IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Effect of Image Distortion on Facial Age and Gender Classification Performance of Convolutional Neural Networks

To cite this article: Choon-Boon Ng and Wei-Haw Lo 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **495** 012029

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 124.207.82.166 on 18/09/2019 at 17:28

Effect of Image Distortion on Facial Age and Gender Classification Performance of Convolutional Neural Networks

Choon-Boon Ng, Wei-Haw Lo

Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman,
43000 Kajang, Selangor, Malaysia.

ngcb@utar.edu.my

Abstract. Significant improvement in the task of age group categorization and gender classification of facial images has been achieved using deep convolution neural networks (CNN). In this paper, we study the effect of image distortions such as blur, noise, rotation and occlusion on the performance of a state-of-the-art CNN. We found that the CNN was more sensitive to noise compared to blurring, especially for age estimation. By studying occlusion, we also identified the salient regions of the face. An interesting result is that the upper half of the face is more important for age estimation, while for gender classification it is the lower half. These insights should prove useful for future development of CNN models for facial age and gender classification.

1. Introduction

The application of machine learning on unstructured data such as images has become increasingly prominent in recent years with the exponential increase of computing power and data. This has attracted interest for many futuristic applications to better serve society, such as surveillance, targeted advertising and human-computer interaction, whether for security or commercial purposes. Soft biometric traits such as the age, gender and ethnicity of a person can play an important role in such systems. Yet, employing computer vision to estimate these attributes remains a challenging problem in practice due to the unconstrained nature of real-life situations.

In recent years, deep learning, in particular convolutional neural network (CNN), has achieved astounding success in problems related to text, speech and visual understanding. The concept, generally speaking, is to train the network to learn from available data the important features to solve the problem. This is in contrast to relying on hand-crafting features based on domain understanding of a problem.

In this study, we are interested in facial age and gender classification. The CNN has also been applied to this problem in recent years with good results. Levi & Hassner [1] trained a simple CNN for facial age and gender estimation of unconstrained images that outperformed state-of-the-art hand-crafted approaches. Sometimes referred to as GilNet, their CNN is relatively shallow by today's standards but still serves as a useful baseline.

While CNNs have achieved improved performance metrics on datasets, their reliability in deployment remains an interesting question. Dodge & Karam [2] showed that for state-of-the-art deep CNNs, their accuracy in object recognition tasks were affected by image distortions, especially noise and blur. Rodner et al. [3] studied the sensitivity of CNNs in fine grain recognition, showing dramatic

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

decrease of performance with just a small amount of noise. Karahan et al. [4] studied how it affected face recognition, in particular, it is most affected by noise, blur and occlusion. These results are enough to raise concern, since, in practice, the cameras used may be of low quality.

So we asked the question, in facial age and gender classification, how will a state-of-the-art CNN be affected by various image distortions such as blur, noise, rotation and occlusion? The results of the experiments are our contribution. From these results, we also gain some insights into the problem which we hope will serve as useful guidance for building better models in the future.

2. Methodology

2.1. Convolutional neural network

In our experiments, we used GilNet, the CNN model for age categorization and gender classification proposed by Levi & Hassner [1], which is now often used as a baseline for deep learning methods in this problem. The model achieved state-of-the-art results for these tasks, outperforming other type of classifiers at that time. Although since then the performance has been surpassed, the main structure of these improved models remains to be a CNN, with the difference mostly being the input to the network (e.g. combining with Gabor filter responses [5]), or in terms of the network depth and learning (e.g. transfer learning [6], residual learning [7]). GilNet model is a plain vanilla CNN, containing most of the typical elements such as max pooling, rectified linear units (ReLU), dropout and local response normalization. Furthermore, their experiment was easily reproducible as the source code was provided on their project webpage.

A description of the GilNet architecture is given in table 1. The input to the network is a 3-channel RGB image of the size 227x227 pixels. This is followed by three stages of alternating convolution and max pooling operation. Next, there are two fully connected (FC) layers that are 512 units in width. ReLU is used as non-linearity after each convolution filtering and fully connected layer. Local response normalization is applied after the first two max pooling layer. Dropout with ratio 0.5 is applied in the FC layers. The difference in the age and gender network structure is only in the softmax regression layer, where the number of units depends on the number of categories.

Table 1. CNN architecture of GilNet [1].

Layer	Description
Input	3x227x227
Layer 1	Conv. with 96x7x7 filters
Layer 2	Max pooling 3x3 stride 2
Layer 3	Conv. with 256x5x5 filters
Layer 4	Max pooling 3x3 stride 2
Layer 5	Conv. with 384x3x3 filters
Layer 6	Max pooling 3x3 stride 2
Layer 7	FC 512 units
Layer 8	FC 512 units
Layer 9	Softmax 2 or 8 units

The dataset used to train and evaluate the CNN was the Adience aligned faces dataset [8]. The dataset contains images of faces in unconstrained situations with a large variety in pose, lighting etc. They have been labeled with eight different age group categories: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and 60 or above. The gender label represents male or female. There are a total of 19538 images of about 2000 plus subjects. These images all have a size of 816x816 pixels.

We reproduced their results by replicating the training and testing protocol exactly as described in their paper. Specifically, the images were divided into training, validation and test sets for five-fold cross-validation, using the same distribution of images. Each image was resized to 256x256 pixels, then cropped to 227x227 pixels for input to the network. Parameters for network initialization, training and testing were also followed exactly. During testing, both single crop and oversample (which averages the prediction from multiple crops) methods were considered. For age classification, results for exact prediction (classification in the correct age category) and one-off prediction (classification off by only one age-category is also considered correct) were found. The results are shown in table 2, which is close to the published results. The slight difference is to be expected since the CNNs were trained from scratch with random weights initialization.

Table 2. Reproduced results for test set accuracy on original images.

Classification task	Accuracy ± standard deviation (%)
Age, exact (single crop)	49.3 ± 4.2
Age, one-off (single crop)	85.3 ± 1.9
Age, exact (oversample)	50.8 ± 5.5
Age, one-off (oversample)	84.9 ± 2.4
Gender (single crop)	85.7 ± 1.2
Gender (oversample)	86.6 ± 1.5

2.2. Applying image distortion

To study the effect of image distortions on the performance of the trained CNN, we next applied distortions to the test images at their original size (before downsizing for the network input). The accuracy of the trained CNN on the age and gender estimation tasks were then determined.

The image distortions were applied using the multi-dimensional image processing package in SciPy, which is an open source Python library for scientific computing [9]. In particular, we experimented with Gaussian blur, Gaussian noise, salt and pepper noise, rotation, and occlusion. For Gaussian blur, a Gaussian filter was applied, varying the standard deviation of the kernel ranging from 0 to 100. For Gaussian noise, random values from a normal distribution were added to each color channel of the image independently, with a mean of 0 and standard deviation ranging from 0 to 100. Since each pixel value is represented by an integer between 0 and 255, the values were clipped to remain in this range. For salt and pepper noise, integer 0 represented pepper noise while integer 255 represented salt noise. The pixel values were replaced by the noise values in each color channel independently by equal amounts with a probability ranging from 0 to 50%. For rotation, the images were rotated in-plane for angles ranging between -40º to +40º.

Various occlusions to the faces, in the form of a black mask, were applied. Figure 1 shows examples of these occlusions. In terms of the facial parts, we experimented with the eyes, periocular region (area including the eye and eyebrows), nose, and mouth to determine which parts affect the accuracy of the classifier. We also experimented with a more general division into the upper and lower half regions of the head, using the tip of the nose to define the horizontal separation line.

To detect these parts in an image, we used the Dlib toolkit [10], in particular, its frontal face detector and implementation of the facial landmark detector of [11]. The facial landmarks were detected and used as the basis to define the facial parts. About 1200 faces were not detected, so we had to manually label the landmarks for these.

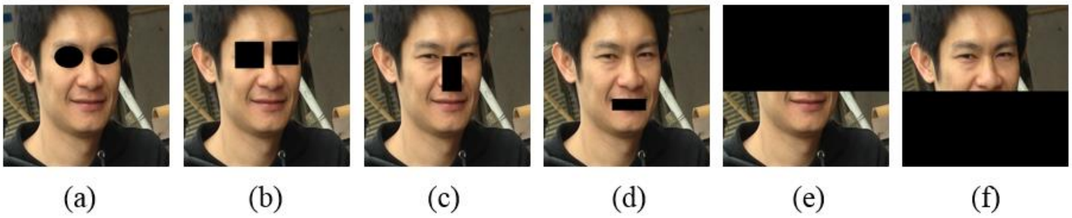

Figure 1. Example of occluded image (a) eyes (b) periocular (c) nose (d) mouth (e) upper half (f) lower half

3. Results

The accuracy performance of the trained CNN on images distorted at various levels is shown in figure 2 (age classification) and figure 3 (gender classification).

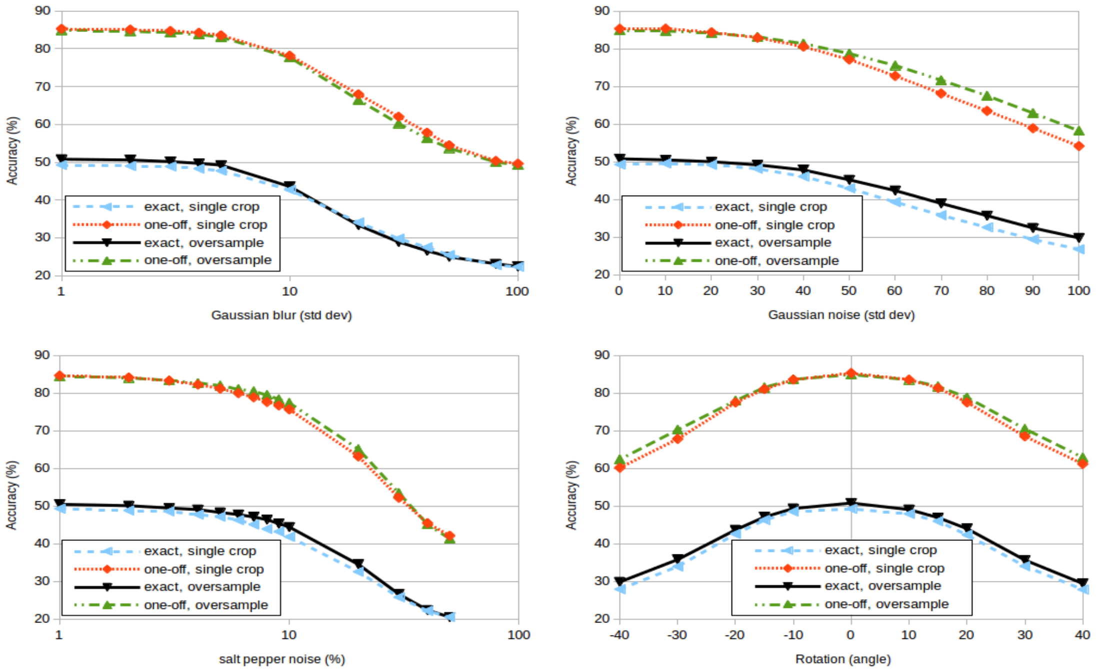

Figure 2. Results for age classification on distorted images.

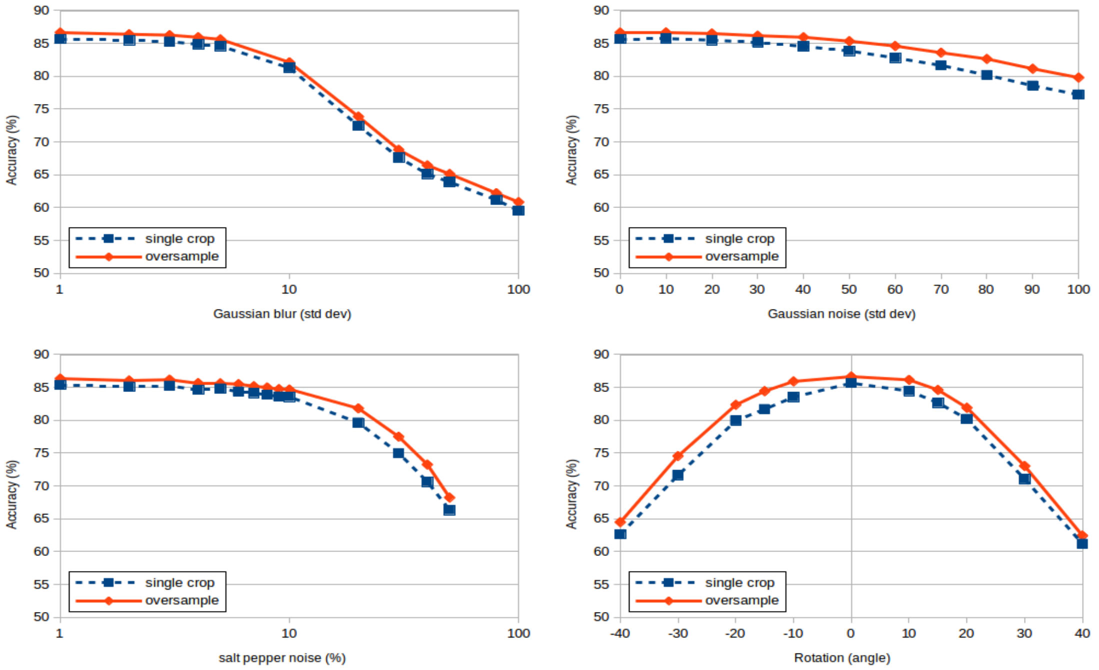

Figure 3. Results for gender classification on distorted images.

As expected, with increasing blur or noise, the accuracy decreases. At low levels, the accuracy drop is still somewhat muted, so it can be said that the CNN has some resistance to small amounts of image distortion. But with higher levels, the deterioration increases. Generally, for blur, accuracy starts to drop rapidly after standard deviation of 10 while for salt and pepper noise, it is at around 10% probability. For Gaussian noise, it is around standard deviation of 50 for age classification, but for gender classification, the drop is less steep.

For rotated faces, the graph is almost symmetrical, meaning rotation clockwise or counter-clockwise has similar effect. For human vision, we know generally that it is not a problem to estimate the age and gender for any amount of rotation, i.e. it is rotation invariant. However, the CNN has only a small degree of rotation invariance.

3.1. Age classification

Let's consider the task of age classification. Take the case when the exact accuracy for single crop test is approximately 30%. At this accuracy, the blur is around standard deviation 30, the Gaussian noise is standard deviation 90 and the salt pepper noise is around 20% probability. Figure 4 shows examples of images with these levels of distortion.

From our observation, we (human vision) see it is easier to estimate the age of the person in the image with noise applied and more difficult for blur. Blurring has removed details and texture information such as wrinkles and creases which give us the natural clue as to the age of the person. Human vision will have more confidence estimating the age of the subject in the noisy image compared to the blurred image. However, the CNN performs with similar accuracy at this level of distortion. From this, it implies that the CNN is more affected by noise than blur.

The possible reason could be that the convolution and max pooling process is sensitive to sudden changes in pixel values introduced by noise. For blurring, pixels values tend to remain similar to its neighbors because of the averaging filter. Neural networks have been known to do well with object recognition from low resolution images (e.g. 32x32 pixels image in [12]), and blurring is akin to reducing the image resolution.

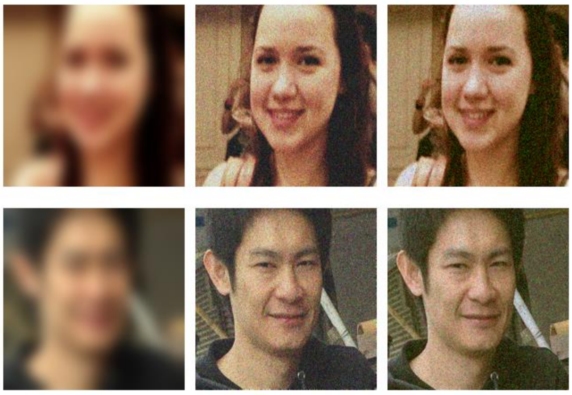

Figure 4. Example of images with blur (left), Gaussian noise (centre) and salt pepper noise (right) at similar levels of age classification accuracy.

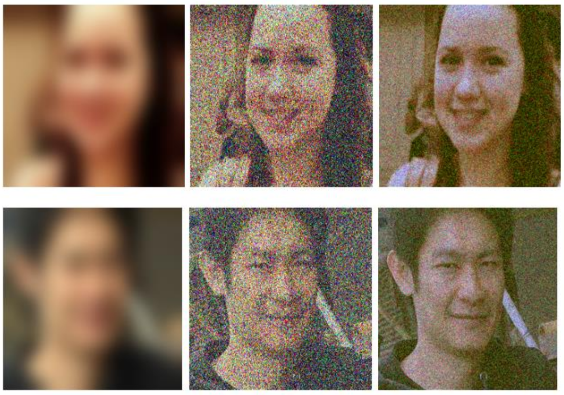

Figure 5. Example of images with blur (left), Gaussian noise (centre) and salt pepper noise (right) at similar levels of gender classification accuracy.

3.2. Gender classification

Next, we consider the task of gender classification. Take the case when the accuracy for single crop test is approximately 65% (which is again an approximately 20% drop from the original). At this accuracy, the blur is around standard deviation 40, and salt pepper noise is around 50% probability. For Gaussian noise, even at standard deviation 100, the accuracy is still above 75%. By extrapolation, we estimate the standard deviation to be 180 for accuracy to drop to 65%. Example of images at these levels of distortion is shown in Figure 5.

Again, human vision can confidently estimate the gender of the person in the image with noise applied and more difficult for blur. Details have been lost in the blurred image, although we can still make out the hairstyle or length, which provides valuable clues to the gender. So again, as implied previously, the CNN is more sensitive to noise than blur.

However, it seems to be less sensitive to noise in gender classification than age classification, since much higher levels of distortion is required for an equal reduction in accuracy. The reason could be that gender classification relies more on coarser features of the head than on the finer details.

3.3. Effect of occlusion

The results for the effect of occlusion on the facial parts and regions are shown in table 3.

Table 3. Effect of various types of occlusion on classification accuracy.

Classification	**Accuracy ± standard deviation (%)**						
task	***None***	***Eyes***	***Periocular***	***Nose***	***Mouth***	***Upper***	***Lower***
Age, exact (single crop)	49.3±4.2	35.9±6.0	33.8±5.3	40.5±3.6	44.5±6.1	25.6±4.0	36.1±4.2
Age, exact (oversample)	50.8±5.5	38.4±6.5	36.7±5.5	43.2±6.1	47.0±6.1	24.8±4.5	35.3±4.3
Age, one-off (single crop)	85.3±1.9	69.5±4.8	68.1±4.5	77.6±1.4	79.3±2.3	54.1±7.1	74.5±5.4
Age, one- off (oversample)	84.9±2.4	71.9±3.6	71.1±3.5	78.8±2.1	81.0±2.1	52.9±7.4	72.0±5.4
Gender (single crop)	85.7±1.2	79.7±0.8	78.2±3.3	83.5±2.5	81.9±1.8	70.4±3.8	65.7±6.1
Gender (oversample)	86.6±1.5	82.3±2.6	81.1±3.9	84.7±2.2	83.3±1.9	70.9±4.7	65.1±5.9

For age classification, the most important facial part is the periocular region (consisting of the eyes and the eyebrows), because occluding this region causes the highest drop in accuracy. This is followed by the eyes, nose and mouth. Masking the upper half of the head causes the accuracy to drop to around 25%, about 10% lower compared to masking the lower half. Masking the periocular region has almost the same effect on accuracy as masking the lower half of the head. These results are hence complementary, implying that the upper half has more discriminative information for estimating age compared to the lower half, particularly the periocular region.

For gender classification, the important facial part is also the periocular region, followed by the eyes, mouth and nose. However the drop in accuracy is not as drastic, which is by less than only 7%, compared to around 15% for age classification. On the other hand, masking the lower half of the head causes the accuracy to drop more than the upper half. Thus, the network may have learnt gender information contained in the lower half such as the hair length or facial hair.

4. Conclusion

In this paper, we have investigated the effect of various image distortions on the performance of a state-of-the-art CNN for facial age and gender classification. Based on our experiment results, noise in an image has more impact than blur on the accuracy of the CNN, especially for age estimation. When occluded, the periocular region of the face causes more loss in accuracy compared to the eyes, nose or mouth. For more general occlusion, the upper half of the face has more impact than the lower half in age classification, while for gender classification it is the opposite. We hope the insights gained will be a useful guide for deploying and improving CNNs in these tasks.

References

[1] Levi G and Hassner T 2015 Age and gender classification using convolutional neural networks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops* (IEEE) pp 34-42

[2] Dodge S and Karam L 2016 Understanding how image quality affects deep neural networks *2016 Eighth Int. Conf. on Quality of Multimedia Experience* (IEEE) pp 1-6

[3] Rodner E, Simon M, Fisher R B and Denzler J 2016 Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches *Preprint* arXiv:1610.06756

[4] Karahan S, Yildirum M K, Kirtac K, Rende F S, Butun G and Ekenel H K 2016 How image degradations affect deep cnn-based face recognition? *2016 Int. Conf. of the Biometrics Special Interest Group* (IEEE) pp 1-5

[5] Hosseini S, Lee S H, Kwon H J, Koo H I and Cho N I 2018 Age and gender classification using wide convolutional neural network and Gabor filter *2018 Int. Workshop on Advanced Image Technology* (IEEE) pp 1-3

[6] Ozbulak G, Aytar Y and Ekenel H K 2016 How transferable are CNN-based features for age and gender classification? *2016 Int. Conf. of the Biometrics Special Interest Group* (IEEE) pp 1-6

[7] Lee S H, Hosseini S, Kwon H J, Moon J, Koo H I and Cho N I 2018 Age and gender estimation using deep residual learning network *2018 Int. Workshop on Advanced Image Technology* (IEEE) pp 1-3

[8] Eidinger E, Enbar R and Hassner T 2014 Age and gender estimation of unfiltered faces *IEEE Trans. on Information Forensics and Security* **9(12)** pp 2170-2179

[9] Jones E, Oliphant T, Peterson P 2001 SciPy: Open Source Scientific Tools for Python *Online* http://www.scipy.org/

[10] King D E 2009 Dlib-ml: A machine learning toolkit *J. Machine Learning Research* **10(Jul)** pp 1755-1758

[11] Kazemi V and Sullivan J 2014 One millisecond face alignment with an ensemble of regression trees *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE) pp 1867-1874

[12] Lin M, Chen Q and Yan S 2013 Network in network *Preprint* arXiv:1312.4400