Multi-modal with Multiple Image Filters for Facial Emotion Recognition *

Thong T. Huynh[+1,2], My M. Nguyen[+1,2], Phong T. Pham[1,2], Nam T. Nguyen[1,2], Tien L. Bui[1,2], Tuong Nguyen Huynh[3], Duc Dung Nguyen[1,2], and Hung T. Vo[*1,2]

[1] Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam
{nddung, vthung}@hcmut.edu.vn
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
[3] Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
huynhtuongnguyen@iuh.edu.vn

Abstract. The need to understand people, especially their behaviors and feelings, is growing in significance in today's quickly moving world. Despite the remarkable progress of science and technology in general and artificial intelligence in particular, facial emotion recognition remains a challenging task. In this research, a unique method for enhancing the accuracy of emotion recognition models is proposed. Through image analysis, the hair area and other facial areas have similar pixels but different intensities. However, to recognize emotions on the face, people only need to focus on facial features. Therefore, areas with the same pixels are not very helpful in accurately recognizing emotions. To solve the above problem, we conducted to eliminate or blur pixels that are the same as on the facial image. In addition, we also demonstrate that the use of the multi-model approach can support the learning process by allowing the sub-models to collaborate and increase accuracy. The experiments showed that this approach offers a valuable contribution to the field of facial emotion recognition and has a significant improvement compared to previous approaches.

Keywords: facial emotion recognition · Human–computer interaction · Convolutional network.

1 Introduction

Facial Emotion Recognition (FER) is a large field in Human-Computer Interaction (HCI) based on two subjects that are emotional psychology and artificial intelligence. Human emotion may be expressed by speech or non-verbal, such as transformations on the face or tone of voice, which are detected by sensors.

* Supported by Ho Chi Minh City University of Technology (HCMUT), VNU-HCM.
+Thong T. Huynh and My M. Nguyen contributed equally.
*Corresponding author: vthung@hcmut.edu.vn

In 1967, Albert Mehrabian - an American psychologist known for his research on the influence of body language and tone of voice in conveying messages [11]. He pointed out that when communicating with others, nonverbal factors such as gestures, facial expressions, and tone of voice play an important role besides language factors. Especially, he showed that 55 percent of emotions were expressed by face, 38 percent by voice, and the rest by speech.

Facial expression recognition has a wide range of applications across various fields, including education, where it can be used to understand learners' responses and engagement with the content of teaching sessions. In examination settings, it can be used to track and predict cheating behaviors by candidates. Besides, marketers can benefit from understanding how buyers react to their product advertisements. Facial emotion recognition can also be applied in the field of security where it can assist in detecting suspicious behavior and prevent potential hazards. The medical field can use FER to automate the care process as well as analyze the mental health of the patients. Finally, in the recruitment process, evaluating the quality of candidates can be more easily achieved with the assistance of this technology.

The importance of facial emotion recognition has attracted significant attention from numerous researchers. Many approaches have been proposed to address this problem, ranging from traditional machine-learning techniques to more advanced deep-learning models.

Shan *et al.* [15] introduced an approach called Boosted-LBP for extracting the most discriminatory features and achieving the highest level of recognition accuracy by employing classic support vector machines (SVM) [3] with the enhanced-LBP features. They carried out some experiments on Cohn-Kanade [9], MMI [12], and JAFFE [10] datasets. They have shown that SVMs based on LBP perform slightly better than SVMs based on Gabor wavelets by using 10-fold cross-validation on each dataset. S L *et al.* [5] proposed a technique using the HAAR classifier to detect faces, extract Local Binary Pattern (LBP) [14] histogram of various block sizes as feature vectors from facial images and use Principal Component Analysis (PCA) [1] to classify categories of facial expression. As different individuals display varied expressions with varying intensity, grayscale frontal face images of individuals were utilized to categorize six fundamental emotions - happiness, sadness, disgust, fear, surprise, and anger. Zhang *et al.* [18] introduced a new facial expression recognition approach using Local Binary Pattern (LBP) [14] and Local Phase Quantization (LPQ) [7] based on the Gabor face image. First, the Gabor filter is adopted to extract features of the face image among five scales and eight orientations to capture the significant visual properties. Then the Garbor image is encoded by the LBP and LPQ, respectively. Considering the size of the combined feature was too large, two algorithms Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [17] are used to reduce its dimension. Finally, the multi-class SVM classifiers based on the JAFFE database were used in the experiment.

In recent years, transformer architecture [16] has emerged as a powerful method. Many researchers have leveraged its ability for self-attention to create

models that perform better in various tasks, including computer vision. Aayushi Chaudhari *et al.* [2] used the ResNet-18 model [6] and transformers to classify facial expression recognition. The experiment underwent associated procedures, including face detection, cropping, and feature extraction using a deep learning model combined with fine-tuned transformer. The purpose of this study was to examine the performance of the Vision Transformer and compare it to their cutting-edge models on hybrid datasets. Roberto Pecoraro *et al.* [13] proposed a self-attention module that can be easily integrated into virtually every convolutional neural network named Local multi-Head Channel (LHC). There are two principal concepts on which the method is based. First, using the self-attention pattern in computer vision, applying the channel-wise application is considered more effective than the traditional approach of using spatial attention. Second, in facial expression recognition, where images have a consistent structure, local attention is referred to as a potentially better method than the global approach in overcoming the limitations of convolution. Compared to the previous state-of-the-art in the FER2013 dataset, LHC-Net is evaluated with significantly less complexity and effect on the underlying architecture regarding computational expense.

In this paper, we propose a new approach to improve emotion recognition model accuracy by eliminating or blurring the same pixels on facial areas. We also illustrate employing a combination of models to enhance accuracy. Our method contributes to facial emotion recognition and outperforms previous approaches.

The rest of this paper is as follows. Section 2 is our approach. Experimental results are given in the section 3. And finally, section 4 is the conclusion and the extended direction.

2 Multiple filter levels for FER

2.1 Multiple model

In machine learning, ensemble learning is a popular approach that combines multiple models in order to improve the predictive accuracy and robustness of a single model. By aggregating the outputs of several models, ensemble learning can reduce overfitting, and capture a wider range of patterns in the data. The architecture of multiple models in this article is built according to the architecture in figure 1.

In the first step, the model receives an image from the data set. The image is processed by the Processor, and the model makes predictions using a set of images that includes an original image and one or more processed images. In this study, there are two approaches were proposed, such as Dropping pixels - Processor, and Blurring images - Processor. Both approaches employ the original image to create special images, and all of them are used for the model. Each image is input to a separate CNN to extract unique features for each image type. The CNNs are the same for each sub-module and are based on the VGG16, which is a well-known and popular architecture. Finally, all features are concatenated and passed through a fully connected block to make predictions.

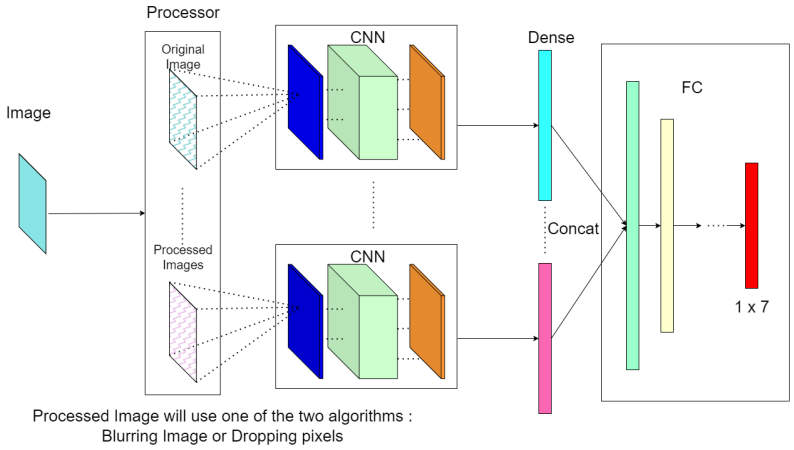


Fig. 1: Multiple model architecture

As previously mentioned, there are two approaches based on the way to process the original input image in the processor block in figure 1. Multi-model following the direction of dropping pixels is named Dropping Pixels Multiple (DPM-FER). And the model is in another direction of blurring image, which is named Blurring Multiple (BM-FER).

2.2 Dropping pixels - Processor

The idea of the algorithm used is to remove pixels that are similar and lie in a certain area (a matrix of size k). Conditions considered are similar: the disparity between the largest and the smallest pixel, and the disparity between the largest pixel and the pixel are under consideration. If the disparity between max-min is too small, we consider that image area to have unchanged pixels (for example, hair area) and give no value to the model.

Dropping pixels algorithm is implemented using a matrix that iterates all the pixels of the image, deciding whether to keep or remove that pixel from the image to reduce unnecessary details in the image. Using two loops to get the submatrix and each sub-matrix will be considered with the given condition.

Figure 2 shows how the *Dropping pixels* algorithm works. A big 4×4 matrix in figure 2 was iterated by a 2×2 matrix, obtaining four submatrices. To determine whether a submatrix is saved or not, take into account the conditions on each submatrix.

In this algorithm, just considering the difference between the 2 largest and smallest pixels is not enough. The $v2$ condition, the difference between the largest and pixel considered, is added to enhance the ability to adjust the granularity reduction between pixels. With the input image as the figure 3a, $v1$, $v2$ is the

Data: $img(48 \times 48), k, v1, v2$
Result: $newImg(48 \times 48)$
$newImg \leftarrow img.copy$
// Iterating the entire images with a matrix of size k, step = k
for $i \leftarrow 0$ **to** $img.shape[0] - k$ **by** k **do**
  for $j \leftarrow 0$ **to** $img.shape[1] - k$ **by** k **do**
    // Considering each matrix of size k in the image
    $M \leftarrow img[i : i + k, j : j + k]$
    // Condition 1: This matrix means that the difference between 2 largest and smallest pixels must be large enough
    $c1 \leftarrow Max(M) - Min(M) >= v1$
    // Condition 2: Each pixel is meaningful when the difference between it and the largest pixel is small enough
    $c2 \leftarrow Max(M) - M <= v2$
    // Save the result of the matrix under consideration
    $newImg[i : i + k, j : j + k] \leftarrow c1 * c2 * M$
  end
end

Algorithm 1: Dropping pixels algorithm

condition that determines whether the pixels are saved or not. Changing the values of $v1$, $v2$ from 0-255, *step* by 15 gives the result as Figure 3b.

Figure 3a is the original input image, also known as the sample image. The algorithm will use this image and create image 3b by adjusting the parameters of the conditions in the algorithm.

2.3 Blurring image - Processor

This algorithm along with the purpose of reducing the detail of the image is deployed by calculating the difference between two adjacent pixels. With such a calculation, a pixel can be counted in terms of another pixel either horizontally, vertically, or diagonally. And to be objective, the result when calculated on a pixel will be the average of the three calculations above for each pixel. The Blurring image algorithm is given in the algorithm 2.

First of all, a row and a column will be added to the right and the bottom of the pixel matrix according to the symmetry mechanism. Then, after iterating the pixels and recalculating each one's value, each pixel will subtract the three adjacent ones—the bottom, right, and bottom right corner. The new value will be the average of the three results that just has been calculated.

3 Experimental and Results

3.1 Dataset

The ICML 2013 Workshop on Challenges in Representation Learning the Facial Expression Recognition 2013 (FER2013) dataset as a main dataset for the facial

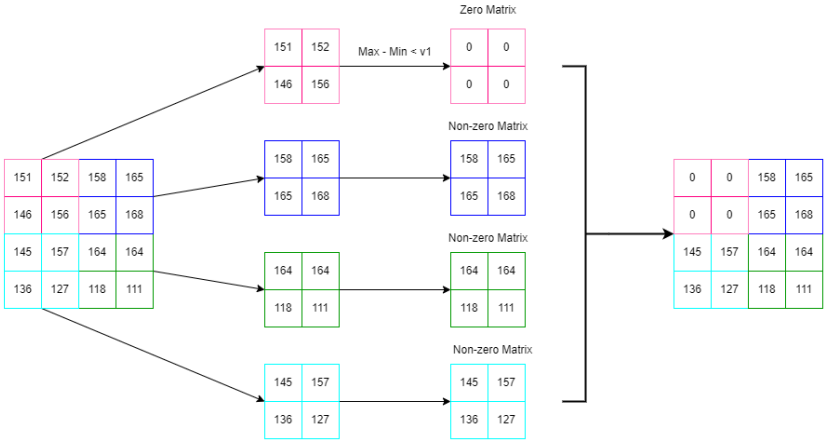
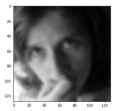


Fig. 2: Dropping pixels Sample



(a) Original Image

(b) Images after using the algorithm

Fig. 3: Illustration of image processing

expression recognition contest [4]. FER2013 was created by Aaron Courville Pierre and Luc Carrier using the Google image search API to search for images of faces with many keywords related to different emotions. These keywords were also combined with words related to ethnicity, age, gender, etc. The authors also used OpenCV to place detect the human face of each image, reject the wrong

Data: $img(48 \times 48)$
Result: $newImg(48 \times 48)$
$newImg \leftarrow img.padding(1, 1)$
`// Iterating the entire images`
for $i \leftarrow 0$ **to** $img.shape[0] - 1$ **by** *1* **do**
  for $j \leftarrow 0$ **to** $img.shape[1] - 1$ **by** *1* **do**
    `// horizontal`
    $h \leftarrow img[i][j] - img[i][j + 1]$
    `// vertical`
    $v \leftarrow img[i][j] - img[i + 1][j]$
    `// diagonal`
    $d \leftarrow img[i][j] - img[i + 1][j + 1]$
    `// Save the result of the matrix under consideration`
    $newImg[i][j] \leftarrow mean(h + v + d)$
  end
end

Algorithm 2: Blurring algorithm

label image then resize and convert it to grayscale. FER2013 is just a small subset of the work for the contest. FER2013 consists of 35,887 48×48-pixel grayscale images, most of which are human faces.

Figure 4 gives some sample images for each emotion category in the FER2013. Each column contains three different images for a category. The inconsistency of face position, face angle, and incomplete faces (e.g some hand parts cover some of the face areas in fear or surprise images) is one of the main challenges in classifying emotions based on the features of each facial region.

Fig. 4: Some sample from FER2013 dataset.

Table 1: Information about FER2013

Category	Training	Public Test	Private Test
Angry	3,995	467	491
Disgust	436	56	55
Fear	4,097	496	528
Happy	7,215	895	879
Sad	4,830	653	594
Surprise	3,171	415	416
Neutral	6,965	607	626
Total	28,709	3,589	3,589

All images in FER2013 were labeled into one of seven main categories: angry, disgust, fear, happy, sad, surprise, and neutral. Table 1 briefly describes some general information about the training set, public test set, and private test set for each of the emotion classes of FER2013. The dataset contains 4,953 Angry images, 547 Disgust images, 5,121 Fear images, 8,989 Happy images, 6,077 Sad images, and 6,198 Neutral images. The training set consists of 28,709 images, the public test set consists of 3,589 images and the private test set consists of 3,589 images. The human accuracy on this dataset was about $65 \pm 5\%$ [4].

3.2 Experimental setup

Spatial transformer networks The spatial transformer network (STN) was introduced by Jaderberg *et al.*[8]. It is used as a small CNN module inserted directly into our models, following the input immediately. STN is a popular approach to increase the spatial invariance of the model. The main idea is pretty simple: its convolution layers capture features of the image and learn how to perform an affine transformation to the original image directly from the back-propagation of the main model. The output image is the original image with canonical orientation, leading to better classification performance. It's a powerful technique to overcome the invariability of CNN.

Dropping Pixels Multiple model - DPM The model is trained on the Fastai[4] platform, using data with multiple dimensions and *learning_ rate*. According to each training with *fit_ one_ cycle*, the learning rate is determined based on the available *lr_ find* function in Fastai.

The trained model uses an original image with a size of 48×48 and the image which is processed with algorithm 1 has a size of 96×96, the *batch_ size* used is 128, *learning_ rate* change through each training phase. Firstly, each branch of the model is trained with its corresponding images. The model's CNNs will load those trained weights and fine-tune them so that the model can converge faster.

[4] Ref: https://docs.fast.ai/learner.html

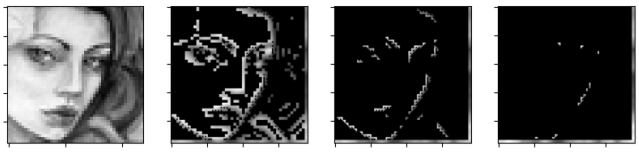


Fig. 5: Example of input data for the DPM

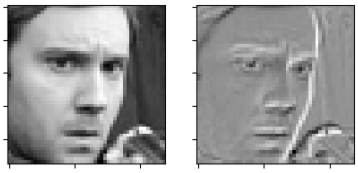


Fig. 6: Example of input data for the BM

The example of input data of the DPM is shown in figure 5. Image data samples for each submodel (Figure 5), the images are processed using algorithm 1 with the parameters described in the table 2. The selected parameters are the parameter sets for each level used to test the model.

Blurring Multiple model - BM The BM-FER is also configured similarly to the DPM. Because the two models follow two different approaches, their input data will differ slightly. The input data will consist of two images: the original image and the image which is transformed by algorithm 2. As a result, both images in this model are the same size: 48×48. The example of the input data of the BM is given in figure 6. On the left of figure 6 is the original image and on the right is the image processed with the algorithm 2.

3.3 Results

Table 3 shows the comparison of the approach in the article compared to the conventional VGG16. Overall, both our approaches outperform the baseline network (VGG16). BM-FER has significant improvement over the base network, with an accuracy of 70.08% on the private test and 70.87% on the public test

Table 2: Table of input image parameters for the DPM

img	**k**	**v1**	**v2**
Image 1	0	0	255
Image 2	2	15	30
Image 3	2	30	30
Image 4	2	45	30

Table 3: Baseline comparison table of results

Model	**Public Test(%)**	**Private Test(%)**	**TTA PV(%)**	**TTA PB(%)**
VGG16	65.27	66.31	64.44	66.01
DPM-FER (2 CNN) + 2STN	66.45	67.87	66.01	67.40
BM-FER + STN	68.23	69.30	70.08	70.87

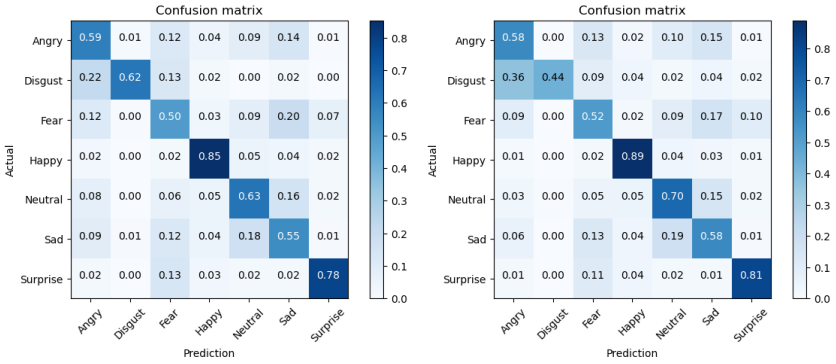
with TTA. The baseline VGG16 only got about 65.27% on the public test set and 66.31% on the private test set. The margin of improvement is more than 7% in related. Compared to the baseline, the DPM-FER model produced results that were a little better than VGG16: 66.45% on the public test set and 67.87% on the private test set. The results also suggest that TTA is one of the useful methods for models, as both two of our approaches could get better performance in prediction with it enabled.

Some different configures of DPM-FER and BM-FER also conduct and results are given in the table 4. For DPM-FER, for the limitation of computing resources, there are only two configures, two filters, and four filters. We also test whether STN blocks are in shared mode. The results show that there is little improvement when using four filters over two filters, 66.05% compared with 65% in two filters. However, four filters of image needed four different of CNNs base network, then the network size is about double, heavily. It looks like the STN block shared enabled does not have a big effect. For BM-FER architecture, we also tested with the larger image size of 64×64 beside STN enabled. Compare with the base of BM-FER, this configuration got a small performance improvement.

Figure 7 is a detailed comparison between two models VGG16 and BM-FER based on their confusion matrix on the Private Test set. For some classes, BM-FER has a large improvement compared to VGG16, such as happy, and neutral. It also decreases the confusion in some pairs of emotions such as fear-angry, fear-sad, and surprise-fear. However, there are some classes that VGG16 is better recognized, such as disgust.

Table 4: Table comparing results of different configurations

Model	**Public Test(%)**	**Private Test(%)**	**TTA PV(%)**	**TTA PB(%)**
DPM-FER (4 CNN)	66.06	66.82	—	—
DPM-FER (2 CNN)	65	—	—	—
DPM-FER (2 CNN) + STN	66.02	68.18	65.85	67.09
DPM-FER (2 CNN) + 2STN	66.45	67.87	66.01	67.40
BM-FER	68.00	68.65	69.12	70.30
BM-FER + STN	68.14	69.07	69.37	70.63
BM-FER + STN + (64×64)	68.23	69.30	70.08	70.87



(a) VGG16 (b) BM-FER

Fig. 7: Confusion matrix of VGG16 and BM-FER on the Private Test

4 Conclusion

In conclusion, facial emotion recognition is one of the significant attention fields in researching things related to human behaviors and emotions. The development of artificial intelligence has experienced a significant leap in recent years, especially with the emergence of cutting-edge deep-learning models. As a result, many researchers and scientists have proposed different methods which aim to optimize the output result of this topic. However, it is still a challenging task that has numerous obstacles in various aspects.

After realizing that certain areas with comparable values may not have meaningful contributions to improve the accurate predictions of the model. To solve the above problem, we decided to focus on regions with greater importance weight and omit those with similar. Simultaneously, we also leveraged the mutual support of sub-models during the self-learning process by combining them together. Through the integration of these concepts, we have introduced two model types, namely DBM-FER and BM-FER, in this study. The experimental outcomes demonstrate a notable enhancement, achieving the best results of 70.78% and 67.40% respectively for BM-FER and DPM-FER on the testing dataset, compared to 66.01% attained by a well-known CNN network, VGG16.

Acknowledgements We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Bro, R., Smilde, A.K.: Principal component analysis. Analytical methods **6**(9), 2812–2831 (2014)

2. Chaudhari, A., Bhatt, C., Krishna, A., Mazzeo, P.L.: Vitfer: Facial emotion recognition with vision transformers. Applied System Innovation **5**(4), 80 (2022)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)
4. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. pp. 117–124. Springer (2013)
5. Happy, S.L., George, A., Routray, A.: A real time facial expression classification system using local binary patterns. In: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI). pp. 1–5 (2012). https://doi.org/10.1109/IHCI.2012.6481802
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Heikkila, J., Ojansivu, V.: Methods for local phase quantization in blur-insensitive image analysis. In: 2009 International Workshop on Local and Non-Local Approximation in Image Processing. pp. 104–111 (2009). https://doi.org/10.1109/LNLA.2009.5278397
8. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
9. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580). pp. 46–53. IEEE (2000)
10. Lyons, M.J.: "Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset (Jul 2021), https://doi.org/10.5281/zenodo.5147170
11. Mehrabian, A.: Silent messages: A wealth of information about nonverbal communication (includes an updated bibliography). Wadsworth, Belmont, CA (1981)
12. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo. pp. 5 pp.– (2005). https://doi.org/10.1109/ICME.2005.1521424
13. Pecoraro, R., Basile, V., Bono, V.: Local multi-head channel self-attention for facial expression recognition. Information **13**(9), 419 (2022)
14. Pietikäinen, M.: Local binary patterns (2010), http://www.scholarpedia.org/article/Local_Binary_Patterns, accessed March 3, 2023
15. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing **27**(6), 803–816 (2009). https://doi.org/10.1016/j.imavis.2008.08.005
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B., Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B.: Linear discriminant analysis. Robust data mining pp. 27–33 (2013)
18. Zhang, B., Liu, G., Xie, G.: Facial expression recognition using lbp and lpq based on gabor wavelet transform. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). pp. 365–369 (2016). https://doi.org/10.1109/CompComm.2016.7924724