Occluded Facial Expression Recognition Based on Generative Adversarial Networks (Postprint)

Authors: Wang Suqin, Gao Yudou, Zhang Jiaqi

Date: 2018-09-12T00:00:00+00:00

Abstract

To address the issue that local occlusion affects facial expression recognition in practical applications, we propose a facial expression recognition algorithm based on Generative Adversarial Networks (GAN) that first inpaints and restores occluded facial images before performing expression recognition. The GAN generator is constructed from a convolutional autoencoder, and adversarial learning with the discriminator renders the generated facial images more realistic. The discriminator, composed of a convolutional neural network, possesses strong feature extraction capabilities; the addition of a multi-classification layer constitutes an expression classifier, avoiding the need to recompute image features. To address insufficient training samples, the celebA facial dataset is utilized for training facial inpainting, while simultaneously pre-training the feature extraction component of the expression classifier. Experiments on the CK+ dataset demonstrate that the inpainted facial images are realistic and coherent, achieving high expression recognition accuracy, particularly improving the recognition rate for faces with large-area occlusions.

Full Text

Abstract

Human emotions are primarily conveyed through facial expressions, making facial expression recognition via image processing techniques highly significant. With advances in computer technology and GPU hardware, expression recognition has achieved substantial progress. Convolutional neural networks (CNNs) have achieved excellent recognition accuracy on standard expression datasets. Beyond the common basic expressions, micro-expressions—which are brief yet reveal hidden true emotions—have become a research hotspot, with applications in lie detection, public security, and criminal investigation. However, in practical applications, captured facial images often suffer from partial occlusion by

objects such as hands, glasses, or masks, which interfere with expression feature extraction and affect recognition accuracy.

Current approaches for occluded facial expression recognition generally employ non-deep learning methods categorized into two main strategies: discard-based and inpainting-based methods. Discard-based methods simplify or discard occluded information through sparse representation, performing recognition based solely on unoccluded regions. While effective, this approach is unreasonable when critical facial regions containing rich expression information (eyes, mouth, nose) are occluded. Inpainting-based methods first reconstruct the occluded regions to approximate the original unoccluded state before performing expression recognition.

1. Introduction

Deep learning methods, particularly CNNs, have achieved remarkable success in image recognition and object classification, largely dependent on large-scale manually annotated training datasets. However, in many applications, labeled data is insufficient for deep model training. Semi-supervised learning addresses this by incorporating unlabeled samples into training. Generative Adversarial Networks (GANs) represent a current research hotspot in artificial intelligence, consisting of a generator (G) and a discriminator (D). The generator learns the distribution of real samples to produce realistic outputs, while the discriminator learns to distinguish real from generated samples. Through adversarial training, both components iteratively optimize, enhancing generation and discrimination capabilities.

This paper proposes a GAN-based approach for occluded facial expression recognition. Given an occluded facial image, the model first synthesizes the missing content to produce a completed image, then correctly classifies the expression. The overall architecture is shown in [Figure 1: see original paper], where black rectangular boxes simulate occlusions. Unlike conventional GANs that use random noise as input, our generator takes occluded facial images as input, functioning as a convolutional autoencoder comprising an encoder and decoder.

2. Methodology

2.1 Generator Architecture

The encoder maps the input occluded image to a hidden representation through multiple convolutional and pooling operations, capturing implicit relationships between known and missing regions. The decoder leverages this information to generate the inpainted content. The generator network structure is symmetric, with encoder and decoder performing inverse operations. The encoder follows the VGG model structure, while the decoder uses fractional-strided convolutions for upsampling. The network is fully convolutional, accepting input images of arbitrary size.

The reconstruction loss is defined as the L2 norm difference between the generated image and the original unoccluded image at corresponding pixel positions. To enhance realism, we introduce adversarial loss: the generator G aims to produce images that fool the discriminator D, while D learns to accurately distinguish real from generated images. The discriminator architecture provides strong feature extraction capabilities, making it suitable for integration with the expression classifier. The adversarial loss follows the standard formulation, where the discriminator distinguishes between real facial images and generated completions. In the classifier, we employ cross-entropy loss to train the expression multi-classifier, which measures the distance between the predicted probability distribution and the ground-truth labels. The total loss function combines reconstruction loss, adversarial loss, and classification loss, with weighting factors to balance their contributions. The experiments are conducted on a standard computing platform with sufficient memory resources.

2.2 Discriminator and Classifier Integration

The discriminator in our model serves a dual purpose: distinguishing real from generated images and extracting features for expression classification. We integrate the feature extraction components of the discriminator and classifier, leveraging the discriminator's strong feature representation capabilities for multi-classification tasks. The generator' s outputs also serve to augment the training data, improving the classification model' s generalization.

The discriminator architecture uses $3\$\times 3 convolution kernels across 16 layers (13 convolutional and 3 fully connect$ kernel networks used in previous work, enables better integration with the expression classifier.

3. Experiments

3.1 Dataset

We evaluate our model on the CelebA dataset, which contains 202,599 facial images of 10,177 celebrities from the Chinese University of Hong Kong. Face regions are cropped and used as our base training dataset, randomly split into 4/5 for training and 1/5 for testing. Due to CelebA' s limited size, we perform data augmentation by introducing occlusions of varying sizes and positions.

3.2 Preprocessing and Occlusion Simulation

To minimize interference from illumination and pose variations, we apply several preprocessing steps: grayscale conversion using weighted averaging, histogram equalization, and face alignment using OpenFace' s algorithm with Ensemble Regression Trees (ERT) to estimate facial landmarks. Images are normalized to $128\$\times\$128$ pixels.

We simulate two types of occlusions: systematic occlusion (fixed positions) and temporary occlusion (random positions). Occlusions are generated by adding

black rectangular boxes covering 10% to 70% of the image area, as shown in [Figure 2: see original paper]. This augmentation expands the training set and improves model robustness.

3.3 Training Procedure

Training proceeds in two stages: first, the face completion network is trained on CelebA using the Adam optimizer with an initial learning rate of 0.0002 and batch size of 64. When the model approaches Nash equilibrium (discriminator accuracy 0.5), we freeze the generator and train the expression classification network on the CK+ dataset.

3.4 Results and Comparison

Qualitative inpainting results are shown in [Figure 3: see original paper], demonstrating that our method produces coherent, realistic facial images across various occlusion types (random 10%, 20%, 40%, eye occlusion, mouth occlusion) and expressions (neutral, contempt, surprise, happiness). The completed images appear natural and continuous, with only minor changes to unoccluded regions.

Quantitative comparison with existing methods—including Sparse Representation-based Classification (SRC), CNN, and DCGAN-CNN—shows our approach achieves superior recognition rates, especially for large-area occlusions. While DCGAN-CNN also performs inpainting, its results lack coherence, negatively impacting expression recognition. Our method maintains over 80% accuracy even with 70% occlusion, significantly outperforming other approaches.

4. Conclusion

This paper presents a GAN-based method for occluded facial expression recognition that integrates face completion and expression classification. By combining the discriminator' s feature extraction with the classification task and leveraging adversarial training for realistic inpainting, our approach substantially improves recognition accuracy for partially occluded faces, particularly in cases of large-area occlusion. Future work will explore further enhancements to model architecture and training strategies.

References

[1] Sun Xiao, Pan Ting, Ren Fuxi. Facial expression recognition using ROI-CNN deep convolutional neural networks. *Computer Applications Research*, 2016, 33(12): 3843-3846.

[2] Li Yijun, Liu Sifei, Yang Jimei, et al. Generative face completion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 5892-5900.

[3] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[4] Denton E, Gross S, Fergus R. Semi-supervised learning with context-conditional generative adversarial networks. *arXiv preprint arXiv:1611.06430*, 2016.

[5] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 2016: 2226-2234.

[6] Lopes RG, Figueira G, Oliveira-Santos T. Facial expression recognition with occlusion. *Technical Report*, 2008: 07-49.

[7] Cottes SF. Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers. *Proceedings of Digital Signal Processing*, 2014.

[8] Yang Shuo, Luo Ping, Loy CC. From facial parts responses to face detection: A deep learning approach. *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 3676-3684.

[9] Zhu Minghan, Li Shutao, Ye Hua. Multi-modal learning for facial expression recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2014: 7890-7898.

[10] Baltrušaitis T, Robinson P, Morency LP. OpenFace: An open source facial behavior analysis toolkit. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016.

[11] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1867-1874.

[12] Liu Jiagang, Ran Junjun, Li Jiang. Facial expression recognition based on principal component analysis and support vector machine applied in intelligent tutoring systems. *Computer Applications Research*, 2012, 29(8): 3166-3168.

Note: Figure translations are in progress. See original paper for figures.