Received: 11 June 2020 | Revised: 2 September 2020 | Accepted: 9 September 2020

DOI: 10.1049/bme2.12004

ORIGINAL RESEARCH PAPER

A robust deep learning approach for glasses detection in non-standard facial images

Saddam Bekhet[1] | Hussein Alahmer[2]

[1]Faculty of Commerce, South Valley University, Qena, 83523, Egypt

[2]Al-Balqa Applied University, Al-Salt, Jordan

Correspondence
Faculty of Commerce, South Valley University, Qena, 83523, Egypt.
Email: saddam.bekhet@gmail.com

Abstract

Automated glasses detection is a cardinal component in facial/ocular analysis that powers forensic, surveillance and biometric authentication systems. Throughout literature, glasses detection was always experimented by either utilizing hand-crafted or deep learning features. Nevertheless, in both cases, highly standard face/ocular images were needed to derive the suggested technique. Both working methods performed reasonably well, but the results were bonded to the quality of the facial image and extracted features, where a slight shift and/or rotation in the input face image negatively affects the results. In addition, the obtained performance is even worse on real-world (non-standard) images, especially when compared to recent achievements in other computer vision research areas. In this paper, we present a robust deep learning approach for glasses detection from selfie photos full/partial frontal body non-standard images captured in real-life uncontrolled environments that do not utilize any facial landmarks. To the best of our knowledge this paper is the first to experiment detecting glasses from selfie photos, using a robust deep learning approach. Experimental results on various benchmark facial analysis datasets demonstrated the superior performance of the proposed technique with 96% accuracy.

1 | INTRODUCTION

Identification of human soft biometrics is a hot computer vision (CV) research filed. This is mostly attributed to the increased reliance on surveillance systems that provides vast amounts of visual data to be analysed [1, 2] in an off-line manner. Smartphones are a key player as well, where these soft biometrics can be acquired using a regular smartphone camera. The majority of soft biometrics identification systems are driven by human facial analysis [3], where the human face has the benefit of being a recognizable demographic core attribute [4]. However, the process is not straightforward, where some technical factors, such as image resolution, presence of hands andillumination could severely affect the system's performance.

The presence of eyeglasses is a technical obstruction in facial analysis systems. This is attributed to the generated shadow, reflection and/or occlusion caused by the glasses' frame that covers the eyes (most important part of the face). This leads to inaccurate output of facial/ocular analysis systems, were the appearance of facial images [5] is changed. The situation is even worse with sunglasses, where the entire ocular area is covered causing failure in eye detection and subsequent systems.

However, in recent years many studies have been proposed to reduce the impact of eyeglasses existence on ocular and facial analysis systems [6]. These studies proposed numerous solutions ranging from detecting eyeglasses' presence up to their removal using sophisticated techniques [7]. The situation is more challenging, especially with real-world unconstrained images [8] that depict full/partial frontal facial images with even extra occlusion under variant viewing angles. These images are typically represented by the modern selfie images, which are very difficult to analyse using standard methods. This is attributed to their high variability in resolution, deformation and occlusion, due to the real-life environment they were captured in. Moreover, it is more difficult when it comes to detect glasses in these selfie photos, as they are not taken in ideal/near-ideal capturing conditions, that is, non-controlled lighting conditions and non-textured background. The standard ways that utilize facial landmarks to detect glasses will not succeed, due to the image non-standard nature, where it is difficult to locate any facial landmarks, for example, nose-piece. Figure 1 depicts an illustration of the different ways

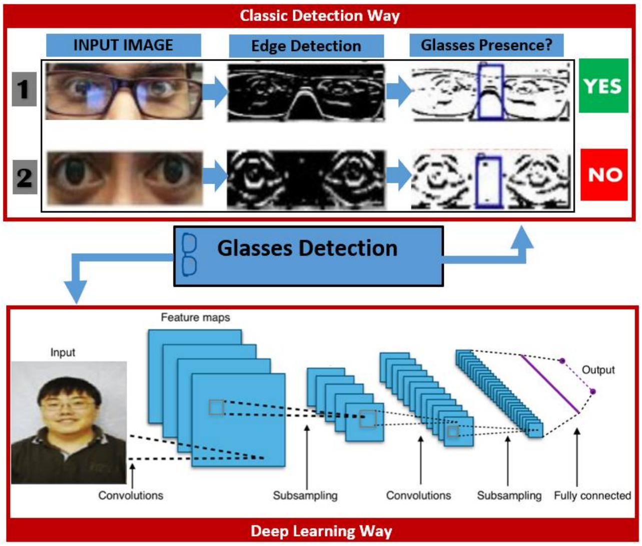

FIGURE 1 Illustration of the different ways to detect glasses. The classic method employs edge detection based on the nose-piece as a famous facial landmark. The deep learning method learns to extract features from input images. Some of the figure parts are adapted from [10, 11]

to detect glasses from facial images. The situation is still unsolvable even with the recent advance in machine learning (ML) approaches, mainly because of ML high sensitivity to the unlimited variations [9] that could be depicted in selfie images.

The contribution of this paper is presenting a robust convolution neural network (CNN) model to effectively detect glasses for the first time from selfie images. The proposed deep learning (DL) model utilizes a transfer learned knowledge from the one-million ImageNet dataset [12], in addition to new features and kernels learned from the selfie images dataset [13]. The utilization of transfer learning in combination with new learned feature maps, compensates the tough uncontrolled nature of selfie images. This enables the network to achieve a higher glasses detection rate. Finally, the reached research results provide a solid baseline for glasses identification from selfie images using DL based approaches, while leaving a room for possible future improvements.

The rest of the paper is organized as follows: Section 2 presents and discuss the related literature work that covers: (1) selfie images and their importance as a new trend in the CV field, and (2)glasses detection recent trends. The proposed CNN model is presented in Section 3. Section 4 presents the experimental part and results. Finally, the paper is concluded in Section 5.

2 | LITERATURE REVIEW

2.1 | Glasses detection

Glasses detection is key problem in CV research, due to its direct relation with facial analysis systems as described earlier. However, there exists a limited research that targets this specific problem. In the past era the problem was commonly approached by localization of the eyes and defining surrounding regions, where glasses are expected and their presence is evaluated using grey-level discontinuities of the frame compared to the face [14]. A combination of edge information (strength and orientation), and geometric features (convexity, symmetry, smoothness, and continuity) were used for eyeglasses detection in [15]. In another research, the task was carried out using edge information within a small area defined between the eyes, that is, nose-piece [15]. Bayes rules were used as well to detect and remove glasses [15]. In the following era, image descriptors were utilized to assist in glasses detection. For example, Local Binary Pattern was used in [10, 16], wavelets in [17], recently HOG in [10] and haar-like features in [18]. Three-dimensional features were also used to detect glasses [19], following their wide effectiveness in describing image landmarks [20]. Recently, using DL based techniques is spreading rapidly through many areas of CV [21]. This is because DL based techniques do not suffer fatigue or moods, thus, they can process huge amounts of data at inconceivable speed outperforming humans in terms of accuracy. For example in [5] Caffe framework [22] was used for glasses detection from iris image dataset. However, this dataset did not depict full faces, but it depicts frontal cropped ocular regions. Later, a two stage CNN was proposed in [6], but, the full experiments were carried out using 23k images for testing and only 1k images for validation. Recently, a different CNN architecture based on adversarial learning was proposed in [23]. This work utilized a separate discriminator and recognizer, where the later can successfully capture the connections among multiple face analysis tasks by sharing feature representations towards a better detection. However, the work utilized standard facial images that does not reflect the real-life current facial images.

Conclusively, key glasses detection literature work is summarized in Table 1. The table data reveals that majority of related work utilized hand-crafted features that were used directly [15, 17] or fused to learning-based methods [10, 16], that is, SVM and AdaBoost or deep learning [5, 18] as well. Although, the reached detection accuracies ranged from 80% to 100%, there are some serious limitations that could be summarized in the following points:

- Some approaches were tested on inhouse datasets that are small in size and typically recorded in ideal conditions [15, 19], moreover they even require the exact location of the eye [18] prior to processing. This led to artificially high lab-accuracy, where such techniques might not achieve the same performance on real-life images.
- Even the utilized larger size datasets does not reflect true variability, as it contains repeated indoor/outdoor sessions for the same person taken for the purpose of experiment, that is, VISOB dataset [24].
- The majority of related work was mostly approached by localizing eyes as a first step and finding the nose-piece as a second step to detect the glasses position [15, 18]. This way requires the facial images to be perfectly standard and frontal, which is not applicable in today's realistic datasets, that is, selfie dataset [13].

TABLE 1 Summary of previous literature on glasses detection

Method	Database	Images	Year	Accuracy (%)
LBP and SVM [16]	LFW	13k	2015	98.65
LNet, ANet [27]	LFW	13k	2015	95
Deep multi-task adversarial learning [23]	LFW	13k	2019	92
Mobilenet CNN [28]	LFW	13k	2019	92
MS cognitive services [29]	LFW	13k	2020	95.8
AdaBoost, haar, Gabor, and SVM [17]	FERET	1k	2004	95.5
Convexity, symmetry, and smoothness with deformable contour [15]	In-house	1k	2000	99.5
Hough transform [19]	In-house	0.5k	2002	80
Cascaded filters, LBP, HOG, SVM, MLP, LDA, and QDA [10]	*FERET + VISOB*	1k + 22k	2017	97.9
Squeezed models of CNNs [25]	*FERET + VISOB*	1k + 120k	2018	94.6
Two stage CNN [6]	*FERET + VISOB*	1k + 120k	2019	100
Haar-like features and AdaBoost [18]	CAS-PEAL-R1	2k	2017	95.11
SVM over Shallow CNN features [30]	NIVE	29k	2017	99.73
Texture, edges and CNN [5]	CASIA-Iris4	20k	2018	99.54

Abbreviation: CNN, Convolution Neural network.

- Most of the training data for ML-based approaches were artificially stamped with eyeglasses [6], as the frame images were aligned for superposition using facial landmarks, that is, eyebrow, eye, ear, and nose [6]. Furthermore, the diversification of real glasses shape, makes the artificial stamping neither accurate nor representative of real training and testing images. Figure 2 depicts sample image with stamped glasses versus real glasses. The images with real glasses are more challenging as they contain natural reflection and shadows.
- For deep learning-based approaches, the majority were either trained on synthetically stamped images [6] or cropped ocular images [6, 25] or iris images [5]. This led to the lab-based 100% accuracy.

In general, the majority of glasses detection related work is still immature in handling fully non-standard images, that is, selfies. This is attributed to the usage of either cropped facial images or ocular images that do not adequately reflect real-world appearance variations (The 120k VISOB [24] dataset is entirely an ocular dataset). Figure 3 depicts sample images from the common glasses detection datasets, that shows their ideal nature that facilitated the reported high detection accuracy. Furthermore, DL based techniques are powerful and suits the job, as it combines feature extraction and classification together in a comprehensive end-to-end model that receives the raw input data and produces the final classification results. Conclusively, there is still much room for work in glasses detection to cover the aforementioned limitations, in terms of using a tailored network architecture and a fully realistic dataset such as the selfie images described in the following section.

2.2 | Selfie images

The emergence of selfies in such enormous volumes granted it a big-data aspect and forced its existence as a new CV research field. Moreover, traditional CV techniques could not effectively handle selfies. This is attributed to two main reasons: (1) their big-data nature, where hand-crafted features are expensive to extract and might not generalize well in such volumes [35], (2) their non-standard capturing way, makes them always prone to extreme occlusion of facial landmarks. Moreover, these images might depict a side view of the face in addition to added emojis or artificial effects, for example, cartoon moustache. Such, problems contribute to the difficulty of glasses detection in selfie images. A group of selfie images that illustrate some of the aforementioned problems are depicted in Figure 4.

From a research perspective, selfie images have been of almost rare usage throughout CV literature. This is because they were more linked to psychology research [36]. From a CV perspective, selfies were studied according to various attributes, that is, senior, youth, Asian, etc., where SIFT [37] and HOG [38] features were employed through SVM to inspect these attributes [13]. However, the reported performance was very poor, that is, <40% accuracy for detecting glasses using SIFT and HOG respectively [13].

Conclusively, selfie images are a new global phenomenon that represents a very sophisticated unique case in CV research, which is worthy of studying and analysing. This will help to unleash the true benefit of such enormous selfie amounts ($\gg$24-billion images [39]). Furthermore, their realistic nature (not being recorded for research purpose) gives them the ultimate diversification and realism for glasses detection work.

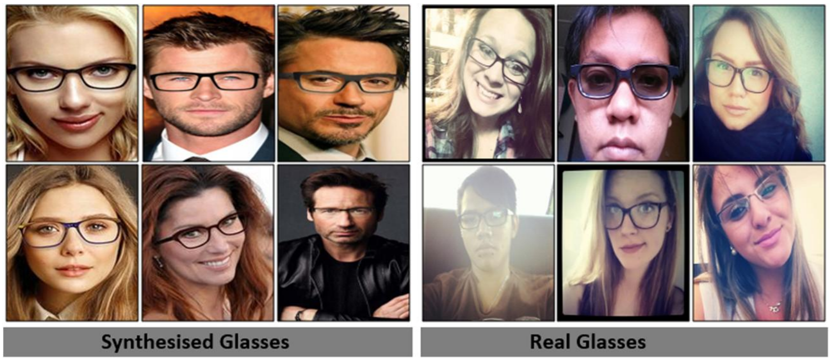

FIGURE 2 Sample images with synthesized [26] and real glasses [13]. The Synthesized glasses are not realistic with no reflection or shadow in the ocular area, and mostly depict similar frames, which eases the task of glasses detection

This could help to improve forensic science and biometric authentication. For biometric authentication, the work is useful in situations where retinal scan is required and the subject is wearing glasses. The system in this case detects the glasses and asks the subject to remove them prior to the scan. This is because glasses act as an obstacle for retinal scan authentication [5]. Regarding the forensic science, it helps in situations where surveillance videos are examined to find a specific suspect, that is, wearing glasses. Figure 4 depicts sample selfie images with/without glasses that reflects the diversification compared to previous datasets depicted in Figure 3.

3 | PROPOSED METHOD

The existence of gigantic labelled data repositories, that is, ImageNet, allowed CNNs to infer and learn rich features representations, which made a huge boost in multiple visual recognition problems [40]. Moreover, for the first time these learned features and representations could be harnessed and transferred to a new problem with some network reshaping. This is often employed when the required consummate amount of sample data needed to train a deep neural network from scratch is simply not sufficient [41]. However, even with the transfer learning approach there are still some major changes and training need to be performed to the original network structure to fit for the new problem. The new training and parametrization could still extend to weeks and months, as the network needs to go through all the training data to learn/remap features and adjust the final layers [41]. Hence, the main objective of this paper is to unleash the power of CNN using transfer learning to better handle selfie images. For this reason we propose the DL architecture illustrated in Figure 5.

The core of DL work is convolutional operations for input images and stacking layers to generate corresponding feature maps. The convolution operation is described as:

$$(X \circledast K)(i,j) = \sum \sum K(m,n)X(i-m,j-n), \qquad (1)$$

where, X is the input image and K is a 2D convolution matrix and $\circledast$ represents the discrete convolution operation. The K matrix slides over the input matrix with stride parameter. The proposed glasses detection CNN architecture is designed based on the famous AlexNet [42] structure. AlexNet is a CNN that was trained over more than million images from the ImageNet database. The original network has eight main layers (5 convolutions + 3 fully connected) and can classify images into

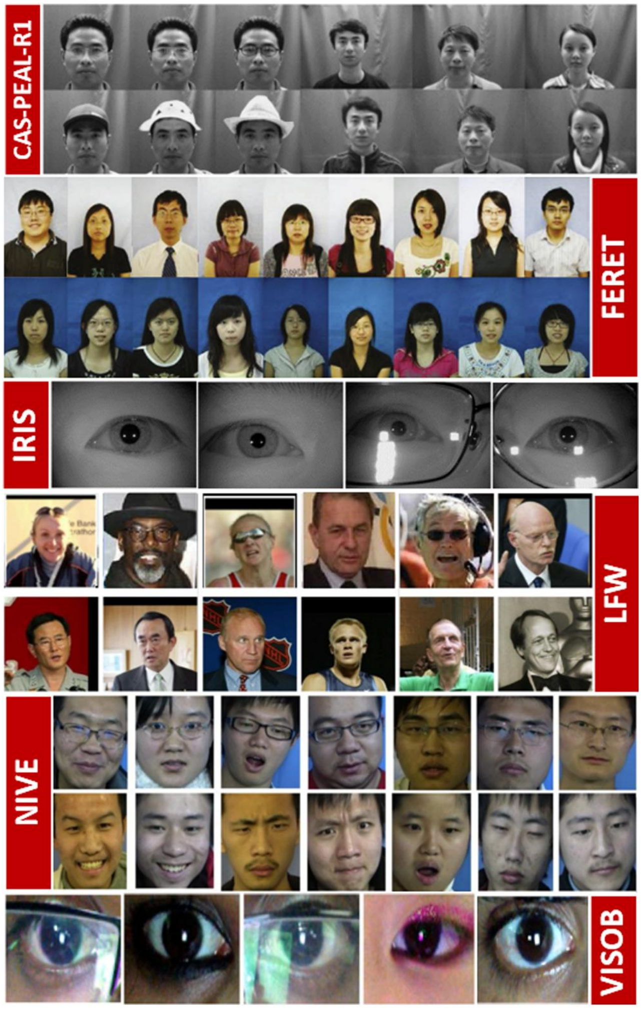

FIGURE 3 Sample facial/ocular images from six common glasses detection datasets, from top to bottom; CAS-PEAL-R1 [31], FERET [11], CASIA-IRIS4 [32], LFW [33], NIVE [34] and VISOB [24]. That images shows full face/ocular images in controlled environment with/without eyeglasses. The photos were taken in textured background, ideal lighting conditions and excellent viewing angles

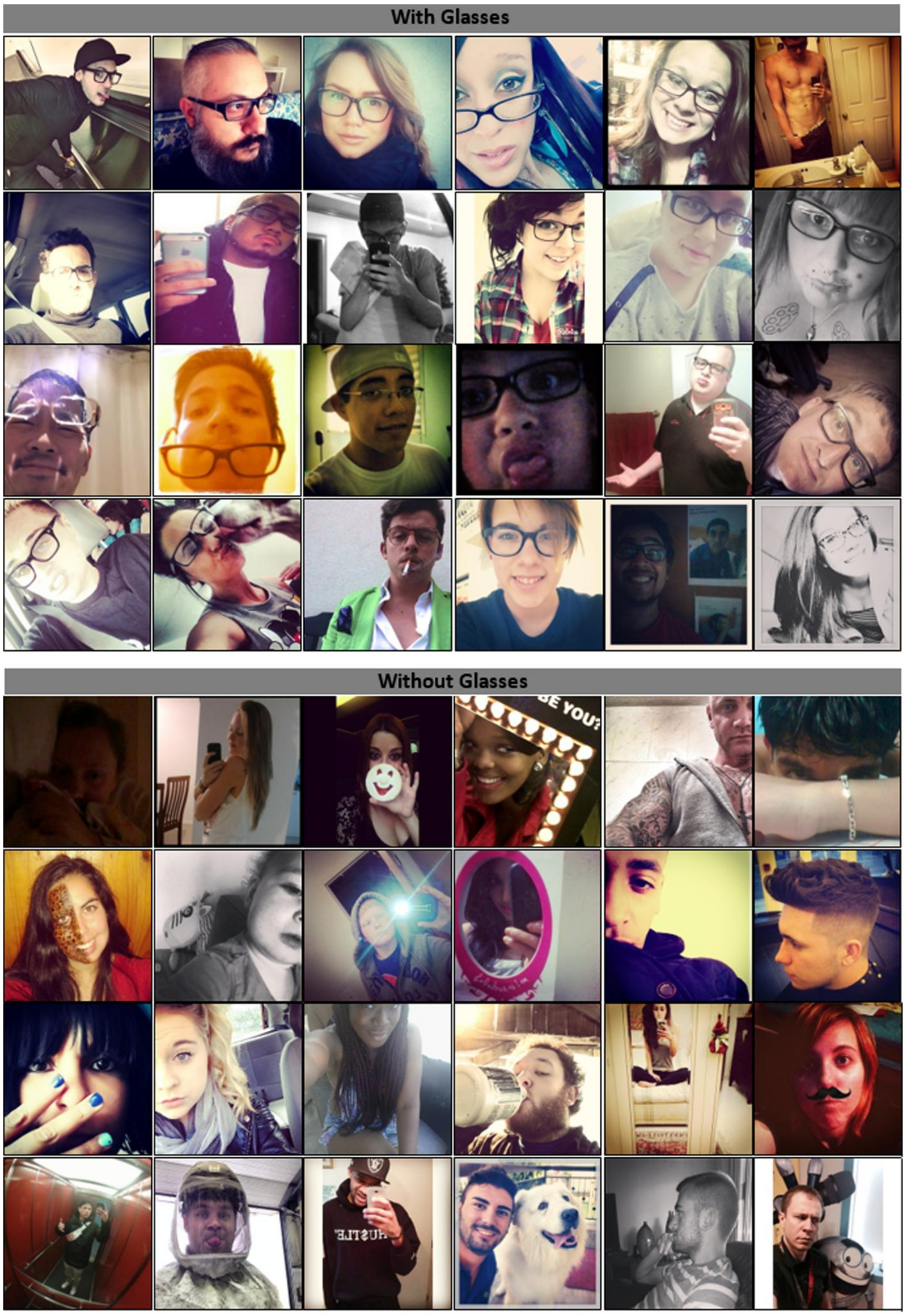

FIGURE 4 Sample selfie images with/without glasses that reflects the dataset problems, for example, occlusion, partial-face/side-face view and artificial effects. The images belong to the selfie dataset [13]

1000 different object categories, such as keyboard, mouse, pencil and many animals. As a result, the network has learned rich feature representations for a wide range of images which makes it a good starting point for the proposed glasses detection CNN.

The transfer learning problem could be mathematically approached by considering a source domain data defines as follows:

$$D_S = \left\{ \left(x_{S_1}, y_{S1}\right), \ldots, \left(x_{Sn_S}, y_{Sn_S}\right) \right\}, \tag{2}$$

where $x_{Si} \in X_S$ is the data instance and $y_{Si} \in Y_S$ is the corresponding class label, and a target domain data as $D_T = \{(x_{T_1}, y_{T1}), \ldots, (x_{Tn_T}, y_{Tn_T})\}$, where $x_{Ti} \in X_T$ is the data instance and $y_{Ti} \in Y_T$ is the corresponding class label. In most cases,

$$0 \leq n_T \ll n_S. \tag{3}$$

where n_S represents the target data which is not available in the same amount as n_T, that is, n_S represents the selfie images data and n_T is the ImageNet data. Transfer learning aims to help

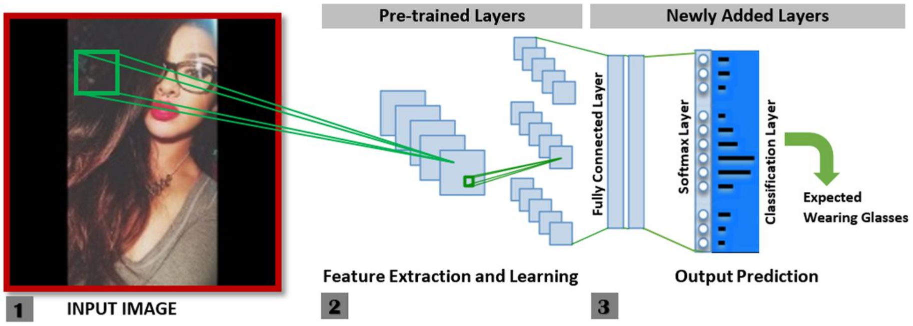

FIGURE 5 Abstract pipeline of the proposed deep learning glasses detection framework. The depicted example is real from the selfie images dataset

improve the learning of a predictive function $f_T(\bullet)$ of the target domain problem D_T using the knowledge from the source domain problem D_S and a learning task T_S. However, neither $D_S \neq D_T$, nor $T_S \neq T_T$. The function $f(\bullet)$ is the objective predictive function that could be learned from the training data pairs $\{x_i, y_i\} \equiv \{\text{feature}, \text{label}\}$, where $x_i \in X$ and $y_i \in Y$. The feature space is represented by X, while the label space is represented by Y.

The original AlexNet CNN has an image input of size 224,×224, which is changed to 227,×227 to fit with the new data image size, and save the resizing step to speed-up training. This is already tackled in the data augmentation step. Moreover, it has an output layer of 1000 softmax-normalized neurons, one for each of the ImageNet [12, 42] object classes. Thus, the last three layers had to be replaced for the proposed glasses detection problem. This was achieved by adding a fully connected layer, a softmax layer and a classification output layer with only two classes, that is, glasses/no-glasses. Such fine-tuning approach allows the network model to pick up the specifics and bias of the selfie images dataset based on the generic features learnt from the first layers. Furthermore, two dropout layers (50% random dropout) were added to counter overfitting because of selfie image data non-big size. Figure 6 depicts the architecture of the proposed glasses detection CNN after adding all of the required layers.

The next section discusses the selfie image dataset specifications and presents the experimental results based on the proposed glasses detection CNN model.

4 | EXPERIMENTS AND RESULTS

This section investigates the performance of the proposed CNN model for glasses detection. The selfie image dataset is introduced in Section 4.1. Section 4.2 highlights the network training phase details, followed by the experimental results and their related analysis in Section 4.3.

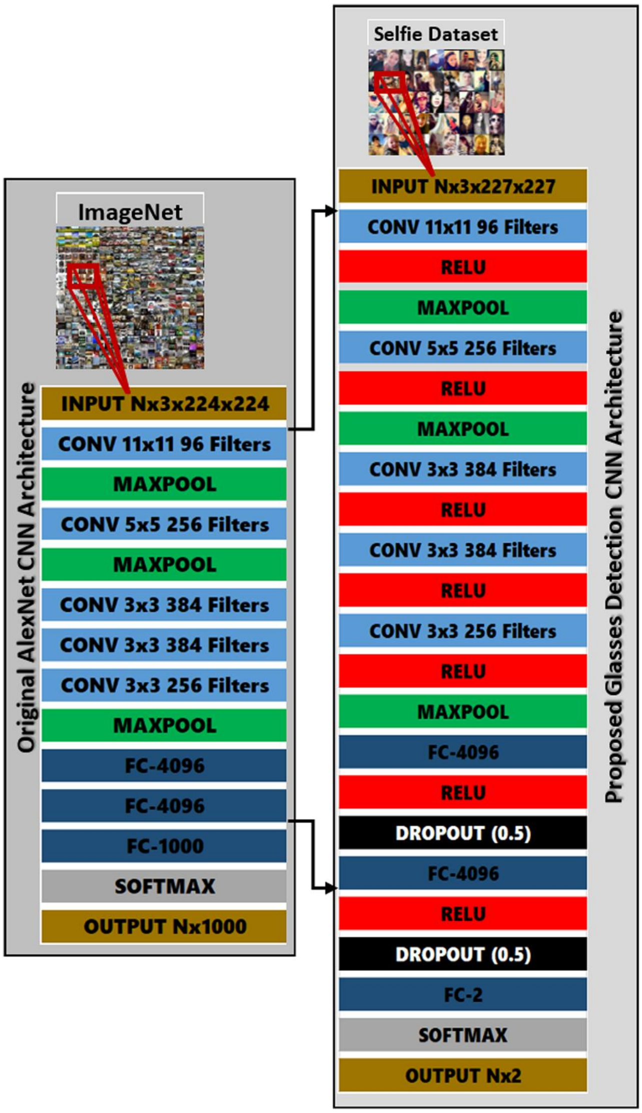

FIGURE 6 The proposed glasses detection CNN architecture, compared to the original AlexNet architecture

4.1 | Datasets

In this paper we had used the first and only selfie images dataset [13, 43] (to the best of our knowledge). The selfie dataset is composed of 46,836 images annotated with 36 different attributes divided into several categories as follows: Accessories: glasses, sunglasses, lipstick, hat, earphone.

Gender: is female. Age: baby, child, teenager, youth, middle age, senior. Race: white, black, Asian. Face shape: oval, round, heart. Facial gestures: smiling, frowning, mouth open, tongue out, duck face. Hair colour: black, blond, brown, red. Hair shape: curly, straight, braid. Misc.: showing cellphone, using mirror, having braces, partial face. Lighting condition: harsh, dim. The important attribute in this dataset for the proposed work is the *glasses/sunglasses* attribute, that is used for the network supervised learning phase. A diverse group of selfie images from this dataset are depicted in Figure 4.

4.2 | Network training phase

The proposed CNN model takes advantage of data augmentation to reduce the effects of overfitting. Before presenting an example image to the network, all dataset images are preprocessed by randomly translating in $(-30, 30)$ range and randomly reflecting the images. The random translation step is necessary to avoid the positional bias in the data. This is because most of the selfie images are tend to be centred. This requires the model to be tested on perfectly centred images as well, which is tackled using this step. The choice of ±30 translation range restricts the effect to ∼10% of the image size, which is common in CNN data training [44, 45]. These preprocessing steps are applied consistently to all images to artificially increase the dataset size using label-preserving transformations [46]. Moreover, to reduce the computational load during the training phase, all of the transformed augmented images are produced from the original ones at runtime without storing them on disk.

The model was trained using a stochastic gradient descent with a batch size of 10 examples, momentum of 0.9 and weight decay of 0.0004. The small value of weight decay is important for the model to facilitate learning, as it helps to reduce the model's training error. Furthermore, the weights of the early layers are preserved from changing, as they had learned the abstract features from the ImageNet dataset. The training and results were obtained using an Intel Core i7, 3.3 GHZ with 16 GB of RAM. The training time extended for a whole 15 days to iterate through all the training data (original + augmented = ∼100k) and fine-tune the network parameters based on the new data. The next section presents and discusses the network performance for the targeted problem.

4.3 | Results and discussion

Following the common experimental setup, the dataset was randomly splitted by assigning 70% of the images to training and 30% to validation (unseen by the network). The split in this case does not affect the results, since there is no established test-set split for the selfie dataset. Regarding the quantitative evaluation, the accuracy measure [47, 48] (Equation 4) is used as it reports the percent of correctly identified images with/without glasses with respect to the whole dataset. Additionally, false rejection ratio (FRR), false acceptance ratio (FAR), and equal error rate (EER) metrics are also utilized, as they are commonly used for biometric application system evaluation [49].

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} I_k(\hat{y}_i)}{\sum_{i=1}^{N} I_k(y_i)}, \tag{4}$$

where K is the number of test set categories, N is the number of testing samples. $I_k(y)$ is an indicator function evaluates to one when $y = k, I_k(y) = 1$, otherwise evaluates to 0. y_i and $\hat{y}_i$ are the true label and predicted label of the i^{th} sample respectively.

The validation loss is also used to provide an extra measure about the model performance, as it indicates how well the model is generalizing to unseen data Equation 5 depicts the loss function. Where $\hat{y}_i$ is the network prediction with the ground truth values y_i and λ is the individual loss function, that is, log-loss in the proposed model.

$$J = \sum_{i=1}^{N} \lambda\left(\hat{y}_i, y_i\right). \tag{5}$$

The network achieved 96% accuracy on the 46k selfie dataset with 0.15 log-loss. Furthermore, Table 2 depicts the standard biometric metric values of the proposed network model. In general, the results are very good considering the challenging realistic nature of the selfie dataset. This result proves that the features learned from the ImageNet dataset are generic enough to generalize to the selfie dataset. However, a high percentage of this accuracy is attributed to the extra feature-maps learned during the network training phase, that enriched the final classification phase.

The proposed CNN model performance is also compared against four baselines to emphasis its effectiveness. The

TABLE 2 Performance of the proposed CNN glasses detection model using standard biometric application metrics

FRR (%)	FAR (%)	EER (%)
8.3	5.5	6.9

Abbreviations: EER, Equal Error Rate; FAR, False Acceptance Ratio; FRR, False Rejection Ratio.

TABLE 3 Performance comparison of the proposed CNN glasses detection model against hand crafted based and CNN-based baselines

Method	Accuracy (%)
Proposed CNN model	**96↑**
Two stage CNN [6]	85.4
Google teachable machine [50]	85
Dense SIFT. VLAD encoding. 256 words Codebook + SVM with linear kernel [13]	37
Dense HOG. VLAD encoding. 256 words Codebook + SVM with linear kernel [13]	35

Note: Higher values are marked in bold.

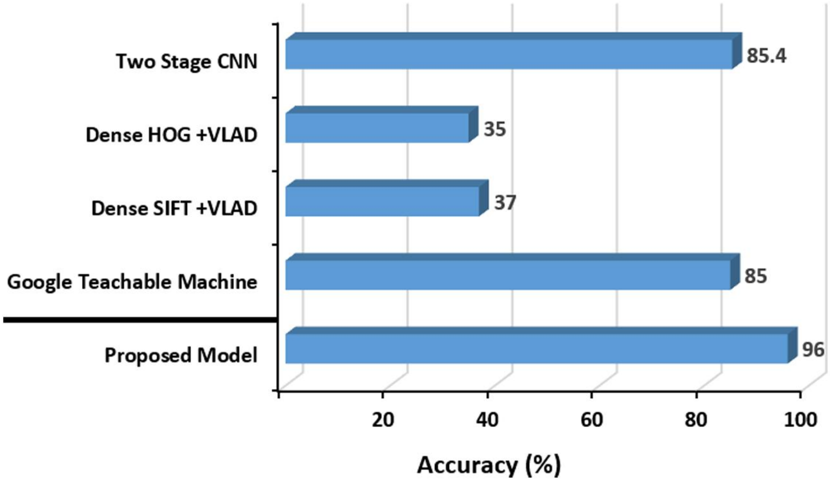

FIGURE 7 Performance comparison of the proposed CNN glasses detection model against recent baselines

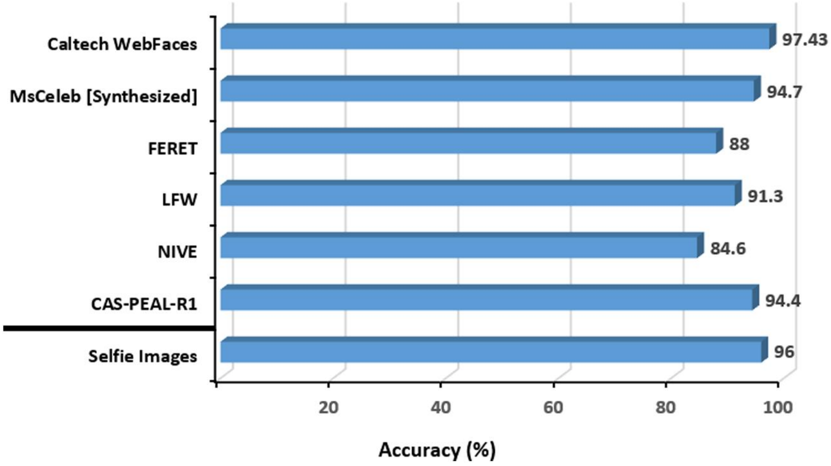

FIGURE 8 Performance of the proposed model on six additional datasets. The selfie dataset is added to the figure for illustration purpose

selected baselines represent the current research themes in glasses detection. Namely hand-crafted features and CNN-based approaches. The hand-crafted approaches are Dense SIFT [13] and Dense HOG3D [13]. For, CNN-based approaches they are represented by the state of art Google Teachable Machine [50] that is based on TensorFlow implementation and the two stage CNN approach [6]. The benchmarking results depicted in Table 3 and visualized in Figure 7 emphasizes the robust performance of the proposed glasses detection CNN model, as it outperformed the hand crafted based approaches with 60 ± 1.4% and 10.8 ± 0.2% for the CNN-based approach.

Regarding the two stage CNN baseline [6], the reported 100% glasses detection accuracy, was obtained based on the VISOB dataset [24]. This dataset only consists of ocular images, that is, not full facial images. The testing data were generated by digitally stamping eyeglasses on the dataset images. The testing and validation were experimented using 23,826 and 1191 images respectively (5% of the dataset for validation). However, in the proposed work, the entire selfie dataset (32,785 images for training and 14,050 images for validation) was used. This is considered 87.22% (without data augmentation) increase in training and validation data compared to [6], which highly contributes towards more robust result. In addition, the proposed work did not utilize any images with digitally stamped frames, as the selfie dataset are entirely realistic.

Furthermore, the performance of the proposed network model was validated on six common facial analysis benchmark datasets. The first dataset, that is, CAS-PEAL-R1 [31] is a Chinese face dataset, composed of 99,594 images of 1040 individuals (595 males and 445 females). The dataset was constructed using nine cameras that were mounted horizontally on an arc arm to simultaneously capture images across different poses. The second, that is, FERET [11] is a standard face recognition dataset composed of 14,126 images that includes 1199 individuals and 365 duplicate sets of images. A duplicate set is a second set of images of a person already in the database and was usually taken on a different day. The third, that is, NIVE [34], is a face expression analysis dataset, collected from 215 test subjects that were captured under different facial expressions. The fourth, that is, LFW [33] is a public face verification benchmark composed of 13,233 images for 5749 people. The fifth, that is, Caltech [51] contains a total of 10,524 faces in 7092 images collected from the web. All of these five datasets depict subjects with glasses, that include dark frame glasses, glasses without frame, and

TABLE 4 The layers and layer parameters of the proposed glasses detection network

Layer type	Output shape	Number of trainable parameters
Conv2D	[217 × 217 × 96]	34,944
Max pooling	[72 × 72 × 69]	0
Conv2D	[68 × 68 × 265]	614,656
Max pooling	[22 × 22 × 256]	0
Conv2D	[20 × 20 × 384]	885,120
Conv2D	[18 × 18 × 384]	1,327,488
Conv2D	[16 × 16 × 256]	884,992
Max pooling	[5 × 5 × 256]	0
Flatten	[2048]	0
Dense	[4096]	26,218,496
Dropout	[4096]	0
Dense	[4096]	16,781,312
Dropout	[4096]	0
Dense	[2]	8194
Total parameters		467,552,02

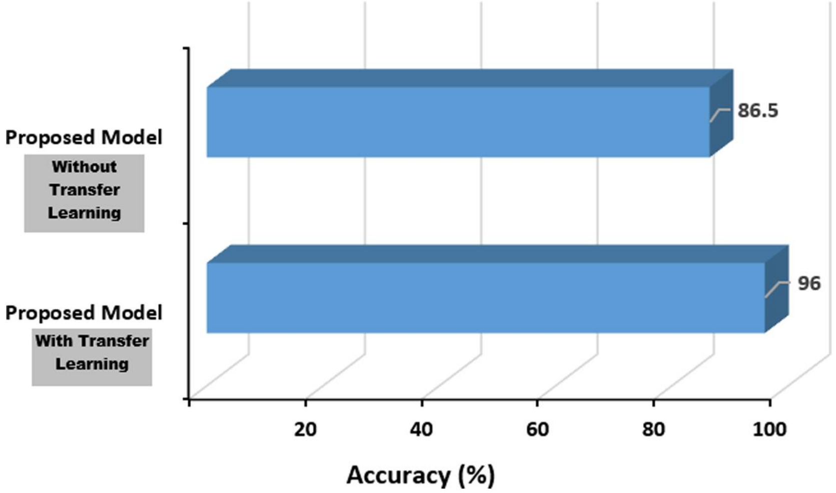

FIGURE 9 Performance of the proposed glasses detection model on selfie dataset with and without transfer learning

sunglasses, which is the main concern of the proposed work. The final dataset is a fully synthesized version [52] of the famous MS-Celeb [53] dataset (47,917 image), that was virtually stamped with thick black-framed eyeglasses. This dataset is very challenging as it optimizes the intra-variations caused by eyeglasses [52].

The proposed model glasses detection accuracy(%) performance on the six aforementioned public datasets is shown in Figure 8. The figure reflects an average performance of 91.7 ± 4.7% over FERET, LFW, NIVE, CAS-PEAL-RI, synthesized MS-Celeb and Caltech WebFaces datasets. In addition, to further confirm the robustness of the proposed CNN model a large diverse group of selfie images were collected from the Internet and were tested on the proposed CNN. The system achieved an accuracy of 97.05%, which consolidates the previous result.

Moreover, to emphasize the benefit of the transfer learning, the full CNN network depicted in Figure 6 was reset and fully trained from scratch on the selfie dataset only. The layers' implementation details are given in Table 4. After a full cycle of training epochs and following the same data setup proposed earlier; the network achieved a detection accuracy of 86.5% on the selfie dataset. Figure 9 shows a 9.5% accuracy less compared to the version that relies on transfer learning. This quantifies the benefit of the transferred knowledge from the ImageNet. However, this result is not bad considering the size of the selfie dataset which is only 4.6% of the ImageNet dataset.

Towards a deeper look into the network learning phase, Figure 10 illustrates the evolving of convolutional kernels throughout the convolution layers till the last image classes learnt by the fully connected layer. The figure shows that the

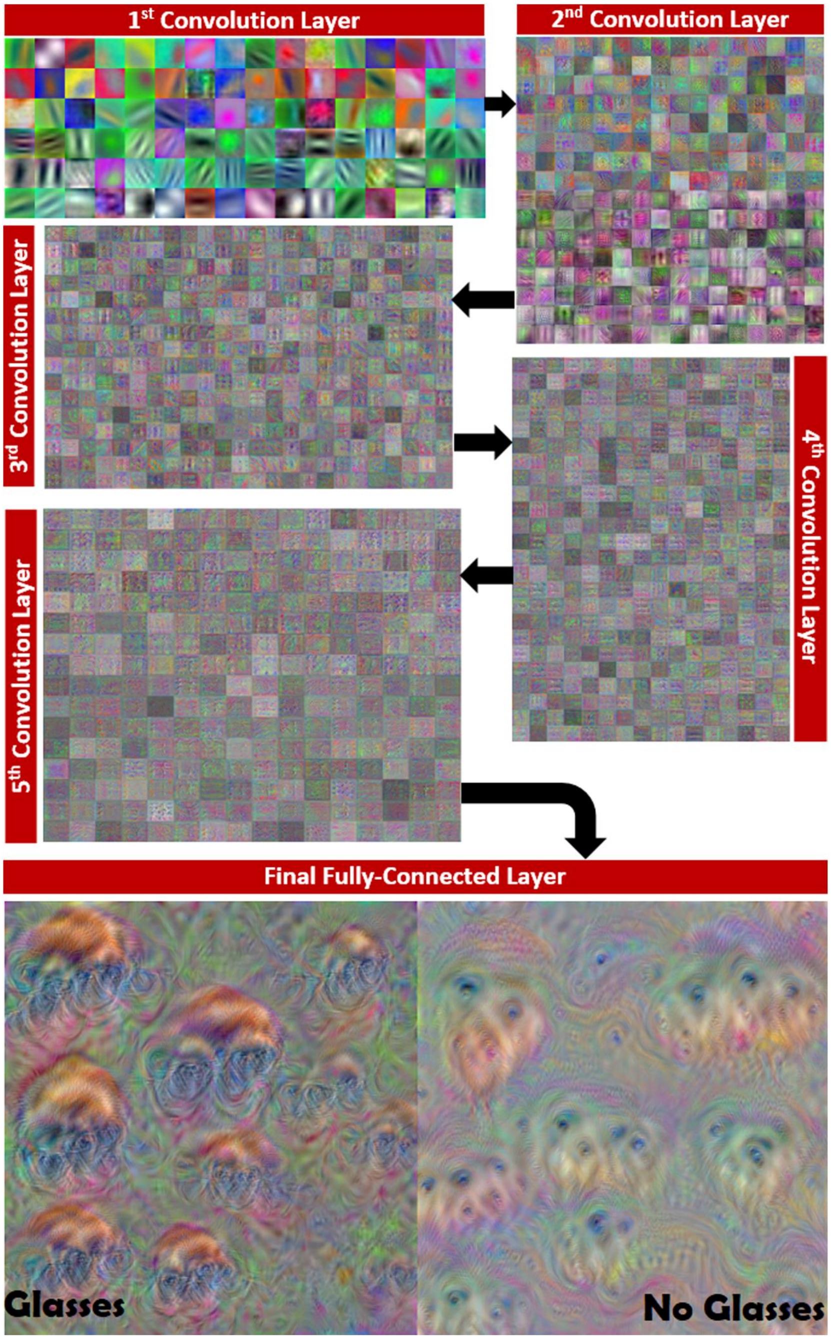

FIGURE 10 Evolution of the convolutional kernels through the glasses detection CNN until the final fully connected layer. conv1 = 96 kernels, conv2 = 256, conv3 = 384, conv4 = 384 and conv5 = 256. The final fully connected layer depicts images that resemble the most closely image class, that is with/without glasses. The glasses shape is clearly depicted in the final layer image class (left image)

network has learned the different components of images like edges and colour blobs, in addition to a group of frequency and orientation filters, where the glasses shape is clearly depicted in the final layer image class. Figure 11 shows activations from the last convolutional layer during the forward-pass based on the input image depicted in the same figure, where the bright areas in the activation channels corresponds to the activation based on the presence of glasses.

Finally, for the qualitative performance of the proposed DL model, Figure 12 depicts a group of challenging selfie images that were classified based on their glasses attribute using the proposed CNN model. The results reflect the classification power of the network, as it has learnt a variety of useful features to identify glasses in such selfie images. A clear example that depicts the developed CNN power is the example image indexed at 4 × 4 (row × column), where the subject is NOT wearing glasses and places it over her head and the network correctly identifies this image as not wearing glasses. However, there are some cases were the network failed to detect the glasses. For example the image in Figure 12, indexed at location 6×2 was mistakenly classified as wearing glasses. This was attributed to the magnifier that covers the eyes, that was mistakenly identified as glasses. Furthermore, the image indexed at location 6×1, was also mistakenly classified because of the small subject to scene ratio, that is the subject is very small in the image.

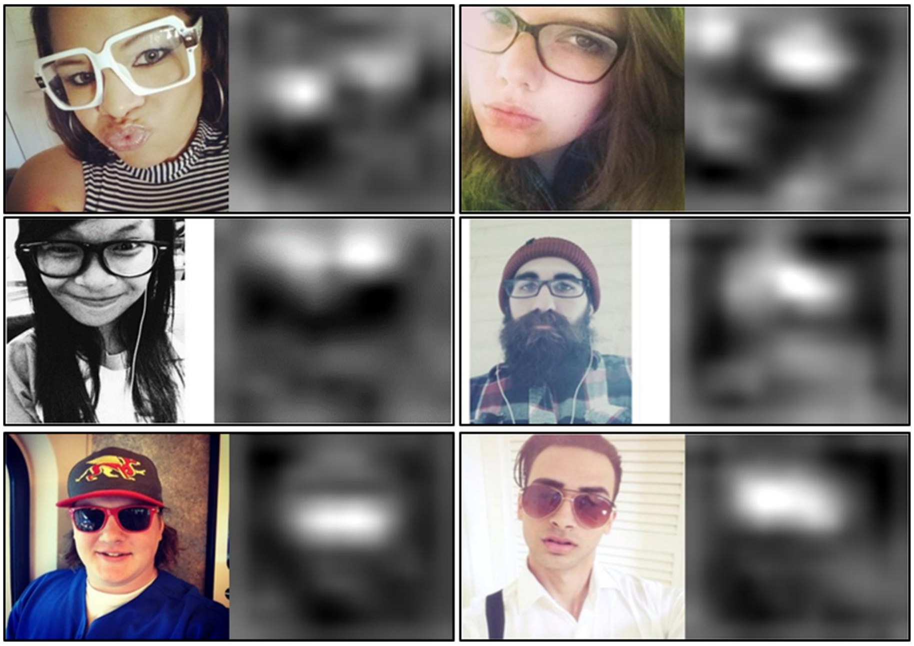

FIGURE 11 Sample input selfie image and their corresponding activation channels from the last convolution layer. The bright areas in the activation channels corresponds to the activation based on the glasses presence

5 | CONCLUSION

This paper presented an effective CNN model for glasses detection from selfie images. Distinct from previous work, this paper utilizes realistic dataset, that is selfie dataset, with non-synthesized glasses and challenging frontal faces with full/partial body. Selfie images is a highly challenging dataset (46k) that was almost of rare usage, due to its unprecedented high variability uncontrolled nature (photos taken by normal users for themselves any time/where). The proposed CNN model achieved 96% accuracy. In order to reach such result, the proposed model was built with transferred knowledge from the ImageNet 1.2 million dataset, this allows raw-data abstraction that expands the analysis to unseen data yet. However, even with such transfer learning the network had to have extra layers and go through full epochs (full training cycle on the whole training data) that extended for almost two weeks. After such extensive training the network had learnt a variety of different image components, that is, edges and colour blobs which are important in detecting the existence of glasses in the input image.

The results presented within this paper are sufficient to the targeted problem. However, there is still room to improve the achieved accuracy through implementation other CNN models or even combining the proposed CNN with a long-short term memory towards a better result.

ORCID

Saddam Bekhet https://orcid.org/0000-0002-3028-6500

REFERENCES

1. Reid, D.A., et al.: Soft biometrics for surveillance: an overview. In: Handbook of statistics, vol. 31, pp. 327–352. Elsevier (2013)
2. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis. 126(2–4), 144–57 (2018)
3. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans Circ Syst Video Technol. 14(1), 4–20 (2004)
4. Golomb, B.A., Lawrence, D.T., Sejnowski, T.J.: Sexnet: a neural network identifies sex from human faces. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, vol. 1, pp. 2 (1990)
5. Drozdowski, P., et al.: Detection of glasses in near-infrared ocular images. In: 2018 International Conference on Biometrics (ICB). pp. 202–208. IEEE (2018)
6. Mohammad, AS., Rattani, A., Derakhshani, R.: Eyebrows and eyeglasses as soft biometrics using deep learning. IET Biometrics (2019)
7. Park, J.S., et al.: Glasses removal from facial image using recursive error compensation. IEEE Trans Pattern Anal Mach Intell. 27(5), 805–11 (2005)
8. Chen, D.Y., Lin, K.Y.: Robust gender recognition for uncontrolled environment of real-life images. IEEE Trans Consum Electron. 56(3) (2010)
9. Rodríguez, P., et al.: Age and gender recognition in the wild with deep attention. Pattern Recogn. 72, 563–71 (2017)
10. Mohammad, A.S., Rattani, A., Derahkshani, R.: Eyeglasses detection based on learning and non-learning based classification schemes. In: 2017 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–5. IEEE (2017)
11. Li, B., Lian, X.C., Lu, B.L.: Gender classification by combining clothing, hair and facial component classifiers. Neurocomputing. 76(1), 18–27 (2012)
12. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int J Comput Vis. 115(3), 211–52 (2015)
13. Kalayeh, M.M., et al.: How to take a good selfie? In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 923–926. ACM (2015)
14. Jiang, X., et al.: Towards detection of glasses in facial images. Pattern Anal Appl. 3(1), 9–18 (2000)
15. Jing, Z., Mariani, R.: Glasses detection and extraction by deformable contour. In: Proceedings 15th International Conference on Pattern Recognition, vol. 2, pp. 933–936. ICPR-2000, Barcelona (2000). https://doi.org/ 10.1109/ICPR.2000.906227

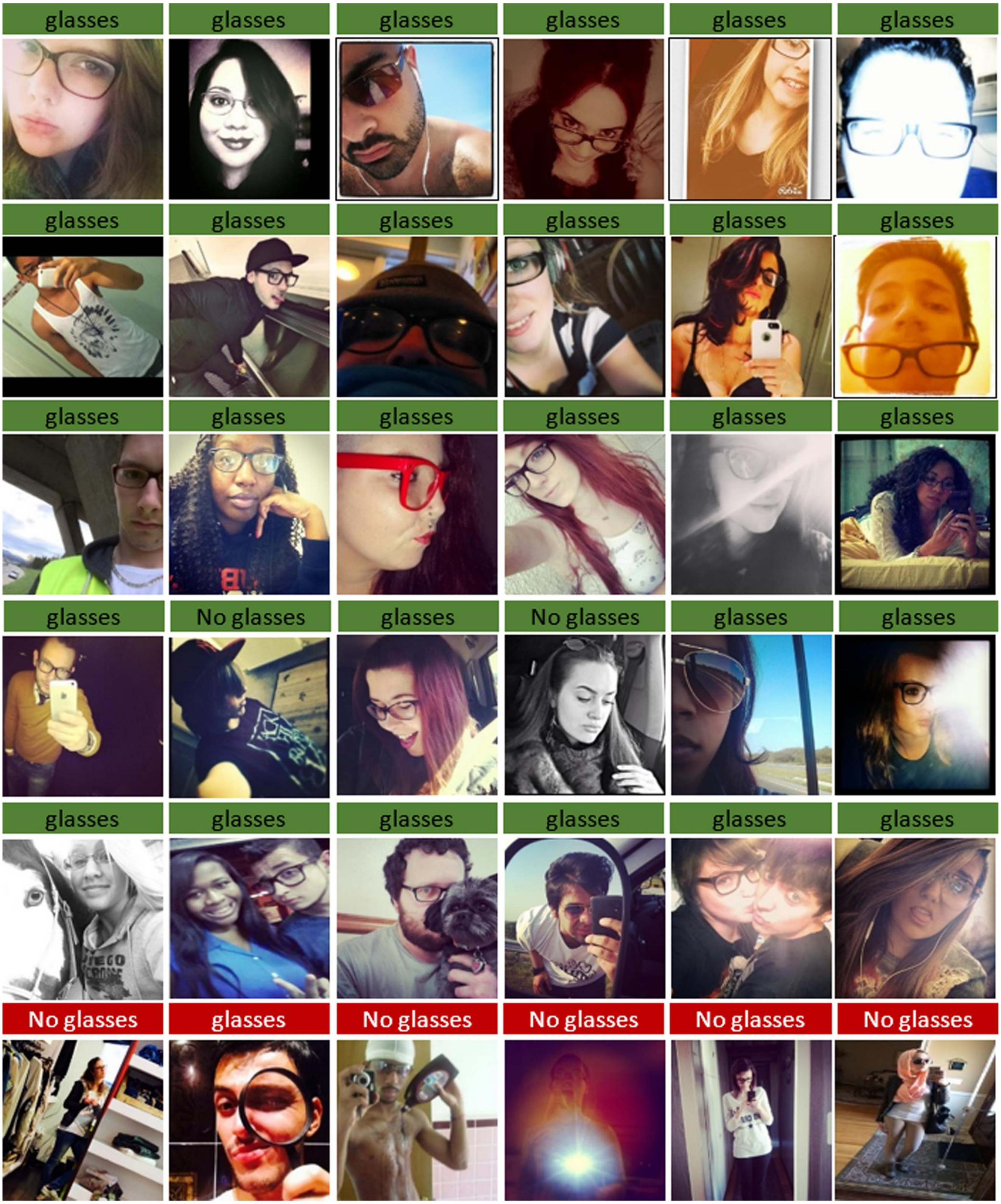

FIGURE 12 Sample images from the selfie dataset that were classified using the proposed deep learning model. Right predications are tagged with a green label and wrong predictions are tagged with a red label

16. Fernández, A., et al.: Glasses detection on real images based on robust alignment. Machine Vision and Applications. 26(4), 519–31 (2015)
17. Wu, B., Ai, H., Liu, R.: Glasses detection by boosting simple wavelet features. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 1, pp. 292–295. IEEE (2004)
18. Du, S., et al.: Precise glasses detection algorithm for face with in-plane rotation. Multimed Syst. 23(3), 293–302 (2017)
19. Wu, H., et al.: Glasses frame detection with 3d hough transform. In: Object recognition supported by user interaction for service robots, vol. 2, pp. 346–349. IEEE (2002)
20. Knopp, J., et al.: European Conference on computer vision. In: Hough transform and 3d surf for robust three dimensional classification, pp. 589–602. Springer (2010)
21. Bekhet, S., Ahmed, A.: An integrated signature-based framework for efficient visual similarity detection and measurement in video shots. ACM Trans Inf Syst. 36(4), 37 (2018)
22. Jia, Y., et al.: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678 (2014)
23. Wang, S., et al.: Multiple face analyses through adversarial learning. arXiv preprint arXiv:1911.07846 (2019)
24. Rattani, A., et al.: Icip 2016 competition on mobile ocular biometric recognition. In: IEEE International Conference on image processing (ICIP), pp. 320–324. IEEE (2016)
25. Mohammad, AS., Rattani, A., Derakhshani, R.: Comparison of squeezed convolutional neural network models for eyeglasses detection in mobile environment. Journal of Computing Sciences in Colleges. 33(5), 136–144 (2018)
26. 'Pinterest'. https://www.pinterest.com (2020)
27. Liu, Z., et al.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp. 3730–3738. (2015)
28. Vasileiadis, M., Stavropoulos, G., Tzovaras, D.: Facial soft biometrics detection on low power devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019)
29. Microsoft azure cognitive services. https://azure.microsoft.com/en-us/services/cognitive-services/. Accessed Aug (2020)
30. Basbrain, A.M., et al.: Shallow convolutional neural network for eyeglasses detection in facial images. In: 2017 9th Computer Science and Electronic Engineering (CEEC). (2017). pp. 157–161
31. Gao, W., et al.: The cas-peal large-scale Chinese face database and baseline evaluations. IEEE Trans Syst Man Cybern Syst Hum. 38(1), 149–61 (2007)
32. Center for Biometrics and Chinese Academy of Sciences' Institute of Automation. 'Casia iris image database'. http://biometrics.idealtest.org (2019). Accessed July 2019
33. Huang, G.B., et al.: Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. Technical Report (2008)
34. Wang, S., et al.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimed. 12(7), 682–91 (2010)
35. Mohamed, O., Mohammed, O., Brahim, A.: Content-based image retrieval using convolutional neural networks. In: First International Conference on Real Time Intelligent Systems, pp. 463–476. Springer, Cham (2017)
36. Schneider, T.M., Carbon, C.C.: Taking the perfect selfie: Investigating the impact of perspective on the perception of higher cognitive variables. Front Psychol. 8, 971 (2017)
37. Lowe, D.G.: Object recognition from local scale-invariant features. In: The proceedings of the seventh IEEE international conference on Computer vision, vol. 2, pp. 1150–1157. IEEE (1999)
38. Dalal, N., Triggs, B: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
39. '24 billion selfie images'. https://www.huffpost.com/entry/24-billion-photos-prove-our-selfie-obsession (2020). Accessed Aug 2020
40. Google tensorflow image recognition'. https://www.tensorflow.org/tutorials/image_recognition Accessed Aug 2020
41. Hussain, M., Bird, J.J., Faria, D.R.: A study on cnn transfer learning for image classification. In: UK Workshop on computational Intelligence, pp. 191–202. Springer (2018)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
43. Selfie images dataset'. https://www.crcv.ucf.edu/data/Selfie/. Accessed Aug 2020
44. Engstrom, L., et al.: Exploring the landscape of spatial robustness In: International Conference on Machine Learning, pp. 1802–1811 (2019)
45. Roelofs, R.: Measuring Generalization and overfitting in Machine learning. UC Berkeley (20192012)
46. Cireșan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. arXiv preprint arXiv:12022745, (2012)
47. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Nat Lang Eng. 16(1), 100–103 (2010)
48. Bekhet, S., Ahmed, A.: 'Evaluation of similarity measures for video retrieval'. Multimed Tool Appl (2019). https://doi.org/10.1007/s11042-019-08539-4
49. Bolle, R.M., et al.: Guide to biometrics. Springer Science & Business Media (2013)
50. 'Google teachable machine'. https://teachablemachine.withgoogle.com/. Accessed Aug 2020
51. Angelova, A., Abu-Mostafam, Y., Perona, P. Pruning training sets for learning of object categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1, pp. 494–501. IEEE (2005)
52. Guo, J., et al.: Face synthesis for eyeglass-robust face recognition. In: Chinese Conference on biometric recognition, pp. 275–84. Springer (2018)
53. Yi, R., et al.: Faces as lighting probes via unsupervised deep highlight extraction. In: Proceedings of the European Conference on computer vision (ECCV), pp. 317–33 (2018)

How to cite this article: Bekhet S, Alahmer H. A robust deep learning approach for glasses detection in non-standard facial images. *IET Biom*. 2021;10:74–86. https://doi.org/10.1049/bme2.12004