Study Impact of Architectural Style and Partial View on Landmark Recognition

Ying Chen
smileyc@stanford.edu

1. Introduction

Landmark recognition in image processing is one of the important object recognition problem. Some of the applications in smart phones with (Global Positioning System) GPS utilize the geo- tag information to recognize landmarks. However, tons of photos on the internet don?t have geo- tag. In addition, the GPS is not always very accurate to provide correct recognition. Thus, recognizing landmarks using machine learning techniques is the important alternative. In addition, these studies are also critical for applications like building virtual tourism, providing travel tour guides and classifying/tagging photos.

Although some recent studies[1][2][3] showed good accuracy of landmark detection and recognition, most of the landmarks in the images were located in Europe and North America. Whether the recognition accuracy will be largely affected or not by Asian style landmark images was not studied. Thus, the mix of Asian style and Europe/North America landmarks were used to understand how this affect recognition accuracy. Furthermore, the impact of partial view of large landmarks on recognition accuracy were be analyzed.

In this study, the input data are images with different landmarks. At the feature detection stage, both HOG (Histogram of Oriented Gradient)[4] and SURF (Speeded Up Robust Features)[5] were used to extract features from input images. The multiclass linear SVM (support vector machines) model was build, and was used to categorize each query image into trained landmark images.

2. Related Work

Carndall et al.[1] performed the research based on large scale image dataset (6.5 million images and 500 landmarks) on Flickr. The geo-tag information was used to derive category labels. SIFT based bag of word features were extracted. Multiclass support vector machine (SVM) was used for classification. In addition, convolutional neural networks (CNN) classifier was also applied, which showed striking improvement over the traditional approaches, especially when the training dataset is large. Carndall et al.[6] also performed another experiment to study the geographic embedding photos in order to recognize landmarks, but the scale size is small.

One of the projects in course CS229[2] of Stanford University compared four classifiers on a small image dataset (193 images) for landmark recognition. The results showed a high accuracy with acceptable processing time. In this study, the dataset is too small to train a good learning model.

A landmark recognition engine was developed by Zheng et al.[3] . Laplacian of Gaussian (LoG) filter and SIFT were used to detect interest point and local descriptor on a large scale image dataset (21.4 million images and 5312 landmarks). Then hierarchical agglomerative clustering was exploited to find similar regions of test images.The detection accuracy of landmarks from other scenes were satisfactory on a small group of query images. However, the recognition rate of individual landmark are low due to local appearance similarity of landmarks. In addition, the number of landmarks in the query image dataset and variation of them were not mentioned. Raguram et al.[7] used similar method to study landmark detection on a large scale images, but there was no recognition study.

Although three of the above study used large scale image dataset, most of the landmarks in the training images were located in Europe and North America according to the statistics and examples provided in the papers. Thus, a interesting question is: that if the recognition accuracy will be largely affected by using image dataset with a good number of asian style landmarks. This forms the main goal of this project. In addition, although some good amount non-landmark building images were used, the impact of partial view of large landmarks on recognition accuracy were not analyzed. It is very hard to be analyzed with huge amount of data.

3. Dataset and Features

Google image search was used to collect the images with keywords of name of landmarks and partial views. Then all the images were visually checked to label the correct category. There are 18 landmarks used in this report. Nine of them are located at Asian countries, and nice of them are located at North America or Europe. So far all the image dataset are global view of landmarks. The name of 18 land-

01	Badshahi Mosque	**02**	Dome of the Rock
03	Forbidden City	**04**	Great Wall
05	Azadi Tower	**06**	Potala Palace
07	Qutub Minar	**08**	Taj Mahal'
09	Temple of Heaven	**10**	Big Ben
11	Colosseum	**12**	Eiffel Tower
13	Golden Gate	**14**	Leaning Tower Pisa
15	Lincoln Memorial	**16**	Sagrada Familia
17	St. Basil Cathedral	**18**	Statue of Liberty

Table 1. Name and index of studies landmarks

marks are shown in Table 1. Landmarks 1 9 are in Asian countries, while landmarks 10 18 are located at Europe or North America. The total number of images are 2104, including both global view and partial view.

Two different feature extraction methods were used - HOG and SURF. Both of them are well known and useful feature descriptor for object detection and recognition in image processing. HOG[4] has been used a lot due to its invariant property to geometric and photometric transformation. The local gradient is computed with predefined cell size. In this study, the images were preprocessed by a resize operation. The resolution of images were resized to either 240x144 or 144x240, depending on the scene type. Then the RGB images were converted to one channel images by using two different ways - binary images and grayscale images.

Comparing to HOG, SURF[5] detects local interest points with aligned orientation at different scales while HOG only works on single scale. Two different methods were analyzed to extract feature points in this study.

In method A, the Hessian matrix is used to detect interest feature points of input image, as shown in the following equation. The determinant of Hessian matrix is used as indicator of local change of certain interest point. The points with maximal determinant are chosen.

$$\mathcal{H}(\mathsf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathsf{x}, \sigma) & L_{xy}(\mathsf{x}, \sigma) \\ L_{xy}(\mathsf{x}, \sigma) & L_{yy}(\mathsf{x}, \sigma) \end{bmatrix} \quad (1)$$

where $L_{xx}(\mathsf{x}, \sigma)$ is the convolution of Gaussian second order derivative with input image in point x. The input image here is grayscale image, which is converted from the original RGB image.

Then the interest points are detected at different scale space, where are a group of image pyramids. In this study, four different scales were used. To find interest points across different scales, the maxima of the determinant of Hessian matrix are interpolated in scale and image space using 3x3x3 neighborhood. One example of interested points is shown in Figure 1. In order to provide a descriptor as image features, one circular region around each interest point

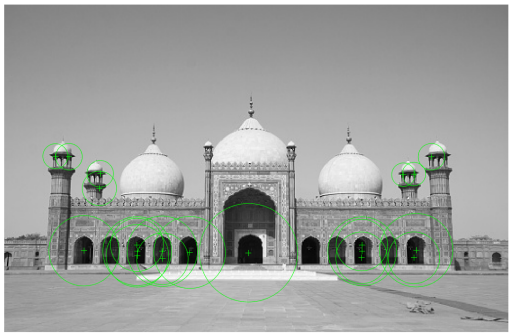

Figure 1. Strongest 20 interest points of an example image.

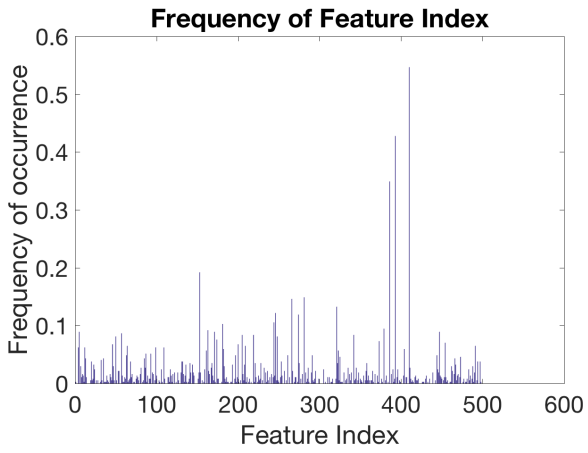

Figure 2. Strongest 500 features selected by K-means method.

is used to fix a reproducible orientation. Then a square region, that aligned to the selected orientation was used to extract SURF descriptor. To do this, each square region is split into 4x4 smaller sub-regions. In each sub-region, the Haar wavelet response are computed on both horizontal and vertical directions. Finally, combining with the absolute values of these responses forms a four dimensional descriptor vector. In this study, if the feature is stronger than a pre-defined threshold, it is selected.

In method B, a fixed grid step with fixed block size and four different scales are used to select interest points. Then the same process is used to get the all the features. Then K-means method was used to get strongest 500 features, which is the descriptor vector. One example is shown in Figure 2.

4. Methods

The SVM (support vector machines) model was used to categorize each test image into trained landmarks. The SVM model was described as the following equation.

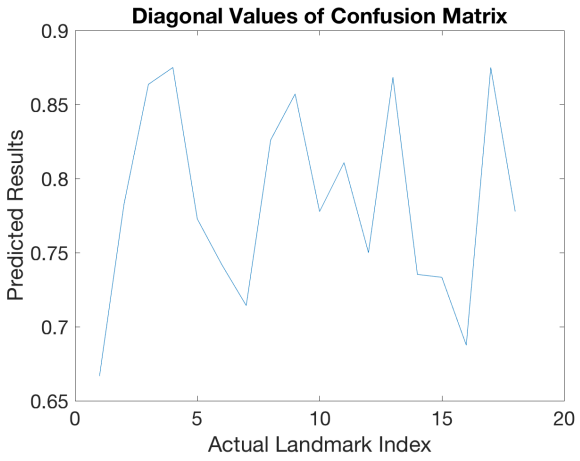

Figure 3. Predicted accuracy of 18 landmarks using HOG features with gray image input.

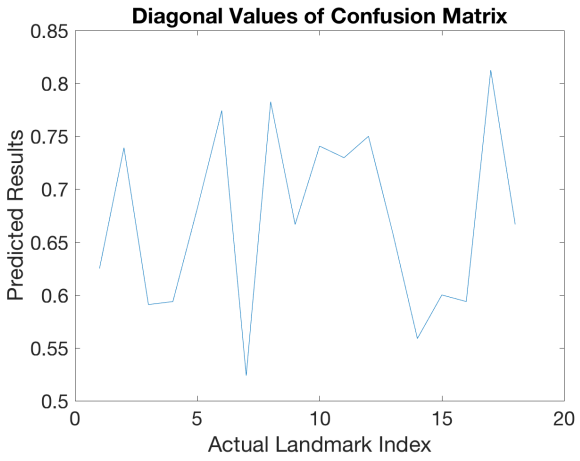

Figure 4. Predicted accuracy of 18 landmarks using HOG features with binary image input.

$$f(x) = sign(\Sigma_i y_i \alpha_i K(x, x_i) + b) \quad (2)$$

where x_i are the training features, y_i is the label of x_i and K is the kernel.

In this study, since multi-class problems need to be modeled, 18 SVMs were trained with linear kernel using the descriptor vectors extracted. Each SVM corresponds to each landmark category. For any test image, it was assigned to the class with largest SVM.

5. Experiment Results and Discussions

5.1. Results - HOG Features

The comparison of recognition accuracy of using binary and grayscale images are shown in Figure 3 and 4. It shows that using grayscale images are obvious better than using binary images for most of the landmarks. This may due to too much information lost when using binary images.

5.2. Results - SURF with Method A

The confusion matrix and prediction accuracy of Method A are shown in Figure 5 and 6. The averaged accuracy of training dataset is 98.34%. The averaged accuracy of test dataset is 90.31%.

Some example of correct and incorrect recognition image results with strongest 20 interest points are shown in Figures 7-13. In Figure 7, it indicates that although with dark lighting, the global view of landmark can still be recognized. In Figure 8, one example with partial view of landmark can be recognized.

However, Figure 9 and 10 show the impact of other objects on landmark recognition, including people. Figure 12 and 13 show the two typical examples with incorrect recognition. The main reason is still the impact of nearby other objects and view angle. Figure 11 shows example of correct recognition.

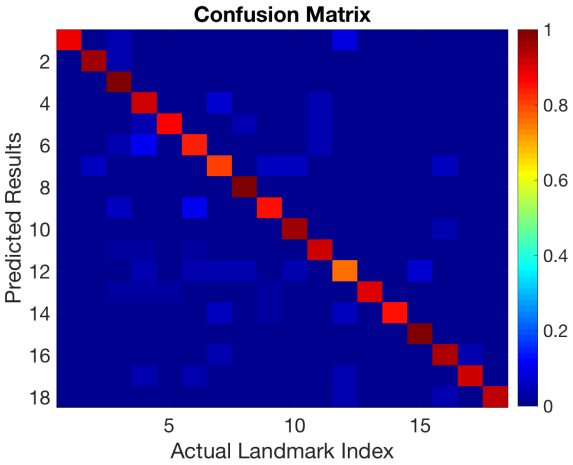

Figure 5. Visualization of confusion matrix using SURF with Method A.

In all 18 landmarks, Eiffel Tower has the lowest recognition accuracy.

5.3. Results - SURF with Method B

The confusion matrix and prediction accuracy of Method A are shown in Figure 5 and 6. The averaged accuracy of training dataset is 93.50%. The averaged accuracy of test dataset is 85.10%. Method B is overall worse than Method A. This indicates that using K-means to selected features is worse than using Method A.

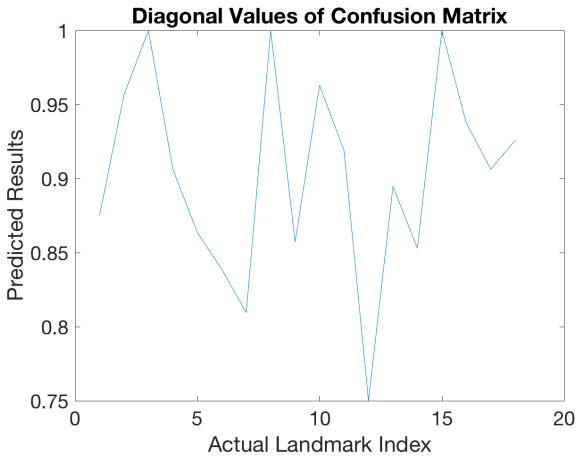

Figure 6. Diagonal values of confusion matrix using SURF with Method A.

Figure 7. Example of correct recognition - dark lighting.

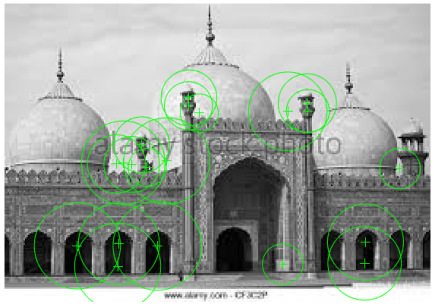

Figure 8. Example of correct recognition - partial view.

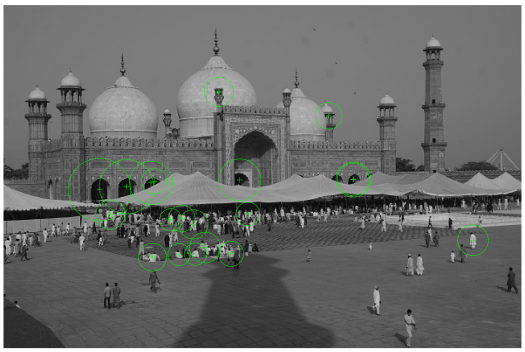

Figure 9. Example of incorrect recognition - people in the scene.

Figure 10. Example of incorrect recognition - other object in the scene.

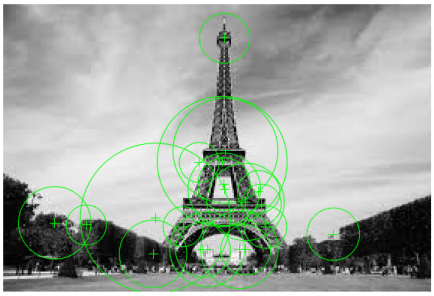

Figure 11. Example of correct recognition - small view.

6. Conclusion and Future Work

Overall the recognition using the SURF with method A is good. It will be interesting the study how deep learning method handle this database.

As mentioned in the results and discussions, the partial

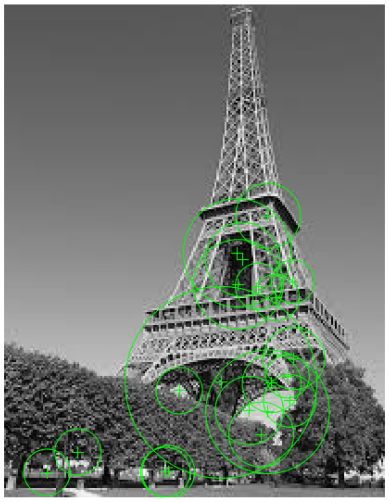

Figure 12. Example of incorrect recognition - partial view.

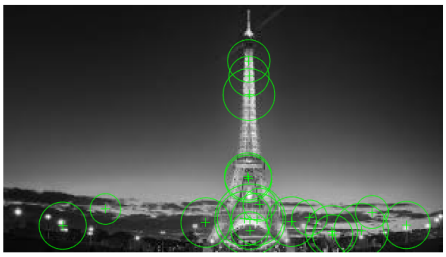

Figure 13. Example of incorrect recognition - small view.

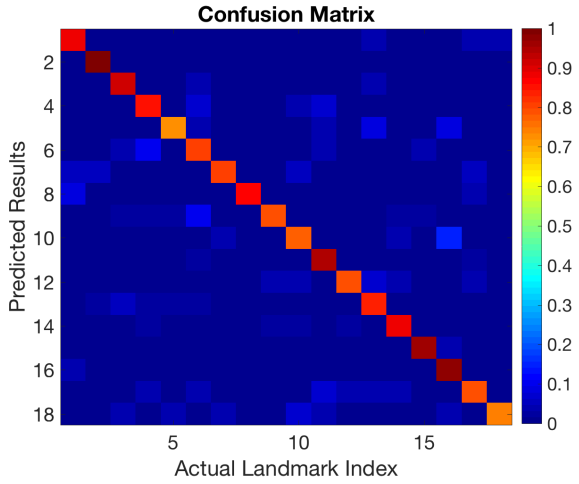

Figure 14. Visualization of confusion matrix using SURF with Method B.

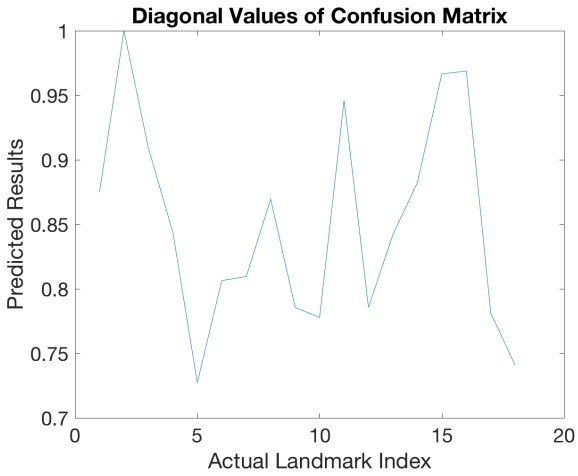

Figure 15. Diagonal values of confusion matrix using SURF with Method B.

and side view issue should be addressed in the future to improve the overall accuracy. In addition, how to detect landmarks with some other objects in the view is also important. For example, people in the landmark scene.

The above two areas will be main directions for future work of this project.

7. References

[1] L. Smith and C. Jones, **The frobnicatable foo filter, a fundamental contribution to human knowledge**. Nature 381(12), 1-213.

[2] A. Crudge, W. Thomas and K. Zhu, **Landmark Recognition Using Machine Learning**, CS229, Project 2014.

[3] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bisacco, F. Brucher, T. Chua and H. Neven, **Tour the World: building a web-scale landmark recognition engine**, 2009.

[4] N. Dalal and B. Triggs. **Histograms of Oriented Gradients for Human Detection**, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1 (June 2005), pp. 886?893.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. **SURF:Speeded Up Robust Features**, Computer Vision and Image Understanding (CVIU).Vol. 110, No. 3, pp. 346?359, 2008.

[6] D. Crandall, L. Backstrom, D.Huttenlocher, J. Kleinberg, **Mapping the world?s photos**. International World Wide Web Conference (2009).

[7] R. Raguram, J. Tighe, J. Frahm, **Improved Geomet-**

ric Verification for Large Scale Landmark Image Collections.