Age and Gender Detection with Crowd Counting for Visually Impaired using Convolution Neural Networks

R N SNEHA PRIYA[1], POOLA CHARAN[2]

Vellore Institute of Technology

ABSTRACT:

Visual Impairment is caused by many reasons, maybe due to diseases or due to age. Vision is the basis of all the navigation tasks, so the visually impaired people have a major disadvantage in having their living as most of the required information in the surrounding is available through vision. With the recent advancements in technology, it is possible to extend a helping hand to the visually impaired people.

The project, Visual Aid for Blind focuses on serving the same. This project serves to provide complete assistance on key aspect Age and Gender Detection along with crowd counting.

KEYWORDS:

Visually Impaired, Convolutional Neural Networks, Age recognition, Gender recognition, Crowd counting, MP3 Output.

1. INTRODUCTION:

Blindness is a condition wherein a person lacks visual senses due to neurological or physiological factors. Age, gender, and crowd count plays a major role in the social interactions. Different languages use different types of salutations and pronouns for different genders or collective crowds, and most probably a wide variety of vocabularies are used to address the elders in a conversation when compared in addressing young people. These basic yet important attributes play a major role in our everyday communication or conversation with others. Providing the visually impaired information about age and gender, will help them communicate better with people.

Existing methods for estimating these features from images depend on changes in the facial features or predefined face identifications. Many of them have used classification methods are specifically designed only for the age detection or gender detections tasks exclusively. Some of them were modelled to handle the challenges which are associated with the restriction less imaging situation. In addition to that, the type of machine learning methods used by these methods don't fully use the huge datasets which consists of tons of image examples which is available in the Internet in order to increase the accuracy and to improve the capabilities of classification.

Our project uses ConvNet also known as CNN or Convolution Neural Networks to detect age, gender and the count of people standing before the blind person. This project captures frames of a live streaming video, sends it to Neural networks for identification of age and gender of the person standing in front of them and converts the output obtained to vocal format and voices it out so that the visually impaired know who the person is standing before him or her.

2. RELATED WORKS:

Before explaining our project, we have reviewed related methods for age and gender detection using different methods.

A. Age Detection:

The main idea behind the interpretation of the age of a person from photographs has started to spread widely and many ideas have been put forth in recent times. A detailed survey related to those methods is mentioned in [1]. Despite of the focus in our project to classify into age groups rather than estimating the exact age (regression), the surveys given below explains both the purposes.

In earlier times, age estimation was dependent on calculating ratios of facial features measurements [2]. Once the facial features like eyes, mouth, nose are extracted, their distance and sizes are measured, and ratios are calculated between them, and they were used to classify them into different age groups by some custom-made rules. [3] used the same kind of approach to detect faces of people under 18 years. As these methods need to detect the facial features accurately for age detection, which is more challenging by itself, it is not suitable to detect ages in an instant which must be the most important feature in our project.

B. Gender Detection:

A very detailed review on gender detection was explained in [4]. In earlier methods, neural networks trained with a smaller data set of front face images were used for gender classification [5] used Webbers Local texture Descriptor which showed perfect performance of FERET benchmark. The intensity, size and texture were used with common information to get almost accurate results in FERET benchmark by [6].

FERET images are usually taken in a controlled environment hence it is easier to detect gender using these when compared to wilder look face images. In addition, the outputs which are obtained in the mentioned method are old and are not adequate for newer methods. Hence it is very difficult to find out the actual use and accuracy of the above-mentioned techniques.

3. SYSTEM DESIGN & ARCHITECTURE:

The system design and architecture which has been proposed below is used all throughout our project which is age, gender detection with crowd counting using CNN. The network consists of two layers which are fully connected and has a small number of neurons and three Convolution Neural Networks layers. The reason for choosing a very small network is that it reduces a lot of risk of overfitting which in turn reduces the accuracy.

All the three channels of colors are being processed directly by the network. First the images which are given as input are reduced their resolution by scaling them to 256 x 256 and the scaled images are then cropped to the size of 225 x 225 for further processing in the three subsequent convolution layers.

The three subsequent convolution layers are:

1. There are 96 number of filters which are of size of 3 x 7 x 7 number of pixels which are then applied and processed to the given input in the first layer of convolution neural network, which is then followed by the RLU also known as rectified linear unit and a max pooling layer that takes the maximum of 3 x 3 which is with a stride of 2 pixels along with a normalization layer which is local for the response.
2. The output from the previous layer i.e., the first layer is of size 96 x 28 x 28 number of pixels which is then processed by the second layer in the architecture which is CNN which contains 256 number of filters which is each of size 96 x 5 x 5 number of pixels. And then, this is again followed by a RLU which

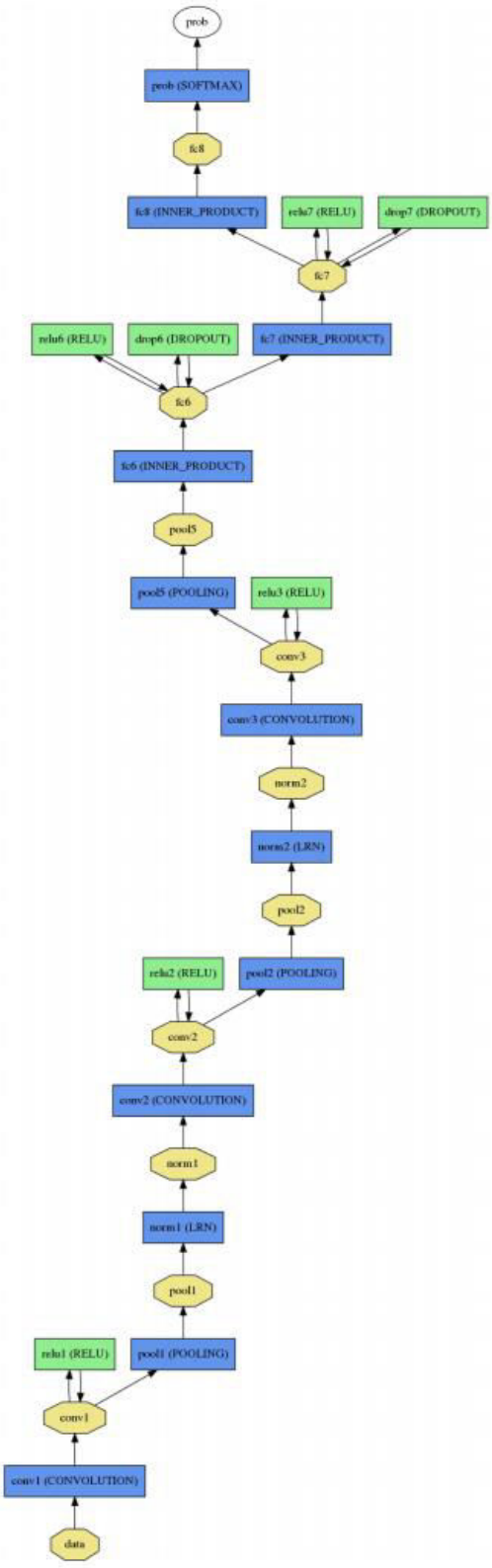
is called as rectified linear unit and a maximum pooling layer that takes the maximal value of 3 x 3 with a stride of 2 pixels and a normalization layer which is local for the response, i.e., with the same hyper parameters similar to the previous layer.

3. The last convolution layer which is the third one has a blob of 256 x 14 x 14 which operates with the help of a set of 384 number of filters which are each of size 256 x 3 x 3, which is in turn followed by RLUs or rectified linear unit and then followed by a maximum pooling layer.

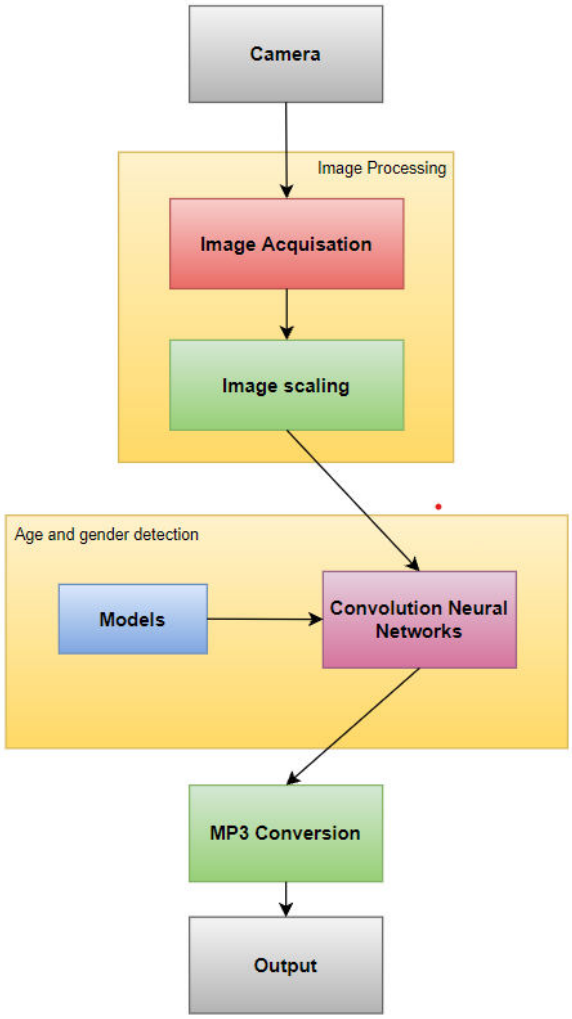
Definition of Fully connected layers:

1. The initial fully connected layer is the one that takes the output which is given by the third CNN layer. This layer contains 512 number of neurons which is in turn followed again by a RLU and which is then followed by a drop out layer.
2. The 2nd fully connected layer is the one which receives the input from the 1st fully connected layer. This layer also has 512 number of neurons which is then followed by a RLU which is then in turn followed by a drop out layer.
3. A 3rd fully connected layer is the one which receives from the 2nd fully connected layer which is then mapped to the classes for the recognition of age and gender of the person.
4. Counting is implemented during the input taken into the Convolution layer.
5. Finally, the output is then connected to the last and final layer of fully connected layer which is fed by a input to the SoftMax layer which assigns the probability of each class and then finally the class with highest probability is selected as the output class given as a recognized age and gender. Then the output will be read aloud so that the blind person can hear the number of persons before them, their age, gender.

System Architecture diagram:



System Design:



Initially, from the camera the video is captured and then converted into frames. Once the we get a frame from the live video stream, image acquisation process is done. Then image scaling is done to the required dimensions and then finnaly fed as the input to the three layer convNet layers. There the age and gender are classified with the help of the age-caffe model and gender-caffe model.
Once the age, gender of the person are detected it is then converted into MP3 and finally read aloud to let the blind person know.

For each person, this process happens, and it also counts the number of persons who are standing by counting the number of faces which undergo the above process.
Finally, the output is read aloud to let the blind person know.

4. EXPERIMENTS CONDUCTED:

Our project is implemented using in Sypder IDE. We implemented it in a PC which has GPU Intel's UHD 620 graphics, with CPU of Intel I5 10^{th} gen processor and 8 GB quad core processor.

5. RESULTS:

Evidently, the methodology we have used have given good results. It outperforms both tasks simultaneously with a very small considerable gap and give the output in the form of MP3. Also it is clear that the contribution made by the over-sampling approach which increases the performance and also enhances by boosting the total network. This means it gives higher accuracy.

Since the audio output can't be placed in the results section, we have just derived a confusion matrix. The results showed that there was a very error rate.

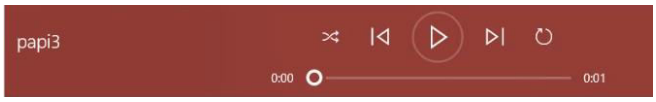
The result is:

Confusion matrix of the output:

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-
0-2	**0.699**	0.147	0.028	0.006	0.005	0.008	0.007	0.009
4-6	0.256	**0.573**	0.166	0.023	0.010	0.011	0.010	0.005
8-13	0.027	0.223	**0.552**	0.150	0.091	0.068	0.055	0.061
15-20	0.003	0.019	0.081	**0.239**	0.106	0.055	0.049	0.028
25-32	0.006	0.029	0.138	0.510	**0.613**	0.461	0.260	0.108
38-43	0.004	0.007	0.023	0.058	0.149	**0.293**	0.339	0.268
48-53	0.002	0.001	0.004	0.007	0.017	0.055	**0.146**	0.165
60-	0.001	0.001	0.008	0.007	0.009	0.050	0.134	**0.357**

Output in the IDE:

Audio Output: The audio with the age and gender and count of the person are heard aloud and this is the screenshot of Windows Music app opened to play the audio file with all the details like age, gender and count.



6. CONCLUSION:

Though there many previous works related to this age and gender detection using many models like R-NET, and many more, they had the problem of images that are taken in the lab settings and those can't adequently predict the images in the real-world conditions. But this project using CNN helped a lot in understand the features of the face and accurately predicting the age and gender in real world situations too.

The easy of availability of the very huge data set collections which are present on the internet are providing the newest machine learning systems and models with highly effectively large never-ending data for training. This data might not be labelled properly according to the needs. Here in this project, we have shown how well CNN can recognize the age and gender recognition more accurately than any other model. The network is very much shallow when compared to some other network architectures, but this always it to avoid the over fitting problem by reducing the number of parameters.

The main conclusion from the project is that CNN is the only model which gives better results of age and gender detection with crowd counting.

7. FUTURE WORK:

We will work for providing complete assistance for the blind people by voicing out the things they come across, like helps them to navigate across places, tells them who the person is standing in front of them, reads out books or wordings they come across, etc. by giving them a device like wearable glasses which has embedded camera, all the required components for the above said features.

8. REFERENCES:

[1] Fu, Y., Guo, G. and Huang, T.S., 2010. Age synthesis and estimation via faces: A survey. IEEE transactions on pattern analysis and machine intelligence, 32(11), pp.1955-1976.

[2] Kwon, Y.H., 1994. da Vitoria Lobo,". In Age classification from facial images," 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA (pp. 762-767).

[3] Ramanathan, N. and Chellappa, R., 2006, June. Modeling age progression in young faces. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 1, pp. 387-394). IEEE.

[4] Reid, D.A., Samangooei, S., Chen, C., Nixon, M.S. and Ross, A., 2013. Soft biometrics for surveillance: an overview. Handbook of statistics, 31, pp.327-352.

[5] Ullah, I., Hussain, M., Muhammad, G., Aboalsamh, H., Bebis, G. and Mirza, A.M., 2012, April. Gender recognition from face images with local wld descriptor. In 2012 19th international conference on systems, signals and image processing (IWSSIP) (pp. 417-420). IEEE.

[6] C. Perez, J. Tapia, P. Estevez, and C. Held. Gender classification from face images using mutual information and feature fusion. International Journal of Optomechatronics, 6(1):92–119, 2012.

[7] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T., 2014, November. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678).