2017 International Conference on Big Data, IoT and Data Science (BID)
Vishwakarma Institute of Technology, Pune, Dec 20-22, 2017

Image based Indian Monument Recognition using Convoluted Neural Networks

Aradhya Saini[1], Tanu Gupta[1], Rajat Kumar[2], Akshay Kumar Gupta[2],Monika Panwar[3], Ankush Mittal[4]
[1]Indian Institute of Technology, Roorkee ,[2,3]Raman Classes ,[4]Graphic Era University
aradhya.saini91@gmail.com, tanugupta.100109@gmail.com, rajat.techy.02@gmail.com

***Abstract*: Monument recognition is a challenging problem in the domain of image classification due to huge variations in the architecture of different monuments. Different orientations of the structure play an important role in the recognition of the monuments in their images. The paper proposes an approach for classification of various monuments based on the features of the monument images. The state-of-the-art Deep Convolutional Neural Networks (DCNN) is used for extracting representations. The model is trained on representations of different Indian monuments, obtained from cropped images, which exhibit geographic and cultural diversity. Experiments have been carried out on the manually acquired dataset that is composed of images of different monuments where each monument has images from different angular views. The experiments show the performance of the model when it is trained on representations of cropped images of the various monuments. The overall accuracy achieved is 92.7%, using DCNN, for a total of 100 different monuments that have been considered in the dataset for classification.**

Keywords - Monument recognition, Convolutional Neural Networks, Deep Learning.

I. INTRODUCTION

A monument implies a structure that has been constructed in order to commemorate a person , event or which has become an important part to a social group as a part of them remembering historic times or cultural heritage, or as an example of the historic architecture [1]. The people belonging to the various cultures, castes, creeds and religions take pride in their culturally rich heritage bestowed upon them in the form of monuments. The term 'monument' is often applied to the buildings or structures that have been considered as examples of an important architectural and/or cultural heritage.

Monuments are also the tourist destinations in any country. They even are representations of great achievements present in art and architecture. It is therefore important to preserve them for the purpose that we can continue to enjoy their majestic views and the future generations too can learn from them. They are a part of India's vast heritage because they show the historical influence of any country with respect to its citizens. These are the important and visual source of analyzing the history of India, very precisely. In India, there are lots of monuments which are connected with the religious feelings of the people. One example is The Sanchi Stupa-incarnating the presence of Buddhism teachings. Second example is of Khajuraho Temples- where both Hinduism and Buddhism feelings have been amalgamated.

Monument recognition deals with classifying the query monument image into its respective label. Monument recognition has many applications in different sections of our society.

The technique used in this paper can be used to retrieve the label of any monument image given to any search engine within a small frame of time. Tourism industry can also grow on the mobile application via recognition of various monuments in different states. But with increase in the dataset, identification becomes difficult and accuracy decreases with basic algorithms. Therefore, better algorithms need to be implemented.

Classifying a monument is difficult because many images of a single monument are to be used to train the system which are very much different from each other in their orientations. The differences in the images of monuments has been shown in Figure 1, which aptly shows the variations.

Monument recognition is a good concept though still not much work has been done in this domain of image classification. Moreover, the noise present in the images in the form of trees, people, animals, decorations etc. often leads to less accuracy. These variations make monument recognition a challenging problem.

Fig 1. Monument images dataset. The variations in the architecture of the monuments can be seen clearly.

978-1-5090-6593-6/17/$31.00 ©2017 IEEE

The organization of the paper is as follows. Section 2 comprises of the proposed approach, methodology and various monument recognition approaches discussed in detail. The dataset description and the results have been covered in the Section 3. Section 4 comprises of conclusions and future work.

II. LITERATURE REVIEW

Monument Recognition is a novel idea. There has not been much work carried out in this field. Some of the authors have shown different strategies for monument recognition. We have adopted a novel strategy to first carry out recognition using handcrafted features and then moving on to better features using CNN for better results.

[2] based on monument recognition is using Graph Based Visual Saliency (GBVS) method to find out saliency in monument images so that better image recognition algorithms work. To achieve this goal, the images are already previously processed which is according to the Graph Based Visual Saliency method, which is in order to either keep these SIFT or SURF features , which are corresponding to the present actual monuments while the background "noise" which has been minimized .

[3] comprises of identification and retrieval of archaeological monuments using visual features. They use Content Based Image Retrieval (CBIR) method in which images have been indexed on the criteria of the low-level features, which are as mentioned: color, texture, and shape that can be automatically can be derived from the present visual content of the visual images.

The noise which comes in monument images has been taken care of already, as manually better cropped images have been used. The handcrafted features gave accuracies but when CNN was used, better accuracy was achieved in monument recognition. CNN features take the best out of monument images and rest of the work is done by the neural network which feeds better features into it. Moreover this method can be thought of as a new idea to recognize different cultural monuments to promote work in Indian culture too.

III. PROPOSED METHOD

The various methods used to recognize monuments have been explained in this section. It comprises of the extraction of features using hand-crafted features and then Convolutional Neural Network (CNN).

The key intuition behind the monument classification was the use of basic hand crafted features like HOG, LBP and GIST to find out whether these features could be used to classify monuments with a better accuracy. The classification can be good only if features which are extracted from images are enough to get a better accuracy. The model can be trained on any features provided it gives better results. The images were already cropped to give better features to our model.

A. Feature Extraction using HOG features

The histogram of oriented gradients (HOG) as shown in Fig. 2 is a feature descriptor which is used in computer vision and image processing basically to serve the purpose of object detection. The technique is used to count occurrences of the gradient orientations present in localized portions of the image.

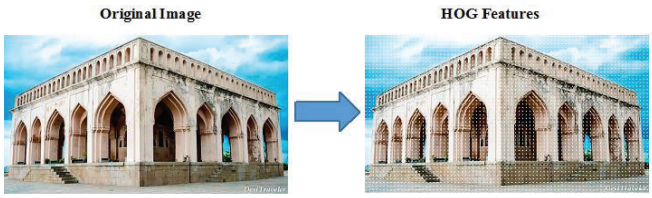

Fig. 2. HOG feature extraction

B. Training using LBP features

Local binary patterns (LBP) as shown in Fig. 3 is a type of visual descriptor which is used for the classification in the field of computer vision. LBP is the particular case of the Texture Spectrum model which has been proposed in 1990 [4, 5]. The first description of LBP was in 1994 [6, 7]. The LBP feature vector, in its simplest form, which is created in the following manner: The examined window are divided into cells (e.g. 16x16 pixels for each of the cells). Then for each of the pixels in a cell, the pixel to each of the 8 neighbors is compared (on its left-top, left-middle, left-bottom, right-top, etc.). The pixels are followed along a circle, i.e. either clockwise or another way is counter-clockwise. Then wherever the center pixel's value is greater in magnitude than the neighbor's value, writing it as "0". Otherwise, writing it as "1". This even provides with an 8-digit binary number (usually converted to a decimal value). Computing of the histogram, over the cell, which is of the frequency of each of the "number" occurring (i.e., each combination of which all pixels are smaller and greater than center) is carried out.

This histogram well then can be seen to constitute of as a 256- dimensional feature vector. Optionally normalize the histogram.

The (normalized) histograms obtained of all cells are concatenated. This then gives us a feature vector for the entire window. The feature vector then can be processed by the classifier methods such as Support vector machine or some other machine-learning algorithm which can be used to classify images. Such classifiers have also been used for in the area of face recognition or the texture analysis.

Fig. 3. LBP feature extraction

C. Training using GIST features

A brief introduction to the GIST descriptors is as shown in Fig. 4. The image is segmented by a 4 by 4 grid for which orientation histograms are extracted to compute the color GIST description. It takes as an input, a square image of fixed size and produces a vector of dimension 960. Most of the works mentioned are using the GIST descriptor [8, 9] resizing of the image in a preliminary stage, for producing of a small square image. According to [10], it's a low dimensional representation of the scene, which does not require any form of segmentation.

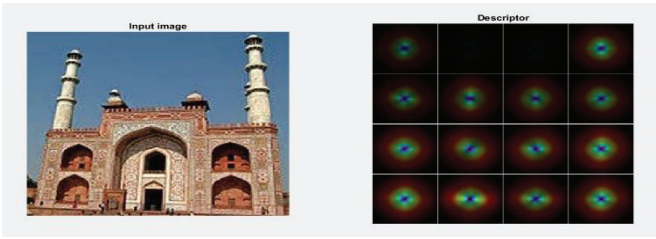

Fig. 4. GIST feature extraction

D. Extraction of Representations using Deep CNN's

CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing [11]. Fig. 5 shows a general DCNN architecture. As shown in Fig.6, Monument

Recognition has been carried out using Deep CNN's fc6 and fc7 layers. Fine tuning has been done for all the different CNN's on the Alex-Net architecture [12]. The input layer is of 227×227×3 , conv of 96×11×11 ,max pooling of 3×3 ,conv of 128×5×5 , max pooling of 3×3 , conv of 256×3×3, conv of 192×3×3 ,conv of 192×3×3 ,3×3 pooling 4096×1 fc-4096×1 fc 1000×1 fc layers. Before being fed into the respective CNN's, monument images were resized to 227 * 227. These CNNs were fine-tuned on the manually acquired dataset so that the representations could be extracted accurately. The initial learning rate was set to 0.001 for the final fc layer and 0.0001 for the remaining layers while fine-tuning. The dropout was 0.5 and the momentum was 0.8 for a total 10,000 iterations. For capturing the different aspects of the monument images, 3 different CNNs were used and 4096 representations were extracted overall from the three CNN's. To form a feature vector comprising of 67840 elements, these representations were concatenated together.

The difference in images is that photos are clicked from different views and at different timings of a day. Hence, the model was finally trained on a total of 100 classes. The division of dataset was 70:30, which means 35 photos per class for training and 15 photos per class for testing. The dataset was collected from different sources across the internet. For experimentation, we used a system with Xeon(R) processor with an NVIDIA K2-quadro gpu.

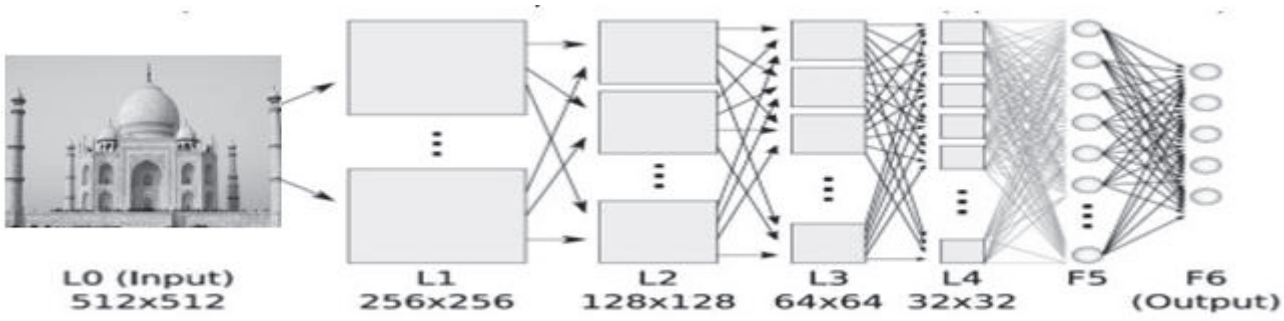

Fig. 5. Deep Learning feature extraction

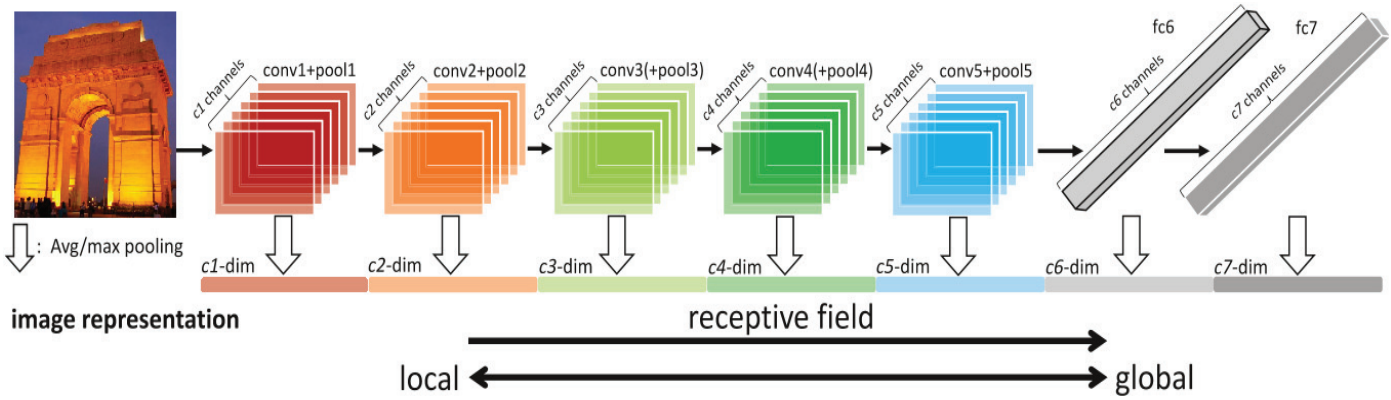

Fig. 6. CNN Architectural Representation

IV. EXPERIMENTAL SETUP

The section includes description of data used for testing and training. Moreover, the different experimental scenarios and their outcomes that were used for testing the performance of the model are also outlined in this section. Fig. 6 clearly explains our experimental setup for monument recognition. MATLAB2016a software was used for performing the experiments.

A. Description of Data-set

The manually acquired dataset comprises of 100 folders with each folder having 50 images per monument. The naming of each folder is done according to the name which corresponds to the monument. Mostly, the famous Indian monuments were considered for dataset. Each folder contains 50 different images of the same monument.

B. Experimental scenario

Different approaches were used to classify monuments. Firstly, HOG features were extracted separately for training and testing. The features were then used for classification purposes in classification algorithms like SVM (Support Vector Machine),KNN(K-Nearest Neighbour) to calculate accuracy for all the 100 monuments and this has obtained an accuracy of 1.47%. Similarly, LBP features were also extracted and then classification was implemented to obtain a 20.09% accuracy. Then GIST features were also used for classification and 1% accuracy was obtained. Finally, Deep Convoluted Neural Networks model was used in order to get a better output. The DCNN gives a very good accuracy of 92.7% using fc6 layer as seen in terms of monument recognition on 100 classes , shown in Table 1.

Table 1. A statistical comparison of various hand-crafted features and deep CNN features when 100 monuments (folders) are taken.

Technique	Accuracy %
HOG+SVM	1.47
HOG+Random Forest	1
HOG+KNN	1
LBP+SVM	7.23
LBP+Random Forest	14.27
LBP+KNN	20.09
GIST+SVM	1
GIST+Random Forest	1
GIST+KNN	1
CNN fc6	**92.7**
CNN fc7	**90.60**
CNN fc6+fc7	**91.82**

A graphical representation as shown in Fig 7. has been used to reiterate the results of Table 1

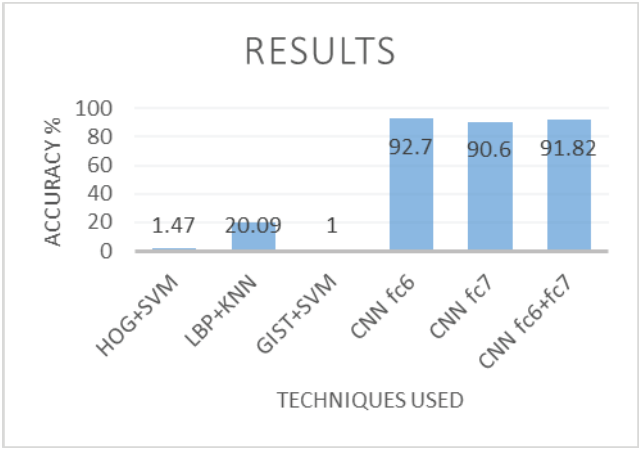

Fig 7. Bar graph showing maximum accuracy reached with respect to particular feature.

V. CONCLUSION

The paper presents a framework for recognizing monument images. The proposed framework relies mostly on DCNN for extracting representations. It was able to achieve a much better accuracy of 92.7% in comparison to the hand crafted features. The experiments performed proved the importance of using representations of monuments images to build an effective monument recognition system. It was seen that performance of model increased with CNN.

The accuracy of the model can be increased by using Graph Based Visual Saliency (GBVS) [13, 14]. This method is used to find the saliency of an image, basically as a preprocessing step. Moreover some other related work in monument recognition are as proposed to be an image retrieval method where shape of an image is extracted by applying mathematical morphology and texture is retrieved by applying GLCM (Grey level co-occurrence method). Grey-level matrix is a matrix whose elements measure the relative frequencies of occurrence of grey level combinations among pairs of pixels with a specified spatial relationship [15]. Other descriptors like MSER (Maximally Stable External Region) can also be used [15]. Therefore, the monument recognition system should also consider the prominent part that monument's clear image should also be extracted from the input image. The future work comprises of increasing the size of the dataset and including monuments with more variations in terms of structures and illuminations.

REFERENCES

[1] Cole, J.Y. and Reed, H.H. eds., 1997. The Library of Congress: the art and architecture of the Thomas Jefferson Building. WW Norton & Company.

[2] Kalliatakis, G. and Triantafyllidis, G., 2013. Image based Monument Recognition using Graph based Visual Saliency. *ELCVIA*, *12*(2), pp.88-97.

[3] Yaligar, S., Sannakki, S. and Yaligar, N., 2013. Identification and Retrieval of Archaeological Monuments Using Visual Features.

[4] He, D.C. and Wang, L., 1990. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, *28*(4), pp.509-512.

[5] Wang, L. and He, D.C., 1990. Texture classification using texture spectrum. *Pattern Recognition*, *23*(8), pp.905-910.

[6] Ojala, T., Pietikainen, M. and Harwood, D., 1994, October. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on* (Vol. 1, pp. 582-585). IEEE.

[7] Ojala, T., Pietikäinen, M. and Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, *29*(1), pp.51-59.

[8] Hays, J. and Efros, A.A., 2007, August. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*(Vol. 26, No. 3, p. 4). ACM.

[9] Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L. and Schmid, C., 2009, July. Evaluation of gist descriptors for web-scale image search. In Proceedings of the ACM International Conference on Image and Video Retrieval (p. 19). ACM.

[10] Oliva, A. and Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), pp.145-175.

[11] Shamov, I.A. and Shelest, P.S., 2017, May. Application of the convolutional neural network to design an algorithm for recognition of tower lighthouses. In *Integrated Navigation Systems (ICINS), 2017 24th Saint Petersburg International Conference on* (pp. 1-2). IEEE.

[12] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[13] Harel, J., Koch, C. and Perona, P., 2007. Graph-based visual saliency. In *Advances in neural information processing systems*(pp. 545-552).

[14] Kalliatakis, G. and Triantafyllidis, G., 2013. Image based Monument Recognition using Graph based Visual Saliency. *ELCVIA*, *12*(2), pp.88-97.

[15] Yaligar, S., Sannakki, S. and Yaligar, N., 2013. Identification and Retrieval of Archaeological Monuments Using Visual Features.