On the Robustness of Deep Learning Based Face Recognition

(removed for review)

ABSTRACT

Identifying persons using face recognition is an important task in applications such as media production, archiving and monitoring. Like other tasks, also face recognition pipelines have recently shifted to Deep Convolutional Neural Network (DNNs) based approaches. While they show impressive performance on standard benchmark datasets, the same performance is not always reached on real data from these applications. In this paper we address robustness issues in a face detection and recognition pipeline. First, we analyse the impact of image degradations (in particular compression) on face detection, and how conceal them in order to improve face detection performance. This is studied both on face samples originating from still image and video data. Second, we propose approaches to improve the classification of faces into "known" and "unknown", in particular to reduce false positives recognitions. We provide experimental results on image and video data and provide conclusions that help to improve the performance in practical applications.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Applied computing** → Digital libraries and archives.

KEYWORDS

robustness, face detection, face recognition, open-set recognition, compression

ACM Reference Format:
(removed for review). 2018. On the Robustness of Deep Learning Based Face Recognition. In *tbd*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The identifiable persons appearing in audiovisual content are one of the most important cues for content understanding and description. In many tasks in media production, media archiving and media monitoring tagging the appearance of known persons via recognition of their faces is needed for describing content, determining its relevance or indexing it for search, to name just a few use cases. Depending on the application, the set of persons of interest may vary, but it is usual a small set compared to the number of faces appearing in video content. This poses face practical face recognition approaches as an open-set recognition problem, i.e., it is a classification into one from a set of known persons or "unknown", with the latter having a high prior probability.

Many tasks for video understanding have shifted from traditional approaches to ones that rely on Deep Convolutional Neural Networks (DNNs) in recent years. This is also true for a face recognition pipeline, which usually consists of a face detection and a face classification step. For example, for face detection it has been shown that multi-task CNNs [**?**] achieve very good performance, including detection of partly occluded faces. In order to enable fast and efficient training of new faces, there are several recognition pipelines that use DNNs as a feature extractor and then use a classifier that can be trained with a moderate number of examples per person, such as support vector machines (e.g., [**?**]) or online random forests (e.g., [**?**]). These are approaches became sufficiently mature for practical use in applications in the media industry and beyond. However, the performance achieved on standard benchmark datasets is not always reached on real data from media production and archiving. There are number of issues related to robustness of the deep learning approaches that need to be considered.

A general issue of neural networks is that the training data has certain characteristics in terms of image quality (noise level, compression artifacts, etc.), and the network may learn some of these characteristics even if they are irrelevant to the task. For practical applications the robustness against variations in these parameters is crucial. The problem is also related to that of using so-call *adversarial samples*, i.e., samples that add noise not visible to humans to the image in order to cause misclassifications by the neural network, exploiting the patters it has learned. In practical applications, some level of compression artifacts is always present, and also other types of impairments might occur.

Another practical problem rarely discussed in the literature is the robust distinction of faces never presented to the classifier from those, already learned [**?**]. This is a non-trivial task often neglected in state of the art face recognition benchmarks that usually focus on an optimization of classification accuracy (true-positives and false-positives values) of "known" faces in the entire database and neglecting the robust separation of "unknown" faces. In a practical application, such as media production and archiving, the majority of faces in the content is likely not be in the dataset, which produces a large number false detections, even at low to moderate false positive rates of the algorithm.

In this paper we address robustness issues in a face detection and recognition pipeline. First, we analyse the impact of image degradations (in particular compression) on face detection, and how conceal them in order to improve face detection performance. This is studied both on face samples originating from still image and video data. Second, we propose approaches to improve the classification of faces into "known" and "unknown", in particular to reduce false positives recognitions.

The rest of this paper is organized as follows. Section 2 discusses related work and 3 presents the experiments we have performed to

AI4TV, October, 2018, Nice, FR

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

analyse the problem and possible solutions. We report and discuss the results in Section 4, and Section 5 concludes the paper.

2 RELATED WORK

The impact of quality degradations in input images has only been studied in few works. [**?**] analysed the impact of different distortions (blur, noise, contrast, JPEG and JPEG2000 compression) on the performance of image classification using GANs. They used rather rather strong compression, i.e., JPEG compression parameters ≤ 20. For all the modifications there is a positive correlation between stronger distortion and decreased DNN performance, with different non-linear relations. [**?**] studied the relation between adversarial attacks and JPEG compression. They argue that JPEG compression tends to reduce high-frequency components in images, and thus has the potential to eliminate noise that could be use in adversarial attacks on the DNN. A somewhat related work is [**?**], which aims to develop image quality metrics targeting alignment with machine performance rather than human perception. They use face detection as example task to derive a metric that aligns with detection performance. However, their face detector is not DNN-based.

Robust distinction of faces never presented to the classifier from those, already learned has been rarely discussed in literature so far [**?**]. One earlier work dealing with the issue of "closed-set" recognition of faces and objects is the work of Scheirer et. al in [**?**]. In particular the authors propose a novel "1-vs-set machine" using modfied marginal distances from a linear, binary SVM. Although the work primarily focuses on object-recognition tasks, there is also an evaluation on Labeled Faces in the Wild (LFW) [**?**] face recognition dataset. More recent work on so called 'open-set' face recognition is the work of Guenther *et al.* [**?**] where the authors compared several algorithms for assessing similarity of deep-feature approaches and concluded, that only extreme value machines (EVM) [**?**] can sufficiently discriminate between 'known' and 'unknown' persons on LFW database. Anyway, the reported performance of EVMs is not sufficient for real-time surveillance applications (only 60% correctly classified faces at an false alarm rate of 0.01) and additional research is recommended for practical applications. Another interesting work reporting also results on the popular Youtube Faces (YTF) [**?**] database is the one from Sun *et al.* in [**?**] where stacked set of novel high-performance deep convolutionalnetworks (25 DeepID2+ networks) has been proposed to achieve new state-of-the-art results. Similar to the observations in the latter work the performance on 'closed-set' evaluations is 99.47% and 93.2% on LFW and YouTubeFaces respectively, but performance for face identification degrades to 80.7% in the 'open-set' evaluation benchmark. Important to note is also the work of Liu *et al.* in [**?**] where the authors proposed a novel loss function (A-Softmax loss) for a CNN architecture (termed 'sphereface') in order to replace Euclidean metrics based margins by a proper Face-manifold metric. This metric can be used to recognize faces with a nearest neighbor classifier while coevally using distance-thresholding in the hypersphere manifold. Reported accuracies on LFW and YTF dataset are 99.42% and 95.0% in a 'closed-set' benchmark protocol respectively. Performance for 'open-set' evaluation is only reported for MegaFace data [**?**], but naturally shows less performance between 72% and 75% depending on the 'SphereFace' variant implemented.

So in summary we can make the following observations from the existing literature. None of the works addresses particularly face detection, and all work entirely on data originating from still images. Also, no concealment strategies have been studied in existing works. Regarding 'closed-set' face recognition performance of approaches we can conclude, that reported performance in literature is below the needs for practical applications and research for novel algorithms is highly recommended.

3 EXPERIMENTS

We use two datasets for the experiments. Labeled faces in the wild (LFW) [**?**] is a commonly used dataset for face recognition, containing 13K still images (JPEG compressed) of faces collected from the web and showing 1,680 labeled persons. Youtube Faces (YTF) [**?**] is a dataset created from web video and thus containing various video compression qualities. The data set contains 3,425 videos of 1,595 different people, which amounts to about 600K frames with face detections.

3.1 Face detection under distortions

For both datasets, we run face detection using multi-task CNNs [**?**] on each of the images. The images usually contain a single face, however, in particular some of the LFW images contain small faces in the background. Thus we limit the number of detected faces to 1, using the largest face only. We apply the detector to the original image, which will indicate where detection fails due to the compression of the source or the insufficient performance of the face detector, and to distorted versions of the images. We apply the following set of distortions:

Blurring. A box filter with size $k \times k$ is applied, with $k = \{3, 6, 9, 12, 15, 18, 21, 27, 30\}$.

Sharpening. A sharpening approach based on unsharp masking is applied. The source image is blurred with a 3×3 binomial filter, and the difference between the source and blurred source image is multiplied the the magnitude m_s an added again to the image. The magnitudes used are $m_s = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60\}$.

JPEG compression. The image is recompressed with a JPEG quality factor of q, with $q = \{90, 75, 60, 45, 30\}$. In contrast to [**?**] we used high to moderate compression settings which better reflect the content in professional media production.

JPEG compression concealment. The analysis of the differences between images compressed with different JPEG compression factors shows that most differences are quantisation noise, with rather small absolute pixel differences. The results of the blurring experiments (for details see Section 4) indicate that moderate blurring never harms the face detection performance. We empirically determined on a small set of sample images, that the compression artifacts can be well suppressed by blurring with a 4×4 box filter, that is applied in two passes. Blurring only once or with a smaller kernel did not sufficiently reduce the artifacts. Thus we apply this blurring to the source image (to address cases where the compression of the source already prevents successful detection) and to each of the recompressed JPEG images.

3.2 Unknown face classification

As shown by *e.g.* [**?**] one of the best performing approaches to 'closed-set' face recognition is FaceNet [**?**]. Recent modifications of this approach for media production and archiving used a combination of FaceNet-features with incremental machine learning approaches to automatically train classifiers for 'unknown' persons [**?**]. In a first experiment we started with their approach and found, that their is a lack of performance for the usage on real practical applications (see section 4.3 and 4.4 for more details).

Improved probabilty measures for the online-random forest.

Pre-processing using correlation-measures in feature-bag .

4 RESULTS

4.1 Face detection results on LFW

The results of LFW are shown in the plots in Figure 1. The first observation is that the face detection rate for the source image is at just above 0.997, i.e., the face in nearly 3% of source images are not detected. For blurring, there is hardly any impact up to a kernel size of $k = 9$, and then the performance start to decline quickly. For sharpening, there is already a small performance loss with the smallest magnitude, and then the performance loss is approximately linear with the magnitude. The results for JPEG compression show a similar behavior of roughly linear performance reduction, but the performance loss is much smaller.

When compression artifact concealment by blurring is applied, the detection rate on the source images increases to 0.9994, i.e., the miss rate drops to one fifth of the one without concealment. When applying concealment to the reencoded JPEG images with lower quality factors, the results oscillate around the detection rate for the source images with concealment. This result indicates that this level of JPEG compression does not cause loss of relevant information on LFW. Any missed detections due to JPEG artifacts are due to the quantisation noise, thus reconstruction to the original performance level is possible.

4.2 Face detection results on YTF

The results of YTF are shown in the plots in Figure 2. Overall they show are similar to those on LFW, with a decline of performance for blurring with kernel sizes above $k = 9$, and roughly linear declines for sharpen and JPEG compression, with a smaller magnitude for the JPEG compression. However, the performance level for the source images is just above 0.995, i.e., the face in nearly 5% of source images are not detected.

When compression artifact concealment by blurring is applied, the detection rate clearly increases. It does not reach the same level as on LFW, but still the miss rate if halved. When applying concealment to the reencoded JPEG images with lower quality factors, the detection performance stays nearly constant, with a very small decline with quality factors below 45. However, the performance stays about 1% below that of the source images without concealment. The main difference is that the source content on YTF has already undergone video compression, while LFW source images are moderately JPEG compressed. Thus not only the overall performance on the source content is lower, but also the effects of further compression cannot be entirely eliminated.

4.3 Classifying unknown faces on LFW

Results for improved probability measures.

Results for correlation-measure preprocessing.

4.4 Discussion

Face detection under distortions. We can gain the following insights from the experiments for face detection on the two datasets.

- Blurring and sharpening causes as expected performance loss that is proportional to the strength of the defect. However, the relation between the parameter of the defect and the performance impact is quite different, as is the effect at small strengths (some tolerance up to certain strength vs. immediate impact).
- Compression does have a non-negligible impact on the performance of face detection, and also the unmodified source images of common datasets are affected.
- At high to moderate JPEG quality factors, this performance loss is not due to a loss of information, but due to quantisation noise that is independent of the quality factor.
- Slight blurring does not cause reduction of detection performance, but can reduce JPEG quantisation noise in at least half of the cases where detection on the original source images fails.
- The findings of [**?**] are only partly confirmed by our analysis. While the compression will remove high-frequency noise from the source content, which could be used for adversarial attacks, the compression process will also produce quantisation noise. This may have an impact on the detection results, though it may be more difficult to use it for an adversarial attack.
- On content with high quality, the additional compression will not cause information loss, and concealment reaches the same performance level whether starting from the source or a compressed version. On content with already higher source compression, concealment always provides improvement, though not beyond that of the source content.

Unknown face classification.

5 CONCLUSION

In this paper, we have proposed stategies for improving robustness of face detection and classification of unknown faces in a face recognition pipeline.

For handling compressed content, the use of slight blurring as a concealment strategy seems useful. In the experiments, the missed detections were at least halved, and no negative impact of the amount of blur could be observed. This is a very efficient and easy to apply preprocessing. Alternative approaches would require retraining of the respective face detector. Data augmentation by providing more compressed samples could be one option. However, given the observations on the nature of the distortion, augmentation would benefit from a large number of differently compressed samples to eliminate any statistical patterns in the quantisation noise rather than covering a broad range of quality factors. The relatedness of the quantisation noise to adversarial samples has been mentioned. Thus a recently proposed approach for adversarial training called

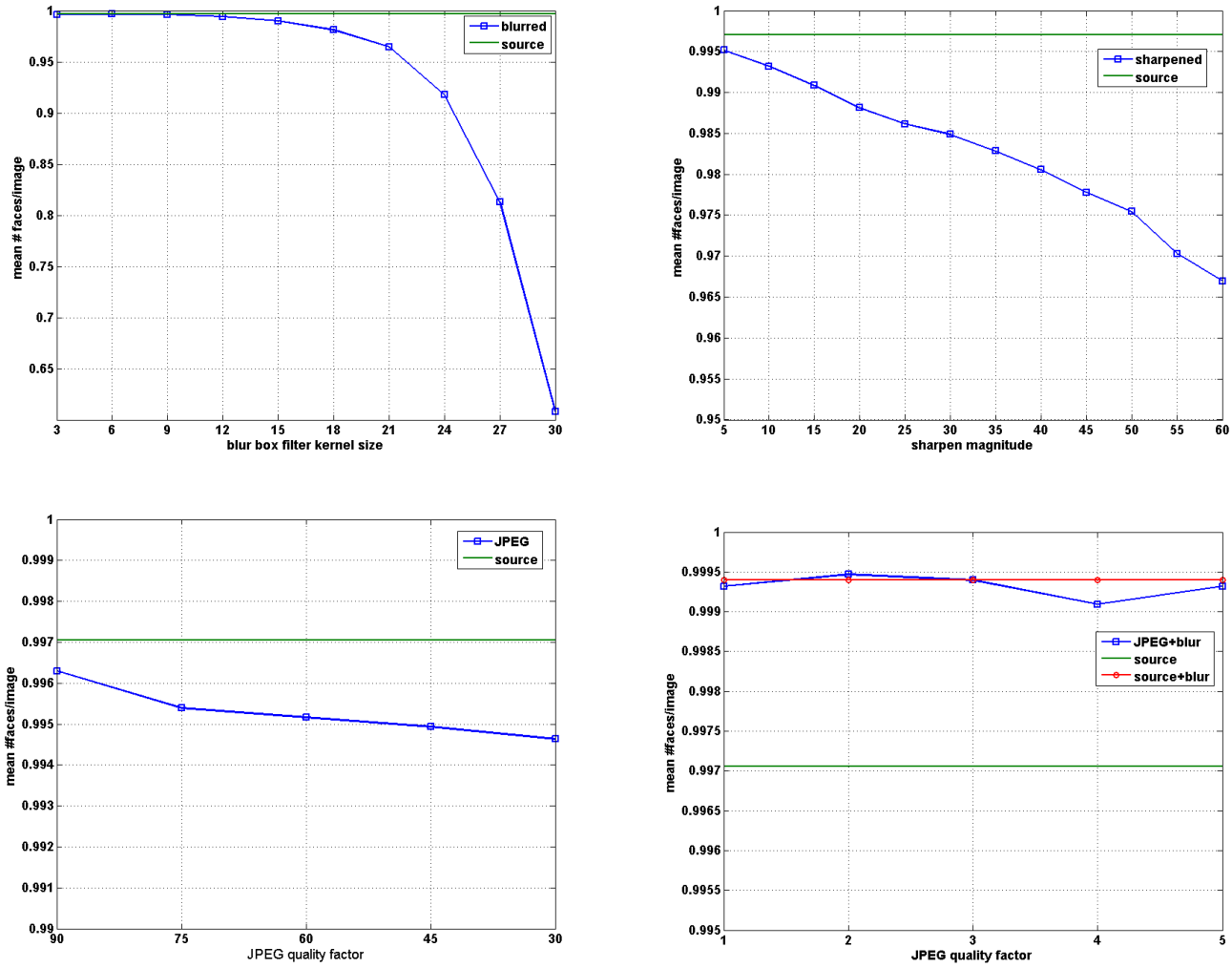


Figure 1: Face detection results on LFW.

ME-Net [1] might be applicable, which using matrix estimation to replace the original training data with an approximated version, thus eliminating noise while preserving larger structures.

ACKNOWLEDGMENTS

(removed for review)

REFERENCES

[] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. 2018. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 196–204.

[] Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, 1–6.

[] Manuel Gunther, Steve Cruz, Ethan M Rudd, and Terrance E Boult. 2017. Toward open-set face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 71–80.

[] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. 2016. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*. Springer, 189–248.

[] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 212–220.

[] Daniel Miller, Evan Brossard, S Seitz, and Ira Kemelmacher-Shlizerman. 2015. Megaface: A million faces for recognition at scale. *arXiv preprint arXiv:1505.02108* (2015).

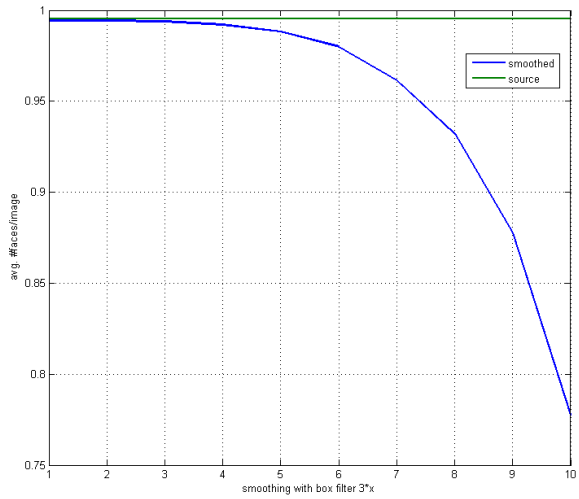
[] Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. 2017. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence* 40, 3 (2017), 762–768.

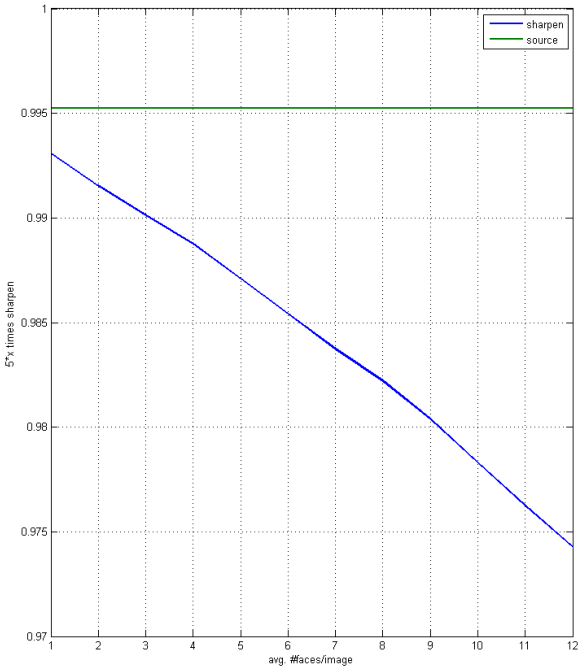
[] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. 2013. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 7 (July 2013), 1757–1772.

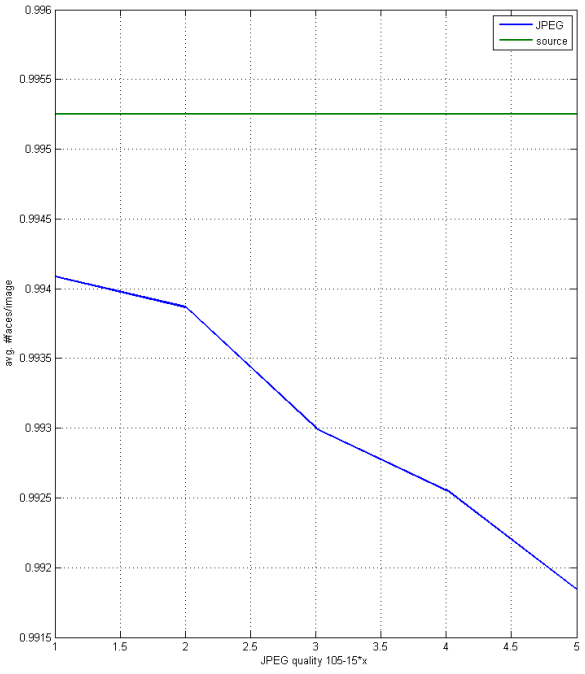
[] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[] Rajiv Soundararajan and Soma Biswas. 2019. Machine vision quality assessment for robust face detection. *Signal Processing: Image Communication* 72 (2019), 92–104.

[] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on*







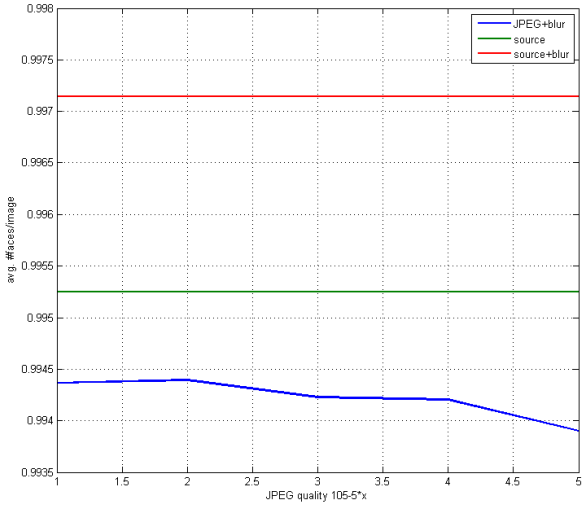


Figure 2: Face detection results on YTF.

computer vision and pattern recognition. 2892–2900.

[] Martin Winter and Werner Bailer. 2019. Incremental Training for Face Recognition. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*. 289–299. https://doi.org/10.1007/978-3-030-05710-7_24

[] L Wolf, T Hassner, and I Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 529–534.

[1] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. 2019. ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation. In *International Conference on Machine Learning*. 7025–7034.

[] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.