Transformed Stereo Vision and Structure Sensor for Development 3D Mapping on "FLoW" Humanoid Robot in Real Time

Ardiansyah Al Farouq[1,2], Raden Sanggar Dewanto[1,2], Dadet Pramadihanto[1,2]
[1]*Electronic Engineering Polytechnic Institute of Surabaya, Indonesia*.
[2]*EEPIS Robotics Research Center, Indonesia*.
alfarouq01@gmail.com

***Abstract*—In this paper, we proposed a method for building 3D Mapping environment in real time. The purpose of this model is to be able to approach the human ability to quickly know the environment. The problem of building a 3D mapping in real time is dependent on the efficiency of the algorithm to transform 2D images into depth data forms, which then transformed into a 3D model. This paper shows a method of transformation which has built an efficient algorithm because it can complete the entire sequential algorithm in real time. The algorithm has successfully run the whole system that only has a depth pixel error average of 18,10% and an average error of the system running in real time.**

***Index Terms*—3D Mapping; Stereo Matching; Robot Automation; Real Time System.**

I. INTRODUCTION

There has been an insufficient development in the research of humanoid robot in building a system that can approach human capabilities. This insufficiency ranges from building mechanical humanoid robot to building intelligence systems. One huge problem is how to build a system that can quickly recognize the environment. The system is said to be able to know if the environment can have complete information based on anything that can be seen so that the robot can move around autonomously. The ability to simultaneously localize a robot and accurately map its environment is considered by many to be a key prerequisite of truly autonomous robots [1,2].

Figure 1 shows the mechanical model of the head and location between the stereo camera and structure sensor used in the proposed method. The stereo cameras used are two cameras Microsoft LifeCam HD cinema.

The objective of the research is to propose a method that builds a system on a robot to map the surrounding environment. In this case, vision sensors to obtain information are needed to build a map. Therefore, the proposed system used a camera stereo, structure sensor and mechanical robot head as its gaze orientation, and it can be run in real time. In the past decade, there have been approaches to find the orientation of the views of the mechanics used [3]. The latest research in building a 3D mapping algorithm tend to focus only on the data detailing depth [5, 6, 8, 11] and 3D shapes without taking into account the real-time accuracy [7, 9, 10].

Stereo camera used in robots can generate in-depth map and it is used to estimate the distance of the object in the robot. The calculation of the distance is the distance between the camera and the object, rather than the distance between the object and the robot as a whole. The calculation allows the robot to know the location of the objects surrounding it, be it roads impassable or obstacle. However, it still cannot represent the actual environmental conditions such as how to look at the man. There are other things to improve in the mapping of details in the environment as it is shifting the actuator's views. Actuators can help the camera to get a wider environmental information. The research conducted by EEPIS Robotic Research Center (ER2C) has built a mechanical head with the corners of the degrees of freedom humanoid [3].

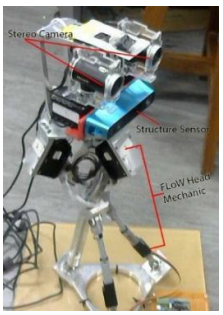

Figure 1: Mechanics of the robot head

The problem in the proposed method is the time to build a 3D mapping on the design resembles a human head that can rotate as the orientation of the view. Hence, the system is built so that it is able to transform the depth of the data model to a 3D model that follows the direction of vision of the robot end effector mechanical head. It is also influenced by the quality of the data depth is formed. The incorporation of a stereo camera and structure sensor is to increase the depth of data. The method will produce a process that can run in real time because it has an efficient algorithm. Besides being able to run in real time, the algorithm of this method has an average error rate of small pixels.

II. RELATED WORK

EEPIS Robotics Research Center (ER2C) has built Humanoid Robot named "FLoW". The "FLoW"'s head has three movement model platforms [3]: In this paper, we discuss the movement of the part of the Flow's eyes only.

The mechanism of the "FLoW"'s eyes used Helmholtz model [4] proposed to imitate the construction of a human's eyes.

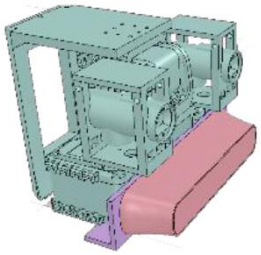

Figure 2: "FLoW" Head Vision Mechanism

III. KINEMATIC TRANSFORMED VISION

In order to solve our research problem, the block system is explained in Figure 3. The vision of "FLoW" Humanoid Robot Research has been developed, which start from stereo camera calibration, and then added to the transformation of the movement mechanism of the head. The movement mechanism in this paper is determined by yaw- axis degree of rotation in the part of the eyes part is used for the direction of the camera's view.

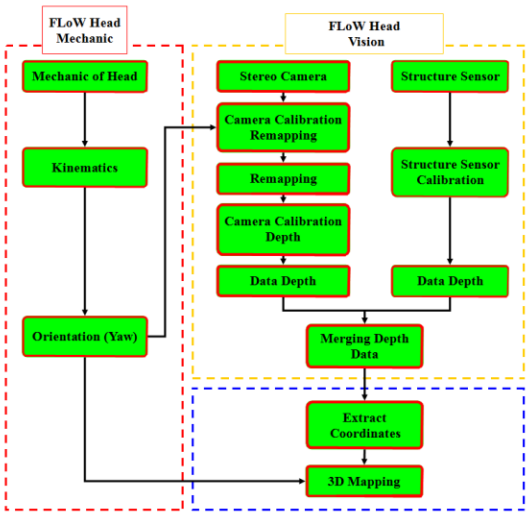

Figure 3: Block System

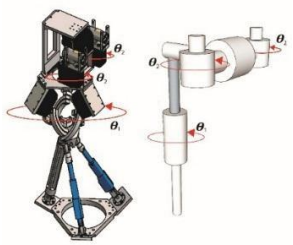

Figure 4: "FLoW" Head Mechanism with DoF

A. Acquire Orientation

The mechanical movement platform on the "FLoW" Eyes can be used to determine the camera angle. The transformation model of "FLoW" Eyes is not described in detail in this paper, because we need only a yaw-axis degree of rotation for moving "FLoW" Eyes. This movement is defined by a matrix to get the orientation between "FLoW" Eyes (θ_1) and each of the camera itself (θ_2).

$$T_M = \begin{bmatrix} cos(\theta_1+\theta_2) & 0 & sin(\theta_1+\theta_2) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(\theta_1+\theta_2) & 0 & cos(\theta_1+\theta_2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

This paper also discusses how the disparity estimation is obtained by some researchers. Joint optimization method that combines the stereo and depth sensor have been used to obtain the depth of the map. Jing et al. [7] explained that they performed fusion stereo DSLR and Kinect depth sensor to propose a method for improving the quality of the dense and depth maps. Jaesik et al. [5,6] attempted to make an upsampling and completion of the Depth Map for RGB-D Cameras.

B. Acquire Vision Data

In order to build 3D environment mapping, data were obtained from the "FLoW" Eye to get the initial orientation of the mechanics. Then, the data sensor camera were obtained from a combination of Stereo Camera and Structure Sensor to get real distance from the depth image.

a. Stereo Calibration

Stereo camera needs to be calibrated first before we get the depth image. Calibration using chessboard pattern is recommended to get the correspondence of image pairs.

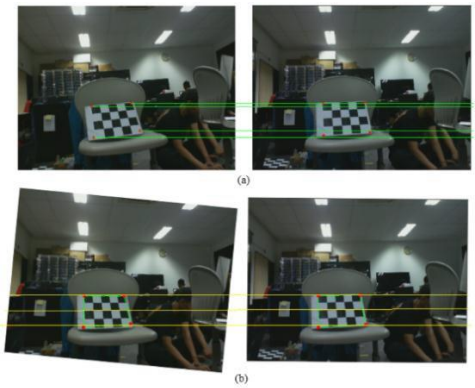

Figure 5: (a) Find Correspondence Point (b) Rotate Frame Following Other Frame

b. Disparity

We should get the disparity (d): Disparity is the difference between the left (x_l) and right (x_r) of the camera.

$$d = x_r - x_l \quad (2)$$

Before we get the depth image, we need to know the focal length, baseline (B) (distance between cameras), and disparity. The focal length (f) was obtained from the real distance (Z_{real}) per $1 + \frac{1}{m}$, where $m = \frac{image\ size}{object\ size}$.

$$f = \frac{Z_{real}}{1 + \frac{1}{m}} \quad (3)$$

c. Depth

Depth value (Z) contains the distance between the cameras to the object in a frame.

$$Z = \frac{fB}{d} \quad (4)$$

C. Merging Depth Data

This section describes the depth combination from each sensor. In this paper, Stereo camera and Structure Sensor

were used. If the depth value of the image sensor structure (S_s) is ≤ 0, then it is turned into a new matrix C_{xy}.

$$C_{x,y} = \begin{cases} 1; S_s(x,y) \leq 0 \\ 0; \end{cases} \quad (5)$$

The output of the structure sensor (S_c) is obtained and matrix C_{xy} multiplied into a new matrix N.

$$N = S_C . C \quad (6)$$

$$N = \begin{bmatrix} S_C.C_{11} & S_C.C_{21} & \cdots & S_C.C_{1x} \\ S_C.C_{21} & S_C.C_{22} & \cdots & S_C.C_{2x} \\ \vdots & \vdots & \ddots & \vdots \\ S_C.C_{y1} & S_C.C_{y2} & 0 & S_C.C_{xy} \end{bmatrix}$$

After the image depth of both sensors are known, we then stored it into a new matrix combination of both sensors (S_{SC}).

$$S_{SC} = N + S_S \quad (7)$$

Our method tries to combine the depth data from both sensors to increase the depth of detail image into the new depth data.

D. Coordinate Extraction

Extracting coordinates from the frame to real is a phase for combining the depth image and the orientation of the camera. This combination result in a 3D Vector that is used for building 3D Mapping. To extract, we have to find the maximum camera angle by the frame orientation (K_h,K_V).

$$K_{h,V} = arctan\left(\frac{(w,h)frame/2}{Z}\right) \quad (8)$$

The actual distance (Z_s) of each object can be known from the real measurement compared to Equation (4). To get the actual distance, the system to find the distance needs to be calibrated. The data need to be compared to a range scaled distance (Z_{Cn}), which has a constant value in each distance measurement, which then reduces it with a real distance (Z_r).

$$Z_S = \begin{cases} \frac{(zc_{n+1} - zc_n) * (z - zr_n)}{(zr_{n+1} - zr_n)} + Z_{Cn} & ; Z_{Cn+1} \\ 0 & ; Z > Z_{Cn} \end{cases} \quad (9)$$

After that, the real width (S_h) and height (S_V) maximum in actual distance (Z_s) are derived:

$$S_{h,v} = tan\,(K_S)\, Z_s \quad (10)$$

Then the depth image is transformed into 3D model,

$$x, y_{Sn} = 2S_{h,v}\left(\frac{x, y_{dn} - x, y_c}{w, h}\right) \quad (11)$$

where :

x,y_{dn} = pixel in depth image
x, y_c = pixel center in depth image
w = width frame
h = height frame

E. 3D Mapping Transformation

The data requirement were calculated to transform them into an environment mapping. This transformation requires data mechanic and data image to be transformed.

The matrix for each pixel data in image (T_s) and matrix that contains pixel data (I_{Ts}), defined by:

$$T_S = \begin{bmatrix} 1 & 0 & 0 & xs \\ 0 & 1 & 0 & ys \\ 0 & 0 & 1 & zs \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (12)$$

$$I_{Ts} = \begin{bmatrix} T_{s11} & T_{s12} & \cdots & T_{s1x} \\ T_{s21} & T_{s22} & \cdots & T_{s2x} \\ \vdots & \vdots & \ddots & \vdots \\ T_{sy1} & T_{sy2} & 0 & T_{sxy} \end{bmatrix}$$

With the transformation of this matrix, the system can build a 3D mapping of the data transformation mechanic (T_M).

$$3Dmap = T_M . I_{Ts} \quad (13)$$

IV. EXPERIMENTAL RESULT

This section discusses the performance of a system built to verify the efficiency of the detail and accuracy of real-time in the proposed method. Several experiments have been conducted to evaluate this method. The first experiment was to compare the system that we have built to systems owned by other researchers. The second experiment was to compare the efficiency of the internal system.

A. Comparison of Systems Based on Time

In this experiment, the percentage of pixels of depth error method that we developed has been calculated and compared with other merger methods. The error was determined by calculating the total pixel that does not have the depth value divided by the size of the frame on each sequential real time that has been specified. The sequential time used is for 15 seconds and during that time, the head of the robot rotates the Y axis of 0° to 126°.

Table 1
Percentage error pixels of each method

Time (second)	Orientation (Yaw)	Our method	Jaesik et al. [4, 5]	Jing et al. [6]
1	0	15.57	100	100
2	9	14.41	100	100
3	18	15.93	100	100
4	27	16.27	16.11	100
5	36	19.31	100	19.21
6	45	21.77	100	100
7	54	19.21	18.79	100
8	63	100	100	100
9	72	100	100	20.04
10	81	21.52	19.05	100
11	90	24.23	100	14.08
12	99	21.21	25.65	100
13	108	15.26	100	100
14	117	16.26	14.03	100
15	126	14.35	11.55	100
Average:		29.02	67.012	83.55533

In Table 1, there is an error that has a value of 100% error. This error was due the failure of the method to complete the algorithm to build the depth on the whole pixels. Thus, it cannot be converted into 3D transformation Map. This can be seen in Figure 8.

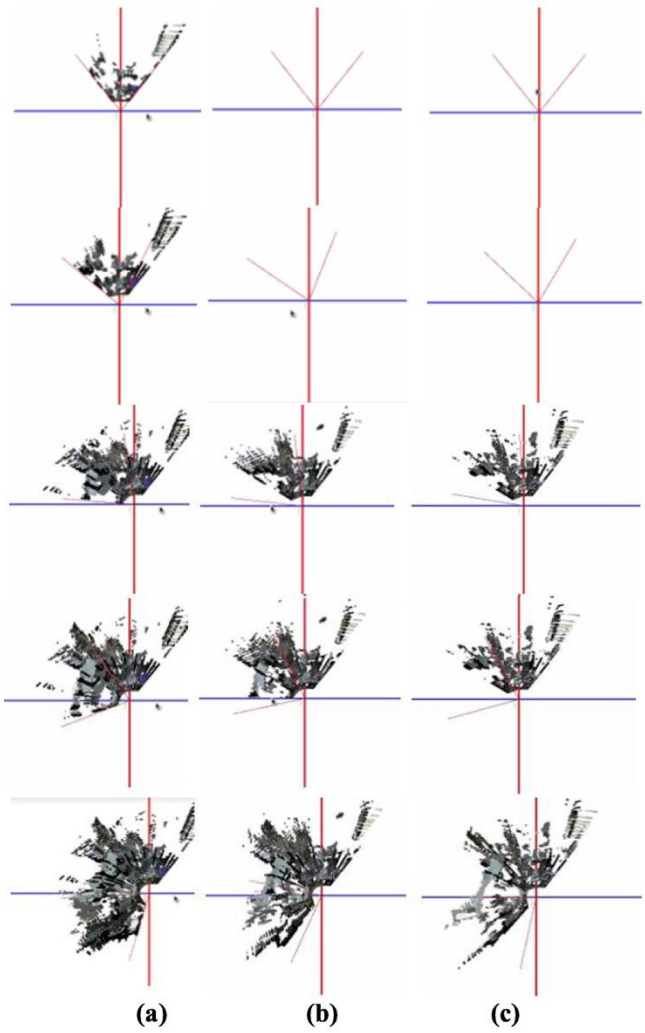

Figure 8: (a) our method; (b) Jaesik et al [5, 6] method; (c) Jing et al [7] method

Figure 8 is a form of transforming 3D mapping from the top view. In Figure 3, each row is oriented by sequential views from 0° to 126°, which run for 15 seconds. From Figure 10, it shows the differences of the detail or information density of each data obtained. Information obtained from our method has an average error of 29.02%. It shows that more detailed information than the methods used by Jaesik et al. [5,6] resulted in an average error of 67.012% and Jing et al [7] with an average error of 83.55%.

Figure 9: Panorama view 126°

In Figure 10. is a first-person view obtained by each method in Figure 1. In view of this, our system proved to have more information than the method results Jaesik et al. [5,6] and Jing et al. [7]. Thus, the shape of each object is more accurate.

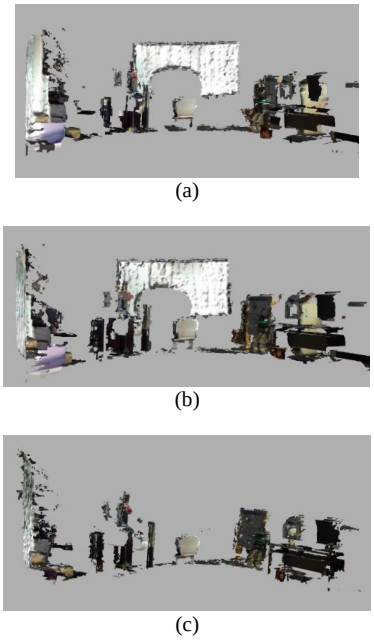

Figure 10: (a)First person view from our method; (b)First person view from Jing et al [5,6] method; (c)First person view from Jing et al [7] method

B. Comparison at the Internal System

The aim of this experiment is to determine the percentage error of data pixel depth of stereo, sensor structure, and methods of merging from both cameras. The error was determined by calculating the total pixel that does not have the depth value divided by the size of the frame on each sequential predetermined orientation. Sequential orientation used is from 0° to 126°.

Table 2
Percentage error pixels of each phase

Orientation (yaw)	Stereo Camera	Structure Sensor	Methods of Merging
0	68.91	22.05	15.08
9	66.10	20.02	14.67
18	62.85	24.47	18.42
27	62.12	24.46	17.86
36	59.80	27.33	17.98
45	59.08	32.30	21.35
54	61.77	32.15	19.41
63	63.55	31.97	19.19
72	67.13	28.73	17.58
81	75.40	28.17	21.66
90	73.84	22.61	20.55
99	69.13	31.22	24.26
108	63.10	30.06	16.85
117	60.40	25.74	15.55
126	54.39	19.64	10.47
Average	64.50	26.72	18.10

Table 2 shows that the error pixel data depth in the methods of system merger between the stereo camera and the structure of sensors has an average error, which is smaller at 18.10%, compared with a stereo camera that has an average error of 64.50% and structure sensor that has an average error of 26.72%.

V. CONCLUSIONS

We proposed a model system that incorporates stereo camera, sensor structure and orientation derived from kinematics end effector at a robot head. “Flow” Humanoid Robots have been able to demonstrate the detailed depth

information and depth data density as information in building a 3D map of environment. Besides, the system can obtain detailed information. Our main contribution is the method of the system built must be able to run in real time. This capability is essential as a system of intelligence in humanoid robotic system should at least approach the human ability to think quickly to find the mapping environment. This is proven, in which the proposed system showed a smaller average error than the other methods, that is 29.02%. This system has been running in real time. The internal system is also great because an average error of 18:10% was obtained by sequential orientation. In the future we will develop this research to add the orientation at the mechanism of mechanical head "FLoW" humanoid more dynamic, in order to identify the overall environment more detail and faster. We will also transform at the system CPU usage models that system is not very high at run time.

REFERENCES

[1] G. Dissanayake, P. Newman, S. Clark, H.F. Durrant-Whyte, and M. Csorba.," A solution to the simultaneous localization and map building (SLAM) problem", IEEE Transactions of Robotics and Automation, (2001).

[2] M. Michael, T. Sebastian, K. Daphne, W. Ben., "FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem", In Proceedings of the AAAI National Conference on Artificial Intelligence, (2002).

[3] Bayu Setiawan, Oxsy Giandi, Dadet Pramadihanto, Raden Sanggar, Dewanto, Sritrusta Sukaridhoto, and Ahmad Subhan Khaillulah., "Flow head: 7 dof mechanism for flow humanoid", In Control, Electronics, Renewable Energy and Communications (ICCEREC), International Conference on, pages 98–102. IEEE, (2015).

[4] Hermann Von Helmholtz. Handbuch der physiologischen Optik, volume 9. Voss, 1867.

[5] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and In So Kweon. "High-quality depth map upsampling and completion for rgb-d cameras". IEEE Transactions on Image Processing, 23(12):5559–5572, (2014).

[6] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. "High quality depth map upsampling for 3d-tof cameras". In 2011 International Conference on Computer Vision, pages 1623–1630. IEEE, (2011).

[7] Jing Liu, Chunpeng Li, Xuefeng Fan, and Zhaoqi Wang. "Reliable fusion of stereo matching and depth sensor for high quality dense depth maps". Sensors, 15(8):20894–20924, (2015).

[8] Zhu, J.; Wang, L.; Yang, R.; Davis, J.E.; Pan, Z. "Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps". IEEE Trans. Pattern Anal. Mach. Intell. (2011), 33, 1400–1414.

[9] Zhang, S.; Wang, C.; Chan, S. A "New High Resolution Depth Map Estimation System Using Stereo Vision and Depth Sensing Device". In Proceedings of the IEEE 9th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, Malaysia, (8–10 March 2013); pp. 49–53.

[10] Wang, Y.; Jia, Y. A "fusion framework of stereo vision and Kinect for high-quality dense depth maps". Comput. Vis. (2013), 7729, 109–120.

[11] Yang, Q.; Yang, R.; Davis, J.; Nistér, D. "Spatial-depth super resolution for range images". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, (17–22 June 2007).