IET Biometrics

Special Issue: Face Recognition and Spoofing Attacks

IET Journals
The Institution of Engineering and Technology

Strengths and weaknesses of deep learning models for face recognition against image degradations

ISSN 2047-4938
Received on 17th May 2017
Revised 14th August 2017
Accepted on 7th September 2017
E-First on 24th October 2017
doi: 10.1049/iet-bmt.2017.0083
www.ietdl.org

Klemen Grm[1] ✉, Vitomir Štruc[1], Anais Artiges[2], Matthieu Caron[2], Hazım K. Ekenel[3]

[1]*Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia*
[2]*Graduate School in Electrical & Computer Engineering and Telecommunications, ENSEA, 6 Avenue du Ponceau, 95015 Cergy, France*
[3]*Department of Computer Engineering, Istanbul Technical University, Maslak, 34469 Istanbul, Turkey*
**First authors with equal contributions.*
✉ *E-mail: klemen.grm@fe.uni-lj.si*

Abstract: Convolutional neural network (CNN) based approaches are the state of the art in various computer vision tasks including face recognition. Considerable research effort is currently being directed toward further improving CNNs by focusing on model architectures and training techniques. However, studies systematically exploring the strengths and weaknesses of existing deep models for face recognition are still relatively scarce. In this paper, we try to fill this gap and study the effects of different covariates on the verification performance of four recent CNN models using the Labelled Faces in the Wild dataset. Specifically, we investigate the influence of covariates related to image quality and model characteristics, and analyse their impact on the face verification performance of different deep CNN models. Based on comprehensive and rigorous experimentation, we identify the strengths and weaknesses of the deep learning models, and present key areas for potential future research. Our results indicate that high levels of noise, blur, missing pixels, and brightness have a detrimental effect on the verification performance of all models, whereas the impact of contrast changes and compression artefacts is limited. We find that the descriptor-computation strategy and colour information does not have a significant influence on performance.

1 Introduction

Recent advances in deep learning and convolutional neural networks (CNNs) have contributed to significant performance improvements in a number of computer vision problems, ranging from low-level vision tasks such as saliency detection and modelling [1, 2] to higher-level problems such as object detection [3, 4], recognition [5–9], tracking [10–12], or semantic segmentation [13–15]. Deep learning-based approaches have been particularly successful in the field of face recognition, where contemporary deep models now report near perfect performance on popular, long-standing benchmarks such as LFW [16], which due to its difficulty, represented the de facto standard for evaluating face recognition technology for nearly a decade.

Most of the ongoing research on deep learning-based face recognition focuses on new model architectures, better techniques for exploiting the generated face representations, and related approaches aimed at improving both the performance and robustness of deep face recognition technology on common benchmark tasks [17–19]. Research in these areas is typically conducted on unconstrained datasets with various sources of image variability present at once, which makes it difficult to draw clear conclusions about the sources of errors and problems that are not addressed appropriately by the existing deep CNN models. Much less work is devoted to the systematical assessment of the robustness of deep learning models for face recognition against specific variations. Considering the widespread use of deep CNN models for face recognition, it is of paramount importance that the behaviour and characteristics of these models are well understood and open problems pertaining to the technology are clearly articulated.

In this paper, we contribute toward a better understanding of deep learning-based face recognition models by studying the impact of image-quality and model-related characteristics on face verification performance. We use four state-of-the-art deep CNN models, i.e. AlexNet [20], VGG-Face [19], GoogLeNet [21], and SqueezeNet [22], to compute image descriptors from input images and investigate how quality-related factors such as blur, compression artefacts, noise, brightness, contrast, and missing data affect their performance. Furthermore, we also explore the importance of colour information and descriptor-computation strategies through rigorous experimentation using the LFW (LFW) benchmark [16]. The deep CNN models considered in this work are representatives of the most commonly employed CNN architectures in use today and were selected due to their popularity within the research community. The studied covariates, on the other hand, represent factors commonly encountered in real life that are known to affect face recognition technology to a significant extent [23] and have not yet been studied sufficiently in the literature in the context of deep learning.

The comprehensive analysis presented in this paper builds on the previous works from [24, 25]. These works both focused on closed-set face identification and investigated the robustness of deep CNN models under facial appearance variations caused by head pose, illumination, occlusion, misalignment in [24], and by image degradations in [25]. Complementing and extending these previous works, we provide in this paper a rigorous and systematical evaluation of the impact of various image- and model-related factors on deep learning-based face verification performance. The goal of this work is to provide answers to essential research questions such as: Are good quality images a must for high verification performance? To what extent does image quality affect the image descriptors generated by contemporary deep models? Are certain model architectures more robust than others against variations of specific covariates? Changes in which quality characteristics are most detrimental to the verification performance? How should image descriptors be computed? Answers to these and similar questions are in our opinion crucial for a better understanding of deep learning-based face recognition technology and may point to open problems that need to be addressed in the future. In summary, we make the following contributions in this paper:

- We study and empirically evaluate the effect of image quality (blur, Joint Photographic Experts Group (JPEG) compression, noise, contrast, brightness, and missing data), and model-related

(colour information and descriptor computation) characteristics on the face verification performance of four state-of-the-art deep CNN models on the LFW dataset.
- We conduct a comprehensive analysis of the experimental results, identify the most detrimental covariates affecting deep CNN models in face verification task, and point to potential areas for improvement.
- We provide a comparative evaluation of the four deep CNN models, namely AlexNet [20], VGG-Face [19], GoogLeNet [21], and SqueezeNet [22], and make the trained models publicly available to the research community through: https://github.com/kgrm/face-recog-eval.

The rest of this paper is organised as follows: in Section 2, we briefly review previous works relevant to our study. In Section 3, we describe the evaluation methodology, models, datasets, and experimental procedures used. In Section 4, we present quantitative results and discuss our experiments. Finally, Section 5 concludes this paper.

2 Related work

Understanding the strengths and weaknesses of machine learning models is of paramount importance for real-world applications and a prerequisite for identifying future research and developments needs. Papers on the analysis of deep models appear in the literature in either (i) work that focuses specifically on the characteristics of deep models or (ii) work that explores the characteristics of deep models as part of another contribution. Papers from the first group such as ours typically explore various models and as the main contribution presents general findings that apply to several deep models, while papers from the second group propose a new deep learning approach and then analyse its characteristics. Both groups of work typically contribute to a better understanding of deep models, but differ in their generality, i.e. the number of models the findings apply to.

An example of work studying the impact of various image-quality covariates on the performance of several deep CNN models was presented by Dodge and Karam in [26]. Here, the authors explored the influence of noise, blur, contrast, and JPEG compression on the performance of four deep neural network models applied to the general image classification task. The authors concluded that noise and blur are the most detrimental factors.

In [27], Chatfield *et al.* compared traditional machine learning models and deep learning models on equal footing by using the same data augmentation and preprocessing techniques that are commonly used with CNNs on traditional machine learning models. The authors also explored the importance of colour information, but focused on the impact of colour on traditional models rather than on its role in deep learning. The main finding of this work was that deep learning models have an edge over traditional machine learning models. However, data augmentation, colour information, and other preprocessing tasks were found to be important, as these approaches also helped to improve the performance of traditional machine learning models.

An alternative view on covariate analyses involving deep models was recently presented by Richard-Webster *et al.* in [28]. In this work, the authors compare and evaluate several deep CNN architectures from the perspective of visual psychophysics. In the context of the object recognition task, they use procedurally rendered images of three-dimensional models of objects corresponding to the ImageNet object classes to determine the 'canonical views' learned by deep CNNs and determine the networks' performance when viewing the objects from different angles and distances or when the images are subjected to deformations such as random linear occlusion of the object bounding box, Gaussian blur, and brightness changes. The main point made by the authors is that model comparison must be conducted under variations of the input data, or in other words, the analysis of the robustness of the models should be used as a methodological tool for model comparison.

Our work builds on the preliminary results reported in [24, 25] and extends our previous results to face verification experiments on the LFW dataset and a wide range of image-quality and model-related covariates. The analysis includes a larger number of deep CNN models and is significantly more comprehensive in terms of amount of analysed factors.

Dosovitskiy *et al.* describe research belonging to the group of model-analysis work in [29]. Here, the authors present an evaluation of the performance of their CNN in the presence of image transformations and deformations in the context of unsupervised image representation learning. They conclude that combining several sources of image transformations can allow CNNs to better learn general image representations in an unsupervised manner. Similar to this work, we study in this paper the effects of image deformations on the learned image representations. However different from [29], we assess several CNNs trained in a supervised manner.

Another work from this group was presented by Zeiler and Fergus in [30]. Here, the authors studied the effects of image covariates including rotation, translation, and scale in the context of interpreting and understanding the internal representations produced by deep CNNs trained on the ImageNet object classification task. In their experiments, the invariance of their CNN to the studied covariates was found to increase significantly with network depth. They also found the deep neural network features to increase in discriminative power with network depth in the context of transfer learning.

More recently, Lenc and Vedaldi in [31] evaluate how well the properties of equivariance, invariance, and equivalence are preserved in the presence of image transformations by various image representation models including deep CNNs. The transformations studied include rotation, mirroring, and affine transformations of the input images. Amongst their findings, representations based on deep CNNs were found to be better than other studied representations at learning either invariance or equivariance to the studied transformations based on training objectives.

3 Methodology

In this section, we first explain the evaluation methodology and introduce the four deep CNN models selected for the analysis. We then proceed by presenting the dataset and procedure used to train the deep models and conclude the section with a detailed description of the covariates considered in this work.

3.1 Evaluation methodology

To assess the robustness of deep CNN models against various image degradations in face verification, we take four pretrained state-of-the-art deep models and use the feature output of each model as the image descriptor of the given input face image, i.e.

$$\boldsymbol{y} = f(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ denotes the input image, $f(\cdot)$ represents the selected deep model, and $\boldsymbol{y} \in \mathbb{R}^{d'}$ stands for the computed image descriptor. The dimensionality of the image descriptor, d', varies from model to model and depends on the design choices made during network construction. Once the descriptors are computed for a pair of face images, a similarity score is calculated based on the cosine similarity between the two descriptors and used to make a verification decision

$$g(\boldsymbol{x}_1, \boldsymbol{x}_2, f, T) = \begin{cases} w_1 & \text{if } \delta(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) = \delta(\boldsymbol{y}_1, \boldsymbol{y}_2) \geq T \\ w_2 & \text{otherwise} \end{cases} \tag{2}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the input images, $\delta(\cdot, \cdot)$ is the cosine similarity, T is a predefined decision threshold, and w_1 and w_2 represent classes of matching and non-matching identities, respectively. Thus, a pair of images should be classified into the

Table 1 Comparison of the quantitative properties of the deep learning models considered

Model	#Parameters	Input size	Output size	#Layers	FLOPS/forward pass
AlexNet [20]	58 282 752	(3, 224, 224)	4096	7	1.1×10^9
VGG-Face [19]	117 479 232	(3, 224, 224)	4096	15	1.5×10^{10}
GoogLeNet [21]	21 577 728	(3, 299, 299)	2048	37	5.6×10^9
SqueezeNet [22]	3 753 856	(3, 224, 224)	2048	12	9.7×10^8

class w_1 if the input images belong to the same identity and into the class w_2 if not.

To assess the robustness of the deep models with respect to different image-quality covariates, we artificially degrade one of the images in (2) by adding different levels of noise, blur, compression artefacts, and the like and leave the second one unaltered. With this procedure, we are able to directly observe the change in verification performance as a consequence of the change in image quality and establish a connection between a given image-quality aspect and the performance of the deep model.

We report our results using the performance metrics introduced by the LFW verification protocol [32], namely the mean and standard deviation of the verification accuracy under a ten-fold cross-validation experimental protocol. As prescribed by the LFW experimental protocol, the decision threshold T is selected separately for each fold.

3.2 Deep CNN models

We consider four recent deep CNN models in our experiments that are representative of the most popular network architectures commonly used for recognition problems, i.e.:

AlexNet: The first model used in our evaluation is the AlexNet [20], which was the first deep CNN to successfully demonstrate performance outperforming the classical image object recognition procedures. The model consists of five sequentially connected convolutional layers of decreasing filter size, followed by three fully connected layers. One of the main characteristics of AlexNet is the very rapid downsampling of the intermediate representations through strided convolutions and max-pooling layers. The last convolutional map is reshaped into a vector and treated as an input to a sequence of two fully connected layers of 4096 units in size. The output of this layer represents the image descriptor produced by AlexNet.

VGG-Face: The second model used in our experiments is the 16-layer VGG-Face network, initially introduced in [19]. The model has a deeper convolutional architecture than AlexNet and exploits a series of convolutional layers with small filter sizes, i.e. 3×3. Each series of convolutional layers is followed by a max-pooling layer, except for the last one, which is followed by two fully connected layers identical to AlexNet. The output of the last fully connected layer represents the VGG image descriptor.

GoogLeNet: Our third model is the GoogLeNet network, which builds on the so-called Inception architecture [9, 21]. Here, we use the third version of the GoogLeNet model, that is, Inception V3 [21], which consists of a hierarchy of complex *inception* modules/blocks that combine channel re-projection, spatial convolution, and pooling operations over different scales in each of the modules. The model reduces the parameter space by decomposing spatial convolutions with larger filter sizes ($n \times n$) into a sequence of two convolutional operations with respective filter sizes of $n \times 1$ and $1 \times n$. The resulting network model is deeper and more complex than AlexNet or VGG-Face, but still has fewer parameters and lower computational complexity than VGG-Face. Unlike other models considered in this work, no fully connected layers are used in GoogLeNet. Instead, the last convolutional map is subjected to channel-wise global average pooling, and the average activation values of each of the 2048 channels are used as the feature vector of the input image.

SqueezeNet: The last model we assess in our experiments is a variant of the SqueezeNet network from [22]. The network features extreme reductions in parameter space and computational complexity via channel-projection bottlenecks (or squeeze layers), and uses identity-mapping shortcut connections, similar to residual networks [8], which allow for stable training of deeper network models. SqueezeNet was demonstrated to achieve comparable performance with AlexNet [20] on the ImageNet large-scale recognition benchmark with substantial reductions in model complexity and parameter space size. The model is comprised of so-called 'fire modules', in which the input map is first fed through a bottlenecking channel-projection layer and then divided into two channel sets. The first one is expanded through a 3×3 convolution and the other through channel projection. The final convolution map is globally average-pooled into a 512 vector and then fed to a fully connected layer with 2048 units. The output of this last layer is the SqueezeNet image descriptor used in our experiments.

Note that using deep models as 'black-box' feature extractors is a standard way of computing (learned) descriptors from input images, as evidenced by the large body of existing research on this topic, e.g. [28, 33, 34]. Furthermore, using distance metrics in the feature space for similarity score calculation is also a standard practise in the field of biometric verification, see for example [17, 19]. All in all, the deep neural network architectures considered in this work are amongst the most popular ones found in the literature and differ greatly in computational complexity, the number of parameters, depth, and representational power. We summarise their key properties including the output feature vector (descriptor) size in Table 1.

3.3 Datasets

We use separate datasets for training and evaluation. We chose the VGG-Face dataset [19] to train our models and the LFW dataset [16] to evaluate their performance.

The VGG-Face dataset was collected during the work on the VGG-Face model [19] and, as reported by the authors, comprises around 2.6×10^6 images of 2622 identities. Using the web addresses and face region coordinates of the images published by the authors, we are able to retrieve $\sim 1.8 \times 10^6$ of the total 2.6×10^6 face images for our version of the dataset. The structure of the VGG dataset, with a uniformly distributed and relatively large number of images per subject, 1000, makes it similar in utility for training deep neural networks to the ImageNet dataset, which is used for image classification [35].

For the experiments, we train the four deep CNN models described in the previous section from scratch using our version of the VGG-Face dataset to attain a fair comparison of their expressivity and other properties given the same training dataset. We train the models by appending a fully connected softmax layer on top of each network and optimising the model weights in accordance with the recognition performance on the VGG data. We use the Adam [36] gradient optimisation method with the categorical cross-entropy loss function. During training, we randomly select 10% of the images of each subject for a hold-out validation set to gauge generalisation performance. Each model is trained to convergence using a GTX Titan X Graphics Processing Unit (GPU). The training takes $\sim$2 days each for the AlexNet and SqueezeNet models, and 1 week each for the GoogLeNet and VGG-Face models.

For testing purposes, we select the LFW [16] dataset, which is among the most popular datasets used to evaluate face recognition models. The dataset consists of 13,233 images of 5749 distinct subjects, and ships with predefined training and evaluation protocols. Images of the dataset were gathered from the web and feature considerable variation in pose, lighting condition, facial expression, and background. We evaluate our models in accordance with the so-called *outside-data verification protocol*, which

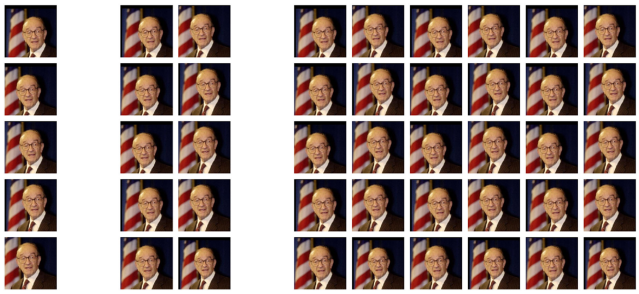

Fig. 1 *Illustration of the three sampling schemes used to study different descriptor-computation strategies (from left to right): the 5-patch (left), 10-patch (centre), and 30-patch (right) schemes*

consists of 6000 image pairs drawn from the dataset equally divided between genuine and impostor pairs, and further equally divided into ten folds for cross-validation. The protocol also allows to use images not part of LFW to train the models being evaluated.

3.4 Performance covariates

The performance of deep face recognition models depends on several factors (or covariates) that can be grouped into different categories. In this paper, we are interested in factors that relate to: (i) the quality of the input images (image-quality covariates) and (ii) the characteristics of the deep models (model-related covariates).

Image-quality covariates: To evaluate the impact of reduced image quality on the performance of our deep models, we apply image distortions of different levels/intensities to the probe images used in our verification experiments. Specifically, we consider the following:

- *Blur:* We simulate blurring effects by applying Gaussian filters with different standard deviations σ to the probe images. We set the filter size in accordance with the selected standard deviations, i.e. $w, h = 2\lceil 2\sigma \rceil + 1$, where $\lceil \cdot \rceil$ is the ceiling operation and w and h stand for the filter width and height, respectively. We vary the value of σ from 2 to 20 and, thus, generate 19 probe sets of different blur levels to investigate the impact of blurring on the performance of our deep models.
- *Compression:* We introduce compression artefacts by encoding the probe images with the JPEG algorithm at different quality presets. Lower-quality presets correspond to more aggressive quantisation of the discrete cosine transform (DCT) coefficients. At the extreme, the quality of 1 corresponds to the setting where all AC components of every minimum coded unit (MCU) block are zeroed out, and each 8×8 pixel block is represented by a constant colour. We generate modified probe sets at quality presets of 1, 3, 5, 10, 15, 20, 25, 30, 35, and 40 for exploring the impact of JPEG compression.
- *Gaussian noise:* To study the impact of noise on the recognition performance of our deep models, we add additive Gaussian noise with a mean of 0 and various standard deviations σ to our probe images. The modified pixel intensities are clipped to the valid dynamic range of [0, 255]. We generate 10 modified probe sets for σ values between 20 and 200, with a uniform step size of 20.
- *Salt-and-pepper noise:* Besides Gaussian noise, we also consider salt-and-pepper noise. Here, we truncate all colour components of each image pixel to zero with a probability of $p/2$ and, similarly, set them to 255 with a probability of $p/2$. We generate 25 modified probe sets for probabilities p between 0.02 and 0.5, with a uniform step size of 0.02.
- *Brightness:* We simulate overexposure effects by changing the brightness level of the probe images. To this end, we multiply the pixel intensities by a brightness factor and clip the resulting pixel values to the valid dynamic range between [0, 255]. We observe the impact of brightness factors between 1.5 and 9 with a constant step size of 0.5 and generate 16 probe sets for our brightness-related experiments.
- *Contrast:* To explore the impact of contrast on the verification performance, we first subtract the central value of the dynamic range from all images. The centred images are then multiplied by a contrast factor and the offset, i.e. the central value, is added back to the image. We evaluate the performance of the models at 15 different contrast factors between 0.03 and 0.79.
- *Missing data:* We simulate missing data (or pixels) by removing contiguous pixel areas from the probe images. Since we set all pixels in the given area to zero, the simulation of missing data is similar in effect to (artificial) partial occlusions of the face. We generate five degraded probe sets with pixels missing around the mouth, nose, periocular, and eye regions. To be able to remove image regions belonging to prominent facial features in a consistent manner, we use the facial landmark detection approach proposed by Kazemi and Sullivan in [37].

Model-related covariates: Among the model-related covariates, we explore the following ones:

- *Model architecture:* Arguably the most important factor affecting the performance of the existing deep face recognition approaches is the architecture of the models and corresponding training procedure used to learn the model parameters. As indicated in the introduction section, a significant amount of todays' research effort related to deep models is, therefore, directed toward this area (see, e.g. [21, 22, 38]). In the experimental section, we account for different architectures by evaluating the four deep CNN models described in Section 3.2.
- *Descriptor computation:* One of the key components of state-of-the-art face recognition systems is the visual descriptor used to encode the input images [27]. With deep learning approaches, the visual descriptor is typically computed directly from the image area returned by the face detector. The predominant approach here is to feed the detected facial area to the trained deep model and use the output of one of the top fully connected layers as the visual descriptor of the input image. An alternative approach is to sample patches from the input image and to combine the corresponding patch representations into the final visual descriptor. Examples of the latter approach include averaging [19] or stacking [39] of patch representations and variants of Fisher vector encoding [40]. For our experiments, we consider four descriptor-computation strategies. The first is a simple approach, where the visual descriptor is computed directly from the facial area found by the face detector. The remaining three approaches are more complex and sample smaller patches from the facial area before averaging the patch representations generated by the models to produce the final image descriptor. We explore three sampling schemes using 5, 10, and 30 patches sampled from the detected facial area. The sampling schemes were implemented based on the suggestions in [19] and are illustrated in Fig. 1.
- *Colour space:* We consider three distinct scenarios relating to the colour information of the target and probe images used in the verification experiments. In the first two cases, given a colour target image, we evaluate the difference in verification performance of the deep models given either colour or grey-scale probe images. In the third case, we evaluate the performance of the models when target and query images are both grey scale. The goal of the colour-related experiments is to investigate the need for colour input images and the capabilities of the models to efficiently handle grey-scale images.

4 Experimental results and discussion

In this section, we describe our experiments aimed at assessing the strengths and weaknesses of the selected four deep models. We first present experiments related to image-quality covariates and then report results pertaining to the model-related covariates described in the previous section.

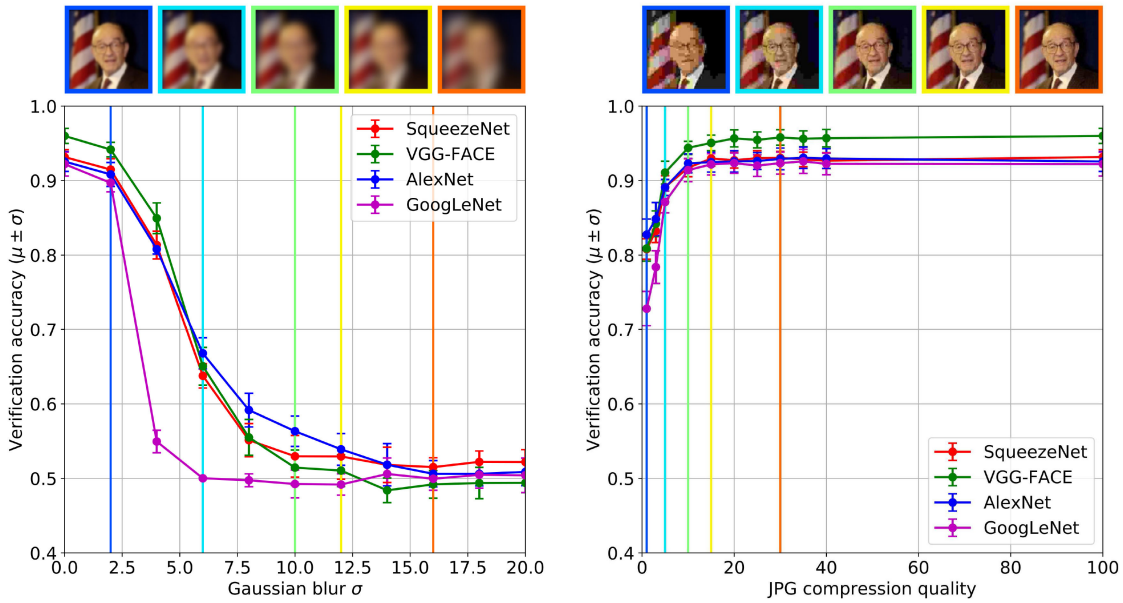

Fig. 2 *Impact of blurring (left) and JPEG compression (right) on the performance of the four deep models. The graphs show the mean and standard deviation of the verification accuracy on the LFW dataset computed over ten folds. The images on top of the graphs show sample images generated with different levels of image distortions. The results are best viewed in colour*

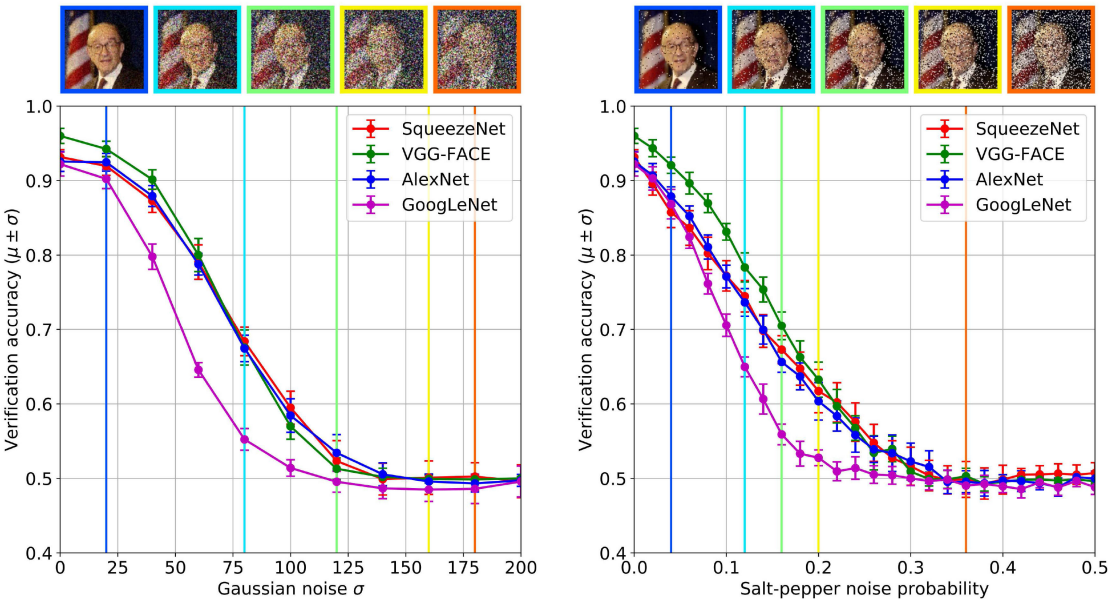

Fig. 3 *Impact of Gaussian (left) and salt-and-pepper (right) noise on the performance of the four deep models. The graphs show the mean and standard deviation of the verification accuracy on the LFW dataset computed over ten folds. The results are best viewed in colour*

4.1 Impact of image-quality covariates

In the first series of verification experiments, we explore the impact of Gaussian blur and JPEG compression. As can be seen in Fig. 2 (left), image blurring has a significant effect on the performance of all deep models, which causes a quick drop in performance with an increase in the standard deviation of the Gaussian filters. Interestingly, the GoogLeNet model loses verification accuracy faster than the other three models. When looking at the impact of JPEG compression in Fig. 2 (right), we see that all models are mostly unaffected by the compression artefacts until the compression quality is at its lowest possible value. Here, a compression quality of 0 corresponds to the scenario where all AC DCT coefficients are rounded to 0. Thus, only the DC components remain unaltered, and consequently every MCU is represented by a constant colour. This is equivalent to uniformly downscaling the image by a factor of 8. We observe that the verification accuracy of all models at the lowest JPEG quality roughly corresponds to the accuracy on the target images degraded with Gaussian blur with $\sigma = 5$, which is consistent with the above interpretation of the JPEG compression process in the sense that the amount of information preserved in the blurred and compressed images is approximately the same.

In the second series of experiments, we investigate the impact of Gaussian and salt-and-paper noise on the verification performance of the four deep models. From the results in Fig. 3, we see that the models behave similarly for both types of noises. The VGG-Face model performs the best and more robustly, followed by the AlexNet and SqueezeNet models, which perform more or less the same, and the GoogLeNet model, which is affected the most by the presence of noise. These results suggest that noise is an important factor affecting the performance of deep models and consequently that sufficiently low levels of noise need to be assured for reliable verification performance.

In the third series of verification experiments, we study the effects of brightness and contrast. We can see from the results in Fig. 4 (left) that the increase in brightness has a significant impact on the verification performance of the deep models and affects all models to more or less the same extent. In relative terms, no model has an edge over the others even at higher brightness factors, which

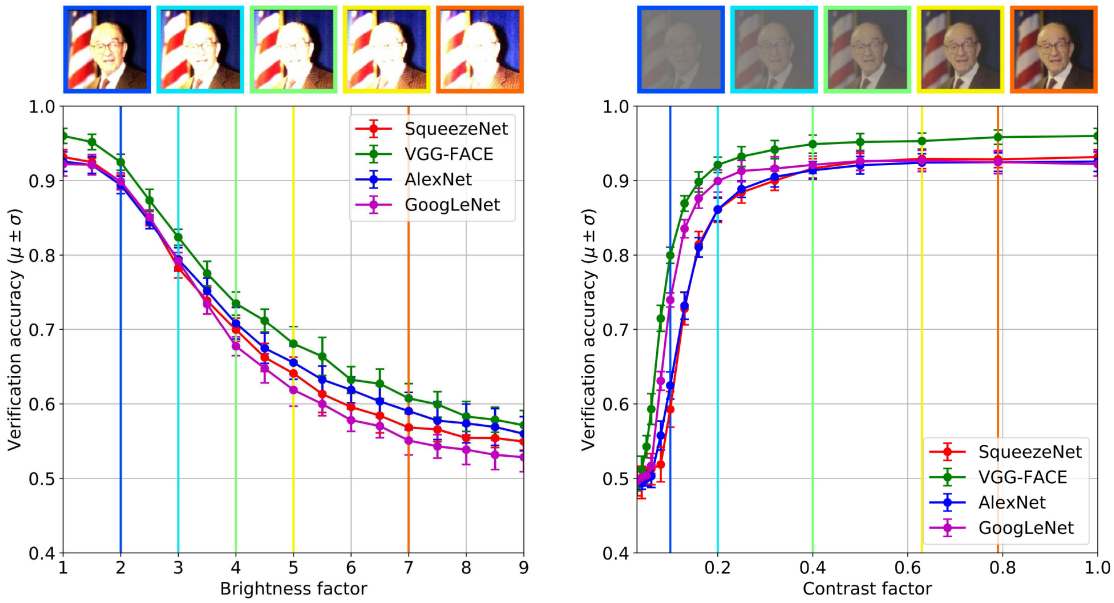

Fig. 4 *Impact of image brightness (left) and image contrast (right) of the verification performance of our four deep models. The graphs show the mean and standard deviation of the verification accuracy on the LFW dataset computed over ten folds. The results are best viewed in colour*

are expected as important discriminative information is lost during the brightening process due to the pixel truncation. However, in absolute terms, the VGG-Face model is the top performer ensuring the highest verification accuracy at all brightness factors. When looking at the results for different contrast factors in Fig. 4 (right), we see that the relative performance of all models degrades similarly as the contrast decreases. In general, the models are not particularly affected by the loss of contrast, as the verification accuracy remains well above 0.9 even when more than 60% of the contrast is removed.

In the last series of experiments pertaining to image quality, we evaluate the effects of missing data on the verification performance. The results are displayed in Fig. 5 in the form of box plots. We can see that the impact of missing data follows the same relative ranking for all models: missing information around the periocular region is the most detrimental for the verification performance, followed in order by the eye, nose, and mouth regions. Interestingly, we can see that the VGG-Face model is the most affected by missing data around the periocular region, whereas the performance degradation for other regions is equal or lower than the degradations of the other models. We can also note that the relative ranking of the tested models changes with respect to the image region, from which textural information was removed. While VGG-Face is the top performer in terms of average verification accuracy on the original images, it falls behind SqueezeNet and GoogLeNet when data around the eye, nose, or periocular regions is missing. All in all, GoogLeNet appears to be the most robust to missing data, as the performance variations are the smallest with this model. Overall, our experiments suggest that image quality is a crucial factor for the performance of existing deep models and that quality assessment of the input images should be an integral part of face recognition approaches based on deep learning. To mitigate problems pertaining to image-quality, image enhancement techniques need to be used or suitable data augmentation approaches need to be integrated into the training procedures to make the models robust against image-quality degradations.

4.2 Impact of model-related covariates

In the first series of experiments pertaining to model-related covariates, we assess the impact of different model architectures on the performance and robustness of the LFW verification task. We present our comparison in the form of radar charts for different probe sets that correspond to the colour-coded sample images at the top of Figs. 2–4. For example, the red curve in each chart corresponds to experiments with the probe images marked red in Figs. 2–4, the green curve to experiments with probe images marked green, and so on. Here, the larger the area covered by a curve, the better the performance of the models across various image-quality covariates and the closer the different colour curves are to each other for a given architecture, the robuster the architecture is to variations of the covariates. While all models perform similarly, the VGG-Face model has overall a slight advantage in terms of robustness over the remaining three models. The SqueezeNet and AlexNet models perform almost the same, whereas our implementation of GoogLeNet is the least robust (Fig. 6).

In the second series of verification experiments of this part, we evaluate the four different descriptor-computation strategies. The results of our experiments are presented in Fig. 7 in the form of box-and-whiskers plots computed from the ten experimental folds defined by the LFW verification protocol. In these experiments, we use the original LFW images without any degradations. We find that the SqueezeNet and VGG-Face models benefit marginally from averaging of the generated patch representations. While the trend shows an increase of 1–2% in verification accuracy by using more than a single patch to generate the image descriptors, the differences in performance are not statistically significant. The AlexNet and GoogLeNet models, on the other hand, do not show any improvements in performance. These results are unexpected as all models were trained with random patches sampled from the base face regions. We also note that while the SqueezeNet and VGG-Face models show some improvement when using 5 patches compared with only the central patch, there is no further improvement from the 10- or 30-patch schemes.

In the last series of experiments on model-related covariates, we explore the impact of colour information. The results of the experiment are shown in Fig. 8 again in the form of box plots. All models exhibit the best performance, when target and probe images are both in colour, which is expected given that they were trained on colour data exclusively. However, with the exception of AlexNet, we note that the accuracy of the models drops only marginally, when either the probe or both the target and probe images are switched to grey-scale. The difference in performance is not statistically significant, which points to a potential degree of redundancy in the models' architecture, observing that eliminating two-thirds of the input information results in nearly identical performance.

5 Conclusion

We have presented a systematic study of covariate effects on face verification performance of four recent deep CNN models. We

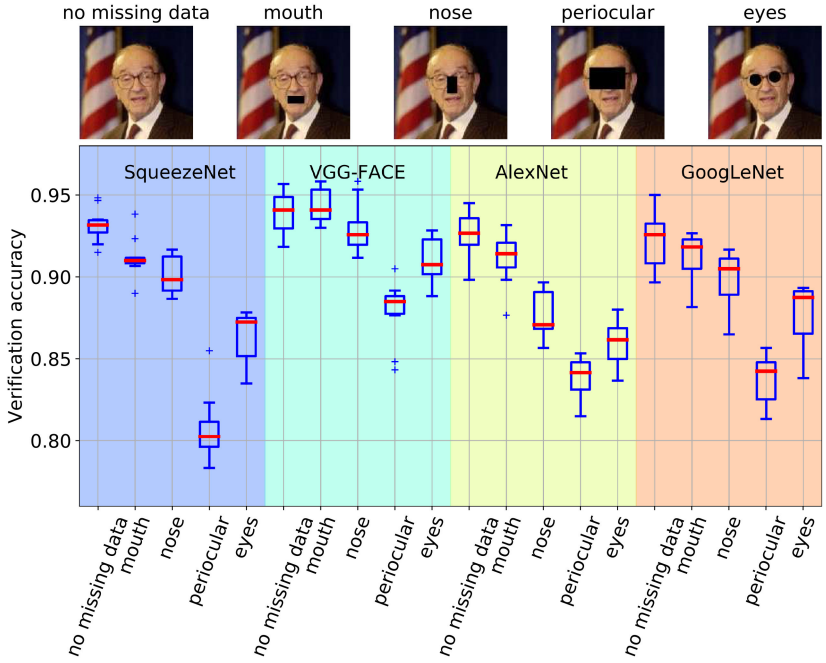

Fig. 5 *Impact of missing data on the performance of the four deep models. The box plots show results for missing data at four different image locations, i.e. around the mouth, the nose, the periocular region, and around the eyes. The box plots were computed from the ten experimental fold defined by the LFW verification protocol*

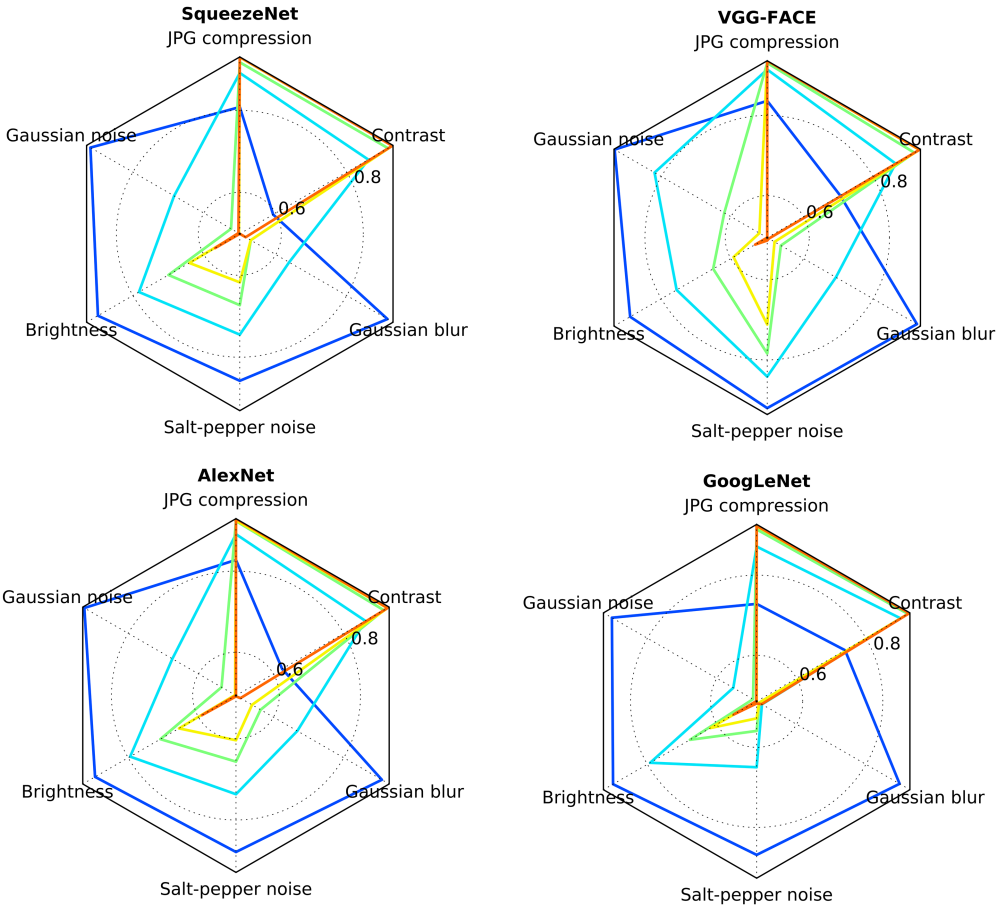

Fig. 6 *Impact of the model architecture on the performance and robustness of the verification procedure. Here, the line colours correspond to the colour-coded sample images on top of Figs. 2–4. A larger area covered by a curve indicates a better performance. The closer the curves of different colours are in a given graph, the more robust the model is to image-quality degradations*

observe that the studied models are affected by image quality to different degrees, but all of them degrade in performance quickly and significantly, when evaluated on lower-quality images than they were trained with. However, given proper architecture choices and training procedures, a deep learning model can be made relatively robust to common sources of image-quality degradations.

We found that the models considered were the most easily and consistently degraded in performance through image blurring, which is similar in nature to real-life scenarios of attempting face recognition from low-resolution imagery. Other covariates found to have a considerable effect on the verification performance were noise, image brightness, and missing data, while image contrast

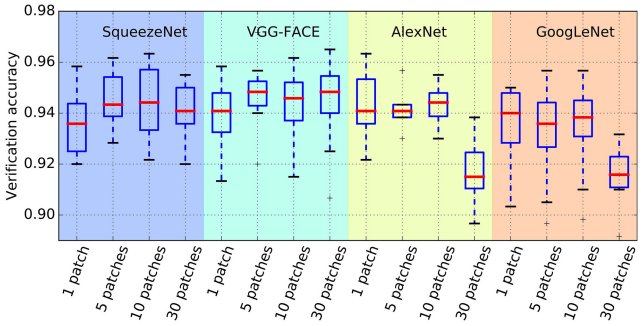

Fig. 7 *Performance evaluation for different sampling schemes. The box plots show results for the four sampling strategies, where image descriptors are computed based on either 1, 5, 10, or 30 face patches. The box plots were computed from the ten experimental fold defined by the LFW verification protocol*

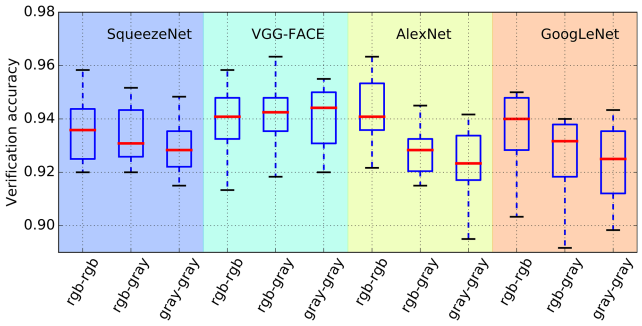

Fig. 8 *Performance evaluation between colour and grey-scale images. For each model, three comparisons were made: colour–colour, colour–grey-scale, and grey-scale–grey-scale*

and JPEG compression impacted the performance of the models only marginally.

Most of the models considered were least affected by changes in input colour space – despite being trained on full colour images – their performance drops negligibly when evaluated on grey-scale images. This finding is also corroborated by the results of the contrast experiments.

No specific architecture was found to be significantly more robust than others to all covariates. The VGG-Face model, for example, was most robust to noise, but performed least well for changes in image brightness. GoogLeNet, on the other hand, performed worst on noise and image blur, but had a slight advantage over the remaining models with images of reduced contrast.

On the basis of our results, we identify the following prospective directions of further research related to deep models:

- *Image enhancement*: Various algorithms exist to enhance the appearance of blurred or low-resolution images for human perception. Given the low face recognition performance on such images, their applicability to automated face recognition systems is likely to be an important research direction for deep face recognition models in the future.
- *Exploitation of colour information*: Given the fact that most of the models we studied retained almost unaltered performance when presented with grey-scale images, it appears to be the case that the architectures considered here do not make proper use of colour information in their input images. It follows that better deep learning models could be developed that either make more efficient use of their input information or that discard colour information altogether in favour of more compact models.
- Recognition from partial data – missing data proved to be a challenge for all evaluated models with performance deteriorating more, when larger contiguous areas of the images known to be important for identity inference were removed, e.g. the periocular region. This observation suggests that research into deep CNN models capable of recognition from partially observed data is needed and should be a focus of future research efforts.

6 Acknowledgments

This research was supported in parts by the ARRS (Slovenian Research Agency) Research Programme P2-0250 (B) Metrology and Biometric Systems, by ARRS through the junior researcher programme and by a Marie Curie FP7 Integration Grant within the 7th EU Framework Programme.

7 References

[1] Bruce, N.D., Catton, C., Janjic, S.: 'A deeper look at saliency: feature contrast, semantics, and beyond'. Proc. the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 516–524

[2] Li, G., Yu, Y.: 'Deep contrast learning for salient object detection'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016

[3] Gidaris, S., Komodakis, N.: 'Object detection via a multi-region and semantic segmentation-aware CNN model'. Proc. IEEE Int. Conf. Computer Vision, 2015, pp. 1134–1142

[4] Ren, S., He, K., Girshick, R., *et al.*: 'Faster R-CNN: towards real-time object detection with region proposal networks'. Advances in Neural Information Processing Systems, 2015, pp. 91–99

[5] Girshick, R., Donahue, J., Darrell, T., *et al.*: 'Rich feature hierarchies for accurate object detection and semantic segmentation'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587

[6] Gidaris, S., Komodakis, N.: 'LocNet: improving localization accuracy for object detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 789–798

[7] Liu, N., Han, J.: 'DHSNet: deep hierarchical saliency network for salient object detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 678–686

[8] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 770–778

[9] Szegedy, C., Liu, W., Jia, Y., *et al.*: 'Going deeper with convolutions'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 1–9

[10] Alahi, A., Goel, K., Ramanathan, V., *et al.*: 'Social LSTM: human trajectory prediction in crowded spaces'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 961–971

[11] Wang, N., Yeung, D.Y.: 'Learning a deep compact image representation for visual tracking'. Advances in neural information processing systems, 2013, pp. 809–817

[12] Wang, L., Ouyang, W., Wang, X., *et al.*: 'STCT: sequentially training convolutional networks for visual tracking'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 1373–1381

[13] Badrinarayanan, V., Handa, A., Cipolla, R.: 'SegNet: a deep convolutional encoder–decoder architecture for robust semantic pixel-wise labelling', arXiv preprint arXiv:150507293, 2015

[14] Chen, L.C., Papandreou, G., Kokkinos, I., *et al.*: 'Semantic image segmentation with deep convolutional nets and fully connected CRFs', arXiv preprint arXiv:14127062, 2014

[15] Sharma, A., Tuzel, O., Jacobs, D.W.: 'Deep hierarchical parsing for semantic segmentation'. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 530–538

[16] Huang, G.B., Ramesh, M., Berg, T., *et al.*: 'Labeled faces in the wild: a database for studying face recognition in unconstrained environments'. Technical Report, 07-49, University of Massachusetts, Amherst, 2007

[17] Taigman, Y., Yang, M., Ranzato, M., *et al.*: 'Deepface: closing the gap to human-level performance in face verification'. CVPR, 2014, pp. 1701–1708

[18] Schroff, F., Kalenichenko, D., Philbin, J.: 'FaceNet: a unified embedding for face recognition and clustering', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823

[19] Parkhi, O.M., Vedaldi, A., Zisserman, A.: 'Deep face recognition'. BMVC, 2015, p. 6

[20] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, 2012, pp. 1097–1105

[21] Szegedy, C., Vanhoucke, V., Ioffe, S., *et al.*: 'Rethinking the inception architecture for computer vision'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 2818–2826

[22] Iandola, F.N., Han, S., Moskewicz, M.W., *et al.*: 'SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 mb model size', arXiv preprint arXiv:160207360, 2016

[23] Jain, A.K., Li, S.Z.: '*Handbook of face recognition*' (Springer, 2011)

[24] Ghazi, M.M., Ekenel, H.K.: 'A comprehensive analysis of deep learning based representation for face recognition'. Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, 2016, pp. 34–41

[25] Karahan, S., Yildirum, M.K., Kirtac, K., *et al.*: 'How image degradations affect deep CNN-based face recognition?'. 2016 Int. Conf. Biometrics Special Interest Group (BIOSIG), 2016, pp. 1–5

[26] Dodge, S., Karam, L.: 'Understanding how image quality affects deep neural networks'. 2016 Eighth Int. Conf. Quality of Multimedia Experience (QoMEX), 2016, pp. 1–6

[27] Chatfield, K., Simonyan, K., Vedaldi, A., *et al.*: 'Return of the devil in the details: delving deep into convolutional nets', arXiv preprint arXiv:14053531, 2014
[28] Richard-Webster, B., Anthony, S.E., Scheirer, W.J.: 'Psyphy: a psychophysics driven evaluation framework for visual recognition', arXiv preprint arXiv:161106448, 2016
[29] Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., *et al.*: 'Discriminative unsupervised feature learning with convolutional neural networks'. Advances in Neural Information Processing Systems, 2014, pp. 766–774
[30] Zeiler, M.D., Fergus, R.: 'Visualizing and understanding convolutional networks'. European Conf. Computer Vision – ECCV, 2014, pp. 818–833
[31] Lenc, K., Vedaldi, A.: 'Understanding image representations by measuring their equivariance and equivalence'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2015, pp. 991–999
[32] Huang, G.B., Miller, E.L.: 'Labeled faces in the wild: updates and new reporting procedures'. Technical Report, UM-CS-2014-003, 2014
[33] Razavian, A.S., Azizpour, H., Sullivan, J., *et al.*: 'CNN features off-the-shelf: an astounding baseline for recognition'. Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813
[34] Chaib, S., Yao, H., Gu, Y., *et al.*: 'Deep feature extraction and combination for remote sensing image classification based on pre-trained CNN models'. Ninth Int. Conf. Digital Image Processing (ICDIP 2017), 2017, pp. 104203D–104203D
[35] Russakovsky, O., Deng, J., Su, H., *et al.*: 'ImageNet large scale visual recognition challenge', *Int. J. Comput. Vis.*, 2015, **115**, (3), pp. 211–252
[36] Kingma, D.P., Ba, J.L.: 'Adam: a method for stochastic optimization'. Int. Conf. Learning Representation, 2015
[37] Kazemi, V., Sullivan, J.: 'One millisecond face alignment with an ensemble of regression trees'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2014, pp. 1867–1874
[38] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778
[39] Sun, Y., Chen, Y., Wang, X., *et al.*: 'Deep learning face representation by joint identification-verification'. Advances in Neural Information Processing Systems, 2014, pp. 1988–1996
[40] Chen, J.C., Sankaranarayanan, S., Patel, V.M., *et al.*: 'Unconstrained face verification using Fisher vectors computed from frontalized faces'. 2015 IEEE 7th Int. Conf. Biometrics Theory, Applications and Systems (BTAS), 2015, pp. 1–8