Evaluation of Generative Adversarial Network Generated Super Resolution Images for Micro Expression Recognition

Pratikshya Sharma[1], Sonya Coleman[1], Pratheepan Yogarajah[1], Laurence Taggart[2] and Pradeepa Samarasinghe[3]
[1]School of Computing, Engineering & Intelligent Systems, Ulster University, Northern Ireland, U.K.
[2]School of Nursing & Health Research, Ulster University, Northern Ireland, U.K.
[3]Department of Information Technology, Malabe, Sri Lanka

Keywords: Micro Expression, General Adversarial Network, Super Resolution.

Abstract: The Advancements in micro expression recognition techniques are accelerating at an exceptional rate in recent years. Envisaging a real environment, the recordings captured in our everyday life are prime sources for many studies, but these data often suffer from poor quality. Consequently, this has opened up a new research direction involving low resolution micro expression images. Identifying a particular class of micro expression among several classes is extremely challenging due to less distinct inter-class discriminative features. Low resolution of such images further diminishes the discriminative power of micro facial features. Undoubtedly, this increases the recognition challenge by twofold. To address the issue of low-resolution for facial micro expression, this work proposes a novel approach that employs a super resolution technique using Generative Adversarial Network and its variant. Additionally, Local Binary Pattern & Local phase quantization on three orthogonal planes are used for extracting facial micro features. The overall performance is evaluated based on recognition accuracy obtained using a support vector machine. Also, image quality metrics are used for evaluating reconstruction performance. Low resolution images simulated from the SMIC-HS dataset are used for testing the proposed approach and experimental results demonstrate its usefulness.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

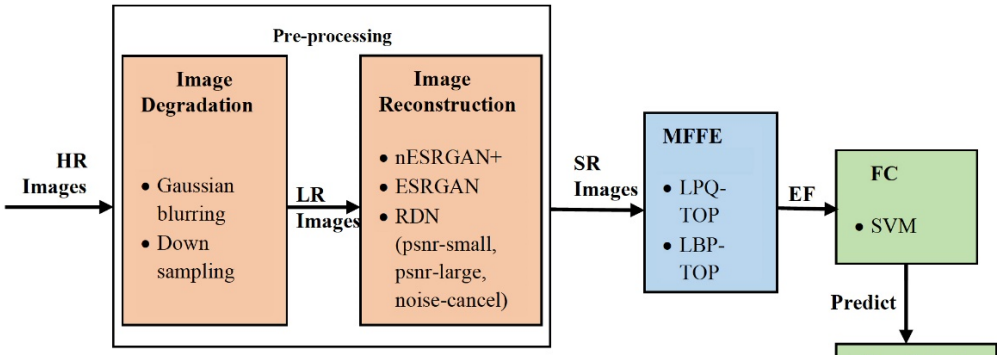
Expressing emotion is a habitual means of putting across one's feelings and generally comes naturally to humans. Verbal or non-verbal expressions of such emotions play a vital role in perceiving a human's mindset. Such emotions when channelled facially are commonly known as facial expressions and identified as non-verbal expressions. By gauging the extent to which an expression lingers on the face, determines whether it is a macro or micro expression. Macro expressions last for longer than micro expressions, typically beyond the micro expression duration of 0.04 to 0.2 of one second (Liong, See, Wong, & Phan, 2018). Controlling the appearance of micro expressions on the face is extremely difficult for an individual which makes it appear more natural, thereby contributing for accurate estimation of one's emotion (Ekman & Friesen,1969) (Ekman,2009). The process involved in recognizing such fleeting expressions is comparatively much more difficult than macro expressions. However, recent trends clearly demonstrate successful advancements of recognition techniques using both shallow and deep learning methods for facial micro expression (Oh, See, Le Ngo, Phan, & Baskaran, 2018). Micro Expression Recognition (MER) is a substantially widespread inter-disciplinary application area. Some application areas include autism, psychology, crowd scenarios, airport security and criminal investigations (Ekman & Friesen,1969) (Ekman,2009) (Oh et al., 2018). Such diverse application areas implicitly spawn situations where images to be analysed are of poor quality. The quality may be affected due to certain realistic situations like images acquired in poor lighting conditions, images taken using low-cost imaging devices, downloaded images, images stored in restricted memory capacity etc. Such images have insufficient resolution which makes it extremely difficult for both humans and machines to utilize the available information. These situations have given rise to a new research direction that needs to cater to the low resolution (LR) problem particularly for micro expressions (Zhao & Li, 2019) (Li, G., Shi,

Sharma, P., Coleman, S., Yogarajah, P., Taggart, L. and Samarasinghe, P.
Evaluation of Generative Adversarial Network Generated Super Resolution Images for Micro Expression Recognition.
DOI: 10.5220/0010820100003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 560-569
ISBN: 978-989-758-549-4; ISSN: 2184-4313
Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

Peng, & Zhao, 2019). Dispersed and loosely aligned pixels in LR images result in comparatively fewer image details within them than standard resolution images. This obviously makes them appear pixelated, less precise, blurrier, and granular. On the other hand, a more concentrated and compact pixel arrangement makes high resolution (HR) images appear crisper and clearer. Implicitly, such images contain denser image details. Broadly, images with LR differ from HR images mainly in terms of pixel density per unit area and degree of coherence. A substantial lack of salient information (e.g., texture details, high frequency information etc.) in LR images makes the process of attribute extraction extremely challenging and laborious. The task of estimating a HR image by reconstructing an image from a LR input image is generally known as image super resolution (ISR) and the reconstructed image is known as a super-resolved image. Several innovative deep convolutional neural networks (e.g., CNNs) are now available with variations that exploit residual dense networks (RDN), residual dense blocks (RDB) and recursive learning architectures (Ledig et al., Jul 2017) (Hung, Wang, & Jiang, 2019) (Wang et al., 2019) (Zhang, Tian, Kong, Zhong, & Fu, Jun 2018), and have been successfully applied to super resolution (SR) problem. The main motive behind super resolving LR facial images is to recover essential facial details. For face SR algorithms, the challenge is not only to reconstruct the face, but also to maintain attribute consistency with the original HR images. Thus, restoring face details in the reconstructed image is vital for face SR algorithms, to facilitate facial expression analysis. Variation in facial expressions among different classes of micro expression is very limited. These expressions are very subtle and have comparatively less distinct inter-class discriminative attributes. Lack of explicit inter-class attributes in micro expression, in addition to inadequate availability of information due to LR, along with the absence of a suitable LR micro expression dataset further increases the difficulty level of the overall MER task (Li et al., 2019). Some popular feature extraction techniques that have been employed for micro expression include Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) (Yan et al., 2014), Histogram of Oriented Gradient on Three Orthogonal Planes (HOG-TOP) (Li, X. et al., 2018), Histogram of Image Oriented Gradient on Three Orthogonal Planes (HIGO) (Li et al., 2018), Local Binary Patterns with Six Intersection Points (LBP-SIP) (Wang, Y., See, Phan, & Oh, 2015) and Local Phase Quantization on Three Orthogonal Plane (LPQ-TOP) (Zong et al., 2018) (Sharma, Coleman, & Yogarajah, 2019; Sharma, Coleman, Yogarajah, Taggart, & Samarasinghe, Jan 10, 2021)

To address some of the problems of LR micro expressions discussed earlier, (Li et al., 2019) proposed reconstructing higher resolution images from LR images by employing a face hallucination algorithm on individual frames. At present datasets available for micro expression contain only HR images. For instance, Spontaneous Micro-expression database (SMIC-HS) (Xiaobai Li, Pfister, Xiaohua Huang, Guoying Zhao, & Pietikainen, Apr 2013) micro expression dataset contains HR images with resolution of 190 x 230 approximately, whereas LR images are usually below 50 x 50 resolution (Li et al., 2019). Hence in their work (Li et al., 2019), LR micro expression image dataset was obtained by simulating three existing HR micro expression image datasets i.e., CASMEII (Yan et al., 2014), Spontaneous Micro-expression database (SMIC-HS) and SMIC-subHS (Xiaobai Li et al., Apr 2013). Through experimental results an improvement on overall classification accuracy was achieved on these datasets. However, low accuracy for individual classes was also observed alongside this. From the results obtained, a drastic decline in the recognition accuracy was observed for expressions with exceptionally low resolution. Datasets from CASMEII and SMIC-HS yielded higher magnitude of misclassification than SMIC-subHS. It was observed that the reliability and validity of any facial expression analysis approach is directly affected by the resolution of the input image used hence acquiring decent resolution for the reconstructed facial micro expression images was crucial when employing SR. Apart from (Li et al., 2019), work involving face SR for expression analysis has employed macro expressions, thus SR on micro expression is a unique concept introduced by (Li et al., 2019).

Taking this concept further, our work attempts to introduce deep learning technique into the LR micro expression recognition framework. Specifically, we propose Generative Adversarial Network (GAN) (Goodfellow et al., 2014) technique and its variant and evaluate its performance in solving the issue of low resolution targeting micro expression. At present GAN has not been applied specifically for low resolution micro expression problem, this work is a first attempt to realise it. The proposed recognition framework aims to combine the best features from handcrafted methods and deep learning techniques. Low resolution ME images obtained by simulating data from SMIC-HS are used to test our proposed approach. SR algorithms can be applied to both videos and images with LR to obtain its



Note:
EF : Extracted features
ESRGAN : Enhanced super resolution generative adversarial network
FC : Feature classification
GAN : Generative adversarial network
HR: High resolution
ISR : Image super resolution
LBP-TOP – Local binary pattern on three orthogonal planes
LPQ-TOP – Local phase quantization on three orthogonal planes
LR : Low resolution
MFFE : Micro facial feature extraction
RDN : Residual dense network
SR : Super resolution
SVM : Support vector machine

Figure 1: Proposed framework employing GAN based image reconstruction before microexpression recognition.

corresponding super-resolved videos and images. Alternatively, SR images can also be estimated using LR video. In our work we consider using LR images only, throughout the reconstruction process. For extracting micro facial features, we use LBP-TOP and LPQ-TOP along with a support vector machine (SVM) (Chang & Lin, 2011) for classifying data into various emotion classes like positive, negative and surprise.

The main contributions of this work are: (1) introducing GAN based model as a solution to the low-resolution MER problem; (2) a comprehensive analysis of the LBP-TOP & LPQ-TOP extraction techniques for SR and LR images at various resolutions; (3) providing an exhaustive performance evaluation of GAN and its variants for MER task.

2 THE PROPOSED MICRO EXPRESSION RECONSTRUCTION & RECOGNITION FRAMEWORK

To deal with the issue of low resolution in images consisting of micro expressions, we propose a framework with GAN based super resolution image reconstruction module along with three other standard modules namely, image degradation, micro facial feature extraction and feature classification as shown in Figure 1. The initial two modules, image degradation and image reconstruction, are also known as pre-processing modules. They are employed to prepare the datasets for the recognition task. This is followed by a micro facial feature extraction module where the essential micro features are extracted from the input facial images. As a final step, the classification module manages the task of assigning appropriate class labels based on the extracted features. This entire process is termed as the micro expression recognition system. All these modules are briefly discussed in this section.

2.1 Image Degradation

To test our proposed framework, we use a popular micro expression dataset: SMIC-HS. As discussed earlier, this dataset contains HR images, therefore the SR algorithm cannot be used on them directly. These algorithms are suitable for LR images hence, we simulate a LR micro expression dataset from these existing HR micro expression datasets. To construct deteriorated images, we apply down sampling and Gaussian blurring on the existing HR micro expression images of SMIC-HS dataset. This process of creating images with loss of quality from its HR images is known as image degradation and can be expressed using equation (1) (Li et al., 2019). The

symbol S denotes down sampling, Y denotes blurring, Z denotes high resolution, x is other additive noise and L represents the LR image obtained.

$$L = SYZ + x \quad (1)$$

Images produced using such a degradation model appear blurrier and lose their quality. These effects along with the extent of loss of image details in the degraded images can be observed in resultant images presented in section 3. Thus, it is evident that both HR and LR images differ in terms of quality, as well as resolution. These newly created SMIC-HS datasets with degraded images, simulated from existing HR datasets are now suitable to be used with SR algorithms.

2.2 Image Reconstruction

Here, we aim to upscale LR images using SR techniques. To do so, we have considered five different models namely Enhanced Super Resolution Generative Adversarial Network (ESRGAN) (Wang et al., 2019), further improving Enhanced Super Resolution Generative Adersarial Network (nESRGAN+) (Rakotonirina & Rasoanaivo, May 2020), basic model based on peak signal to noise ratio (PSNR) namely, psnr-large and psnr-small (Zhang et al., Jun 2018) (Ledig et al., Jul 2017) and artefact cancelling Generative Adversarial Network model (noise-cancel) (Zhang et al., Jun 2018). In our work these super resolution algorithms employ LR images, simulated from the SMIC-HS dataset. Ideally GAN based super resolution algorithm consists of three vital components namely generator, discriminator, and a loss function. The generator function estimates a HR image for the corresponding LR image whereas, the discriminator estimates if the generated image is realistic enough. Additionally, the loss function essentially helps in optimizing the overall GAN framework. A brief description of these five models is presented in this section.

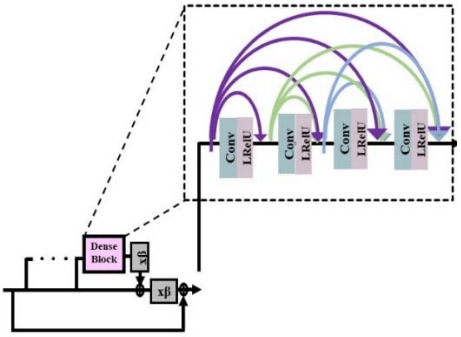


Figure 2: Residual in Residual Dense Block architecture (RRDB) (Wang et al., 2018).

2.2.1 Enhanced Super Resolution Generative Adversarial Network (ESRGAN)

During implementation, to achieve upscaling, a deeper network is employed in ESRGAN using residuals in the residual dense network (RRDN) with its architecture as described in (Wang et al., 2019). Fundamentally RRDN is built using several residuals in a residual dense block (RRDB) (see Figure. 2). Further, each RRDB consists of several residual dense blocks (RDB). The architecture used to build the residual dense network (RDN) consisting of RDB is adapted from (Zhang et al., Jun 2018) (see Figure. 3 & 4). The RDN was introduced for exploiting the hierarchical features from all its convolution layers which had not been accomplished by previous CNN based SR algorithms. RDN is composed of RDB which is a densely connected CNN capable of extracting ample local features. Also, due to a contiguous memory (CM) mechanism between several RDBs, it is capable of continuously learning from more improved features. To achieve this, every layer of the current RDB is fed with the state of the preceding RDB. If F_{d-1} denotes the input and F_d denotes the output for d^{th} RDB with G_0 feature maps for both input and output, then the output of its c^{th} convolution layer is given by (Zhang et al., Jun 2018):

$$F = \sigma\left(W_{d,c}\left[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}\right]\right) \quad (2)$$

In equation (2), the activation function is rectified linear unit (ReLU) denoted by σ, the weight for c^{th} convolutional layer is represented by $W_{d,c}$, and the feature maps produced by (d-1) RDB in its concatenated form is given by $\left[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}\right]$. The convolutional layers in the d^{th} RDB are 1, …., (c-1) which produce a feature map G_0+(c-1) x G, where G represents feature map for $F_{d,c}$. Thus, a direct connection is established between successive layers and the output of each layer and RDB. To downsize the extracted features the d^{th} RDB is directly fed with (d-1)th RDB feature maps in concatenated form. Also, the output information is regulated using a 1x1 convolutional layer. Since the architecture is a combination of local residual learning and dense connections, this architecture is often known as a residual dense block. To extract shallow features RDN uses two convolutional layers. A composite function used in RDN involves operations like convolutional and rectified linear unit (ReLU) (Zhang et al., Jun 2018) (see Figure 3). Meanwhile, global features are also extracted exploiting the hierarchical structure using similar procedures. The discriminator network used consists of a custom architecture from (Ledig et al., Jul 2017)

and the deep features are extracted using Keras VGG19 network. Upscaled images obtained from their corresponding LR images with the process described here are all super-resolved images. Instances of such super-resolved images obtained are presented in section 3.

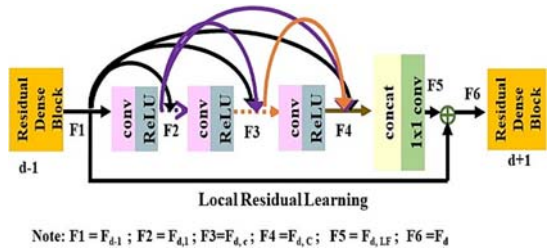


Figure 3: Architecture of Residual Dense Block (RDB) (Zhang et al., Jun 2018).

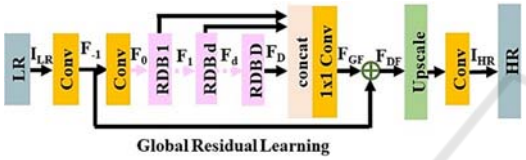


Figure 4: Architecture of Residual Dense Network (RDN) (Zhang et al., Jun 2018).

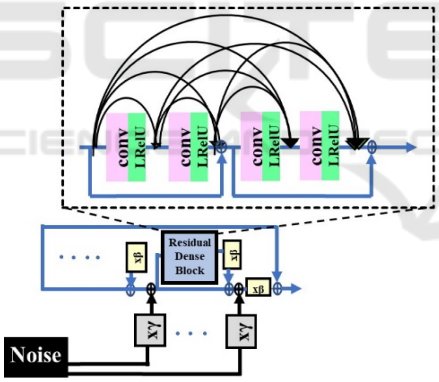


Figure 5: Architecture used in nESRGAN+ (Rakotonirina & Rasoanaivo, May 2020).

2.2.2 Further Improving Enhanced Super Resolution Generative Adversarial Network (nESRGAN+)

To improvise the existing ESRGAN model, two changes were introduced. Firstly, in the existing dense block of ESRGAN model a residual learning was added to enhance the network capacity (see Figure 5). Secondly, the architecture of the generator was modified by injecting noise at input to explore the impact of stochastic variations (see Figure 5). This improvised model was named as nESRGAN+ which aimed at producing more realistic super-resolved images with further image details. This SR model for our work is taken from (Rakotonirina & Rasoanaivo, May 2020).

2.2.3 Basic Model with Residual Dense Network (RDN)

The architecture of this model consists of residual dense block (RDB) where a preceding RDB is directly connected to current RDB (see Figure 3). This type of dense network utilizes hierarchical features from all the convolutional layers and is termed as residual dense network (RDN) (see Figure 4). To extract low-level features RDN uses two convolutional layers. After extraction of local and global features up-sampling is achieved. An abstract view of the procedure followed in this model is depicted in Figure 3 and Figure 4 (Zhang et al., Jun 2018). For creating the RDN model using a PSNR driven approach two different training methods were performed namely psnr-small and psnr large. In the psnr-small method RDN model was built by training the network on image patches with low PSNR value. Similarly, in the psnr-large method the network was trained on image patches with high PSNR value. These methods will be referred to as psnr-large and psnr-small in our work hereon. The architecture employed in this method is taken from (Zhang et al., Jun 2018).

2.2.4 Artefact Cancelling Generative Adversarial Network

In addition to PSNR driven approach, another set of RDN model was built, trained on adversarial and VGG features losses known as artefact cancelling GANs model (noise-cancel) (Ledig et al., Jul 2017; Zhang et al., Jun 2018). Here, an adversarial component is added to the loss function that exploits GAN based training approach. The final weights are obtained by combining weights produced in different training sessions. The training is performed on different datasets along with VGG19 perceptual loss. Unlike PSNR driven approach which focuses on pixel level reconstruction, this method focuses its reconstruction at perceptual level. The discriminator network employed in this model is taken from (Ledig et al., Jul 2017). This method will be referred to as noise-cancel in our work hereon.

2.3 Micro Facial Feature Extraction

For extracting micro facial features, we have employed two methods: Local Binary Pattern on Three Orthogonal Plane (LBP-TOP) and Local Phase

Quantization on Three Orthogonal Plane(LPQ-TOP). LBP-TOP is widely popular for micro expression and has been used for baseline evaluation (Yan et al., 2014). LPQ-TOP method was implemented in (Zong et al., 2018) for designing cross database micro expressions. Exploiting this idea further, (Sharma et al.,2019; Sharma, Coleman, Yogarajah, Taggart, & Samarasinghe, Jan 10, 2021) tested its effectiveness as a micro feature extraction technique with a positive outcome. Thus, building on these groundworks, in this research micro facial features were extracted by conducting two sets of experiments. In the first set of experiment, using LPQ-TOP technique features were extracted from images with micro expression. As the name suggests this technique uses local phase information computed using Short Term Fourier Transform to describe the image textures defined by equation (3) for a given rectangular M x M neighbourhood, $\mathcal{N}x$ (Heikkila & Ojansivu, Aug 2009; Ojansivu & Heikkilä, 2008; Päivärinta, Rahtu, & Heikkilä, 2011):

$$F(u,\mathrm{x}) = \sum_{\mathbf{y}\in N_x} f(x-y)e^{-j2\pi u^T y} = w_u^T f_{\mathrm{x}} \quad (3)$$

For a frequency u its basis vector is denoted by w_u and f_x represents the vector containing M x M samples taken from $\mathcal{N}_x$. At four frequency points the local Fourier coefficients are computed denoted as : u_1=[a , 0]^T , u_2=[0 , a]^T , u_3=[a , a]^T , u_4=[a , -a]^T , with a as a scalar frequency. Ultimately resulting in a vector F_x for every pixel position, given as:

$$F_x = [F(u_1,x), F(u_2,x), F(u_3,x), F(u_4,x)] \quad (4)$$

For each component in F_x, its real and imaginary parts are examined for its sign by applying a scalar quantizer given in equation (5).

$$q_j = \begin{cases} 1, & \text{if } g_j \geq 0 \text{ is true} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

With binary coding given in equation (6), the resultant 8-bit binary coefficient $q_j(x)$ is then represented in the form of integers.

$$f_{LPQ}(x) = \sum_{j=1}^{8} q_j 2^{j-1} \quad (6)$$

The LPQ features are extracted from three planes i.e., XT,YT and XY and stacked into a histogram to be used later.

In the second set of experiments the LBP-TOP method was used to extract the desired micro facial features. For a center pixel, c with cordinates (x_c, y_c), with P neighbouring pixel at R radius, the LBP is computed using equation (7) and (8) (Yan et al., 2014).

$$LBP_{\mathrm{P,R}}(x_c, y_c) = \sum_{p=0}^{\mathrm{P}-1} s(\mathrm{g}_p - \mathrm{g}_\mathrm{c}) 2^p \quad (7)$$

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad (8)$$

In equation (7), g_p and g_c denote grey values for the neighbour pixel and centre pixel respectively , 2^p represents the weight on the neighbouring pixel at p^{th} location where, $p = 0, \ldots, \mathrm{P}-1$ and $s(x)$ in equation (8) manages the sign issue. The LBP computation is performed on all three planes i.e., XY, YT and XT planes, which are later concatenated to form a single feature vector.

2.4 Feature Classification

To keep the classification process simple, in our work we have chosen to use a supervised learning-based method named Support Vector Machine (SVM) (Chang & Lin, 2011). In our experiment, datasets were separated into two sets namely training and testing. Individual instances of data in the training set contain several attributes and class labels. Based on this information and test data attributes, the SVM builds a model capable of predicting class labels for each instance of such test data. Throughout the experiments conducted in this work classification is achieved using a multi-class SVM classifier. To identify the right hyperplane that best differentiates various classes, it uses the kernel trick. The three most popular kernel functions used with SVM are linear, polynomial, and radial basis function denoted by equation (9), (10) and (11) respectively (Chang & Lin, 2011).

$$K(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{x}_i^T \mathrm{x}_j \quad (9)$$

$$K(\mathrm{x}_i, \mathrm{x}_j) = (\gamma\, \mathrm{x}_i^T \mathrm{x}_j + r)^d, \gamma > 0 \quad (10)$$

$$K(\mathrm{x}_i, \mathrm{x}_j) = \exp(-\gamma\, ||\mathrm{x}_i - \mathrm{x}_j||^2), \gamma > 0 \quad (11)$$

The symbols d, r and γ used in these three equations are called the kernel parameters, $(\mathrm{x}_i, \mathrm{x}_j)$ represent training samples and T denotes the transpose operation. In equation (10), d refers to polynomial degree, r is coefficient, and γ in equation (10) and (11) is the gamma parameter that describes the scale of influence of each training sample.

3 EXPERIMENTS RESULTS AND DISCUSSION

As stated earlier, experiments were performed using the SMIC-HS datasets which contains 164 micro

expression samples identified as 51 positive, 70 negative and 43 surprise class labels. The facial resolution of these HR images is 190 x 230 approximately. To create images suitable for the super resolution algorithm we simulate LR dataset by applying degradation method described in section 2.1. Details of implementations and parameters used for various SR models and feature extraction methods in our work are presented in this section.

(a) (b) (c)

Figure 6: (a) Instance of HR 128x128.(b) LR image at 64x64 and (c) LR image at 32x32 simulated from SMIC-HS dataset by applying image degradation.

3.1 Implementation Details

Following the work by (Li et al., 2019) the resolution of the original HR images is set to a standard size of 128 x128 (see Figure 6. (a)) for obtaining LR images. Particularly, these HR images are down sampled by 2 and 4 times to obtain images of size 64x64 and 32x32 (see Figure 6. (b)&(c)), referred to as LR64 and LR32 respectively from hereon. Instances of these LR64 and LR32 are then upscaled using different super resolution algorithm with scaling factor set to two (x2) and four (x4) respectively as described in section 2.2 to obtain 128 x 128 super resolved images (see Figure.7), to be referred as SR64 and SR32 respectively from hereon.

In the basic model with RDN architecture parameter D refers to number of RDB, C refers to number of convolutional layers that are stacked inside a RDB, G refers to number of feature maps of every convolution layer that exists in RDBs, G_0 refers to the output filters i.e., number of feature maps for convolutions that are outside of RDBs and of every RDB output. The values for these parameters employed in the psnr-large model is C=6, D=20,G=64, G_0=64 and x2. For psnr-small model parameter values used were C=3,D=10, G=64, G_0=64 and x2. In the noise-cancel model parameters were set to C=6, D=20,G=64, G_0=64 and x2. Thus, using psnr-small, psnr-large, and noise-cancel models three sets of SR64 datasets was obtained.

To achieve super resolution with ESRGAN, four convolution layers are used in each RDB(C=4) , three RDB inside each RRDB (D=3) and ten RRDB(T=10). Each RDB further consists of 32 convolutional output filters (G) along with 32 output filters (G_0) for each RDB. For ESRGAN training, the learning rate is set to 0.0004, with a decay frequency of 100 and 0.5 decay factor.

Table 1: PSNR and SSIM.

Method	Resolution	PSNR	SSIM
ESRGAN	SR64	35.79	0.9826
psnr-large	SR64	36.58	0.9789
psnr-small	SR64	**37.41**	**0.9827**
noise-cancel	SR64	30.38	0.9412
ESRGAN	SR32	29.5	0.8502
nESRGAN+	SR32	23.08	0.7601

Note: SR64 denotes upscaling from 64 to 128 (x2 scale factor) and SR32 denotes upscaling from 32 to 128 (x4 scale factor).
Bold indicates best value obtained in our work.

Table 2: Micro expression recognition accuracy obtained at various resolutions using different super resolution models on SMIC-HS dataset.

Resolution		SR Method	Accuracy % (our)		Others (Li et al., 2019)
			LBP-TOP	LPQ-TOP	
HR	128	-	50.06	**52.43**	50.00
SR	64	ESRGAN	51.43	**52.43**	52.44
	64	psnr-large	50.67	52.00	
	64	psnr-small	**51.45**	**52.43**	
	64	noise-cancel	49.39	51.82	
	32	ESRGAN	49.82	50.60	51.83
	32	nESRGAN+	49.24	50.00	
LR	64	-	49.2	49.39	50.00
	32	-	44.25	**48.17**	46.95

Note: SR64 denotes upscaling from 64 to 128 (x2 scale factor) and SR32 denotes upscaling from 32 to 128 (x4 scale factor).
Bold indicates best values obtained in our work.

The weight of the loss function is set to 1 for the generator and 0.003 for the discriminator during training. The GAN is optimized using Adam with β1 set to 0.9 and β2 set to 0.999 during this training phase. The discriminator is implemented with a kernel size set to 3 and α set to 0.2 in LeakyReLU. The size of all convolution layers is kept as 3x3 throughout the experiment. However, for local and global feature fusion its size is set to 1x1. Using this ESRGAN model two sets of super resolved datasets were obtained i.e., SR64 and SR32 by setting scale factor to 2 and 4 respectively. For basic model and

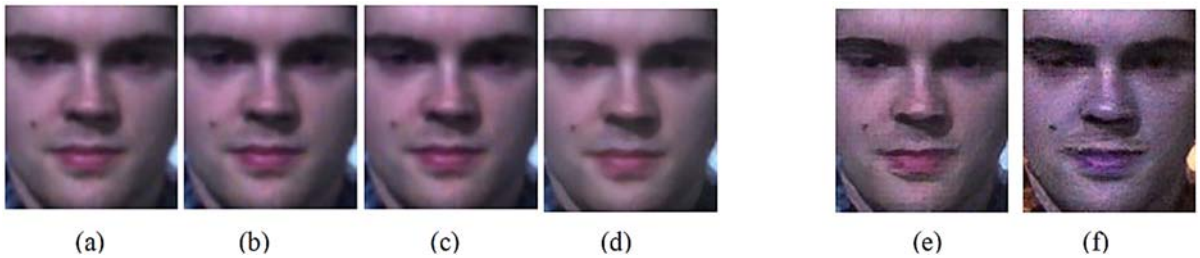


Figure 7: Super resolved images (128 x128) generated by ESRGAN model (a) & (e), psnr - large basic model (b), psnr-small basic model (c), artefact cancelling GAN model (d) and nESRGAN+ model (f).

ESRGAN implementation the parameter settings are taken from (Francesco et al., 2018).

The nESRGAN+ uses Adam optimizer with initial learning rate set to 10^{-4}, 3x3 convolutional kernels, 10 residual blocks and x4. Most of the parameter settings of ESRGAN were kept intact while implementing nESRGAN+. With this model one set of super-resolved datasets was obtained i.e., SR32. The performance of the SR algorithm can be determined by evaluating two quality metrics i.e., peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) (Horé & Ziou, Aug 2010; Wang et al., 2019). Fundamentally, a high PSNR value (in decibels, dB) and SSIM value closer to 1 indicates better quality reconstructed images. The PSNR and SSIM values are obtained by comparing HR images with the reconstructed images. The average PSNR and SSIM values obtained for resultant image sequences at various resolutions are presented in Table I for various SR models. The performance of all reconstruction models based on these values are discussed in section 3.2. along with the reconstructed images presented in Figure 7. In addition, overall recognition performance of proposed framework is also discussed in the same section.

3.2 Results & Discussion

From the results obtained in Table 1, we can ascertain that the results for super resolved images, constructed using the proposed approach, is proportional to the input resolution used. Numerically, this is evident since PSNR and SSIM values at SR64 is higher by 6.29 and 0.1324 than SR32 using ESRGAN method. Likewise, both these image metrics are higher for SR64 than SR32 for all the models. The best reconstruction performance for micro expression is given by psnr-small model with 37.41dB PSNR and 0.9827 SSIM image metric considering both the scale factor. Using x2, both the PSNR driven approach have better PSNR values for super resolved micro expression images than ESRGAN method. For the same scale factor, the lowest reconstruction performance was given by noise-cancel model, but its image quality metrics are still higher than those images produced at x4 with ESRGAN and its variant. However, when comparing the SSIM metric at x2, the performance of psnr-small and ESRGAN is almost similar, with ESRGAN behind by only 0.0001. For x4, reconstruction using ESRGAN method is better than that of nESRGAN+ with PSNR higher by 6.42dB and SSIM higher by 0.0901. The reconstructed images obtained using both these methods for x4 are not as good as those obtained at x2 with other models. In overall we can say that images reconstructed using LR64 at x2 is better than those obtained using LR32 at x4. Thus, this strengthens the fact that quality of input images does affect the super resolution reconstruction process for micro expression as well. Visually the performance of these methods can be observed through the reconstructed images presented in Figure7.

The recognition performance obtained for all five super resolution methods employed with both feature extraction methods is recorded in Table 2. In line with the performance observations made for reconstruction through PSNR and SSIM results, the ME recognition performance also reflects the same notion by producing comparatively higher accuracy values at SR64 in comparison to SR32. This observation holds true for both the feature extraction techniques. Best recognition performance on SMIC-HS dataset recorded was 52.43%. This accuracy was obtained for both psnr-small and ESRGAN reconstructed images, with features extracted using LPQ-TOP. For LBP-TOP method at x2 the highest recognition accuracy observed was 51.45% using psnr-small super resolution model. Similarly, for the same feature extraction method at x4 the highest accuracy observed was 49.82%, obtained using ESRGAN super resolution model. For the same scale factor recognition accuracy of 50.60% was obtained

for ESRGAN reconstructed images, extracted using LPQ-TOP. Thus, at x4 the combination of ESRGAN and LPQ-TOP seemed to work well whereas, at x2 both ESRGAN and psnr-small with LPQ-TOP produced higher results. The framework seemed to produce higher accuracy when features were extracted using LPQ-TOP than LBP-TOP. It must be noted that recognition accuracy obtained using ESRGAN and psnr-small model are almost similar at x2. This could be due to the quality of reconstructed images obtained which is almost similar for these two methods as reflected by SSIM values. Similarly, the image quality metrics obtained at x2 was lowest for noise-cancel model hence same observation is reflected during recognition performance as well. Ideally nESRGAN+ is an improvement over ESRGAN method and expected to produce better reconstructed images at perceptual level, however in our implementation ESRGAN performed much better. At this stage it is difficult to say what influenced this, however we assume the noise injection could have further diminished the image quality affecting the overall recognition performance. More in-depth investigation is required to make this inference which will be an interesting development for the proposed framework. With the recognition accuracy achieved through this work, we can say that GAN based SR methods have helped in recovering micro expression image details to some extent. Though more robust training maybe required to achieve a decent improvement for micro expression images especially for noise-cancel and nESRGAN+ model. It must be mentioned here that at present only one work exist that have addressed the LR cases for micro expression (Li et al., 2019), but have not considered deep learning-based method however, we have considered their work for comparison.

4 CONCLUSIONS

Through this work we present a comprehensive investigation of a GAN based approach for recognizing micro expressions in low resolution images. With a target to achieve a quality upscaled image, we introduce deep learning-based reconstruction into the MER framework. The method was successful in recreating micro facial image details and implicitly helped to boost the overall recognition accuracy at all instances. The aim of this extensive study was to eventually provide a novel deep learning-based pipeline for solving low resolution MER problem and also to provide an analysis of GAN based methods specifically for low resolution problem for ME. The results achieved clearly imply that GAN and its variants can be exploited further for optimizing this specific problem targeting micro expression. Though our work is currently at an elementary stage, yet the results achieved are promising. Experimenting with low resolution micro expressions was a challenge due to the unavailability of such a dataset. It must be mentioned here that the current work does not address the dataset imbalance that exists for most of the available ME datasets, overcoming this limitation can certainly be another target for future work to achieve more robust performance.

REFERENCES

Chang, C., & Lin, C. (2011). LIBSVM. ACM Transactions on Intelligent Systems and Technology, 2(3), 1-27. doi:10.1145/1961189.1961199.

Ekman, P., & Friesen, W.V. (1969) Nonverbal Leakage and Clues to Deception, Psychiatry, 32:1, 88-106, DOI: 10.1080/00332747.1969.11023575.

Ekman P, “Telling Lies: Clues to deceit in the marketplace, politics, and marriage”, (Revised Edition), WW Norton & Company, 2009.

Francesco,C. & Dat, T .,(2018), “Image Super Resolution”, 2018,https://github.com/idealo/image-super-resolution.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. Communications of the ACM, 63(11), 1406-2661.

Heikkila. J., & Ojansivu V.,(2008) “Blur insensitive texture classification using local phase quantization”, International conference on image and signal processing. Springer, pp. 236–243, 2008. doi:10.1007/978-3-540-69905-7_27.

Heikkila, J., & Ojansivu, V. (Aug 2009). Methods for local phase quantization in blur-insensitive image analysis. Paper presented at the pp. 104-111. doi:10.1109/LNLA.2009.5278397.

Horé, A., & Ziou, D. (Aug 2010). Image Quality Metrics: PSNR vs. SSIM. Paper presented at the pp. 2366-2369. doi:10.1109/ICPR.2010.579.

Hung, K., Wang, K., & Jiang, J. (2019). Image interpolation using convolutional neural networks with deep recursive residual learning. Multimedia Tools and Applications, 78(16), 22813-22831. doi:10.1007/s11042-019-7633-1.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., et al. (Jul 2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. CVPR, pp. 105-114. doi:10.1109/CVPR.2017.19.

Li, G., Shi, J., Peng, J., & Zhao, G. (2019). Micro-expression Recognition Under Low-resolution Cases, 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and

Applications - Volume 5: VISAPP, ISBN 978-989-758-354-4,pages427-434,2019.doi:10.5220/0007373 604270434.

Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., et al. (2018). Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods. IEEE Transactions on Affective Computing,9(4),563-577. doi: 10.1109/TAFFC.2017 .2667642

Liong, S., See, J., Wong, K., & Phan, R. C. -. (2018). Less is more: Micro-expression recognition from video using apex frame. Signal Processing. Image Communication,62,82-92. doi:10.1016/j.image.2017 .11.006.

Oh, Y., See, J., Le Ngo, A. C., Phan, N. W., & Baskaran, V. M. (2018). A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges Frontiers Media SA.doi:10.3389/fpsyg. 2018.01128.

Päivärinta, J., Rahtu, E., & Heikkilä, J. (2011). Volume Local Phase Quantization for Blur-Insensitive Dynamic Texture Classification. Image Analysis (pp. 360-369). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-21227-7_34.

Rakotonirina, N. C., & Rasoanaivo, A. (May 2020). ESRGAN+ : Further Improving Enhanced Super-Resolution Generative Adversarial Network. Paper presented at the pp. 3637-3641. doi:10.1109/ICASSP40776.2020.9054071.

Sharma, P., Coleman, S., & Yogarajah, P.(2019). Micro Expression Classification Accuracy Assessment Irish Machine Vision & Image Processing, Dublin, Ireland,August28-30, 2019.doi:10.21427/kbny-0a41.

Sharma, P., Coleman, S., Yogarajah, P., Taggart, L., & Samarasinghe, P. (Jan 10, 2021). Magnifying Spontaneous Facial Micro Expressions for Improved Recognition. Paper presented at the pp. 7930-7936. doi:10.1109/ICPR48806.2021.9412585.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., et al. (2019). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. Computer Vision – ECCV 2018 Workshops (pp. 63-79). Cham: Springer International Publishing. doi:10.1007/978-3-030-11021-5_5.

Wang Y, See J, Phan RC-W, Oh Y-H (2015) Efficient Spatio-Temporal Local Binary Patterns for Spontaneous Facial Micro-Expression Recognition. PLoSONE10(5):e0124674. https://doi.org/10.1371/ journal.pone.0124674

Xiaobai Li, Pfister, T., Xiaohua Huang, Guoying Zhao, & Pietikainen, M. (Apr 2013). A Spontaneous Micro-expression Database: Inducement, collection and baseline. Paper presented at the pp. 1-6. doi:10.1109/FG.2013.6553717.

Yan, W., Li, X., Wang, S., Zhao, G., Liu, Y., Chen, Y., et al. (2014). CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. PloS One, 9(1), e86041. doi:10.1371/journal.pone.0086041.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (Jun 2018). Residual Dense Network for Image Super-Resolution.CVPRpp.2472-2481. doi:10.1109/CVPR. 2018.00262.

Zhao, G., & Li, X. (2019). Automatic Micro-Expression Analysis: Open Challenges Frontiers Media SA. doi:10.3389/fpsyg.2019.01833.

Zong, Y., Zhang, T., Zheng, W., Hong, X., Tang, C., Cui, Z., et al. (2018). Cross-Database Micro-Expression Recognition: A Benchmark. Retrieved from https://arxiv.org/abs/1812.07742.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS