Deep Learning-Based Filtering of Images for 3D Reconstruction of Heritage Sites

Ramesh Ashok Tabib, Sujaykumar Kulkarni, Abhay Kagalkar, Vaishnavi Hurakadli, Abhijeet Ganapule, Rohan Raju Dhanakshirur, and Uma Mudenagudi

Abstract In this chapter, we propose a deep learning-based pipeline for filtering of internet-sourced images towards 3D reconstruction of heritage sites. The 3D reconstruction of heritage sites facilitates creation of virtual walk-through, digital museum and augmented reality. Using internet-sourced images for 3D reconstruction of heritage sites is challenging, as these images may contain blur, text, occlusion, shadow and many other noises. We propose to include pruning and selection of images in the pipeline to select a suitable set of images for 3D reconstruction. We propose a method for pruning of images using learning-based classification models to eliminate the contribution of unwanted images in 3D reconstruction. We also propose a method to select a suitable set of images using a combination of mean-shift and hierarchical clustering algorithms. We demonstrate the proposed pipeline by generating various 3D models of cultural heritage sites.

Keywords Deep learning · 3D reconstruction · Pruning · Filtering · Classification

1 Introduction

In this chapter, we propose a deep learning-based filtering pipeline to address the issues in 3D reconstruction of heritage sites using internet-sourced images. With time and climatic changes, the texture, shape and colour of monuments fade or lose information. Monuments are ruined due to attacks (during war) and natural calamities. In order to preserve our cultural heritage and pass it on to the next generation, there is a need to store all information in digital format. One of the effective ways of storing information on heritage sites is through 3D models. 3D reconstruction using images collected from the internet is challenging due to

R. A. Tabib · S. Kulkarni · A. Kagalkar · V. Hurakadli · A. Ganapule · R. R. Dhanakshirur (✉) · U. Mudenagudi
KLE Technological University, Hubballi, Karnataka, India
e-mail: ramesh_t@kletech.ac.in

J. Mukhopadhyay et al. (eds.), *Digital Techniques for Heritage Presentation and Preservation*, https://doi.org/10.1007/978-3-030-57907-4_8

varying captured conditions and with different sensors. Most of the image-based reconstructions for the generation of a detailed and informative 3D model rely on the images chosen [1]. The input images to the reconstruction algorithm may contain artefacts like occlusion and shadow. These artefacts influence 3D reconstruction and result in distortion of shape and texture in reconstructed 3D models.

Most of the works in the literature on 3D reconstruction carry out pre-processing of input data towards better reconstruction [2–4]. The authors in [2] propose enhancement of images before 3D reconstruction and use the tone-mapping approach with Contrast Limited Adaptive Histogram Equalization (CLAHE). This results in amplification of local contrast adaptively and prevents amplification of local noise. The authors in [3] propose colour balancing, denoising of image and enhancement of raw images using adaptive median filters before 3D reconstruction. The authors in [4] use Semi-Global Matching (SGM) as an image matching technique, which is applied to Unmanned Aerial Vehicle (UAV) images to generate dense point cloud. These methods find challenges when applied to images with text, blur, occlusion and shadow. The following are proposed to address these challenges:

- We propose a deep-learning pipeline for filtering internet-sourced images towards better 3D reconstruction.
- We propose to prune internet-sourced images to eliminate unwanted images towards better 3D reconstruction.
- We propose to select a suitable set of images for a given query image by combining mean-shift and hierarchical clustering algorithms towards 3D reconstruction.
- We demonstrate our results using internet-sourced images and compare with existing reconstruction methods.

In Sect. 2, we discuss the proposed pipeline for filtering of images towards 3D reconstruction. In Sect. 3, we demonstrate the results of the proposed pipeline and its effect on 3D reconstruction and conclude in Sect. 4.

2 Filtering of Images Towards 3D Reconstruction

In this section, we discuss the proposed learning-based pipeline for filtering of internet-sourced images towards 3D reconstruction. The proposed pipeline includes pruning, selection of images and 3D reconstruction modules as shown in Fig. 1. Internet-sourced images with blur, shadow, text and occlusion are eliminated during the pruning process. A subset of filtered images is further selected for 3D reconstruction.

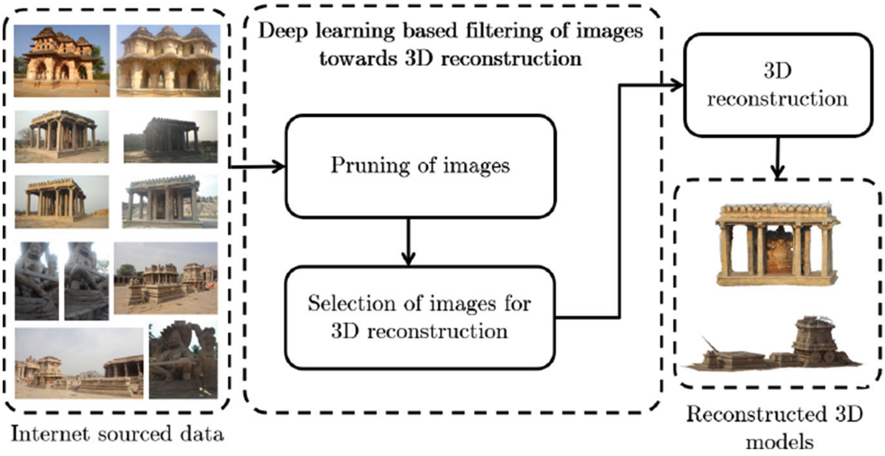

Deep learning based filtering of images towards 3D reconstruction
Pruning of images
Selection of images for 3D reconstruction
3D reconstruction
Internet sourced data
Reconstructed 3D models

Fig. 1 Pipeline for filtering of images towards 3D reconstruction

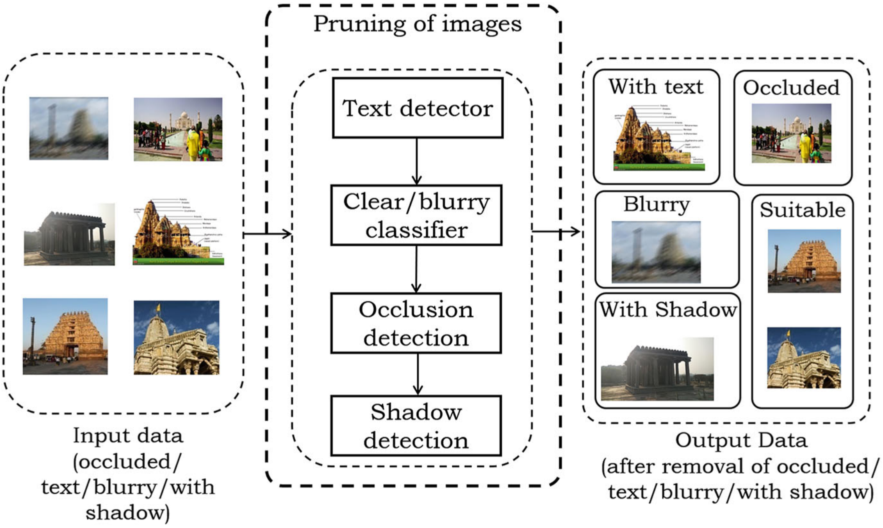


Fig. 2 Pruning of images

2.1 Pruning of Images

Internet-sourced images with text, blur, occlusion and shadow are input to pruning module as shown in Fig. 2. Internet-sourced images with text are filtered using a text detection algorithm (Tesseract Optical Character Recognition [OCR]) [5].

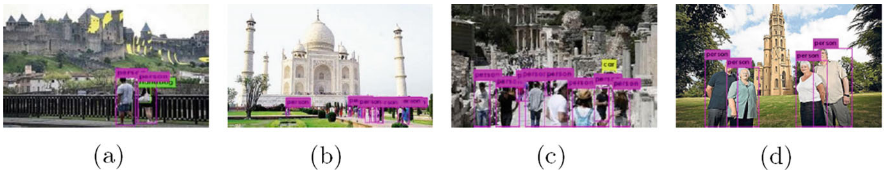
Blur detection in images using traditional methods is computationally expensive [6]. In order to reduce the complexity of blur detection, we use a binary classifier to classify the input images into blur and non-blur images. This classification includes feature extraction using stacked autoencoders [7, 8] and using the features as input for the binary classifier. The encoder consists of 64 × 64 × 3 nodes as input, two intermediate stacked layers with 1024 and 512 nodes and 64 nodes as output. The input data is encoded to 64 nodes. These 64 nodes are decoded with the intermediate stacked layer of 512 and 1024 nodes to output 64 × 64 × 3 nodes. The decoder reconstructs the input image. Initially, the stacked autoencoder is trained to extract the features, and then the decoder is replaced with a binary classification layer for classifying blur and non-blur images. The non-blur images are given as input to the occlusion detection module.

Most of the internet-sourced images with respect to cultural heritage comprise occluded objects in front of the monuments, which might affect the 3D reconstruction. Thus, we propose to detect occluded portions in order to eliminate these images. You Only Look Once (YOLO) [9] is used to generate bounding boxes on each object over input images. Our proposed algorithm computes the area of the bounding box, and depending on the effect of the area, the percentage of occlusion is calculated. If the percentage is greater than the particular threshold (heuristically we set the threshold to 20%), the algorithm discards the images. If there are overlapping multiple bounding boxes, we find the union of all multiple bounding boxes given as

$$\cup_{i=1}^{N} A = \{x \in U :\ni i \in \{1, 2, 3 \ldots N\}, x \in A_i\}. \tag{1}$$

In Fig. 3, we observe that the threshold does not affect the 3D model if occlusion is less than 20%. In Fig. 3a and b, we show that the occlusion percentage is small, and by experiment it is observed that there is no significant effect on 3D models. In Fig. 3c and d, we see that the occlusion percentage is greater than 20% and covers the major part of the monument area, and by experiment it is observed that there is significant effect on 3D models. If the occlusion percentage is greater than 20%, then the 3D model contains hole in the occlusion area resulting in an incomplete 3D model (see Fig. 8).

The images with shadow usually affect the texture of reconstructed 3D models, which is an open problem to be solved. Some of the shadow detection algorithms in the literature are detailed in [10, 11]. However, these techniques do not provide desirable 3D models. Thus, we propose to use a convolutional autoencoder [12] as shown in Fig. 4 to eliminate the shadow images by classifying the images into shadow and non-shadow images. The convolutional autoencoder has 64 × 64 × 3 size



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

Fig. 3 (**a**, **b**) Images retained and (**c**, **d**) images discarded on set threshold. (**a**) 7% occlusion. (**b**) 10% occlusion. (**c**) 28% occlusion. (**d**) 36% occlusion

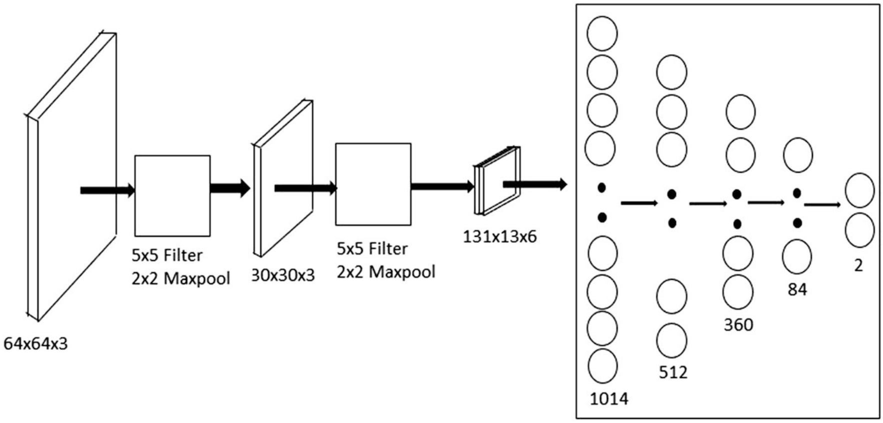


Fig. 4 Classification of shadow and non-shadow images

input layer, which is convolved by a 5 × 5 kernel with three channels. The convolved output is max-pooled with kernel size 2 × 2 and stride 2. This max-pooled output of size 30 × 30 × 3 is convolved by a 5 × 5 kernel with six channels. The convolved output is max-pooled with kernel size 2 × 2 and stride 2. This max-pooled output of size 13 × 13 × 6 is flattened and provided as input to the fully connected network with three hidden layers of size 512, 360 and 84 nodes, respectively. The last layer, i.e. fully connected output, is fed to the binary classifier to classify the images into shadow and non-shadow classes.

2.2 Selection of Images for 3D Reconstruction

The filtered images are processed to choose appropriate images as shown in Fig. 5 for 3D reconstruction. The autoencoder is trained to extract features that are mapped to the latent space. We use stacked autoencoder in Sect. 2.1 to represent data as latent points towards clustering. The two types of clustering algorithms considered are Meanshift [1] and Hierarchical [13]. We use content-based image retrieval (CBIR) [14] technique with considered clustering algorithms and compare the clusters with the input query image. The query image is obtained from curator or user. The intersection of obtained image clusters from mean-shift and hierarchical algorithms is considered for 3D reconstruction as shown in Fig. 5.

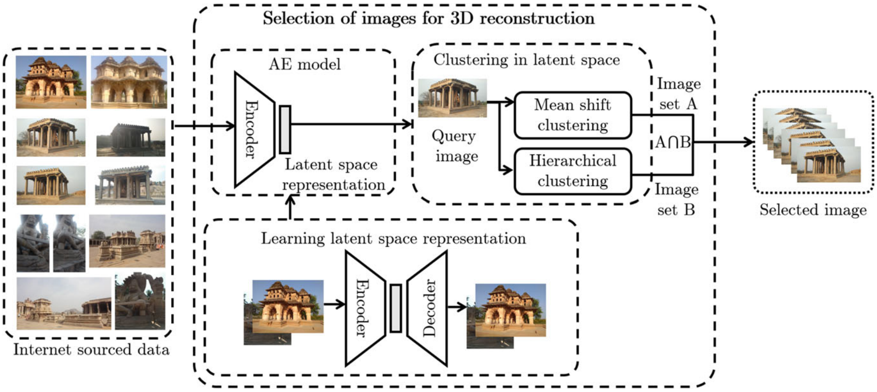


Fig. 5 Selection of images for 3D reconstruction

3 Results and Discussions

The implementation is carried out in the Intel Xeon i5 processor, 64 GB RAM, Nvidia Quadro K5000 graphic processor. In this section, we demonstrate the results of our pipeline and compare the results with existing 3D reconstruction techniques. We used openMVG and openMVS[15] pipeline for 3D reconstruction.

We used 300,000 internet-sourced images as dataset, which comprises 60 heritage sites in India. Approximately 150,000 were discarded by our pipeline. We used 50,000 synthetically generated blurred and real blurred images for training stacked autoencoder. We obtained 95.432% test accuracy and 97.213% cross-validation accuracy from Stacked Autoencoder used for blur and non-blur classification (Table 1). We used standard shadow detection dataset [10] for training convolutional autoencoder, and we obtained 96.156% cross-validation accuracy and 94.591% testing accuracy from convolutional autoencoder used for shadow and non-shadow classification. We used openMVG and openMVS pipeline for 3D reconstruction. We performed subjective analysis for the obtained results with 100 volunteers and the ratings (rating between 1 and 5, 1 being the least and 5 being the highest) are recorded, as shown in Table 2.

Table 1 Results of the stacked autoencoder and convolutional autoencoder

Algorithm	Learning rate	No. of epochs	Time	Accuracy
Stacked autoencoder	0.001	1200	120 h	Cross-validation: 97.213%
				Testing: 95.432%
Convolutional autoencoder	0.0001	1500	216 h	Cross-validation: 96.156%
				Testing: 94.591%

Table 2 Subjective quality analysis for the obtained results with 100 volunteers and the corresponding ratings (rating between 1 and 5, 1 being the least and 5 being the highest)

Heritage site	Figure reference	Overall rating	Deviation	Spread
Pattadakal	Fig. 6a	3.23	0.83	1.9
	Fig. 6b	3.79	0.18	1.5
	Fig. 7a	3.16	0.23	1.4
	Fig. 7b	4.01	0.14	1
	Fig. 8a	2.38	0.15	0.7
	Fig. 8b	3.47	0.19	0.9
Hampi	Fig. 9a	1.46	0.21	1.35
	Fig. 9b	2.69	0.18	0.65
Pattadakal	Fig. 10a	2.68	0.20	1.4
	Fig. 10b	3.9	0.37	2.2
Sasivekal Ganpati	Fig. 10c	3.9	0.17	0.75
	Fig. 10d	4.2	0.2	1
Stone Chariot	Fig. 10e	3.4	0.16	0.8
	Fig. 10f	3.9	0.16	0.9

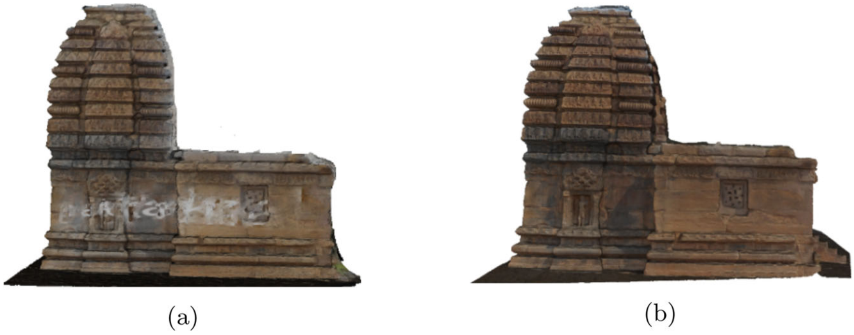


Fig. 6 (**a**) 3D model of Pattadakal obtained from images containing text. (**b**) 3D model of Pattadakal obtained after discarding images with text

In Figs. 6, 7, 8 and 9, we show the results of the individual stages of the proposed pipeline with the 3D reconstruction of the Pattadakal temple situated in Badami (Tq), Karnataka, India, and Stone Chariot, Hampi, Bellary, Karnataka, India. Figure 6a shows the 3D model reconstructed using 100 sample images of Pattadakal from the dataset among which 36 images contain text. Figure 6b corresponds to the 3D model reconstructed using 64 images after eliminating images with text. Figure 7a shows the 3D model reconstructed from 200 sample images of Pattadakal from the dataset among which 42 images contain blur. Figure 7b shows the 3D model reconstructed with 158 images after eliminating images with blur. Similarly, Figs. 8a and 9a correspond to the 3D models reconstructed using 200 and 150 images among which 14 and 27 images are occluded and contain shadow, respectively. Figures 8b and 9b represent the 3D models reconstructed using 186 and

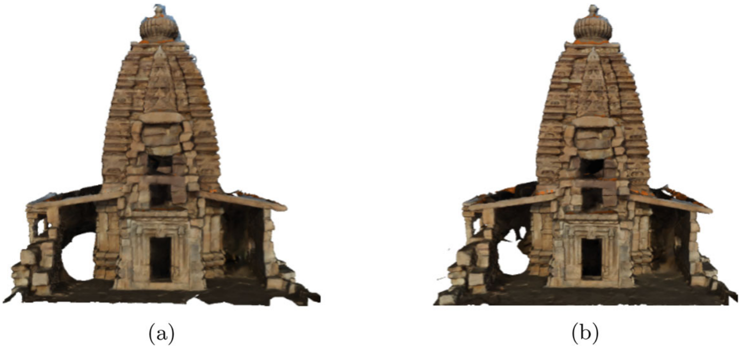


Fig. 7 (**a**) 3D model of Pattadakal obtained from blurred images. (**b**) 3D model of Pattadakal obtained after the removal of blurred images

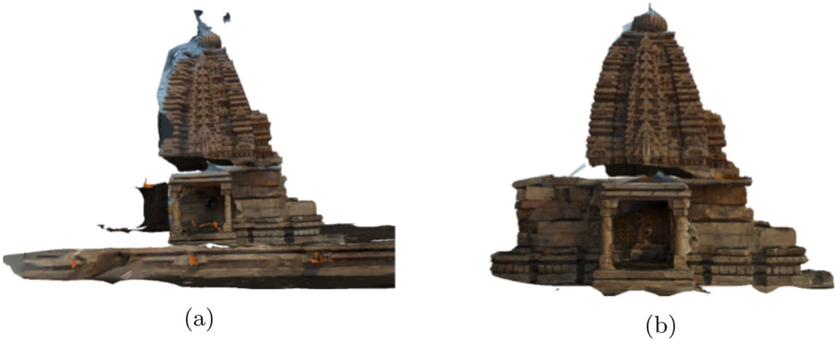


Fig. 8 (**a**) 3D model of Pattadakal obtained from images containing occlusion. (**b**) 3D model of Pattadakal obtained after discarding images containing occlusion

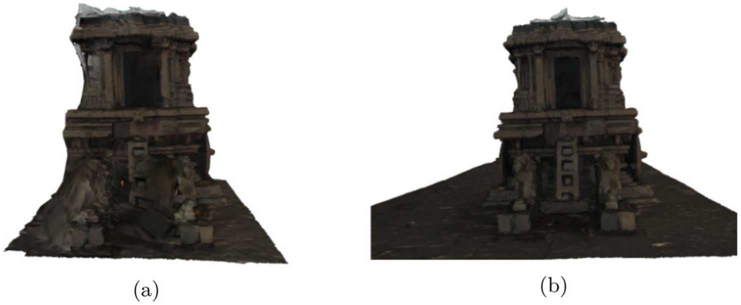


Fig. 9 (**a**) 3D model of Hampi obtained from images containing shadow. (**b**) 3D model of Hampi obtained after removing images containing shadow

123 images after eliminating occluded and shadow images, respectively. Figure 10 shows a comparison of 3D models reconstructed with and without the proposed filtering pipeline.

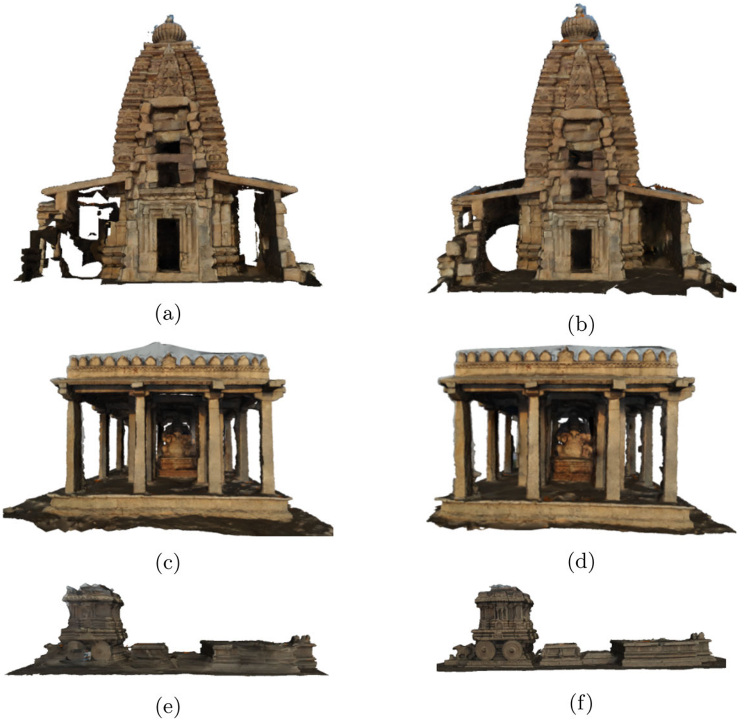


Fig. 10 (**a**, **c**, **e**) 3D models obtained without applying filtering pipeline and (**b**, **d**, **f**) 3D models obtained after applying filtering pipeline

4 Conclusions

In this chapter, we have proposed a deep learning-based filtering pipeline for processing internet-sourced images of heritage sites for better 3D reconstruction. 3D reconstruction of heritage sites, using images collected from the internet, is challenging since images may contain blur, text, occlusion and shadow artefacts with reported pre-processing methods. To improve the results for the internet-sourced images, we have proposed a pipeline with pruning and selection modules in order to select a suitable set of images for 3D reconstruction. We have also proposed a method to select the suitable set of images using a combination of mean-shift and hierarchical clustering algorithms. We have demonstrated the results of the proposed pipeline by generating various 3D models of cultural heritage sites and have performed subjective qualitative analysis on the obtained results.

References

1. Xiao, C., Liu, M.: Efficient mean-shift clustering using Gaussian KDTree. Comput. Graph. Forum **29**, 2065–2073 (2010)
2. Aldeeb, N., Hellwich, O.: Reconstructing textureless objects – image enhancement for 3D reconstruction of weakly-textured surfaces. In: Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 5: VISAPP, ISBN 978-989-758-290-5, pp. 572–580 (2018). https://doi.org/10.5220/0006628805720580
3. Ballabeni, A., Apollonio, F., Gaiani, M., Remondino, F.: Advances in image pre-processing to improve automated 3d reconstruction. In: ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2015). XL-5/W4.315-323. https://doi.org/10.5194/isprsarchives-XL-5-W4-315-2015
4. Alidoost, Fatemeh & Arefi, Hossein. (2015). An image-based technique for 3D building reconstruction using multi-view UAV images. In: ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XL-1-W5, pp. 43–46. https://doi.org/10.5194/isprsarchives-XL-1-W5-43-2015
5. Smith, R.: An overview of the Tesseract OCR engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), September, vol. 2, pp. 629–633 (2007)
6. Landge, R.Y., Sharma, R.: Blur detection methods for digital images – a survey. Int. J. Comput. Appl. Technol. Res. **2**, 495–498 (2013). https://doi.org/10.7753/IJCATR0204.1019
7. Matsumoto, K., Tajima, Y., Saito, R., Nakata, M., Sato, H., Kovacs, T., Takadama, K.: Learning classifier system with deep autoencoder. In: 2016 IEEE Congress on Evolutionary Computation (CEC), July, pp. 4739–4746 (2016)
8. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2015). CoRR, abs/1506.02640
10. Al-Najdawi, N., Bez, H., Singhai, J., Edirisinghe, E.: A survey of cast shadow detection algorithms. Pattern Recogn. Lett. **33**, 752–764 (2012). https://doi.org/10.1016/j.patrec.2011.12.013
11. Sharma, P., Sharma, R.: Shadow detection and its removal in images: a review. Res. Cell **17**, 2229–6913 (2016)
12. Turchenko, V., Chalmers, E., Luczak, A.: A deep convolutional auto-encoder with pooling–unpooling layers in caffe (2016). CoRR, abs/1701.04949
13. Nazari, Z., Kang, D., Asharif, M.R., Sung, Y., Ogawa, S.: A new hierarchical clustering algorithm. In: 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), November, pp. 148–152 (2015)
14. Beaudoin, J.E.: Content-based image retrieval methods and professional image users. J. Assoc. Inf. Sci. Technol. **67**(2), 350–365 (2016)
15. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: OpenMVG: Open multiple view geometry. RRPR@ICPR (2016)