SmileNet: Registration-Free Smiling Face Detection In The Wild

Youngkyoon Jang[†‡], Hatice Gunes[‡], Ioannis Patras[†]
[†]Queen Mary University of London, London, UK
[‡]University of Cambridge, Cambridge, UK
{youngkyoon.jang, i.patras}@qmul.ac.uk, {yj293, Hatice.Gunes}@cl.cam.ac.uk

Abstract

*We present a novel smiling face detection framework called SmileNet for detecting faces and recognising smiles in the wild. SmileNet uses a Fully Convolutional Neural Network (FCNN) to detect multiple smiling faces in a given image of varying resolution. Our contributions are threefold: 1) SmileNet is the first smiling face detection network that does not require pre-processing such as face detection and registration in advance to generate a normalised (cropped and aligned) input image; 2) the proposed SmileNet is a simple and single FCNN architecture simultaneously performing face detection and smile recognition, which are conventionally treated as separate consecutive pipelines; and 3) SmileNet ensures real-time processing speed (*21.15 *FPS) even when detecting multiple smiling faces in a given image (*300×300*). Experimental results show that SmileNet can deliver state-of-the-art performance (*95.76%*), even under occlusions, and variances of pose, scale, and illumination.*

1. Introduction

Among the basic emotions, happiness is the most universally recognised emotion that represents a positive state such as joy or satisfaction [5, 6]. Therefore, an increasing number of studies have focused on smile detection instead of the recognition of the six basic emotions of happiness, sadness, surprise, anger, fear and disgust. Research on smile detection is mainly divided into two categories, using learned convolution filters [3, 11, 31] based on deep learning, or using hand-crafted visual features [10, 13, 14, 21]. In the best performing methods, the extracted features are combined with deep learning (CNN) classifiers [11, 31] or other machine learning techniques, such as Support Vector Machine (SVM) [10, 14], AdaBoost [3, 21], and Extreme Learning Machine [2].

Although there are many existing works on facial expression recognition and smile detection, the following challenges remain unaddressed:

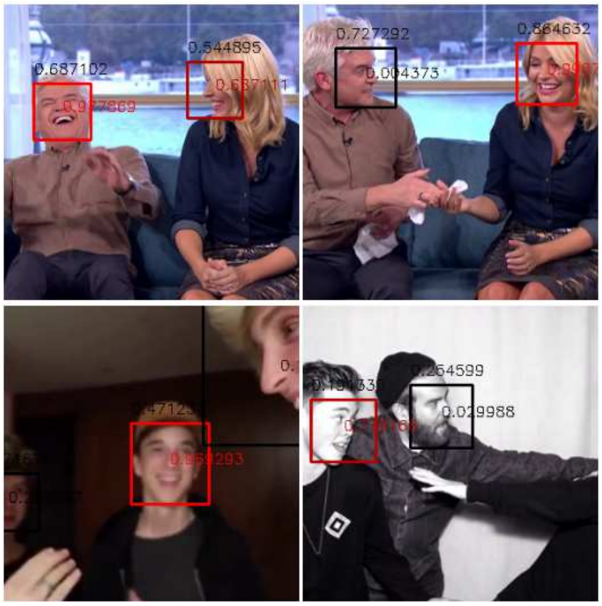

Figure 1. Our system, which we refer to here as SmileNet, detects faces and recognises smiles in the wild. When detected faces are determined as smiling faces, the black bounding box colour changes to red. The probability that appears at the top of the box indicates the face confidence score, and the one appearing in the middle of the box is about smiles. The intensity of red corresponds to the level of confidence.

Unconstrained conditions: When analysing human behaviour in the wild, computer vision algorithms are challenged by variances in pose, scale and illumination.

Real-time performance: Existing methods cannot guarantee real-time performance because they require time-consuming preprocessing steps such as face detection and registration before performing smile recognition.

Addressing the above mentioned challenges, we propose SmileNet (see Fig. 1), which performs simultaneous face detection and smile recognition in a single architecture. SmileNet aims not only to detect faces in a given colour image, but also to estimate the confidence score of a smiling face associated with a detected face. To achieve this, the proposed SmileNet first inherits the feature extraction capability of the pre-trained VGG16 network [22] by utilising the parameters of this network. Following this step, we add multi-scaled convolution layers at the end of the base structure of the VGG16 network (convolution layers) to perform both classification (face and smile classification in pixels) and regression (boundary box localisation) tasks. To the best of our knowledge, SmileNet is the first smiling face detection network that can process naturally captured images without a pre-normalisation step which causes processing time latency.

After building the proposed architecture, we train SmileNet on well-known datasets for face detection (AFLW [12]) and smile (CelebA [16]) recognition. Utilising the properties of the SmileNet design discussed in Sec. 3.1.1, we obtain a set of matched default boxes as originally proposed by Liu et al. [15]. Then, we train SmileNet by recursively calculating losses (including face classification, bounding box regression, and smile classification) and updating parameters in the network. During training, we randomly apply one of the data augmentation strategies such as shrinking, cropping and gamma correction. Finally, we perform quantitative evaluation on the CelebA and GENKI-4K datasets, which contain images captured in the wild.

The main contributions of our work are three-fold:
1) Unconstrained processing: SmileNet takes the original image captured in the wild as an input to produce smiling face detection results. It does not require a pre-normalisation step including face detection and registration.
2) Simple multi-task learning: SmileNet is able to learn both face detection (including face classification and bounding box regression) and smile recognition in a single architecture, while achieving the state-of-the-art results.
3) Real-time processing: SmileNet has the scalability to process more face-related tasks without requiring additional processing time.

2. Related Work

General methods on facial affect recognition. Sariyanidi et al. [20] discusses the-state-of-the-art methods for face registration, representation, dimensionality reduction and recognition, which are the common components of a generic pipeline for performing automatic facial affect analysis. Depending on the target application, the generic pipeline might have to be changed to some degree. Nonetheless, the first two steps of face localisation and 2D / 3D registration steps have been necessary for most of the face analysis tasks such as face and gender recognition, age prediction, and head pose estimation. See [25, 28, 30] for details.

Smile detection. Despite significant technological advances in the field of affective computing, automatic facial expression recognition still faces major challenges caused by occlusions and variances of head pose, scale, and illumination. These challenges are the main reason why every state-of-the-art approach to smile detection (see Table 2) requires a pre-normalisation step involving face detection and registration (rotation, scaling, and 2D/3D transformation). Some of the previous methods manually process the input image to detect the face (when an automatic detector fails) [13] or register the face based on the eye positions [21]. Prior studies that do not provide the details of the methods they utilise for face detection [21, 31] and registration [3, 10, 13] remain questionable in terms of manual interventions to the input image.

Approaches without pre-normalisation. In the field of affective computing, there exist works that process the original input image without pre-normalisation steps. Liu et al. [16] combines LNet localising a face and ANet to predict facial attributes including smiles. However, they use EdgeBox [32] that proposes a number of candidate windows to determine the final facial region among the multiple predicted positions scattered by LNet. Before feeding the output of LNet to ANet, this process for narrowing the potential face region is performed several times through several LNet stages.

Ranjan et al. [18] proposed a deep neural network consisting of multiple branches to handle various face-related tasks. The proposed network uses Selective Search [26] to generate multi-region proposals. Although the proposed network has both face classification and smile recognition branching, face classification and smile recognition are performed as separate continuous pipelines. Most of the previous works that do not require a pre-normalisation step follow the same mechanism as [16, 18], which require region proposal steps in the middle of the process. These region proposal steps typically increase the overall processing time.

Object (face) detection in the wild. Thinking of a face as one type of object, there are many hints for designing a novel smiling face detection network that performs face detection and smile recognition in a single architecture without going through a pre-normalization step. SmileNet inherits the structural benefits of the latest methods with the concept of default box [15] or anchor box [19].

3. The Proposed Framework: SmileNet

The proposed SmileNet framework is shown in Fig. 2. SmileNet M is a fully convolutional neural network consisting solely of convolution and pool layers. The basic structure of SmileNet consists of six output layers (total number of output layers S is 6). Each output layer corresponding to a scale s provides a set of predicted output pairs that include the face confidence score, bounding box parameters and the smile confidence score $\{c_f, b, c_e\}_s$ at every pixel location, as shown in Fig. 2(b). As the input to SmileNet M, a colour

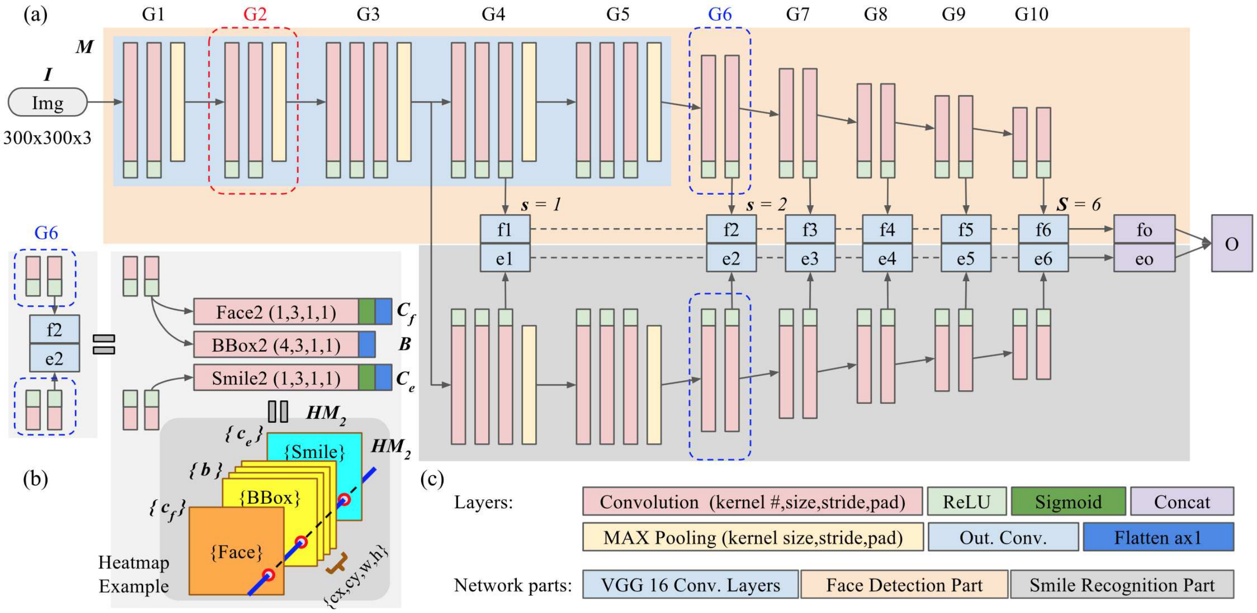

Figure 2. The architecture of SmileNet. (a) The entire architecture of SmileNet consisting solely of convolution and pool layers. (b) Example of the concatenated output convolution layer for the second scale ($s = 2$) that produces a heatmap volume. (c) Legend for layers and parts of SmieNet.

image I is given.

Our framework simultaneously learns the labels of the face class α_f, the bounding box β, and smile class α_e at each output layer. Based on the face confidence score $\{c_f\}$ that appears on each pixel in the face heatmap C_f, SmileNet first filters several candidates of the bounding box and smile confidence score. Then, SmileNet deduces a set of representative bounding boxes and the corresponding smile confidence scores based on Non-Maximum Suppression (NMS) [17]. The following sections describe how to configure SmileNet (Sec. 3.1), how to train face detection and smile recognition in a single architecture (Sec. 3.2), and how to combine the face detection and smile recognition results during testing (Sec. 3.3).

3.1. Model Construction

SmileNet M mainly consists of feature extraction (VGG16 Conv. Layers), face recognition, and smile recognition parts as shown in Fig. 2(a). $G[1:10]$ represents convolution and pool layer groups with the same input resolution. For example, G2 consists of two convolution layers and one pool layer, whereas G6 consists of two convolution layers. Similar to SSD [15], SmileNet outputs six-scale ($S = 6$) heatmap volumes generated by multiple output convolution layers [(f1, e1):(f6, e6)]. f[1:6] is produced by the face detection part, while e[1:6] is produced by the smile recognition part. The output convolution layers of the two different parts are eventually aligned and concatenated.

Each concatenated output convolution layer outputs a pixel-wise heatmap volume consisting of six heatmap planes. For example, the concatenated output convolution layer for the second scale ($s = 2$) outputs a three-dimensional volume ($HM_2 \times HM_2 \times 6$) consisting of six heatmap planes having the same resolution ($HM_2 \times HM_2$) of the second scale, as shown in Fig. 2(b). The first plane indicates the existence of a face. The last one shows confidence of the smile in pixels. The remaining four planes output the offset position of centre coordinates (cx, cy) relative to each pixel position, width w, and height h relative to the current heatmap scale s that make up the bounding box, respectively.

All of the convolution layers are followed by ReLU activation function except for the output convolution layer. For the output convolution layer, the sigmoid function comes after the layer for face and smile binary classification. The layers for bounding box regression use linear values like SSD [15]. The detailed parameters for layers in SmileNet are summarised in Table 1. The parameters of the convolution layer are denoted in the order of number of kernels, kernel size, stride and padding, while the parameters of the pool layer follow the order of kernel size, stride and padding.

During training, the output values that appear in heatmaps responsible for the bounding box and smile are examined only when the corresponding face label exists in the pixel (see details in Sec. 3.2.1). During testing, the values for the bounding box and the smile are examined only when the corresponding face confidence score exceeds

Group ID	Conv. ID: Parameters	Pool
G1	[1:2]: (64, 3, 1, 1)	(2, 2, 0)
G2	[1:2]: (128, 3, 1, 1)	(2, 2, 0)
G3	[1:3]: (256, 3, 1, 1)	(2, 2, 0)
G4	[1:3]: (512, 3, 1, 1)	(2, 2, 0)
G5	[1:3]: (512, 3, 1, 1)	(3, 1, 1)
G6	1: (1024, 3, 1, 1) 2: (1024, 1, 1, 0)	· ·
G7	1: (256, 1, 1, 0) 2: (512, 3, 2, 1)	· ·
G8	1: (128, 1, 1, 0) 2: (256, 3, 2, 1)	· ·
G9	1: (128, 1, 1, 0) 2: (256, 3, 1, 0)	· ·
G10	1: (128, 1, 1, 0) 2: (256, 3, 1, 0)	· ·
Out. Conv.	C_f: (1, 3, 1, 1) B: (4, 3, 1, 1) C_e: (1, 3, 1, 1)	· · ·

Table 1. The detailed parameters of SmileNet layers (see text).

a threshold.

3.1.1 Properties of the SmileNet Design

Handling multi-scale faces: We train SmileNet following similar mechanisms, including the loss functions, of SSD [15]. SSD is a scale-invariant object detector that uses multi-scale output layers.

Simplified model for the face modality: We utilise only one aspect ratio configuring a default box to assign a ground truth (face or smile) label to a pixel position in a heatmap, as shown in Fig. 3. Face deformations, caused by expression and pose, do not create many different shapes between faces. Therefore, a square bounding box fits generally well into the shape of a face. For the same reason, Hao et al. [8] uses Single-Scale RPN utilising one anchor box.

Inheriting characteristics of pre-trained models: Liu et al. [16] confirmed that a model taking the parameters of a pre-trained model trained on an object dataset (e.g., ImageNet [4]) is suitable for localising faces. In addition, the model that uses parameters of a pre-trained network based on a face identity dataset (e.g., CelebFaces [24]) is useful for capturing more detailed level of face attributes. Therefore, we copy the pre-trained parameters of the VGG16 network [22] (originally trained for object classification) to finetune the face detection part of SmileNet by training on a face dataset (e.g., AFLW). Then, the finetuned parameters of the face detection part of SmileNet are inherited by the smile recognition part (capturing the details of face attributes) (see details in Sec. 3.2). This selective inheritance of characteristics of the models trained with different datasets helps to make the best of SmileNet performance.

3.2. Training

Training of SmileNet follows the following four steps: 1) Copying parameters of the VGG16 network [22] (convolution layers) to the VGG16 (feature extraction) part $G[1:5]$ of SmileNet and subsampling the parameters from fully connected layers ($fc6$ and $fc7$) of VGG16 network to the $G6$ layers of SmileNet, as described in SSD [15], 2) freezing the smile recognition part and finetuning the face detection part by using the AFLW (face) dataset [12], 3) copying the parameters of the layers $G[4:10]$ constituting the face detection part to the corresponding layers of the smile recognition part, and 4) freezing the face detection part and finetuning smile recognition part by using CelebA (smile) dataset [16]. The first and second steps are very similar to the initialisation and end-to-end learning process of SSD network [15]. In particular, we use the same cost function of the SSD to finetune the face detection part of SmileNet. However, in this paper, only one square aspect ratio ($a_r \in \{1\}$ in [15]) is used to match the default box [15] (see Fig. 3), which ultimately reduces the kernel size of the output convolution layers, as shown in Fig. 2(b). The details of the initialisation and training procedure are similar to [15].

3.2.1 Face Detection

As described above, finetuning of the face detection part is based on the use of an objective loss function L_{face} (similar to SSD [15]), which is a weighted sum of the face classification loss L_{cls} and the bounding box regression loss L_{reg} defined as:

$$L_{face}(x_f, c, l, g) = \frac{1}{N}(L_{cls}(x_f, c) + \lambda x_f L_{reg}(l, g)), \quad (1)$$

where N is the total number of matched default boxes. For the regression loss L_{reg}, Smooth L1 loss [7] is used for calculating the distance between predicted $l = \{l_{cx}, l_{cy}, l_w, l_h\}$ and the ground truth $g = \{g_{cx}, g_{cy}, g_w, g_h\}$ bounding boxes [15], as shown in Eq. 2 and 3. The regression loss is activated only when the indicator $x_f \in \{1, 0\}$ for matching the default box $d = \{d_{cx}, d_{cy}, d_w, d_h\}$ to face existence is identified as $True$ ($x_f = 1$), and is disabled otherwise $x_f = 0$.

$$\begin{aligned} L_{reg}(l, g) &= \sum_{m \in \{cx, cy, w, h\}} smooth_{L_1}(l_m - \hat{g}_m), \\ \hat{g}_{cx} = (g_{cx} - d_{cx})/d_w, &\quad \hat{g}_{cy} = (g_{cy} - d_{cy})/d_h, \\ \hat{g}_w = \log(g_w/d_w), &\quad \hat{g}_h = \log(g_h/d_h), \end{aligned} \quad (2)$$

where

$$smooth_{L_1}(k) = \begin{cases} 0.5k^2, & \text{if } \|k\| < 1 \\ \|k\| - 0.5, & \text{otherwise} \end{cases} \quad (3)$$

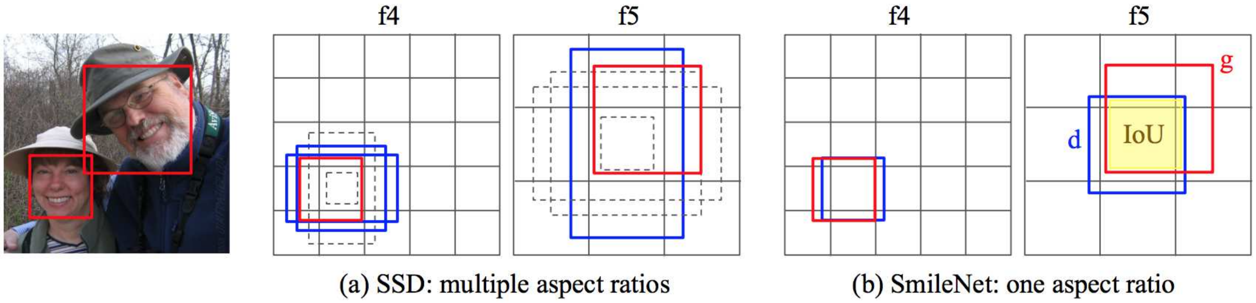
f4
f5
(a) SSD: multiple aspect ratios
f4
f5
g
d
IoU
(b) SmileNet: one aspect ratio

Figure 3. Example of matched default box for the face confidence heatmaps $C_{f_{[4:5]}}$, produced by $f4$ and $f5$ output convolution layers (see Fig. 2). (a) Dotted boxes (grey) represent multiple candidate default boxes with different aspect ratios. In this case, when calculating the regression loss of the bounding box in the process of SmileNet training, the various shapes of the matched default box (blue) do not help converge the solution. Thus, (b) SmileNet uses only one aspect ratio in the matching process of the default box d. The example image is one of the sample images of AFLW dataset [12].

The face classification loss L_{cls} is based on binary cross entropy over face confidence scores c, as shown in Eq. 4.

$$L_{cls}(x_f, c) = -x_f log(c) - (1 - x_f) log(1 - c) \tag{4}$$

A positive matched box $x_f = 1$ for a face is indicated based on the ground truth label. However, a negative matched box $x_f = 0$ that represents the background region is indicated in the process of Hard Negative Mining (HNM), as described in [15]. In the process of HNM, we sort the calculated losses only in the background region ($\neg(x_f = 1)$) in descending order. Then, we select and use the highest losses as the classification result for the negative region. The loss-balancing weight λ (in Eq. 1) is set to 1 by experiment. If we want to bias towards better box locations, we can set the weight to a higher value (e.g., to 2).

3.2.2 Smile Recognition

For calculating the smile classification loss L_{smile}, we use the same binary cross entropy over smile confidence scores e, as shown in Eq. 5. However, in this case, all positive and negative matched boxes $x_e = \{1, 0\}$ are indicated based on the ground truth smile label. That means that the process for finetuning of the smile recognition part does not require Hard Negative Mining (HNM).

$$L_{smile}(x_e, e) = -x_e log(e) - (1 - x_e) log(1 - e) \tag{5}$$

By freezing the face detection part of SmileNet, finetuning the smile recognition part does not impair the face detection performance. Therefore, if we are able to annotate the same location of the face bounding box across the dataset, it is possible to train the smile recognition task (or other face-related tasks) without going through a pre-normalisation step, such as face detection and registration. Based on this, this section focuses on calculating the misclassification losses to finetune the smile recognition part.

However, note that the AFLW (face) dataset [12] and the CelebA (smile) dataset [16] have different bounding box positions and shapes. To solve this problem, we empirically adjusted the bounding box position of the CelebA dataset to create a square box that surrounds the entire face area centred on the nose, as shown in Fig. 4. To do this, we used the five landmark locations provided by CelebA dataset.

3.2.3 Data Augmentation in Training

SmileNet uses a 300 × 300 resolution and 3 channel colour input image. Prior to data augmentation, all pixel values for the R, G, and B channels of a sample image are normalised based on the mean and standard deviation values of the entire dataset. Each sample image is first flipped in the horizontal direction with a probability of 0.5. In the training session, we randomly select one of data augmentation mechanisms (shrinking, cropping and gamma correction) to create noise-applied data samples for each epoch. Detailed description of the augmentation strategy is described as follows:

Shrinking: We maintain the ratio of the image width and height, however, we randomly select a real number between 0 and 1 to make the face smaller. The chosen number represents the percentage of the original size. For example, if the number is 0.5, the resolution of the augmented image is half the size of the original image. If the augmented image is smaller than 300 × 300, we fill the background with random colour.

Cropping: To enlarge the face size in the image, we randomly select a partial area of the original image (the selected subregion is a square). After cropping, we rescale the cropped image to 300 × 300 pixels. The face of the original image is then enlarged. If the subregion does not contain more than half of the original face region, we repeatedly

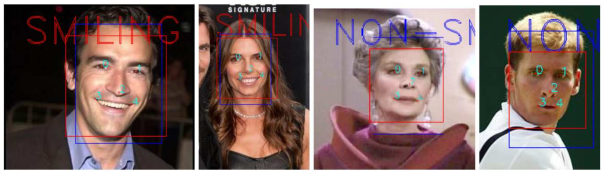

Figure 4. Examples of adjusted (red) square bounding boxes of the original CelebA (blue) rectangle ones.

select random subregions to find an appropriate region.

Gamma correction: We apply gamma correction to each colour channel to make the SmileNet invariant to illumination. The input image of SmileNet is a three-colour channel image. For gamma correction, we randomly select one of R, G, or B colour channels. Then, we apply a random gamma correction value to adjust the grayscale in the selected image plane, and we combine the individually corrected image planes to create a colour image with different lighting.

3.3. Testing

The smiling face detection is based on face and smile confidence scores. If the face confidence score exceeds the threshold th_{face} (0.5, defined empirically), we classify the set of face candidate pixel positions as a face. We then make a set of bounding box candidates based on the pixel positions. Finally, the representative faces among the candidates are detected using the Non-Maximum Suppression (NMS) method (with jaccard overlap value 0.5), as shown in Fig. 5 (black squares). In addition, if the smile confidence score associated with a representative bounding box exceeds the smile threshold th_{smile} (0.5, defined empirically), the bounding box indicates a smiling face (red square).

As mentioned in Sec. 3.1, each output layer of SmileNet generates three categorical heatmaps that represent the existence of a face, the face bounding box and smiling face, as shown in Fig. 2(b). Specifically, Fig. 5 visualises the heatmap generated by SmileNet's second-scale output layer ($s = 2$), which handles the second smallest size of the face that appears in the image. Thus, the pixels in the heatmap are highlighted or activated only when a specific size of a face is detected. The forefront heatmap highlights two clusters of pixels, indicating that two faces exist. The rearmost heatmap highlights the corresponding pixel only when the detected face is smiling.

The reason for the smile heatmap highlighting the background area rather than the face area (in Fig. 5) is because we consider only the face region when calculating the smile recognition loss as mentioned in Sec. 3.2.2. In this case, although the training samples of the dataset are used for distinguishing between smile and non-smile, the background area including non-face texture is not considered in the training process for smile recognition that ultimately outputs the random prediction value for the non-face region. The predicted confidence score of the smile appearing in the background area is not considered in the smiling face detection process.

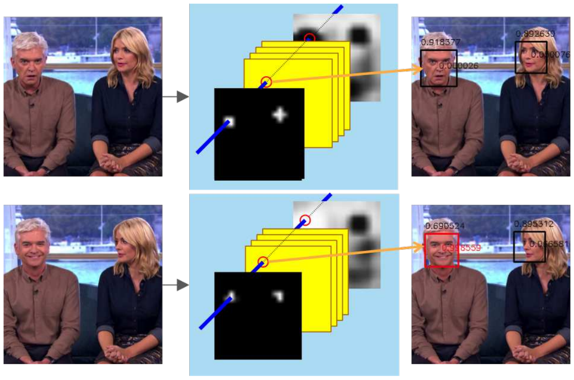

Figure 5. Examples of face detection and smile recognition.

4. Experiments and Results

SmileNet is inspired by SSD [15], which promises real-time detection performances. Thus, the parameter values used in the process of finetuning the face detection and the smile recognition parts of SmileNet are initialised with the values used for training the base network of SSD [15]. We used SGD with initial learning rate 10^{-3}, 0.9 momentum, 0.0005 weight decay, and batch size 10. We used the 10^{-3} learning rate for the first $40K$ iterations, then continued training for $40K$ with 10^{-2}. We continuously reduced the learning rate every $40K$ iterations until it reached 10^{-5}. Increasing the learning rate for the second $40K$ iterations speeds up the optimisation. However, we first started training with the learning rate of 10^{-3}, because the optimisation process diverged if we used a larger learning rate in the beginning.

In the next sections, we detail the experiments we have conducted to evaluate two main performance factors of SmileNet: the smiling face detection accuracy and the processing time.

4.1. Smiling Face Detection Performance

For smiling face detection, the accuracy refers to the smile recognition performance including the face detection results. If face detection fails, the result of smile recognition is considered to be a non-smile. Beginning with [29], which performed the first extensive smile detection study, most of the subsequent studies used the GENKI-4K dataset for performance evaluation [1][1]. In this paper, the smiling face detection experiments were performed not only on the

[1] The GENKI-4K dataset is a subset of the GENKI dataset used in [29]. This dataset consists of 4,000 face images, each labelled with smile and head pose (yaw, pitch, roll). Only the GENKI-4K dataset is publicly available.

Method	Feature	Classifier	Detection	Registration	Input ($W \times H \times C$)	Accuracy (%)
[21]	Pixel comparison	AdaBoost	Y	Eyes (manual)	$48 \times 48 \times 1$	89.70 ± 0.45
[14]	HOG	SVM	[27]	Eyes	$48 \times 48 \times 1$	92.26 ± 0.81
[10]	Multi-Gaussian	SVM	[27]	Y	$64 \times 64 \times 1$	92.97
[11]	LBP	SVM	[27]+[23] / ori.	$5+6$ Pts	$96 \times 96 \times 1$	93.20 ± 0.92
[2]	HOG	ELM	[27]	Flow-based [2]	$100 \times 100 \times 1$	88.20
[31]	CNN	Softmax	Y	Face Pts	$90 \times 90 \times 1$	94.60 ± 0.29
[13]	Gabor-HOG	SVM	[27] / manual	Y	$64 \times 64 \times 1$	91.60 ± 0.89
[3]-I	CNN	SVM	[16]	Y	$64 \times 64 \times 1$	92.05 ± 0.74
[3]-II	CNN	SVM	[16]	·	$64 \times 64 \times 1$	**90.60** ± 0.75
[3]-III	CNN	SVM	·	·	$64 \times 64 \times 1$	78.10 ± 0.56
SmileNet	**CNN**	**Sigmoid**	·	·	$\mathbf{300 \times 300 \times 3}$	**95.76** ± 0.56

Table 2. A comparison with the state-of-the-art methods on the GENKI-4K dataset [1]. We summarise the features, classifiers, detection / registration methods and input image resolution (width, height, and channel) that were used in previous studies in published order. All previous studies require a normalised (cropped and aligned) input image, which necessarily require face detection and registration steps in advance (except [3]-II and III). Some works [21, 10, 31, 13, 3] do not specify how to detect and align a face (in this case, 'Y'), while [11] mentions that the original image is used if the face detection fails.

GENKI-4K dataset but also on the CelebA dataset which also contains smile labels.

Testing on the GENKI-4K dataset: Experiments that use this dataset are conventionally based on four-fold validation procedures. The four-fold validation utilises one of four combinations of training and testing samples. Each fold uses 75% of the dataset for training and the remaining 25% for testing.

However, as GENKI-4K dataset contains a relatively small number of data samples ($4,000$), we initially utilised the CelebA dataset that contains a rich set of images for training. When SmileNet was trained on the CelebA dataset, we used the entire GENKI-4K dataset for testing. We obtained a smiling face detection accuracy of 95.23%, as shown in Fig. 6. Despite being trained on a completely different dataset with different characteristics, SmileNet has already surpassed all the latest methods that use the GENKI-4K dataset for testing, as shown in Table 2.

Additionally, to provide a fair comparison with other methods that use the four-fold validation strategy, we used the GENKI-4K dataset together with the bounding box annotations obtained with our method (see Sec. 4.3) to finetune the SmileNet, which was trained on the CelebA dataset. In this case, the smiling face detection accuracy is improved further. This is due to the fact that the training samples in GENKI-4K dataset are relatively similar to the testing samples as compared to CelebA dataset. Although the training and testing samples do not overlap, using the same dataset (GENKI-4K) for training helps SmileNet learn the test sample characteristics of the same (GENKI-4K) dataset. Our four-fold validation results were 96.33%, 96.30%, 95.30% and 95.10%, as shown in Fig. 6. We obtained the best results (mean: **95.76**%, standard deviation: 0.56%) compared to the accuracies reported by existing works listed in Table 2.

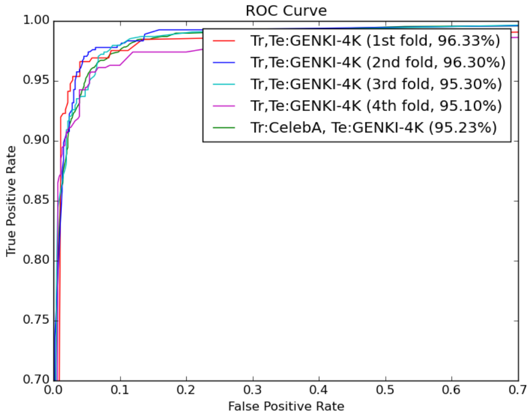

Figure 6. Receiver Operating Characteristic (ROC) curve for smiling face detection accuracy using GENKI-4K [1] dataset. Tr and Te represent training and testing, respectively.

Although SmileNet does not require separate steps for face detection and registration, SmileNet's smiling face detection results rely on the face detection performed in parallel on the same architecture. Among the existing works listed in Table 2, Chen's work ([3]-II) reports testing accuracy when the registration process is not used. We therefore compare SmileNet's smiling face detection performance more closely to the method of Chen ([3]-II). Our experimental results show that SmileNet outperforms (**95.76**%) the most recently reported smile detection result (90.60%) based on a deep learning architecture ([3]-II).

Testing on the CelebA dataset: In the second experiment, we used the CelebA dataset to train and test SmileNet. In this experiment, we randomly selected 75% of the dataset

Method	RP	Acc. (%)	Time (ms.)
Liu et al. [16]	EB [32]	92.00	139.00
Ranjan et al. [18]	SS [26]	93.00	3,500.00
SmileNet	·	**92.81**	**47.28**

Table 3. Comparison to the state-of-the-art methods on the CelebA dataset. RP, EB and SS represent Region Proposal, EdgeBox [32] and Selective Search [26], respectively.

for training and used the remaining 25% for the testing. We performed several experiments using different combinations of randomly selected training and test samples. Our experimental results show that SmileNet detects smiling faces accurately (mean: **92.81**%), similarly to the state-of-the-art methods ([16]: 92.00% and [18]: 93.00%), as shown in Table 3. However, SmileNet is much faster (**47.28** *ms*) than the other methods ([16]: 139 *ms*, [18]: 3,500 *ms*) that require region proposal methods for smile recognition (see Table. 3).

4.2. Computational Speed

The average processing time of SmileNet was **47.28** *ms* (**21.15** *FPS*) during testing, in our experimental environment consisting of an Intel Core i7-6700HQ CPU processor and an NVIDIA GeForce GTX 960M GPU with 23.5GB of DRAM. The state-of-the-art method [16] requires 35 *ms* to generate the face confidence heatmap and 14 *ms* to classify the attributes. In addition, this method requires another 90 *ms* to find the candidate bounding box (EdgeBox [32]) for localising the final bounding box that ends up with a total processing time of 139 *ms* (7.19 *FPS*). Another state-of-the-art method [18] takes an average of 3,500 *ms* (0.29 *FPS*) to process an image. Ranjan et al. [18] explains that the main bottleneck for speed is the process of proposing regions (Selective Search [26]) and the repetitive CNN process for every individual proposal.

To ensure a fair comparison of the processing times, we should measure the time in the same experimental environment. However, Liu et al. [16] does not provide detailed information about the experimental environment, except that they use GPUs. Ranjan et al. [18] implemented their all-in-one network using 8 CPU cores and GTX TITAN-X GPUs. The processing speed of the all-in-one network is 74 times slower, even with a more powerful experimental environment.

4.3. Discussion

Although SmileNet is significantly faster than other smile detection methods, the processing speed is lower than the base object detection (SSD) model [15] as the complexity of SmileNet is nearly twice that of SSD. Placing more layers to perform smile recognition increased the number of parameters in SmileNet. However, the structure of the all-in-one network [18] shows that sharing more convolutional features does not degrade the performance of various tasks. Capitalising on this idea, we expect to further reduce the complexity of SmileNet by sharing more layers and assigning a relatively small number of layers to other face-related tasks (e.g., smile recognition).

SmileNet requires bounding box and smile labels for training. However, GENKI-4K dataset does not provide bounding box labels. Therefore, we annotated the face bounding box of the images by using SmileNet trained on the CelebA dataset (with bounding box label). When detection failed, we manually annotated the bounding box. We then used the annotated bounding box label when training SmileNet with the GENKI-4K dataset. The bounding box labels for the GENKI-4K dataset will be made publicly available.[2]

The face detection performance of SmileNet inherits the performance of SSD [15], which is typically poor for detecting small faces. However, the GENKI-4K and CelebA datasets do not contain extremely small faces. To handle extremely small faces, we can extend the model by applying the idea that takes advantage of context patterns surrounding a face area, as proposed in [9].

5. Conclusions

In this paper, we tackled the problem of smiling face detection in the wild without a pre-normalisation step (face detection and registration). To this end, we proposed SmileNet which performs face detection and smile recognition simultaneously in a single framework. For fast and scale-invariant detection, SmileNet inherits the benefits of the state-of-the-art object detection network SSD. In addition, we used pre-trained parameters of two different networks (those trained for object classification and trained for face detection) to learn the face and smile patterns. Consequently, we built a single framework that enables real-time scale-invariant smiling face detection in the wild. Our experimental results show that SmileNet (95.76%) outperforms the state-of-the-art methods while maintaining real-time speed (21.15 *FPS*).

Acknowledgment

This work is supported by the Technology Strategy Board / Innovate UK project Sensing Feeling (project no. 102547).

References

[1] The MPLab GENKI Database, GENKI-4K Subset. `http://mplab.ucsd.edu`. Accessed on Jun. 23, 2017.

[2] Project page: https://sites.google.com/view/sensingfeeling/

[2] L. An, S. Yang, and B. Bhanu. Efficient smile detection by extreme learning machine. *Neurocomput.*, 149(PA):354–363, Feb. 2015.

[3] J. Chen, Q. Ou, Z. Chi, and H. Fu. Smile detection in the wild with deep convolutional neural networks. *Machine Vision Applications*, 28(1-2):173–183, Feb. 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[5] P. Ekman. Universal and cultural differences in facial expression of emotion. *Nebr. Sym. Motiv.*, 19:207–283, 1971.

[6] P. Ekman. Strong evidence for universals in facial expressions: A reply to russells mistaken critique. *Psychol. Bull.*, 115(2):268–287, 1994.

[7] R. Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[8] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu. Scale-aware face detection. In *CVPR*, 2017.

[9] P. Hu and D. Ramanan. Finding tiny faces. In *CVPR*, pages – , 2017.

[10] V. Jain and J. L. Crowley. Smile Detection Using Multi-scale Gaussian Derivatives. In *12th WSEAS International Conference on Signal Processing, Robotics and Automation*, Cambridge, United Kingdom, Feb. 2013.

[11] S. E. Kahou, P. Froumenty, and C. J. Pal. Facial expression analysis based on high dimensional binary features. In *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, pages 135–147, 2014.

[12] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[13] J. Li, J. Chen, and Z. Chi. Smile detection in the wild with hierarchical visual feature. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 639–643, Sept 2016.

[14] M. Liu, S. Li, S. Shan, and X. Chen. Enhancing expression recognition in the wild with unlabeled reference data. In *Computer Vision - ACCV 2012, 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part II*, pages 577–588, 2012.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 2016.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[17] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, ICPR '06, pages 850–855, Washington, DC, USA, 2006. IEEE Computer Society.

[18] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *12th IEEE International Conference on Automatic Face and Gesture Recognition FG 2017, Washington, DC, USA, May 30-June 3*, 2017.

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2016.

[20] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.

[21] C. Shan. Smile detection by boosting pixel differences. *Trans. Img. Proc.*, 21(1):431–436, Jan. 2012.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[23] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3476–3483, Washington, DC, USA, 2013. IEEE Computer Society.

[24] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1891–1898, Washington, DC, USA, 2014. IEEE Computer Society.

[25] G. K. L. Tam, Z. Q. Cheng, Y. K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X. F. Sun, and P. L. Rosin. Registration of 3d point clouds and meshes: A survey from rigid to nonrigid. *IEEE Transactions on Visualization and Computer Graphics*, 19(7):1199–1217, July 2013.

[26] K. E. A. van de Sande, J. Uijlings, T. Gevers, and A. Smeulders. Segmentation as Selective Search for Object Recognition. In *IEEE International Conference on Computer Vision*, 2011.

[27] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[28] N. Wang, X. Gao, D. Tao, and X. Li. Facial feature point detection: A comprehensive survey. *arXiv*, 2014.

[29] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, Nov 2009.

[30] C. Zhang and Z. Zhang. A Survey of Recent Advances in Face Detection. *Microsoft Research, Tech. Rep.*, 2010.

[31] K. Zhang, Y. Huang, H. Wu, and L. Wang. Facial smile detection based on deep learning features. In *3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015*, pages 534–538, 2015.

[32] L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. In *ECCV*. European Conference on Computer Vision, September 2014.