

IV. CONCLUSIONS

This research presents an initial development for object recognition as the first step to build a tool to help people with low vision and blindness. The results show that the tool can achieve a high percentage of detection (93,46%). This remarkable precision of the system is affected by the combination of three datasets that are used for the process of object recognition. Nevertheless, this approach makes the system runs slow. Hence, it is almost impossible to be implemented as a real-time system for now. In addition, this system only able to detect one object (single-detection). Objects with almost identical figure and the feature will produce several estimation results. For example, Figure 6 have different levels of confidence percentage. Moreover, the current system is still limited to the context-based problem. Therefore, if there are a figure and feature that is not common on a certain object, the system will be difficult to recognize. For future works, the development towards scene recognition system will be interesting to complete this project. Moreover, a prediction in real time will be the focus in the next step of research.

REFERENCES

Anjum, S. (2012). *Place recognition for indoor blind navigation*. Retrieved February 23rd, 2017 from http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mjs/ftp/thesis-program/2011/theses/qatar-anjum.pdf

Bujacz, M., & Strumillo, P. (2006). Stereophonic representation of virtual 3D scenes-a simulated mobility aid for the blind. *New Trends in Audio and Video*, *1*, 157-162.

CALTECH. (2017). *Caltech256: Image datasets*. Retrieved February 22nd, 2017 from http://www.vision.caltech.edu/Image_Datasets/Caltech256/images/

Carreira, J., Li, F., & Sminchisescu, C. (2012). Object recognition by sequential figure-ground ranking. *International Journal of Computer Vision*, *98*(3), 243-262.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

Chowanda, A., Blanchfield, P., Flintham, M., & Valstar, M. (2014). Erisa: Building emotionally realistic social game-agents companions. In *International Conference on Intelligent Virtual Agents* (pp. 134-143). Springer International Publishing.

Chowanda, A., Blanchfield, P., Flintham, M., & Valstar, M. (2016). Computational models of emotion, personality, and social relationships for interactions in games. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 1343-1344). International Foundation for Autonomous Agents and Multiagent Systems.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., & Hebert, M. (2009). An empirical study of context in object detection. In IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. (pp. 1271-1278).

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 20-25 June 2009* (pp. 1778-1785).

Gu, C., Lim, J. J., Arbeláez, P., & Malik, J. (2009). Recognition using regions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 20-25 June 2009* (pp. 1030-1037).

Jakarta Globe. (2012). *Indonesia has second-highest rate of blindness in world.* Retrieved February 23rd, 2017 from http://jakartaglobe.id/archive/indonesia-has-second-highest-rate-of-blindness-in-world/

Hochbaum, D. S., & Singh, V. (2009). An efficient algorithm for co-segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on 29 Sept.-2 Oct. 2009* (pp. 269-276).

Kim, W., Park, J., & Kim, C. (2010). A novel method for efficient indoor-outdoor image classification. *Journal of Signal Processing Systems*, *61*(3), 251-258.

Li, L. J., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 20-25 June 2009* (pp. 2036-2043).

Li, Y., Crandall, D. J., & Huttenlocher, D. P. (2009). Landmark classification in large-scale image collections. In *Computer vision, IEEE 12th International Conference on 29 Sept.-2 Oct. 2009* (pp. 1957-1964).

Manduchi, R., & Coughlan, J. (2012). (Computer) vision without sight. *Communications of the ACM*, *55*(1), 96-104.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145-175.

PERTUNI. (2009). *Resolusi Munas VII PERTUNI 2009*. Retrieved November 23rd, 2016 from http://pertuni.idp-europe.org/Resolusi2009/

Philip, B., & Updike, P. (2001). *California Institute of Technology SURF project for summer*. Retrieved February 23rd, 2017 from http://www.vision.caltech.edu/html-files/archive.html

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput Biol*, *4*(1), e27.

Pinto, N., Barhomi, Y., Cox, D. D., & DiCarlo, J. J. (2011). Comparing state-of-the-art visual features on invariant object recognition tasks. In *2011 IEEE workshop on Applications of Computer Vision (WACV)* (pp. 463-470).

Ran, L., Helal, S., & Moore, S. (2004). Drishti: an integrated indoor/outdoor blind navigation system and service. In *Pervasive Computing and Communications, 2004. PerCom 2004. Proceedings of the Second IEEE Annual Conference on 17-17 March 2004* (pp. 23-30).

Sutoyo, R., Prayoga, B., Suryani, D., & Shodiq, M. (2015). The implementation of hand detection and recognition to help presentation processes. *Procedia Computer Science, 59*, 550-558.

Sutoyo, R., Harefa, J., & Chowanda, A. (2016). Unlock screen application design using face expression on android smartphone. In *MATEC Web of Conferences* (Vol. 54). EDP Sciences.

Sutoyo, R., Lesmana, T. F., & Susanto, E. (2017). KINECTATION (Kinect for Presentation): Control presentation with interactive board and record presentation with live capture tools. *Journal of Physics: Conference Series, 8*(1), 1-6.

Yao, B., Yang, X., & Zhu, S. C. (2007). Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition* (pp. 169-183). Springer Berlin Heidelberg.