Journal of Advanced Computing and Intelligent Systems (JACIS)
https://journal.chainknowpub.com/index.php/JACIS/index
ISSN Online: –
ISSN Print: –

Simulated and Cropped Masked FER2013: A Novel Simulated Dataset for Masked Facial Expression Recognition

Norlina Mohd Sabri

College of Computing, Informatics and Media, Universiti Teknologi MARA, Selangor, Malaysia, shuzlina@uitm.edu.my

Jie Zhang

College of Computer and Mathematics, Xinyu University, Jiangxi, P. R. China, 28966321@qq.com

Abstract: Wearing masks has become the norm after the COVID-19 pandemic, obscuring key facial regions and posing significant challenges to Facial Expression Recognition (FER) systems. Existing FER datasets, such as FER2013, lack masked face images and consist of low-resolution samples, resulting in reduced model robustness under complex occlusions. To address these limitations, this paper introduces an innovative approach that integrates image processing and data augmentation techniques to simulate realistic mask occlusions in the FER2013 dataset. Leveraging the MaskTheFace algorithm, the proposed method automatically generates the Simulated and Cropped Masked FER2013(SCMFER2013) dataset. OpenCV is utilized for facial area detection, Dlib for facial key point localization, and the Albumentations library for data augmentation. Experimental results demonstrate that models trained on the SCMFER2013 dataset achieve substantial performance improvements, particularly under challenging low-resolution and occlusion conditions.

Keywords: Computer Vision; Data Augmentation; Simulated and Cropped Masked FER2013; MaskTheFace

1. Introduction

The COVID-19 pandemic has led to widespread adoption of mask-wearing, which has profoundly affected human interaction and posed unique challenges to computer vision systems. Facial Expression Recognition systems, which rely on detecting facial features to interpret emotions, encounter significant difficulties when masks obscure key facial regions, such as the mouth and nose. This occlusion reduces the accuracy and robustness of FER models, limiting their effectiveness in real-world, post-pandemic environments[7].

Existing FER datasets, such as FER2013, were created before mask-wearing became widespread and do not include masked faces. Additionally, FER2013 consists of low-resolution images (48x48 pixels), further hindering the performance of FER models under complex occlusion conditions . These limitations highlight the need for a dataset and training methodology capable of handling mask occlusion and low-resolution inputs, particularly as mask-wearing continues in many parts of the world[5].

To address these challenges, this paper proposes the creation of a SCMFER2013 dataset, which incorporates realistic mask occlusions into the original FER2013 dataset. The main contributions of this study are as follows:

· Development of SCMFER2013: A simulated dataset incorporating masked faces using MaskTheFace, OpenCV, and Dlib.

· Empirical Validation: Experimental results demonstrating that models trained on SCMFER2013 significantly outperform those trained on the original FER2013, particularly under conditions of mask occlusion and low resolution.

2.Masked FER Dataset Review

Most Masked face recognition datasets have been constructed to address the challenges posed by face recognition technology during the pandemic era. These datasets contain a substantial number of face images with masks, enabling the training of accurate masked face detection models that subsequently serve the task of masked face recognition[2].

MFDD is a comprehensive masked face recognition dataset sourced from both related research and internet crawling, with 24,771 labeled images indicating mask-wearing status and mask position coordinates. It facilitates the training of accurate masked face detection models for subsequent recognition tasks and aids in determining mask compliance, crucial during pandemics. In contrast, RMFRD is the world´s largest real-world masked face dataset, consisting of 5,000 masked and 90,000 unmasked images of public figures, meticulously filtered and annotated using semi-automatic tools. Lastly, SMFRD expands dataset diversity and volume by automatically applying masks to existing large-scale face datasets, such as LFW and Webface, utilizing a custom software based on the Dlib library, resulting in a simulated masked face dataset of 500,000 images.

All of these datasets aim to address the challenges of masked face recognition by providing a comprehensive and diverse collection of masked and non masked facial images. Whether from relevant research, Internet crawling, or through enhancing existing data sets with simulated masked faces, they are tailored to promote the development and training of accurate masked detection and recognition models. Their use in the original extensive and practical face recognition has little significance, and they have not solved the problem of gray scale or pixel difference face recognition.

3.Dataset Generation

3.1Basic Dataset Selection

According to the expression types and data sources of different data sets, we selected the FER2013 dataset as the basic dataset in this study. FER2013 is a facial expression dataset released in the 2013 Kaggle competition "Challenges in Representation Learning: FER Challenge". It contains 35,887 facial images with expression categories annotated. The dataset is divided into three categories, including 28,709 training sets, 3,589 public test sets, and 3,589 private test sets. Each image is composed of a grayscale image with a fixed size of 48×48. There are 7 expressions in total, corresponding to digital labels 0-6, including: 4953 Angry, 547 Disgust, 5121 Fear, 8989 Happy, 6198 Neutral, 6077 Sad, 4002 Surprise. This dataset does not directly give pictures, but saves the expression, picture data, and purpose data into csv files. When using it, the picture data value can be synthesized into pictures through code.The example of each expression in the FER2013 dataset is shown in Figure 1.

The FER2013 dataset, which was sourced from the Internet, is subject to variations in factors such as camera angles, lighting conditions, and other inconsistencies that can influence the dataset´s overall quality. According to a survey conducted on the Kaggle forum, the average recognition accuracy of different methods applied to the FER2013 dataset tends to be approximately 65 ∫ 5%.

3.2 Simulated Masked Dataset Generation

The faces in the FER2013 dataset are all faces without masks. In order to expand the scope of the dataset, MaskTheFace is used to simulate wearing masks to form simulated Masked FER2013 dataset(SMFER2013). MaskTheFace[3] is a computer vision script designed to apply masks to faces in images. It utilizes a face landmarks detector based on dlib to recognize the tilt of the face and six key facial features required for mask placement. Depending on the face´s orientation, an appropriate mask template is chosen from a pre-

existing library. This template is then adjusted using the six key facial features to ensure a precise fit. The flowchart of MaskTheFace is illustrated in the Figure2.

Figure 1 Seven samples of FER2013 dataset

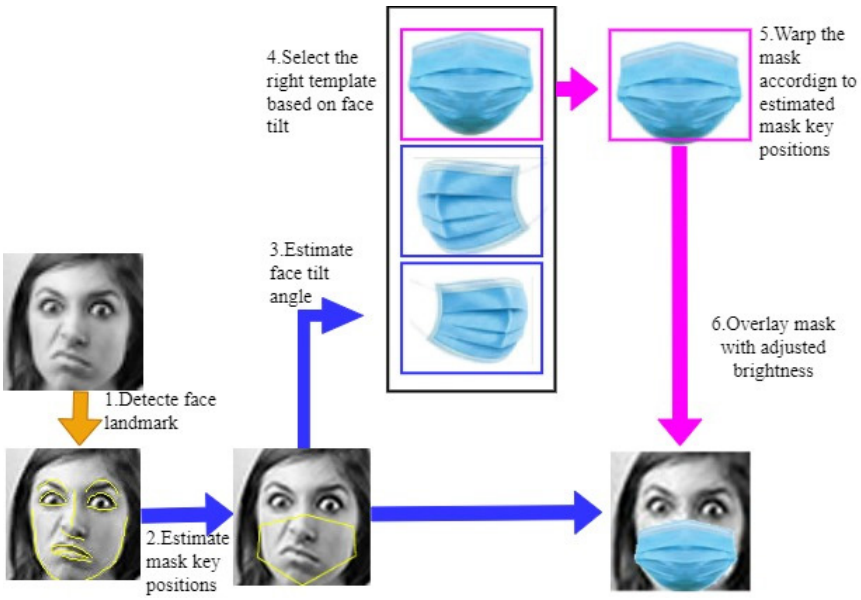

Figure2 The flowchart of MaskTheFace

Additionally, a manual threshold was implemented in the MaskTheFace algorithm, where images with confidence values below this threshold were deemed too difficult to mask and were excluded from the dataset. This explains why the final dataset is smaller than the original one.

Figure3 Seven samples of SMFER2013(not cropping) dataset

The SMFER2013 dataset contains 4314 Angry, 516 Disgust, 4228 Fear, 8248 Happy, 5514 Neutral, 4622Sad, 3672 Surprise emoticons.

3.3 Simulated and Cropped Masked Dataset Generation

To enhance the diversity and generalizability of the SCMFER 2013 dataset, a series of data augmentation techniques will be systematically applied to SMFER2013 dataset. These techniques aim to simulate real-world variations, improve robustness against data imperfections, and address the challenges posed by facial occlusions.

i) Geometric Transformations

Geometric transformations increase the diversity of facial poses and perspectives, thereby reducing model overfitting. Images will be randomly rotated within a range of -15 ° to +15 ° to simulate slight head tilts. Random scaling (between 0.8x and 1.2x) will mimic variations in the distance between the camera and the subject. Random shifts along the horizontal and vertical axes (up to 10% of the image size) will emulate imperfect facial alignments.

ii) Region-Specific Cropping

Region-specific cropping focuses on the upper facial area, which remains visible in masked images. This technique enhances the network´s ability to extract meaningful features from unmasked regions. Identify the upper facial regions, including the eyes, eyebrows, and forehead, as areas of interest. Perform random cropping within the defined upper facial area, with cropped portions covering approximately 70%-90% of the original image. Resize the cropped regions to match the original image dimensions for compatibility with the network㈱s input requirements. The procedure of region-specific cropping is shown in Figure4.

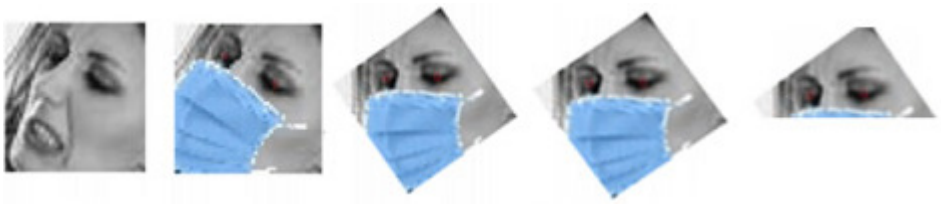

Figure4 the procedure of region-specific cropping

iii) Color Adjustments

Color adjustments enhance robustness to variations in illumination and color conditions[1]. Brightness levels will be randomly altered within ∫ 30% to simulate changes in lighting intensity. Image contrast will be modified within a ∫ 20% range to enhance clarity and variability. Random adjustments within ∫ 15% will account for color distortions in real-world settings.

Figure5 Seven samples of SCMFER2013(cropped) dataset

iv) Noise Injection

Adding noise improves the network㈱s resilience to real-world imperfections, such as sensor errors and compression artifacts. Generate Gaussian noise with a mean of 0 and a standard deviation of 0.01 to simulate pixel-level distortions. Add the generated noise matrix to each pixel of the original image, ensuring pixel values remain within the valid range (0-255). Adjust the intensity of noise injection

based on the dataset㈱s characteristics and augmentation objectives.

The example of each expression in the SCMFER2013 dataset is shown in Figure5.

3.4Comparison of emotion classification of three datasets

After 3.2 we got the mask-wearing dataset SFER2013, and after 3.3 we got the final masked versions of the FER2013 dataset(SCMFER2013).It contains 3048 Angry, 418 Disgust, 3104 Fear, 6510 Happy, 4357 Neutral, 3287 Sad, 2721 Surprise emoticons. Comparison of emotion classification of FER2023,SMFER2013 and SCMFER2013 datasets are reported in Table 1.

Table1 Comparison of emotion classification of FER2023,SMFER2013 and SCMFER2013 datasets

Dataset	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
FER2013	4953	547	5121	8989	6198	6077	4002
SMFER2013	4314	516	4228	8248	5514	4622	3672
SCMFER2013	3048	418	3104	6510	4347	3287	2721

4. Experiments

4.1 Experimental Setup

A.Architecture

Xception network[4] is another improvement to Inception V3 proposed by Google Company after Inception V3. It mainly uses depth-separable convolution to replace the classical convolution operation in Inception V3, and also refers to the residual separable convolution of ResNet. And the structural unit of the network can increase the number of layers of the network while reducing the amount of parameters, which not only reduces the storage space, but also enhances the expressive ability of the mode. Xception mainly includes 36 layers of convolution. it can be divided into 3 flows, namely Entry flow, Middle flow, and Exit flow. it is divided into 14 blocks, including 4 blocks in Entry flow, 8 blocks in Middle flow, and 2 blocks in Exit flow. Xception model is composed of some linear stacks of depth-separable convolution block and ResNet-like residual separable convolution blocks .

The following experiments are developed under four settings: (a) networks are learned on FER2013; (b) networks are learned on SMFER2013; and, (c) networks are learned on SCMFER2013.

B.Training details

The experimental software and equipment configuration parameters are as follows:Ubuntu 18.04, 64-bit Linux operating system, Tesla T4 graphics card with 32 G memory, PyTorch 1.10 and CUDA 10.1.We set the basic parameters:epoch = 70, learning rate = 0.001, batch_size = 64, weight_decay = 0.001, step_size = 25,and gamma = 0.01.

C.Dataset Splitting and Balancing

All datasets will be divided into training, validation, and testing subsets:

· Training Set: 70% of the data, augmented and balanced, will be used for model training.

· Validation Set: 15% of the data will guide hyperparameter optimization and overfitting prevention.

· Test Set: 15% of the data will evaluate the final model´s performance on unseen samples.

To address class imbalance, strategies include weighted loss functions, targeted augmentation for underrepresented categories, and Synthetic Minority Over-sampling to generate additional samples for minor classes.

4.2 Quantitative Analysis.

Table 2 compares the performance of various models on the SCMFER2013 dataset, highlighting the accuracy, precision, recall, and F1-score of each model. Among the networks, Xception achieves the highest accuracy of 71.5%, followed by ResNet-18[6] at 67.2%.

MobileNetV2 performs the weakest with an accuracy of 64.7%. In terms of computational efficiency, Xception remains lightweight, with 0.68M parameters and 30.1M FLOPs, offering a balance between high performance and low resource usage.

Table 2 Comparison of Performance of different networks on SCMFER2013 dataset

Network	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)	FLOPs (M)
Xception	71.5	71.0	71.3	71.2	0.68	30.1
MobileNetV2	64.7	63.5	63.2	63.3	2.23	56.8
ResNet-18	67.2	66.4	66.1	66.3	11.69	98.7

5. Conclusion

In summary, the COVID-19 pandemic has necessitated widespread mask-wearing, posing significant challenges for Facial Expression Recognition (FER) systems due to the obscuring of facial features. The inadequacy of existing FER datasets, such as FER2013, which lack masked face images and contain low-resolution samples, undermines model robustness under complex occlusion conditions. To address these limitations, this paper introduces a novel approach that integrates image processing and data augmentation techniques to simulate realistic mask occlusions within the FER2013 dataset, resulting in the creation of the Simulated and Cropped Masked FER2013 (SCMFER2013) dataset. Experimental results indicate that models trained on the SCMFER2013 dataset exhibit substantial performance improvements, particularly in challenging scenarios characterized by low resolution and occlusion, thereby contributing to the advancement of FER systems in the context of widespread mask-wearing.

Acknowledgement

This work was supported by the Science and Technology Research Foundation of Jiangxi Province Department of Education under Grant No.GJJ212308😞Research on Facial Expression Recognition based on Lightweight Convolutional Neural Network😊

References

[1] Buslaev, A., Iglovikov, V., & Kalinin, A. (2020). Albumentations: Fast and flexible image augmentations. arXiv preprint arXiv:1809.06830.

[2] Cheng, C., Lee, J., & Chen, S. (2021). A study on facial expression recognition with partial occlusion due to mask wearing. IEEE Transactions on Affective Computing, 12(4), 899-909.

[3] Chen, X., & Li, Y. (2020). MaskTheFace: A tool for simulated mask occlusion. Journal of Computer Vision, 24(6), 200-211.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

[5] Li, Y., Deng, Z., & Zhao, Y. (2020). Robust facial expression recognition using deep learning under mask occlusion. Journal of Computer Vision and Image Understanding, 192, 102978.

[6] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[7] Mokhtar, M. R., Yang, X., & Wang, Z. (2022). Enhancing facial expression recognition systems under low-resolution and occlusion conditions. Journal of Visual Communication and Image Representation, 91, 103472.