International Journal of Research in Engineering, Science and Management
Volume 4, Issue 9, September 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Structure from Motion

Subhrangshu Adhikary*
Department of Computer Science and Engineering, Dr. B. C. Roy Engineering College, Durgapur, India

***Abstract*: With increasing computational capacity, it is becoming easier to simulate much of the human brain activity. Similar to left and right eyes coordination to perceive shape and distance of the surrounding objects, we can use cameras to get the images of an object from different angles and perform a popular Computer Vision technique known as Structure from Motion (SFM) and utilizing this, we can compute a 3D point cloud and visualize the object in 3 Dimensions. Through this article, we have reviewed the latest advancements in SFM techniques, experimented with dataset, attempt to find limitations for low textural variance using Grey Level Co-Occurrence Matrix methods causing missing pixels problem in the reconstructed 3D point cloud and suggest methods to overcome them based on recently discovered techniques. SFM technology can reduce substantial workload while designing virtual reality environment for commercial domain, engineering surveys and medical assistance.**

***Keywords*: Augmented reality, Computer vision, Structure from motion, 3D Mapping, Pointless SFM.**

1. Introduction

Our human brain deciphers the shape and distance of an object by calibrating the differences between the images obtained by our left and right eye [1]. This helps us to estimate a 3D structure about the objects in front of us in our mind. Scientists over the years have tried to understand the working mechanism of this left-right eye calibration. The major challenges for this are computational power. Consisting of 100 billion neurons, our brain process at up to terahertz scale [2]. To mimic the 2D image to 3D deciphering process of the human brain, we would need gigahertz scale computational power which was not very common until a couple of decades [3]. Now with the technological improvements, the processing speed of gigahertz has become common and along with this SFM techniques has been enhanced substantially [4].

Now the main ideology behind a 2D to 3D reconstruction is to first determine the similarity and differences between different frames of images of the same object captured from different angles as shown in fig. 1. For this purpose, firstly unwanted images are filtered out through several techniques like grey level co-occurrence matrix, canny edge detection techniques, K-Means Clustering to cluster colours, etc. From these methods, the pattern between the images is extracted and compared to filter out the contrasting images. After this, a method called tf-idf (term frequency-inverse document frequency) weighting is implemented to assign weights to various images which would be later compared to find the similarity. This is given by,

$$t_i = \frac{n_{id}}{n_d} log \frac{N}{n_i} \quad (1)$$

where n_i and n_j represents the common features within image i and j.

JacS (Jaccard Similarity) methods are also often used along with tf-idf to eliminate any noise in image-pair set. It is computed by,

$$s_{ij} = \frac{n_{ij}}{n_i + n_j - n_{ij}} \quad (2)$$

The generated sets n_i, n_j and s_{ij} are then passed through K-Means Clustering process to filter out the closest image pairs k which accumulates the set into k number of clusters by forming centroid [5], [6].

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (3)$$

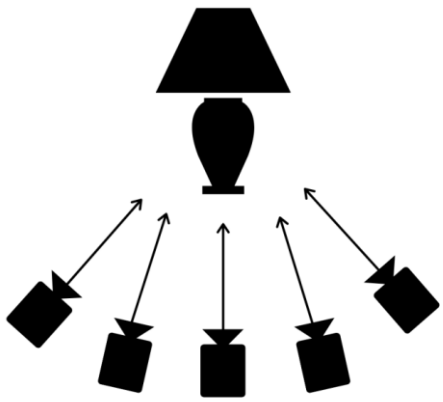

Fig. 1. Ideology behind Structure from Motion. An object is scanned from different angles and then the images are compared to construct a 3D point cloud

After the filtration process through K-Means Clustering to find the closest pair for the images, next comes the estimation of camera points based on the image pairs and the process is called camera calibration [7], [8]. 2D line-to-line matching technique is utilized to compute the overlap between the image-pair with the help of common 3D points. The similarities and

*Corresponding author: subhrangshu.adhikary@spiraldevs.com

differences between the overlapping images are then estimated to find the camera position responsible for both the images. Also during this, the overlapping points, based on the change in texture, orientation, depth, etc., the points of both the images are mapped to a 3D point cloud matrix. The process is repeated over and over again for several numbers of images until a 3D point cloud matrix is formed consisting of as little missing pixels as possible [9]. Now, this 3D point cloud is the visual representation of the structure reconstructed from 2D images. The process can generate 3D structures in both scenarios, firstly where the subject is constant and the camera is moving and secondly where the camera is constant and the subject is moving. The accuracy or density of the 3D point cloud depends on the frames the camera can capture along with the resolution [10].

The technique could be utilized for easier 3D mapping in architectural domains to construct environments and buildings, in machinery domains to create 3D models of the machines used, during health diagnosis to assist during surgery, in the entertainment field for creating inexpensive visual effects, etc. Through this review article, we have tried to highlight the latest advancements in the study area of Structure from Motion for the creation of augmented reality through artificial intelligence consisting of deep learning and computer vision. We would also like to test the latest SFM technologies and discuss their performances.

2. Related Works

The SFM ideology comes from the motion parallax phenomenon [11], [12]. Motion parallax is the phenomenon that two moving objects moving with the same velocity along the same direction at the same side of the point of reference, the object closer to the point of reference appear to move faster than the other object. The angle created by the initial and final position of the objects with the point of reference is larger for the closer object and thus causing the motion parallax effect. In both scenarios, first, where the point of reference is fixed and second, where the subject is fixed and the point of reference is moved, the motion parallax effect could be exploited to determine the depth, size, distance or velocity of the objects where there exist a relative motion between the object and point of reference [13], [14]. The features extracted by motion parallax estimation methods could be utilized for the structure from motion technique to establish a point cloud matrix in 3D space [15]. The closest images could be studied for similar pixels whose difference in the given two images could be used to estimate a depth map which could further be combined to a point cloud matrix for 3D projection [16], [17].

Advancements in neural network drastically improved computer vision techniques [18] Convolution neural network is very popular neural network model which can handle image data very well by extracting features from image matrix [19]. Deep learning is a form of neural network formed by combining several layers of perceptron's to optimize the capabilities of the network. The feature extraction process with help of deep neural network made it possible to extract several features from images and utilizing these features to estimate a 3D point cloud from several nearby images [20]. With this strategy, researchers have made several successful 3D structure reconstruction with image or video data [21]. Along with convolution neural networks, other popular deep learning models like auto-encoders are also proven to be effective to build a 3D point cloud [22].

With the perspective of the dataset, SFM could be constructed with multiple techniques. For example, the camera could have no additional features apart from the pixels matrix [23]. In this process, the computational power requirements are low however the model faces difficulty reconstructing 3D with low variance in texture. These are performed mainly when the camera points are not very far apart, most likely all camera points are within a meter range to each other [24]. The other could be to use additional geotagging parameters such as altitude, latitude and longitude [25]. Both manned and unmanned aerial vehicles based sensors including satellites are majorly utilized for the purpose and much larger structures like buildings, landscapes or terrains are reconstructed with this technique facilitating a low-cost aerial survey [26], [27].

SFM often fails to differentiate pixels from the highly similar surface. For this reason, we need to study the Grey Level Co-Occurrence Matrix (GLCM) and its properties for the images [28]. GLCM traverse through the array of image and find similar patterns which could be found in different places of the matrix [29]. This is a very useful technique for determining the similarity between different images. Same could also be used to study the variations among the image. Glossy surfaces have lower variance and rough surfaces have higher variance. SFM often mixes pixels of highly similar texture and we would like to test this through our work as well [30]. This motivated us to review the SFM techniques and test its limitations to generate a depth map for low variance GLCM.

3. Methodology

A. Data Collection and Preprocessing

The most crucial part of any experiment is the data collection procedure. For our experiment, we have used a mobile phone camera to record video of the object of concern. After video collection, we have broken the video into individual frames. We have removed very similar frames with almost no visible differences. Then we have resized all images to 264x264 pixels. Followed by this, we have masked out the background of the object to avoid as much distortion as possible.

B. Reconstruction Technique

1) Feature extraction and pairing

The images are now paired up based on their closest appearing images with tf-idf and JacS methods introduced earlier in introduction section. For this experiment we have used Lukas-Kanade tracker for image matching to avoid any unwanted image pairs [31]. Now our next step is to transform the model to make it suitable for the SFM reconstruction. To do this, we need to transform the normalize image coordinates to pixel coordinates. For and image with height h and width w, pixel coordinates is given by,

$$H = \begin{pmatrix} \max(w,h) & 0 & \frac{w-1}{2} \\ 0 & \max(w,h) & \frac{h-1}{2} \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

2) Camera Calibration

The camera calibration could be made to visualize the point cloud in three formats. Perspective Camera, Fisheye Camera and Spherical Camera.

For and image with coordinates *x*, *y* and *z*, the point cloud could be projected into *u* and *v* coordinates. For this, perspective camera is given by,

$$x_n = \frac{x}{z} \quad (5)$$

$$y_n = \frac{y}{z} \quad (6)$$

$$r^2 = x^2 + y_n^2 \quad (7)$$

$$d = 1 + k_1 r^2 + k_2 r^4 \quad (8)$$

$$u = f d x_n \quad (9)$$

$$v = f d y_n \quad (10)$$

Fisheye Camera is given by,

$$r^2 = x^2 + y^2 \quad (11)$$

$$\theta = arctan(r/z) \quad (12)$$

$$d = 1 + k_1 \theta^2 + k_2 \theta^4 \quad (13)$$

$$u = f d \theta \frac{x}{r} \quad (14)$$

$$v = f d \theta \frac{y}{r} \quad (15)$$

Finally, spherical camera is given by,

$$lon = \arctan(\frac{x}{z}) \quad (16)$$

$$lat = arctan(\frac{-y}{\sqrt{x^2+z^2}}) \quad (17)$$

$$u = \frac{lon}{2\pi} \quad (18)$$

$$v = -\frac{lat}{2\pi} \quad (19)$$

The projections could be utilized in all these three camera models to spectate various forms of the point cloud.

3) Depth map and Point Cloud Estimation

The image pairs formed then further process to generate the depth map and make the 3D point cloud. This is done by overlapping the images on top of each other and measuring the shifting of a pixel due to change in camera position. All the shifts are estimated to find the depth of the points and generate a depth map for the purpose. Also during the same process, the points which are completely overlapped are directly mapped to its corresponding 3D point cloud matrix and the pixels which undergone a shift due to the camera movement are then mapped to the 3D point cloud based on the depth map generated. The next closest image to the pair is also compared with this depth map to fill the missing pixels in the point cloud and replace and erroneous pixels in the point cloud to make it denser [32]. Now finally the 3D point cloud is inversed from pixel coordinates to normalized coordinates to make it easier to spectate the 3D reconstructed structure. From eqn. (4), which was given for image matrix in 2D plane, could also be represented for a 3D structure as if each layer of the 3D structure represents a 2D plane, therefore the inversion of the point cloud matrix could be given by,

$$M = H^{-1} = \begin{pmatrix} 1 & 0 & -\frac{w-1}{2} \\ 0 & 1 & -\frac{h-1}{2} \\ 0 & 0 & \max(w,h) \end{pmatrix} \quad (20)$$

Now, this 3D Point Cloud matrix M could be visualized as a 3D structure generated from 2D images.

4. Results and Discussions

A. Discussion on Reconstruction Observations

For the reconstruction mechanism, we have replicated the latest methods as discussed in the earlier section. We have captured video footage of an object "Swan" from different angles. Followed by this, we have split the video into individual frames and selected 16 images of the objects with contrasting angles. We have masked out the background to minimize distortion and noises, fig. 2.

Fig. 2. Masked out background as a part of pre-processing to reduce noises and avoid distortions

Fig. 3. Perspective projection of 3D Reconstructed point cloud generated by Structure from Motion based Computer Vision

Following this step, we have applied structure from motion techniques discussed earlier on the given set of images. After final processing, we can observe the 3D reconstructed point cloud in perspective projection fig. 3.

B. Performances and Limits Evaluation for SFM

From the 3D reconstructed image fig. 3, we can see that the body of the swan is having firm reconstruction at rough surfaces like the feathers. However, distortions or missing pixels are visible near the neck and head region which have a smoother surface. Therefore, SFM with this technique fails to perform well with a smoother surface. From this, we can guess that the texture impacts on the SFM performance. To find the textural differences, we can study the Grey Level Co-Occurrence

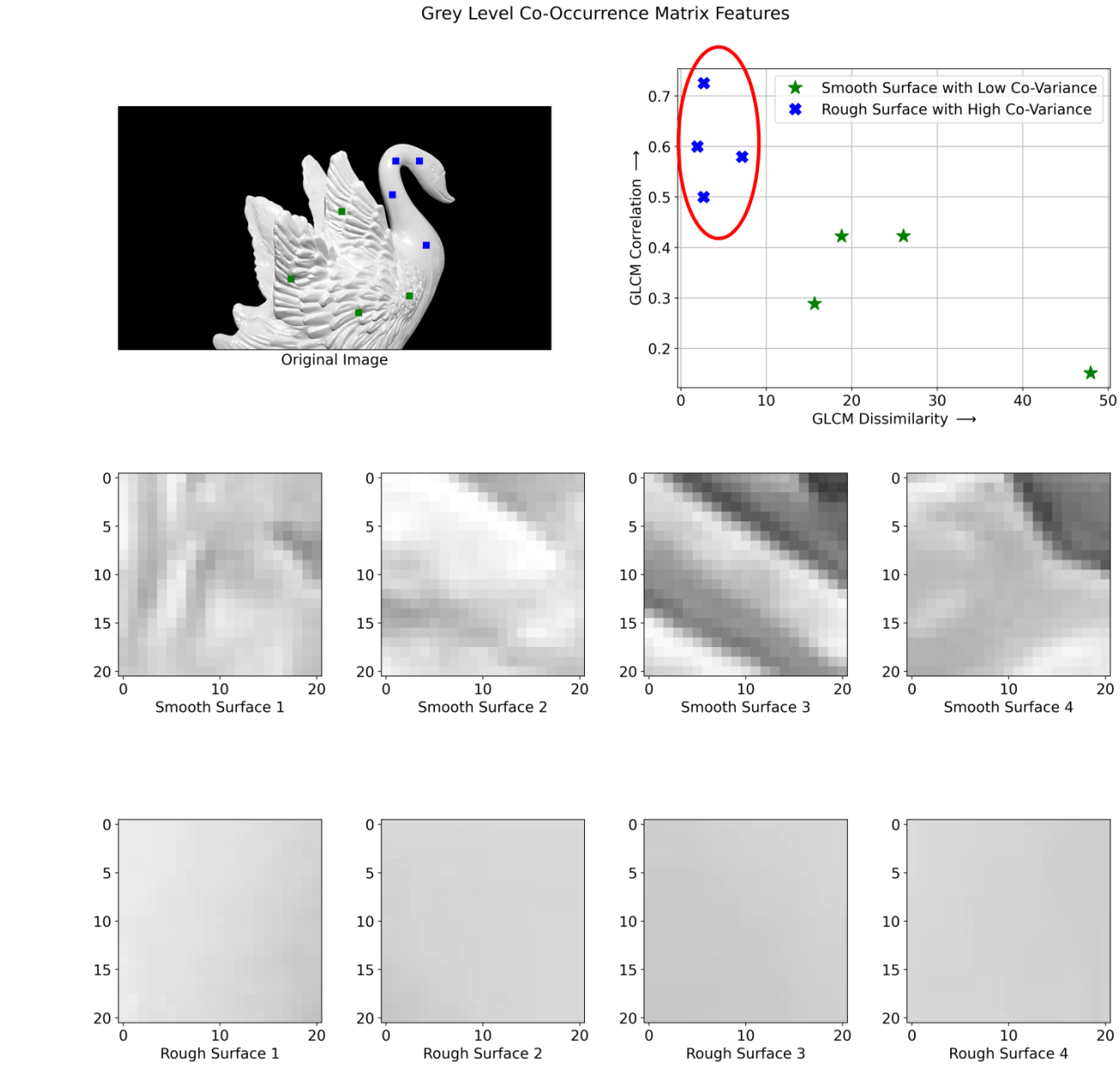

Fig. 4. Random points of two categories, smooth and rough surfaces, have been studied from the raw image with GLCM and the graph for Dissimilarity v/s Correlation has been plotted where we can see that blue points encircled by red mark are the regions where SFM causes missing pixel problem where SFM fails to compute depth map of the surrounding pixel

Table 1
Comparison Metrics for GLCM Matrix

Metric	Rough				Smooth			
	Point 1	Point 2	Point 3	Point 4	Point 1	Point 2	Point 3	Point 4
Dissimilarity	15.651	18.827	47.940	26.062	7.181	1.943	2.708	2.678
Correlation	0.288	0.422	0.151	0.422	0.579	0.599	0.725	0.499
Contrast	395.758	655.357	3530.488	1508.318	57.848	8.014	9.482	9.684
Homogeneity	0.0626	0.0398	0.0135	0.0569	0.0330	0.4428	0.2361	0.2373
ASM	0.00174	0.00177	0.00155	0.00191	0.00605	0.03324	0.01318	0.01442
Energy	0.0417	0.0420	0.0394	0.0437	0.0778	0.1823	0.1148	0.1200

Matrix (GLCM). For this purpose, we have considered 20x20 pixel points from different regions of the perspective projection of the 3D reconstructed image. Figure 4 shows the points that we have considered for the texture comparison. Green dots belong to regions from denser cloud points and blue dots belongs to regions from missing pixels. From fig. 4, for GLCM Dissimilarity v/s GLCM Correlation sub-plot, the green dots are highly dissimilar whereas blue dots scatter within a small range and are highly similar. As the points are highly similar, the algorithm performs badly to detect the differences in the pixels and therefore the depth map generation for those points fails and are hence excluded from the point cloud matrix. This makes it clear that SFM performs poorly for highly similar points, that is, a low variance of co-occurrence matrix and vice versa.

For GLCM histogram P with, μ being GLCM mean and σ being the intensity variation , dissimilarity metric is given by,

$$Dissimilarity = \sum_{i,j=0}^{levels-1} P_{i,j} |i-j| \quad (21)$$

Correlation is given by,

$$Correlation = \sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}}\right] \quad (22)$$

The GLCM matrix properties have been recorded in table 1. From this, we can see that the dissimilarity score for rough surface is over 10 for each and it is lower than 10 for smooth surfaces. Similarly, correlation is lower than 0.5 for the rough surface and higher than that for smoother surfaces. Based on this conditions, a decision boundary could also be established between the rougher and smoother surface to detected which points could be excluded from SFM procedure as an early detection machine learning algorithm and images with a higher density of such points, the images could be skipped from the feeding into the model which can ultimately reduce the processing cost.

Contrast is given by,

$$Contrast = \sum_{i,j=0}^{levels-1} P_{i,j} (i-j)^2 \quad (23)$$

The contrast for rough surface was found to be in scale of hundreds where smooth were found within a range of tens and ones. This could be another prominent clustering feature with linear decision boundary.

C. Introduction to Pointless SFM to Improve Missing Pixel Problem for Low Variance GLCM Matrix

To reduce this low textural variance-based missing pixel problem, a newer technique is introduced called Pointless SFM. The ideology behind this is to introduce 3D curve refining camera positions to construct the point cloud. The model works on bundle adjustment method and hence requires an initial estimate of the structure. For this purpose, two kinds of cameras are used, one for initialization and the other for ground truth. This has significantly low error rates, however, requires a higher setup cost ultimately reducing feasibility. Development of Pointless SFM to minimize cost is very essential to ensure feasibility. Else missing pixel problem would require state-of-the-art technology to produce a high-density point cloud.

5. Conclusion

Structure from Motion method for 3D mapping from 3D images captured from different angles has become a popular technique which mimics left and right eye coordination to compute depth map. SFM techniques first compare all images to find the closest pairs, and then the images are overlapped, the differences in pixels are estimated to calculate a depth map for the image pair and finally, the depth maps are mapped to form a 3D point cloud which could be visualized in perspective, fisheye or spherical camera projection schemes. Through this article, we have reviewed recent advancements in the SFM technology and recreated the process to test them. We have captured images of a test subject from different angles, masked background and performed SFM techniques to reconstruct a 3D model for the object. In the process, we have noticed that the rough surfaces are reconstructed well however glossy or smooth surfaces have missing pixels in the 3D point cloud. To test this, we have used GLCM matrix and its properties. We have observed that for the given image, dissimilarity of rough and smooth surface are higher and lower than 10 respectively and correlation for rough and smooth surface are lower and higher than 0.5 respectively. Based on this, we can cluster the two classes to create a decision boundary to facilitate machine learning as a means of a prediction model to filter out images from the dataset which won't help much in SFM technique and ultimately reduce computational cost. The missing pixel problem for glossier surface has been reduced with a newly introduced technique called Pointless SFM which is still in preliminary phases and requires high setup cost, the model could further be made feasible by reducing the complexity of the setup.

Acknowledgement

The authors would like to thank Spiraldevs Automation Industries Pvt. Ltd. West Bengal, India - 733123 and CubicX, West Bengal, India 700070. The work is a part of The Gyanam Project which is a joint venture between the two companies to promote development of scientific community.

References

[1] Zenil, H, Hernandez-Quiroz, F. On the possible computational powerof the human mind. In: Worldviews, Science and Us: Philosophy andComplexity. World Scientific; 2007, p. 315–337.

[2] Azevedo, FA, Carvalho, LR, Grinberg, LT, Farfel, JM, Ferretti, RE,Leite, RE, et al. Equal numbers of neuronal and nonneuronal cells makethe human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology 2009;513(5):532–541.

[3] Nousias, S, Lourakis, M, Bergeles, C. Large-scale, metric structure frommotion for unordered light fields. In: Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. 2019, p. 3292–3301.

[4] Banerjee, Saikat, Sudhir Kumar Chaturvedi, and Surya Prakash Tiwari. "Development of Speed Up Robust Feature Algorithm for aerial image feature extraction." INCAS Bulletin 11, no. 4 (2019): 49-60.

[5] Kato, T, Shimizu, I, Pajdla, T. Selecting image pairs for sfm by introducing jaccard similarity. IPSJ Transactions on Computer Vision and Applications 2017;9(1):12.
[6] Likas, A, Vlassis, N, Verbeek, JJ. The global k-means clustering algorithm. Pattern recognition 2003;36(2):451–461.
[7] Griffiths, D, Burningham, H. Comparison of pre-and self-calibratedcamera calibration models for uas-derived nadir imagery for a sfm application. Progress in Physical Geography: Earth and Environment2019;43(2):215–235.
[8] Adhikary S., Ghosh R., Ghosh A. (2021) Gait Abnormality Detection without Clinical Intervention Using Wearable Sensors and Machine Learning. In: Muthukumar P., Sarkar D.K., De D., De C.K. (eds) Innovations in Sustainable Energy and Technology. Advances in Sustainability Science and Technology. Springer, Singapore.
[9] Nurutdinova, I, Fitzgibbon, A. Towards pointless structure from motion:3d reconstruction and camera parameters from general 3d curves. In:2015 IEEE International Conference on Computer Vision (ICCV). 2015, pp. 2363–2371.
[10] Holynski, A, Geraghty, D, Frahm, JM, Sweeney, C, Szeliski, R. Reducing drift in structure from motion using extended features.
[11] Davidson, P, Mansour, M, Stepanov, O, Pich´e, R. Depth estima-tion from motion parallax: Experimental evaluation. In: 2019 26th SaintPetersburg International Conference on Integrated Navigation Systems(ICINS). IEEE; 2019, p. 1–5.
[12] Basu, S, Chaturvedi, SK, Banerjee, S, Adhikary, S. An analytical review on electric propulsion system for space satellites. INCAS Bulletin2020;12(4):3–11
[13] Wang, P, Sang, X, Yu, X, Gao, X, Yan, B, Liu, B, et al. Demonstration of a low-crosstalk super multi-view light field display with natural depth cues and smooth motion parallax. Optics Express 2019;27(23):34442–34453.
[14] Hataji, Y, Kuroshima, H, Fujita, K. Dynamic corridor illusion in pi-geons: Humanlike pictorial cue precedence over motion parallax cue insize perception. i-Perception 2020;11(2):2041669520911408.
[15] Liu, L, Zhang, T, Leighton, B, Zhao, L, Huang, S, Dissanayake, G. Robust global structure from motion pipeline with parallax on manifold bundle adjustment and initialization. IEEE Robotics and Automation Letters 2019;4(2):2164–2171.
[16] Adhikary S., Chaturvedi S., Chaturvedi S.K., Banerjee S. (2021) COVID-19 Spreading Prediction and Impact Analysis by Using Artificial Intelligence for Sustainable Global Health Assessment.
[17] Siddiqui N.A., Bahukhandi K.D., Tauseef S.M., Koranga N. (eds) Advances in Environment Engineering and Management. Springer Proceedings in Earth and Environmental Sciences. Springer, Cham.
[18] Klodt, M, Vedaldi, A. Supervising the new with the old: learning sfm from sfm. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 698–713.
[19] Cai, S, Zhou, S, Xu, C, Gao, Q. Dense motion estimation of particle images via a convolutional neural network. Experiments in Fluids2019;60(4):73.
[20] Yan, S, Pen, Y, Lai, S, Liu, Y, Zhang, M. Image retrieval for structure-from-motion via graph convolutional network.
[21] Lucieer, A, Jong, SMd, Turner, D. Mapping landslide displacements using structure from motion (sfm) and image correlation of multi-temporarily photography. Progress in Physical Geography 2014;38(1):97–116.
[22] Kong, C, Lucey, S. Deep non-rigid structure from motion. In: Proceedings of the IEEE International Conference on Computer Vision. 2019, pp. 1558–1567.
[23] Yu, H, Chen, X, Shi, H, Chen, T, Huang, TS, Sun, S. Motion pyramid networks for accurate and efficient cardiac motion estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2020, pp. 436–446.
[24] Adhikary S., Chaturvedi S.K., Banerjee S., Basu S. (2021) Dependence of Physiochemical Features on Marine Chlorophyll Analysis with Learning Techniques. In: Siddiqui N.A., Bahukhandi K.D., Tauseef S.M., Koranga N. (eds) Advances in Environment Engineering and Management. Springer Proceedings in Earth and Environmental Sciences. Springer, Cham.
[25] Skarlatos, D, Kiparissi, S. Comparison of laser scanning, photogram-metry and sfm-mvs pipeline applied in structures and artificial surfaces. International Society for Photogrammetry and Remote Sensing Congress, 2012.
[26] Ruggles, S, Clark, J, Franke, KW, Wolfe, D, Reimschiissel, B, Martin, RA, et al. Comparison of sfm computer vision point clouds of a land-slide derived from multiple small uav platforms and sensors to a tls-based model. Journal of Unmanned Vehicle Systems 2016;4(4):246–265.
[27] Chesley, J, Leier, A, White, S, Torres, R. Using unmanned aerial vehicles and structure-from-motion photogrammetry to characterize sed-imentary outcrops: An example from the morrison formation, utah, usa. Sedimentary Geology 2017;354:1–8.
[28] Chaturvedi, SK, Banerjee, S, Basu, S, Yadav, M, Adhikary, S. Mathematical review of the attitude control mechanism for a spacecraft. INCAS Bulletin 2020;12(3):33–48
[29] Yang, C, Chen, J, Xia, C, Liu, J, Su, G. A sfm-based sparse todense 3d face reconstruction method robust to feature tracking errors. In:2013 IEEE International Conference on Image Processing. IEEE; 2013, pp. 3617–3621.
[30] Woodget, AS, Austrums, R. Subaerial gravel size measurement using to-pographic data derived from a uav-sfm approach. Earth Surface Processesand Landforms 2017;42(9):1434–1443.
[31] Gao, CC, Hui, XW. Glcm-based texture feature extraction. Computer Systems & Applications 2010;6(048).
[32] Chen, ZW, Chiang, CC, Hsieh, ZT. Extending 3d lucas–kanade tracking with adaptive templates for head pose estimation. Machine Vision and Applications 2010;21(6):889–903.
[33] Javidnia, H, Corcoran, P. Accurate depth map estimation from small motions. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017, p. 2453–2461.