k-Shot Learning for Face Recognition

Omkar Ranadive
Department of Computer Science
K.J Somaiya College of Engineering
Vidyavihar, Mumbai - 400077

Dhiti Thakkar
Department of Computer Science
K.J Somaiya College of Engineering
Vidyavihar, Mumbai - 400077

ABSTRACT

There have been many recent advancements in the field of artificial intelligence and machine learning. Nevertheless, the problem of learning from a few examples persists. The process of learning from just an example is easy for humans but not for a computer. Learning from a small number of samples is especially necessary in the case of facial recognition systems as the number of samples per person is limited.

The aim is to explore, analyze and improve the different techniques which can be used for Face Recognition where the algorithm is fed with a few examples of faces i.e**. the process of k shot learning for Face Recognition** has been explored using the LFW and FEI datasets. The techniques of transfer learning have been used along with the famous Dlib library with some improvements using methods of deep learning.

General Terms

Deep Learning, Face Recognition, Transfer Learning

Keywords

k-shot Learning, resnet, Dlib, skip connections

1. INTRODUCTION

In human beings, the areas associated with vision in the temporal lobe of the brain interpret the meaning of visual stimuli and establish object recognition. These areas appear to be involved in the high-level visual processing of complex stimuli such as faces and scenes. The neurons of the temporal lobe respond to certain features of the face and store them, eventually leading to face identification. This makes a human require very few examples to remember a particular object and when given an unknown object, a human can classify it accurately.

Today, machine learning systems are fed with colossal amounts of data which are interpreted and stored. However, it is better if a system can recognize a particular face with very few examples because huge amounts of data are not available consistently. For face recognition, the system should be efficient enough to be able to recognize a person's identity by just feeding few pictures to the system.

This engenders the concept of "k-shot" learning, which consists of learning a class from only 'k' labeled examples. The work aims at recognizing faces with maximum accuracy by using few(k) examples of faces of each class.

2. LITERATURE SURVEY

Face recognition being vital for various applications like entertainment, smart cards, information security and law enforcement, has massively attracted researchers and developers in the field of machine learning. A holistic view can be stated in three steps - Face Detection, Extraction and Recognition [1]. Understanding the concepts of face detection before extraction and recognition build a foundation for further research. Thus, the survey given in [2] forms a basis of the techniques used here. Face detection can be classified as feature-driven or image-driven where the former includes edges, color, gray-scale, etc and the latter includes neural networks, subspaces and statistical approaches. Here, the image-driven neural networks have been used. A simple method for face detection was used in [3] where a neural network was just trained to identify whether a there is face in a picture. But the model implemented here focuses on the detection, extraction and recognition of faces. Firstly, the concepts of transfer learning [4] are to be learnt. Thus, a complete survey on transfer learning was referred to, which acknowledges the fact that training and testing data do not necessarily belong to the same feature space or distribution. Transfer learning or knowledge transfer, in such situations, is the most desirable. Here, the focus is on transfer learning for classification. Lake et al. [5] used a cognitive-based approach to solve the problem of one(k=1)-shot learning where they used a generative model to show how knowledge from previous characters helps to infer the latent strokes in novel characters. A more recent advancement are the Matching Nets [6], which use attention and memory to enable rapid learning. This neural net focuses on the non-parametric structure to make it more adaptable to new training sets. Another recent method is the Memory-Augmented Neural Networks [7] which uses the concepts of meta learning and rapidly assimilates new data and makes accurate predictions even if there are very few examples. These networks hold the information about the predicted output in their 'short term memory' until an actual sample class information is received in the next quantum.

After which the information is bounded and stored in 'long term memory'. Moreover, even the Siamese Neural Network [8] architecture has been greatly used in the field of k-shot learning by exploiting the power of convolutional models to generalize on the predictive powers not just to new data but also data of different distributions. A CNN feature extractor with a learnable parametric function was used in the Deep Face Recognition [9]. Moreover, the DeepID3 [10] achieved an amazing accuracy on the LFW dataset for face verification and identification by using a very deep convolutional architecture with inception nets. The strategy used here underscores the use of deep neural networks for extracting feature representations and using face-verification as supervisory signals and they are input to intermediate layers, thus boosting the accuracy.

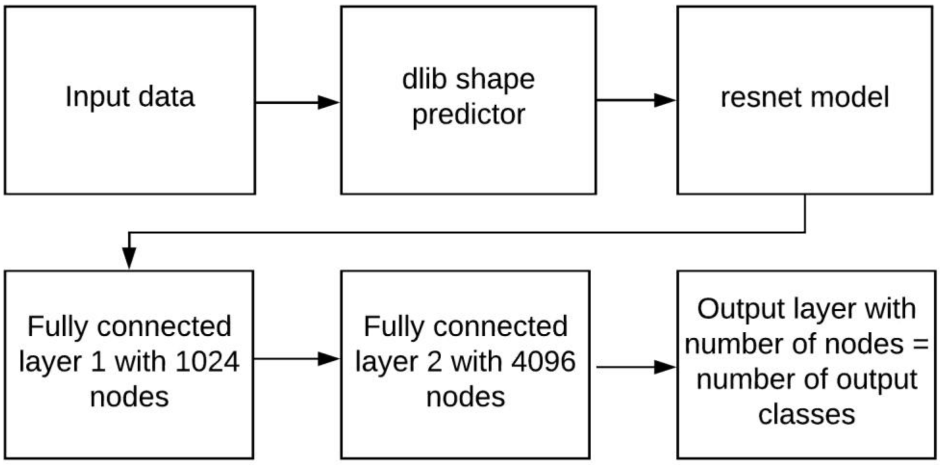
Input data
dlib shape predictor
resnet model
Fully connected layer 1 with 1024 nodes
Fully connected layer 2 with 4096 nodes
Output layer with number of nodes = number of output classes

Figure 1: Model Architecture of the Face Recognition System

In the endeavor to solve the problem of k-shot learning for Face Recognition, the initial KNN [11] and SVM/Grid Search models with PCA have also been used and compared with the dlib-resnet model [12, 13] with some modifications by adding fully connected layers, thus exploiting the advantages of neural networks.

Initial PCA

PCA, Principal Component Analysis, is an important method for feature extraction which is greatly used for dimensionality reduction. PCA, through the covariance matrix, helps in understanding how one variable is related to every other variable. The eigenvectors which give the direction and the eigenvalues which give magnitude are vital components of this technique. Through these components, the technique causes dimensionality reduction and thus dropping certain features and only capturing the most important ones. Thus, the PCA is also used for Face Recognition [14]. Eigenfaces are formed for all the images in the dataset by these weighted eigenvectors which are obtained by the covariance matrix of the image set. After the most important eigenfaces are obtained, a classification algorithm for Face Recognition can be used. Here, the KNN (K nearest neighbors using Euclidean distance) and Support Vector Machines algorithms have been used. But, since the aim is to get maximum accuracy for a k-shot recognition problem where k is either less than or equal to 3, these techniques give a very poor performance because there is insufficient information given by only 3 or less images, thus, giving way to newer techniques to tackle the problem of k-shot learning for Face Recognition. In this paper, the PCA method has been used to compare it with the advanced methods and focus is more on the latter.

3. MODEL

3.1 Architecture

As the model is supposed to work on a very limited amount of data (k samples), the concept of transfer learning has been used. The input data is fed through the dlib library to detect 68 facial landmarks. This is now passed on to a resnet model which outputs a 128-feature vector. This 128- feature vector is passed on through the fully connected 2-layer neural net architecture with 1024 nodes in the first layer and 4096 nodes in the second layer. The architecture is shown in figure 1.

3.2 Extraction of features using Dlib

The Dlib C++ library is widely used for Face and Facial Landmark Detection. This facial landmark detection is actually an implementation of the 'One Millisecond Face Alignment with an Ensemble of Regression Trees' paper by Kazemi and Sullivan [15]. Its face detection is based on the concept of Histogram of Gradients with a linear classifier and sliding window detection. Here, the pre-trained weights of the 68 facial landmarks detector were used. These 68 landmarks include the mouth, right and left eyebrows, right and left eyes, nose and jaw. The training data actually includes 68 (x,y) coordinates which are manually labelled. Not only was this algorithm trained on the iBUG 300-W dataset for 68 landmarks, but also trained on the HELEN dataset for a 194-point model. At the end, the result is a very powerful face landmark detector with highly accurate predictions. In the model, the dlib library is used to find 68 facial landmarks which are passed on to the resnet model along with the image. The visual position of these landmarks is shown in figure 2 as follows:

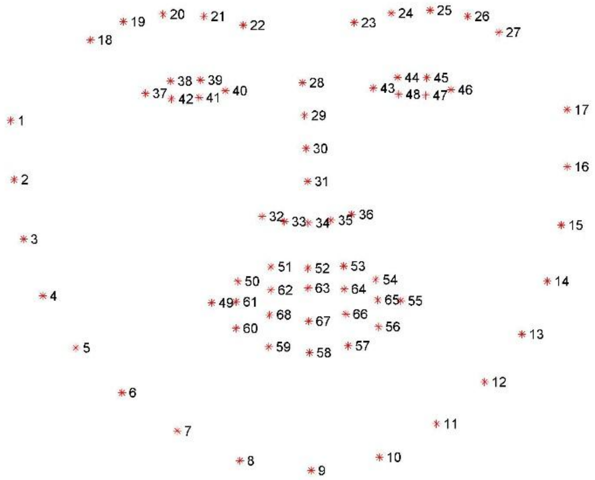

Figure 2: 68 facial landmarks extracted by Dlib

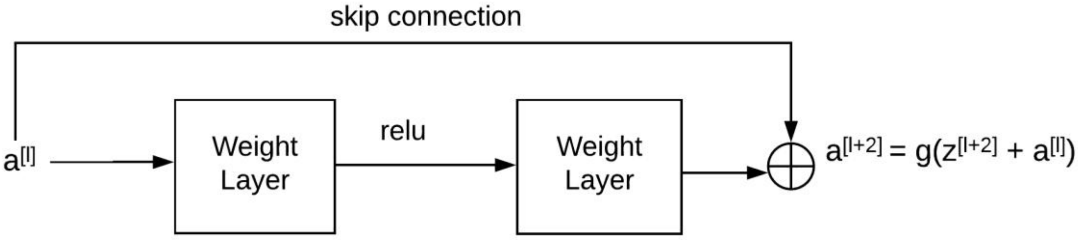

Figure 3: Residual Block

3.3 Generating facial features using Resnet model

The Resnet model used is a Resnet model with 27 convolution layers, that is, a variation of Resnet-34 used in [12], with few layers removed and number of filters reduced by half. The deeper the neural net, the harder it becomes for a neural network to generalize well and at a certain point the training error starts increasing with the increase in number of layers. A Resnet solves this problem by introducing skip connections. That is, the output of some layer ($a^{[l]}$) not only goes to its next layer as input but also goes directly to some layer k ahead (where k > l). For example, a residual block is shown in figure 3 where a layer $a^{[l]}$ is having a skip connection to $a^{[l+2]}$. The skip connections go to the layer ahead after the z calculation is done for that layer but before passing it through the activation function of that layer. That is, $a^{[l+2]} = g(z^{[l+2]} + a^{[l]})$. Thus, a residual neural network works well because it can nullify the effect of intermediate layers by setting the weight vectors close to zero.

In the model, the residual neural network is generating a 128-feature vector. The resnet model chosen is already trained on a large number of faces and thus can produce a sufficiently different 128-feature vector for different people. Basically, different people will get mapped to a different output space and if the same person (but a different image) is given as input, the model will produce a similar 128-feature vector. This 128-feature vector is then passed on to a fully connected layer.

3.4 Fully Connected Layers

The 128-feature vector is passed on to a 1024 node fully connected layer which is then passed on to a 4096 fully connected layer. The actual retraining for k-shot learning task is done on these last two layers only. Due to this, the training phase of k-shot learning is small, and the model can be quickly deployed. Training was performed on 1000 epochs and with a learning rate of 0.0001. An extra parameter epsilon was chosen which was adjusted as per the dataset to avoid the problem of vanishing/exploding gradients.

4. EXPERIMENTAL RESULTS

4.1 LFW Dataset

The Labelled Faces in the Wild (LFW) dataset is a set of face photographs designed for studying the problem of unconstrained face recognition. The 8 subjects were chosen from the dataset and k (small number) of samples of each subject were fed into the algorithm.

- **Uneven distribution (k <= 7) :** Firstly, the LFW dataset was tested on an uneven distribution of training samples. That is, some subjects had only 1 training image, some had 2 and so on. The results can be seen in table 1. The samples with a single training example (Allyson and Amelia) were misclassified in some instances. To alleviate this problem, a same number of samples (even distribution) per class should be used.

- **One-shot Learning (k = 1):** The model was then trained on a single sample each instead of an uneven number of training samples. This one-shot learning approach gave a better training and testing accuracy as seen in table 2. From the 23 testing images, only a single side-faced picture was misclassified.

- **k-Shot Learning (k = 3):** The model was now trained on three images per subject instead of a single image. In this scenario, a perfect training and testing accuracy was achieved as seen in table 3.

- **k-Shot Learning (k = 3) with a larger testing set:** The model with K = 3 per subject was now tested on a larger testing set. Even with more variations in the testing examples including selfies and low-resolution shots, a perfect accuracy was achieved as seen in table 4. The larger dataset (testing images = 63 and training images = 24) was also trained and executed on some existing facial recognition techniques like SVM, KNN and Grid Search. The results are shown in table 5. As expected, the old techniques perform poorly when supplied with meagre training data.

4.2 FEI Dataset

The FEI face database is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil.

Table 1: Uneven distribution of k (k <= 7)

Total training images:	27
Total testing images:	23
Total Output Classes	8
Images per class	
Agnes	1
Alexander	4
Allyson	5
Amelia	7
Angelo	4
Anibal	3
Anthony	2
Yekaterina	1
Train accuracy	96.2963%
Test accuracy	82.608%

Table 2: One shot learning (k=1)

Total training images:	8
Total testing images	23
Total Output Classes	8
Images per class	
Agnes	1
Alexander	1
Allyson	1
Amelia	1
Angel	1
Anibal	1
Anthony	1
Yekaterina	1
Train accuracy:	100 %
Test accuracy	95.65 %

Table 3: Even distribution of k. (k = 3)

Total training images:	24
Total testing images	23
Total Output Classes	8
Images per class	
Agnes	3
Alexander	3
Allyson	3
Amelia	3
Angelo	3
Anibal	3
Anthony	3
Yekaterina	3
Train accuracy:	100 %
Test accuracy	100 %

Table 4: k = 3 and a larger testing set

Total training images:	24
Total testing images	63
Total Output Classes	8
Images per class	
Agnes	3
Alexander	3
Allyson	3
Amelia	3
Angelo	3
Anibal	3
Anthony	3
Yekaterina	3
Train accuracy:	100 %
Test accuracy	100 %

Table 5: Testing the larger set on old facial recognition techniques

Total training images	24
Total testing images	63
Total output classes	8
KNN	4/63 correct, 6.34% accuracy
SVM	10/63 correct, 15.87% accuracy
Grid Search	7/63 correct, 11.11% accuracy

- **k-Shot Learning (k = 3) with large number of output classes:** The model was trained on a dataset with a large number of output classes (50 different subjects) with 3 training images per subject. The results are shown in table 6. As the number of output classes exceeds a certain limit, the model is no longer able to classify with a high accuracy. A 10% increase in both training and testing accuracy was achieved after performing data augmentation on the initial 3 images per subject.

- **k-Shot Learning (k = 3) with 30 and 20 output classes:** The model gave a better accuracy when the number of output classes was reduced to 30 from 50. The results are shown in table 7. By reducing the number of classes down further to 20, the results obtained are even better. [table 8]

- **k-Shot Learning (k = 3) with 10 output classes:** When the number of output classes was reduced down to 10, the model was able to perform perfectly and gave a 100% testing and training accuracy as seen in table 9.

The graph of number of subjects against training and testing accuracy is shown in figure 4.

5. CONCLUSION

The proposed model was able to give 100% accurate results as long as the number of subjects was limited i.e. maximum accuracy with minimum number of subjects. For a large number of subjects, the model fails to separate out the subjects correctly and the accuracy falls to 58.66% for training and 52% for testing. i.e. low accuracy for many subjects. However, this is still in accordance with how humans perform one-shot learning.

Table 6: 50 subjects with k = 3

Total training images:	150
Total testing images:	200
Total Output Classes	50
Images per class	
All 50 subjects:	3 images per subject
Train accuracy	58.666%
Test accuracy	52%
Train accuracy (after data augmentation)	68.14%
Test accuracy (after data augmentation)	60%

Table 8: 20 subjects with k = 3

Total training images:	60
Total testing images:	80
Total Output Classes	20
Images per class	
All 20 subjects:	3 images per subject
Train accuracy	90%
Test accuracy	92.5%

Table 7: 30 subjects with k = 3

Total training images:	90
Total testing images:	120
Total Output Classes	30
Images per class	
All 30 subjects:	3 images per subject
Train accuracy	88.89%
Test accuracy	77.5%

Table 9: 10 subjects with k = 3

Total training images:	30
Total testing images:	40
Total Output Classes	10
Images per class	
All 10 subjects:	3 images per subject
Train accuracy	100%
Test accuracy	100%

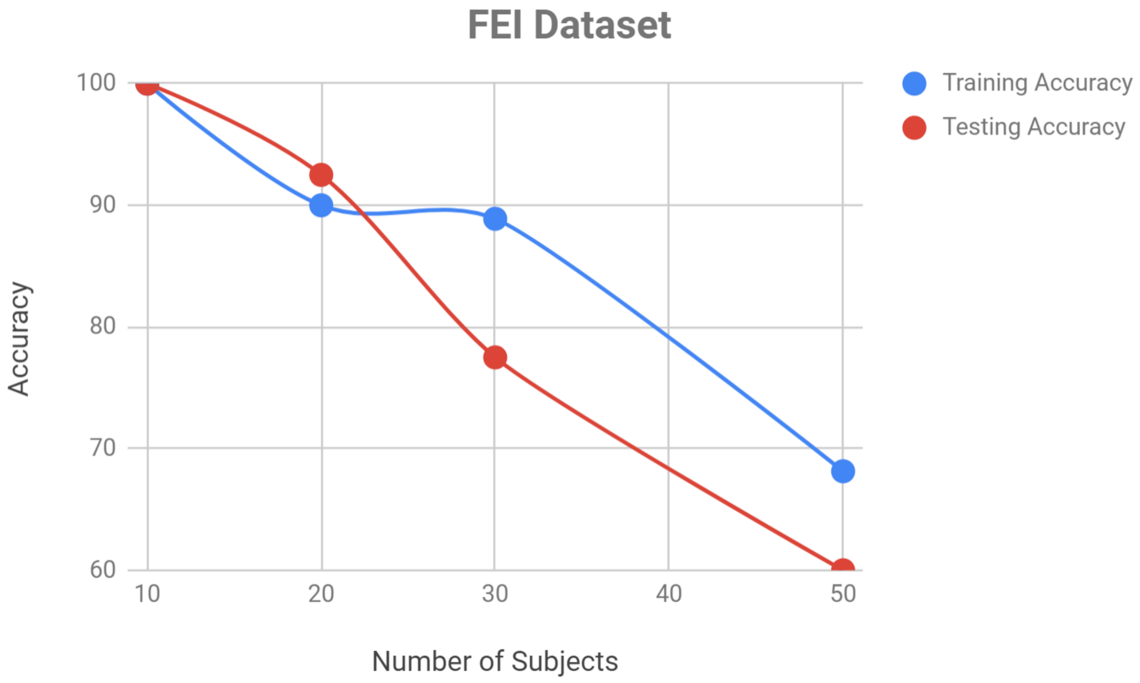

Figure 4: Train and test accuracy on different number of output classes

If humans are given a large number of faces and told to remember them, they would most likely forget them. For example, in a daily commute, people come across many different faces, but they do not end up remembering all of those faces by seeing them once. This implies that even humans are bound to make errors in recognising other unknown humans by just one example or just one look. Thus, a sound conclusion — results using the method used in this paper are in accordance to humans in terms of k-shot Facial recognition — can be made.

6. FUTURE SCOPE

The main problem with the proposed model is that the accuracy starts falling as the number of subjects increase i.e. when the algorithm has too many faces to remember or learn. This problem can be alleviated by combining this proposed model with the current advancements in the field of one-shot learning like Neural Turing Machines, Memory Augmented Networks and Matching Networks. Thus, this combined model will be able to outperform even humans in the task of One-Shot Facial Recognition.

7. ACKNOWLEDGMENTS

The research and experiments were performed under the guidance of Mrs. Grishma Sharma, Assistant Professor at K.J Somaiya College of Engineering. We thank Mrs. Grishma Sharma for her guidance and support.

8. REFERENCES

[1] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips: Face Recognition: A Literature Survey

[2] B.K. Low, E. Hjelmas: Face Detection: A Survey

[3] H.A. Rowley, S. Baluja, T. Kanade: Neural Network Based Face Detection

[4] Sinno Jialin Pan and Qiang Yang Fellow, IEEE. "A survey on Transfer Learning".

[5] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. "One shot learning of simple visual concepts".

[6] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra. "Matching Networks for One-Shot Learning".

[7] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, Timothy Lillicrap. "One-shot Learning with Memory-Augmented Neural Networks".

[8] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov. "Siamese Neural Networks for One-shot Image Recognition".

[9] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. Proceedings of the British Machine Vision 1(3), 6 (2015)

[10] Y. Sun, L. Ding, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks.

[11] Kilian Q. Weinberger, John Blitzer and Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification

[12]] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep Residual Learning for Image Recognition".

[13] Davis E. King. "Dlib-ml: A Machine Learning Toolkit".

[14] Liton Chandra Paul, Abdulla Al Sumam. "Face Recognition Using Principal Component Analysis Method".

[15] Vahid Kazemi, Josephine Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees".

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October 2007.