Learning Based Image Selection for 3D Reconstruction of Heritage Sites

Ramesh Ashok Tabib(✉), Abhay Kagalkar, Abhijeet Ganapule, Ujwala Patil, and Uma Mudenagudi

KLE Technological University, Hubballi, India
ramesh_t@kletech.ac.in

Abstract. In this paper, we propose learning based pipeline with image clustering and image selection methods for 3D reconstruction of heritage site using cleaned internet sourced images. Cleaned internet sourced images means the images that do not contain an image with text, blur, occlusion, and shadow. 3D reconstruction of heritage sites is one of the emerging topics and is gaining importance as efforts are made to digitally preserve the heritage sites. 3D reconstruction using internet-sourced images is challenging as they often contain thousands of images taken from the same viewpoint. We propose to use autoencoders to extract robust features from images to cluster similar parts of heritage sites. We propose to use the image selection algorithm to select images from each cluster with the removal of redundant images. We demonstrate the proposed pipeline using available 3D reconstruction pipeline for a variety of heritage sites which contain one cluster to eight clusters and obtain better visual 3D reconstruction.

Keywords: Learning · Clustering · Image selection · 3D reconstruction

1 Introduction

In this paper, we propose learning based image clustering and selection methods for 3D reconstruction of heritage sites. We use cleaned images which do not have an image with watermark, blur, and occlusion as input to the proposed pipeline. When a particular search for a heritage site is given on the internet we get a few millions of images. For Example Hampi - a UNESCO world heritage site is spread over 16 sq mi and has about 1600 monuments. There are more than three million images on Google Image Search under the search label "Hampi". The search result contains various temples, fountains, stupas, paintings, sculptures etc. Most of these images are captured from hundreds or thousands of viewpoints and illumination conditions and may contain images with blur, occlusion, watermark etc. We use the method proposed in [11] to clean the images before giving into our proposed pipeline. This method proposed in [11] removes the images which contain blur, watermark, and occlusion. But even after the removal, the acquired

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 499–506, 2019.
https://doi.org/10.1007/978-3-030-34869-4_54

subset has redundant images which are not useful for 3D reconstruction and may affect the quality of 3D reconstruction as shown in Fig. 1. Typically heritage sites contain different parts like stupas (A hemispherical structure containing relics that is used as place of meditation), pillars, mandapas (It is a pillared outdoor hall), when 3D reconstruction of the whole space is carried without selecting images the reconstruction is not good due to a large number of outliers coming images with very near viewpoints which led to the selection of representative images for 3D reconstruction. For the selection process, we categorize the image into different parts of the heritage site and select a set of images from each part.

Most of the image clustering methods in literature talk about clustering as a tool to parallelize the sparse bundle adjustment step as including this step reduces the computational time. For large scale, 3D reconstruction efforts [1,8] are made to cluster the input image set with respect to viewpoints, but we propose to cluster the similar parts of the heritage site and a select subset of images for reconstruction. In [12] Thorsten Thormhlen et al., proposed an algorithm that selects the images having a large number of feature points by estimating the error of initial camera motion and object structure. In [3] Yasutaka Furukawa et al., proposed an approach for enabling existing multi-view stereo methods to operate on extremely large unstructured images and to remove an image if the conditions hold good even after the removal. By using these proposed approaches we propose to select images by removing redundant images that do not contribute to 3D reconstruction.

In this paper, we propose a pipeline with image clustering and selection modules to select images for 3D reconstruction. We summarize our contribution as follows:

- We propose learning based unsupervised image clustering to cluster similar-parts of the heritage site. An autoencoder is trained on the cleaned internet sourced images to extract robust features which are fed to a clustering algorithm to obtain image clusters containing similar parts of the heritage site.
- We propose an image selection module to get a subset of images from each cluster by removing redundant images. We use camera parameters and reprojection error factor in images to calculate the contributed value of the image for 3D reconstruction and by using a preset threshold we decide whether to retain or reject the image.
- We demonstrate the proposed pipeline using different heritage containing one to eight clusters and also compare with 3D models obtained using cleaned images.

In Sect. 2 we discuss the methodology of the proposed pipeline. In Sect. 3 we demonstrate the results of the proposed pipeline. In Sect. 4 we provide concluding remarks.

2 Learning Based Image Selection

We propose a pipeline as shown in Fig. 2, which consists of image clustering and selection modules. We assume our input images to be free from the text,

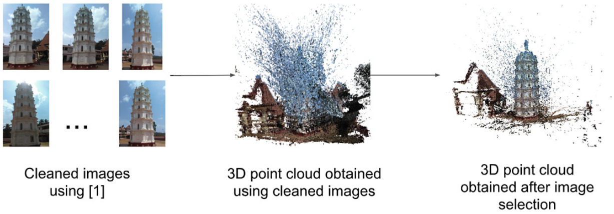


Fig. 1. Essence of image selection for 3D reconstruction

blur, occlusion etc. As it is evident from Fig. 1 even after the removal of these images the obtained point cloud from cleaned images is noisy and contain a large number of outliers. In the clustering module, images are clustered into respective similar-parts of the heritage site. Images from all clusters are given to the image selection algorithm, which removes redundant images. The selected set of images are then used to 3D reconstruct the heritage site. In what follows we explain in detail the image clustering and selection modules.

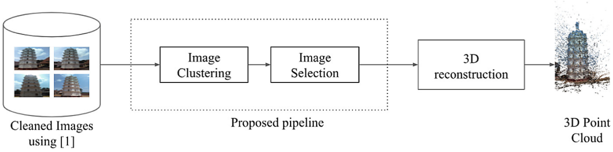


Fig. 2. Learning based image selection for 3D reconstruction

2.1 Image Clustering

Internet-sourced images contain all parts of the heritage site stored under a single label. Hence it makes clustering an important step before images are given to the image selection algorithm. In many clustering problems use handcrafted features like SIFT [5], SURF (Speeded-Up Robust Features) etc. But not all of the selected features are useful and relevant. In such a case choosing robust features leads to better performance. In unsupervised learning, class labels are not defined and choosing the right set of features becomes challenging because all features are not important. Redundant and irrelevant features may misguide clustering results. Hence we use autoencoders to extract robust features from input images.

Autoencoders [4,10,13] work by compressing the input into a latent-space representation and then reconstructing the output from this representation. It consists of an Encoder and a decoder as shown in Fig. 3. By pre-training an autoencoder network we except it to learn salient features of the input data. This feature vector is given as input to the clustering algorithm to obtain desired clusters.

We propose a network that consists of convolution layers stacked at the beginning to extract hierarchical features from the input images then flatten all units in the last convolution layer to form a vector followed by an up-sampled fully connected layer. Finally, it is down-sampled using subsequent fully connected layers with only 25 units which are called an embedded layer. The input image is thus transformed into a 25-dimensional feature vector as shown in Fig. 4 This obtained feature vector is given to the Mean-shift clustering algorithm [2] to obtain the clusters as shown in Fig. 5. In what follows we explain in detail the Image selection module.

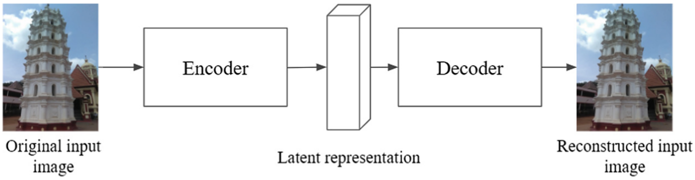


Fig. 3. Encoder-decoder network to learn robust features

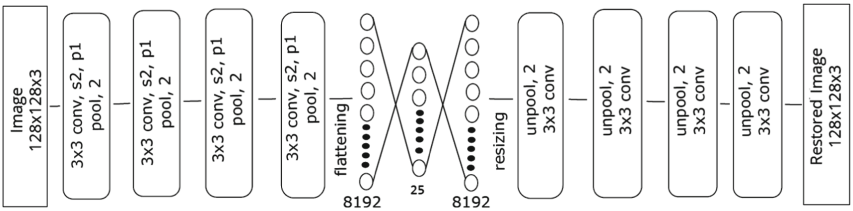


Fig. 4. Convolutional Autoencoder architecture trained for feature extraction in our proposed pipeline

2.2 Image Selection

Most of the image-based reconstruction for the generation of a detailed and informative 3D model relies on the images chosen. So image selection plays an important role in 3D reconstruction. Internet-sourced images are unstructured and redundant in nature. Many Multi-view Stereo algorithms aim to reconstruct a 3D model using all the images available [7]. Such an approach is not feasible as the number of images grows. Hence, it becomes an important step to select the right subset of images. After clustering, the clusters may contain redundant images that may not contribute to 3D reconstruction. So we use image selection to remove redundant images. The state of art Structure from Motion(SfM) [9] provides precise camera poses and 3D coordinates of the scene. A measure function proposed in [15] is used to estimate. The contributed value of two matching images to 3D reconstruction.

$$f_{jk} = fac\_angle_{jk} * fac\_error_{jk} \tag{1}$$

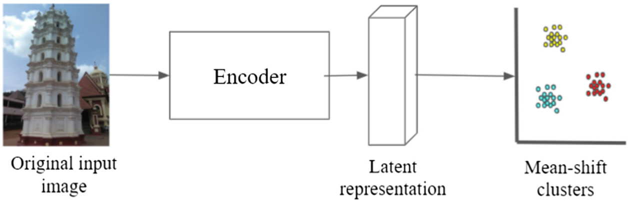


Fig. 5. Clustering using the features extracted from encoder

where f_{jk} represents the contributed value to 3D reconstruction of the combination of image j and image k. Let b be threshold and in our pipeline b is set to 7 heuristically. By using Algorithm 1 we remove the redundant images from acquired image subsets. The retained images are used for 3D reconstruction as shown in Fig. 6.

Algorithm 1. Image Selection

```
Input: Images $i_1 \ldots i_n$
for $i \leftarrow 1$ to $n$ do
    for $j \leftarrow 1$ to $n$ do
        $f_{ij} += fac\_angle_{ij} * fac\_error_{ij}$
for $k \leftarrow 1$ to $n$ do
    calculate contributed value without $k^{th}$ image;
    for $p \leftarrow 1$ to $n$ do
        for $q \leftarrow 1$ to $n$ do
            $f_{pq} += fac\_angle_{pq} * fac\_error_{pq}$
    if $f_{pq}/f_{ij}$ less than 0.97 then
        Retain $k^{th}$ image;
    else
        Remove $k^{th}$ image;
```

3 Results and Discussions

In this section, we present the results of our proposed pipeline. The proposed pipeline is implemented on NVIDIA DGX-1 server with 500 GB RAM, 32 GB NVIDIA Tesla V100 graphics processor. We used internet sourced and passive crowd-sourced images to evaluate our proposed pipeline. We use OpenSfM to compute depth maps and generate the 3D point cloud.

Note that a heritage site containing more number of clusters or no viewpoint in common then the clusters are reconstructed separately. For the demonstration of results, we use three heritage site datasets. Shanta Durga Temple, Bankapura

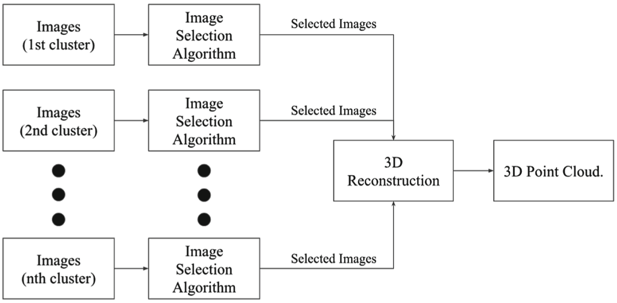
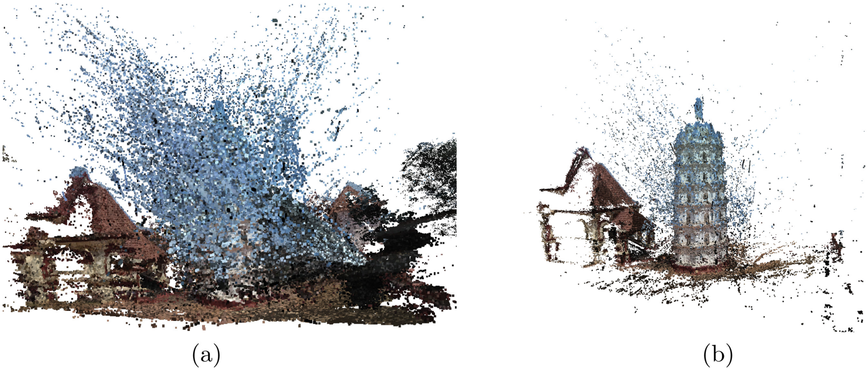


Fig. 6. Image selection pipeline

Table 1. Statistics

Dataset		Shantadurga temple	Bankapur Nageshwar	Pattadkal
No. of images	Input	611	703	236
	Cleaned	579	647	220
	After image selection	305	139	92
No. of clusters		2	1	2
No. of SFM points	Input	7,038,446	16,869,227	6,559,431
	Cleaned	5,879,238	10,788,004	4,401,547
	After image selection	3,007,136	3,033,418	1,544,580


(a)

(b)

Fig. 7. 3D point cloud of (a) Shanta Durga Deepa Stambha from input images (579 Images). (b) Shanta Durga Deepa Stambha with our pipeline (305 Images).

Nageshwar Temple, Pattadkal are chosen for the demonstration. Figure 7 is 3D point cloud of Shanta Durga temple situated in Ponda (Tq), Goa, India. Figure 8 is the 3D point cloud of Bankapura Nageshwara temple situated in Haveri (Dist.), Karnataka, India. Figure 9 is the 3D point cloud of Pattadkal temple situated in Bagalkot (Dist.), Karnataka, India. Figure 7a illustrates 3D point cloud generated using 579 images. After image selection, 274 images were rejected. Note that the image selection discarded nearly 47.32% of images. We can observe that

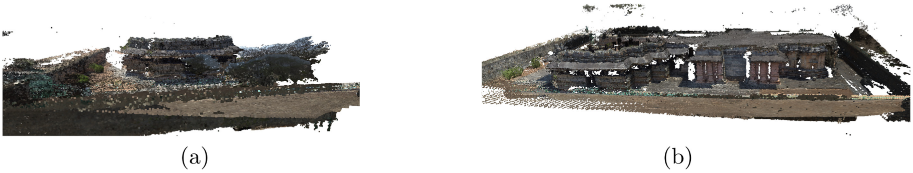


Fig. 8. 3D point cloud of (a) Bankapur Nageshwar Temple from original images (647 Images) (b) Bankapur Nageshwar Temple with our pipeline (139 Images).

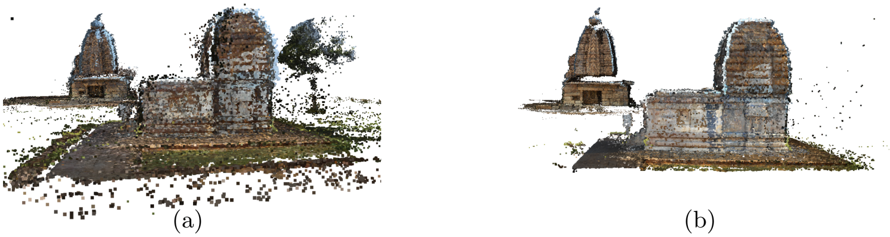


Fig. 9. 3D point cloud of (a) Pattadkal Temple from original images (220 Images) (b) Pattadkal Temple with our pipeline (92 Images).

the 3D point cloud is noisy. Figure 7b illustrates 3D point cloud using 305 images after Image Selection. The obtained point cloud is free from noise and clear. This result indicates that Image selection is useful.

Similarly Figs. 8a and 9a illustrates 3D point cloud generated using 647 and 220 images respectively. And Figure. 8b and 9b illustrates the 3D point cloud generated using 139 and 92 images respectively after Image selection. Table 1 provides more information about number of clusters obtained and SFM points in generated 3D point clouds.

4 Conclusions

In this paper, we have proposed learning based image clustering and selection methods for 3D reconstruction of heritage site using cleaned internet sourced images. Cleaned internet sourced images means the images that do not contain an image with text, blur, occlusion, and shadow. 3D reconstruction of heritage sites is one of the emerging topics and is gaining importance as efforts are made to digitally preserve the heritage sites. 3D reconstruction using internet-sourced images is challenging as they often contain thousands of images taken from the same viewpoint. We have proposed to use autoencoders to extract robust features from images to cluster similar parts of heritage sites. We have proposed to use the image selection algorithm to select images from each cluster with the removal of redundant images. We have demonstrated the proposed pipeline using available 3D reconstruction pipeline for a variety of heritage sites which contain one cluster to eight clusters and obtain better visual 3D reconstruction.

Acknowledgement. This project is partly carried out under Department of Science and Technology (DST) sponsored project Indian Heritage in Digital Saces (IHDS).

References

1. Agarwal, S., Snavely, N., Simon, I., Sietz, S.M., Szeliski, R.: Building Rome in a day. In: Twelfth IEEE International Conference on Computer Vision (ICCV 2009). IEEE, September 2009
2. Carreira-Perpiñán, M.Á.: A review of mean-shift algorithms for clustering. CoRR abs/1503.00687 (2015)
3. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1434–1441, June 2010
4. Jiang, Y., Yang, Z., Xu, Q., Cao, X., Huang, Q.: When to learn what: deep cognitive subspace clustering. In: Proceedings of the 26th ACM International Conference on Multimedia, MM 2018, pp. 718–726. ACM, New York (2018)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
6. Peng, X., Feng, J., Lu, J., Yau, W.Y., Yi, Z.: Cascade subspace clustering (2017). https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14442
7. Pons, J.P., Keriven, R., Faugeras, O.: Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. Int. J. Comput. Vision **72**(2), 179–193 (2007)
8. Sawyer, S.M., Ni, K., Bliss, N.T.: Cluster-based 3D reconstruction of aerial video. In: IEEE Conference on High Performance Ext Computing, pp. 1–6, September 2012
9. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comput. Vision **80**(2), 189–210 (2008)
10. Song, C., Liu, F., Huang, Y., Wang, L., Tan, T.: Auto-encoder based data clustering. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) CIARP 2013. LNCS, vol. 8258, pp. 117–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41822-8_15
11. Sujay, A., Vaishnavi, A., Tabib, R., Mudenagudi, U.: Deep learning-based filtering of images for 3D reconstruction of heritage sites. In: 2018 11th Indian Conference on Computer Vision, Graphics and Image Processing, November 2018
12. Thormählen, T., Broszio, H., Weissenfeld, A.: Keyframe selection for camera motion and structure estimation from multiple views. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 523–535. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24670-1_40
13. Xie, J., Girshick, R.B., Farhadi, A.: Unsupervised deep embedding for clustering analysis. CoRR abs/1511.06335 (2015). http://arxiv.org/abs/1511.06335
14. Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M.: Towards k-means-friendly spaces: simultaneous deep learning and clustering (2016). http://arxiv.org/abs/1610.04794
15. Yang, C., Zhou, F., Bai, X.: 3D reconstruction through measure based image selection. In: 2013 Ninth International Conference on Computational Intelligence and Security, pp. 377–381, December 2013