Face RGB-D Data Acquisition System Architecture for 3D Face Identification Technology

Aldi Bayu Kreshnanda Ismail
Dept. of Informatics and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
aldibayu@ce.student.pens.ac.id

Ihsan Fikri Abdurahman Muharram
Dept. of Informatics and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
ihsan@ce.student.pens.ac.id

Adnan Rachmat Anom Besari
Dept. of Informatics and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
anom@pens.ac.id

Dadet Pramadihanto
Dept. of Informatics and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
dadet@pens.ac.id

***Abstract*–The three-dimensional approach in face identification technology had gained prominent significance as the state-of-the-art breakthrough due to its ability to address the currently developing issues of identification technology (illumination, deformation and pose variance). Consequently, this trend is also followed by rapid development of the three-dimensional face identification architectures in which some of them, namely Microsoft Kinect and Intel RealSense, have become somewhat today's standard because of its popularity. However, these architectures may not be the most accessible to all due to its limited customisation nature being a commercial product. This research aims to propose an architecture as an alternative to the pre-existing ones which allows user to fully customise the RGB-D data by involving open source components, and serving as a less power demanding architecture. The architecture integrates Microsoft LifeCam and Structure Sensor as the input components and other open source libraries which are OpenCV and Point Cloud Library (PCL). The result shows that the proposed architecture can successfully perform the intended tasks such as extracting face RGB-D data and selecting out region of interest in the face area.**

***Keywords*–three-dimensional approach, face identification technology, RGB-D data, RGB-D architecture**

I. INTRODUCTION

As three-dimensional face identification technology has been adapted into many applications, there is a high demand of a computationally economic and free-to-customise architecture. Although there are various pre-existing architectures in this particular field, fulfilling the aforementioned criteria is not the case for most of them. High computational cost, and limitation for the developer to access and to experiment with all of the available tools and functions within the architecture are the most common premises which could be seen in most of those architectures. Therefore, to allow more freedom and capability in the development of three-dimensional face identification system, an alternative architecture is needed. This architecture should meet the aforementioned criteria, which is to have a properly optimised performance, while at the same time to remain as customisable as possible to the developer. Therefore, by lowering the computational requirement it will allow more developers to take part in the three-dimensional face technology development. Thus, this paper attempted to provide this alternative by building one of its foundations, which is a data acquisition system. This data acquisition system will rely on open source resources, including the tools used and the other components utilised in the system.

II. RELATED WORKS

At its early stage of development, the three-dimensional approach is quite common to be adapted as the standard for many face identification architectures. Zhang et al [1], for example, attempted to propose a better approach to identify facial expression and its variations based on two-dimensional face image. Dadet et al [2], on the other hand, attempted to promote a hypothetically more effective method in estimating facial landmarks which is also applied to two-dimensional face identification system. However, as the requirements and the challenges in the society getting more complicated and demanding, the two-dimensional approach in face identification system have become irrelevant and obsolete. Ruiz and Illingworth [3] revealed that the facial expression change which occurs in face, is actually a challenge for face identification system since it will introduce deformation in the face and will make the system performance suffer. Zhu et al [4] pointed out that besides facial expression, the variety of head pose will also affect the overall face identification system performance because most identification systems use frontal face as the standard identification reference. Moskovich and Osadchy [5] in their research defined the variety of illumination as one of major issues in face identification. Therefore, due to its capability to address the mentioned issues, three-dimensional approach has gained its popularity and have been featured in many researches and regarded as a somewhat today's standard for face identification technology, resulting many researches and papers revolved around three-dimensional approach as the topic of interest [6].

Fig. 1. (a) Structure Sensor [13]; (b) Microsoft LifeCam [14]

Three-dimensional approach in face identification system requires a proper architecture to obtain RGB-D data. RGB-D data is basically a recent development in three-dimensional

approach which is a combination of colour information, commonly in RGB (Red, Green, Blue) format, and depth or three-dimensional information of the face (i.e. the XYZ coordinate of face points). The combination of these data will introduce more flexibility to the system, for example, less sensitive to the illumination variations since the depth data is unaffected by the illumination change. Some of the most commonly used architectures used in recent researches are Intel RealSense as shown in [7][8][9] and Microsoft Kinect as shown in [10][11][12]. These architectures, while being used often, may not be the most accessible architectures out there since it limits user to do complete customisation on its data and its functions due to its nature of being a commercial product. Also, these architectures demand a relatively high computational resource to run properly.

This research aims to propose a state-of-the-art architecture of obtaining RGB-D data of the face for three-dimensional (3D) face identification system. The proposed architecture also aims to be customisable to the user by producing RGB-D data of the face which can be adapted for many purposes at will, while remaining to be accessible to all systems with a reasonable demand for computational resource. This architecture is built upon the principal of open source product by involving open source libraries and features so that user may fully optimise all of its functions. The proposed architecture will include RGB-D data acquisition custom hardware, data structure and also methods. It is expected to be an alternative to the pre-existing architectures and contribute to the development of three-dimensional face identification technology.

The architecture of the RGB-D face data acquisition overall system is constituted from three main components which are the input component, the processing component and the output component. Each component has a distinct role which indicated by its respective name. Figure 2 shows the architecture diagram and the relationship between elements.

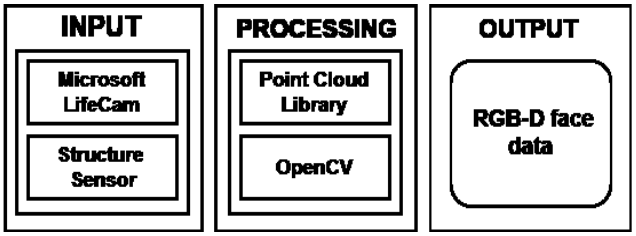


Fig. 2. Architecture diagram

The input component involves the hardware used to acquire RGB-D data of the face, which are Microsoft LifeCam and Structure Sensor. The processing component involves the libraries and functions which are used to process the RGB-D data. The output component is the representation of RGB-D of face data which is the product of this architecture.

III. HARDWARE ARCHITECTURE

The input component used in this research is a custom built RGB-D data acquisition device which is a product of integration between Microsoft LifeCam and Structure Sensor. Custom built means that such integration has never been attempted to do before. Both of these devices are selected due to their customisable nature which make its functions and features available for user to modify optimally [15]. Microsoft LifeCam is utilised to obtain the colour information from the face in RGB (Red, Green, Blue) format. On the other hand, Structure Sensor is used to obtain the three-dimensional or depth information of the captured face. In this system, both devices will be integrated and operated simultaneously so that the RGB and the depth data of the same scene could be obtained. It is important to make sure that system really does so, since a difference in the capture time between both devices will cause a mismatch in the RGB and Depth data integration process afterward, resulting in an invalid RGB-D data.

Consequently, since both devices have different perspective, it is also required to align the perspective of these devices so that they will be able to capture the exact same scene at the same time. In doing so, a customised acrylic body as shown in Figure 3 is designed to attach the two devices together and lock them in one position. This body is required to fixate the perspective position of both devices so that the following alignment process could be done easier. It can also be seen that Microsoft LifeCam is placed on the top-centre of the Structure Sensor intentionally because it is expected that the perspectives of both devices will be fixated on the centre of the body. The finalised hardware then attached onto a tripod for a stabilisation

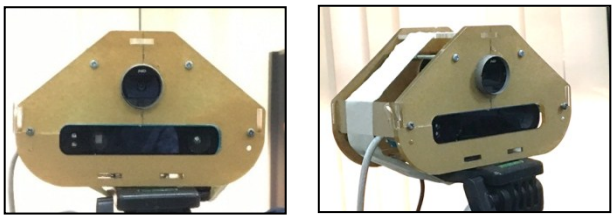


Fig. 3. Front and side view of finalised hardware

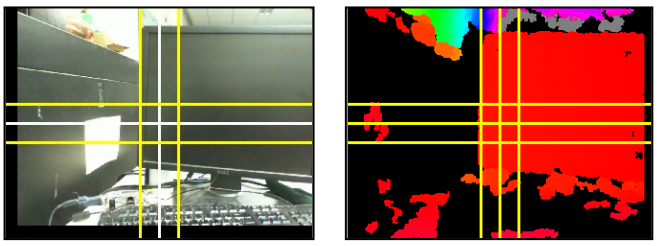
IV. CALIBRATION METHOD

Before any processing could be done, a calibration and transformation process must be carried out to the depth and RGB data. Due to the data similarity used in [16], a similar transformation process is also performed in this research. Although both input devices have been fixated by attaching them to an acrylic body as shown previously, still there is a slight difference of perspective between the two devices. Therefore, a further calibration toward the hardware must be carried out. The calibration in this research is exclusively defined as an attempt to make sure that the perspective of the two devices are perfectly aligned to each other. The difference of perspective merely occurs in the horizontal or x-axis of the captured RGB image from the Microsoft LifeCam. Therefore, utilising OpenCV library, the two-dimensional translation operation is be conducted on the captured RGB image from the Microsoft LifeCam to comply with the captured depth data from the Structure Sensor.

An image is fundamentally a matrix with particular dimension. In this case, the RGB image obtained from Microsoft LifeCam is a two-dimensional image hence a two -dimensional matrix. Consequently, the translation operation is a common operation which done to a matrix and in this case such operation is employed to perform translation on the RGB image so that it can be aligned to that of the Structure Sensor. This operation can be described by the following equation:

$$Mx = \begin{bmatrix} x1 \\ y1 \end{bmatrix} + \begin{bmatrix} x2 \\ y2 \end{bmatrix} \quad (1)$$

Where *Mx* is the result of translation which is a shifted RGB image, *x1 & y1* are the horizontal dan vertical position of the initial image RGB respectively, and *x2 & y2* are the destined horizontal and vertical position of the RGB image. In this case, the value of *x2* and *y2* may vary depending on the hardware architecture itself, which further means depends on how the input devices are configured and installed in the architecture. As for this, the degree of difference between the perspective of both devices will be different for each architecture. After the translation operation has been performed, the result of the translation must be examined thoroughly. A simple method is used to check the result of this translation operation. Identical grids are drawn to overlap the scenery obtained from both devices as a reference to identify whether or not the perspective between the devices differ to one another. Scenery which consists of many linear or rectangular objects (e.g. a box with sharp linear edges) should be chosen since it is easier to identify where the overlapping happens. If the grids on both perspectives overlap at the same particular part of the scenery, it means that the two perspectives have been successfully aligned to one another. Figure 4 shows the captured scenery from the Microsoft LifeCam and Structure Sensor after the translation operation which indicated that it has been successfully performed.



(a) (b)

Fig. 4. Side-by-side perspective comparison: (a) The perspective from Microsoft LifeCam after translation; (b) The perspective from Structure Sensor.

V. RGB-D DATA STRUCTURE

After the perspective of the two devices has been aligned successfully, the hardware now is ready to be implemented in performing the RGB-D data acquisition. However, a proper data structure that can handle the RGB-D data must be established first. In the proposed architecture, the RGB-D data of the face is stored in the point cloud. Point cloud is a data structure which is quite popular among three-dimensional technology researchers, due to its great functionality in terms of representing three-dimensional information of an object [17].

Point Cloud Library (PCL), a novel and powerful open source library to establish point cloud data structure in computing, is utilised within this architecture. Essentially. the RGB-D data is obtained through integrating the colour information of the face, and the depth or three-dimensional information of the face. In this architecture, the colour information is represented by using RGB (Red, Green, Blue) format. On the other hand, the depth information is basically the three-dimensional coordinate information of all the points that represent facial surface. Therefore, the integration of both data will result in a custom data type which is RGB-D data. The structure of such data could be visualised as shown in Figure 5.

Vertex-1	X_1	Y_1	Z_1	B_1	G_1	R_1
	X_2	Y_2	Z_2	B_2	G_2	R_2
Vertex-n	X_n	Y_n	Z_n	B_n	G_n	R_n

Fig. 5. Structure of face RGB-D data in point cloud

The RGB-D data type proposed in this architecture is basically a matrix containing RGB and depth data for each vertices (points). The depth data contains the pixel coordinate in X, Y Z (3D space). Whereas, the RGB data contains the Red, Green, and Blue colour information of each pixel from the image. This data type is to be implemented in point cloud. From a single face capture using the architecture's hardware, thousands of points could be obtained at once. By employing this custom data type, the colour information (RGB) for every point could be acquired too.

VI. EXPERIMENT RESULT

The experiment in this paper is conducted to demonstrate how the proposed architecture works in reality and to see if it really fulfils the research expectation. The experiment covers the process of obtaining individual RGB and depth data with architecture hardware, until the integration process between both data resulting a proper face RGB-D data, which eventually will be the system output.

A. Face RGB-D data acquisition

This experiment began with the first step in the system workflow, which the acquisition of RGB data from a calibrated Microsoft LifeCam and depth data from Structure Sensor. The purpose of this experiment is to make sure that the hardware in this architecture could function as intended by acquiring RGB data and depth data respectively.

These data will then be integrated into one custom data type, which is RGB-D data type. The approach of integrating depth data and RGB data is inspired by a similar experiment done in [18], where the difference is instead of using stereo camera, a single RGB camera is used. This experiment is done by situating a subject in the testing environment. This environment is an environment where its environment parameters (e.g. illumination) have been predetermined, or in the other word, with pre-set parameters. A scan of a subject is taken with the architecture hardware until a valid or proper data could be obtained. The definition of valid here means that on both RGB and depth data, there is no failure in acquisition or a missing part in scanning, which are obtained is aligned in perspective. The experiment which has been conducted shows the acquisition of RGB image and depth image could be done successfully as shown in Figure 6.

(a) (b)

Fig. 6. Face data acquisition: (a) Face RGB data; (b) Face depth data.

B. Integrating RGB and depth data.

The next experiment is integrating the RGB data and depth data to create RGB-D data of face. The purpose of this experiment is to prove that the architecture can produce RGB-D data of face with only the presence of RGB data and the depth data. Using the calibration method which has been explained previously, the difference of perspectives in this architecture is calculated and translation operation is performed with the following equation:

$$Mx = \begin{bmatrix} x1 \\ y1 \end{bmatrix} + \begin{bmatrix} 20 \\ 13 \end{bmatrix} \quad (2)$$

It can be observed from the equation that, in this architecture, the RGB image needs to be translated 20 pixels to the right and 13 pixels upward. The result from this image is a shifted image as shown in Fig 7.

Fig. 7. Translation operation: (a) Initial RGB image; (b) Shifted RGB image.

After the calibration has been conducted and a shifted RGB image has been acquired, this next phase in this architecture is joining this RGB data and the depth data from the Structure Sensor.

The integration process is done by merging the two data into a new data type which will be stored in the point cloud. Figure 8 shows the visualised RGB-D data which has been integrated successfully. It can be observed that the RGB-D data resembling the output from Structure Sensor, only this time there is a colour information for each point which is extracted from the RGB image. It is also possible to perform three-dimensional transformation to this data, as the example in this experiment, the data is rotated to resemble a side view. This result showed that the integration process between the RGB and the depth data using the custom hardware used in this architecture could be achieved and well implemented.

(a) (b)

Fig. 8. RGB-D integration: (a) Face RGB data; (b) Face RGB data (rotated)

C. Face detection and local area segmentation

The last experiment in this research is to do a detection and segmentation operation which is applied toward the face RGB-D data which have been carried out previously so that the local face area from the face could be selected out and the rest of the data could be eliminated. The face detection and segmentation are performed by making use of the functions that comes within OpenCV and also Point Cloud library. The face detection method used in this system employs OpenCV HAAR Classifier library [19] whereas the segmentation process done in this experiment is facilitated through the Euclidean Cluster Extraction method which is come equipped with the Point Cloud Library [20]. The purpose of this experiment is to show that the face RGB-D data which is acquired through the proposed architecture is flexible. It means that this data is not only limited to certain procedure to be succeed, but it is possible to implemented in other various method as it is one of the purposes of this research. Figure 9 shows that the local area of face could be selected out from the rest of the RGB-D data which is unnecessary for the processing and may increase the burden of computation since it makes the most part of the RGB-D data.

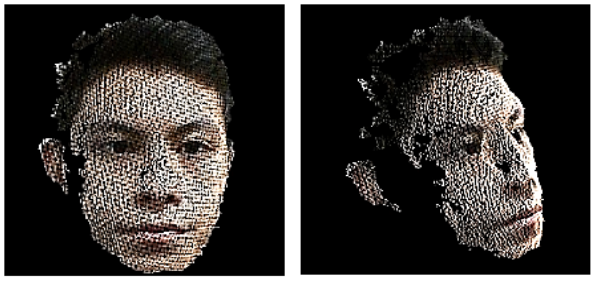


Fig. 9. The localised face RGB-D data in multiple viewpoints

VII. PERFOMANCE EVALUATION

After successfully conducted the experiments, the performance of the algorithms featured in the system must also be evaluated. Since it is important to see how each algorithm could handle a different kind of dataset, and whether it could still maintain its performance, thus, the performance evaluation is carried out to see the performance consistency of each algorithm despite variation within the dataset.

The evaluation is done by employing specific variety of datasets from several subjects to be processed by the system as the input, and then evaluated. These subjects are consisted of male and female participants to introduce variation into the datasets. The list of subjects used in the evaluation could be seen in Table 1.

TABLE. 1. LIST OF SUBJECTS FOR EVALUATION

Data	RGB Data	Depth Data	Localised Face (RGBD)
A			
B			

Data	RGB Data	Depth Data	Localised Face (RGBD)
C			
D			

The RGB image and depth image of the subjects in neutral expressions are taken. The each of these data are consequently being processed through each algorithm until the localised face of each subject could be obtained. Using the same system which is used to perform the previous experiment, now the processing time of each step is being measured. The system specification is shown in Table 2.

TABLE. 2. SYSTEM SPECIFICATION FOR EVALUATION

Hardware	Specification
Processor	Intel Xeon® CPU E3-1225 V2 @ 3.20GHz x 4
RAM	6 GB
VGA	NVIDIA GF108GL (Quadro 600) x64 clock 33Mhz 25.6 GB/s DDR3

The system performance of each dataset as seen in Table 3 is varied depending on the dataset. However, it must be noted that the number of point cloud being processed are also quite different. Therefore, it is important to consider that the amount of point cloud processed will affect directly upon the length of the processing time.

TABLE. 3. TIME ELAPSED FOR EACH ALGORITHM

Data	Capture RGBD	RGBD Integration	Face Segmentation	Point Cloud
A	25ms	1046ms	5395ms	14577pts
B	27ms	997ms	6995ms	20320pts
C	25ms	1013ms	5362ms	15283pts
D	27ms	1044ms	7505ms	21348pts

The result further revealed that the time to capture RGBD images is consistent with an average of 26ms processing time.

VIII. CONCLUSION

In this paper, the RGB-D data acquisition of the face has been presented. This architecture has been proven to be functioning as intended, which is to acquire the RGB-D data from the face. The proposed architecture also exhibited a good integration between the Microsoft LifeCam and Structure Sensor as the custom hardware used to obtain the input. The integration of open source libraries which are OpenCV and Point Cloud Library also enable the architecture to be modified to meet various preferences since there is no limitation to its features and functions. More importantly, the proposed architecture has also proven that a proper RGB-D data type could be well established and eventually resulting a proper face RGB-D data.

According to the samples, the average time taken to capture to RGB-D data is 26ms, the average time to perform RGBD integration is 1045ms, the average time to perform face segmentation is 4815ms, and the average amount of point cloud obtained and processed in each data acquisition are 17817 points. It shows that the performance of each algorithm is still acceptable despite still needing improvement. One way to do this is by performing downsampling on the RGBD data to reduce the amount of point cloud which could enhance the process speed.

Due to its flexibility and uniqueness, the proposed architecture is expected to be present as an alternative to three-dimensional face identification technology as a stand-alone system, or as a module which could be integrated with other pre-existing 3D face identification architectures.

IX. FUTURE WORKS

In the following research, we will build general face RGB-D database which will be used to perform various task, including a more advanced testing scenario and identification system, to be equipped within the proposed architecture. This database will include various face RGB-D data including face with expression variance, and pose variance and can be used for various kind of experiments. The overall performance of the system is also expected to be improved.

ACKNOWLEDGMENT

We would like to thank the Ministry of Research, Technology and Higher Education of Indonesia and EEPIS Robotics Research Centre (E2RC) of Electronics Engineering Polytechnique Institute of Surabaya, for providing the funding and facilities for this research so that this research could be realised and put into implementation.

REFERENCES

[1] L. Zhang, N. Ye, E.M. Marroquin and T. Sim, "Expressive deformation profiles for cross expression face recognition," *in 2012 21st Internationals Conference on Pattern Recognition (ICPR)*, Tsukuba, Japan, 11-15 November 2012, pp. 1582– 1585.

[2] D. Pramadihanto, H. Wu, M. Yachida, "Invariant face recognition by Gabor wavelets and neural network matching" *in 1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems* Vol. 9 (2017) Vol 1 (1996), pages: 59-63.

[3] M.C. Ruiz and J. Illingworth, "Expression classification of 3D faces using local deformations," *in IET Conference on Image Processing (IPR 2012),* London, UK, 3-4 August 2012, pp. 1-6.

[4] X. Zhu, Z. Lei, J. Yan, D. Yi and S. Z. Li, "High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild," *in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachussets, USA, 7-12 June 2015, pp. 797-796.

[5] B. Moskovich and M. Osadchy, "Illumination Invariant epresentation for privacy preserving face identification," *in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops,* San Francisco, CA, USA, 13-18 June 2010, pp. 154-161.

[6] D. Kim, M. Hernandez, J. Choi and G. Medioni, "Deep 3D face identification," *in 2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, 1-2 October 2017, pp. 133 - 142

[7] Zhang, Z. Lu, W. Li and Q. Liao, "A robust and fast 3D face reconstruction method using realsense camera," in 2017

International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22-24 March 2017, pp. 2691-2695.

[8] J. V. Patil and P. Bailke, "Real time facial expression recognition using RealSense camera and ANN," *in 2016 International Conference on Inventive Computation Technologies (ICICT),* Combaitore, India, 26-27 August 2016, pp. 1-6.

[9] C. Chih, Y. Wan, Y. Hsu and L. Chen, "Interactive sticker system with Intel RealSense," *in 2017 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 8-10 January 2017, pp. 174-175.

[10] R. Siv, I. Ardiyanto, and R. Hartanto, "3D human face reconstruction using depth sensor of Kinect 2," *in 2018 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 6-7 March 2018, pp. 355-359.

[11] Z. Cheng, T. Shi, W. Cui, Y. Dong and X. Fang," 3D face recognition based on kinect depth data," *in 2017 4th International Conference on Systems and Informatics (ICSAI),* Hangzhou, China, 11-13 November 2017, pp.555-559.

[12] N. N. Kaashi and R. Safabakhsh, "3D constrained local model-based face recognition using Kinect under variant conditions," *in 2014 7th International Symposium on Telecommunications (IST),*

[13] H. Lee, J. Seo and H. Jo, "Gaze tracking system using structure sensor & zoom camera, " in 2015 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, South Korea, October 2015, pp. pp. 830-832.

[14] Lievendag, N. 2017, "Structure Sensor 3D Scanner Review", www.3dscanexpert.com/structure-sensor-review-part-1. Accessed in: 26 May 2018.

[15] K. Kadir, M.K Kamarudin, H. Nasir, S.I. Safie and Z.A.K. Bakti, "A comparative study between LBP and Haar-like features for Face Detection using OpenCV," in 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T), 2014, pp. 335-339.

[16] A. Alfarouq, R. Sanggar Dewanto, D. Pramadihanto, "Transformed Stereo Vision and Structure Sensor for Development 3D Mapping on "FLoW" Humanoid Robot in Real Time" *in 2017Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 9 (2017) No. 2-5, pages: 129-133.

[17] W. Li, X. Ren, F. Li and W. Wang, "Study on the method of high-precision vehicle-borne lidar point clouds data acquisition in existing railway survey," in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-28 July 2017, pp. 1704 – 1707.

[18] D. Pramadihanto, A. Alfarouq, S. A. Waskitho, S. Sukaridotho "Merging of Depth Image Between Stereo Camera and Structure Sensor on Robot FLoW Vision" *in 2017 IEEE International Journal on Advanced Science Engineering Information Technology*, Vol. 7 (2017) No. 3, pages: 1014-1025.

[19] L. Cuimei, Q. Zhiliang, J. Nan and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, 20-22 October 2017, Yangzhou, China, pp. 483-487.

[20] R. B. Rusu, "Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments". In *KI - Künstliche Intelligenz,* Vol 24 (2010) No. 4, pp. 345-348.