

Archaeological Site Image Content Retrieval and Automated Generating Image Descriptions with Neural Network

Sathit Prasomphan^(✉)

Department of Computer and Information Science, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand
ssp.kmutnb@gmail.com

Abstract. This research presents a novel algorithms for generating descriptions of stupa image such as stupa era, stupa architecture by using key points generated from SIFT algorithms and learning stupa description from the generated key points with artificial neural network. Neural network was used for being the classifier for generating the description. We have presented a new approach to feature extraction based on analysis of key points and descriptors of an image. The experimental results for stupa image content generator was analyze by using the classification results of the proposed algorithms to classify era and architecture of the tested stupa image. To test the performance of the purposed algorithms, images from the well-known historical area in Thailand were used which are image dataset in Phra Nakhon Si Ayuttha province, Sukhothai province and Bangkok. The confusion matrix of the proposed algorithms gives the accuracy 80.67%, 79.35% and 82.47% in Ayutthaya era, Sukhothai era and Rattanakosin era. Results show that the proposed technique can efficiently find the correct descriptions compared to using the traditional method.

Keywords: Image content retrieval · Neural network · SIFT algorithms · Feature extraction

1 Introduction

Thailand or Siam is one of a country that has a long history in the Southeast Asia. Several temples, palaces, or residences had been developed in each era. However, in present, because of the long time of that place some importance things are broken. Some parts of that place still remain which is interesting to the new young generation who interested in an archaeological site. One of archaeological site which is most importance for studying is stupa. Several architecture of the stupa was created. If we know the stupa architecture, more details of that stupa can be described.

To study the stupa architecture, the shape of each stupa will be considered by finding the shape similarity between the target image and all of images in the database. The purpose of these algorithms is to get information from the interested image. These algorithms also known as image content retrieval algorithms. Nowadays, there are several techniques for getting information from an image [4–6]. For example, a machine learning algorithm was

applied such as recursive neural network [2] or convolution neural network [3] which introduced by Richard Socher et al. [2]. In this technique the relevance between image and the sentence was used. Another rule which used the recurrent neural network was presented by Andrej Karpathy and Li Fei-Fei [1]. The language and image relationship was studied. Different sources of image such as Flickr8K, Flickr30K and MSCOCO were used in the research. The problem of using these categories is the complicated of model and input attribute to be used to find the relationship between language structure and image. Another group of model was suggested by using the scoring method to find the relationship between sentence space, image space, and meaning space. This technique was suggested by Ali Farhadi et al. [8]. However, there is some confusion to create the relationship between all of these spaces.

From the limitation of these techniques, the new algorithm for generating image description is developed. The archaeological site image especially stupa image was studied. The research aims to find the architecture of the stupa image which can be used this architecture to generate other descriptions. For example, era of the stupa, when the stupa was builds or where the stupa location frequently occurs. In this research, the combination between SIFT algorithms and neural network was studied. SIFT algorithms was used for generating key points of an image. Neural network was used for classifying the architecture of stupa image. The input attributes of neural network comes from key point.

The remaining of this paper is organized as follows. At first, we show the stupa characteristics. Next, the combination between SIFT algorithms and neural network is discussed in Sect. 2. Results and discussion are discussed in Sect. 3. Finally, the conclusions of the research are presented in Sect. 4.

2 Proposed Algorithms

In this section, the characteristics and examples of stupa image occurred in Thailand were described. The SIFT algorithms for generating key points was detailed. The neural network for classifying architecture of stupa was shown. Finally, the process for generating description to the stupa image was explained.

2.1 Characteristics of Stupa

The stupa or known as chedi is a Buddhist memorial monument usually housing holy relics associated with the Buddha. The stupa shape was created from the shape of an ancient Indian burial mound [7]. Several styles of stupa in Thailand occurred. The architecture's category of stupa in Thailand can be divided into three styles. These categories are divided based on the period of that architecture [7]. However, the overlapped architecture between periods also occurs. The stupa architecture in the Sukhothai period is grouped into the lotus blossom style, the bell-shaped style, the Prang style, etc. The stupa architecture in the Ayutthaya period is classified into the bell-shaped style, the Prang style, etc. Finally, The stupa architecture in can be classified into these categories: the square wooden chedi style, the Prang style, etc. The example of stupa architecture in each period (Sukhothai period, Ayutthaya period, and Rattanakosin period) are shown in Figs. 1, 2, and 3 in orderly.



Fig. 1. Example of stupa in the Sukhothai era (a) the lotus blossom style (b) the bell-shaped style (c) the Prang style (d) the Chomhar style [7].



Fig. 2. Example of stupa in Ayutthaya era [7].



Fig. 3. Example of stupa in Rattanakosin era [7].

2.2 Scale Invariant Feature Transforms (SIFT) Algorithms

Scale-invariant feature transform (SIFT) [9, 10] is an algorithm in computer vision to detect and describe local features in images. The algorithm was published by David Lowe in 1999 [9]. SIFT [9, 10] is an algorithm for calculating the interesting point (key point) in an image and determining the features of key point. The characteristics of these features are invariant to image scale and rotation. They are robust to addition of noise, distortion or brightness. SIFT technique has an advantage that it is not based on a scale or the orientation angle position, which can be used to compare the features more easily and accurately, more precisely. Key point generally refers to a pixel in the image which is changed the orientation from two-dimensional of brightness levels surrounded a key point. The algorithm for findings SIFT key point in the picture is as follows. Firstly, detect key point from the input image (key point detection). In this step, we get a series of x, y coordinates of a key point which is used to provide a description of the key point. For the next step, part of the explanation of key point (key point description) is calculated in form of a vector explain (descriptor vector), which is calculated from the brightness of the pixels in the area surround key point. These vectors are used to describe the series of identity when it appears in photos. After the key point is generated, the process of learning or learning phase is performed. It is the process for matching between the most two similar images. Following are the major stages of computation used to generate the set of image features:

1. **Scale-space extrema detection:** The process for detect the most importance feature of image which not depends on the size or orientation of an image. The process is done by burring image with Gaussian function in each octave. In each octave consists of several burring scale by burring normal scale and increasing scale parameter which effect to the burring image. It is done with octave which in each step reduce the size to half of old octave.
2. **Key point localization:** From the previous process each scale space image will be used to find the key point of image by using the octave with the following equation. The process for getting key point is local maxima/minima in DoG image and finds sub pixel maxima/minima. In the previous process, it will get many key points. The process for reduce number of key points is removing low contrast feature and removing edges.
3. **Orientation assignment:** After specify the key points; it will calculate magnitude and orientation of gradient surrounded the key points for becoming descriptors of key points. After that, this magnitude and gradient of pixel surround key points will be used for generating histogram which x is degree and y is gradient.
4. **Key point descriptor:** Create 16×16 windows and divided to 4×4 window with 16 sets which in each to calculate magnitude and gradient and create histogram with 8 bins. After finished this process the feature vector will be 128 and will be used to the next processes.

The process in each step can be shown in Figs. 4, 5 and 6.

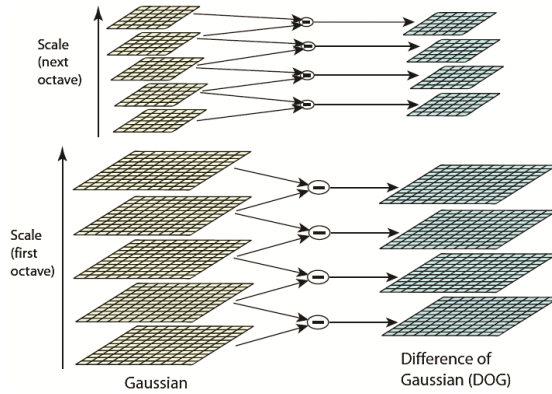


Fig. 4. Difference of Gaussian is obtained as the difference of Gaussian blurring of an image with two different σ , let it be σ and $k\sigma$ [9].

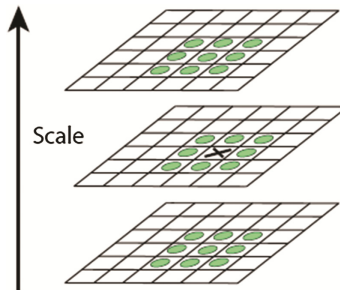


Fig. 5. Once this DoG are found, images are searched for local extrema over scale and space. One pixel in an image is compared with its 8 neighbours as well as 9 pixels in next scale and 9 pixels in previous scales. If it is a local extrema, it is a potential keypoint. It basically means that keypoint is best represented in that scale [9].

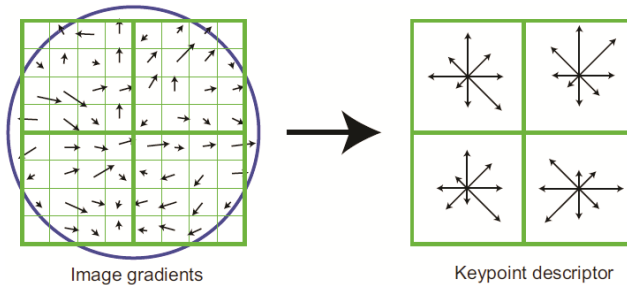


Fig. 6. After specify the key points; magnitude and orientation of gradient surrounded the key points for becoming descriptors of key points will be calculated [9].

2.3 Image Content Generator System

The processes for generating stupa image contents as shown in Fig. 7 are described in the following section.

1. **Pre-processing process:** In this step, the training stupa images and the testing stupa image is collected in which all of training stupa image come from the well-known historical area in Thailand (Phra Nakhon Si Ayutta province, Sukhothai province and Bangkok). The testing stupa image is taken from camera. The similarity between the training stupa images and the testing stupa image are compared. The transformation from RGB image to gray scale image is necessary. Moreover, to improve the image quality by using the image enhancement is required in this step.
2. **Laplacian algorithm for edge detection:** In this research, the Laplacian algorithm is used for being the technique for detecting edge in a stupa image. The different of intensity of nearest points is measured. Finding the line surrounding the object inside an image is required. We use this process for finding the edge that pass through or near to the interested point. In low quality image, the different of intensity of nearest point is low which effect to the process of finding edge. The foreground and background of brightness may be not covering all of image. In this case the edge may be blurred compared to using the high different of intensity. The information from edge detection is used for the next process.
3. **Feature extraction:** The key point generating from the SIFT algorithm is the main input features in this research. In each stupa image, the vector of key point is used. The key point can be identified the property in each image. Magnitude and orientation of gradient surrounded the key points for becoming descriptors of key points is calculated. 128 attributes from the 128 key point's descriptor are generated. Our assumption in this step is the most two similar images will have the same key point.
4. **Stupa architecture classification:** In this paper, neural networks have been created in order to distinguish for key points of an image. Use the input attribute from the key points and descriptors generated from SIFT algorithms for becoming the input to neural network. We use 128 attributes from the 128 key point's descriptor to be input attribute. Train on only the training set by setting the stopping criteria and the network parameters. Feed-forward multilayer neural network with back propagation learning algorithms was used. The network consists of one input layer, one hidden layer, and one output layer. Set of inputs and desired output belongs to training patterns were fed into the neural network to learn the relationship of data. The process in hidden layer is to adjust weight which connected to each node of input. The root mean square error is calculated from desired output and its calculated output. If the error is not satisfied with the predefined values, it will propagate error back to the former layer. This will be done from the direction of the upper layer towards the input layer. This algorithm will adjust weight from initial weight until it gives the satisfied mean square error. We use this technique for matching the architecture of the stupa between the testing stupa image and all of stupa reference images in database. Inside the stupa reference images database, it will contain stupa image in several architectures, the descriptions of the reference image, stupa architecture,

stupa era and other importance details of stupa. The algorithm will select the most matching architecture of testing stupa image and the reference stupa images.

5. **Stupa description generating process:** After the matching process finished, there is the stupa description generating process. The algorithms will use image descriptions inside the database to show and set description to the testing image.

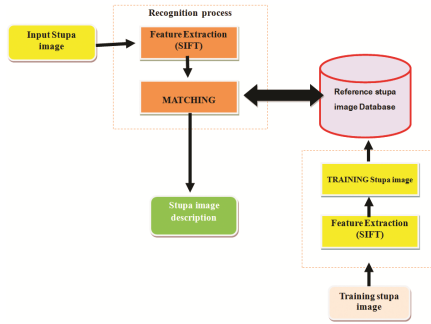


Fig. 7. The process of image content generator system.

3 Experimental Results and Discussions

3.1 Experimental Setup

The stupa image datasets in this research came from the historical area at Phra Nakhon Si Ayuttha province, Sukhothai province and Bangkok in which the most stupa architecture in that area is Ayutthaya period, Sukhothai period and Rattanakosin period in ordering. The constraint in each image must have least 80% of image coverage. The brightness of image will be effect to the classification results. So, the pre-processing of image was required. The number of image using in the research was shown in Table 1.

Table 1. Number of samples in the experiments

Era	Number of image
Ayutthaya	67
Sukhothai	44
Rattanakosin	46
All	157

3.2 Performance Indexed

3.2.1 Confusion Matrix

Tables 2, 3 and 4 are the confusion matrices, which are widely used graphical tools that reflect the performance of an algorithm. Each row of the matrix represents the instances of a predicted class, while each column represents the instances of an original class. Thus, it is easy to visualize the classifier’s errors while trying to accurately predict each

original class' instances. Percentage of the test data being classified to the original stupa image was shown inside the table.

Table 2. Confusion matrix of the proposed algorithms by using sift with neural network

Architecture	Image recognition		
	Ayutthaya	Sukhothai	Rattanakosin
Ayutthaya	80.67	4.43	7.35
Sukhothai	4.07	79.35	3.29
Rattanakosin	7.43	3.79	82.47

Table 3. Confusion matrix of the KNN algorithm

Architecture	Image recognition		
	Ayutthaya	Sukhothai	Rattanakosin
Ayutthaya	60.82	12.69	26.49
Sukhothai	20.15	62.78	17.07
Rattanakosin	15.67	21.02	63.27

Table 4. Confusion matrix of euclidean distance

Architecture	Image recognition		
	Ayutthaya	Sukhothai	Rattanakosin
Ayutthaya	70.28	12.88	16.84
Sukhothai	13.58	75.22	11.20
Rattanakosin	17.57	8.79	73.64

3.2.2 Comparing Algorithms

In this paper, Euclidean distance, SIFT algorithms with neural network, and k-nearest neighbors were used for comparing. We compared each of this method with the proposed algorithms which used to predict the era of that architecture for example Ayutthaya era, Sukhothai era and Rattanakosin era. We used the combination of SIFT algorithms and neural network algorithms.

4 Experimental Results and Discussions

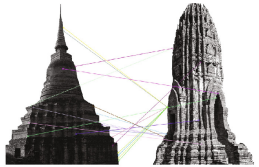
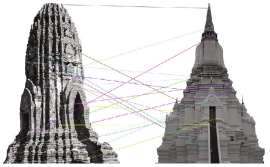
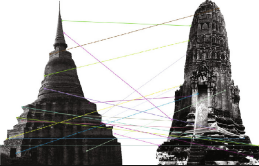
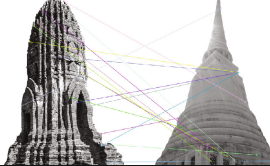
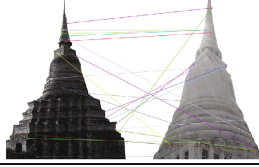
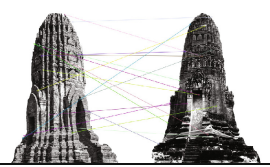
4.1 Experimental Results

To show the accuracy of the proposed stupa image content generator, the stupa architecture classification results were analyzed. Euclidean distance, SIFT algorithms with neural network, and k-nearest neighbors were used for comparing. The classification accuracy was shown in Tables 2, 3 and 4. Table 2 showed the confusion matrix of the proposed algorithms which was the combination between SIFT algorithms and neural network. It gives the accuracy 80.67%, 79.35% and 82.47% in Ayutthaya era, Sukhothai

era and Rattanakosin era. Table 3 showed the confusion matrix of using k-nn algorithms. This algorithm gives the accuracy 60.82%, 62.78% and 63.27% in Ayutthaya era, Sukhothai era and Rattanakosin era. Table 4 showed the confusion matrix of using the Euclidean distance. This algorithm gives the accuracy 70.28%, 75.22%, and 73.64% in Ayutthaya era, Sukhothai era and Rattanakosin era. The experimental results confirms that using key points of image which is generated from the SIFT algorithms and using the neural network for training the key points to get the period of a stupa architecture inside an image can be successfully used. Also this predicted architecture’s era can be used for generating the description of image as shown in table. The proposed algorithms give the accuracy about 80–85% in average.

Example of using SIFT algorithms for generating key point to match the architecture of stupa with different comparing algorithms for example, Euclidean distance, neural network, and k-nn algorithm can be shown in Table 5.

Table 5. Examples of using sift algorithms for generating key point.

Algorithms	Example	
	Rattanakosin era	Ayutthaya era
SIFT with Euclidean distance		
SIFT algorithms with k-nearest neighbors		
SIFT algorithms with neural network		

5 Conclusions

This research proposes a new algorithm for getting image descriptions via key point descriptor from SIFT algorithms. The key point descriptors of an image are used to distinguish identity of an image. The important feature was extracted from a stupa image. A basic architecture of neural network to reduce the difficulty of the classification process is used. Number of features to be sent to neural network is reduced. The algorithm was tested with a stupa image getting from the real world in historical area in Ayutthaya province, Sukhothai province and Bangkok. The confusion matrix of the proposed algorithms gives the accuracy 80.67%, 79.35% and 82.47% in

Ayutthaya era, Sukhothai era and Rattanakosin era. Results show that the proposed technique can efficiently find the correct descriptions compared to using the traditional method.

Acknowledgment. This research was funded by King Mongkut’s University of Technology North Bangkok. Contract no. KMUTNB-59-GEN-048.

References

1. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: The 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–14 (2015)
2. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. *TACL* **2**, 207–218 (2014)
3. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint [arXiv:1409.2329](https://arxiv.org/abs/1409.2329) (2014)
4. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *TACL* **2**, 67–78 (2014)
5. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Intell. Res.* **47**, 853–899 (2013)
6. Su, H., Wang, F., Yi, L., Guibas, L.J.: 3D-assisted image feature synthesis for novel views of an object, CoRR <http://arxiv.org/abs/1412.0003> (2014)
7. Charuwan, C.: *Buddhist Arts of Thailand*, Buddha Dharma Education Association Inc., Tullera (1981)
8. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010. LNCS*, vol. 6314, pp. 15–29. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15561-1_2](https://doi.org/10.1007/978-3-642-15561-1_2)
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proceedings of 7th International Conference on Computer Vision (ICCV 1999)*, Corfu, Greece, pp. 1150–1157 (1999)
10. Lowe, D.G.: Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)