DEPTH FACE RECOGNITION THROUGH DEEP LEARNING NETWORKS FINE-TUNING

Conference Paper - May 2018

1 author:

Yaser Saleh
Petra University
3 PUBLICATIONS 2 CITATIONS

All content following this page was uploaded by Yaser Saleh on 13 August 2018.

The user has requested enhancement of the downloaded file.
DEPNTH FACE RECOGNITION THROUGH DEEP LEARNING NETWORKS FINE-TUNING

YASER SALEH
Department of Software Engineering, University of Petra, Amman, Jordan
Email: Yaser.Saleh@uop.edu.jo

Abstract: Face recognition still has many challenges, and with the appearance of the Microsoft Kinect device, new potentials of research were discovered, trying to use the Kinect depth maps as a data source to recognize human faces was considered an interesting research area, mainly because the Kinect could provide the required data in an affordable and accurate way, but till this day no research managed to utilize the new and popular deep learning techniques to achieve higher accuracy and better results on face recognition using any deep learning technique and Kinect 2 depth map images. In this paper, with the goal of enhancing face recognition through the use of depth map images and the lack of a depth map datasets that can be used to train a deep learning network, we introduce a new practice for network fine-tuning, as the methodologies were applied with face depth maps to achieve high face recognition accuracy. The process of building a convolutional neural network and loading weights of a similar ImageNet trained network was introduced, where the network was fine-tuned and trained to work with depth map images of faces. The results of the new training procedure presented superior performance compared to any previous methods where depth maps were the source of data.

Keywords: Face recognition; Kinect 2; Deep Learning; Fine tuning, Convolutional neural network.

I. INTRODUCTION

Over the past years, one of the most common areas for research in computer vision an image processing was object recognition, more specifically face recognition. Taking that in consideration, when Microsoft released a new device with the name of The Microsoft Kinect [1], much interest was shown in utilizing this device for different research purposes, at first, the Kinect was created to be a gaming add-on for the Xbox console, but later on, research in the area of face recognition and computer vision overall showed interest in the device since it had many appealing features, one of the features was the depth sensor, and the ability to capture depth and color images at the same time.

With the recent developments in Convolutional Neural Networks (CNNs), improvements on the overall capabilities of multiple image processing areas was shown. Recently, the top leading network model for the ImageNet challenge was the Inception-v4 model architecture, which was able to achieve 3.08% top error on the ImageNet challenge through using 75 trainable layers [2], overcoming the ResNet and GoogleNet that were the previous champions in image classification [3]. However, one of the main downsides of latest state of the art is the increasing training run time, and the very large datasets required for training and testing.

The objective of this paper is to utilize face depth map images captured using the Kinect 2, which were not utilized alone with deep learning networks before, and to present how transfer learning can help in training a network with a small batch of depth images. This paper shall also explain the importance of depth maps and how fine tuning was used instead of normal training techniques.

This paper is divided as follows: Section 1 contains the introduction, Section 2 comments on some related work, Section 3 explains the Kinect device and CNNs used, Section 4 explains the procedure of the experiments, Section 5 discusses the results, and finally Section 6 presents the conclusion.

II. RELATED WORK

TheIn this section, some research that relates to this paper is presented and studied, going through face recognition with depth maps and with deep learning approaches.

Although face recognition using depth maps is not a very popular area where many other applications for depth maps can be worked on, some work can be found that utilizes depth maps for various recognition approaches, [4], [5] discusses the area of face recognition using depth maps.

The introduction of CNNs has gained substantial attention in the field of face recognition in recent years, and reviewing some research done on face recognition with deep learning [6], it can be found in [7], an 8-layered CNN called Deepface was trained, using a large face database that contained 4 million face images for 4,000 subjects, furthermore, the authors presented a 3D alignment method using affine camera model. Looking at another paper [8], a CNN-based face representation, called Deep hidden Identity feature (DeepID), was produced, where unlike the previous Deepface approach which had features learned by on single CNN, DeepID leaned by training an ensemble of multiple small CNNs, where the input would be the patches or crops of facial images, and then, a concatenation of the outputted
features would be produced as a powerful feature. DeepID CNNs 4 convolutional layers, 3 max pooling layers and 2 fully connected layers, presenting a large feature set of 19, 200 compared to the 4096 of the Deepface. Later on, an extension of DeepID called DeepID2 [9] was presented, where to train a CNN, both identification and verification information were used, compared to the DeepID where only the identification information was used.

The goal in the approach presented by DeepID2 was to “maximize the inter-class difference but minimize the intra-class variations” [6]. In [10] further improvements on the performance of DeepID and DeepID2 were presented under the name DeepID2+, where the supervision information was added to all convolution layers instead of the top layers as done in DeepID and DeepID2. Also, DeepID2+ enhanced the number of filters for every layer, and using a much larger training set. The most important properties of DeepID2+ were presented as being sparse, selective, and robust. The last update on the DeepID series was DeepID3 [11], were two deep neural network architectures were presented, which were built considerably deeper than the previous DeepID2+ architecture. DeepID3 networks were built from basic elements of the VGG net and GoogleNet.

Described in the training process how “joint face identification-verification supervisory signals” were added to the last feature extraction layer, as well as some intermediate layers of each network. Furthermore, weights in top layers of some of the networks were unshared in order to learn a richer pool of facial features. Since DeepID3 was trained on the same dataset as DeepID2+, it showed improvement on the face verification accuracy from 99.47% to 99.53%. Looking at another work in [12], the authors presented CASIA-WebFace, a large scale dataset containing around 500,000 images for 10,000 subjects. Furthermore, an 11-layer CNN was used on the database to learn discriminative representation and obtain state-of-the-art accuracy on LFW and YTF. Looking at the process used in [13], a network that was introduced after the appearance of the VGG Network was introduced, where the architecture of VGG-16 was used for training only face images, the network was trained on 2.6M images for 2622 different identities. The network, called VGG-Face, had produced results comparable to the state-of-the-art networks, while using fewer images.

Although a number of CNNs already are published and produce great facial recognition accuracy, it is still found that there are no networks that are built to work only with depth map images, or face depth maps in particular, which if presented, as far as our knowledge, can be the first time CNNs have been tested with face depth map images without the need of color images.

III. BACKGROUND

This section presents the technologies and techniques that are required to understand the work and concepts presented within the rest of this paper. At first, giving an introduction to the Microsoft Kinect 2, since it was the main capturing device for the data used later on, then, the networks utilized are explained.

3.1. Microsoft Kinect 2

when Microsoft released a new device with the name of The Microsoft Kinect [1], much interest was shown in utilizing this device for different research purposes, at first, the Kinect was created to be a gaming add-on for the Xbox console, but later on, research in the area of face recognition and computer vision overall showed interest in the device since it had many appealing features, one of the features was the depth sensor, and the ability to capture depth and color images at the same time, added to that, many capabilities where presented for the Kinect in some published work [14]. The Kinect mainly consists of two synchronized components, the hardware, and the software set up on top of the hardware. Looking at both versions of the Kinect, the hardware contains three main components, the RGB camera, the 3D depth sensors, and an array of microphones installed along the Kinect, these components can be seen in Fig. 1. These components will be able to provide RGB images, depth maps, and audio signals at the same time. With these hardware components comes the software giving the developers the ability to create applications for different uses. These software tools provide the ability to handle the RGB images, capture the motion of the human, do some face recognition, and even recognize voice commands [14]. The most interesting improvement on the Kinect was that the depth sensor was improved in the new Kinect 2 with a higher resolution of 512x424, much less noise, and higher ability to detect smaller objects clearly [1].

![Fig 1. Kinect 2 components](image)

Depth maps were able to provide a detailed pixel-by-pixel depth of any scene facing the Kinect, a sample of face depth images can be seen in Fig. 2, which made research around different aspects of computer vision to take notice of the sensor. One of the most
3.2. VGG16 and VGG-S

A convolutional neural network (CNN) works by allowing an image to pass through as input, then go through a set of layers containing convolutions, pooling and other fully connected layers, to finally provide an output of a single class of a set of possible classes for the image.

One of the already built and trained models that have been widely utilized in the past couple of years has been the VGGNet model architecture, which was the winner of the ImageNet Challenge in 2014. It was very beneficial to have fewer layers than the current state-of-the-art while still providing significantly good results with an error of 7.3% on the overall accuracy [17]. The VGGNet models are also published and widely distributed on the Internet for various deep learning frameworks, making it the perfect model to start working with, noting that convolutional layer parameters are denoted as “conv( receptive field size)-(number of channels)” the structure of the used (16 layers) model can be described as follows:

- input (224 × 224 RGB image)
- conv3-64; conv3-64
- maxpool
- conv3-128; conv3-128
- maxpool
- conv3-256; conv3-256; conv3-256
- maxpool
- conv3-512; conv3-512; conv3-512
- maxpool
- conv3-512; conv3-512; conv3-512
- maxpool
- Fully Connected (FC)-4096
- FC-4096
- FC-1000
- soft-max

Another VGG model called VGG-S is also utilized, where the architecture can be explained as follows:

- input (224 × 224 RGB image)
- conv7-96
- maxpool
- conv5-256
- maxpool
- conv3-512; conv3-512; conv3-512
- maxpool
- Fully Connected (FC)-4096
- Dropout Layer
- FC-4096
- Dropout Layer
- FC-1000
- soft-max

IV. PROPOSED PROCEDURE

4.1. Data Collection

Since the Microsoft Kinect 2 is still considered a new device, although it is possible to find multiple face databases online, most if not all of them was recorded using the older Kinect 1, also, in this research, we aimed at showing the capabilities of depth maps in poor illumination, for those reason, we created a custom face dataset that would contain the depth maps for 10 subjects in good and poor illumination.

For the process of data capturing, a specific scene was created, the scene was built to extract the images in good and poor illumination, and to capture the faces in different poses but at the same time have approximate directions for each pose, the approach used in [15] was applied as a technique to minimize errors in the process of capturing the data.

Following the recording of all the data, the data can be explained such as for each person a good and poor illumination conditions where applied, with nine different face direction for both illuminations, and finally two different expressions for each of the nine poses, this resulted in 360 depth map images. Finally, the cropped face areas where passed through an algorithm to create a grayscale image, with no addition enhancement on the data.

4.2. Fine-Tuning

As described previously, two VGG models were chosen for the fine-tuning process, were matching models was created, then, the pre-trained model weights were loaded into the layers of each network, and finally, our own soft-max classification layer was added as the final layer after the last fully connected layer, this process prepared the networks for fine-tuning which is implemented by running further training iterations.

Following the construction of the networks, the face depth map images dataset was used to further train the network, where 80% of the images were used for training and the remaining 20% for testing.
V. EXPERIMENTS RESULTS

After running the 20 epochs on the VGG-16 network, the highest validation accuracy reached was 98.61%, and it was reached after the fourth epoch. While considering the VGG-S model, after running 20 epochs, 100% accuracy was reached after the 19th epoch. Examining the results of the training and testing procedures, it can be seen that depth map images where suitable to use for fine-tuning the networks and very good accuracy was produced when trying to recognize faces from depth maps of the subjects. Although more examination with larger datasets could produce a slight different accuracy, the fact that there is a lack of large face depth datasets recorded with the Kinect 2 shows great potential to the presented results. Also, looking at previous research done on face recognition with deep learning, it can be seen that multiple networks produce very high accuracy when using only color images, nevertheless, depth maps ability to work in poorly illuminated scenes can give a major improvement to the face recognition field if applied properly.

CONCLUSIONS

In this paper we have presented a novel technique that will help in evolving the research on the use of depth maps with deep learning methodologies. Our first goal was to show the capabilities of deep learning techniques to work with depth map images that were captured with an affordable device. However, the lack of depth map datasets, captured with the Kinect 2 that can be used to train a new deep learning network provided a challenge, to try and solve this problem a dataset was created and published in order to have some depth data for further use. Since to train a CNN requires a large amount of data utilizing transfer learning, more specifically fine-tuning, was the only option as a solution. The VGG-16, and VGG-S networks were constructed and fine-tuned with the dataset already created, and accuracy results reached 98% and 100%.

The next aim was to show the capabilities of the Kinect 2 depth maps captured in poorly illuminated scenes, where the data was mixed before the training and testing procedures to see if the poorly illuminated images would affect the accuracy, and results shown the great potential of the use of depth maps.

To sum up the approach used not only had shown the great potential of using depth maps for face recognition, but also introduced the Kinect 2 data to the deep learning world. However, it is important for depth maps to gain more attention, as with more interest more public dataset can be published and could provide enough information for further investigation on the depth maps and their uses with deep learning techniques.

REFERENCES