Project Proposal:Instance Based Image Retrieval using Deep Learning

Akanksha Shrimal
Indraprastha Institute of Information and Technology
MT20055
akanksha20055@iiitd.ac.in

Shivam Sharma
Indraprastha Institute of Information and Technology
MT20121
shivam20121@iiitd.ac.in

Shivank Agrahari
Indraprastha Institute of Information and Technology
MT20096
shivank20096@iiitd.ac.in

Pradeep Kumar
Indraprastha Institute of Information and Technology
MT20036
pradeep20036@iiitd.ac.in

Sudha Kumari
Indraprastha Institute of Information and Technology
MT20098
sudha20098@iiitd.ac.in

1. PROBLEM STATEMENT

We propose an approach to solve the problem of instance retrieval. In instance retrieval task, we retrieve images from the database based upon the query or input image. There has been a lot of progress in the image classification and image retrieval domain in recent years. Recently, Instance retrieval has also gained a lot of attention. Its importance is felt in medical sciences, astronomy, security, autonomous vehicles among others. The task is to retrieve images from the database based upon the scene or object in the target image. It deals with the content inside the image such as color, shape and image structure. It can be of particular use when the required feature is partially obscured or the feature is inseparable in the reference image in the database.

2. MOTIVATION

Considering expansion of web data at cosmic level over time, image retrieval is a very essential problem. There are plenty of image retrieval techniques which are evolved. We found that state of the art techniques used for image retrieval(used by tech giants like Google, Bing etc.) perform well while finding relevant images. But we experienced that, the techniques followed there for retrieval consider image as a whole and automatically decide weights to different aspects of the image. There is no user intervention in deciding the part to focus on in the input image while retrieval. We tried to perform some queries found that it does not consider part of the image which is of importance to the user, instead just returns the images which are visually similar to input image at large. Some of the example image query performed on Google Image search are shown in 1a where we have given the following as input images. 1) An animated house 2) A fortress 3) Samsung store. While Google is able to detect the key feature\concept, it overlooked the remaining features. 1b is the proposed retrieval system which allows the user to decide upon the key feature(s) in the query image. The images on left targets a particular subsection of the image and following three images are the expected retrievals.

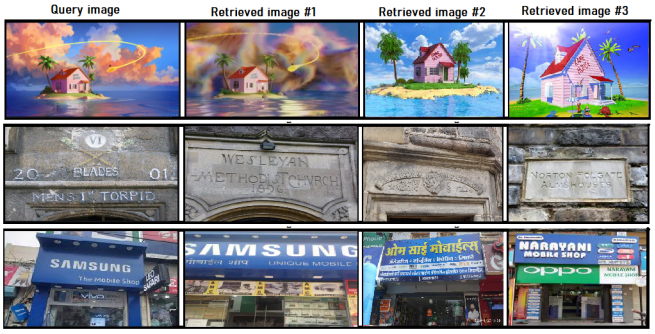

Figure 1a: Current

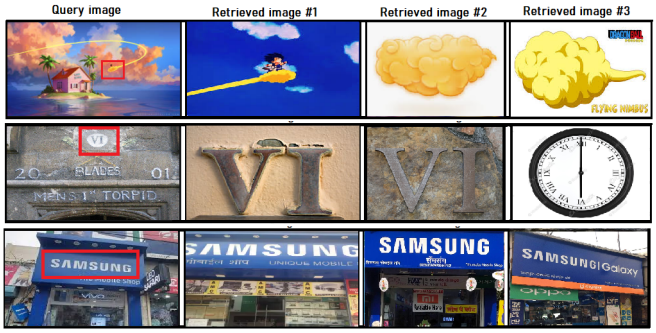

Figure 1b: Proposed

3. LITERATURE REVIEW

Image Retrieval is an important research area in computer vision in which similar images are retrieved from data base w.r.t. a given query image. Basically, the similarity between the query image and the database images is used to rank the database images in decreasing order of similarity.
In the past decade, a number of image retrieval types have been explored which include cross-modal retrieval, sketch based retrieval, multi-label retrieval, instance retrieval, object retrieval,semantic retrieval, fine-grained retrieval and asymmetric retrieval [1].

Image instance retrieval is the problem of retrieving images from a large database that contain or depict similar objects or scene to a target image. Recently, Instance Retrieval has gained popularity as users can retrieve images from database focusing on particular region of image.
Early Techniques on Instance Retrieval task used hand-crafted features for image matching but from past decade there is a shift to deep learning based features[1][2] .

Zheng, Liang and Yang, Yi and Tian, Qi [2] emphasize on a comparison between Scale Invariant Feature Transform(SIFT) based hand-crafted features and Convolutional Neural Networks(CNN) based features for Instance Retrieval. [3]. They stated that SIFT-based methods does not perform well for a specific object retrieval and for common object retrieval it is not as competitive as the CNN models are. On the other hand CNN-based methods with fixed-length representations have advantages in nearly all the benchmarking datasets. When sufficient training data is provided, the ability of CNN embedding learning can be fully utilized. Also the pre-trained CNN modals are competitive. Thus with CNN based methods, the results achieved for instance retrieval are far better then the traditional SIFT based methods and there is shifting from SIFT based methods to CNN based methods.

Followed by hand-crafted approaches, bag-of-features representations were largely used for instance retrieval. In [3] input image is first fed to several CNN-Pooling layers, this results a convolutional feature map of the image, Now, on application of K Means clustering on this feature map several clusters are generated, the quantity of feature-pixels in clusters are represented as a histogram(no. of pixel vs clusters), this representation of image is called Bag of Words(BoW) or Bag of local convolutional features. For finding similar images, BoW of input image is compared with that of other images in the database and images are ranked on the basis of highest value of localization score.

Chandrasekhar, Vijay and Lin, Jie and Morère, Olivier and Goh, Hanlin and Veillard, Antoine [4] stated that current State-of-the-art image instance retrieval pipelines consist of two major blocks: first, a subset of images similar to the query are retrieved from the database, and then geometric consistency checks are applied to select the relevant images from the subset with high precision. The first step is based on the comparison of global image descriptors thus global descriptors are key to improving retrieval performance.
[5] uses the above SOTA pipeline for instance retrieval. They used object detection CNN features from Faster R-CNN to extract both local and global features for given query image. Image wise pooling strategy is used to obtain image descriptors for both query and database images and a initial ranking is obtained based on cosine similarity between the two. After filtering stage, top N elements are locally analyzed using region-wise features from Faster R-CNN and re-ranked. This paper provides a simple baseline that uses off-the-shelf Faster R-CNN features to describe both images and their sub-parts.

One of the baseline model in instance retrieval is R-MAC[6] but it considers fixed spatial pooling which leads to higher computations as fixed grids considered to extract features may not even contain objects one is interested in. To overcome this problem [6] extended the R-MAC [7] approach for Instance Retrieval. Extended R-MAC produced a global image representation by aggregating the activation features of a CNN using a variable region pooling mechanism. They used three-stream Siamese network to optimize the weights of the R-MAC representation for the image retrieval task by using a triplet ranking loss.

Most of the approaches for instance retrieval were not able to produce promising results with low resolution images. *Razavian, Ali Sharif and Sullivan, Josephine and Carlsson, Stefan and Maki, Atsuto* proposed a solution to this in [8]. The paper [8] highlighted the availability of image representations based on convolutional networks and an efficient pipeline to extract local features by taking geometric invariance into explicit account. Variable footprints were used based on the size of the dataset and memory requirements. They made use of the last convolutional layer for the instance retrieval instead of taking the output of the fully connected layer which would have required to crop or edit the image again. Dimensionality of the features is reduced using spatial max-pooling and PCA which increases the efficiency of the model. Multi-resolution search and jittering is used to improve the instance retrieval capability. Using Multi-resolution search helps when the query image has a lower scale in the image and alternatively jittering helps when the query image has a larger scale in the image than the reference images in the database. Similarity between the feature vectors is calculated by calculating the sum of the distance of each query sub-patch.
The datasets used for Instance retrieval include INRIA Holidays, Oxford Buildings, UKBench and Graphics.

4. AVAILABLE DATASETS

In our analysis, we found following datasets that we can use for training and testing our Deep Learning model:

a) Paris Dataset: This dataset contains 6412 images of 12 different landmarks of Paris, it was extracted from Flickr.

b) Oxford Dataset: This dataset contains 5042 images of 11 different Oxford buidings, collected from Flickr.

c) INSTRE Dataset: This is set of Datasets containing diiferent pictres of architectures, buildings,toys, designs, paintings etc. There are three divisions of this datset S1, S2 and M.INSTRE-S1 and INSTRE-S2 contains 11011 images 12059 images respectively. Different objects/designs/scenes are annotated using bounding boxes. In these divisions, a single image may contain multiple annotations(useful for complex Queries). On the other hand INSTRE-M dataset contains 5473 images strictly having only one annotation.

5. PLAN OF WORK

We want to unravel above given issues in the current image retrieval systems by doing these tasks in chronological order:
1. Reproduce existing instance based models on newly available datasets and analyzing its performance.
2. Estimating limitations in performance of existing models on input query of low resolution images and fine tuning the

model for improvement.
3. Development of a final working instance based image retrieval system which can take a region of importance annotation in the query image, and retrieve images by emphasising enclosed region.

Our tentative schedule for completing project work is as follows:

Plan of Work	
Milestones	**Estimated Timeline**
1.Literature Analysis and Feasibility Study	Before Project Proposal Deadline
2.Reproduction of Existing Models	Till Midsem review
3.Analysis of Existing Models	Till Midsem review
4.Development of Proposed Model	After Midsem Review
5.Comparison of proposed model on different Datasets	After Midsem review
6.Final changes And Presentation	On Endsem review

6. REFERENCES

[1] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval using Deep Learning," *arXiv:2012.00641 [cs]*, Nov. 2020.

[2] L. Zheng, Y. Yang, and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval," *arXiv:1608.01807 [cs]*, May 2017.

[3] E. Mohedano, A. Salvador, K. McGuinness, F. Marques, N. E. O'Connor, and X. Giro-i-Nieto, "Bags of Local Convolutional Features for Scalable Instance Search," *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327–331, Jun. 2016.

[4] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A Practical Guide to CNNs and Fisher Vectors for Image Instance Retrieval," *arXiv:1508.02496 [cs]*, Aug. 2015.

[5] A. Salvador, X. Giro-i-Nieto, F. Marques, and S. Satoh, "Faster R-CNN Features for Instance Search," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 394–401.

[6] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning global representations for image search," *arXiv:1604.01325 [cs]*, Jul. 2016.

[7] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," 2016.

[8] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual Instance Retrieval with Deep Convolutional Networks," *arXiv:1412.6574 [cs]*, May 2016.