Synchronization and calibration of a stereo vision system

Xin Chen, Xiangfei Wu, Shiqi Gao, Xiaomei Xie
Institute of Astronautics & Aeronautics
University of Electronic Science & Technology of China
Chengdu, China

Ya Huang*
School of Mechanical & Design Engineering
University of Portsmouth
Portsmouth, UK
ya.huang@port.ac.uk

***Abstract*—In this paper, we design a stereo image and video acquisition system with adjustable baseline and relative angle. Stereo vision has been used both for field robotic navigation and 3D wave reconstruction. However, there has been inadequate documentation about the synchronization performance of such system. For a dynamically moving scene, such as the wave, any delay between the two cameras will affects the accuracy of the disparity map. The present study presents a physical process to measure the delay between the two cameras. Firstly, we obtain the maximum frame rate of camera when it is triggered by external signal. Then we adopt a free fall experiment to measure the delay between trigger intervals at maximum frame rate. The relationship between baseline distance, calibration checker box tile size and target distance is discussed. Two calibration tools, i.e. MATLAB and WASS are compared using a range of distances.**

Keywords—stereo vision, camera synchronization, stereo calibration

I. INTRODUCTION

Reconstruction of 3-dimensional (3D) maps of dynamically moving environment is crucial for the design of a robot's guidance, navigation and control system (GNC). Seascape is by far the most challenging context for such task. The seakeeping and ride quality of a surface vehicle are dominated by the understanding of the dynamic wave prorogation direction and speed under the influence of tidal currents, rapids (due to local geography of water bed) and windage. X-band radar system is able to construct detailed surface wave 3D map but the cost and complexity of such system often prohibits its application on a budget and agile mobile system [1]. Vision, especially stereo, based approach remains one of the most effective to recover depth map of the seascape at close proximity of up to 40 meters [2]. Many studies published results of stereo systems working on fixed structures and on moving platforms i.e. surface vehicles with fused inertial measurement units (IMU) [3] or without [4], but rarely there has been adequate information to reproduce and validate each setup and configuration step by step. In particular, through the configuration and calibration processes for stereo cameras, one needs to consider their lighting at the scene, different scene photomatrics, depth measuring range, and the delay between the camera pair.

For relatively static scene, the depth can be estimated even if the image pair is captured at different times. However, the vehicles are not usually stable. When the scene changes at high speed, the salient points could move dozens of pixels between the two images captured. This prevents the stereo image calibration process to stablish the correspondence between the pair of images. Synchronization between the two captured images is crucial for the accuracy in the disparity and depth estimation. A key validation step for all stereo system used for dynamic scene is to physically quantify the delay between the two cameras due to any electronic hardware and underlying software. One cost-effective way of measuring the delay between two nominally 'synchronized' cameras is to use the camera pair to capture an object moving at a known velocity. With the help of gravity and alignment tools, it is possible to level the baseline of the camera pair perpendicular to a vertically dropping ball and measure the vertical displacement of the ball between two timed instances. By comparing the two nominally synced cameras their vertical differences in pixel at a set interval in time, one can effectively measure the delay between the two cameras. With a drop height of less than 2 meters and steel ball of negligible aerodynamic resistance, a constant gravitational acceleration at sea level can be assumed to derive instantaneous velocity. The study intends to compare the real-life delay between the two cameras.

The combination of baseline distance, tile size of the calibration checkerboard, and target distance impact on the accuracy of the intrinsic and extrinsic parameters of the camera pair. The study will investigate these combinations in relation to the actual measuring range from 5 to 40 m.

In this paper, we present a stereo vision acquisition system with adjustable-baseline and camera relative pose to capture indoor and outdoor dynamic scenes. The two cameras are triggered by external trigger signal generated by STM32 (a microcontroller programmed to generate pulse signals), which produces the maximum acquisition rate of the camera. A free fall ball experiment is designed to measure delays between the two cameras. Factors such as baseline and target distance which may affect calibration accuracy are investigated. The study comprises the following objectives: 1) design a stereo vision acquisition system with adjustable-baseline and camera pose to according to the scene. 2) Quantify physical delays between the two synced cameras at a sampling rate of 100Hz using a free fall test. 3) Establish the relationship between checkerboard tile size, baseline distance, and measuring distance.

The remaining part of the paper proceeds as follows: Section II is an overview of related works. Section III shows our system for the synchronization test and stereo calibration. Section IV discusses experimental results. Section V concludes.

* Correspondence author: Ya Huang (ya.huang@poet.ac.uk)

II. Related Work

The research on synchronization validation focuses on accurate acquisition of time stamps and its impact on the standard stereo calibration process. A stereo system with adjustable baseline is developed to acquire images indoors and outdoors [5]. The design uses external trigger pulse generated by a Raspberry Pi microprocessor at 20Hz giving rise to a delay between the two acquisition channels in microseconds. The fixed low frequency external trigger pulse may have caused delays, resulting in correspondences of some key points lost between consecutive frames.

The maximum synchronised frame rate governed by the external trigger can be measured by acquiring a video sequence on a display [6]. The frequency of external trigger signal needs to be kept the same as the video frame rate. By gradually increasing this frequency until the images captured by the left and right cameras become different, it is possible to identify the maximum frequency.

Baseline distance of stereo cameras can affect 3D projection accuracy of target object at different distances [7]. However, reported results were drawn from a short range of target distance within 700mm with the checkerboard tile size unchanged. In fact, the baseline distance, tile size and calibration distance could all affect the intrinsic and extrinsic parameters of the stereo system. The present study intends to investigate these variables.

III. Implementation

In this section, we introduce our hardware design with adjustable-baseline and camera pose. The camera data cables are directly connected to the computer, and the trigger cables are connected to the STM32 via a PWM unit. The acquisition interface is programmed to be multithread with the captured images saved to the computer hard disk. Fig. 1 shows the structure of acquisition system, the STM32 module generates the PWM trigger signal and the corresponding I/O pin connects to the GPIO port of both cameras. The computer runs multithread image acquisition programmes and stores the acquired images.

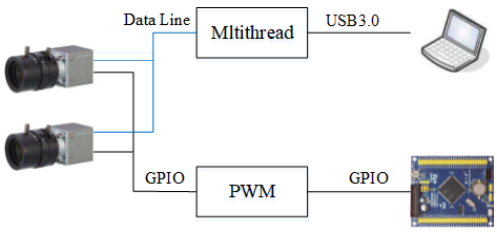

Fig 1 The stereo acquisition hardware

A. Image Acquisiton System Design

Two VCXU_23C Baumer industrial cameras with adjustable baseline and pose are installed on a bespoke rail strut (Fig 2). The STM32 development board is mounted on top of the camera but does not interfere with the adjustment. It generates the external trigger signal. The camera sensor is Sony IMX174 with a resolution 1920×1200 pixels and pixel size 5.86×5.86 µm. For our experiment, the binning technique is applied so that the resolution is limited to 960×600 pixels with a maximum sampling rate of 165 frame per second (fps). The computer runs on Intel i7 CPU with 16G RAM.

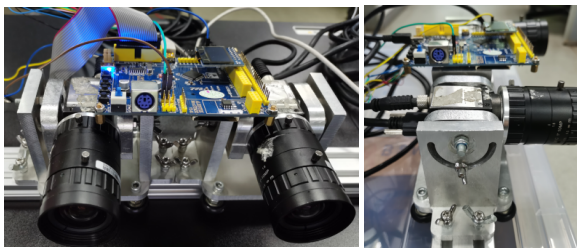

Fig 2 Camera adjustable baseline and pose

The non-potential-free general purpose input output (GPIO) cable requires trigger signals of 2.8 V or higher. The STM32 is able to output 3.3 V via the PWM signal generated from script. The two GPIO trigger cables from the camera are joint and connected to the I/O pins of the STM32. With a square wave rising edge, the cameras initiate, and the captured images are stored first in the camera buffers. The images are then saved on the computer as TIFF format.

For the software interface, multithreaded image acquisition script is developed on Linux. The trigger source is set as 'Line1'. The GPIO port PA8 can be programmed via timer T1M1 to produce Pulse Width Modulation (PWM). The STM32's APB1 bus clock Tclk (36MHZ) is set as the clock frequency of TIM1. The frequency of the PWM wave output by the PA8 pin is determined by configuring the value of the automatic reload register ARR and the Prescaler register (PSC) using:

$$f = Tclk / ((ARR+1)*(PSC+1)) \qquad (2)$$

For example, we configure ARR=359 and PSC=999 to generate the PWM wave with a frequency of 100Hz.

B. Synchronization Performance Test

In this section, two methods are provided to test the synchronization performance of the stereo cameras: 1) video sequence using a display; 2) a physical free fall experiment. Before that, it is necessary to identify any delay from the trigger signal itself. The STM32's PA4 and PD2 pins are connected to the oscilloscope with a frequency of 12.51Hz. In Fig 3, the difference between the rising edges is 512.8 ns – this shows a negligible delay between the input and output trigging signals, suitable for stereo camera hardware trigger.

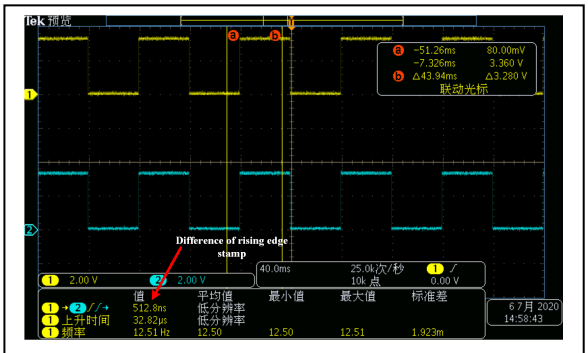

Fig 3 Difference between trigger signals 1 and 2

1) Video sequence display

Based on the hardware and software setup described so far, the trigger signal is set from 15 to 30Hz, and the camera exposure time is 2000 μs. The two cameras are configured to capture a sequence of images from a display playing a video sequence with dynamic movement. The video playback frame rate and the camera acquisition frame rate (i.e. the trigger signal frequency) remain the same, and gradually increase until the images captured by the left and right cameras are out of sync i.e. the sequence number of any image pairs becomes different. The video plays at 30 fps, but it is modified to suit lower rate such as 15 and 16 fps. The Adobe Premiere Pro CC® is used to insert and display the frame sequence number to each frame so as to determine if the captured image pair is from the same frame of source (Fig 4). If the sequence number is the same, the acquisition is synchronized. The recording is carried out indoors.

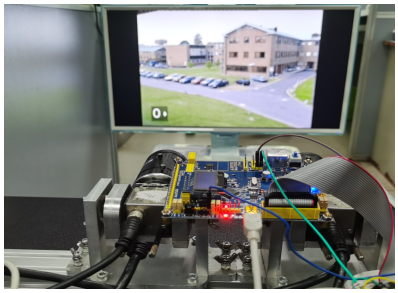

Fig 4 Image acquisition of a video sequence

2) Free fall experiment

The free fall experiment is designed to measure the relative delay between the stereo cameras. A 1-kg yellow ball with negligible air resistance is attached to a thin thread and later released by using a heat torch to burn the thread. A 200x30 cm black-white horizontal stripe background sheet with 1 cm thickness is positioned just behind the falling ball as a visual reference (Fig. 5).

Based on the illumination level of a 'sunny' day in Chengdu, the exposure time is adjusted to 800 μs. The baseline of the stereo pair is initially set to 15 cm. The middle of the camera baseline is positioned 300 cm away from the vertical gravity trajectory of the falling ball. The alignment of the camera baseline relative to the vertical gravity trajectory of the falling ball is achieved by using an infrared level. First, the baseline is adjusted to be horizontal. Then one of the two cameras is adjusted so that the optical axis of both are approximately in parallel and the plane formed by these two parallel axes is perpendicular to the vertical line of gravity. This can be achieved by first observing the boundaries of the camera fields of view in the live view window, and then compare the centroid pixels in the two cameras with an horizonal line in the view. The ball is released at the centreline between the cameras projected to the stripe sheet.

The camera pair is triggered to capture at the maximum frame rate of 100 Hz. In order to pinpoint the precise ball position in the left-right frame of the cameras, the Hough Transform is used to detect the top line of each image, and the HSV filter is applied to extract the yellow ball area. The image 'open operation', i.e. erode and then dilate is applied to both images to remove any noise separated from the ball area. The connected area is then identified, and the ball position is determined as the centroid of the area. The generated rectangle representing the footprint of the ball can be expressed in the image coordinates in pixel. The vertical ordinate y value in the image indicates the ball vertical position. By calculating the pixel difference of the top edge of the rectangle in two subsequent images from one camera, one can derive the time interval from this pixel difference. The difference in the 'interval measured in pixel' between the nominally synchronized left and right cameras is the physical measure of delay between the two cameras.

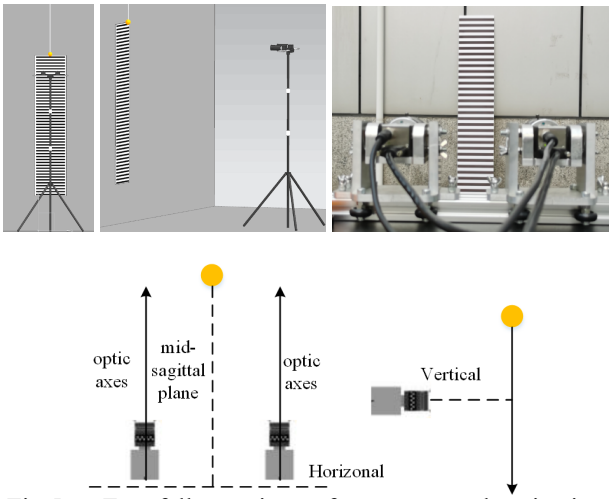

Fig 5 Free fall experiment for stereo synchronization evaluation

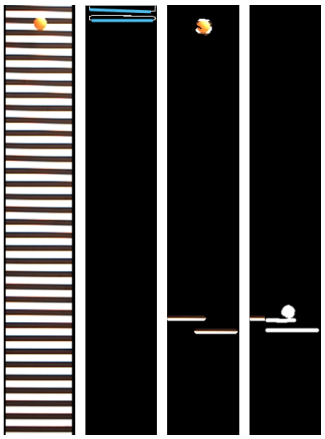

Fig 6 Ball extraction during the free fall

C. Stereo calibration

Two stereo calibration workflows are adopted: 1) Zhang's method based on bundle adjustment implemented in the MATLAB 2020a® stereo calibration toolbox on Windows 64 bit [9]; 2) feature matching [10] and optimization using sparse bundle adjustment [11] implemented in the wave acquisition stereo system (WASS) [2]. The WASS pipeline runs on Linux. The Baumer® VCXU_23C cameras have fixed focus (12 mm) and aperture (f3.6). Baseline setup is investigated at 70, 138 cm with different sets of checker box tile sizes (10, 15 and 20 cm) and calibration distances (4, 8, 12, 16, 20, 25, 30, 35, and 40 m).

IV. Results and Discussions

Outdoor scenes are used in the calibration process. Since the calibration scenes are stationary, the software trigger is used.

A. Video sequence display

The video sequence lasts for 89 seconds, each time with a different frame rate varying from 15 to 30 Hz. The external trigger frequency is the same as the video frame rate (Fig 4). Depressing the start button will initiate the external trigger, the video, and the image acquisition process. 16 sequences of stereo images are collected with one example shown in Fig 7. The ordinate represents the synchronized time – the time length of the recorded sequence stereo pair that is in sync – until 89 s. The abscissa indicates the acquisition rate or trigger signal frequency. It is apparent that the image pair is in sync from 15 to 30 Hz.

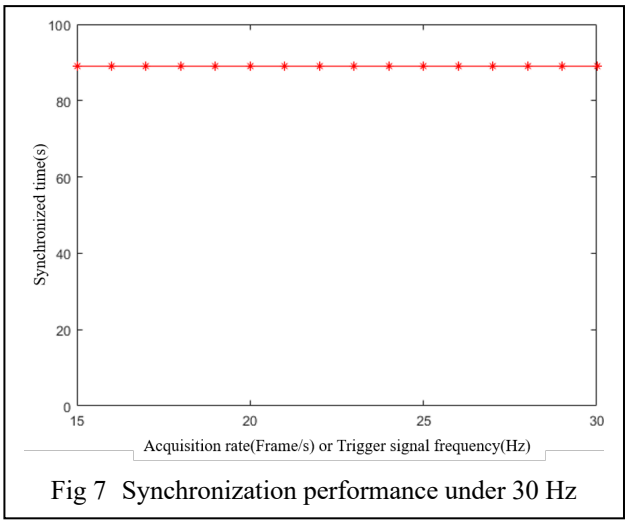

Fig 7 Synchronization performance under 30 Hz

By using the maximum display refresh rate of 60Hz to play the video, the stereo images are still in sync. So the maximum synchronous frame rate is still unknown using this method. Fig 8 shows the working sequence of the camera using external trigger signal. The exposure time $t_{exposure}$ can be adjusted but not the readout time of the camera sensor $t_{readout}$. If the rising edge time interval is too short, the external trigger would be invalid. To prevent this, a method to confirm the maximum external trigger frequency is required.

The real image acquisition frame rate f_{real} is defined as:

$$f_{real} = n_{image} / (t_{end} - t_{begin}) \quad (2)$$

where n_{image} is the number of cycles executed by the acquisition function; t_{begin} and t_{end} are the acquisition start and end timestamps respectively.

The f_{real} depends on $t_{exposure}$ and $t_{readout}$. In the lab, the $t_{exposure}$ is set to 2000 μs in order to obtain the maximum synchronous frame rate. The relationship between trigger signal frequency and real image acquisition frame rate is shown in Fig 9. The f_{real} remains the same until the trigger signal frequency reaches 120 Hz. As the trigger frequency increases, f_{real} settles at 120 Hz – the maximum external trigger frequency.

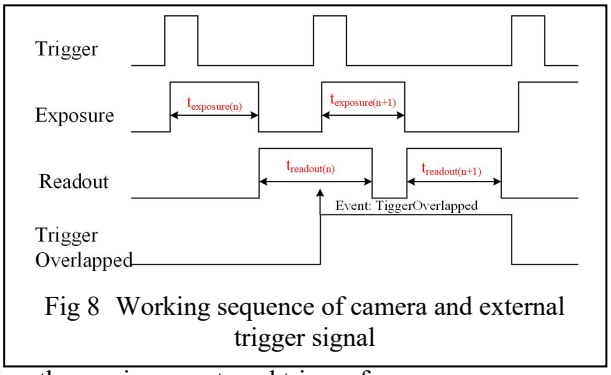

Fig 8 Working sequence of camera and external trigger signal

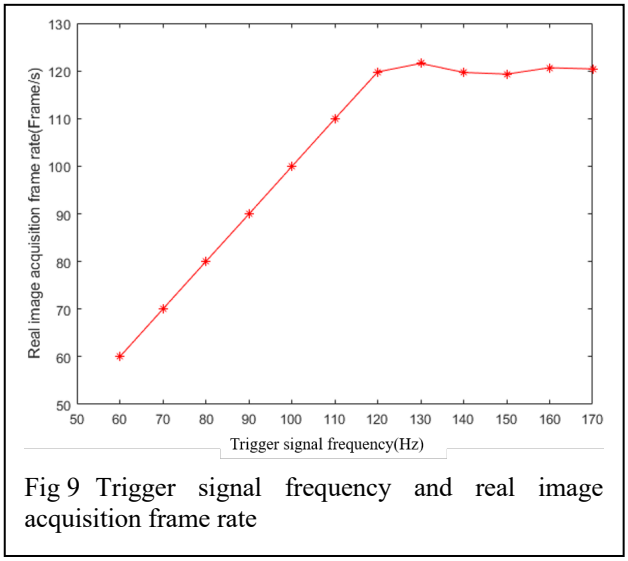

Fig 9 Trigger signal frequency and real image acquisition frame rate

B. Free fall experiment

Once the maximum image acquisition frame rate is determined in the lab, we can proceed to find out the delay between the stereo cameras. Using the free fall experiment (Fig 5). The camera exposure time is set to 800 μs after testing, the maximum external trigger frequency is set to 100Hz. This frequency is related to exposure time and readout time defined in the previous section, so different scene may result in different frequency. 23 sequences of the stereo images are acquired with the free falling ball positions extracted. In Fig 6, the left image is the original image. The Hough Transform is used to detect the top line with the position Ytop in pixel. Then the HSV filter and morphological operations are applied to extract the ball position with the ordinate of the centroid of the ball Yball.

The RDistBall represents the relative distance of the ball from the top, RDistBall = Yball-Ytop. The IdistBall represents the interval descending distance of ball in 1/100 s – the interval of trigger signal, IdistBall(n) = Yball(n+1)-Yball(n), where n is frame number. Fig 10 shows the relative position error representing the difference between the left and right cameras' RdistBall; the interval error is the difference between the left and right cameras' IdistBall(n). The mean error of relative position is 0.904 pixel, and the mean interval error is 0.475 pixel. It is possible that uncertainties in Ytop

lead to the higher error in the relative position error. The interval measured pixel error shows that during two consecutive triggers, the difference between the pictures captured by the left and right cameras is on average 0.475 pixels, about 2.8 μm. The salient points on the image would not be affected, and the image pair is regarded synchronized.

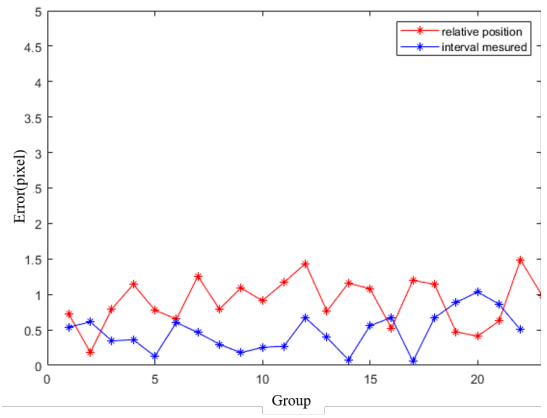

Fig 10 Error in relative position and time interval measured

C. *Stereo camera configuration of baseline*

The maximum baseline between the camera pair is 138cm. The MATLAB Stereo Camera Calibrator is used to calibrate the stereo cameras. Two baseline distances (70, 138cm), nine target distances (4, 8, 12, 16, 20, 25, 30, 35, 40 m), and three checker box tile sizes (10, 15 and 20 cm) are investigated. This produces 54 sets of combinations using the naming convention 'baseline-distance-square'. So the name '70-4-10' represent a trial with baseline 70 cm, target distance 4 m and tile size 10 cm. Each collection has 25 image pairs. The intrinsic and extrinsic parameters of the cameras are calculated and shown in Fig 11.

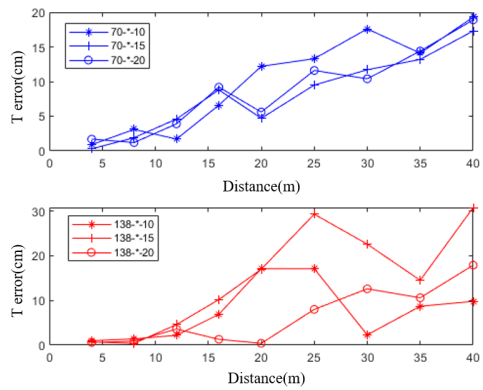

Fig 11 Estimated T error with baseline distance of 70 cm and 138 cm

Fig 11 shows the T error i.e. the difference between the estimated baseline distance and the actual baseline distance. The offset matrix TanslationofCamera2 generated in MATLAB measures the translational offset of camera 2 relative to camera 1 (left). Our camera is positioned horizontally and their optic axes in parallel. The first element in the matrix is the estimated baseline distance. It is used to assess the accuracy of parameters in our experiment. The blue curve is the error using 70 cm baseline, and the red curve is the error using 138 cm baseline. The tile size has little influence at the 70 cm baseline – the mean error is less than 0.7 cm when the calibration distance is less than 8 m. The error gradually increases with the distance increases. For 138 cm baseline, 15 cm tile size seems to outperform others as the calibration distance increases.

The semi-global block matching (SGBM) algorithm deployed by MATLAB is used to reconstruct disparity map and target depth. It also supplies intrinsic calibration parameters to the WASS pipeline. Once done, wass_match and wass_autocalibrate are used to calculate and optimize the extrinsic parameters. The wass_stereo is then deployed to generate disparity map. Different filtering steps and sparse bundle adjustment are implemented in the WASS pipeline. In WASS, the camera poses relative to the perpendicular axis to the mean sea surface is 20 degree. In our experiment, the camera optic axis is horizonal, so this parameter is set to 90 degree. In theory, longer baseline would produce wider field of view. The baseline of 138 cm and a tile size of 10 cm are chosen to compare the two methods. The reprojection errors are shown in TABLE I. The average reprojection error using WASS is much smaller than that produced by MATLAB at the target distance of 30 m. At greater distances, the depth map quality degrades. The filters and the sparse bundle adjustment used in the WASS could reduce reprojection error within a certain distance.

TABLE I. AVERAGE REPROJECTION ERROR

Real Depth(m)	Matlab average reprojection error (pixel)	WASS average reprojection error(pixel)
4	0.131	1.91e-12
8	0.0956	3.52e-12
12	0.0638	2.70e-13
16	0.0764	3.06e-12
20	0.0947	4.85e-13
25	0.0629	1.352-12
30	0.0449	NaN
35	0.0477	NaN
40	0.0496	NaN

V. CONCLUSION

The paper presented some preliminary work trying to physically measure the delay between two stereo cameras intended to produce wave depth map. A stereo vision system is configured to allow varies baselines and camera poses. These variables in combination with different target distances and calibration checker box tile sizes provide the readers with some insights into the practical requirement of a stereo vision system for outdoor wave depth acquisition. Two existing stereo calibration packages are compared.

ACKNOWLEDGMENT

The authors like to acknowledge the technical support provided by students from the Institute of Astronautics and Aeronautics at UESTC.

REFERENCES

[1] A Benetazzo, F Serafino, F Bergamasco, G Ludeno, F Ardhuin, P Sutherland, M Sclavo, F Barbariol, “Stereo imaging and X-band radar wave data fusion: An assessment,” Ocean Engineering, vol. 152, pp. 346-352, May 2018.

[2] F Bergamasco, A Torsello, M Sclavo, F Barbariol, A Benetazzo, “WASS: An open-source pipeline for 3D stereo reconstruction of ocean waves,” Computers & Geosciences, vol. 107, pp. 28-36, Oct 2017.

[3] B Bovcon, R Mandeljc, J Perš, M Kristan, “Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation,”. Robotics and Autonomous Systems, vol. 104, pp. 1-13, June 2018.

[4] A Benetazzo, “Measurements of short water waves using stereo matched image sequences,”. Coastal Engineering, vol. 53, pp. 1013-1032, Dec 2006.

[5] P Hu, X Hao, J Li, C Cheng A Wang, "Design and Implementation of Binocular Vision System with an Adjustable Baseline and High Synchronization," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, 2018, pp. 566-570

[6] D Lin, A Zhang, R Shen, P Wang, L Yang, X Chen, "Dual-camera synchronous acquisition method for binocular vision pulse measurement system," Journal of Jilin University (Engineering and Technology Edition), vol. 6, pp. 40, 2015.

[7] W Boonsuk, “Investigating the effects of stereo camera baseline on the accuracy of 3D projection for industrial robotic applications,” International Journal of Engineering Research and Innovation, vol. 8, no. 2, pp94-98, 2016.

[8] A Brandt, JL Mann, Rennie, SE Rennie, AP Herzog, TB Criss, “Three-dimensional imaging of the high sea-state wave field encompassing ship slamming events,” Journal of atmospheric and oceanic technology, vol. 27, no. 4, pp. 737-752, 2010.

[9] Z Zhang, “A Flexible New Technique for Camera Calibration,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.11, pp. 1330-1334, Nov 2000.

[10] A Albarelli, E Rodolà, A Torsello, “Imposing semi-local geometric constraints for accurate correspondences selection in structure from motion: a game-theoretic perspective," International journal of computer vision, vol. 97, no. 1, pp. 36-53, 2012.

[11] M Lourakis, AA Argyros, “SBA: a software package for generic sparse bundle adjustment,”. ACM Transactions on Mathematical Software (TOMS), vol. 36, no. 1, pp. 1-10, 2009.