Mau Le-Tien
Khoi Nguyen-Tan
Romain Raffin

AUTOMATIC INDEXATION OF CULTURAL HERITAGE 3D OBJECT

Abstract *There has been significant evolution in the fields of 3D digitization thanks to the development of 3D reconstruction and geometry processing. The results of digitization researches have been widely applied in many fields, especially in Cultural Heritage and Archaeology. Reconstruction, characterization and annotation of components forming 3D objects have become an effective tool for research, conservation and promotion of archaeological relics. The aim of this paper is to propose a process of 3D model reconstruction, segmentation and annotation on the basis of a enhanced corresponding 2D dataset. A machine learning method is used for the semantic segmentation of 2D images, thereby label, annotate and reconstruct a 3D model based upon links between distinctive invariant features, orientation of images, and depth map of images. The initial result as a data basis for research, reconstruction and identification of parts in 3D objects is applied in the reconstruction of archaeological relics, object identification, 3D printing, etc. Our work uses the data collected from the Museum of Cham Sculpture DaNang and the Myson QuangNam sanctuary in VietNam, to carry out the proposed method.*

Keywords 3D reconstruction, cultural heritage, semantic annotation, 3D digitization

Citation Computer Science 26(1) 2025: 5–25

1. Introduction

In Vietnam and neighboring countries, the Champa cultural heritage domain is very important. There are a variety of models of (tangible) cultural heritages and archaeological objects [6]. Management and preservation of the Champa archaeological remains are of particular significance in terms of their religion and origin. However, over time these remains have become extinct, and they are in needs of urgent protection. Cultural heritage items are priceless and of very high historical values. Maintaining and conserving these assets is a top priority for all nations around the world [2]. The purpose of preserving digital cultural heritage is not limited to the persistence of historical archives, as it can provide a good reference for creating applications on cultural heritage data with the purpose of information sharing and cultural communication.

Initially, 3D digital mockups of cultural heritage items were used for conservation and archives. As 3D digitization becomes cheaper and more affordable, processing 3D data (measuring, comparing, indexing...) has started a trend in archaeological studies. On the other hand, the development of reconstruction and analysis of 3D model researches has grown rapidly for applications such as computer vision, virtual reality, digital heritage and many others [3, 14].

The 3D model enables us to have different views, from overview to details, and to analyze objects according to many different criteria. There are many methods and tools for reconstruction of models and samples in various ways such as using scanners, magnetic resonance imaging, and reconstruction from 2D images. The outcomes are used to segment, re-identify, and rebuild objects. Models and artifact samples must be kept and conserved digitally in light of recent natural disasters, climate change, and looming devastation, therefore it is enabling the capture and digitization of 3D models of historical and archaeological artefacts.

We propose the method of data collection for 3D model reconstruction from 2D images and create a set of model training data for segmentation, analysis and annotation of parts of 2D objects. Our method is based upon the simultaneous combination of recognition, segmentation, and annotation on 2D images and conversion of these segmentations and annotation to 3D images of the same object. Then, we will carry out the annotation and segmentation of 3D objects based upon the analysis of features and links between them. The experimentation data is a set of Cham statues from Champa culture in the central region of Vietnam (DaNang, MySon). These statues are similar in material, color, and shapes.

This paper has the following layout and structure: Section 1: Introduction, Section 2: Introduction to some researches related to 3D model reconstruction, machine learning methods Section 1: Proposing methods for 2D data collection for model training and reconstruction. In this stage we propose a method for analysis, segmentation and annotation of 3D objects on the basis of 2D images for labeling and segmentation. Section 4: The experimental results from the Cham and MySon statues of Vietnam, and the last Section farther is the conclusion and discussion.

2. Related works

In the field of cultural heritage studies, it is necessary to collect images for 3D reconstruction for visual representations and analyze their characteristics. There have been many studies where several ways to achieve this goal have been explored. Authors in [20] analyzed current optical 3D measuring sensors and 3D modeling methodologies, as well as their prospective and actual 3D surveying and modeling of cultural sites. And the ones in [26] presented a novel approach to reconstruct the 3D shape of scene via a single camera.

The paper proposed that it was possible to simultaneously retrieve the structure of dynamic surfaces and static scene geometry in the nature. The study conducted in [4] proposed a method to make it possible for automatically annotating 2D textures of heritage objects and visualizing them onto 3D geometries based on supervised machine learning methods. In a similar context, the authors of [13] presented a method for doing semantic annotations on 2D photos with automated transmission of these annotations between different associated representations of the object (either 2D or 3D).

More recently, works conducted by [15, 21] based on the precise identification and matching of homologous image features, which could be separated into two major parts: image orientation and dense image matching, this method used the camera pose estimation and sparse point cloud generation techniques. However, this approach can still result in a noisy cloud and a number of mistakes. Based on the precise identification and matching of homologous image features, which could be separated into two major parts: image orientation and dense image matching, [25] used the camera pose estimation and sparse point cloud generation techniques. However, this approach can still result in a noisy cloud and a number of mistakes.

The study in [1] performed segmentation incorporating information from both the 3D and 2D space based upon the same mathematical surface. In [12] divided the different components of the building facade without giving the discovered segments semantic names. These authors were concerned with small-scale objects and have a view of a single image. They did not have semantic annotation for models. In [7], the experimental results of the paper involve annotating the 3D model based on feature points from corresponding 2D images. The advantage of the method lies in using a binary mask to filter all the feature points of the 3D model of the same object.

However, the feature points and the binary mask on the 2D images were manually created by the authors. Similarly, in [8], the authors employed a similar idea to semantically annotate the 3D model based on 2D image features. The approach of this paper involves extracting and recognizing features from a pre-recognized and pre-trained model (such as faces) on the 2D images.

The combination of recognition on 2D and 3D models of the objects was limited. We therefore propose an approach for 3D reconstruction, detection, and semantic annotation of a statue based upon information from the 3D model and a set of 2D images. Then, we present an approach to segment objects based upon 2D/3D combination and multimodal approach, working on both 2D and 3D information. This approach detects the information based upon a combination of 2D images and 3D models to detect features in 2D and 3D objects.

Then, we present an approach to segment objects based upon 2D/3D combination and multimodal approach, working on both 2D and 3D information. This approach detects the information based upon a combination of 2D images and 3D models to detect features in 2D and 3D objects.

3. Methodology

3.1. Overview

The goal of the proposed method is semantic annotations for 3D objects based on segments of corresponding 2D objects. To do that, we used the machine learning method [5,22] to train a model, then identify the components that created the 2D object. The data for training here is a set of images collected from many objects and a variety of sources. And to obtain 3D models, with each set of photos of the same object performing 3D model reproduction (helped by [17,19]), a range of processing data are available).

The important stage of this process is to determine orientation, position and calibrate the set of images to detect the relative orientations and the depth map of the oriented images. Markers linking images are obtained according to SIFT features.

All relations between 2D pixels set and 3D model must be maintained along the reconstruction process. Suppress this and the results of the stage are a semantic segmentation image, which are created from an image segmentation converted into a collection of regions of pixels represented in a labeled image. Thus, the propagation of annotations between images is implemented with a 2D and 3D relative projection. We process only the important segments of the image instead of the entire image Based on the SIFT points, homologous points between pairs of images, direction of images and stage of establishing is the image depth map. The state of model reconstruction combined with image segmented on the same object. The flowchart of our approach is shown in Figure 1.

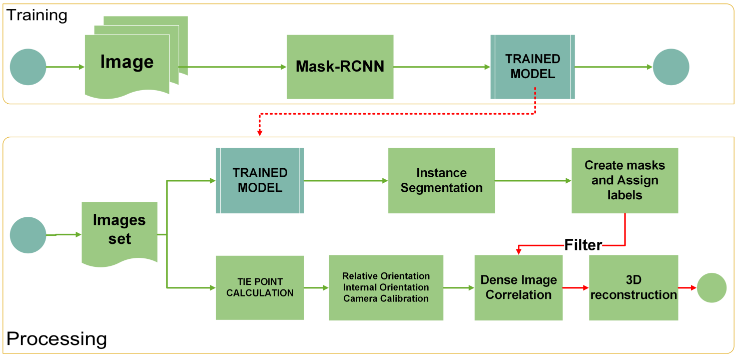

Figure 1. Overview of our processing stages

3.2. Data collection

In the field of 3D model reconstruction from images, to transform 2D image measurements into 3D information, image data needs a mathematical framework. Most of the time, at least two photos are needed to construct 3D data, and projective geometry or perspective techniques can be used [20]. However, in order to improve the efficiency of reconstruction and training, each pair of images needs to have an overlap, and the parameters of the camera such as luminosity, focal length, aperture, and movement speed should be stable [19].

We construct our dataset by digitizing sculptures at the Museum of Cham in DaNang and MySon statues, by taking a set of images around each item. Figure 2 is an example describing a digitization with various camera positions:

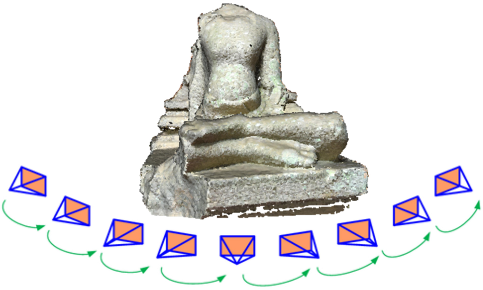

Figure 2. Describing camera positions

With each set of the multi-view images obtained, in the next section we introduce the processing method and reconstruction of 3D model for each set of images.

3.3. Processing and reconstruction of 3D model

In the field of photogrammetry, to calculate depth maps and build 3D dense point clouds, stereo matching techniques are typically utilized in pairs. Through the process of interpolating the 3D sparse point cloud created by the picture orientation process, an initial depth map is essentially produced [25]. This conversion is based on a projection in object space of each pixel of a master image according to the image orientation parameters and the associated depth values. During 3D reconstruction, all images are considered the master image. They are selected and processed in a sequence. In addition, if we want to obtain color features of the object, we just need to choose a master image, the one that can cover the whole object. Each 3D point is associated with an RGB attribute from the master image [18]. In [24] describes a method to recover depth points based on the disparities of corresponding image points and the surface model produced.

Finding a mapping is the goal of matching $\mathcal{F}_{px} : \tau \otimes \varepsilon_{px}$. The point cloud is created in the 3D euclidean space τ. The depth of the Euclidean space is ε_{px} and $\mathcal{Z}$, in there $\mathscr{Z}$ is the image disparity or the Euclid distance between two images. As in Figure 3 and Equation (1).

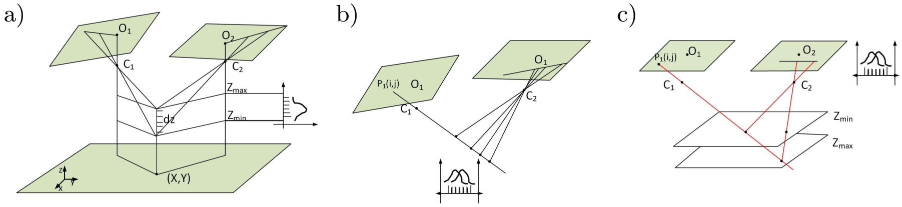

Figure 3. MicMac uses three restitution geometries: a) ground, euclidean space; b) image geometry that is discretized along the ray; c) epipolar geometry resampled in image space

MicMac provides a sufficiently flexible formulation of the matching cost function in terms of the optimization strategy and the dimension of the match.

$$\mathcal{E}(\mathcal{F}_{px}) = \iint_{\mathcal{T}} \mathcal{A}(x, y, \mathcal{F}_{px}(x, y)) + \|\nabla(\mathcal{F}_{px})\|^{reg} \tag{1}$$

The similarity metric between pixels is $\mathcal{A}(x, y, \mathcal{F}_{px}(x, y))$, and $\|\nabla(\mathcal{F}_{px})\|^{reg}$ is the gradient's norm and determined as follows:

$$\|\nabla(\mathcal{F}_{px})\|^{reg} = \alpha_1 * |\nabla\left({\mathcal{F}_{px}}^1\right)| + \alpha_2 * |\nabla\left({\mathcal{F}_{px}}^2\right)| \tag{2}$$

In which α_1, α_2 are the regularization of the disparity's first and second elements, respectfully. The a priori in the disparity space is controlled by value of $\|\nabla(\mathcal{F}_{px})\|^{reg}$. Micmac used a global reduction using the Min-Cut/Max-Flow approach to identify the disparity map that optimizes the energy [23]. This stage provides a map of correlation coefficients between pairs of images to determine the depth for each pixel in the depth map. Each 3D point, after reconstruction, is always associated with a specific pixel. The main processing steps on the stage of reconstruction [10,16,19] are described as follows (see Fig. 4):

1. Take pictures from different positions and directions to cover the entire subject.
2. Calculate tie points and the match between all pairs of image using the SIFT algorithm.
3. Use the tie point set of observations in the bundle adjustment and determines the element of image orientation (external orientation and camera calibration).
4. Use orientation of image to measure the likelihood for pixels from two images to belong to a unique three-dimensional point. And the image dense matching performed in the image's geometry with the pair image. As a result, one depth map is generated for each image. These depth maps are merged into one single point cloud model covering the entire area.
5. Finally, extract the global point set by using an energetic approach to minimize, on the whole considered space, a sum made up of correlation coefficients for each pair of images and smoothing term in order to homogenize the point set.

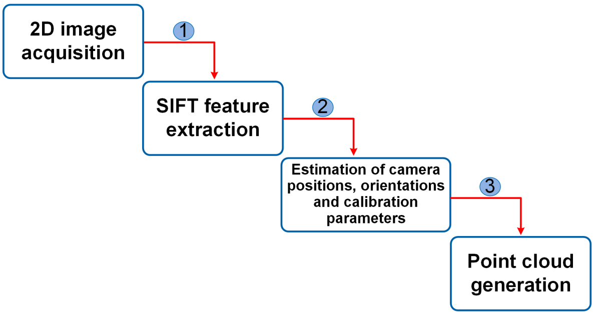
2D image acquisition
1
SIFT feature extraction
2
Estimation of camera positions, orientations and calibration parameters
3
Point cloud generation

Figure 4. Overview of point cloud generation process

In this stage, key points (interest points) are detected from image, then the key points are computed for descriptors by Sift algorithm. By comparing the two sets of stereo image descriptors, they will be matched to obtain corresponding points. The point matches will be utilized to calibrate and align the photos in 3D coordinates, as well as to acquire camera views automatically. Additionally, a well-known method called RANDOM SAmple Consensus (RANSAC) is utilized to eliminate outliers, or unreliable point matches, depending on camera settings. A bundle adjustment used to compute the camera parameters is referred to [17].

Depth map merging: The multiple depth maps that focus on the same area of the image are combined to eliminate the duplicate depth values for each back-projected 3D point and project to the 3D space. This results in a smooth and distinctive thick cloud. While in the process of fusing, similar to the depth filtering step before it, pixels are once again projected to the three dimensional space and then back projected to the neighboring views, merging only the depth values that are determined to be close enough [25].

3.4. Semantic segmentation and annotation on 2D image

Convolutional Neural Networks (CNNs) are a type of deep learning algorithm highly effective for visual data analysis. Designed to process pixel data, CNNs excel in tasks like image classification, object detection, and image segmentation by learning spatial hierarchies of features through layers such as convolution, pooling, and fully connected layers.

Convolutional neural networks are one of the most advanced models, having been widely used in the deep learning method. From the research results of CNNs, there have been many applications using the research results in the identification and image classification. The paper uses the identification and segmentation results to annotate corresponding 3D object models. Figure 5 presents the process using machine learning [5, 22] on our dataset, for identification and image classification. The proposed method includes the following main stages: The first stage is to generate a set of proposals candidate object bounding boxes that have the higher probability of an object called a Region Proposal Network (RPN). The second stage extracts features using these proposals from each candidate box, performs classification and bounding-box regression. And the third stage presents mask segmentation on each Region of Interest (RoI) as Figure 5.

1. Use convolutional neural networks to extract input features.
2. Characteristic zones are routed through RPN to find RoIs and return them to bounding boxes in zones containing objects.
3. The RoIs are separated from the feature map and through the RoI pooling layer to adjust and stack into blocks on the same size.
4. The RoIs are passed through the connection layer for classification and boundary box prediction.
5. In step 3, RoI continue to be processed through 2 convolutional steps to create binary masks for objects.

From the method, we have gained a model, after training, for object identification and segmentation. Hereby, we propose a method of separating zones which have been identified, to mark and label different parts, respectively, for each image in the dataset. The purpose of reuse in the filter extraction of specific zone and establish a depth map for each image in section 3.4 and mark, label the 3D point set to create semantic annotations for the objects.

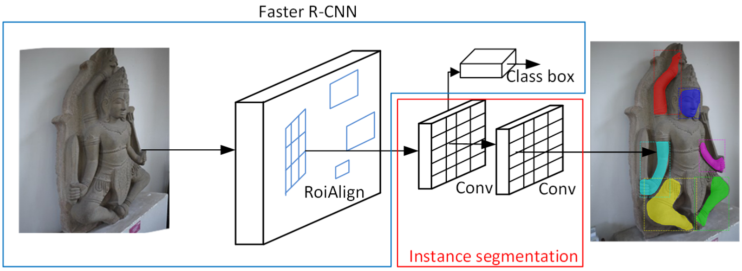

Figure 5. The architecture of the Mask-RCNN model used to segment for one point of view

In the training step, authors [22] defined a loss function on each sampled RoI as:

$$L = L_{cls} + L_{box} + L_{mask} \tag{3}$$

Where, the classification loss $L_{cls}(u,p) = log(p_u)$ is the log loss for each class u with network predicted probability p. The box loss L_{box} is defined over a tuple of true bounding box regression targets and the center of box, dimensions for class u, $v = (v_x, v_y, v_w, v_h)$, and a predicted tuple $t_u = (t_x^u, t_y^u, t_w^u, t_h^u)$, for class u is:

$$L_{box} = \sum_{\in x,y,w,h} smooth_{L_1}(t_i^u - v_i) \tag{4}$$

in which:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if|x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \tag{5}$$

L_{mask} is defined as the average binary cross entropy loss:

$$L_{mask} = -(ylog(p) + (1-y)log(1-p)) \tag{6}$$

After performing the model training, the segment results of some statues will be shown in Figure 6. Thereby, each segmented image will be processed to attach a corresponding label and ID. For the same segment type, the same label will be used. As the Figure 7 shows, although the images have different shooting directions, if the part of the head, arms and legs are the same, they will all have the same label.

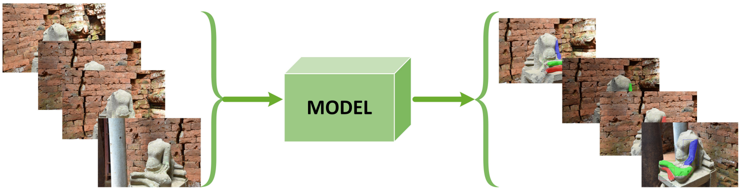

Figure 6. Using pre-trained models for annotation and segmentation

The aim of this step is to build an instance segmentation model for cultural heritage items segmentation by fine-tuning the state of the art Mask-RCNN algorithm. The model is performed using heritage dataset which is prepared and collected from Vietnamese statues.

3.5. Multi-modal merging step : automatic indexing 3D points sets

The 3D model obtained from one point of view after reconstruction is an unstructured of points and its information is discrete. However, the 3D points that are characterized are obtained from the feature sets, and each point is added with depth information. From that results, our method annotates the 3D object based on the identified and segmented set of images. Based on the masking of the segmented images to mark and filter the corresponding 3D point cloud, with each segmented part the creation of a 3D mask, labeling and indexing 3D points.

The sequential steps are described as follows:

1. Using the trained model to segment each image, set corresponding labels for each segmented part. Each segmented part must contain pixels with a different color and classes.
2. For each segmented image, we need to extract information as orientation, position and depth map that have been created in the reconstruction step.
3. Create a binary mask and apply corresponding labels to each segmented part.
4. Based on the depth map of each image, create the sub 3D point cloud and the corresponding binary mask to extract the set of 3D points for each segmented part.
5. Each obtained 3D point set for each segment part on the same label is merged.

In this step 4, the images are oriented in relation to each other, and the depth map of each image created. Each depth map is converted into a 3D, metric point cloud by projecting each pixel of the image in space according to the image orientation parameters and pixel depth by [24] and method [7, 11], based on the binary mask, to extract the set of pixel depth for each mask part according to the segmented composition. With each data set collected, we perform the model reconstruction to obtain 3D point cloud. As shown in Figure 7, the results on some models will have 3 different perspectives after reconstructing.

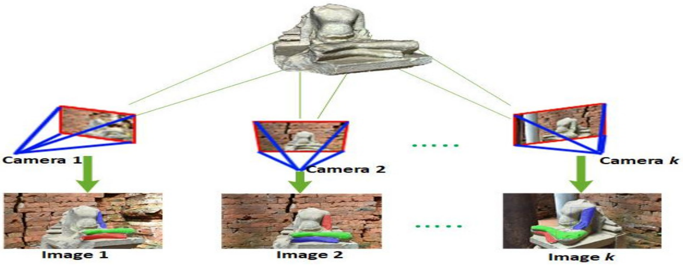

Figure 7. Annotation 3D model based on segmentation and orientation of the corresponding images

In Figure 7 one can see the generation of a point cloud for a segment, we made: For each image of the dataset segmented and labeled has similar labels, then group together (as arm, legs, face and so on). These segments have the same label mapped to 3D point cloud based on extracting the depth map corresponding to each part. Finally, merge all subset of point cloud with the same label. Because all the generating point clouds are in the same reference system, they can be easily mixed together, based on the position and orientation of segmented image and the depth images created point cloud. As described in section 3.3, these images used in the steps for the point cloud generation. It has an oriented, defined position in the MicMac tool chain. Figures 8 and 9 present 3D segments which are annotated and merged on one image.

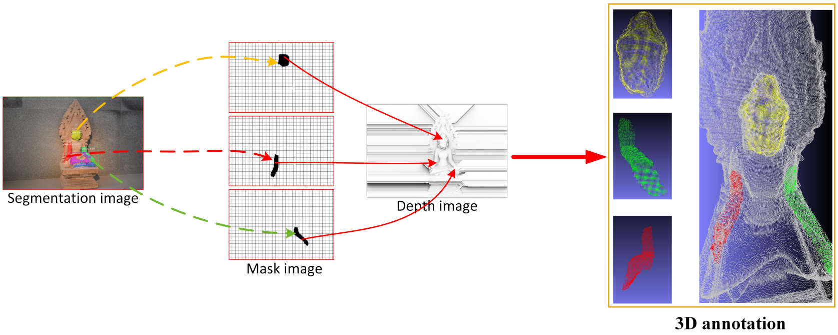

Figure 8. Use the mask of one segmented images to filter the corresponding 3D point cloud

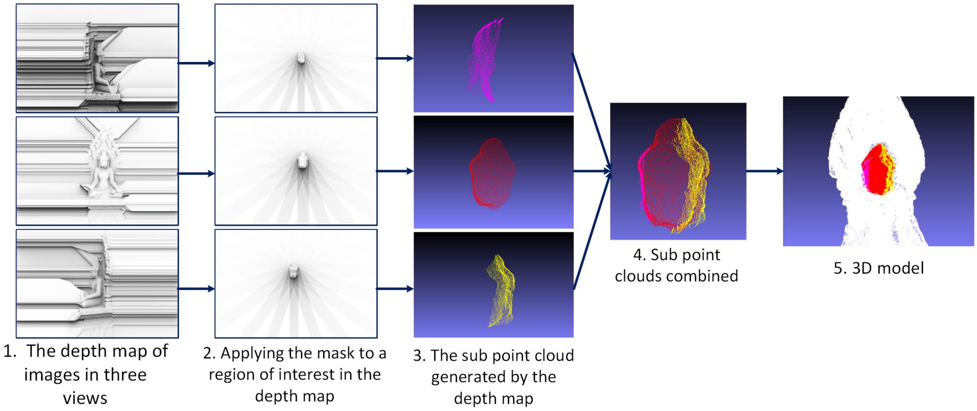
1. The depth map of images in three views
2. Applying the mask to a region of interest in the depth map
3. The sub point cloud generated by the depth map
4. Sub point clouds combined
5. 3D model

Figure 9. The example of depth map in 3D view and the merged point cloud

4. Experimental results

The image dataset used for training stage were obtained from heritage objects in museum of Vietnam such as Da Nang, MySon. The sample images are summarized in Figure 10, but they do not consist of the number of acquired images. The images are also used for 3D modeling obtained from heritage objects such as sculpture, statue. To provide data for training the CNN network for automatic semantic segmentation, the data labeling was done manually. The labeled images include five classes sharing certain similarities, namely "left hand", "right hand", "leg right", "leg right", "face". We train the statue model architecture with backbone CNNs namely ResNet-101 and using our own dataset to fine-tune all layers directly on the pre-trained weights obtained from MSCOCO dataset [5, 9].

As Figure 10 shows, various patterns of images were obtained from Champa museum for various statue type Shiva lingam, Garuda, Visnu, Tusi, Myson, Nuthan, Linhvat.

The Figure 11 shows training and validation loss value and the Region Proposal Network loss value while training one of the segmentation models.

Figures 12, 13 and 14 show the results of the instance segmentation framework on sample statues. Our research results are also limited by training and evaluation elements to arms, legs, face of the statue. The next step will be: given each set of 2D multi view images, 3D model is reconstructed and represented with texture. With the [24] method, we can obtain a 3D model from the set of images taken around the object from the featured point set analysis of the image and use the geometric model to reconstruct the model. All relations between 2D pixels set and 3D model must be maintained along the reconstruction process. The sample images and 3D model of Cham statue are shown in Figures 16 and 15.

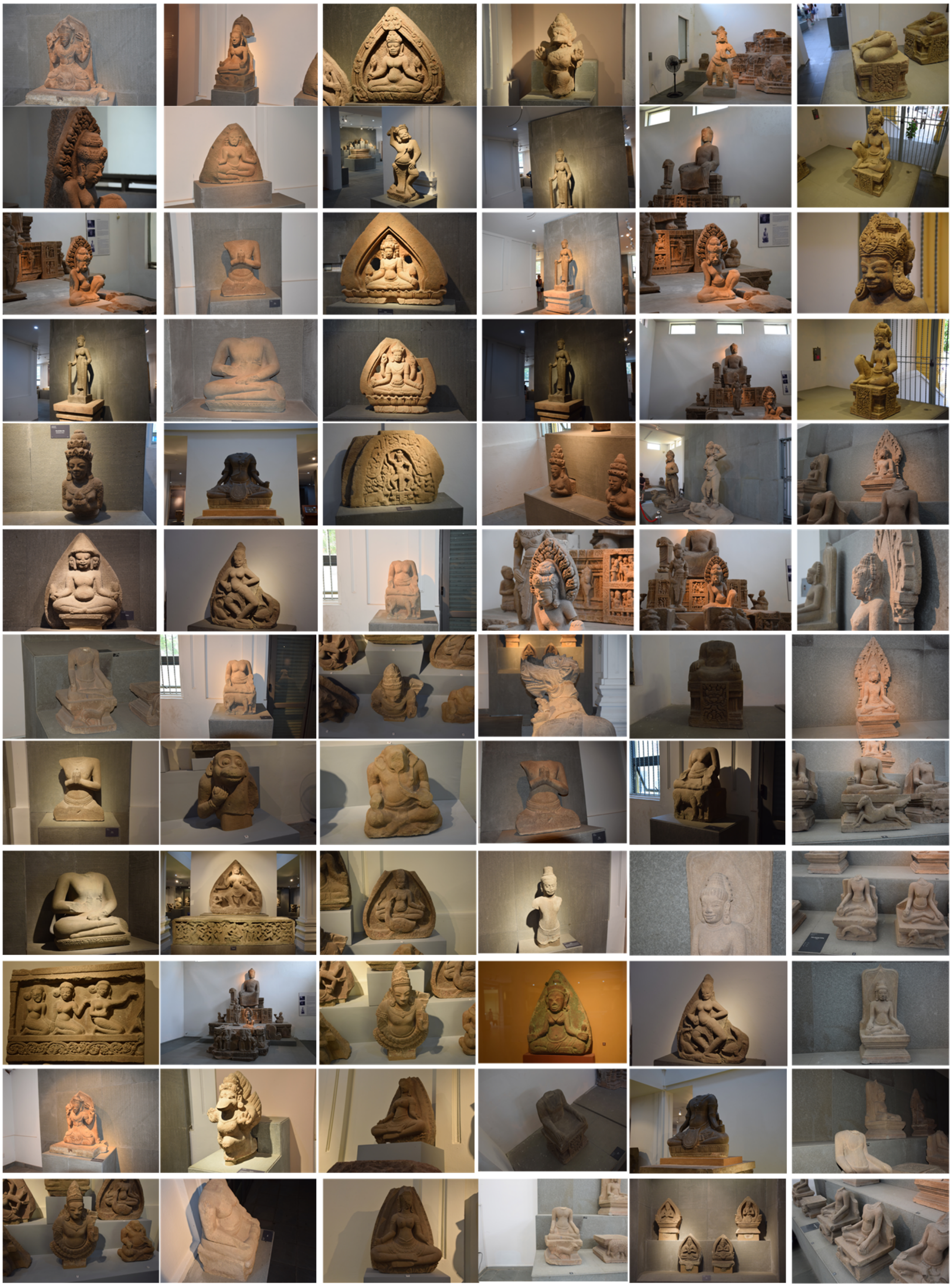

Figure 10. The training 72 sample images

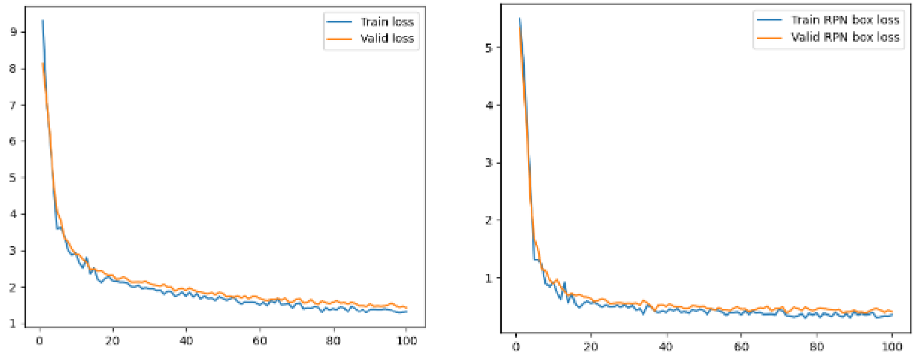

Figure 11. Training set loss (blue) and validation set loss(red) per epoch during the training epochs

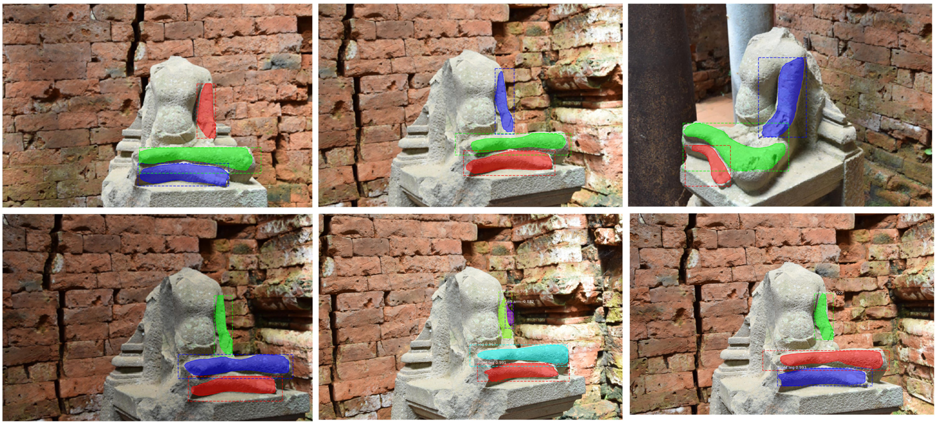

Figure 12. Segmentation images with MySon statue

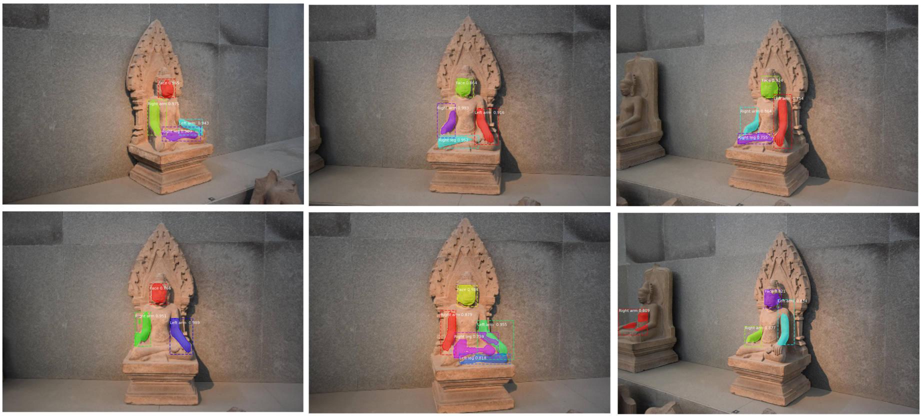

Figure 13. Segmentation images with Nuthan statue

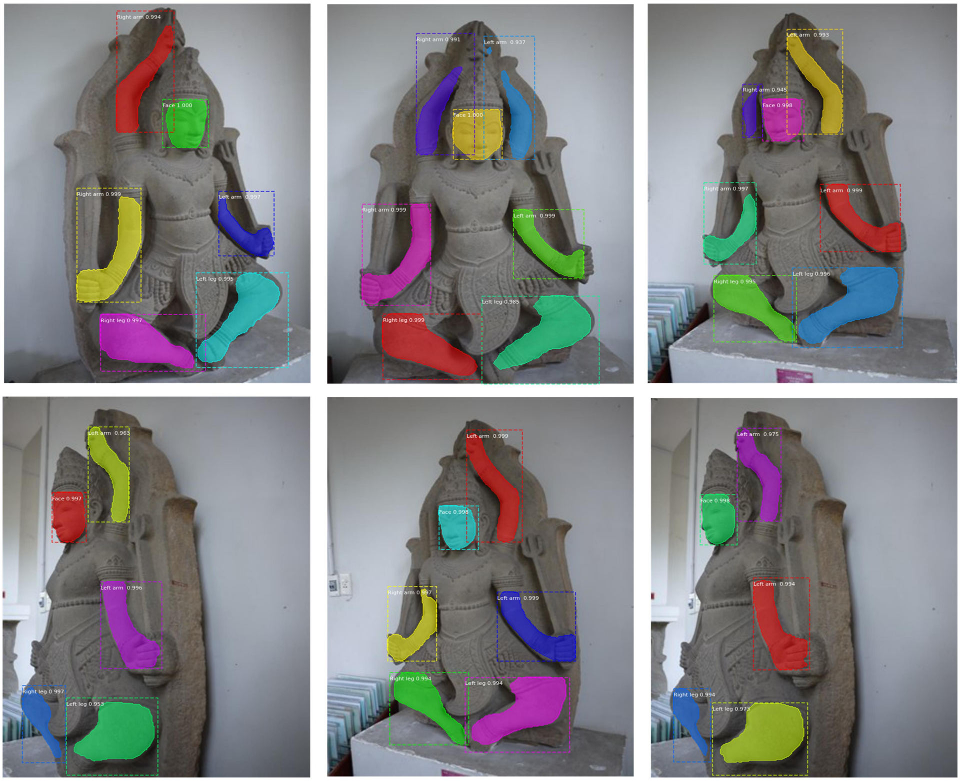

Figure 14. Segmentation images with Champa

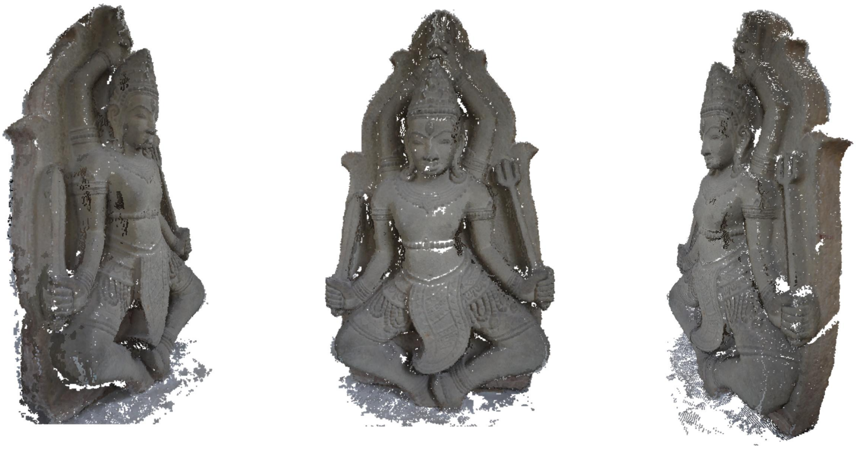

Figure 15. 3D point cloud of Cham statue with texture

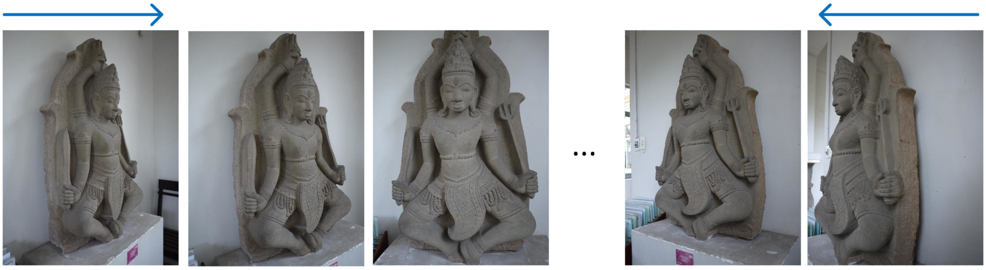

Figure 16. 2D multi-view images of Cham statue

Another sample images and 3D point cloud of Myson statue are shown in Figure 17 and 18. And sample images and 3D point cloud of Nuthan statue are shown in Figure 19 and 20.

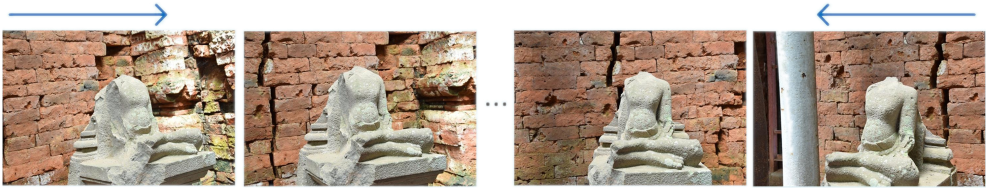

Figure 17. 2D multi-view images of MySon statue

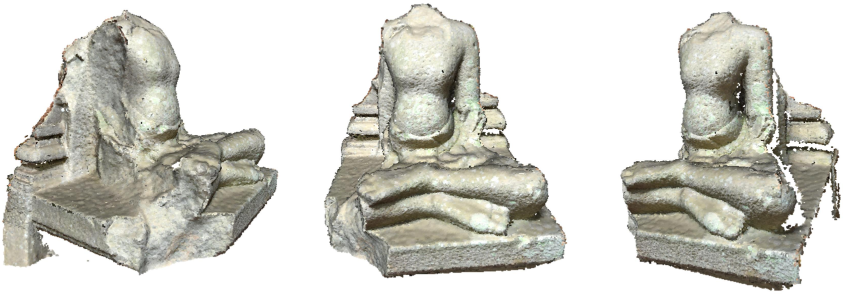

Figure 18. 3D point cloud of MySon statue with texture

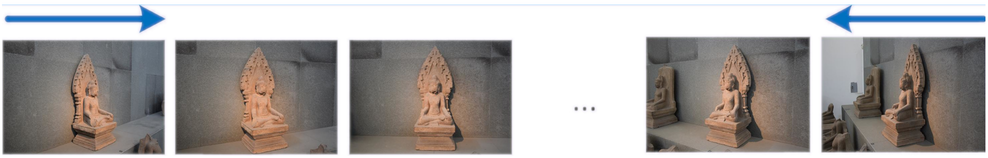

Figure 19. 2D multi-view images of Nuthan statue

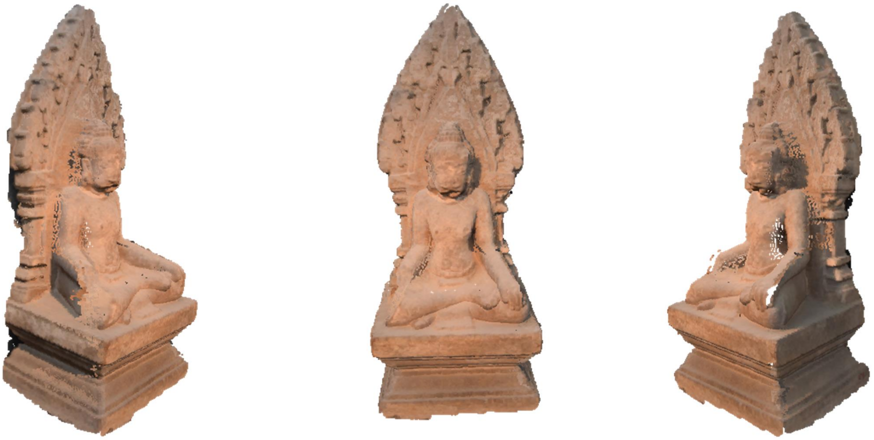

Figure 20. 3D point cloud of Nuthan statue with texture

As mentioned at Section 3.4, we trained the network with batches of size 50, a learning rate $\alpha = 2.5 \cdot 10^3$ over 100 epochs and the training data comprises an image set of cultural heritage. For the loss function, we used the binary cross entropy, which is defined as 3. As Figure 11 is an illustration to record the results of training set loss and validation set loss per epoch during the training epochs.

In previous processing step, 3D point clouds were created by using MicMac tool chain. Feature detection and feature matching were the key concepts to produce 3D point cloud from an image sequence. Figures 18, 15, 20, are showing the point cloud model of the statue. From the results, we see that the raw point clouds are often noisy and have holes, because it depends on the image acquisition technique as they represent the input data. So, the context must be exhaustively analyzed, including the lighting conditions of the scene, the values of exposure aperture and shutter speed of the camera.

Also in this period, we used [5] method to detect the part of statue as face, hands, legs on the 2D image, and from these the segmented image extraction is converted to binary mask to create an annotation area on point cloud, the 3D mask is created by a 2D binary mask. So, a white pixel coordinates will be extracted from the associated corresponding point cloud. And each point cloud set of different image orientations, after reconstructing, obtains a point cloud set of the same method [24]. So they all maintain the same scale and resolution, the combination between them are done by MeshLab open source software. Each point cloud is labeled with the annotated 2D image label.

The obtained results are shown on Figure 21, the left image creates 3D semantic annotations. The middle images and right image shown with non-texture and texture of the point cloud.

The experimental results demonstrate that the method achieves more detailed effectiveness in recognition, segmentation, and semantic labeling compared to previous studies. For instance, in [7], manually segmenting and identifying features in images posed limitations due to the requirement for manual region creation. Conversely, [8] employed a pre-trained model to identify parts of object faces. Our approach involves data collection, model training, and segmentation recognition on 2D images, followed by semantic labeling applied to corresponding 3D point clouds. However, the efficacy of identifying regions and features heavily relies on the quality of training data. The method also faces limitations, such as unclear areas in 2D datasets posing challenges during labeling. Regarding point clouds, the quality of 3D model reconstruction depends on the quality of collected image data, with a risk of increased noise levels in low-quality data.

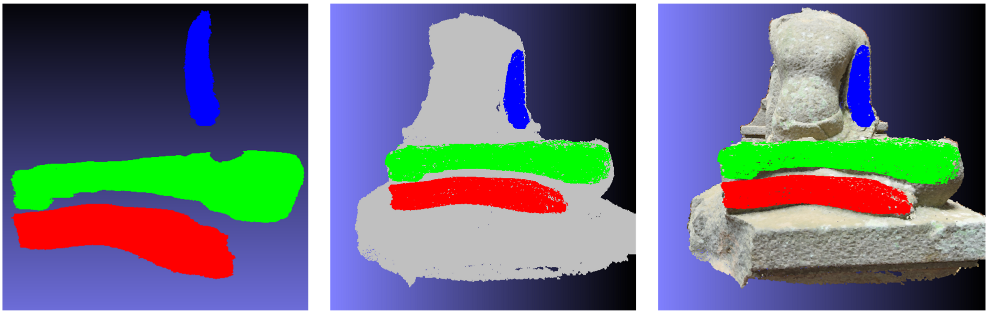

Figure 21. 3D indexed parts of point cloud of MySon statue

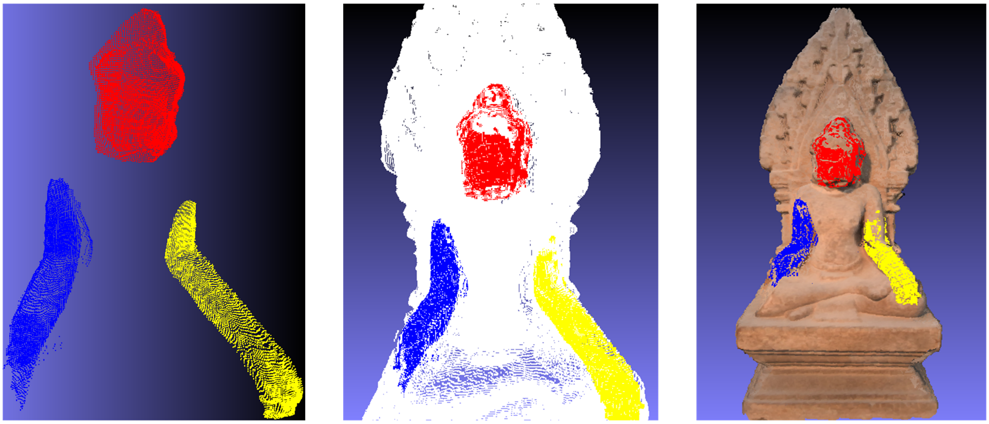

Figure 22. 3D indexed parts of point cloud of Nuthan statue

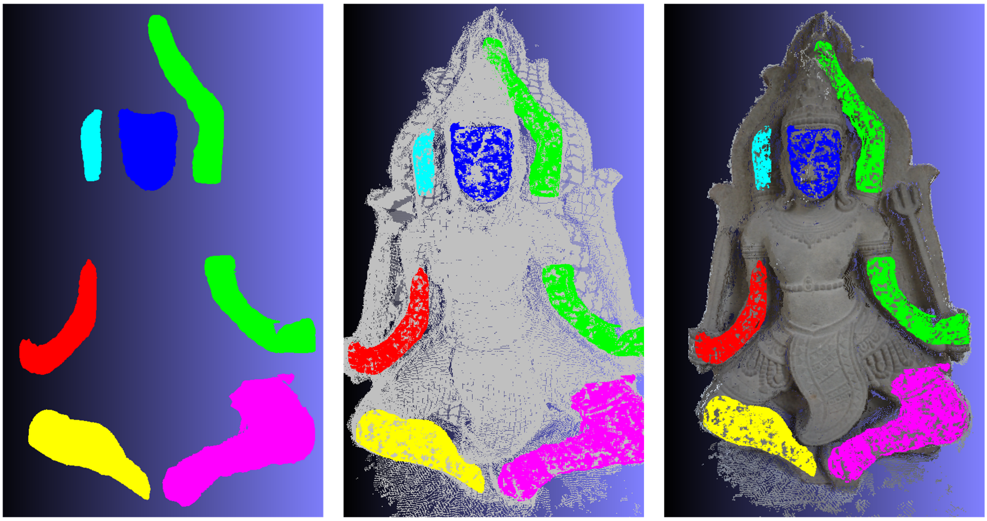

Figure 23. 3D indexed parts of point cloud of Nuthan statue

5. Conclusion and discussion

The process of reconstructing 3D from 2D images compared to digitizing 3D objects by focusing on collecting 3D points presents both advantages and challenges. While reconstructing from 2D images is often more feasible and convenient, it also faces challenges regarding accuracy and detail. Despite occasionally not achieving high precision, this method offers flexibility and lower costs, thus fostering potential development in cultural heritage preservation and research. The outcomes of the paper involve collecting and reconstructing images of several ancient sculptures at the Da Nang Museum and the My Son relics in Vietnam. Test results provide a foundation for segmenting, identifying, and analyzing various parts of 2D/3D objects for digital storage and preservation. With the proposed method, initial results include reconstructing and identifying parts of sculptures such as heads, hands, and feet. The paper‘s future research direction aims to identify all other components comprising the objects, creating a dataset of related parts for research and reconstructing damaged 3D models. In the future, we will propose reconstructing 3D models to identify and analyze damaged models, leading to their complete reconstruction. This will serve as a foundation for restoring ancient models and providing a database for research and preservation. We will reconstruct 3D models of various archaeological sculptures in Vietnam, potentially providing public access through a website and establishing a web-based annotation system. Highly accurate 3D models can serve as reference data for heritage, archaeological objects requiring restoration due to aging or natural disasters.

References

[1] Adam A., Chatzilari E., Nikolopoulos S., Kompatsiaris I.: H-RANSAC: A hybrid point cloud segmentation combining 2D and 3D data, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2, pp. 1–8, 2018. doi: 10.5194/isprs-annals-iv-2-1-2018.

[2] Belhi A., Foufou S., Bouras A., Sadka A.H.: Digitization and Preservation of Cultural Heritage Products. In: J. Ríos, A. Bernard, A. Bouras, S. Foufou (eds.), *Product Lifecycle Management and the Industry of the Future*, pp. 241–253, Springer, Cham, 2017. doi: 10.1007/978-3-319-72905-3_22.

[3] El-Hakim S.F., Beraldin J.-A., Picard M., Godin G.: Detailed 3D reconstruction of large-scale heritage sites with integrated techniques, *IEEE Computer Graphics and Applications*, vol. 24(3), pp. 21–29, 2004. doi: 10.1109/MCG.2004.1318815.

[4] Grilli E., Dininno D., Petrucci G., Remondino F.: From 2D to 3D supervised segmentation and classification for cultural heritage applications, *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2, pp. 399–406, 2018. doi: 10.5194/isprs-archives-XLII-2-399-2018.

[5] He K., Gkioxari G., Dollár P., Girshick R.: Mask R-CNN, *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/iccv.2017.322.

[6] Hubert J.-F.: *The art of Champa*, Parkstone International, 2011.

[7] Le-Tien M., Nguyen-Tan K., Raffin R.: Matching correspondence between images and 3D model in a reconstruction process, *Journal of Science and Technology: Issue on Information and Communications Technology*, vol. 2(1), pp. 64–69, 2016. doi: 10.31130/jst.2016.29.

[8] Le-Tien M., Nguyen-Tan K., Raffin R.: A Method to Determine the Characteristic of Object Based on 2D/3D Correspondence. In: *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2019. doi: 10.1109/RIVF.2019.8713732.

[9] Lin T.Y., Maire M., Belongie S., Bourdev L., Girshick R., Hays J., Perona P., *et al.*: Microsoft COCO: Common Objects in Context, *CoRR*, vol. abs/1405.0312, 2015. doi: 10.48550/arXiv.1405.0312.

[10] Lowe D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.

[11] Manuel A., Gattet E., De Luca L., Véron P.: An approach for precise 2D/3D semantic annotation of spatially-oriented images for in situ visualization applications. In: *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 1, pp. 289–296, 2013. doi: 10.1109/DigitalHeritage.2013.6743752.

[12] Manuel A., M'Darhri A.A., Abergel V., Rozar F., De Luca L.: A semi-automatic 2D/3D annotation framework for the geometric analysis of heritage artefacts. In: *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, pp. 1–7, 2018. doi: 10.1109/DigitalHeritage.2018.8810114.

[13] Manuel A., Véron P., De Luca L.: 2D/3D Semantic Annotation of Spatialized Images for the Documentation and Analysis of Cultural Heritage. In: C.E. Catalano, L. De Luca (eds.), *14th EUROGRAPHICS Workshop on Graphics and Cultural Heritage*, Eurographics, Genova, Italy, 2016. doi: 10.2312/gch20161391.

[14] Noh Z., Sunar M.S., Pan Z.: A Review on Augmented Reality for Virtual Heritage System. In: *Learning by Playing. Game-based Education System Design and Development*, vol. 5670, pp. 50–61, 2009. doi: 10.1007/978-3-642-03364-3_7.

[15] Özyeşil O., Voroninski V., Basri R., Singer A.: A survey of structure from motion, *CoRR*, vol. abs/1701.08493, 2017. doi: 10.1017/s096249291700006x.

[16] Pierrot-Deseilligny M.: Micmac Interface, http://logiciels.ign.fr/IMG/pdf/docinterface.en.pdf.

[17] Pierrot-Deseilligny M., Clery I.: Apero, An Open Source Bundle Adjusment Software for Automatic Calibration and Orientation of Set of Images, *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVIII-5/W16, pp. 269–276, 2011. doi: 10.5194/isprsarchives-XXXVIII-5-W16-269-2011.

[18] Pierrot-Deseilligny M., De Luca L., Remondino F.: Automated Image-Based Procedures for Accurate Artifacts 3D Modeling and Orthoimage Generation, *Geoinformatics FCE CTU*, vol. 6, pp. 291–299, 2011. doi: 10.14311/gi.6.36.

[19] Pierrot-Deseilligny M., Jouin D., Belvaux J., Maillet G., Girod L., Rupnik E., Muller J., *et al.*: Micmac, apero, pastis and other beverages in a nutshell, *Institut Géographique National*, 2014.

[20] Remondino F.: Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning, *Remote Sensing*, vol. 3, pp. 1104–1138, 2011. doi: 10.3390/rs3061104.

[21] Remondino F., El-Hakim S., Girardi S., Rizzi A., Benedetti S., Gonzo L.: 3D Virtual Reconstruction and Visualization of Complex Architectures – The "3D-ARCH" Project, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, 2009.

[22] Ren S., He K., Girshick R., Sun J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39(6), pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[23] Roy S., Cox I.J.: A maximum-flow formulation of the N-camera stereo correspondence problem. In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 492–499, 1998. doi: 10.1109/ICCV.1998.710763.

[24] Rupnik E., Daakir M., Pierrot-Deseilligny M.: MicMac – a free, open-source solution for photogrammetry, *Open Geospatial Data, Software and Standards*, vol. 2, pp. 1–9, 2017. doi: 10.1186/s40965-017-0027-2.

[25] Stathopoulou E.K., Remondino F.: Multi view stereo with semantic priors, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W15, pp. 1135–1140, 2019. doi: 10.5194/isprs-archives-xlii-2-w15-1135-2019.

[26] Xiong J., Heidrich W.: In-the-Wild Single Camera 3D Reconstruction Through Moving Water Surfaces. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12558–12567, 2021. doi: 10.1109/iccv48922.2021.01233.

Affiliations

Mau Le-Tien
Danang University of Science and Technology, VietNam, tienmauqn@gmail.com

Khoi Nguyen-Tan
Danang University of Science and Technology, VietNam, ntkhoi@dut.udn.vn

Romain Raffin
University of Burgundy, LIB, France, romain.raffin@u-bourgogne.fr

Received: 18.12.2022
Revised: 6.02.2024
Accepted: 7.04.2024