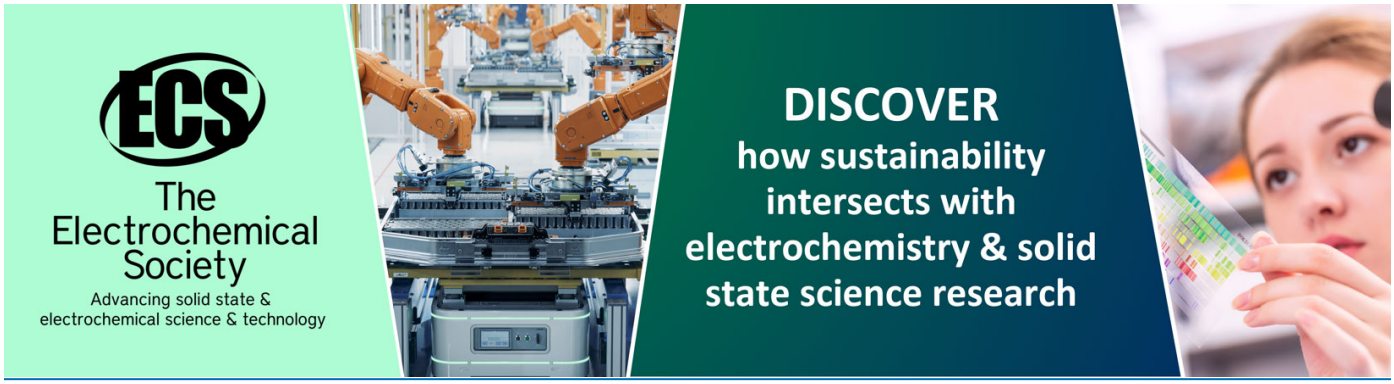
3D reconstruction using Structure From Motion (SFM) algorithm and Multi View Stereo (MVS) based on computer vision



3D reconstruction using Structure From Motion (SFM) algorithm and Multi View Stereo (MVS) based on computer vision

M Kholil*, I Ismanto and M N Fu'ad

Akademi Komunitas Negeri Putra Sang Fajar Blitar, East Java, Indonesia

*moch.kholil89@gmail.com

Abstract. The development of the Information and Computer Technology (ICT) sector, three-dimensional (3D) technology is also growing rapidly. Currently, the need to visualize 3D objects is widely used in animation and graphic applications, architecture, education, cultural recognition and Virtual Reality. 3D modeling of historic buildings has become a concern in recent years. 3D reconstruction is an attempt to document reconstruction or restoration if the building is destroyed. By using the 3D model reconstruction using Structure from Motion (SFM) and Multi View Stereo (MVS) algorithm based on Computer Vision, it is hoped that the results of this 3D modeling can be utilized as an effort to preserve 3D objects in the Penataran Temple cultural heritage area. This research was conducted by taking as many as 61 images of objects in the Blitar Penataran Temple area. The photos obtained were reconstructed into a 3D model using the Structure From Motion algorithm in the meshroom. This research a trial of the original image with a compressed image for reconstruction is used to compare the 3D reconstruction process from the two input data. From 61 images processed using the Structure Form Motion algorithm, 33 poses of camera pose and 3D points were improved, both original and compressed images. The number of iterations compresses 1.4% less than the original image and takes 43.53% faster than the original image.

1. Introduction

Law of the Republic of Indonesia Number 11 Year 2010 article 1 concerning Cultural Heritage explains that cultural preservation is a material cultural heritage in the form of cultural heritage objects on land and / or in water that needs to be preserved because they have important values for history, science, education, religion, and / or culture through the process of determination. In the Law of the Republic of Indonesia Number 11 Year 2010 also stated that what is referred to as preservation is a dynamic effort to maintain the existence of cultural heritage and its value by protecting, developing, and utilizing it.

The development of the Information and Computer Technology (ICT) sector, three-dimensional (3D) technology is also developing rapidly. At present, the need to visualize 3D objects is widely used in animation and graphics applications, architecture, education, cultural recognition and Virtual Reality [1]. 3D modeling of historic buildings has become a concern in recent years. 3D reconstruction is an attempt to document reconstruction or restoration if the building is destroyed.

Law No. 11 Article 5 section 4 also stipulates that the preservation of cultural heritage must be supported by documentation activities before activities that can cause a change in authenticity. The documentation is not only limited to knowing the geometry dimensions of cultural heritage, but also

related to how large changes in the dimensions of the geometry that occur within a certain time span. One method of documenting cultural heritage which is currently undergoing development is a three-dimensional modeling method. Utilization of documentation methods by making three-dimensional models of objects or areas of cultural heritage provides many advantages including documentation data that has the shape and dimensions of objects that are precise and easy to store. Therefore, currently making three-dimensional models for the benefit of the documentation of objects and areas of cultural heritage is very necessary in conservation activities so as to be able to maintain the elements of cultural works that are in a state quite complete in such a way that it is still able to provide a complete picture of cultural heritage exists and reflects the important values they contain [1].

The selection of the Penataran Temple area as a research site to model 3D objects in the area is based on the Penataran Temple, one of the grandest and most extensive temples in East Java, located on the western slope of Mount Kelud, north of the city of Blitar with an altitude of 450 meters above sea level. 3D object reconstruction techniques are divided into 2 categories, namely active techniques and passive techniques. Active technique (object scanning) requires control of structured light. Some researchers use a projector or viewer to produce structured light [2]. Other researchers used a laser beam and a video camera. Passive technique is done by taking using two or more images of an object from various positions with the camera [3,4]. This technique is often known as the adoption of photogrammetry or structure from motion. By using the 3D model reconstruction approach based on Multi View Stereo (MVS) using the Structure From Motion (SFM) algorithm, it is hoped that the results of this 3D modeling can be utilized as an effort to preserve 3D objects in the Penataran Temple cultural heritage area.

2. Literature

2.1. Structure Form Motion (SFM)

Structure From Motion (SFM) is the process of estimating the 3D structure of a scene from a series of 2D images. SFM is used in many applications, such as 3D scanning and Augmented Reality [5]. SFM can be calculated in various ways. The way to approach a problem depends on various factors, such as the number and type of cameras used. If the picture is taken with a single calibrated camera, the 3D structure and camera movement can only be restored to an existing scale, changing the structure of the scale, the magnitude of the camera's movement or still maintaining observation. For example, if you place the camera close to an object, you can see the same image as when zooming in and moving the camera away. If you want to calculate the actual scale of structures and movements in the area, additional information is needed, such as:

- The size of the object in the scene.
- Information from other sensors, for example, odometer.

For most applications, such as robots and autonomous drivers, SFM uses more than two displays.

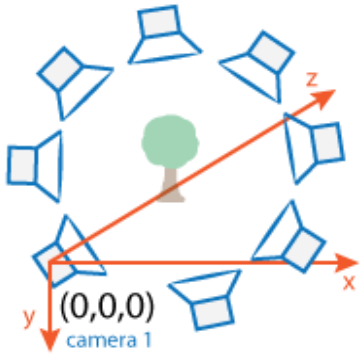


Figure 1. SFM multiple view.

The approach used for SFM from two views can be extended to several views. The set of displays that are used for SFM can be ordered or canceled. The approach taken here assumes a regular sequence of views. SFM from multiple displays requires the corresponding point in several images, called tracks. A typical approach is to calculate tracks from correspondence of paired points.

The SFM algorithm takes as input one set of images and produces two things: the camera parameters of each image, and a set of 3D points seen in images that are often encoded as tracks. Tracks are defined as 3D coordinates from reconstructed 3D Points and a list of corresponding 2D coordinates subset of input images. Most of the cutting-edge SFM algorithms share the same basic processing pipeline.

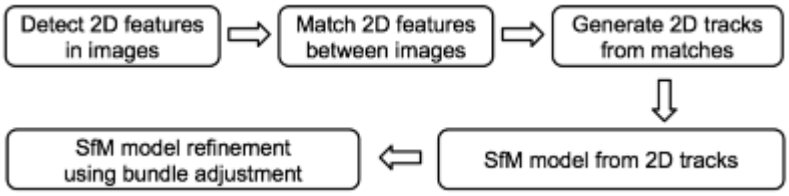


Figure 2. SFM stages.

The effort then focused on two key components of the SFM algorithm:

- Calculate Euclidean reconstructions (up to scale) from multiple cameras, i.e. estimating camera parameters and 3D positions of tracks, and
- Build 2D tracks that are longer.

2.2. Multi View Stereo (MVS)

The origins of stereo multi-view can be traced back to human stereopsis and the first attempt to solve the stereoscopic matching problem as a matter of calculation [6]. To this day, the Multi-View Stereo algorithm has become a very active and fruitful area of research [7]. The Multi-View Stereo version originated as a natural enhancement for the two-display case. Instead of taking two photos from two different viewpoint photos, Multi-View Stereo will capture more angles in between to increase durability, for example for image noise or surface texture [8,9].

Although MVS has the same principles as the classical stereo algorithm, the MVS algorithm is designed to handle images with more varied perspectives, such as sets of images that surround objects, and also handles very large numbers of images, even in millions of orders. The difference in the nature of the MVS problem eventually results in a significantly different algorithm from the classic stereo partner. For example, industrial applications for 3D mapping [6], processing millions of photos over hundreds of kilometers at a time, effectively reconstructing large metropolitan areas, countries and eventually the entire world.

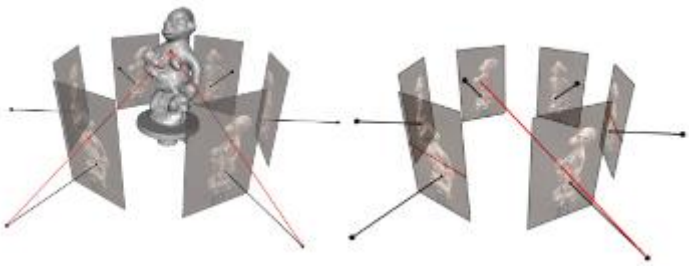


Figure 3. SFM stages.

Matching pixels in an image is a difficult problem unique to stereo or multi-view stereo. In fact, optical flow is another area that is very active in Computer Vision, overcoming the problem of correspondence that is dense across images [10]. The main difference with the existence of MVS is that optical flow is

usually a matter of two images (similar to two stereo displays), the camera is not calibrated, and its main application is image interpolation rather than 3D reconstruction.

3. Methodology

This research was carried out sequentially and systematically (sequence) starting from data collection, process to produce output in order to facilitate the process. In this study using the Photogrammetry Pipeline model [11]. The following stages of research are carried out to produce 3D objects from 2D images.

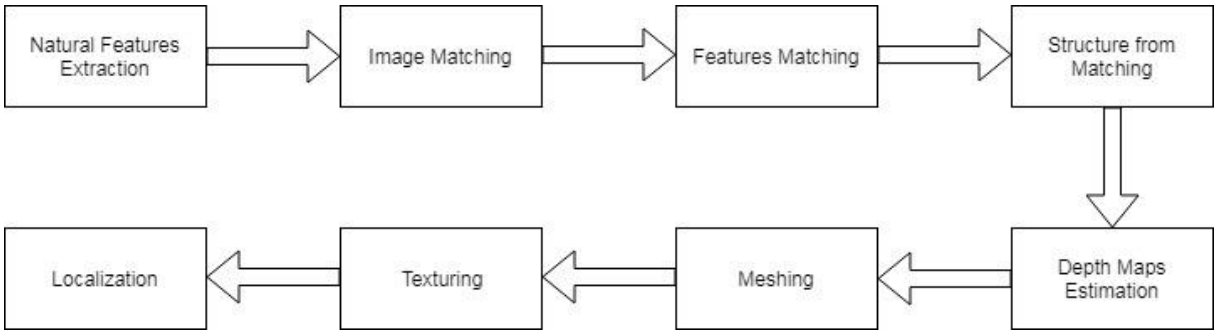


Figure 4. Photogrammetry Pipeline.

3.1. Natural features extraction

The purpose of this step is to extract distinct groups of pixels that, to some extent, do not change at the changing camera viewpoints during image acquisition. Therefore, the features in the scene must have similar feature descriptions in all images.

3.2. Image matching

The purpose of this section is to find images that are looking into the same area of the scene. To do this, we use shooting techniques to find images that share some content without the cost of completing all of the feature matches in detail. His ambition is to simplify the image in a concise image descriptor which makes it possible to calculate the distance between all image descriptors efficiently.

3.3. Features matching

The purpose of this step is to match all features between candidate pairs of images. First, a photometric match was made between the descriptors set of 2 input images. For each feature in figure A, get a list of candidate features in figure B. Because the descriptor space is not linear and well-defined space, it cannot rely on absolute distance values to find out whether the match is valid or not (can only have a higher absolute bound distance). To remove a bad candidate, it is assumed that there is only one valid match in the other image. So for each feature descriptor in the first image, look for the 2 closest descriptors and use the relative threshold between images. This assumption will turn off features in the repetitive structure but has proven to be a strong criterion. It provides a list of candidate matching features based only on photometric criteria. Find the 2 nearest descriptors in the second image for each feature that is computationally intensive with the brute force approach, but there are many optimized algorithms. The most common is the Estimated Nearest Neighbor, but there are alternatives such as, Cascading Hashing.

3.4. Structure from motion

The aim of this step is to understand the geometric relationships behind all the observations provided by the input image, and deduce the rigid structure of the scene (3D dots) with poses (position and orientation) and internal calibration of all cameras using the non-linear bundle adjustment method to improve structure and motion as well as minimize reprojection errors.

$$E(P,X) = \sum_{i=1}^{m}\sum_{j=1}^{n} D(X_{ij}, P_i X_j)^2 \quad (1)$$

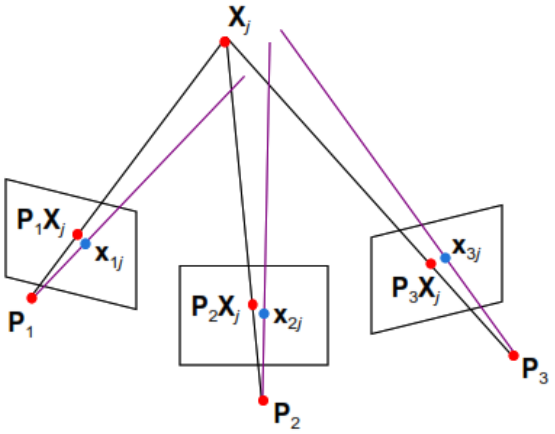


Figure 5. Bundle adjustment process.

Furthermore, the reconstruction step will be calculated automatically from the two initial views which are extended iteratively by adding a new view or what is called Incremental SFM.

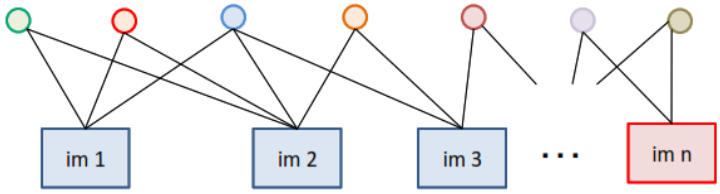


Figure 6. Incremental SFM.

3.5. Depth maps estimation
For all cameras that have been completed by SFM, we want to take the depth value of each pixel using the Semi-Global Matching (SGM) or ADCensus approach applied in AliceVision in Meshroom [11].

3.6. Meshing
The purpose of this step is to make a geometric representation of the dense surface of the scene.

3.7. Texturing
The purpose of this step is for the resulting mesh texture. If the mesh has no associated UVs, it will calculate the UV map automatically. AliceVision implements a basic UV mapping approach to minimize texture space.

3.8. Localization
Based on the results of SFM, camera localization can be performed and take animated camera movements at the 3D reconstruction site.

3.8.1. Camera calibration. The internal camera parameters can be calibrated from several checkered displays. This makes it possible to take parameters of focal length, main point and distortion.

3.8.2. Single camera localization. Use the algorithm presented in the image matching section to localize the nearest image in the SFM result. Then do the matching features with these images as well as with the previous N frames. Then immediately get the 2D-3D association, which is used to localize the camera.

3.8.3. Lots of cameras. If a camera rig is used, it can calibrate the rig. Localize the cameras one at a time in the whole sequence. Then use all valid poses to calculate relative poses between the camera rig and choose a more stable value throughout the image. Then initialize the relative pose of the rig with this value and make a global Bundle Adjustment on all rig cameras. When the rig is calibrated, it can use it to directly localize the pose of the rig from a multi-camera system that is synchronized with the approach.

4. Results and discussion

The 3D object reconstruction research was built from photos as many as 61 images from various angles. Furthermore, the image is reconstructed using the Structure From Motion algorithm in the Meshroom application [11] to produce a complete 3D model with textures. While the photos themselves were taken using the Canon EOS-100M camera. Figure 7 shows the 3D Model Reconstruction using the Multi View Stereo and Structural From Motion method.

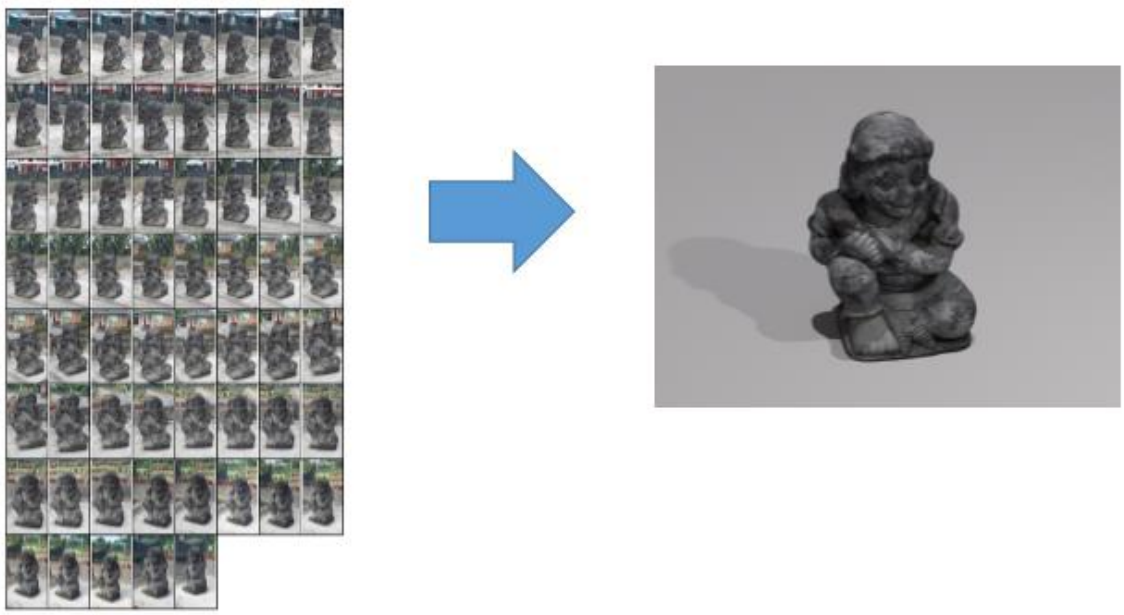


Figure 7. 3D reconstruction of Dwarapala statues in Penataran temple.

Table 1 is a comparison of data processing data from shooting using the Canon EOS-100M camera on the original image or before the data is compressed and after the image is compressed in order to compare the optimal process and results at the Structure Form Motion Budle Adjustment stage.

Table 1. Structure From Motion Bundle Adjustment.

No	SFM Bundle Adjustment	Original Picture	Compress Picture
1	Refine Pose	33	33
2	Iteration	223	220
3	Time (s)	50.30	6.77

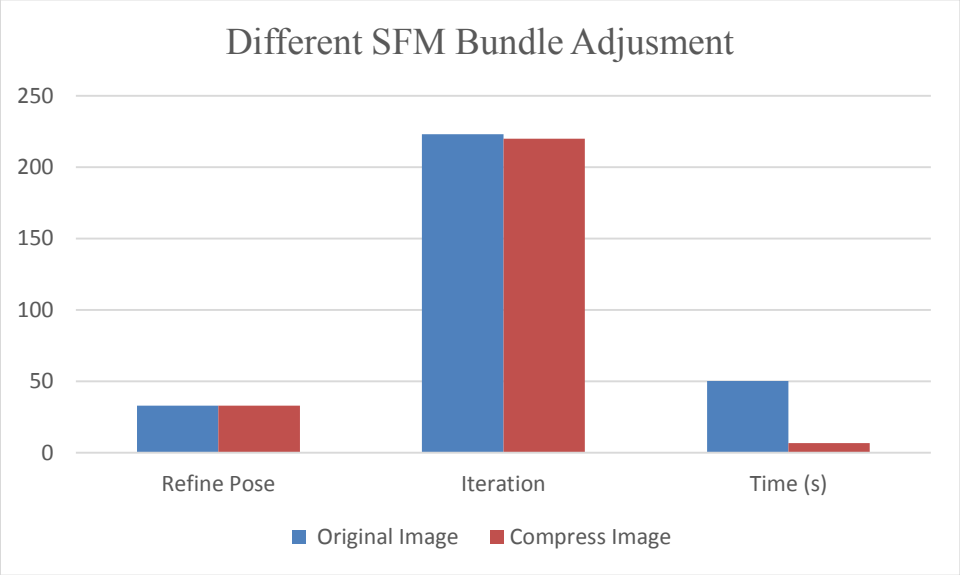


Figure 8. SFM Bundle Adjustment chart.

5. Conclusion

This research was conducted by taking as many as 61 images of objects in the Blitar Penataran Temple area. The photos obtained were reconstructed into a 3D model using the Structure From Motion algorithm in the meshroom. In this study a trial of the original image with a compressed image for reconstruction is used to compare the 3D reconstruction process from the two input data. From 61 images processed using the Structure Form Motion algorithm, 33 poses of camera pose and 3D points were improved, both original and compressed images. The number of iterations compresses 1.4% less than the original image and takes 43.53% faster than the original image.

References

[1] Mulyadi Y 2012 Mengoptimalkan Zonasi Sebagai Upaya Pelestarian Cagar Budaya *Buletin Somba Opu Balai Pelestarian Peninggalan Purbakala Makassar* **15**
[2] Dipanda A and Woo S 2005 Towards a real-time 3D shape reconstruction using a structured light system *Pattern recognition* **38**(10) 1632-1650
[3] Prakoonwit S and Benjamin R 2007 3D surface point and wireframe reconstruction from multiview photographic images *Image and Vision Computing* **25**(9) 1509-1518
[4] Park J S 2005 Interactive 3D reconstruction from multiple images: A primitive-based approach *Pattern recognition letters* **26**(16) 2558-2571
[5] Hwang J T and Ting-Chen C 2016 3D Building Reconstruction By Multiview Images and the Integrated Application With Augmented Reality *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **41** 1235
[6] Google Inc. Google maps. http://maps.google.com.
[7] Scharstein D and Szeliski R 2002 A taxonomy and evaluation of dense two-frame stereo correspondence algorithms *International journal of computer vision* **47**(1-3) 7-42
[8] Okutomi M and Kanade T 1993 A multiple-baseline stereo *IEEE Transactions on pattern analysis and machine intelligence* **15**(4) 353-363
[9] Tsai R Y 1983 Multiframe image point matching and 3-d surface reconstruction *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2) 159-174
[10] Baker S, Scharstein D, Lewis J P, Roth S, Black M J and Szeliski R 2011 A database and evaluation methodology for optical flow *International journal of computer vision* **92**(1) 1-31
[11] https://alicevision.org/#photogrammetry (Agustus 1, 2019).