Heritage Conservation Through AI: Caves Object Recognition in Monuments Using STYLE GAN and Faster RCNN Object Detector

Deepak Kumar[1] and Vinay Kukreja[1,*]

ARTICLE INFO

Article history:
Received: 9 May 2023
Revised: 31 August 2023
Accepted: 3 November 2023
Online: 20 Jue 2024

Keywords:
Heritage monuments
Object detection
Generative adversarial networks (GAN)
Faster RCNN
Cultural heritage
Mitigation

ABSTRACT

To explore Indian monuments would require too much time and information. Working on monuments, one of India's most outstanding industries will undoubtedly have a significant impact on the country's economy and growth. Any culture is distinctive in terms of its various monuments, writing, and music. It is difficult to examine methods for identifying these monuments, particularly when considering the histories and tales associated with each monument. Among several heritage monuments, Caves are one of the heritage monuments that represent the nation of any country. Cave identification in heritage monuments through experienced experts takes a lot of time and money for recognition purposes. To overcome the issues of the experienced experts, a combined Style generative adversarial network (SGAN) based Faster Region-based convolutional neural network (FRCNN) has been employed to detect the caves in heritage monuments. The main aim of this study is to recognize the caves along with their boundary location in the form of an anchor box in a real-time manner. For recognizing the caves in heritage monuments, Indian heritage monument images are taken from the UNESCO website. The heritage monument images have been augmented through the SGAN model. The SGAN model increases the dataset size which increases the booster speed as well as the training speed of the FRCNN model. A total number of 3500 images have been used for training and testing purposes in the FRCNN model. Three different overlapping conditions between ground and truth value in terms of Intersection over union (IoU) are determined. The determination result of IoU (50%) produces high Mean average precision (mAP) (94.9%) than other overlapping IoU (60%, 70%, 80%) for caves recognition in heritage monuments. The experimental findings indicate that the model is highly likely to finish the initial screening for Caves heritage recognition.

1. INTRODUCTION

Monuments are extremely complex three-dimensional buildings that are built to honor a person or an occasion and that describe [1] their historical, political, aesthetic, or architectural value, constitute an essential component of cultural heritage. The preservation and promotion of Indian monuments [2] varied with cultural and historical heritage is crucial in the modern, fast-paced world. By visiting the locations and conducting first-hand observations, archaeologists and historians have invested a great deal of time and energy into researching the many monuments and architectural styles. Even monuments describe the art of a historical nation. The Caves [3] are one of the historical monuments. In terms of the shapes and patterns, materials and composition, places and settings of paintings, and rock-cut buildings, caves are authentic. Computer vision techniques have now been used to recognize heritage applications [4] including the classification of monuments, the segmentation of different architectural styles. The process of recognizing and sub-classifying images of cave monuments based on their architectural style is known as monument classification. Classification of caves falls comes under the category of landmark identification. Recognition of the Caves monument [5] is substantially hampered by these significant obstacles in various countries such as India. In previous studies [6], a conventional methods of caves recognition is typically based on in-person visual assessment has been used to identify different types of monuments. Thus, the image processing techniques are just contemporary techniques created to solve traditional methods' challenges and provide important [7] advantages for heritage monument recognition. The machine learning methods use feature extraction techniques for recognizing the monuments and these monuments' features are classified using image-based machine learning methods. Even, the image features are missing while feature extraction

[1]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.*
[*]*Corresponding author:* Vinay Kukreja; Email: onlyvinaykukreja@gmail.com.

techniques. Many researchers also [8] identify monuments by combined image processing and machine learning methods. The authors [9] describe a CBIR approach that uses various regional factors to classify Indian monuments. For the extraction of form features, morphological procedures are used. Texture features are also extracted by gray level co-occurrence matrix (GLCM). The study looked at 500 images from 5 categories like mosques, churches, Hampi temples, Kerala temples and southern temples and found that the top 10 images from each category had a retrieval accuracy of 78-90% for monument identification. Researchers [10] used a genetic algorithm based method to create a classification system for Indian monuments. To train the model, they use 100 images out of which 25 images will be of Taj Mahal, India Gate, Golden Temple and Qutub Minar. Many authors [11] even employed the DCNN approach to extract image features. Using a dataset of 6102 monuments, SVM and KNN are used as classifiers after feature extraction. The purpose of this study is to develop a monument recognition model based on combined machine learning and feature extraction technique. During monument recognition, images of Indian heritage places are categorized as Indo-Islamic, Architecture styles include Colonial, Temple, Rock-Cut, and Cave. Researchers [12] also developed another classifier that uses a method to classify temple-related images in the dataset into Dravidian, Nagara, and Vesala styles. Additionally, we developed a third model using the same dataset to classify images that reflect structure. For the classification of Indian churches and Mughal buildings, authors [13] showed the application of DCNN model. DCNN models are built using the Tensor Flow framework. The DCNN model is applied to 5000 images. The image dataset includes local churches, Taj Mahal, shrines, and attractive minarets, which were selected as leaf nodes. They found that DCNN's local weight sharing plays an important role in achieving 80% accuracy in identifying cultural artifacts in India. Even the author [14] evaluated the performance of ResNet and VGGNet transfer learning models to identify historical monuments. The main goal of the researchers is to develop a proof-of-concept application for automated heritage monuments in Egypt. The ResNet achieves the highest accuracy (88%) than VGG16 transfer learning model for Egyptian monument recognition.

The major contribution of this paper: The main purpose of this study is to recognize the accurate Caves along with its boundary boxes in each heritage image through a combined approach of STYLEGAN (SGAN) and Faster RCNN (FRCNN) object detector network model.

- In the context of cave recognition, SGAN can be used to generate images that help to identify caves that are hidden or partially obscured in heritage monument images.
- In the context of cave recognition, FRCNN can be used to identify caves in images of heritage monuments. The algorithm can be trained on a dataset of images that contain caves and then used to identify caves in new images.
- By combining the SGAN and FRCNN models, it is possible to recognize caves in heritage monuments with high accuracy. Together, they provide a powerful tool for cultural preservation and archaeological research.

The structure of this paper is assembled as: Section 2 presents the methodology which introduces the description of the SGAN and FRCNN model along with its dataset details. Even, the detailed summary of the dataset for experimental setup and results has been described in section 3. The conclusion of this paper for recognizing caves in heritage monuments has been interpreted in section 4.

2. METHODOLOGY

This section defines the methodology for recognizing caves in Indian heritage monuments. The proposed methodology comprises of SGAN and FRCNN model. The proposed methodology in a diagrammatic way has been presented in Figure 1.

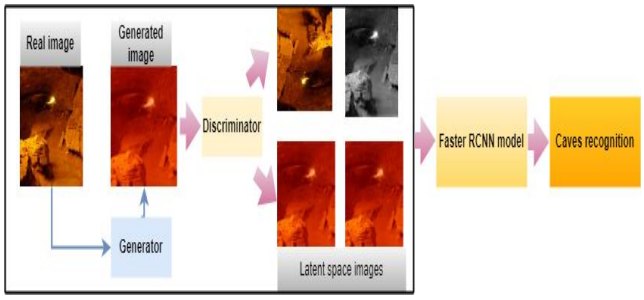

Fig. 1. Proposed methodology for caves recognition.

2.1. Dataset description

For cave recognition in heritage monuments, different regions of India have been selected for image gathering. India is one of the most ancient wonders of most other countries because it has one of the oldest civilizations in the world. There are a significant number of natural and wild caves in India [2] that are tucked away in the deep and dense forest of the country's uncharted valleys and landscapes. The history of vast India, including the regions of Ashoka the Great and the Mauryan, Chalukyan, and Pallava dynasties, is preserved in these caves. Each region has several caves so the heritage caves have been collected with the same type of image dimensions. Images have been collected from secondary sources. As a result of secondary sources [15], a total of 350 heritage monument images have been gathered from different regions of India. The UNESCO tentative list also features a number of the city's immovable cultural assets. All the images have been gathered with the size of 512*512 pixels dpi. Each image has been gathered from different states of India region. The gathered Cave images

have been used for training and testing objectives in SGAN and FRCNN models. The samples of gathered images have been shown in Figure 2.

Fig. 2. Samples of gathered images.

2.2. Style GAN (SGAN)

The Style Generative Adversarial Network [14] is an extension of GAN architecture that suggests defines symbolic changes to the generator model, including the use of a mapping network to map points in latent space to an intermediate latent space. StyleGAN design is based on a modified version of the progressive growing GANs (PGGANs) architecture, which trains the network with a generator and a discriminator. StyleGAN, on the other hand, introduces several critical advances to improve the quality and diversity of the generated images.

To begin, SGAN employs a mapping network to convert a random vector is considering into a styled vector, which influences the overall aesthetic of the generated image. This enables more precise control over the image production process. Second, SGAN injects the style vector into the generator network using a technique known as adaptive instance normalization (AdaIN). AdaIN adjusts the mean and variance of the features to match the style vector by applying an affine transformation to the feature mappings of each layer in the generator network. This contributes to the resulting image having the desired style. Third, SGAN has a multi-scale generator and discriminator architecture, with each scale comprised of a succession of convolutional layers that increase in size sequentially. As a result, the network can produce high-resolution images with precise features.

Finally, SGAN introduces a truncation trick that allows the network to generate more diverse images by restricting the latent code range. The network is forced to generate images that are closer to the mean of the training data by truncating the latent coding. The generator and discriminator are the two main phases of SGAN. The generator network is responsible for making new images from random noise vectors, whilst the discriminator network tries to differentiate between real and artificially generated images.

The mathematical equations of StyleGAN can be quite complex, but the basic structure of the generator network can be represented by the following equation:

$$G(I,V) = A\sum_{i=0}^{n} Vn * T(z) + Bn \quad (1)$$

where, I and V are the input noise vector along with the style vectors that control how the image is generated. The combined T(z) multiplication of the activation function and mapping is performed on the convolutional layer. The bias term is denoted with Bn.

StyleGAN employs the same discriminator network as other GAN architectures, which can be described by the following equation:

$$D(X) = A\sum_{i=0}^{n} Vn * T(z) + Bn \quad (2)$$

The generated image (X) is applied to the discriminator for real/fake determination.

A truncation trick introduced by StyleGAN allows the network to produce more diverse images by limiting the latent code range. It helps prevent the creation of extreme or unrealistic images by truncating the latent code and forcing the network to generate images closer to the mean of the training data. The objective function of SGAN consists of two loss functions: adversarial loss and perceptual loss.

- **Adversarial loss:** The adversarial loss ensures that created images are indistinguishable from real images. It is based on a discriminator network that attempts to determine if a picture is real or fake. The generator network is trained to generate images that the discriminator classifies as real, hence minimizing adversarial loss.
- **Perceptual loss:** The perceptual loss ensures that the created images have the appropriate visual characteristics, such as style and structure. It is built on a pre-trained perceptual network that compares the generated images to a set of target images in terms of high-level visual attributes. The representation of all parameters has been shown in Table 1.

Table 1: Network parameters of SGAN

Parameter name	Type	Value
Filter size	Input	3
Image size	Input	256*256
Reshape	Reshape	128*128
Leaky relu	Leaky relu	Max(0,x)
Activation function	Input	Sigmoid
Stride	Input	2*2

2.3. Architecture of FRCNN model

Through a selective search technique [16], FRCNN extracts the regions from the image. The extraction of regions from a picture using the selective search method takes a long time. The Resnet model serves as the foundation for feature

extraction in the FRCNN model, which also classifies object suggestions. The FRCNN model's performance is negatively impacted by the selective search strategy that it uses for region extraction. The target classless object is the sliding window. The structure of the faster-RCNN model has been shown in Figure 3.

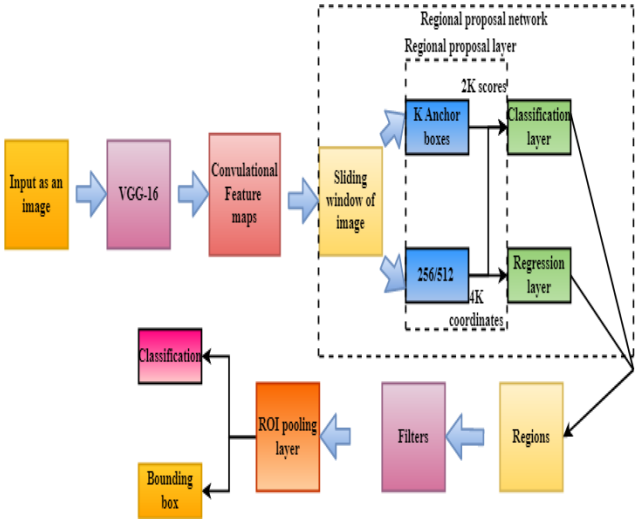

Fig. 3. Structure of FRCNN model.

- **Region proposal network:** The creation of a set of proposals is RPN's principal objective. The RPN module produces a label for each object as well as a probability of class for each object. When first RPN object locations are anticipated, these are in charge of supplying a specified set of bounding boxes with varying widths and dimensions used as a reference. In the RPN, anchor box bounding boxes and objectness scores are predicted using fully convolutional networks. The RPN calculates the objectness score and bounding box offsets for each anchor box [17]. The objectness score is calculated through the probability of each anchor (An) containing an object and represented as:

$$P(An\{i\}) = Sig(Os\{i\}) \quad (3)$$

An represents the anchor of i objects and Os shows the objectness score obtained in the region proposal network. The sigmoid (Sig) is applied to the objectness score. Once, the objectness score has been calculated, it is easily determined the bounding boxes offsets of predated regions which have been calculated as:

Let t_{i} = (t_{x, i}, t_{y, i}, t_{w, i}, t_{h, i}) be the predicted offsets for anchor i, where t_{x, i} and t_{y, i} are the predicted translations and t_{w, i} and t_{h, i} are the predicted widths and heights. Then, the predicted bounding box for anchor i is given by:

B_{i} = (x_{i}, y_{i}, w_{i}, h_{i}) (4)

where, x_{i} = x_{a,i} + t_{x,i}w_{a,i}, y_{i} = y_{a,i} + t_{y,i}h_{a,i}, w_{i} = w_{a,i}e^{t_{w,i}}, and h_{i} = h_{a,i}e^{t_{h,i}}. Here, (x_{a,i}, y_{a,i}, w_{a,i}, h_{a,i}) are the coordinates and size of anchor i.

- **Detection network:** The detection network classifies the RPN's component proposals into several object categories and adjusts their bounding boxes. The detection network performs the classification as well as bounding box regression on each region that has been extracted in the RPN region. The formulated equation of classification as well as bounding box regression have been as follows as:

$$Os\{i, c\} = Softmax(p\{i, c\} \quad (5)$$

where, p{i,c} is the probability of an object belonging to category c. The bounding boxes offsets of each object are calculated as:

q_{i} = (q_{x,i}, q_{y,i}, q_{w,i}, q_{h,i}) (6)

The q_{i} the predicted offsets for object *i*.

The faster R-CNN model is constructed up of layers that perform feature extraction, region proposal creation, feature pooling, object classification, and bounding box regression, followed by post-processing processes for removing redundant detections. The parameters of FRCNN have been represented in Table 2.

Table 2: Outline of FRCNN model

Parameter name	Type	Size
Batch size	Input	8
Learning rate	Input	0.001
Epochs	Input	100-500
Self non maximum suppression (NMS) threshold	Input	0.3
Self-score threshold	Input	0.7
Gradient descent	Optimizer	Adam
Image size	Convolution layer	256*256

3. RESULTS AND DISCUSSION

3.1. Experimental setup and implementation

The experiments of SGAN and FRCNN have been performed on an NVIDIA RTX GPU server in python based Jupyter notebook. On each image, the feature extraction method was performed for Caves recognition. For speedier computations, all of the images in the dataset have been scaled to 256*256 sizes. The feature extraction module employed an object identification method was used to classify the acquired feature vector using a Resnet model. During classification, the FRCNN approach extracts Caves features from each image.

3.2. Evaluation metrices

In this study, two different types of experiments have been

performed for image generation. The SGAN has been trained with 500 epochs to generate more than 10 times the original dataset. The generated images dataset has been used for training and testing purpose in the FRCNN model. After each epoch, the weights of the generator and discriminator models were modified to generate synthetic images that were as near to real images as possible. Several evaluation metrics can be used for evaluating the performance of an SGAN model.

- **Precision and Recall (P&R):** Precision measures the percentage of generated images that are correctly classified as generated, while recall measures the percentage of real images that are correctly classified as real. A good model should have high precision and recall values. The equations of precision and recall is written as:

  Precision= ((Tp)/(Tp+Fp)) (7)

  Recall= ((Tp)/(Tp+Fn)) (8)

- **Intersection over union (IoU):** IoU scores range from 0 to 1 and represent the overlap between predicted bounding boxes and ground truth boxes.

3.3. Performance of SGAN model

The SGAN obtains more than ten times images for improving recognition accuracy. Several layers of neural networks have been embedded to improve the Caves recognition accuracy in heritage monuments. The images have been trained through a generator with different image translations. Even, the 3500 images have been generated in the training phase of SGAN, but due to noise filters in the discriminator, the one-part means 10:1 images have been reduced. The discarded images have a low noise ratio in SGAN. Before using the AdaIN approach, each activation map has Gaussian noise applied to it. Based on the scaling parameters of that layer, each block's unique noise sample is assessed. The SGAN uses five different convolution layers with different stride sizes. As the SGAN generates more than 10 times the original dataset. The batch size for generating images is set to 1 with a learning rate of 0.001ms. The samples of SGAN images have been shown in Figure 4. The performance of the SGAN model in terms of their accuracy has been shown in Figure 5 and the accuracy of SGAN with the different number of epochs has been shown in Figure 6.

3.4. Performance of FRCNN model

For Caves recognition in heritage monuments generated images, the FRCNN model is employed. The object detection network model helps to recognize the Cave in the image along with bounding box size (32*32) pixels. Two different transfer learning models of residual network models have been employed in FRCNN for feature maps. Numerous epochs and learning parameters have been used to measure the performance of the FRCNN model. On the backbone feature map, a 3 x 3 convolution with 512 units is first applied to provide 512-d feature maps for each place. Among from generated images, 2450 and 1050 images have been used for training and testing purposes respectively. Among 2450 images, 300 images were selected as random in manual annotation which is considered as ground truth image. Even, the Caves predicted image with anchor box is considered as predicted image. The FRCNN model extracts the patches directly through annotated images which have been used to calculate the threshold IoU. A total number of 6000 patches along with the anchor box have been extracted by the FRCNN model which has been shown in Figure 7.

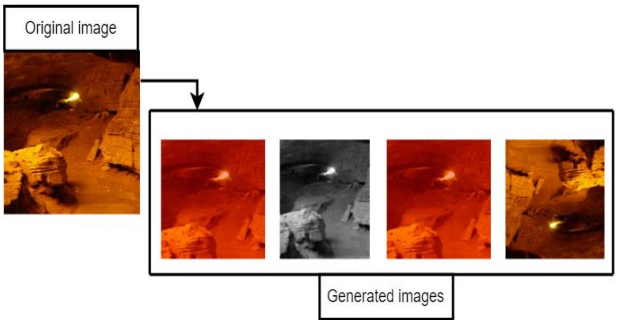

Fig. 4. Samples of SGAN images.

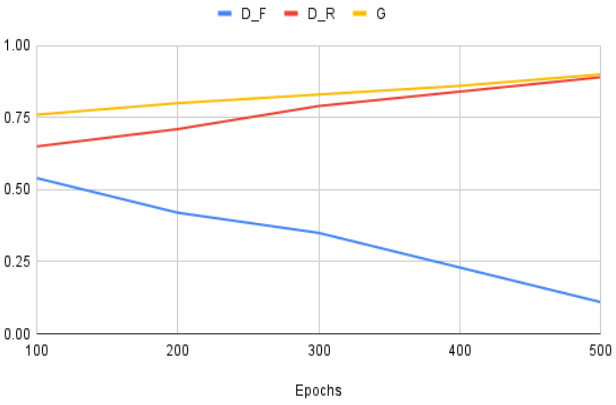

Fig. 5. Performance of SGAN model where D_F/R defines the discriminator of fake and real images whereas G shows the Generator.

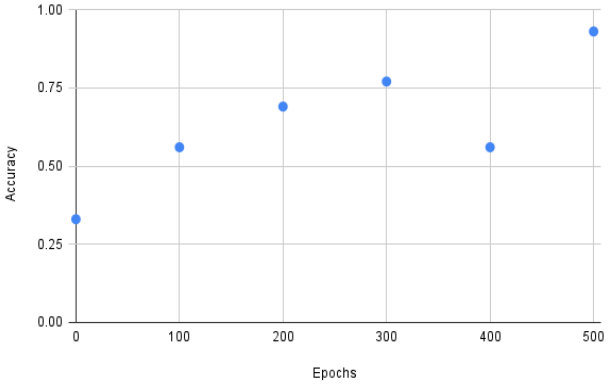

Fig. 6. Accuracy of SGAN model with different numbers of epochs.

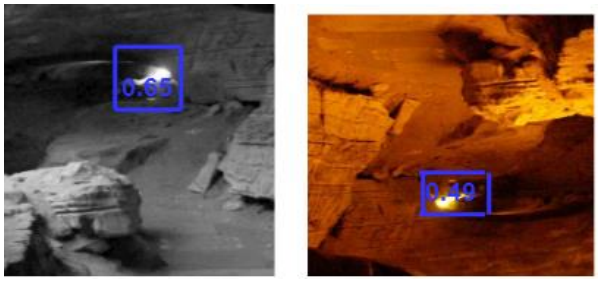

Fig. 7. FRCNN model results along with bounding boxes.

The performance of FRCNN is measured through ground truth and predicted image which is known as mean average precision (mAP). Several IoU values for measuring the performance of FRCNN in caves recognition. The threshold value of IoU with 0.50 achieves high mAP (94.9%) than the three different threshold values in IoU. The batch size in the FRCNN model is set to 8. Thus, the caves detected in the generated image with a threshold value of 0.50 achieve high performance than other threshold values of IoU. The performance of the FRCNN model in terms of mAP with different numbers of IoU has been shown in Figure 8.

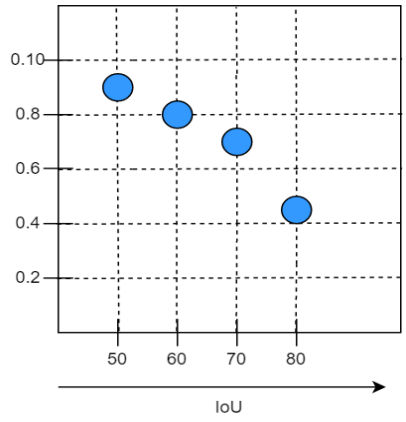

Figure 8: Different IoU values for measuring mAP.

4. CONCLUSIONS

This paper proposed a combined approach of SGAN and FRCNN object detection network model for Caves detection in heritage monuments. The combined approach is employed on real-time datasets which are gathered from several regions of India. The heritage monument images have been gathered from secondary sources. The secondary sources are too less and decrease the training as well validation accuracy of the FRCNN model. To overcome the dataset issues, the SGAN is applied for generating images. Even, SGAN adds some noise filters for generating images. The SGAN employed 10:1 for generated as well as discriminated datasets. Even the SGAN generates 500 images which have been useful for training and testing purposes in the FRCNN model for Caves recognition. Our model was first trained before moving on to the cross-validation phase, where we employed the two cross-validation procedures with 25 iterations per image. Three different overlapping conditions between ground and truth value in terms of Intersection over union (IoU) are determined. The determination result of IoU (50%) produces high Mean average precision (mAP) (94.9%) than other overlapping IoU (60%, 70%, 80%) for caves recognition in heritage monuments. StyleGAN may help researchers and historians in illustrating how a cave might have appeared in the past. This can help us understand how ancient traditions used the cave and how it changed through time. The combined SGAN and FRCNN model can aid in the conservation and preservation of cave paintings and other artifacts by allowing accurate digital representations of them to be created. Furthermore, the integrated technique aids in the creation of detailed maps of heritage monuments, including the position and distribution of caves within them, which can be useful for conservation and management.

REFERENCES

[1] V. Barrile, E. Bernardo, and G. Bilotta. An Experimental HBIM Processing: Innovative Tool for 3D Model Reconstruction of Morpho-Typological Phases for the Cultural Heritage. 2022. Remote Sensing 14(5):1288–1302.

[2] A. D. P. Paul, A. Jyoti, S. Ghose, K. Aggarwal, N. Nethaji, S. Pal. 2021. Machine learning advances aiding recognition and classification of Indian monuments and landmarks. In Proceedings of the IEEE conference on International Conference on Electrical, Electronics and Computer Engineering. Dehradun, India, 11-13 November.

[3] D. A. R. Moroni, A. Victoria, V. Gnezdilova. 2015. Geological heritage in archaeological sites: case examples from Italy and Russia. Proceedings of the Geologists Association 126(2): 244–251.

[4] S. I. Sharma, Saurabh, P. Aggarwal, A. N. Bhattacharyya. 2018. Classification of Indian monuments into architectural styles. In Proceedings of the Springer conference on National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics. Mandi, India, 22-24 December.

[5] C. G. A. Giuseppe, F. Falchi. 2015. Fast image classification for monument recognition. Journal on Computing and Cultural Heritage 8(4): 1–25.

[6] N. Q. UOC, N. T. Duong, L. A. Son, and B. D. Thanh. 2022. A Novel Automatic Detecting System for Cucumber Disease Based on the Convolution Neural Network Algorithm. GMSARN International Journal 16: 295-301.

[7] I. I. Hatir, M. Ergün, M. Barstuğan. 2020. Deep learning-based weathering type recognition in historical stone monuments. Journal on Computing and Cultural Heritage 45: 193–203.

[8] F. H. I. Hesham, S. R. Khaled, D. Yasser, S. Refaat, N. Shorim. 2021. Monuments recognition using deep learning vs machine learning. In Proceedings of the IEEE conference on 11th annual computing and communication workshop and conference. NV, USA, 27-30 January.

[9] V. K. P. Desai, P. J. Pujari, N. H. Ayachit. 2013. Classification of archaeological monuments for different art forms with an application to CBIR. In Proceedings of the IEEE conference on International Conference on Advances in Computing, Communications and Informatics. Mysore,

India, 22-25 August.

[10] S. I. Kumar, A. S. Bhowmick, N. Jayanthi. 2021. Improving Landmark Recognition using Saliency detection and Feature classification. In Proceedings of the Springer conference on National Conference on Digital Techniques for Heritage Presentation and Preservation. Delhi, India 16-19 Novmber.

[11] M. S. Bhatt, T. P. Patalia. 2015. Genetic programming evolved spatial descriptor for Indian monuments classification. In Proceedings of the IEEE conference on International Conference on Computer Graphics, Vision and Information Security. Bhubaneswar, India, 02-03 November.

[12] P. Shukla, B. Rautela, A. Mittal. 2017. A computer vision framework for automatic description of Indian monuments. In Proceedings of the IEEE conference on 13th International conference on signal-image technology & internet-based systems. Jaipur, India, 04-07 December.

[13] K. Saliankham, B. Douangngeune, and Z. Bin. 2014. Tourist Satisfaction on Tourism Growth and Tourism Site Development in VangVieng District of Vientiane Province, Lao PDR. GMSARN International Journal 8(3): 85-88.

[14] F. H. I. Yasser, A. M. Salama, A. Amr, L. E. Yehia, S. Refaat. 2022. Egyart_classify: an approach to classify outpainted Egyptian monuments images using GAN and ResNet. In Proceedings of the IEEE conference on 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference. Cairo, Egypt, 08-09 May.

[15] B. Launge. 2020. Primary vs Secondary Data [On-line serial]. Retrieved October 09, 2022 from https://www.formpl.us/blog/primary-secondary-data.