Smart City Surveillance System Using Facial Recognition

Yuvraj Choudhary
School of Computer Science and Engineering
Galgotias University
Greater Noida India
yuvraj78612305@gmail.com

Vinit Rai
School of Computer Science and Engineering
Galgotias University
Greater Noida India
vinitrai425@gmail.com

Yashvi Nagpal
School of Computer Science and Engineering
Galgotias University
Greater Noida India
yashvinagpal98@gmail.com

Abstract--

In the last few years, there has been some brilliant work in the field of facial recognition [1,2,3,4], it has been quite difficult to perform this task efficiently. In this paper, we are proposing a system where the distance of a face from a 3-d surface is a direct measurement of the facial similarities. Once we generate the 3-d space then it's easy to perform various tasks like clustering using feature vector's which is very similar or a basic idea in machine learning. Our method works on a CNN trained to provide optimized results and we tried to implement facial recognition by using only 128 bits pers head/face.

IJSER

Introduction--

In this paper, we tried to solve a big problem with the smart city surveillance that was facial recognition and we tried to verification (if it is the same person) and recognition (who is this person). We tried to learn using 3-d embedding for each image by using a CNN. So we can say that the faces of the same person have the same distance from the 3-d surface and different persons have different or can say that difference btw distances of the same person should be a small value and for different persons, it should be higher value.

Once we are done with the embeddings then the remaining tasks can be performed very easily. Face verification can be understood as thresholding distances between two faces and recognition can be understood as the KNN classification problem

Related Work--

- Similar to other recent works that employ deep networks [3, 4], our approach may be a purely data-driven method that learns its representation directly from the pixels of the face. Rather than using engineered features, we use a large dataset of labeled faces to attain the appropriate invariances to pose, illumination, and other variational conditions. In this paper, we explore two different deep network architectures that are recent won't too great success in
- the computer vision community. Both are deep convolutional networks [6, 5]. The first architecture is based on the Zeiler&Fergus [5] model which consists of multiple interleaved layers of convolutions, non-linear activations, local response normalizations, and max-pooling layers. We additionally add several 1×1×d convolution layers inspired by the work of [6]. The second architecture is based on the Inception model of Szegedy et al. which was recently used
- as the winning approach for ImageNet 2014 [6].
- These networks use mixed layers that run several different convolutional and pooling layers in parallel and concatenate their responses. we've found that these models can reduce the
- The number of parameters by up to 20 times and have the potential to reduce the number of FLOPS required for comparable performance. There is a huge corpus of face verification and recognition works. Reviewing it's out of the scope of this paper so we will only briefly discuss the foremost relevant recent work.
- The works of [6,3,4] all employ a posh system of multiple stages, that mixes the output of a deep convolutional network with PCA for dimensionality reduction and an SVM for classification.
- Zhenya et al. [5] employ a deep network to "warp" faces into a canonical frontal view then learn CNN that classifies each face as belonging to a known identity. For face verification, PCA on the network output in conjunction with an ensemble of SVMs is used.

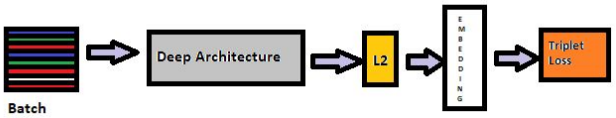
4. Problem Statement--

Now while we talk about smart cities one thing that is important is surveillance and that leads us to cameras and some fancy face recognition software but the problem is most facial recognition systems use labeling and thus they didn't know faces who are not in the database and ends with giving wrong information.

For example, Sam is using a face recognition system for his house main door and trained it for all 5 faces of his family member's but the problem is that when someone unknown came it would try to give the most probable results and will open the door and this may lead to there is no use of using this.

Solution --

Firstly we have to create a model to identify faces or facial recognition. We used a deep conventional network. Now the model can be understood as below figure:



In this model we have a batch input layer that goes through the deep CNN followed by an L2 normalization which gives face embeddings and that is followed by triplet loss during training.

Now what triplet analysis does is, it minimizes the difference between the anchor and positive and maximizes the distance between the anchor and negative.

Now Important thing in this approach is an end to end learning. Here triplet loss gives us or reflects something that we want during recognition. We try to embed a function f(x) on some image x in a space R, so that squared distance between all faces, ignoring the conditions of the environment while taking pictures, for the same person is small whereas for different faces it will give us a large value.

We believe that triplet loss is more suitable for face recognition. It tries to create a margin between each pair of images whether it is of the same person or one another. Let's take a look at triplet loss:

Triplet Loss --

It attempts to bring close the Anchor (current record) with the Positive (A record that is in principle comparable with the Anchor) beyond what many would consider possible from the Negative (A record that is not quite the same as the Anchor).

The real formula for this is:

$$Loss = \sum_{i=1}^{N} \left[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha \right]_+$$

So as long as the negative value is farther than the positive value + alpha there will be no addition for the calculation to gather the positive and the stay. Here is the thing that I mean:

$$\sum_{i=1}^{N} \left[(f_i^a - f_i^p)^2 - (f_i^a - f_i^n)^2 + N \right]$$

This will give us too many triplet values which are easily satisfied. Now as we know these values are already satisfied so this is gonna pass during testing and will slow the process of training and testing so we should select hard triplets to improve the model.

So let's take a look at triplet selection.

IJSER

Triplet Selection --

It is infeasible to register the argmin(hard negative) and argmax(hard positive) over the entire preparing set. Also, it may lead to helpless preparation, as mislabelled and inadequately imaged faces would rule the hard positives and negatives. There are two clear decisions that stay away from this issue:

• Generate triplets disconnected each n steps, utilizing the most ongoing system checkpoint and figuring the argmin furthermore, argmax on a subset of the information.

• Generate triplets on the web. This should be possible by choosing the hard certain/negative models from inside a smaller than usual group.

Here, we center around the online age and utilize huge smaller than usual groups in the request for a couple of thousand models and just register the argmin and argmax inside a little cluster.

To have a significant portrayal of the anchor positive separations, it should be guaranteed that an insignificant number of models of any one character is available in each smaller than normal clump. In our tests, we test the preparation information with the end goal that around 40 appearances are chosen for every personality per minibatch. Furthermore, haphazardly inspected negative countenances are added to every smaller than expected clump.

Rather than picking the hardest positive, we utilize all anchor positive sets in a small scale clump while choosing the hard negatives. We don't have a one next to the other examination of hard grapple positive sets versus all stay positive combines inside a scaled-down cluster, yet we found by and by that the all anchor positive strategy was more steady and met marginally quicker toward the start of preparing.

Deep Convolutional Network--

We trained CNN using g Stochastic Gradient Descent. Initially, we started training with a learning rate of 0.05. The margin α is set to 0.2. We utilized two sorts of models and investigated their compromises in more detail in the trial segment. Their down to earth contrasts lies in the distinction of boundaries and Flops. The best model might be distinctive relying upon the application. For example, a model running in a data center can have numerous boundaries and require an enormous number of Failures, while a model running on a cell phone needs to have boundaries, so it can fit into memory. All our models utilize corrected direct units as the non-straight enactment work.

The table below shows the working of the CNN model to make the model accessible :

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

Now, this model is used with other libraries like OpenCV and other imaging libraries to get input and output for the user.

Then a front-end system will communicate with this model and show a user-friendly interface using HTML, CSS, and Django.

.Performance evaluation:--

Our model can be evaluated as these two:

1. Fixed center crop to create thumbnails of images.
2. Then run the model on these thumbnails till the face is aligned.

We achieved a classification accuracy of almost 92% by fixed center crop and 94% of standard error mean, this reduces the error reported for deep face by more than a factor of 5.

All images used for training and testing were more than 10,000 per slot. We tried it for 7,8 rounds which is a decent amount of data and that gives us an accuracy of 92% which is good. We manage to reduce the error rate by 50% and up to 30% or less can be achieved by training it on large datasets..

Summary:--

We provide a method to directly learn an embedding into a 3d space for face verification with an accuracy of 92% and more and that would be pretty much useful for surveillance systems for cities and crowded places. Our end to end training both simplifies setup and shows that by directly optimizing and showing results.Future improvements will be done to reduce the system requirements and finding all the cases or errors and optimize it for stable uses.

References--

[1] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with gaussianface. CoRR, abs/1404.3840, 2014. 1

[2] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by join identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3

[3] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 8

[5] D. Yanga , Abeer Alsadoona , P.W.C. Prasad*a , A. K. Singhb , A. Elchouemic: An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment . 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

[6] Nicole Martinez-Martin, JD, PhD: What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care. AMA Journal of Ethics® February 2019, Volume 21, Number 2: E180-187

IJSER