VIDEO-BASED POINT CLOUD GENERATION USING MULTIPLE ACTION CAMERAS

Tee-Ann Teo *

Dept. of Civil Engineering, National Chiao Tung University, Hsinchu, Taiwan 30010. – tateo@mail.nctu.edu.tw

WG IV/7, WG V/4

KEY WORDS: Action cameras, Image, Video, Point clouds

ABSTRACT:

Due to the development of action cameras, the use of video technology for collecting geo-spatial data becomes an important trend. The objective of this study is to compare the image-mode and video-mode of multiple action cameras for 3D point clouds generation. Frame images are acquired from discrete camera stations while videos are taken from continuous trajectories. The proposed method includes five major parts: (1) camera calibration, (2) video conversion and alignment, (3) orientation modelling, (4) dense matching, and (5) evaluation. As the action cameras usually have large FOV in wide viewing mode, camera calibration plays an important role to calibrate the effect of lens distortion before image matching. Once the camera has been calibrated, the author use these action cameras to take video in an indoor environment. The videos are further converted into multiple frame images based on the frame rates. In order to overcome the time synchronous issues in between videos from different viewpoints, an additional timer APP is used to determine the time shift factor between cameras in time alignment. A structure form motion (SfM) technique is utilized to obtain the image orientations. Then, semi-global matching (SGM) algorithm is adopted to obtain dense 3D point clouds. The preliminary results indicated that the 3D points from 4K video are similar to 12MP images, but the data acquisition performance of 4K video is more efficient than 12MP digital images.

1. INTRODUCTION

1.1 Motivation

Three-dimensional geospatial information of indoor environment can be generated from cameras and laser scanners. Laser scanners obtain 3D points directly while camera indirectly obtains 3D points via stereo image matching. Digital still cameras and digital videos are two possible ways to collect digital images for image matching. Nowadays, a lightweight action camera such as GoPro Hero 4 Black Edition is able to collect digital still images up to 12Mp (4000 x 3000) resolution and video up to 8.3MP (3840 x 2160) resolution at 30 frames per second. Although the spatial resolution of a digital still camera is higher than a digital video, the sampling rate of a digital video is better than a digital still camera. As the video data can be converted to frame images like digital still camera, these highly overlapped frame images from video provide high similarity and high redundancy for image matching. In addition, action camera is able to acquire both video and image (5 seconds per frame) simultaneously. Therefore, there is a need to compare these two strategies for indoor point clouds generation.

1.2 Action Cameras

With the development of camera technology, most action cameras provide both image and video functions. To compare the traditional consumer digital camera and action camera, the action camera, such as GoPro (GoPro, 2015), emphasizes on: light weight, small dimensions, waterproof, large field-of-view (FOV), 4K video recording and high burst frame rate. The comparison of up-to-date action cameras can be found at (Crisp, 2014; Staub, 2015). The action cameras are originally developed for sports and underwater usage. The user uses the action camera to record their activities during extreme sports or special events. Due to the light weight, low cost and high spatial resolution of video mode, the usage of action cameras are extended to unmanned aerial vehicle (UAV), mobile mapping system (MMS), and other photogrammetric purposes.

1.3 Related Works

The digital video devices record sequence images and these dynamic sampling images can be used for different applications. The traditional photogrammetry is mostly relied on high spatial resolution images. Due to the improvement of video's resolution and frame rate, the use of video technology for collecting geo-spatial data becomes an important trend. Many video-related applications are presented in different geoinformation-related domains. For example, the space borne Skybox™ constellation is capable of acquiring sub-meter satellite imagery and high-definition panchromatic video for earth monitoring; the video collected by UAV can be used to produce geospatial data via Full Motion Video (FMV) in ArcGIS™ software or other commercial software; the video of car cam recorder can be used for crowdsourced street level mapping via Mapillary.com or other online-mapping services.

Several photogrammetry studies used GoPro action cameras for 3D measurement purposes. Balletti et al. (2014) discussed different camera calibration methods using GoPro for 3D measurement purposes. Kim et al., (2014) construct the 3D point clouds of building façade using GoPro 1080P super-view stereo video. As the needs of stereo vision, the GoPro Company provide accessories (i.e. dual cameras stereo housing, synchronization cable, software) to capture and produce 3D movie. Because of water proof housing, this technology has also applied in underwater stereo vision. For example, Schmidt and Rzhanov (2012) used dual GoPro cameras to measure seafloor

* Corresponding author.

micro-bathymetry. The 4K stereo videos are able to generate 3mm resolution grid of seafloor at 70cm distance. Nelson et al., (2014) combined the sonar scanner and dual GoPro cameras in a remotely operated vehicle for underwater 3D reconstruction. The results showed the potential of combining 3D sonar data and 3D surface from image matching for underwater archaeological application. The previous studies indicated that GoPro stereo videos are suitable for close-range photogrammetry purposes.

1.4 Research Purposes

The objective of this study is to compare the image-mode and video-mode of multiple action cameras for 3D point clouds generation. Frame images are acquired from discrete camera stations while videos are taken from continuous trajectories. The proposed method includes five major parts: (1) camera calibration, (2) video conversion and alignment, (3) orientation modelling, (4) dense matching, and (5) evaluation. As the action cameras usually have large FOV in wide viewing mode, camera calibration plays an important role to calibrate the effect of lens distortion before image matching. A black and white chess box pattern and Brown equation are adopted in camera calibration. Once the camera has been calibrated, the author use these action cameras to take video in an indoor environment. The videos are further converted into multiple frame images based on the frame rates. In order to overcome the time synchronous issues between videos from different viewpoints, the author manually identify image scene to calculate the time shift factor between cameras in time alignment. A structure form motion (SfM) technique is utilized to obtain the image orientations. Then, semi-global matching (SGM) algorithm is adopted to obtain dense 3D point clouds (Remondino et al., 2014).

2. EXPERIMENTS AND RESULTS

2.1 System Specifications

This study uses five GoPro Hero4 Black cameras for point clouds generation. These five cameras are integrated in a Freedom360^{TM} mount to obtain data 360 degrees panorama image and controlled by a GoPro Remote Controller. The size of this multi-view camera is about 10cm x 10cm x 10cm cube (see Figure 1). The camera provides both camera and video modes. The highest spatial image resolution for a digital still image is 12MP (4000 x 3000) while the finer spatial image resolution for a digital video is 4K (3840 x 2160) at 30 frames per second (fps). As the shutter of 4K video (1/30 sec) might produce blur images, this study also consider 1080P (1920 x 1080) at 120fps to avoid image blur. Table 1 shows the related camera parameters.

The spatial resolution of action camera is usually lower than digital single-lens reflex (DSLR) cameras. In order to understand the suitability of using action camera in close-range photogrammetry, this study analyse the spatial resolution of action camera at different distances and different modes. Figure 2 summaries the spatial resolution of image and video at nadir and diagonal points. The action camera usually has large FOV and consequently the point near to image boundaries has larger spatial resolution. This issue should be taken into consideration in 3D measurement. To obtain at least 5cm resolution, the maximum distance for 12MP image and 4K video should be less than 20m. The action camera might not suitable for long-range photogrammetry, but it is suitable for indoor environment at near range distance (<20m). Therefore, the scope of this study is to use the multiple action cameras in an indoor environment.

Table 1. Related parameters for GoPro Hero Black

Item	Description
Size	41mm x 59mm x 30mm
Weight	89g
CCDsize	1/2.3”
Nominal focal length	3mm
Image size (digital still image)	4000 x 3000
Image size (4K video)	3840 x 2160 (max 30fps)
Image size (1080P video)	1920 x 1080 (max 120fps)

Figure 1. Multiview GoPro System.

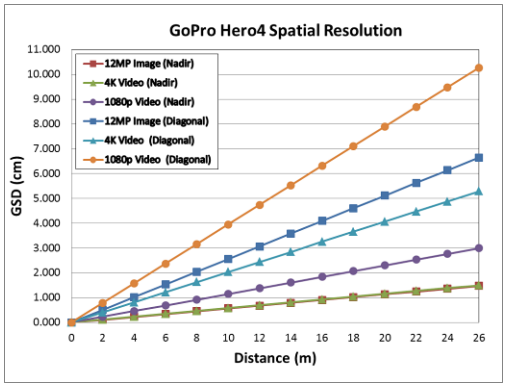


Figure 2. Ground sampling distances in different modes.

2.2 Camera Calibration

As the action cameras usually have large FOV in wide viewing mode, camera calibration plays an important role to calibrate the effect of lens distortion for image matching. This study uses Brown distortion model (equations (1) to (4)) (Brown, 1971) to determine the lens distortion. PhotoScan (Agisoft, 2015) and PhotoModeler (EOS System, 2015) are used to evaluate the results. PhotoScan uses regular chessboard pattern to obtain a large number of conjugate points in camera calibration. PhotoModeler uses circular signalized targets and self-calibration to determine the lens distortion parameters. Notice that, the radial distortion parameters K3 is needed for a large FOV camera.

$$\Delta x = \Delta x_r + \Delta x_d$$
$$\Delta y = \Delta y_r + \Delta y_d \quad (1)$$

$$\Delta x_r = \bar{x} \times (K_1 r^2 + K_2 r^4 + K_3 r^6)$$
$$\Delta y_r = \bar{y} \times (K_1 r^2 + K_2 r^4 + K_3 r^6) \quad (2)$$

$$\Delta x_d = P_1(r^2 + 2\bar{x}) + 2P_2\bar{x}\bar{y} \qquad (3)$$
$$\Delta y_d = P_2(r^2 + 2\bar{y}) + 2P_1\bar{x}\bar{y}$$

$$r = \sqrt{(\bar{x})^2 + (\bar{y})^2} = \sqrt{(x - x_0)^2 + (y - y_0)^2} \qquad (4)$$

Where, (Δx, Δy) are total lens distortion; (Δx_r,Δy_r) are radial distortion; (Δx_d, Δy_d) are tangential distortion; (K_1~K_3) are coefficients of radial distortion; (P_1~P_2) are coefficients of tangential distortion; r is radial distance; (x,y) are photo coordinate ; and (x_0, y_0) are principal points.

This study performs the camera calibration for a 12MP image, a 4K video and a 1080P video separately. In video calibration, this study uses video mode to shoot the target code at different view angles and positions. Then, these video frames are converted into images at 1 image per second. Besides, the initial focal length and frame size (Kolor, 2015) are also written at EXIF for calibration purpose. The total errors of PhotoModeler are smaller than 2 pixels in all modes. However, the PhotoScan does not provide accuracy index in lens distortion correction. Table 2 shows the results of camera calibration for camera id 2 using Photomodeler. Figure 3 show the distortion curves of radial and tangential distortions. The impact of radial distortion is significantly larger than the tangential distortion. To compare the digital still image and video, the results of PhotoModeler show high consistence in radial distortion except the tangential distortion for 1080P.

To compare the results of PhotoModeler and PhotoScan, the radial distortion of PhotoModeler is larger than PhotoScan. This study also generates two undistorted images using these two methods (See Figure 4). The behavior of these two methods is similar at the center area. But for straight lines near to the corner area, the result of PhotoModeler is better than PhotoScan. Therefore, this study uses the lens distortion parameters from PhotoModeler.

Table 2. Results of camera calibration using Photomodeler

Type	Image	Video	Video
Resolution	12MP	4K	1080P
Width (pixel)	4000	3840	1920
Height (pixel)	3000	2160	1080
F (mm)	2.752130	2.654308	2.821934
xp (mm)	3.171405	2.945599	3.139957
yp (mm)	2.412459	1.620624	1.717661
Fw (mm)	6.246702	5.881419	6.243885
Fh (mm)	4.686000	3.308000	3.514000
k1	3.887E-02	3.659E-02	3.568E-02
k2	6.895E-04	2.429E-03	8.153E-04
k3	1.491E-04	0.000E+00	1.098E-04
p1	-1.193E-04	4.979E-04	3.126E-04
p2	1.622E-04	-4.025E-04	0.000E+00
PixelSize	0.001562	0.001532	0.003252

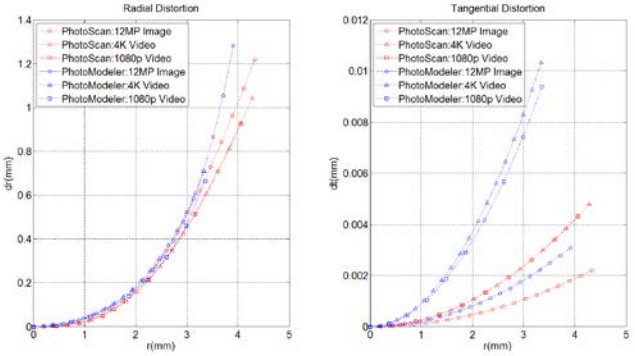


Figure 3. Results of lens distortions using different modes and different methods.

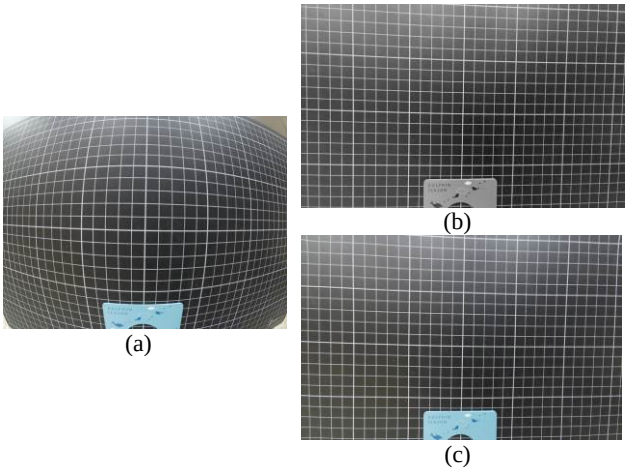


Figure 4. Original and undistorted images: (a) original image; (b) undistorted image using Photomodeler parameters; (c) undistorted image using Photoscan parameters.

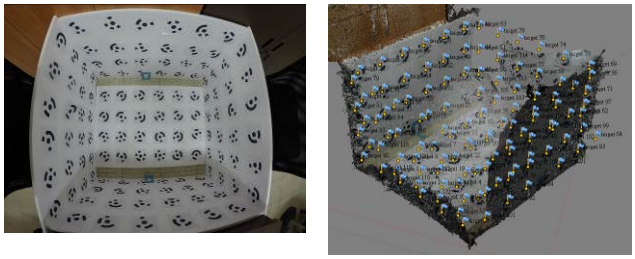
2.3 Cameras Alignment

These five cameras are fixed together in a mount and a camera alignment is needed to determine the geometrical relationship between cameras. This study uses camera 1 as the master camera while the other 4 are the slave cameras. The transformation between master and slave cameras is descripted by two sets of parameters, i.e. lever-arms (dx, dy, dz) and boresight-angles (dω, dφ, dκ). In this study, 120 signalized targets (markers) are distributed on a 90cm x 90cm x 65cm box (see Figure 5a). Then, 80 images are taken from 16 stations by 5 cameras. These 80 images are used for bundle adjustment and determine their exterior orientations in mapping frame (see Figure 5b). The lever-arms and boresight-angles are calculated by equations (5) and (6) using exterior orientations (Rau et al., 2011).

$$R_{Slave}^{Master} = R_{Mapping}^{Master} \times R_{Slave}^{Mapping} \qquad (5)$$
$$d_{Slave}^{Master} = R_{Master}^{Mapping}(d_{Slave}^{Mapping} - d_{Master}^{Mapping}) \qquad (6)$$

Where, R_{Slave}^{Master} is the rotation matrix between master and slave camera frames; $R_{Slave}^{Mapping}$ is the rotation matrix between slave camera frame and mapping frame from bundle adjustment; $R_{Master}^{Mapping}$ is the rotation matrix between master camera frame and mapping frame from bundle adjustment; d_{Slave}^{Master} is the lever-

arms between master and slave camera; $(d_{Slave}^{Mapping}, d_{Master}^{Mapping})$ are the vectors from slave/master to mapping frame.



(a) (b)

Figure 5. Configuration of cameras alignment: (a) distribution of markers; (b) result of bundle adjustment.

Table 3 summaries the results of cameras alignment. The standard deviations of boresight-angles are better than 0.6 degrees except for camera 5 on the top. The standard deviations of lever-arms are less than 1.9cm in all cases. It is about 19% of the size of this camera system (see Figure 1). In other words, the variation of lever-arms is around 1.9cm. These parameters can only be treated as initial values in orientation modelling and further investigation is needed.

Table 3. Estimated lever-arms and boresight-angles

	Lever-arms			Boresight-angles		
	dx (m)	dy (m)	dz (m)	dω (deg)	dφ (deg)	dκ (deg)
Cam1 & 2	-0.024 ±0.019	0.081 ±0.006	-0.001 ±0.006	-0.917 ±0.247	-0.121 ±0.499	179.518 ±0.510
Cam1 &3	0.042 ±0.007	0.041 ±0.006	-0.003 ±0.007	89.123 ±0.597	-0.906 ±0.178	90.501 ±0.510
Cam1 &4	-0.052 ±0.015	0.041 ±0.015	-0.009 ±0.011	-93.035 ±0.420	1.599 ±0.449	-90.166 ±0.501
Cam1 &5	-0.006 ±0.009	0.031 ±0.008	-0.024 ±0.014	16.646 ±3.750	86.004 ±0.256	96.406 ±3.863

2.4 Data Synchronous

These five cameras are controlled by a remote control and no cables are connected between cameras. The author found a slightly time lag when triggering the camera to take image or video. This time lag does not affect the digital still image on a fixed tripod, but it might cause the data asynchronous in the video mode. As there is no cable to connect these cameras for synchronous purpose, the only way is using an additional timer to align the videos. Figure 6 shows the same timer taken from different cameras using video mode. A timer APP which has 1/100 sec precision in time alignment is used. All videos are shot to a same timer separately and the videos recorded times are shifted to the reference time of time. Although the timer may provide 1/100 sec precision, the time alignment precision is restricted by frame rate. For example, the time interval of 4K video frame is 1/30 sec. This method can only ensure 1/30 sec time synchronous for a 4K video.

2.5 Point clouds generation

After camera calibration, cameras and time alignment, video are converted into image frames at different sampling interval for 3D point clouds generation. The procedure includes: (1) structure from motion (SfM) technique for image orientations; (2) absolute orientation using control points; (3) semi-global matching (SGM) algorithm for dense point matching. This study utilizes a commercial Agisoft PhotoScan in 3D point clouds generation.

3. EVALUATION

The evaluation includes two cases, one is a stair and the other is a lobby.

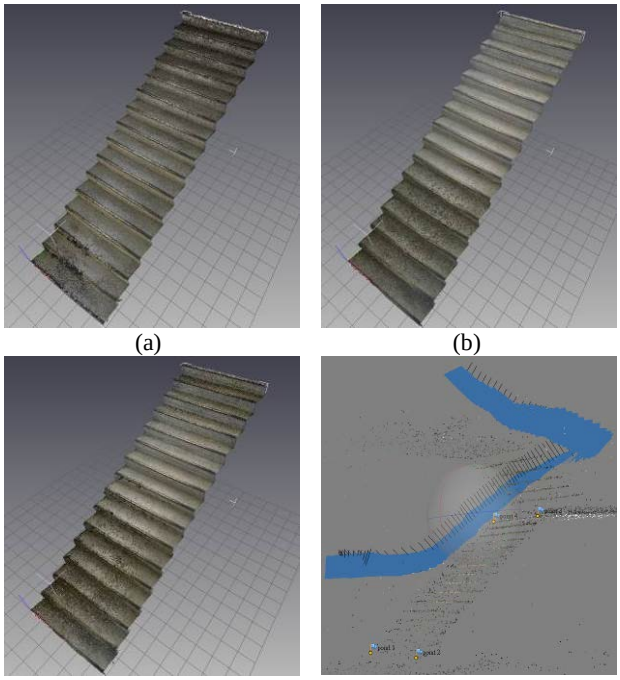
3.1 Case 1. Stair

The 3D stair modelling is a challenging task in indoor modelling. A discrete digital still image usually cannot provide favourite intersection geometry due to the limited camera station. In the contrary, a digital video is able to take multi-view images effectively. The aim of this session is to compare the performance of a 12MP image, a 4K video and a 1080P video. Only one action camera is adopted in this section. For a digital still image, the author take the images for every steps of the stair. The duration of images is about 150 seconds. However, the duration of video is only 25 seconds for the same area. The data acquisition of video mode is much effective than image mode. Besides, the standard deviations of camera baseline are 15.1cm for 12MP, 7.9cm for 4K and 3.8cm for 1080P. The digital video may provide more uniform camera station than digital camera. To compare the 4K and 1080P videos, the resolution of 4K is higher than 1080P while the sampling rate of 4K (i.e. 30fps) is lower than 1080P (i.e. 120fps). Hence, the image quality (e.g. effect of motion blur) for 1080P is better than 4K visually.

The author use the image and video in relative orientation modelling and 4 control points are manually selected in absolute orientation. The residual of control points are less than 5cm in the three cases. Then, high density image matching is used to obtain point clouds of a stair. Table 4 summaries the results of these three modes. The point density of 12MP is the highest one, but the result of 4K video is similar to the results of 12MP. Figure 7 is a section of stair for comparison. The section includes 18 steps and the size of the stair is about 1.5m width, 4m length and 2.4m height. The shape of these three results shows high consistency. In other words, the 4K video is possible to produce similar results like 12MP images.

Table 4. Comparison of images and videos for a stair

	12MP	4K video	1080P video
Number of pixel	12MP	8.29MP	2.07MP
Duration of data acquisition	150sec	25sec	25sec
Sampling rate	3sec (manual)	0.5sec	0.25sec
Number of image	46	50	100
Estimated processing time	1hr	1hr	2hr
Camera distance (m)	38.2±15.1	27.5±7.9	15.0±3.8
Point density (pts/cm^2)	65	60	44



(a) (b) (c) (d)

Figure 7. Results of a stair: (a) results of 12MP images; (b) results of 4K video; (c) results of 1080P video; (d) perspective centre of cameras for 1080P video.

3.2 Case 2. Lobby

In Case 2, the author uses the multi action cameras system to reconstruct the point clouds of a lobby. The test area is about 20m width, 15m length and 3m height. In order to have multi-view images for image matching, a tripod is used to take digital still images at five different heights (i.e. 1.0m, 1.25m, 1.50m, 1.75m, and 2.00m). The distance between cameras for the same station is about 0.25m while the distance between different stations is about 3m. The duration of image acquisition for these five stations is about 10 minutes. The duration of 4K video is just 22 seconds for the same area. Table 5 summaries the results of these two modes. The video mode obtains continuous image frames. The average camera centre of video mode is 32.2cm. Therefore, the number of frame from video is larger than traditional digital image (i.e. 376 images > 125 images). However, the video mode needs more computational time to produce point clouds (i.e. 4hrs > 2hrs).

Table 5. Comparison of images and videos for a lobby

	12MP	4K video
Number of pixel	12MP	8.29MP
Duration of data acquisition	600sec	47sec
Sampling rate	-	0.5sec
Number of image	125	376
Estimated processing time	2hrs	4hrs
Camera distance (m)	Same station: 25cm Between station: 300cm	32.2cm±11.6cm
Point density (pts/cm^2)	78.6	72.8

Figures 8a and 8b show the distributions of 12MP images and 4K video frames. The numbers of camera station for 12MP and 4K video are 5 and 94. The video mode provides high overlapped multiview images for space intersection. Figures 8c and 8d compare the generated points from the same view point. The point clouds of 4K video are more complete than the 12MP image. Due to the limitation of image matching, the area without texture does not have 3D points after image matching.

(a) (b) (c) (d)

Figure 8. Results of a lobby: (a) perspective centres of 12MP images; (b) perspective centres of 4K video frames; (c) points from 12MP image; (d) points from 4K video.

4. CONCLUTIONS AND FUTURE WORKS

This research proposed a multiple action cameras system for indoor mapping. The characteristic of this system is 360 degrees panorama imaging and 4K high resolution video. It is beneficial for data acquisition in an indoor environment as well as 3D point clouds generation. This study also demonstrated the results of camera calibration for image and video modes. The maximum radial distortion of a4K video reached 500 pixels at image boundary. The lens distortion should be pre-calibrated as the impact of lens distortion was significant in related to image frame. These five cameras were mounted together and the lever-arms and boresight-angles were calculated by cameras alignment. The results of cameras alignment can be used as the initial orientations in orientation modelling. The time synchronous was implemented by an additional timer in video mode. It can adjust the time tag issue of this system. Finally, the

3D point clouds were generated by orientation modelling and dense matching.

The preliminary result indicated that the 3D points from a4K video were similar to 12MP images. Besides the data acquisition performance of a4K video was faster than 12MP digital images, the limitation of this video-based point clouds generation is the huge computational time for large data set and low image quality caused by video compression and motion blur. Future works will evaluate the system in different scenarios and different parameters. As the radiometric performance of action camera will influence the geometrical performance, future works will focus on the radiometric performance for action cameras in image and video modes.

ACKNOWLEDGEMENTS

This investigation was partially supported by the National Science Council of Taiwan under project number NSC 101-2628-E-009-019-MY3.

REFERENCES

Agisoft, 2015. PhotoScan, URL: http://www.agisoft.com

Balletti, C., Guerra, F., Tsioukas, V. and Vernier, P., 2015. Calibration of action cameras for photogrammetric purposes, Sensors, 14: 17471-17490.

Brown, D.C., 1971, Close-range camera calibration, Photogrammetry Engineering, 37:855-866.

Crisp, S. 2014. 2014 Action camera comparison guide, Gizmag, URL: http://www.gizmag.com/compare-best-action-cameras-2014/34974/

EOS System, 2015, PhotoModeler Motion, URL: http://www.photomodeler.com

GoPro, 2015, GoPro Hero4 Black, URL: http://gopro.com

Kim, J.H., Pyeon, M.W., E.O, Y.D., and Jang, I.W., 2014. An experiment of three-dimensional point clouds using GoPro, International Journal of Civil, Architectural, Stuctural and Construction Engineering, 8(1): 82-85.

Kolor, 2015, About GoPro focal length and FOV, URL: http://www.kolor.com/wiki-en/action/view/Autopano_Video_-_Focal_length_and_field_of_view

Nelson, E.A. Dunn, I.T., Forrester, J., Gambin, T., Clark, C.M. and Wood, Z.J. 2014. Surface reconstruction of ancient water storage systems: an approach for sparse 3d sonar scans and fused stereo images. GRAPP, 161-168.

Rau, J.Y., Habib, A.F., Kersting, A.P. Chiang, K.W., Bang, K.I., Tseng, Y.H. and Li, Y.H. 2011. Direct sensor orientation of a land-based mobile mapping system, Sensors, 11: 7243-7261.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F. and Nex, F. 2014. State of the art in high density image matching, Photogrammetric Record, 29 (6): 144-166.

Schmidt, V.E. and Rzhanov, Y., 2012 Measurement of micro-bathymetry with a GoPro underwater stereo camera pair, IEEE Ocean 2012, 1-6.

Staub, D., 2015. 2015 Best action camcorders review, Top Ten Reviews, URL: http://action-camcorders-review.toptenreviews.com/