A computer vision framework for automatic description of Indian monuments.

Pushkar Shukla
Department of Computer Science
University of California, Santa Barbara
pushkarshukla@umail.ucsb.edu

Beena Rautela
Department of Electronics and Communication,
Women Institute of technology,
Dehradun, India
beenarautela5@gmail.com

Ankush Mittal
Graphic Era University
Dehradun, India
dr.ankush.mittal@gmail.com

***Abstract*—Monument recognition and description has emerged as a promising area of research. For any given image of a monument a question arises that *up to what extend can a computer model describe the monument from that image?* The main objective of the paper is to propose a framework which is capable of identifying multiple attributes from a single image of a monument. Four different attributes i.e. the class of the monument, the style of the architecture, the time period in which the monument was constructed and the type of the monument are taken into consideration. The paper proposes a framework that relies on Deep Convolutional Neural Networks (DCNN) for describing the monument in terms of the aforementioned attributes. The experiments have been performed on a dataset comprising of 6102 images of 117 Indian monuments. The model was able to achieve an accuracy greater than 80% for all the different set of experimentation's. The results clearly indicate the usefulness of the framework.**

Keywords:- deep learning, Convolutional neural network, Monument recognition, Landmark recognition.

I. INTRODUCTION

Automatically describing different attributes of a monuments having religious, cultural and historical significance can have numerous applications. Models capable of describing different attributes such as the age of the monument or the architectural style of the monument from a given image can be used for e-preservation of the historical cites. These systems can also be useful in the tourism industry and can aid a tourist to learn more about a given monument from images. A sample of such a system has been shown in Figure 1.

India is country with a rich cultural heritage of over 2000 years. Due to the differences in regions,culture and civilizations during these 2000 years Indian monuments differ greatly depending upon the style of architecture, era of construction and other factors. The prime focus of paper is to propose a frame-work that is capable of describing different attributes of Indian monuments.

Through this paper we have tried investigating the answers to the following questions.

- What information can be inferred from images of Indian monuments?
- Can deep representations be used for building a unified framework capable of predicting multiple attributes for a given image of an Indian monument ?

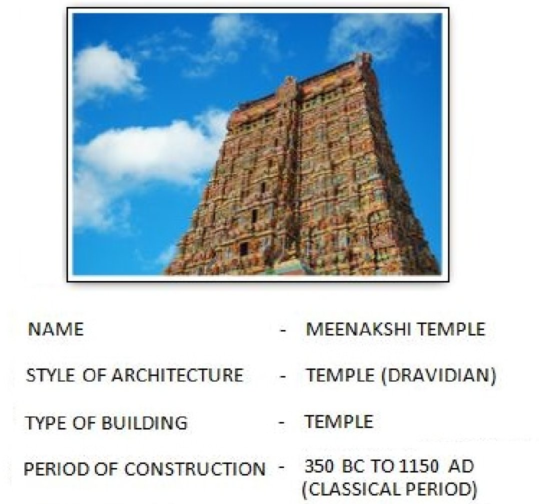

Fig. 1. Attributes of a monument predicted by the system for a sample test image.

- If yes, then with what success can the framework predict these attributes?

It was observed that the previous works on monument description have focused only upon a single aspect of monument description. There have been models that have focused upon monument identification[15] while some approaches have also focused upon recognizing architectural styles [8]. As per our knowledge, no common framework has been proposed that is capable of describing different aspects of a monument. In case of Indian monuments, little work has been done on identifying monuments or recognizing the styles of architecture for different monuments. No prior work has focused upon on identifying the age of different Indian Monuments. It must be noted that most of these experiments have been carried out on a relatively small data-set which is insufficient to draw a strong conclusion.

The major contribution of the paper lies in proposing a unified frame-work that is able to predict the following aspects of historical monuments in India from their images.

- The class of the monument.
- The architectural style of the monument.

- The era in which the monument was constructed,
- The purpose of the monument.

These experiments have been conducted on a sufficiently large, newly introduced data-set comprising of images of monuments of historical , cultural and religious importance from all over India. The recent success of Deep Convolutional Neural Networks for a variety of image recognition tasks [16], [13], [11], [18] extracting representations. Further, Graph based Visual Saliency is employed on the images in order to extract distinct parts of the image and remove the unnecessary background which may be common to other images as well.

The rest of the paper has been organized as follows. A detailed description of the previous models on monument recognition has been presented in Section 2. Section 3 describes the proposed methodology. The data-set that has been used for experimentation has been described in Section 4. The experimental results has been presented in Section 5. The analysis of these results has been done in Section 6. Section 7 concludes the paper and discusses the future scope of this work.

II. Literature Review

The initial work on monument recognition and retrieval relied mostly on the use of basic descriptors for texture like Grey Level Co-occurrence matrix(GLCM) . Gradually, more sophisticated descriptors like Scale Invariant Feature Transform (SIFT), Speed-ed up robust Features (SURF) and Oriented FAST and Rotated BRIEF (ORB) [1] were used for monument recognition. Saliency based models were also used for monument recognition in order to remove the unnecessary background from the images[12]. Many monument recognition models have also made use of the Bag of Visual words [10]. Recently, [3] used Deep Convolutional Neural networks for monument recognition.

Recognizing or retrieving monuments with similar architectural styles is a relatively new area of research. [2] proposed a novel descriptor for matching the images of building with the similar style of architecture. [6] made use of patch based methods to extract repetitive patterns across similar styles of architecture. [8] also proposed a similar framework for recognizing architectural styles of European Monuments which relied on extracting repetitive patterns from each of these monuments. The motive of the paper was to show that buildings constructed during the same era and within a region are similar to each other and can be thought of as instances of one another.

Previous works on Indian Monument recognition have relied mostly on hand crafted descriptors for extracting representations. ORB(Oriented Brief and Fast) descriptors were used by Sahay et.al. [17]. The data-set was divided into 5 different classes i.e. 'Ancient', 'British', 'Indo-Islamic', 'Maratha' and 'Sikh' and the experiments were carried out on a data-set of 500 images. Bhatt and Patila [4] proposed a genetic programming evolved spatial descriptor using GLCM for monument classification.[5] proposed a model that relied on GLCM and Canny and Sobel methods for retrieving different types of monuments from a data-set of 500 images belonging to 5 different classes. Ghosh et.al[7] proposed an approach for generating narratives of heritage tours from egocentric videos. Further, [7] proposed a frame work that relied on Bag of Visual Words for describing different parts of archaeological site.The framework was tested for two different monuments in India.

III. Proposed Methodology

Images of different monuments may have similar backgrounds. Therefore, models may get confused by the irrelevant parts of monument's background like the sky, trees etc. Visual saliency methods are used in order to remove the irrelevant aspects from the images. The proposed framework relies on two CNN architectures. The complete image is given as input to the first CNN whereas the salient parts of the monument are extracted from the same image and are then fed into another CNN. Representations *F1* and *F2* are extracted from both these networks and are then concatenated into a single vector *F* such that

$$F = \{F1, F2\} \tag{1}$$

where *F* is the final representation vector. The final representation vector is then fed into the classification module that uses either SVM or KNN for classification, depending upon the experimental scenario. A block diagram representation of the proposed approach has been drawn in Figure 2.

A. Saliency based region extraction

Saliency refers to parts of the image that are from the other parts of the image and are more appealing to the human eye.Highly salient parts of the image correspond to parts that are unique whereas lower saliency regions generally represent the background of an image. Graph based visual saliency is a bottom -up saliency model proposed by [9]. The approach can be divided into two parts.The first part focuses on computing activation maps on the basis of certain feature channels whereas the second part focuses on combining these activation maps with other maps. In the proposed frame-work all regions having a threshold greater than 0.25 were extracted from the feature map. Figure 3 shows an example of a sampled region extracted from the feature map.

B. Using Deep CNN's for extracting representations

Deep representations were chosen over other hand crafted representations primarily due to the superiority of DCNN's over traditional descriptors for complex image recognition tasks. The limited size of the data-set was the primary reason for choosing a pre-trained CNN model over hand crafted CNN's as these CNN's require a lot of data to train.The Alex-Net model [14] was used for extracting representations. The representations were extracted from the *fc6* and *fc7* layers of each of the two CNN's. The dimensions for each layer is 4096*1. Therefore, the dimensions of the final feature vector *F* for each sample is 16376*1.

C. Classification

The extracted representations are then trained and classified with the aid of SVM and KNN classifiers. A multi-class SVM is trained for the model. A one-vs-all strategy is used for training the SVM. The SVM is trained on a linear kernel. The SVM is trained with a sequential minimum strategy and the value of the kernel offset was set to 0.

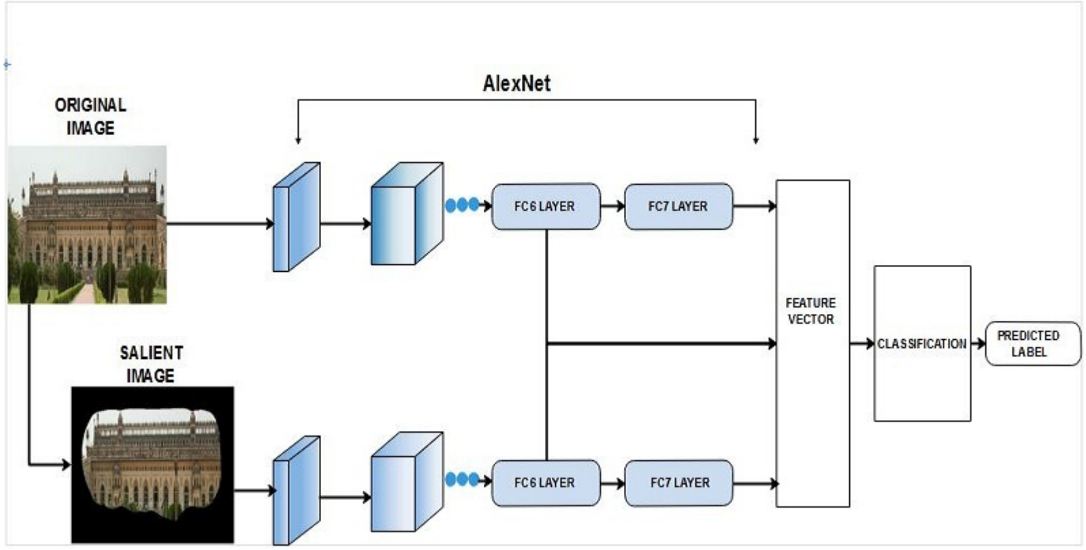

Fig. 2. A block diagram representation of the proposed approach.

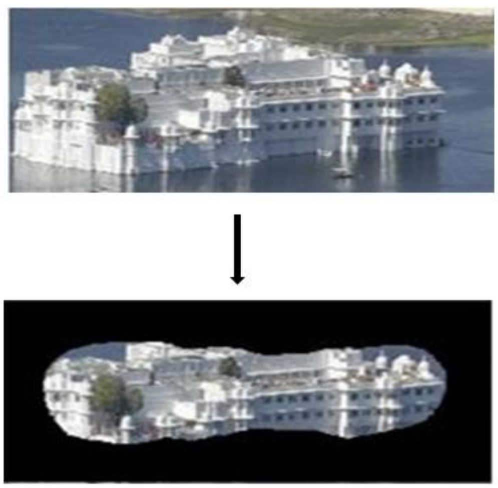

Fig. 3. Salient Regions of the monument extracted after applying Graph Based Visual saliency.

For KNN, the euclidean distance metric is used for calculating the distance between the training and testing samples. The value of K was set to 3 for all test samples

IV. DATA-SET DESCRIPTION

A detailed description of the data-set used for experimentation has been presented in this section. The data set comprises of a total of 6102 images of 117 historical, religious and culturally important sites in India. There are approximately 50 images per class. The data-set was collected from various sources across the Internet.

A. Annotation in terms of Architectural Styles

Monuments can be classified according to the different styles of architectures. Different monuments built in the same era in the same region generally tend to follow a similar architectural style. The different monuments in the data-set have been annotated in terms of various styles of architecture. A total of six different architectural styles were considered for labeling these monuments. These are **Cave, Colonial, Indo-Islamic, Rock-Cut and Temple** style of architecture. The details of the annotations have been described in Table 1 while a single image for each style of architecture has been shown in Figure 4 .

Fig. 4. Sample images for different styles of architectures.The different styles architectures from left to right and top to bottom are in the order Temple, Rock-cut, Colonial, Hindu, Indo- Islamic and Cave architecture

The temple architecture has been further sub divided into three more classes namely, the **Mix, Dravidian** and **Nagara** styles of Architecture. The details of the annotations are given in Table 2. It must be noted that there were certain monuments

Type of Architecture	Number of Images
Colonial	715
Cave	385
Rock- Cut	354
Temple	1589
Indo-Islamic	1828
Hindu	673

TABLE I. DATA-SET ANNOTATION ACCORDING TO THE DIFFERENT STYLES OF ARCHITECTURE.

Type of Architecture	Number of Images
Dravidian	511
Mix	255
Nagara	823

TABLE II. ANNOTATIONS FOR DIFFERENT STYLES OF TEMPLE ARCHITECTURE.

that did not have a specific style of architecture. Hence, those monuments are not annotated in terms of architectural style.

B. Annotation in terms of Age

Recognizing the age of a given monument from their images is a problem that has not received much attention till date. The complete data-set has been annotated in terms of the era in which different monuments were constructed The details of the annotations are described in Table 3. The annotations are loosely based on the different era's of the Indian history. The time period has been classified as the classical period from -350 BC to 1150 AD , the early medieval period from 1150 AD up to 1550 AD before Akbar's coronation, the Golden Mughal period starting from 1550AD to 1787 AD, the early colonization era from 1787 to 1887 AD , the late colonization era from 1887 AD up to the independence in 1947 and the post independence era from 1947 to the present. Sample images of monuments from each era are shown in Figure 5 .

C. Annotation in terms of the type of the monument.

The complete data-set has been annotated in terms of the type of the monument.The purpose of this type of annotation is to teach the model the purpose for which a given monument has been constructed. A total of 7 different classes were considered. The monuments were annotated as **churches, tombs, mosque, temples, palaces, natural landscapes and forts.** The details of the annotations are described in Table 4. Sample images of monuments of each type are shown in Figure 6 .It must be noted that there were certain monuments that did not fit into any of these categories. These monuments were not annotated.It should be noted that the places included in natural landscapes were not included because of the scenic beauty of these landscapes but due to their cultural , religious or historical importance.

Time of Construction	Number of Images
350 BC to 1150 AD	1401
1150 AD to 1550 AD	776
1550 AD to 1787 AD	1621
1787 AD to 1887 AD	771
1887 AD to 1947 AD	754
1947 AD to present	361

TABLE III. DATA-SET ANNOTATION ACCORDING TO THE DIFFERENT ERA'S IN WHICH DIFFERENT MONUMENTS WERE CONSTRUCTED.

Fig. 5. Sample images of monuments constructed in different era's.

Fig. 6. Sample images of monuments constructed in different era's.

V. EXPERIMENTS

A. Experimental Scenarios

A number of experimental scenarios were considered depending upon the type of annotations. The experiments were performed using MATLAB 2016a on a q2 quadro GPU.The

Type of Building	Number of Images
Church	256
Fort	395
Mosque	446
Palace	888
Temple	1390
Tomb	522
Natural Landscape	613

TABLE IV. DIVISION OF THE DATA-SET ACCORDING TO THE DIFFERENT TYPES OF BUILDING CONSTRUCTED FOR DIFFERENT PURPOSES.

Method	Accuracy (%)
HOG+SVM	46.19
HOG+KNN	33.12
HOG+Random Forest	41.74
GIST+SVM	54.59
GIST+KNN	51.49
GIST +Random forests	52.56
Alex-Net +SVM	85.24
Proposed approach +SVM	**86.67**
Proposed Appraoch +KNN	**78.05**

TABLE V. A STATISTICAL COMPARISON OF THE PERFORMANCE OF THE PROPOSED FRAME-WORK WITH OTHER TECHNIQUES.

details of the experimental scenarios are given below.

1) **Scenario 1:** In the first experimental scenario we wanted to evaluate the performance of the model for identifying monuments from a total of 117 classes.The performance of the frame-work was also compared with other existing models for Indian monument recognition.

2) **Scenario 2:** The frame-work was then trained for recognizing different the architectural styles of different monuments.The performance of the model was evaluated for these styles of architectures.

3) **Scenario 3:** Once the model has been evaluated for recognizing monuments and architectural styles, the authors were then interested to see if the model could be used for predicting the era in which the monument was constructed. In this scenario the aim was to see that to how well would the model perform when it comes to recognizing the age of any monument.

4) **Scenario 4:** The framework is now tested to predict the type of the monument.

B. Experimental Results

The experimental results for all the four different experimental scenarios have been presented in this section. It should be noted that that in each set of experiments the ratio of the training data to the testing data was 7:3.

1) Experimental Scenario 1: The frame work was evaluated on a total of 117 monuments. SVM when used for classification resulted in an accuracy of **86.67 %** whereas KNN resulted in an accuracy of **78.05%** . The model has been compared with other approaches that were previously used for monument recognition. The statistical comparison has been presented in Table 6. Top 8 Samples of instances of monuments with the highest and the lowest recognition ac-curacies have been shown in Figure 7 and Figure 8 respectively.

2) Experimental Scenario 2: The framework achieved an accuracy of **86.45%** with SVM and **89.44 %** with KNN . Table 6 shows the confusion matrix for different styles of architecture. A statistical comparison of the Mean Average Precision (MAP) and Mean Average Recall (MAR) has been presented in Figure 9.The highest mean average precision of **92.33%** was achieved by monuments belonging to the **Hindu** type of architecture whereas monuments of the **Cave architecture** achieved the highest mean average recall of **93.61%**

Further, the temple architecture was subdivided into three classes.An accuracy of **89.30%** was achieved by SVM whereas

	T	RC	II	H	CO	CA
T	**0.913**	0	0.022	0.043	0	0.022
RC	0	**0.908**	0.019	0.034	0.008	0.031
II	0.02	0.05	**0.817**	0.07	0.005	0.04
H	0.002	0.026	0.026	**0.923**	0.002	0.022
CO	0.107	0	0	0.036	**0.76**	0.102
CA	0.006	0.025	0.034	0.04	0.002	**0.893**

TABLE VI. CONFUSION MATRIX FOR ARCHITECTURAL STYLE RECOGNITION. .THE DIFFERENT STYLES OF ARCHITECTURES ARE TEMPLE (T), ROCK-CUT (RC), INDO-ISLAMIC(II), HINDU (H), COLONIAL (CO), AND CAVE (CA) TYPE OF ARCHITECTURE.

	Mix	Nagara	Dravidian
Mix	**0.9281**	0.0719	0.0
Nagara	0.0810	**0.9150**	0.0040
Dravidian	0.0390	0.0909	**0.8701**

TABLE VII. CONFUSION MATRIX FOR DIFFERENT STYLES OF TEMPLE ARCHITECTURES.

KNN achieved an accuracy of**91.19%** .The confusion matrix for these three architectural types has been shown in Table. A comparison of the mean average precision and the mean average recall has been shown in Figure 10. The highest precision of **92.81%** was achieved by the **Mix** style of temple architecture whereas the highest recall of **98.52%** was achieved by the **Dravidian** style of architecture.

3) Experimental Scenario 3: The model was trained to classify images on the basis of their age or the time period in which a given monument was constructed. Both SVM and KNN were used for classification . SVM was able to achieve an accuracy of **70.12%** whereas KNN recorded an accuracy of **84.64%**. The confusion matrix for the different classes has been shown in Table 8. The statistical comparison of the mean average precision and the mean average recall for each of these classes has been shown in Figure 11. The highest mean average precision of **88.80 %** was achieved by the monuments that were constructed between **350 BC to 1150 AD** whereas monuments constructed between **1887AD to 1947AD** achieved the highest mean average recall of **93.51%**.

4) Experimental Scenario 4: The model was trained to classify monuments depending upon the type of building. SVM's were able to achieve an accuracy of the **90.19%** while KNN achieved an accuracy of **88.41%**. A statistical comparison of the mean average precision and mean average recall for each class has been shown in Figure 12. The highest mean average precision of **96.73%** was achieved by images of **natural landscapes** whereas **Palaces** achieved the highest mean average recall of **96.4%**.The confusion matrix for various classes as been presented in Table Y.

	1	2	3	4	5	6
1	**0.888**	0.049	0.039	0.013	0.008	0.003
2	0.159	**0.764**	0.047	0.021	0.009	0
3	0.051	0.023	**0.885**	0.025	0.012	0.0041
4	0.056	0.052	0.048	**0.827**	0.004	0.013
5	0.056	0.052	0.037	0.028	**0.816**	0.01
6	0.092	0.009	0.057	0.037	0	**0.806**

TABLE VIII. CONFUSION MATRIX FOR RECOGNIZING THE TIME PERIOD IN WHICH A GIVEN MONUMENT WAS CONSTRUCTED.THE NOTATIONS ARE 350 BC TO 1150 AD (1) 1150 AD TO 1550 AD(2) 1550 AD TO 1787 AD (3) 1787 AD TO 1887 AD (4) 1887 AD TO 1947 AD (5) 1947 AD TO PRESENT (6).

Fig. 7. Samples of images of the top 8 classes having the highest recognition rate.

Fig. 8. Samples of images of the top 8 classes having the lowest recognition rate.

	Church	Fort	Mosque	Landscapes	Palace	Temple	Tomb
Church	**0.844**	0.013	0.013	0	0.065	0.052	0.013
Fort	0	**0.817**	0	0.1	0.058	0.017	0.008
Mosque	0.007	0.007	**0.888**	0.008	0.03	0.022	0.037
Landscapes	0	0.005	0.011	**0.978**	0	0.005	0
Palace	0.007	0.003	0.011	0.011	**0.94**	0.007	0.002
Temple	0.01	0.002	0.017	0.029	0.005	**0.919**	0.019
Tomb	0.006	0.006	0.051	0.019	0.045	0.025	**0.8471**

TABLE IX. CONFUSION MATRIX FOR THE DIFFERENT TYPES OF BUILDINGS..

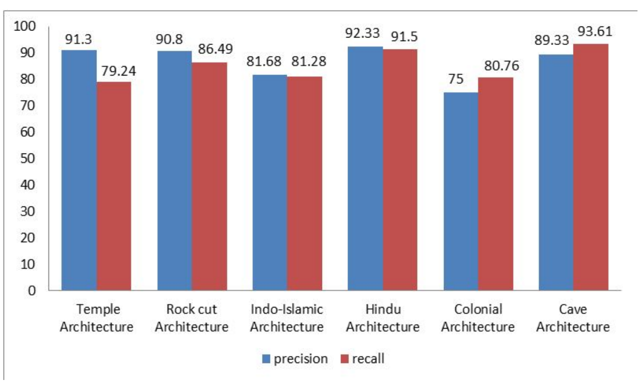

Fig. 9. Mean Average Precision and Mean Average Recall for different styles of architectures .

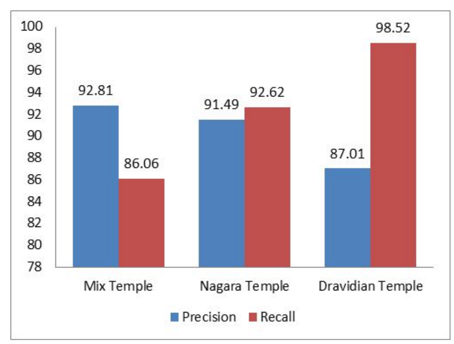

Fig. 10. Figure 8. Mean Average Precision and mean average recall for different styles of temple architecture.

VI. EXPERIMENTAL ANALYSIS

An experimental analysis of the outcomes of the different experimental scenarios has been presented in this section. It should be noted that both SVM and KNN were used for classification in 4 different scenarios . SVM outperformed KNN for two scenarios whereas KNN outperformed SVM for the remaining two scenarios.

A. Scenario 1

The proposed model was able to achieve a reasonable accuracy of 86.67 % for a total of 117 classes. The approach outperformed other existing approaches for monument recognition by a considerable margin.

B. Scenario 2

A considerable recognition accuracy was achieved using both the classifiers. KNN was slightly better than SVM for classifying monuments according to the different styles of architecture. The model also performed well when used for classifying different styles of temple architectures.

C. Scenario 3

KNN was able to perform significantly better than SVM and was able to achieve an accuracy of 84.64%. The results proved that deep learning frameworks could be used to predict the era in which a given monument was constructed.

D. Scenario 4

SVM performed marginally better than KNN and achieved an accuracy of 90.19%. Natural Landscapes were able to achieve the highest precision. one the possible reasons might be that the features for natural landscapes greatly from other monuments.

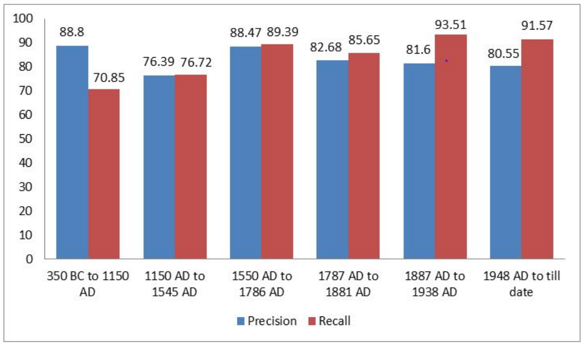

Fig. 11. Mean average precision and recall for different era's in which Indian monuments were constructed.

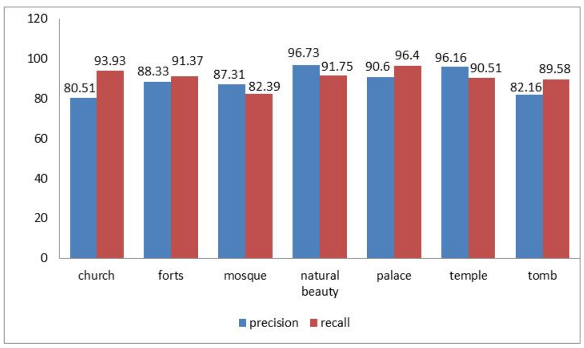

Fig. 12. Mean Average Precision and Mean Average Recall for different classes of buildings .

VII. Conclusion

The paper proposes a unified framework for describing Indian monuments. Four different attributes were taken into consideration in order to describe a given monument from it's image.The frame-work makes use of Deep CNN's for extracting representations. We also introduce a new data-set of 117 different monuments in India annotated in terms of age, architectural style and type of the monument.

The results for each of the four experimental scenarios clearly indicate the usefulness of the framework.One interesting applications of the models is that it can be trained to predict the era in which a given building was constructed thus aiding in the preservation and restoration of monuments. The frameworks can also have huge application in the tourism industry.

References

[1] G. Amato, F. Falchi, and P. Bolettieri. Recognizing landmarks using automated classification techniques: Evaluation of various visual features. In *Advances in Multimedia (MMEDIA), 2010 Second International Conferences on*, pages 78–83. IEEE, 2010.

[2] G. Amato, F. Falchi, and C. Gennaro. Fast image classification for monument recognition. *Journal on Computing and Cultural Heritage (JOCCH)*, 8(4):18, 2015.

[3] P. Bezak. Building recognition system based on deep learning. In *Artificial Intelligence and Pattern Recognition (AIPR), International Conference on*, pages 1–5. IEEE, 2016.

[4] M. S. Bhatt and T. P. Patalia. Genetic programming evolved spatial descriptor for indian monuments classification. In *Computer Graphics, Vision and Information Security (CGVIS), 2015 IEEE International Conference on*, pages 131–136. IEEE, 2015.

[5] P. Desai, J. Pujari, N. Ayachit, and V. K. Prasad. Classification of archaeological monuments for different art forms with an application to cbir. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1108–1112. IEEE, 2013.

[6] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[7] A. Ghosh, Y. Patel, M. Sukhwani, and C. Jawahar. Dynamic narratives for heritage tour. In *European Conference on Computer Vision*, pages 856–870. Springer, 2016.

[8] A. Goel, M. Juneja, and C. Jawahar. Are buildings only instances?: exploration in architectural style categories. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, page 1. ACM, 2012.

[9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[10] N. Hascoët and T. Zaharia. Building recognition with adaptive interest point selection. In *Consumer Electronics (ICCE), 2017 IEEE International Conference on*, pages 29–32. IEEE, 2017.

[11] D. C. Hauagge and N. Snavely. Image matching using local symmetry features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 206–213. IEEE, 2012.

[12] G. Kalliatakis and G. Triantafyllidis. Image based monument recognition using graph based visual saliency. *ELCVIA*, 12(2):88–97, 2013.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1957–1964. IEEE, 2009.

[16] P. Shukla, T. Gupta, A. Saini, P. Singh, and R. Balasubramanian. A deep learning frame-work for recognizing developmental disorders. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 705–714. IEEE, 2017.

[17] S. J. Tanvi Sahay, Ankita Mehta. Architecture Classification for Indian Monuments. Technical report, University of Massachusetts Amherst.

[18] C. Wang. Video classification with convolutional neural network. *Proceedings of the 9th Ph. D. retreat of the HPI Research School on service-oriented systems engineering*, 100:245, 2017.