Journal of Electronic Imaging

JElectronicImaging.org

Robust smile detection using convolutional neural networks

Simone Bianco
Luigi Celona
Raimondo Schettini

SPIE. IS&T imaging.org

Simone Bianco, Luigi Celona, Raimondo Schettini, "Robust smile detection using convolutional neural networks," *J. Electron. Imaging* **25**(6), 063002 (2016), doi: 10.1117/1.JEI.25.6.063002.

Robust smile detection using convolutional neural networks

Simone Bianco,* Luigi Celona, and Raimondo Schettini
University of Milano-Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione, Viale Sarca 336, 20126 Milano, Italy

Abstract. We present a fully automated approach for smile detection. Faces are detected using a multiview face detector and aligned and scaled using automatically detected eye locations. Then, we use a convolutional neural network (CNN) to determine whether it is a smiling face or not. To this end, we investigate different shallow CNN architectures that can be trained even when the amount of learning data is limited. We evaluate our complete processing pipeline on the largest publicly available image database for smile detection in an uncontrolled scenario. We investigate the robustness of the method to different kinds of geometric transformations (rotation, translation, and scaling) due to imprecise face localization, and to several kinds of distortions (compression, noise, and blur). To the best of our knowledge, this is the first time that this type of investigation has been performed for smile detection. Experimental results show that our proposal outperforms state-of-the-art methods on both high- and low-quality images. *© 2016 SPIE and IS&T* [DOI: 10.1117/1.JEI.25.6.063002]

Keywords: smile detection; deep learning; convolutional neural networks; face detection; face alignment.

Paper 16715 received Aug. 24, 2016; accepted for publication Oct. 24, 2016; published online Nov. 14, 2016.

1 Introduction

Smiling is an expression denoting happiness, pleasure, satisfaction, or amusement. It is characterized by the upward movements of the lip corners and of the cheeks. In the framework of the facial action coding system,[1] smile can be seen as the combination of the facial muscles corresponding to the action unit 6 and 12 (AU6 and AU12).

The first works on smile detection used databases taken under constrained laboratory environment; Shinohara and Otsu[2] used higher-order local autocorrelation features and Fisher weight map and achieved good performance on their own database consisting of only four people. Bai et al.[3] extracted pyramid histogram of oriented gradients (HOGs) features from the region of the mouth on the Cohn–Kanade AU-Coded Facial Expression Database. The first comprehensive work for smile detection in unconstrained scenarios was proposed by Whitehill et al.[4] At the same time, they also made publicly available a new dataset (GENKI) with content from the web for smile detection in the wild. Using this dataset, Shan[5,6] proposed a very efficient smile detection approach by simply comparing intensities of a few pixels in a face image.[3] Zhang[7] demonstrated the effectiveness and efficiency of mouth features (MFs) for smile detection. More recently, An et al.[8] proposed a fully automated smile detection approach. They adopted three popular feature descriptors (local binary patterns,[9] local phase quantization,[10] and HOGs[11]) and achieved the best results on both the GENKI-4K database and their own collected MIX databases. Gao et al.[12] proposed a semiautomated smile detector, which achieved the best performance on the GENKI-4K database using a combination of features [raw pixel values, HOG, and self-similarity of gradients (GSS)] combining multiple classifiers.

In this paper, we present a fully automated approach for smile detection in digital images. According to our proposal, the input image is processed in order to detect faces using a face detector inspired by Farfade et al.[13] The faces are then aligned using an eye-based approach using a facial landmarks detector[14] that does not require any manual labeling. Then a convolutional neural network (CNN) is exploited to predict smiling of the detected faces. The CNN architecture has been designed to be trained even when the amount of learning data is limited. We evaluate the performance of the proposed pipeline on the GENKI-4K database,[15] the only publicly available dataset in unconstrained scenarios. The proposed pipeline achieves very good results in smile detection accuracy and is more robust to various image distortions and transformations in comparison with the state of the art.

2 Proposed Approach

The main steps of the pipeline are shown in Fig. 1. Given an image, we detect the faces and align them fixing the eyes position. Then a CNN is used to understand whether it is a smiling face or not.

Given an image, we detect the faces using a multiview face detector inspired by Farfade et al.[13] The detected faces are aligned fixing the eyes position and then rescaled to a common size. In more detail, we compute the (x, y) coordinates of 49 facial landmarks, obtained using the publicly available implementation of Chehra.[14] Among the detected landmarks, we consider only the two landmarks corresponding to the eyes, corner locations. These are used in our eye-based face alignment method, which consists of fixing the eye corner distance to 85 pixels using an affine transform matrix, which is composed only of rotation and scaling. Facial images are then obtained by cropping and scaling the transformed images to 36×36 pixels.

Given the cropped and aligned 36×36, a central 32×32 patch is extracted and given as input to a CNN to classify it as smile or nonsmile. Different CNN configurations are tested in this paper. They are designed to be trained even when the amount of labeled data is limited. Their configurations are summarized in Table 1. In the following, we will refer to the CNNs by their names (A to C). The input to our CNNs is a fixed-size 32×32 RGB image. The image is passed through a stack of convolutional (conv) layers, where we use filters with a variable number of 5×5 and 3×3 kernels. The convolution stride is fixed to 1 pixel and the spatial padding is such that the spatial resolution is preserved after convolution, i.e., $(\text{kernel size} - 1)/2$. Spatial pooling is carried out by a max-pooling (maxpool) layer, which follows the

*Address all correspondence to: Simone Bianco, E-mail: simone.bianco@disco.unimib.it

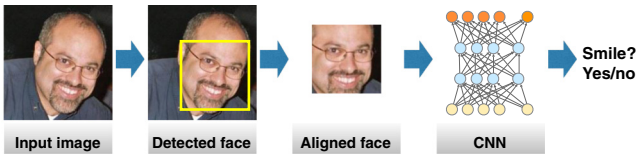

Fig. 1 Outline of the proposed method for smile detection.

Table 1 CNN configurations investigated (shown in columns). The ReLU activation function is not shown for brevity.

CNN Configuration		
A	B	C
Four weight layers	Five weight layers	Five weight layers
input (32 × 32 RGB image)		
conv3-32	conv3-32	conv3-32
Maxpool		
LRN		
conv5-32	conv3-32 conv5-32	conv5-32
Avgpool		
LRN		
conv5-64	conv5-64	conv3-64
Avgpool		
		FC-1024
FC-2		
Soft-max		

first convolutional layer, and two average-pooling (avgpool) layers after the other convolutional layers (not all the convolutional layers are followed by spatial pooling). All the spatial pooling layers are performed over 3×3 pixel windows, with stride 2. All hidden layers are equipped with the rectified linear unit (ReLU).[16,17] Local response normalization (LRN) layers follow the maxpool layer and the first avgpool layer. The LRN layer is applied on spatial regions of size 3×3 independently for each channel. The previous stack of layers is followed by a fully connected (FC) layer performing two-way classification, i.e., smile/nonsmile. The final layer is the soft-max layer, which produces a probability distribution over the two class labels.

3 Experimental Setup

In the experiment, we use the GENKI-4K database.[15] It is the most challenging and largest available database for the smile detection task in the unconstrained scenario. It contains 4000 facial images of a wide range of subjects with different ethnicity, age, facial appearance, pose, illumination, and imaging conditions. All the images are labeled by human coders, 2162 images are labeled as smile and the remaining 1828 images are labeled as nonsmile. Although a few images are, in our opinion, incorrectly labeled, we did not make any change to the groundtruth labels.

We train our CNNs from scratch using data augmentation to regularize our CNNs and reduce the risk of overfitting. It consists of artificially enlarging the dataset using label-preserving transformations. In this paper, data augmentation consists of generating image translations and horizontal reflections. We do this by randomly extracting five 32×32 patches as well as their horizontal reflections from 36×36 facial images. This increases the size of our training set by a factor of 10.

In our experiments, fourfold cross-validation is performed on the GENKI-4K dataset, meaning that we randomly partitioned the dataset into four subsets. For each round of cross-validation, we used a subset for testing and the other three subsets as training. Results are reported in terms of average accuracy over the four rounds of cross-validation.

The prediction made by the CNNs's soft-max is computed cropping the central 32×32 patch from the 36×36 facial image. In addition to this single patch prediction, we compute the prediction by oversampling the facial image; in this case, the prediction is made by considering five 32×32 patches (the four corner patches and the center patch) as well as their horizontal reflections and averaging the predictions made by the CNNs's soft-max layer on the 10 patches. We also investigate the performances obtained by combining the predictions of the three proposed CNN configurations and the influence of the face alignment step on the overall accuracy.

The average accuracy of the different instantiations of the proposed pipeline is reported in Table 2. From the results, it is possible to notice that using a single CNN the best results are obtained with CNN-A and using face alignment. It can be seen that performance can be slightly improved by oversampling the input image and combining the predictions of different CNNs.

To the best of our knowledge, the best performance on GENKI-4K database is obtained by Gao et al.,[12] who, exploiting a semiautomatic procedure (i.e., manual face alignment), report an average accuracy of 94.61%. For sake of comparison, we have therefore reimplemented their method within our processing pipeline. The comparison with other fully automatic smile detection methods in the state of the art[6–8,12] is reported in Table 3. It is possible to see that the proposed method is able to outperform the best method in the state of the art, i.e., the reimplementation of Gao et al.[12] in

Table 2 Smile detection accuracy results using the proposed CNN configurations.

CNN configuration (see Table 1)	Accuracy (%)	
	Without face alignment	With face alignment
A	92.60	93.13
B	92.18	92.80
C	92.70	92.75
A (oversampled)	90.45	93.35
A+B+C (oversampled)	92.53	93.77

Table 3 Comparison with state-of-the-art methods on the GENKI-4K database.

Method	Features	Classifier	Accuracy (%)
An et al.[8]	HOG	ELM	88.50
Zhang[7]	MFs	AdaBoost	89.21
Shan[6]	Pixel difference	AdaBoost	89.70
Gao et al.[12]	Raw pixels+HOG+GSS	Linear SVM	91.20
Proposed	CNN	CNN	93.35

our pipeline, by 2.15%. Some examples of misclassified images are reported in additional material.[18] Since one of the sources of error is the facial landmarks localization, we now investigate the robustness of the CNN to bad face alignment and image distortions.

3.1 *Classification Robustness to Face Alignment*

Imprecise face alignment can be caused both by inaccurate face detection and bad facial landmarks localization. As seen in Table 2, the removal of the face alignment step causes a drop in performance for all the CNN configurations investigated in this paper. The same is true also for the best algorithm in the state of the art, i.e., Gao et al.,[12] whose average accuracy without face alignment drops to 87.78%. To investigate this issue, given the aligned cropped faces of the GENKI-4k database, we create a dataset applying some geometric transformations on the 36×36 facial images. Specifically, we use rotation of the face around its center with different angles $(-30 \text{ deg}, -20 \text{ deg}, \ldots, 30 \text{ deg})$, scaling with different scale factors $(0.80, 0.90, \ldots, 1.20)$, and translation with various pixel ofsets $(-8, -6, \ldots, 8)$. For all the transformations, zero-padding is used for pixels falling outside the image window.

We run a set of three experiments considering a single geometric transformation at a time. The transformed images are classified using the (transformation-free) trained CNN-A. The results of the performed experiments are reported in Fig. 2. In the same plots, we also report the results obtained by our implementation of the method by Gao et al.[12] From the plots, it is possible to notice that CNN-A shows a very high level of robustness against scaling. The performance remains almost unaltered except when the object of interest is small. Regarding translation and rotation, the CNN shows a lower level of robustness, with performance significantly decreasing, respectively, for offsets larger than 5 to 10 pixels and for a rotation angle larger than 10 deg to 20 deg. Comparing our results to those by Gao et al.,[12] we notice that both methods show a similar trend for the robustness to scale changes, while our method results in more robustness to rotations and translations.

3.2 *Classification Robustness to Image Artifacts*

Images available to consumers usually undergo several stages, namely acquisition, compression, transmission, and reception, and they may suffer multiple distortions.[19] In this set of experiments, we test the robustness of the proposed CNN with respect to four of the most common image artifacts in real-world digital photos: JPEG compression at different quality indexes, Gaussian noise, Gaussian blur varying the filter size and variance, and motion blur with fixed angle and different pixel lengths. We run two different experiments: in the first one, we consider a single artifact at a time and in the second one, images are corrupted by multiple artifacts together. In both cases, artifacts are applied on the detected faces after the alignment step. Faces are classified at the increase of the strength of the artifacts using the (distortion-free) trained CNN-A. The results of the single-artifact experiment are reported in additional material. In the multiple-artifacts experiment, we evaluate the robustness of the proposed pipeline at six different distortion levels obtained by combining blur, noise, and JPEG compression. We run an experiment considering a single distortion level at a time for each face. Specifically, artifacts are applied in the same order they generate in typical imaging pipelines[20]: motion blur varying pixel lengths (5, 10, 15, 20, 25, and 30) and fixed angle 45 deg, Gaussian noise with zero-mean and different variances ($\sigma^2 = 0.01, 0.02, 0.03, 0.04, 0.05$, and 0.06), and JPEG compression at different quality indexes (95%, 75%, 60%, 40%, 20%, and 0%). In total, we consider six different distortion levels that can be divided into three distortion groups: low distortion (levels 1 and 2), medium distortion (levels 3 and 4), and high distortion (levels 5 and 6). Figure 3 shows the results of the performed experiment both on our pipeline and our implementation of the method by Gao et al.[12] From the plots, it is possible to see that our method has a higher robustness for all distortion levels except for the highest one, where the difference between our method and that by Gao et al.[12] is less than 1%. For intermediate distortion levels, the accuracy of our method is higher than that achieved by Gao et al.,[12] with an improvement higher than 9% for distortion levels from 1 to 5 (with a peak 13.6% improvement for distortion level 3).

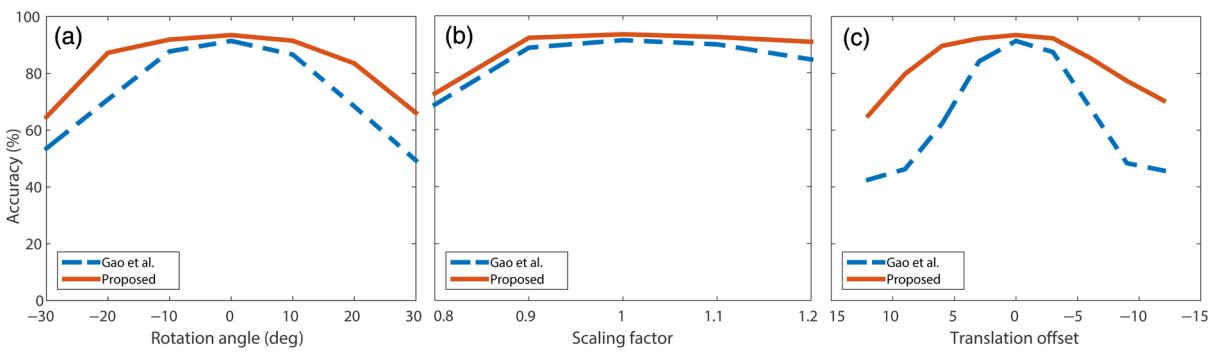

Fig. 2 Rates varying (a) the rotation angle, (b) the scaling factor, and (c) the translation offset.

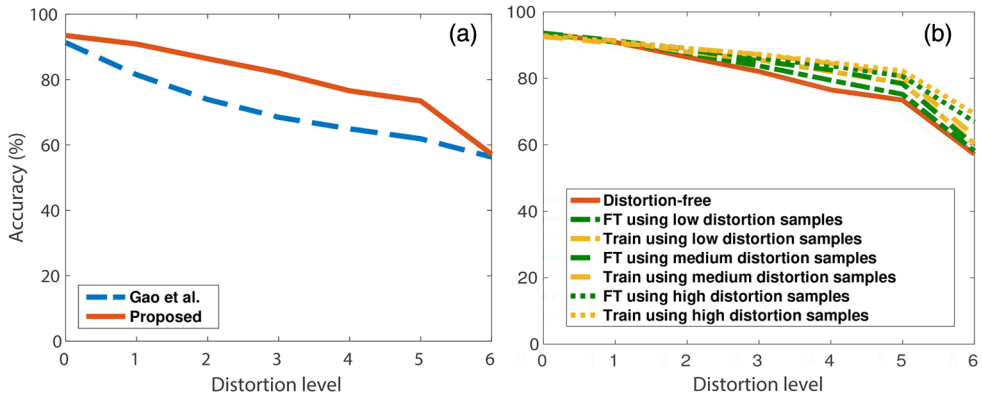

Fig. 3 Classification rates: (a) applying a combination of three artifacts and (b) including artifact-affected images in the training set.

3.3 *Training Adding Artifact-Affected Images*

In this section, we add to the distortion-free training set the artifact-affected images belonging to one of the three aforementioned distortion groups at a time and measure classification robustness across all six distortion levels considered. As for the previous experiments, we use the CNN-A architecture. Two different training setups are considered: in the first one, the already trained CNN-A is fine-tuned,[21] while in the second one, the CNN is trained from scratch. We fine-tune the CNN-A by chopping and retraining from scratch the FC-2 layer.

Results are reported in Fig. 3. From the plots, we want to highlight that: (i) adding images with artifacts does not affect the performance on distortion-free images and on images with low distortion levels, showing the robustness of the CNN to such training data; (ii) adding distorted images in the training set is able to increase robustness with respect to low distortion levels up to 2.7% for both fine-tuned and trained CNNs; and (iii) robustness increases up to 7.3% and 8.3% for medium level distortion levels for fine-tuned and trained CNN, respectively.

4 Conclusion

In this work, we proposed a robust processing pipeline for smile detection in face images acquired in unconstrained scenarios. The proposed pipeline involved detecting faces using a multiview face detector, aligning facial image using an eye-based approach, and predicting whether it is a smiling face or not using an *ad hoc* designed CNN. We investigated the robustness of the method to different kinds of geometric transformations (rotation, translation, and scaling) and to several kinds of distortions (JPEG compression, Gaussian noise, Gaussian blur, and motion blur). On the basis of this evaluation, we foresee several ways for further improvement. Concerning face detection and alignment, we plan to test different face alignment approaches such as from Ref. 22, including also color information.[23] Concerning classification, we would like to refine the GENKI-4k database and investigate a fusion approach with the method by Gao et al.[12] Starting from Ref. 24, we will also consider the simultaneous classification of smile and subject identity.

References

1. P. Ekman, W. V. Friesen, and M. O'Sullivan, "Smiles when lying," *J. Personality Soc. Psychol.* **54**(3), 414–420 (1988).
2. Y. Shinohara and N. Otsu, "Facial expression recognition using Fisher weight maps," in *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 499–504, IEEE (2004).
3. Y. Bai et al., "A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition," in *Int. Conf. on Image Processing (ICIP)*, pp. 3305–3308 (2009).
4. J. Whitehill et al., "Toward practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 2106–2111 (2009).
5. C. Shan, "An efficient approach to smile detection," in *Face and Gesture 2011*, pp. 759–764, IEEE (2011).
6. C. Shan, "Smile detection by boosting pixel differences," *IEEE Trans. Image Process.* **21**(1), 431–436 (2012).
7. Y. Zhang, "A novel approach to detect smile expression," in *IEEE. Int. Conf. on Machine Learning and Applications*, pp. 482–487 (2012).
8. L. An, S. Yang, and B. Bhanu, "Efficient smile detection by extreme learning machine," *Neurocomputing* **149**, 354–363 (2015).
9. T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European Conf. on Computer Vision*, pp. 469–481 (2004).
10. V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, pp. 236–243, Springer (2008).
11. O. Déniz et al., "Face recognition using histograms of oriented gradients," *Pattern Recognit. Lett.* **32**(12), 1598–1603 (2011).
12. Y. Gao et al., "A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios," *Neurocomputing* **174**, 1077–1086 (2015).
13. S. S. Farfade, M. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval*, pp. 643–650, ACM (2015).
14. A. Asthana et al., "Incremental face alignment in the wild," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1859–1866 (2014).
15. "The MPLab GENKI Database, GENKI-4K Subset," http://mplab.ucsd.edu (2009).
16. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. of the 27th Int. Conf. on Machine Learning*, Vol. 3, pp. 807–814 (2010).
17. D. C. Cirean et al., "High-performance neural networks for visual object classification," in *Advances in Neural Information Processing System*, Vol. 12 (2011).
18. S. Bianco, L. Celona, and R. Schettini, "Robust smile detection using convolutional neural networks," www.ivl.disco.unimib.it/activities/smile-detection/ (06 October 2016).
19. D. Jayaraman et al., "Objective quality assessment of multiply distorted images," in *Asilomar Conf. on Signals, Systems and Computer*, pp. 1693–1697, IEEE (2012).
20. S. Bianco et al., "Color correction pipeline optimization for digital cameras," *J. Electron. Imaging* **22**(2), 023014 (2013).
21. A. S. Razavian et al., "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 512–519 (2014).
22. T. Hassner et al., "Effective face frontalization in unconstrained images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE (2015).
23. S. Bianco and R. Schettini, "Adaptive color constancy using faces," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(8), 1505–1518 (2014).
24. K. Zhang et al., "Gender and smile classification using deep convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 34–38 (2016).