Face Retrieval using SIFT and HoG features

Jeet Jivrajani
SEAS - AU
jeet.ji.btechi15@ahduni.edu.in

Priya Mehta
SEAS - AU
priya.m.btechi15@ahduni.edu.in

Harshil Shah
SEAS - AU
harshil.s.btechi15@ahduni.edu.in

Vishwa Saparia
SEAS - AU
vishwa.s.btechi15@ahduni.edu.in

Hiren Galiyawala
SEAS - AU
hiren.galiyawala@ahduni.edu.in

Abstract*— *This paper describes various techniques in order to retrieve best possible facial image for the given test image from the database. Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HoG) algorithms are used for feature extraction from face images present in the dataset. Euclidean distance, Support Vector Machine (SVM) and Artificial Neural network (ANN) based classifiers are used to compare the performance of the retrieval. Convolutional Neural Network (CNN) based approach uses AlexNet and it achieves good accuracy for face retrieval over other classifiers.

Keywords*— *Scale Invariant Feature Transform, Histogram of Oriented Gradients, Convolutional Neural Network, Face recognition, Feature extraction

I. INTRODUCTION

Image retrieval has been one of the most prominent problems in various computer vision systems for identification of any entity or finding similarity between two objects. In particular face image retrieval has very essential applications about the person identification for various purpose of security or authorization. Image retrieval is basically the feature extraction from the images, classifying it by any classifier or network and then testing the model on a test image. There are various feature extraction techniques in the current state of art for image processing such as edge detection, corner detection, blob detection, ridge detection, scale invariant feature transform, histogram of oriented gradients. Here we use the Scale Invariant Feature Transform and the Histogram of Oriented Gradients techniques for feature extraction on the face images dataset.

For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination. Such points usually lie on high-contrast regions of the image, such as object edges. Another important characteristic of these features is that the relative positions between them in the original scene shouldn't change from one image to another. SIFT detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors. SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, illumination changes, and partially invariant to affine distortion.

Histogram of oriented gradients (HoG) is a feature descriptor which counts occurrences of gradient orientation in localized portions of an image. It is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. Essential thought behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is the concatenation of these histograms. For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing. The HoG descriptor has a few key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation. Such changes would only appear in larger spatial regions. Coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HoG descriptor is thus particularly suited for human detection in images.

This paper briefs us through different approaches for face image retrieval thereby giving a comparison across methods. At first, SIFT with Euclidean distance approach is shown, followed by feature extraction with HoG and SIFT and then passing them as parameters to classifiers like SVM and Artificial Neural Networks. We also discuss Convolutional Neural Network approach for the deep networks in the approaches section with all the approaches supported by results and graphs in the results section. Finally in conclusion we derive an inference as to which approach is better for what density of image and database.

II. OUR APPROACH

In the face image database of original and sketch images, total 35 persons are selected to create a face recognition system. There are 16 images for each person, so total 36 x 16 = 576 images. Out of 16 images, 12 images per person are used for training i.e. 36 x 12 = 432 images for training. Testing is performed on remaining 144 images, which are not used for training. All the images were resized to 200 X 200 resolution.

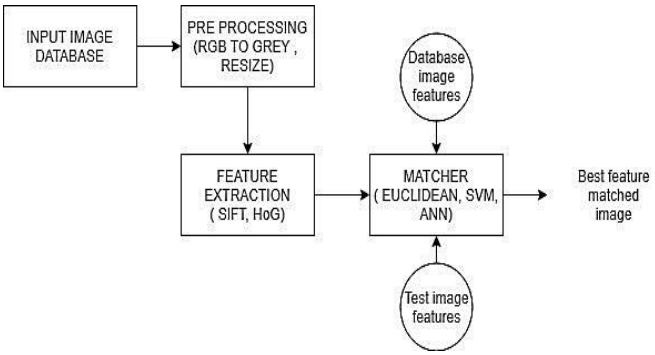

Fig 1. Block Diagram of Face Image retrieval

Here the input images are captured in various emotions and conditions and are a collection of the database. For the Pre-processing, input color image is converted in to gray scale image. Database contains images with various resolutions like 1570 x 1172, 969 x 727, 4313 x 2426 etc. Thus, all the images are resized to the same resolution of 200 X 200. For the feature extraction we are using the SIFT and HoG techniques to obtain the essential and valuable features of the images in the database, we are using 432 images for the training purpose. Now for the models for predicting the best possible image matching to the test image we use various methods and classifiers given below.

A. Euclidean Distance

After getting the important features of the images to be trained and the test image, the best possible match for the test image is retrieved by simply taking the Euclidean distance between the features

$$ED = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Thus it is just a distance measure between a pair of samples *p* and *q* in an *n*-dimensional feature space. We have implemented this algorithm for SIFT and HoG features. The class giving the minimum Euclidean distance is selected as the class for the image. With SIFT features we got an accuracy of 60% and with HoG features we got an accuracy of 80%.

B. SVM classifier

SVM categorizes data by building a hyper plane to separate the data points belonging to different classes. SVM is used for multiclass classification. In our implementation we have used a linear kernel with gamma set as auto and penalty parameter of error term C as 100. Varying these two features we can achieve considerable accuracy on the non-linear classifier. Support vectors for each of the images is computed when fed in the SVM. We have trained the SVM for HoG and SIFT features. The SIFT SVM model gave an accuracy of 65%. The HoG features, when fed in the SVM classifier gave an accuracy of 85%.

C. Artificial Neural Network (ANN)

One of the reasons ANN supersedes SVM is because SVM support vectors grow linearly with the train dataset. This leads to a costly and time consuming operation, even though the accuracy is maintained. Such scenarios demand the need for networks like ANN. Since our dataset is not too large, we have created an ANN with an input layer, two hidden layers and an output layer. The input layer has 256 nodes, the hidden layer has 72 nodes and the output layer has nodes equal to the number of classes which is 36. We have used Keras sequential model with sparse categorical crossentropy loss, Adam optimizer, batch size 70, validation split equal to 20 percent, 150 epochs and dropout equal to 10 percent. We have used both SIFT and HoG features for training the network. With SIFT features an accuracy of 52% was achieved. With the HoG features, an accuracy of 58% was achieved.

D.Convolutional Neural Network (CNN)

ANN operates on the feature vector or an image represented in the form of the vector. Spatial structure is not preserved if the image is fed in vector form. Convolutional Neural Network (CNN) is used to preserve the spatial structure of the image. The approach uses the AlexNet [4] to create the model for face retrieval. Implementation of AlexNet is accomplished using pretrained weights on images from the ImageNet [3] dataset. The last three layers of AlexNet namely (fc6, fc7, and fc8) are fine-tuned for face recognition. It is trained for 30 epochs with learning rate of 0.0001, batch size of 64 and dropout of 0.30. CNN based retrieval uses 29 person out of 36. Thus, total 29 x 12 = 348 images are available for training which may overfit the network. Thus, training set is augmented with various operations like rotation, flipping and brightness increment. Each train image is horizontally and vertically flipped, rotated with 10 angles {1^o, 2^o, 3^o, 4^o, 5^o, -1^o, -2^o, -3^o, -4^o, -5^o} and brightness increased with gamma factor of 1.5. The AlexNet training is accomplished on workstation with Intel Xeon core processor and accelerated by NVIDIA Quadro K5200 of 8 GB GPU. The overall fine-tuning process took 3-4 hours for accomplishment.

III. RESULTS

The face database contains images of 36 persons which are of various resolutions. 16 images are captured for each person. 8 images are original images and other 8 images are sketch version of original images. In the eight original images, 6 images with various expressions (joy, fear, disgust, surprise, sadness and anger) and two images with accessories (hat / cap and goggles / spectacles). For the training purpose

12 images are used, six of which are original images and six the sketch versions of the original images. Testing is done for the remaining 4 images of each person. Training and testing set is given in Table 1.

Table 1 Number of images used for training and testing table

No. of person	Training	Testing		
	No. of image per person	Total Images	No. of image per person	Total Image
36	12	36*12=432	4	36*4=144

Fine-tuning of AlexNet uses 29 person out of 36. Total 29 x 12 = 348 (original + sketch) images are augmented and 4,872 images are used for training the network.

In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. Precision (P) is defined as the number of true positives (tp) over the number of true positives plus the number of false positives (fp). Recall (R) is defined as the number of true positives (tp) over the number of true positives plus the number of false negatives (fn). Precision, Recall and accuracy equations are as follow:

$$Precision\ (P) = \frac{tp}{tp + fp}$$

$$Recall\ (R) = \frac{tp}{tp + fn}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Where $(tp) =$ true positive, tn $=$ true negative, fp $=$ false positive and fn $=$ false negative.

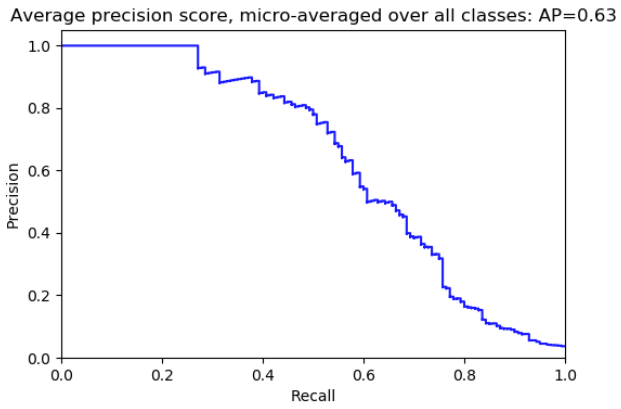

Fig: Scale Invariant Feature Transform using Artificial Neural network

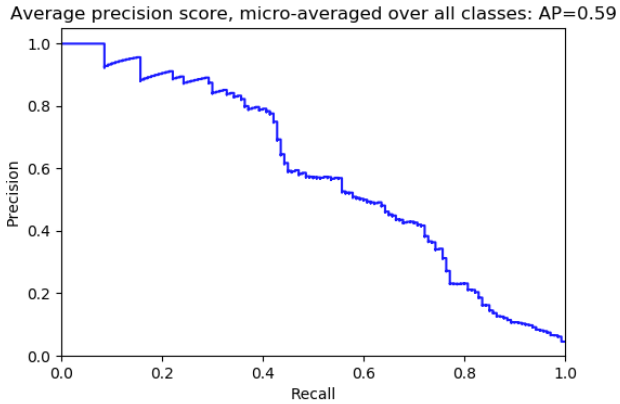

Fig: Histogram of Gradients using Neural network

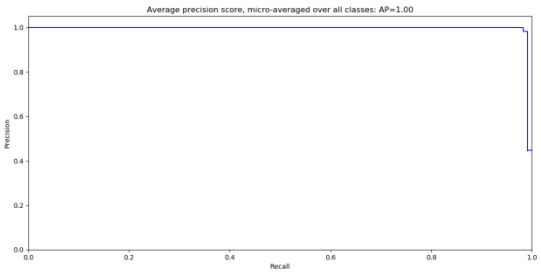

FIG: CONVOLUTIONAL NEURAL NETWORK

IV. CONCLUSION

We have implemented image classification by various methods using the SIFT and HoG features. Evaluating the accuracies and the precision recall graphs, it could be concluded that out of Euclidean Distances, SVM, ANN and CNN, CNN performs the best for the classification of images. We have implemented the same algorithms for images which we rotated by 180 degree and by applying affine transform. We saw roughly the same similarity for the rotated images, but the algorithms did not perform well for affine transformation. From the results obtained, we also inferred that HoG performs better for image classification purposes than SIFT.

V. ACKNOWLEDGMENT

The authors would express their sincere gratitude to Dr. Mehul Raval for his influence on the course of this work and for his suggestions in the improvement of results. We would also like to acknowledge the work done in the state of art which was very helpful during the research.

VI. REFERENCES

[1] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 2004

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, pp. 886-893 vol. 1.

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009, pp. 248-255.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks", in 25th International Conference on Neural Information Processing Systems 2012 (pp. 1097 – 1105)

[5] H. Gao, L. Dou, W. Chen and J. Sun, "Image classification with Bag-of-Words model based on improved SIFT algorithm," *2013 9th Asian Control Conference (ASCC)*, Istanbul, 2013, pp. 1-6.

[6] Hongyan Zhang, Jiazhen Luo, Zihao Wang, Long Ma and Yifan Niu, "An accelerated matching algorithm for SIFT-like features," *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, 2017, pp. 103-107.

[7] K. T. Islam, R. G. Raj and A. Al-Murad, "Performance of SVM, CNN, and ANN with BoW, HOG, and Image Pixels in Face Recognition," *2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE)*, Rajshahi, 2017, pp. 1-4.

[8] S. G. Bhele and V. H. Mankar, "Recognition of Faces Using Discriminative Features of LBP and HOG Descriptor in Varying Environment," *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, 2015, pp. 426-432.

[9] J. Yu and C. Li, "Face Recognition Based on Euclidean Distance and Texture Features," *2013 International Conference on Computational and Information Sciences*, Shiyang, 2013, pp. 211-213.