Figure 9: Keyword based image search of the web application

The image could also be uploaded by users with the file path or URL as the query images to search for similar images within the database (Figure 10).

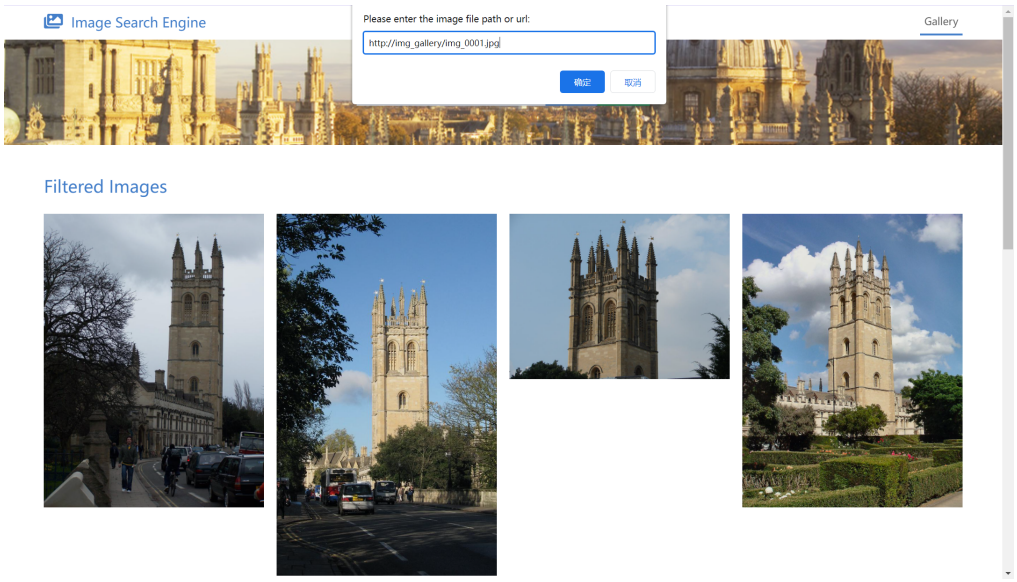

Figure 10: Upload images for the web application

4 Evaluation

Five query images from different landmarks are used to test and compared the performance of each image retrieve method. The result is plotted with the web application model order: the first is the ResNet18 with 430 training images, the second is the ResNet18 with 189 training images, the third is the Autoencode model, and the fourth is the SIFT and VLAD algorithm. The result for the 5 query images are shown in Figure 11 to Figure 15.

The ResNet18 with 189 training images performed the best on the five query images, with an accuracy of 90%; SIFT and VLAD algorithm has a 20% loss on the query images; whereas the ResNet18 trained with 430 images got a surprising only 55% accuracy, the autoencoder method yield very poor search results of only 15% accuracy.

Compared to the result for ResNet18 with 189 and 430 training images in Figure 11, the 430 training model produces the building object with a different point of view, this may be due to more various training images contained in the 430 images dataset, the feature learning obtains more diverse angle information of the buildings. However, more noise images may be introduced in the dataset, the result of the 430 models in Figure 12 and Figure 13 demonstrated that the feature extraction may be interfered with by the texture and color information (the stripe in Ashmolean and the brown color in Balliol).

The SIFT and VLAD algorithm performed quite well on the query images, but it may focus on the non-key points due to the large area of a dark shadow on the query image or the special position of the object. For example, the mistake output image with all souls query image has a large shadow at a similar position to the query image. Similarly, as shown in Figure 14, the white building with the large area tree shares a similar position with the query image, thus may be mistakenly regarded as a similar image.

For the autoencoder method, it produces poor results on the query images except for the christ church image, a possible explanation for this is the large extracted feature vector, the PCA dimensionality reduction could be applied on the extracted feature vector to reduce the effect of non-obvious objects. Furthermore, the depth of the CNN may be inappropriate, more layers could be added for better feature extraction. The advantage of the autoencoder method is that it uses non-labeled images, and the time and cost for image labeling could be significantly reduced.

Moreover, the execution time for SIFT and VLAD methods is more than the CNN-based methods. Thus, for a time-critical real-time image searching system, CNN-based methods may be better than the traditional methods.

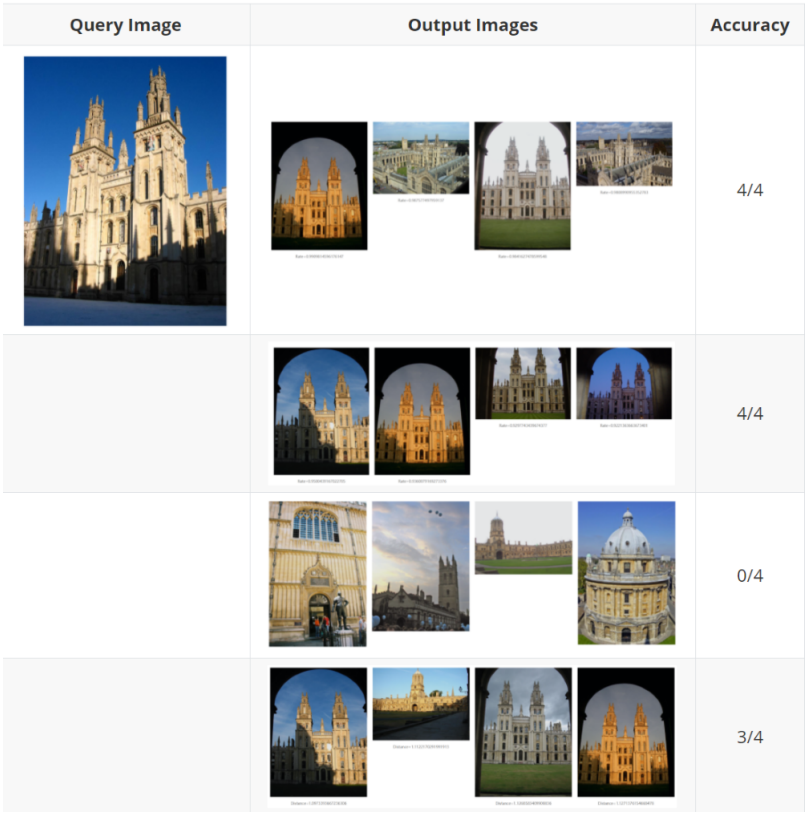

Figure 11: Query image: All Souls

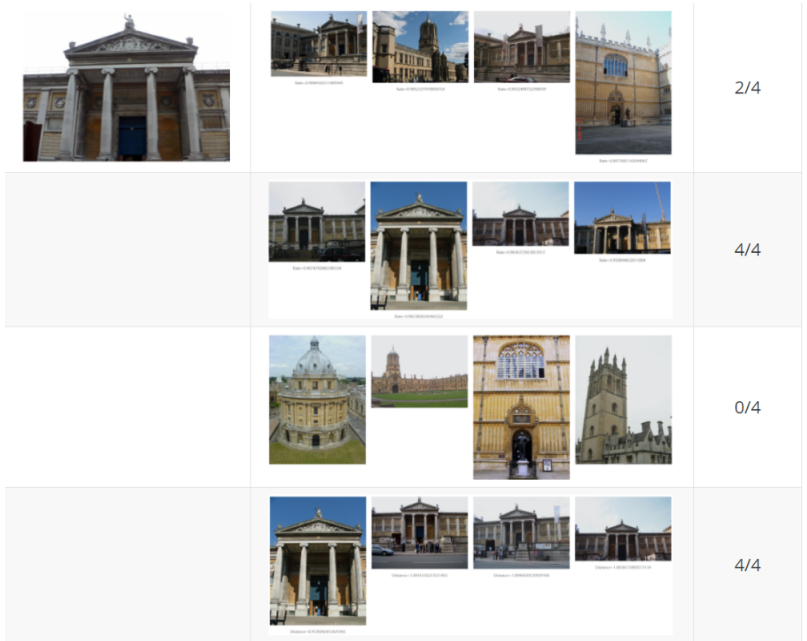

Figure 12: Query image: Ashmolean

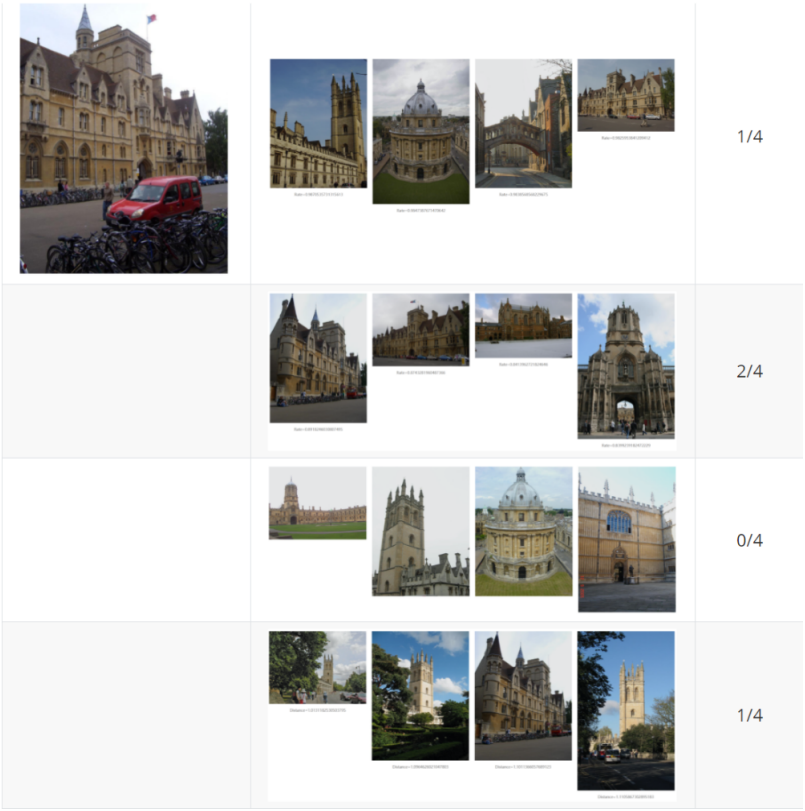

Figure 13: Query image: Balliol

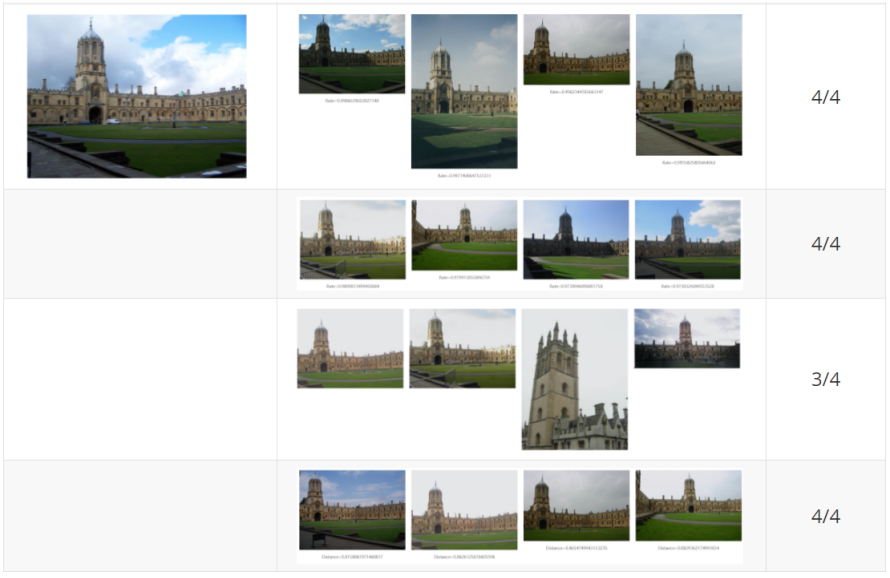

Figure 14: Query image: Christ Church

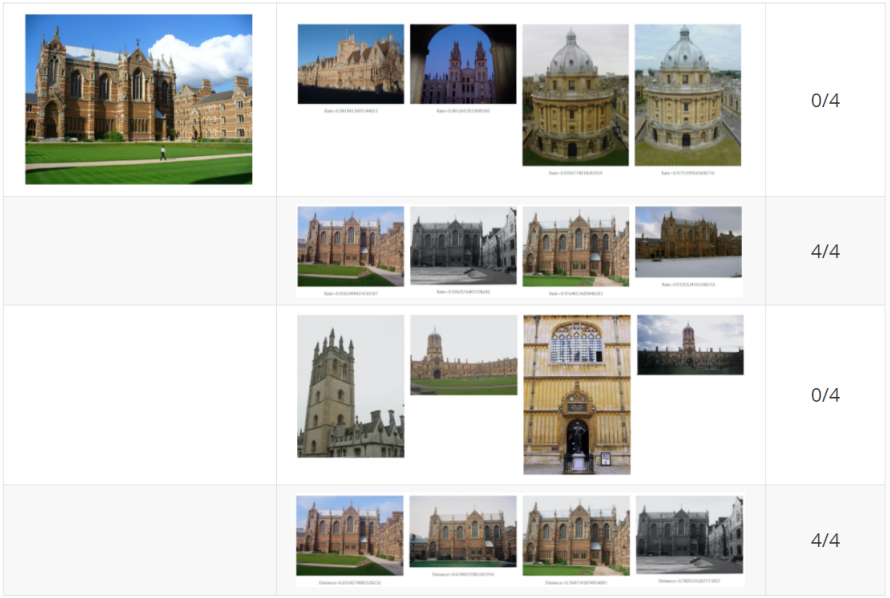

Figure 15: Query image: Keble

5 Conclusion

To summarize, at least six features have been achieved in total for the image search project. The key features implemented are: content-based image retrieval with ResNet18 models trained on 189 images as feature descriptor, content-based image retrieval with ResNet18 models trained on 430 images as feature descriptor, content-based image retrieval with Autoencoder models as feature descriptor, content-based image retrieval with SIFT and VLAD algorithm for feature extraction, text-based image retrieval with keyword search, the web-based application for visualizing the image search results, image search with query images among database or with user upload images.

5.1 Future Work

In addition, there are several future directions for this project:

1. Add search box function, where users could frame the area as the query patch to search the similar results;
2. Apply the GeM layer instead of Max pooling layer for the CNN based image retrieval method [17].

References

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008.

[2] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1):262–282, 2007.

[3] Linda H Armitage and Peter GB Enser. Analysis of user need in image archives. *Journal of information science*, 23(4):287–299, 1997.

[4] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[6] William T Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301. Zurich, Switzerland, 1995.

[7] Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 29(2):300–312, 2007.

[8] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[10] T Karthikeyan, P Manikandaprabhu, and S Nithya. A survey on text and content based image retrieval system for image mining. *International Journal of Engineering*, 3, 2014.

[11] B Dinakaran, J Annapurna, and Ch Aswani Kumar. Interactive image retrieval using text and image content. *Cybern Inf Tech*, 10:20–30, 2010.

[12] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[13] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[14] Alexy Bhowmick, Sarat Saharia, and Shyamanta M Hazarika. Fhvlad: Fine-grained quantization and encoding high-order descriptor statistics for scalable image retrieval. *Multimedia Tools and Applications*, 80(28):35495–35520, 2021.

[15] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[16] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[17] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.