Face Recognition using One-shot Learning

Nikhil Thakurdesai
Department of computer engineering
K. J. Somaiya College of Engineering
Mumbai, India

Nikita Raut
Department of computer engineering
K. J. Somaiya College of Engineering
Mumbai, India

Anupam Tripathi
Department of computer engineering
K. J. Somaiya College of Engineering
Mumbai, India

ABSTRACT
Face Recognition represents one of the attractive research areas. It has drawn the attention of many researchers due to its varying applications such as security, healthcare, marketing, identity authentication, surveillance etc. In this order, different face recognition algorithms have been proposed however, one algorithm that stands out in the event of limited dataset is one shot learning. "One shot" means learning from a single training item. This paper discusses a way for solving this problem. Neural networks are notorious for requiring extremely large datasets to reach a considerable accuracy. This paper proposes a method to solve this problem for the face recognition domain by bringing down the number of training samples required to just one and still achieving a decent accuracy close to 90%.

General Terms
Face Recognition, Deep Neural Network, Resnet

Keywords
CNN, Transfer learning, fully connected network.

1. INTRODUCTION
The subject of face recognition is as old as computer vision, both because of the practical importance of the topic and theoretical interest from cognitive scientists. It is a method of identifying or verifying the identity of an individual using their face. Despite the fact that other methods of identification (such as fingerprints, or iris scans) can be more accurate, face recognition has always remained a major focus of research because of its non-invasive nature and because it is people's primary method of person identification.

The great progress of face recognition in recent years has made large-scale face identification possible for many practical applications. It is being used for Security purposes like fraud detection. They have also reduced the need for traditional passwords, and improved the ability to distinguish between a human face and a photograph. Another such application of face recognition is Healthcare where it is used to accurately track patient medication consumption and support pain management procedures. Fraught with ethical considerations, marketing is also a burgeoning domain of facial recognition innovation.

Building a face recognizer is not an easy task especially when the dataset is limited. One of the major challenges is caused by the highly imbalanced training data i.e. some of the classes in the dataset may have very less images while other classes may have abundant images. In real world applications this paper has limited training samples for some specific persons, especially when the number of persons to be recognized is extremely large. Additionally, there are other challenges also, introduced by the fact that different persons may have very similar faces, and the fact that the faces from the same person may look very different due to lightning, pose, and age variations.

Furthermore, in the case of mobile face ID unlocking, the use of most machine learning algorithms would require a lot of time, energy consumption, and impractical availability of training data of different faces. It is not very feasible to ask a person to upload a million of his/her images just so that his/her phone can learn to recognize their face for authentication purposes. Thus, this approach is not viable. In such cases, the use of one-shot learning can be employed by learning an object class from only a single piece of data.

2. LITERATURE REVIEW
Jia et al. [1] propose the use of Bayesian reasoning to infer an object category from a few examples; however, in [1] the full, large-scale training set is available during training. Such a system might fail if the available training set is limited.

Transfer learning is a promising technique that promotes the use of deep CNNs in different fields with limited amounts of data. In their paper Xiaogang Li et al. [2] show that convolutional neural network based transfer learning can achieve better classification results in our task with small datasets (target domain), by taking advantage of knowledge learned from other related tasks with larger datasets (source domain). To solve the problem of limited data in [3], Zhongling et al. propose a transfer learning based method, making knowledge learned from sufficient unlabeled images transferrable to labeled target data where pre-trained convolutional layers are reused to transfer knowledge to target classification tasks. According to the experimental results demonstrated in [3], transfer learning leads to a better performance in the case of scarce labeled training data.

In their paper, Gregory Koch et al. [4] present a novel supervised metric-based approach for character recognition with siamese neural networks, then reused that network's features for one-shot learning without any retraining. They have employed large siamese convolutional neural networks which are capable of learning generic image features useful for making predictions about unknown class distributions even when very few examples from these new distributions are available. Then standard optimization techniques are used to train the pairs sampled from the source data; and provide a competitive approach that does not rely upon domain-specific knowledge by instead exploiting deep learning techniques. The model learns to identify input pairs depending on whether they belong to the same class or different classes. This model can then be used to evaluate new images, in a pairwise manner against the test image. The pairing with the highest is then awarded the highest probability for the one-shot task. If the features learned by the model are sufficient to confirm or

deny the identity of characters from one set of alphabets, then they ought to be sufficient for other alphabets, provided that the model has been exposed to a variety of alphabets to encourage variance amongst the learned features. They found that the proposed model outperformed many existing models and have suggested extending one-shot learning tasks in other domains.

Poor generalization ability of the one-shot classes is mainly caused by the data imbalance problem, which cannot be effectively addressed by multi-nominal logistic regression that is widely used as the final classification layer in convolutional neural networks. To solve this problem, Yandong Guo and Lei Zhang [6] propose a novel supervision signal called underrepresented-classes promotion (UP) loss term, which aligns the norms of the weight vectors of the one-shot classes (or underrepresented-classes) to those of the normal classes. Their experimental results show that the proposed UP term significantly helps improve the recognition coverage rate from 25.65% to 77.48% at the precision of 99% for one-shot classes, while still keep an overalltop-1 accuracy of 99.8% for normal classes.

Most machines fail to recognize novel object categories from very few examples. Bharath Hariharan et al. [5] present a low-shot learning benchmark on complex images that mimics challenges faced by recognition systems in the wild. They have further proposed the use of representation regularization techniques to hallucinate additional training examples for data-starved classes. Their method have found to improve the effectiveness of convolutional networks in low-shot learning, improving the one-shot accuracy on novel classes by 2.3x on the challenging ImageNet dataset.

Aishwarya et al. [7], propose the use of deep attribute based representation for one-shot face recognition. They performed fine-tuning of a deep CNN for face recognition on data for specific attributes such as gender and shape of face. The experimental results illustrated that when the face features were further adapted by various attributes, it improved the accuracy for one-shot recognition. This was observed for two different methods, one-shot recognition using Exemplar-SVM based and one-shot similarity kernel based techniques.

Facial landmarks are used to localize and represent salient regions of the face, such as: Eyes, Eyebrows, Nose, Mouth. Dlib framework has proved to be an important framework to extract facial landmarks from an image. In [8], Brandon et al. have used dlib's pre-trained face detector for higher accuracy than OpenCV's detector. Similarly, after comparing the Viola Jones detector with the dlib detector, Banzhaf [9] found that dlib is expensive but higher in accuracy than Viola Jones

3. DATABASE

One of the major drawbacks of most of the machine learning algorithms is that they require training on hundreds or thousands of images and very large datasets. One shot learning is an object categorization problem which takes care of this problem by learning information about object categories from one, or only a few, training images.

Initially images of the project members were used to prepare a dataset of 3 classes. Then the number of classes were increased to 10 by adding classes for celebrities. FEI dataset was used to further increase the number of classes to 50. creating a robust database with 50 classes.

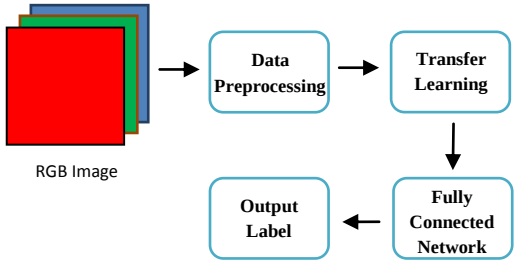

Fig 1. Block diagram of the entire methodology

Furthermore, dataset included photos taken with various backgrounds, head rotation, lean, tilt, scale (distance from the camera), angles and different contrasts. Also, the subjects selected were from different races, cultures and ethnicities. The subjects were also both male and female and of different age ranges.

4. METHODOLOGY

This paper discusses implementing face recognition using 1-shot, 2-shot, 5-shot and 10-shot learning which implies that only 1 training image was used in the first method, 2 training images in the second method and so on. The number of faces the system could detect were also varied between 10, 20 and 50 and the results were compared. The block diagram of our method is shown in Figure 1. The procedure used in each block is discussed in the following sections.

4.1 Data Preprocessing

The library used for preprocessing the data is dlib. It is a modern toolkit containing Machine Learning algorithms and tools for creating complex software to solve real world problems. The image of a person's face is passed into the dlib object to locate the face in the image followed by creating a bounding box around it. The co-ordinates of the box are then retrieved and stored. The process is shown in Figure 2.

A 68 landmark predictor file was used to locate the predefined 68 landmarks, i.e. 68 points, on a person's face once the face has been located. The image along with the co-ordinates of the bounding box were passed in the dlib's shape predictor method. Thus, the 68 predefined points were located on a person's face. The 68-points are shown in Figure 3.

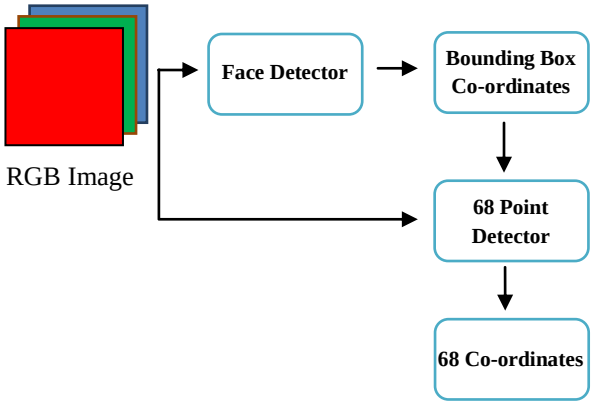

Fig 2. Data Preprocessing

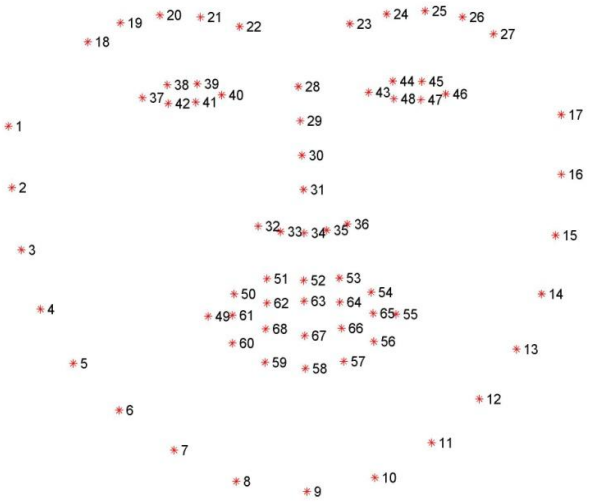

Fig 3. 68-point markup

4.2 Transfer Learning

Transfer learning is a research problem in Machine Learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained when learning to identify a cat may be used when trying to recognize other members of the cat family. This helps us to avoid unnecessary repeated training and obviate the need for excessive hardware for training. This also helps in achieving better results. In transfer learning, usually the last layer is fine tuned to make the system identify classes of your own instead of the classes it was originally trained on.

After the preprocessing step, we have the co-ordinates of the 68 predefined points on a person's face. We used a resnet model for transfer learning. The resnet model was pretrained on the triplet loss function. The triplet loss function and the resnet models are discussed in the following sub-sections.

4.2.1 Triplet Loss

In face recognition, triplet loss is used to learn good embeddings for faces. It was introduced in the Facenet paper by Google [10]. The triplet loss takes into consideration three images. The images are called "anchor", "positive" and "negative". The "anchor" image is the image of the person whose face is to be recognized by the system. The "positive" image is another image of the same person. The "negative" image is an image of some other person. We need to make the network learn that the distance between the encodings of "anchor" image and the "positive" image is less than the distance between the encodings of "anchor" image and the "negative" image. If we denote the "anchor" by "A", "positive" by "P", "negative" by "N", then the equation can be written as:

$$d(A,P) \leq d(A,N)$$

This can be rewritten as:

$$d(A,P) - d(A,N) \leq 0$$

This equation however, has a drawback that, if the distance encoding is always set to output zero, then the equation will still be satisfied leading us to the wrong result. Thus, we can keep a margin, alpha, to prevent this from happening. The equation can be correctly re-written as:

$$d(A,P) - d(A,N) + \alpha = 0$$

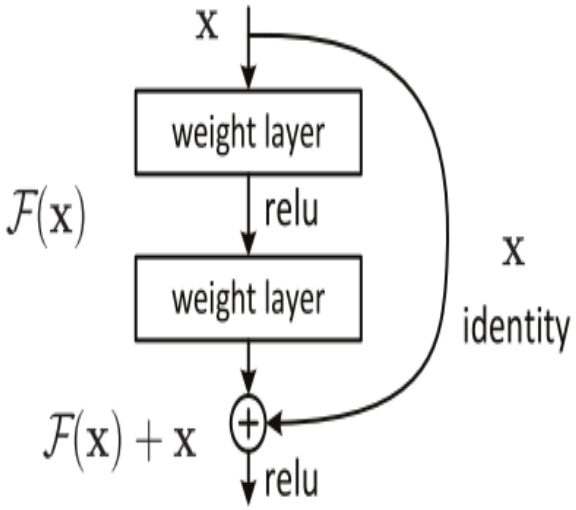

Fig 4. Resnet architecture

The network is now trained on triplet loss which helps it in getting better at distinguishing positive from negative samples. The closer the negative sample is to the positive sample, the better the training of the network in distinguishing between the two.

4.2.2 Resnet

Resnet stands for residual network, a type of neural network used in machine learning. A resnet model has been used for transfer learning. The resnet was trained on the triplet loss function. The basic structure of a resnet is shown in Figure 4.

The 68 facial points extracted at the end of data preprocessing step were given as input into the resnet along with the image of the person. The resnet outputs a 128-feature vector. This 128-feature vector is sent to the fully connected layer for classification. The process is shown in Figure 5.

4.3 Fully Connected Network

This is the only part of our model whose weights are trained by us. The resnet is fine-tuned and connected to the fully connected layers. The 128-feature vector output of the resnet is input to the fully connected network. It is a 2-layer network with the number of hidden layer nodes equal to 68. The number of output classes were changed to see the corresponding effect on the accuracy of the model. The fully connected network is shown in Figure 6. A learning rate of 0.03 was used and the model was trained for 100 epochs.

5. RESULTS

Many experiments were conducted with different models and hyperparameters to see which model fits the data well and more so, generalizes to images it hasn't seen before. The

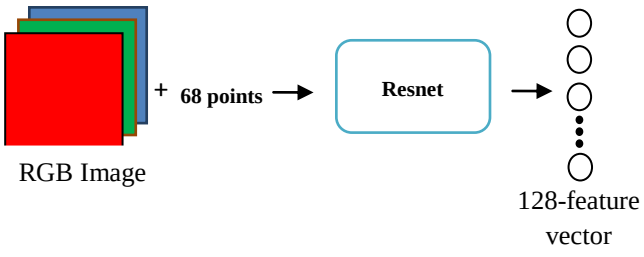

Fig 5. Function of the Resnet

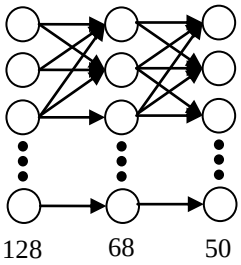

Fig 6. Fully Connected Network

model proposed in the above section gave us the best results. An accuracy more than Omkar et al. [11] was achieved which couldn't generalize well to images the model hadn't seen before as the number of their output classes increased. The accuracy of the model proposed in this paper does not vary by much, no matter the number of output classes. A reasonable accuracy was achieved using just one training image and the accuracy rises even greater for five training images, which is still a feasible number. The performance of the model proposed, for different number of training images and different number of output classes is shown in the following tables.

Table 1. 1 shot 30 classes

Number of output classes	30
Number of training images	30
Number of testing images	308
Training Accuracy	86.62 %
Testing Accuracy	88.64 %

Table 2. 2 shot 30 classes

Number of output classes	30
Number of training images	60
Number of testing images	278
Training Accuracy	91.44 %
Testing Accuracy	96.04 %

Table 3. 5 shot 30 classes

Number of output classes	30
Number of training images	150
Number of testing images	188
Training Accuracy	97.14 %
Testing Accuracy	96.28 %

Table 4. 1 shot 50 classes

Number of output classes	50
Number of training images	50
Number of testing images	553
Training Accuracy	88.30 %
Testing Accuracy	90.96 %

Table 5. 2 shot 50 classes

Number of output classes	50
Number of training images	100
Number of testing images	503
Training Accuracy	97.30 %
Testing Accuracy	89.07 %

Table 6. 5 shot 50 classes

Number of output classes	50
Number of training images	250
Number of testing images	353
Training Accuracy	99.42 %
Testing Accuracy	97.73 %

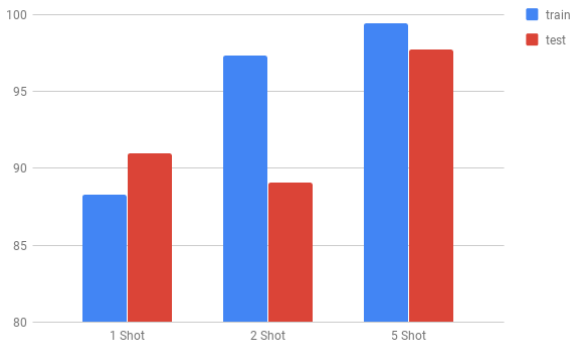

Fig 7. 30 Classes

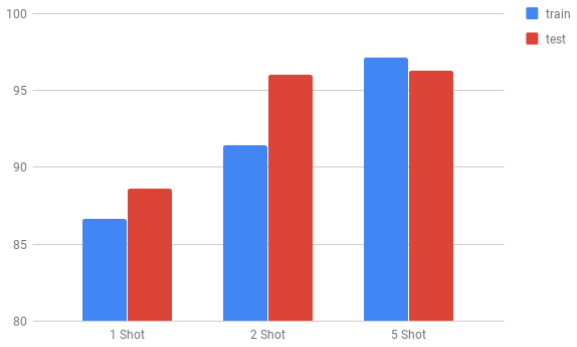

Fig 8. 50 Classes

Figure 7 and 8 show the graphical representation of these results for 30 and 50 classes respectively. Thus, this method

can be used to create a face authorization software which would only allow entry to authorized personnel. The accuracies show that, if the people are asked to just upload five images of their own into the database, the software will be trained to a testing accuracy of 97.73% which is pretty reliable.

6. FUTURE WORK

The above implementation still has a lot of shortcomings. It does not provide a reality check, whether the person standing in front of the camera is a real person or not, or is it just an image of that person. This is still a very challenging task for computers to differentiate between images of real people and just still images of people. Although, we believe that our work will serve as a foundation to other researchers who want to make further improvements to the liveness detection challenge and also in improving the overall face recognition software.

7. REFERENCES

[1] Jia, Yangqing, and Trevor Darrell. "Latent task adaptation with large-scale hierarchies." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[2] Li, Xiaogang, et al. "Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification." Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on. IEEE, 2017.

[3] Huang, Zhongling, Zongxu Pan, and Bin Lei. "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data." Remote Sensing 9.9 (2017): 907.

[4] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." ICML Deep Learning Workshop. Vol. 2. 2015.

[5] Hariharan, Bharath, and Ross B. Girshick. "Low-Shot Visual Recognition by Shrinking and Hallucinating Features." ICCV. 2017.

[6] Guo, Yandong, and Lei Zhang. "One-shot face recognition by promoting underrepresented classes." arXiv preprint arXiv:1707.05574 (2017).

[7] Jadhav, Aishwarya, Vinay P. Namboodiri, and K. S. Venkatesh. "Deep attributes for one-shot face recognition." European Conference on Computer Vision. Springer, Cham, 2016.

[8] Amos, Brandon, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. "Openface: A general-purpose face recognition library with mobile applications." CMU School of Computer Science (2016).

[9] Banzhaf, Clint. Extracting facial data using feature-based image processing and correlating it with alternative biosensors metrics. MS thesis. 2017.

[10] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[11] Omkar Ranadive and Dhiti Thakkar. k-Shot Learning for Face Recognition. International Journal of Computer Applications181(18):43-48, September 2018.