Automatic facial expression recognition under partial occlusion based on motion reconstruction using a denoising autoencoder

Abdelaali Kemmou[1], Adil El Makrani[1], Ikram El Azami[1], Moulay Hafid Aabidi[2]
[1]Laboratory of Research in Informatics, Faculty of Science, Ibn Tofail University, Kenitra, Morocco
[2]Department of Computer and Mathematical Engineering, Higher School of Technology, University Sultan Moulay Slimane (USMS), Khenifra, Morocco

Article Info

Article history:

Received Aug 27, 2023
Revised Dec 30, 2023
Accepted Jan 3, 2024

Keywords:

Autoencoder
Facial expressions
Facial occlusion
Motion
Recognition

ABSTRACT

Automatic facial expression recognition (FER) plays a valuable role in various fields, including health, road safety, and marketing, where providing feedback on the user's condition is crucial. While significant progress has been made in controlled environments (such as frontal, unconcluded, and well-lit conditions), recognizing facial expressions in unconstrained environments (natural settings) remains challenging. The presence of occlusions poses a particular difficulty as they obscure parts of the facial information captured in the image. To address this issue, researchers have proposed different solutions, broadly categorized into two approaches: those focusing on visible regions of the face and those attempting to reconstruct hidden parts. Currently, most solutions rely on texture or geometry-based methods, with only a few utilizing motion-based approaches. However, incorporating motion appears to be particularly promising in adapting to occlusions due to its unique characteristics, such as close-range propagation and local coherence. In this paper, our focus lies on leveraging motion to overcome the challenges posed by occlusions in FER tasks.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Abdelaali Kemmou
Laboratory of Research in Informatics, Faculty of Science, Ibn Tofail University
Kenitra, Morocco
Email: kemabd@gmail.ma

1. INTRODUCTION

In natural environments, various factors such as accessories or the person's actions can frequently lead to the concealment of certain facial features. For instance, scarves, surgical masks, or even a person's hand can obscure the lower part of the face. Similarly, sunglasses, hair, or hats can obscure the upper part of the face as shown in Figure 1. Furthermore, these occlusions can be further complicated by unfavorable lighting conditions, low image resolution, or intentional alterations like kinetic blur or anonymizing elements. All these factors contribute to the challenge of facial recognition, as they result in the loss of critical information and the introduction of noise caused by these occlusions.

In response to this challenge, researchers have put forward various solutions as documented in the literature. These solutions can be broadly categorized into two main groups: firstly, there are methods that concentrate solely on the visible regions of the face, disregarding any obscured or hidden areas. These approaches primarily exploit the information readily available on the face's exposed regions. On the other hand, the second category comprises methods that address the issue of occlusions by reconstructing the missing information. By adopting these techniques, it becomes possible to recover and integrate data encompassing all the details of a complete facial structure, even those concealed by occlusions.

Within the first category, the initial proposed solutions [1], [2] involved dividing the face into different regions and primarily focusing on the visible regions. More recently, these methods have advanced towards neural approaches, employing attention mechanisms that learn to focus on the visible parts of the face. These mechanisms assign weights based on the significance of each region in the recognition process [3], [4]. When occlusions occur, the weights naturally become higher for the unoccluded areas of the face.

In the second category, the early approaches [5], [6] were based on tracking facial minutiae. In cases where certain areas were not visible due to occlusions, statistical techniques were employed to deduce the undetected minutiae. With the emergence of appearance descriptors, the subsequent solutions shifted towards directly reconstructing the facial appearance using algorithms based on robust principal component analysis (RPCA) [7] or generative algorithms [8]. These generative algorithms have experienced significant advancements in recent years with the rise of deep learning architectures [9]. The persistent challenge in facial recognition is exacerbated by occlusions, hindering the accurate analysis of concealed facial features.

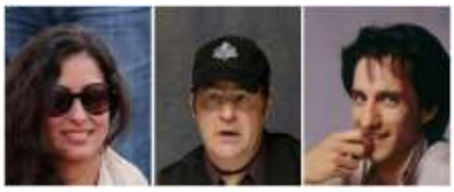


Figure 1. Different occlusions, extracted from CelebA dataset

2. BACKGROUND AND RELATED WORKS

2.1. Impact of occlusions

Stadies by Kemmou *et al.* [10] and Kotsia *et al.* [11] on the CK database offer the most comprehensive exploration of the impact of occlusions on facial expression recognition (FER). This study delves into the significance of facial regions in the identification of facial expressions. The research offers an overview of facial regions whose obstruction significantly influences expression recognition.

It becomes evident from various studies that determining the importance of facial regions and assessing the effects of occlusions pose challenges. These investigations reveal that the impact varies based on the descriptors employed. However, it is evident that the eyes and mouth play crucial roles, and occlusions in these areas notably hinder FER. Furthermore, the study highlights that occlusions yield distinct effects depending on the specific facial expressions being analyzed.

2.2. Exploitation of visible areas of the face

The first category of solutions addresses the issue of facial recognition by concentrating solely on the visible regions of the face, neglecting the information present in the obscured or hidden areas. These methods aim to analyze and extract features exclusively from the observable portions of the face, disregarding any potential insights that could be derived from the concealed regions. While this approach may be computationally efficient and straightforward, it inherently misses out on vital cues and details that could be valuable for achieving a more comprehensive and accurate understanding of facial characteristics and expressions.

Several solutions have been suggested that involve dividing facial regions based on a sparse representation of facial images. These approaches draw inspiration from sparse representation classifier (SRC) classifiers [12], which involve constructing dictionaries from facial images. Each dictionary comprises images belonging to the same class, specifically those representing identical facial expressions in our context. In the testing phase, input data are represented by computing a linear combination of the training data within the same dictionary. This computation aims to generate a linear combination closely resembling the initial data. The dictionary used for representing the test data inherently facilitates classification based on the dictionary label.

Cotter [13], [14] proposed advancing this classifier by segmenting the face into distinct regions and constructing dictionaries for each of these regions. The final classification is then accomplished by amalgamating the classification results from each region. When calculating the linear combination, the disparity between the original image and the linear combination, which comprises unoccluded images, is more pronounced in the presence of occlusion. To leverage this characteristic, Cotter [13], [14] suggest assigning weights to different facial regions during the fusion of classifiers. This observation was also employed by Huang *et al.* in 2012 [15] for occlusion detection. Nevertheless, solutions based on image dictionaries of this kind necessitate training data closely resembling test data to compute a meaningful linear combination.

Dapogny *et al.* [16] present an approach centered on computing region weights. In their study, the authors employed an autoencoder (AE) trained to assign confidence weights to different facial regions. These confidence scores automatically determine the relevance of each region and implicitly indicate whether a region is occluded.

More recently, there has been a trend towards automating these techniques through attention mechanisms designed to concentrate on the most pertinent facial regions [3], [4]. Figure 2 illustrates the attention maps generated by these mechanisms. The colored indications in the figure highlight the areas of focus, with the reddest regions signifying the highest levels of attention. The importance weights derived from these mechanisms enable the weighting of pixels in the image based on their significance. In the context of occlusions, these mechanisms are employed to emphasize the visible regions of the face.

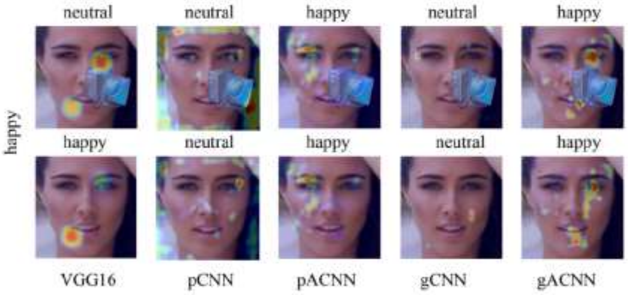


Figure 2. Attention maps from various methods on original (second row) and occluded images (first row) are shown, with image labels (“happy”) on the left and method classification results above each image [3]

2.3. Reconstruction of hidden areas of the face

A second set of solutions revolves around the restoration of concealed facial areas, with various methods proposed in this regard. These approaches may involve reconstructing feature points or texture. The former relies on tracking facial feature points, as demonstrated by Bourel *et al.* [5], [6], who utilized the Kanade-Lucas tracker to recover lost feature points. Subsequently, static solutions emerged, employing techniques like principal component analysis (PCA) proposed by Towner and Slater [17], and a combination of iterative closest point (ICP) and fuzzy-C-means (FCM) suggested by Zhang *et al.* [18]. This methodology heavily relies on feature point detection, which remains challenging, especially in the presence of occlusions.

The second set of methods involves the direct reconstruction of texture, bypassing the challenges associated with feature extraction in the presence of occlusions. These reconstruction techniques often leverage RPCA [19], proposed as an enhancement of PCA [20] for increased robustness against occlusions. In this approach, an RPCA-based reconstruction is computed, and the occluded portion of the original image is substituted with the reconstructed segment. However, this method tends to introduce artifacts into the original image, as depicted in Figure 3 from the paper by Cornejo and Pedrini [7]. Panel Figure 3(a) displays the original images without occlusion, Figure 3(b) presents the occulted images, Figure 3(c) illustrates the RPCA reconstruction of the occluded image, and Figure 3(d) reveals the original reconstructed image with the occluded area replaced by the reconstruction. These artifacts can distort facial expressions, thereby complicating the recognition process.

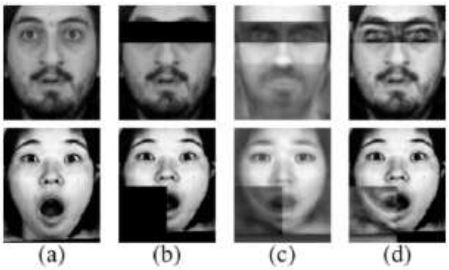


Figure 3. Reconstruction process via RPCA from Cornejo and Pedrini [7] article; (a) original images without occlusion, (b) occluded images, (c) RPCA reconstruction of the occluded image, and (d) original image reconstructed by substituting the occluded area with the reconstruction

In recent times, researchers have increasingly explored solutions that rely on neural network architectures for addressing the problem. Two prominent types of architectures being utilized are AEs and generative adversarial networks (GANs). The AE, illustrated in Figure 4, comprises an input layer, a hidden layer, and an output layer. This network is designed to achieve data compression at the output of the encoder and then reconstruct the original data from its compressed representation at the output of the decoder, typically with a similar number of neurons at the input and output layers. By employing AEs, the system aims to effectively capture the essential features of the facial data and subsequently reconstruct the complete face from its compressed representation, thereby mitigating the impact of occlusions on facial recognition.

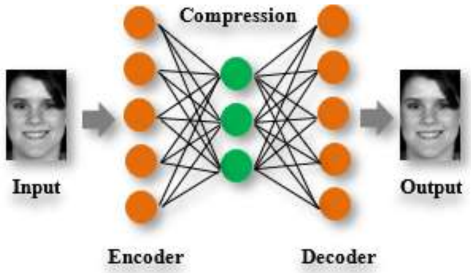


Figure 4. Illustration of a classical autoencoder architecture

The architecture of the AE comprises an input layer with neurons corresponding to the data dimension. Following this input layer is a hidden layer with a reduced number of neurons. The output of this layer serves as a compressed representation of the initial data, containing essential elements for its reconstruction. The decoder's output consists of the same number of neurons as the input layer, completing the data reconstruction process. The incorporation of nonlinearity enhances the AE's capabilities, rendering it more powerful [21]. Various AE variants have been proposed, including the denoising AE depicted in Figure 5, designed to reconstruct noise-free data from noisy input.

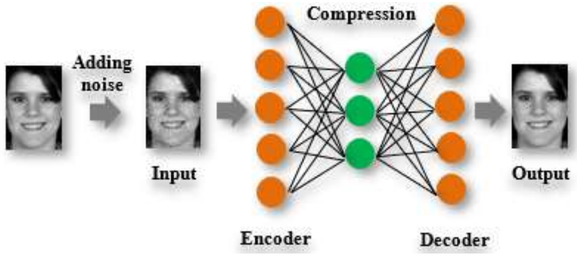


Figure 5. The denoising autoencoder involves introducing noise to the original data before training the autoencoder for data reconstruction. This design enhances the autoencoder's robustness to noise and minor variations

In the presence of occlusions, occlusion can be viewed as a form of noise. Consequently, the denoising AE is trained to reconstruct the image without occlusion. For instance, Gondara *et al.* [22] applied a denoising AE architecture to eliminate noise, including capture artifacts, from medical images. These AE architectures can be stacked to further enhance the reconstruction of occluded areas. This approach was employed by Zhang *et al.* [23], who suggested reconstructing occluded facial regions in the context of person recognition.

In recent advancements, generative adversarial networks (GANs) have emerged as a prominent class of reconstruction architectures and have been increasingly integrated into state-of-the-art research in various fields, including facial recognition [8], [24], [25]. Originally introduced by Goodfellow *et al.* in 2014 [26], GANs were primarily designed as generative algorithms to produce new data. However, researchers have also discovered their potential in handling noisy data reconstruction tasks. The architecture of a GAN involves two interconnected networks, namely the generator and discriminator, as depicted in Figure 6. The generator is trained to generate new data samples, while the discriminator's role is to evaluate the realism of the generated data. Interestingly, the output of the discriminator is directly utilized by the generator as a cost function, creating a competitive feedback loop that leads to the refinement of the generated data. This iterative process fosters the generation of increasingly realistic data, ultimately enhancing the quality of reconstructed information, even in the presence of occlusions.

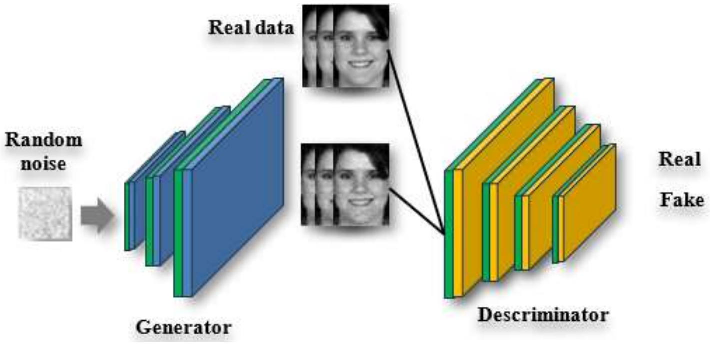


Figure 6. Illustration of a GAN architecture

GANs are frequently employed for tasks such as image editing and completion [24], [25], [27]–[29]. However, these innovative architectures are not yet widely utilized in the realm of FER for reconstructing occluded areas. In the context of FER, Ranzato *et al.* [30] pioneered the introduction of a deep architecture for the reconstruction of occluded faces. More recently, Lu *et al.* [8] presented a GAN architecture featuring two discriminators: a traditional discriminator distinguishing real data from generated data, and another assessing the model’s ability to recognize facial expressions in the generated images. This approach explicitly considers the proficiency in reconstructing facial expression information. In section 3, we present our unique solution, centered around motion reconstruction using an AE architecture. Finally, section 4 discusses the experimental protocol employed to evaluate the effectiveness of this solution.

3. RESEARCH METHOD

Our proposed solution aims to address the challenges posed by occluded faces during analysis by adopting a reconstruction-based approach. By compensating for the effects of occlusions, we endeavor to restore the faces to their ideal conditions for accurate analysis. To achieve this, we intend to leverage the motion similarity property illustrated in Figure 7. By reconstructing the information affected by occlusions directly in the motion domain, we can effectively mitigate the impact of occlusions and enhance the quality of facial data for subsequent analysis, thereby improving the facial recognition and expression interpretation processes.

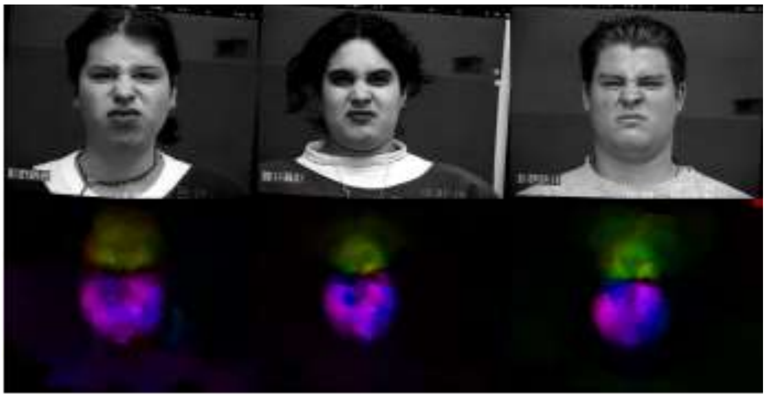


Figure 7. Individuals showing expressions of disgust from the CK+database and their optical flows calculated using the DeepFlow method, comparing neutral, and apex images

Our proposed approach involves leveraging the power of reconstruction techniques to restore the computed optic flow obtained from video sequences featuring occluded faces, as illustrated in Figure 8. By reconstructing the optic flow data, we aim to recover the hidden or distorted motion information caused by occlusions. This reconstructed optic flow can then be used to enhance the analysis and understanding of facial expressions, thereby improving the accuracy and robustness of facial recognition systems when faced with challenging conditions such as occlusions. Ultimately, this approach holds the potential to enable more comprehensive and reliable facial analysis in real-world scenarios.

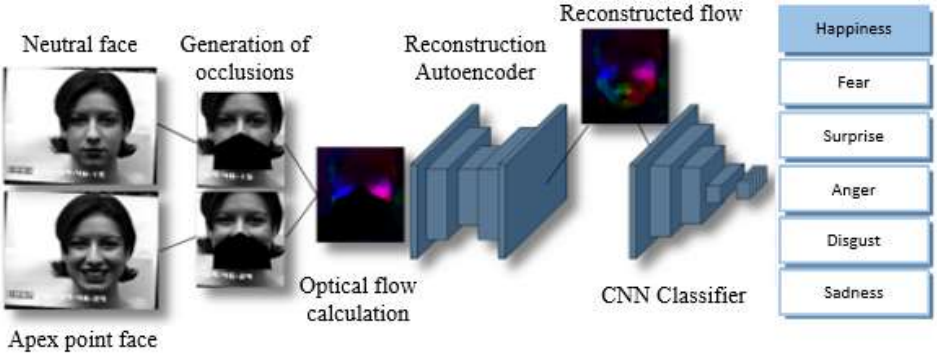


Figure 8. Complete process of the optical flow reconstruction method

3.1. Data preparation

To train a denoising AE effectively, two essential components are required: firstly, the ground truth optical flow values calculated from unoccluded data, which serve as the reference for accurate motion information. Secondly, the optical flow values computed from the same set of occluded images are needed. By having both datasets, the denoising autoencAEoder can learn to map the noisy or occluded optical flow data to its corresponding ground truth, effectively reducing the impact of occlusions and enhancing the quality of the reconstructed motion information. This process enables the denoising AE to develop a robust ability to handle occlusions and improve the overall accuracy of optical flow estimation for occluded facial sequences.

3.2. Generation of occlusions

To facilitate the selection of the most crucial occluded regions for FER, our approach involves introducing different types of occlusions specifically targeting the eyes and mouth, as illustrated in Figure 9. The eyes and mouth are commonly chosen as the regions of interest since they are frequently affected by occlusions in various evaluation scenarios. To simulate these occlusions, we apply static black boxes onto all the images within a video sequence, thus obscuring the respective regions. This method allows us to systematically evaluate and analyze the performance of proposed solutions for addressing the challenges posed by occlusions, particularly in critical facial areas, and provides valuable insights into the effectiveness of these methods in real-world applications.

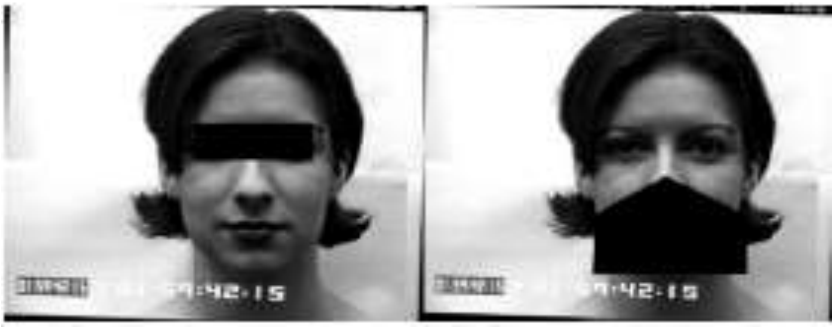


Figure 9. Selected occlusions to evaluate the approach, applied on the CK+database

3.3. Optical flow calculation

Given the databases available in the literature, we currently do not have large volumes of data to evaluate the approach. Therefore, it is necessary to limit the number of parameters to be learned by using shallow architectures. In addition, we also work on optical flows of reduced size at the input of the system. Figure 10 shows the used process for calculating and reducing size of the calculated optical flows. In order to preserve as much as possible, the quality of the flow computation, we compute the optical flow on high resolution images before applying a reduction process to obtain a standard size that allows to satisfy the normalized input criteria of the AE.

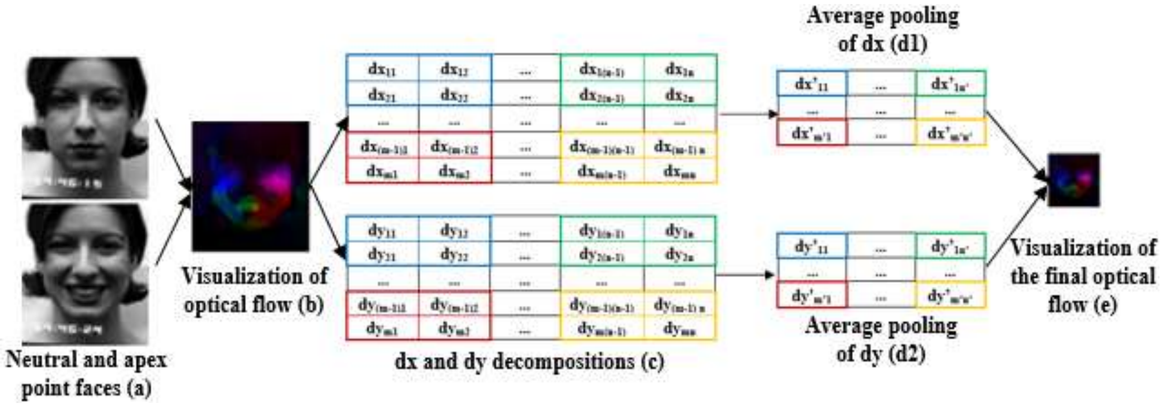

Average pooling of dx (d1)
dx_{11} | dx_{12} | ... | $dx_{1(n-1)}$ | dx_{1n}
dx_{21} | dx_{22} | ... | $dx_{2(n-1)}$ | dx_{2n}
... | ... | ... | ... | ...
$dx_{(m-1)1}$ | $dx_{(m-1)2}$ | ... | $dx_{(m-1)(n-1)}$ | $dx_{(m-1)n}$
dx_{m1} | dx_{m2} | ... | $dx_{m(n-1)}$ | dx_{mn}
dx'_{11} | ... | $dx'_{1n'}$
... | ... | ...
$dx'_{m'1}$ | ... | $dx'_{m'n'}$
dy_{11} | dy_{12} | ... | $dy_{1(n-1)}$ | dy_{1n}
dy_{21} | dy_{22} | ... | $dy_{2(n-1)}$ | dy_{2n}
... | ... | ... | ... | ...
$dy_{(m-1)1}$ | $dy_{(m-1)2}$ | ... | $dy_{(m-1)(n-1)}$ | $dy_{(m-1)n}$
dy_{m1} | dy_{m2} | ... | $dy_{m(n-1)}$ | dy_{mn}
dy'_{11} | ... | $dy'_{1n'}$
... | ... | ...
$dy'_{m'1}$ | ... | $dy'_{m'n'}$
Visualization of optical flow (b)
Visualization of the final optical flow (e)
Neutral and apex point faces (a)
dx and dy decompositions (c)
Average pooling of dy (d2)

Figure 10. Proposed process for calculating and reducing the size of the calculated optical flows for use in the method

3.3.1. Optical flow reconstruction

Our approach is based on a denoising AE architecture to reconstruct the optical flow computed on occluded data. The optical flows computed by the methodology described in the previous section are then directly used as input to the architecture. In this section, we first describe the auto-encoder architecture used for the reconstruction. Secondly, we discuss the different cost functions that we believe are suitable for training the AE.

3.3.2. Autoencoder architecture

The AE architecture employed in our approach draws inspiration from the Hourglass and U-Net architectures [31], [32], which both propose symmetric AEs equipped with skip connections. These skip connections facilitate the seamless transfer of information between encoder and decoder layers, aiding in better feature representation and information preservation during the reconstruction process. Additionally, the U-Net architecture has been particularly designed to excel with limited training data, making it well-suited for scenarios where the availability of facial data may be constrained. By combining the strengths of these architectures, our AE-based approach aims to achieve robust and efficient learning while effectively handling occluded facial data for improved FER.

The proposed architecture is composed of successive layers of convolutions with kernels of size 3×3 which is the minimum possible size of a convolution kernel allowing to characterize the spatial changes. Given the small size of the inputs, we limit ourselves to a size of 3×3 in order to characterize the local changes. These convolutions are followed by a layer of rectified linear unit (ReLU) and max-pooling for the encoder and successive layers of convolutions and Up-sampling for the decoder.

The optical flow contains information about the x and y displacements of each point, which can be positive or negative based on the direction of movement. To accommodate this possibility of negative values in the reconstructed optical flow, the last convolution layer is deliberately not followed by a ReLU activation function. By doing so, the AE can accurately reconstruct optical flow data, even with negative elements. As illustrated in Figure 11, the AE architecture is designed to take the computed optical flows between two occluded images as input and then generate the corresponding reconstruction as its output. This process enables the AE to effectively learn and capture the essential motion patterns, aiding in the accurate recovery of facial expressions despite occlusions.

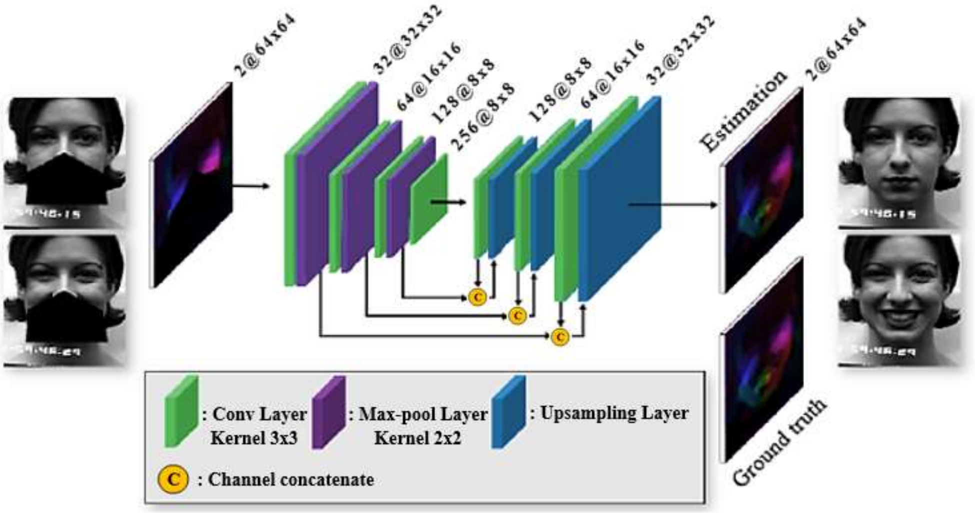


Figure 11. Illustration of the reconstruction autoencoder which takes in input optical flows noisy by the presence of occlusions on the original images and which calculates a reconstruction of these optical flows

3.3.3. Cost function

In our approach, the cost function plays a crucial role in assessing the quality of reconstruction achieved by the AE by comparing it with the ground truth. We have explored various cost functions to identify the most suitable one that effectively preserves the critical information associated with facial expressions during the reconstruction process. To formulate the cost function, we consider two sets of optical flows, namely U (representing the predictions) and V (representing the ground truth), each containing n optical flows. Each optical flow in these sets consists of x and y displacements, which represent the movements of points in the facial region. By evaluating the cost function on these optical flow sets, we can quantify the accuracy and fidelity of the AEs reconstruction, thus refining our approach to enhance FER under occlusions.

MSE: mean square error is a classical cost function. It calculates the square of the Euclidean distance between the reconstruction and the ground truth at each point.

$$\mathrm{MSE}\ (\mathrm{U},\mathrm{V}) = \frac{1}{2\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{u}_{\mathrm{ix}}-\mathrm{v}_{\mathrm{ix}})^2+(\mathrm{u}_{\mathrm{iy}}-\mathrm{v}_{\mathrm{iy}})^2 \quad (1)$$

Wing: initially proposed to locate feature points [33], the wing cost function further penalizes medium and small errors by using a log-based error for errors below a certain threshold.

$$\mathrm{loss}\ (\mathrm{U},\mathrm{V}) = \frac{1}{2\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\ \mathrm{wing}(\mathrm{u}_{\mathrm{i}}-\mathrm{v}_{\mathrm{i}}) \quad (2)$$

$$\mathrm{wing}\ (\mathrm{U},\mathrm{V}) = \begin{cases}\omega \ln\left(1+\frac{|\mathrm{u}-\mathrm{v}|}{\epsilon}\right), \mathrm{if}|\mathrm{u}-\mathrm{v}| < \epsilon \\ |\mathrm{u}-\mathrm{v}| - \mathrm{C}\ \mathrm{otherwise}\end{cases} \quad (3)$$

Where ω defines the nonlinear part, ϵ the curvature of the function and C a smoothing constant between the linear and nonlinear parts.

Endpoint: evaluate optical flow calculations [34] which analyzes the optical flow calculation error at each point in the x and y directions by calculating Euclidean distances.

$$\mathrm{endpoint}(\mathrm{U},\mathrm{V}) = \frac{1}{2\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\sqrt{(\mathrm{u}_{\mathrm{ix}}-\mathrm{v}_{\mathrm{ix}})^2+(\mathrm{u}_{\mathrm{iy}}-\mathrm{v}_{\mathrm{iy}})^2} \quad (4)$$

4. RESULTS AND DISCUSSION

The evaluation of our approach primarily revolves around its efficacy in recognizing facial expressions using the reconstructed data, rather than solely focusing on the accuracy of the reconstruction itself. To initiate the evaluation, we establish a well-defined experimental protocol, including the selection of an appropriate database and classifier, to measure the influence of our reconstruction technique on FER. Additionally, we introduce an optimization process for fine-tuning the various meta-parameters involved and proceed with a comprehensive evaluation in multiple stages to ensure the robustness and effectiveness of our proposed approach.

4.1. Experimental protocol

The CK+ database stands as one of the most utilized databases in the literature for evaluating recognition methods under the challenge of partial facial occlusions. Its popularity stems from two key factors. Firstly, CK+ offers a fully controlled environment, making it particularly suitable for studying occlusions. Secondly, it aligns well with our proposed approach due to its dynamic nature, containing 374 annotated video sequences, which makes it convenient for computing optical flows. This alignment allows us to effectively assess and demonstrate the effectiveness of our approach in handling occlusions and enhancing FER on a well-established and relevant dataset.

4.2. Experimental protocol for automatic FER

In our proposed approach, we opt to employ the architecture introduced by Allaert *et al.* [35], which has already demonstrated its efficacy in a FER framework centered around optic flow-based learning, as depicted in Figure 12. To evaluate our approach's performance, we employ a rigorous experimental protocol involving a 10-fold cross-validation. This process entails testing on each fold, validating on a different fold, and learning on the remaining folds successively. This systematic cross-validation ensures the robustness and generalization capabilities of our approach by thoroughly assessing its performance on various data subsets.

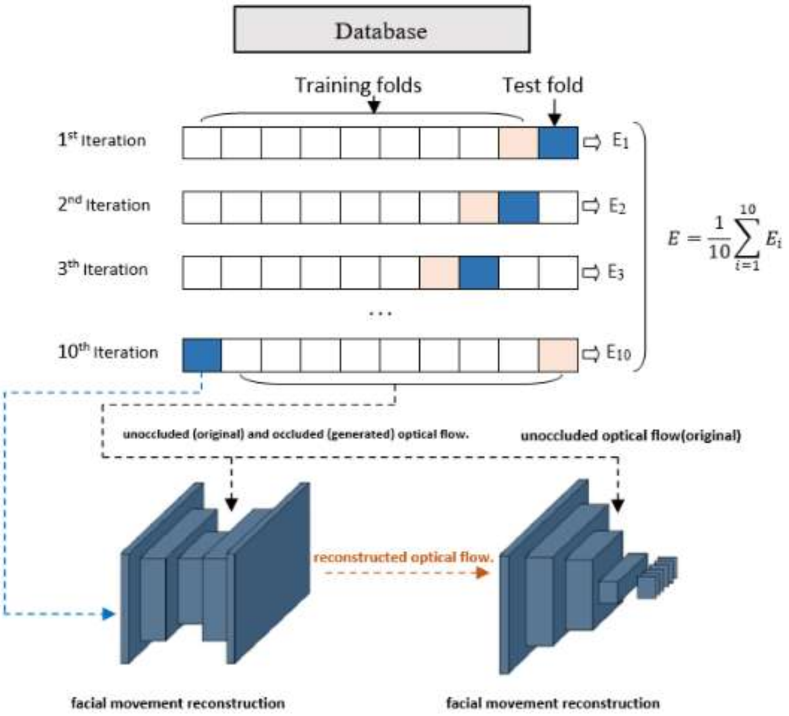

Database
Training folds
Test fold
1st Iteration
2nd Iteration
3th Iteration
10th Iteration
E_1
E_2
E_3
E_{10}
...
$E = \frac{1}{10}\sum_{i=1}^{10} E_i$
unoccluded (original) and occluded (generated) optical flow.
unoccluded optical flow(original)
reconstructed optical flow.
facial movement reconstruction
facial movement reconstruction

Figure 12. Evaluation of the approach with an experimental protocol in cross-validation in 10 folds. The results are obtained by testing successively on each fold, validating on a second fold, and learning on the 8 remaining folds

4.3. Recognition process parameterization

To enhance the classification step, we conduct a thorough examination of the optical flow's size with the goal of achieving a balance between satisfactory recognition scores and computational complexity reduction. The investigation involves testing various optical flow sizes during flow normalization, ranging from 24×24 to 128×128 pixels. The results presented in Figure 13 indicate that the most favorable size for obtaining the best recognition scores is 64×64 pixels.

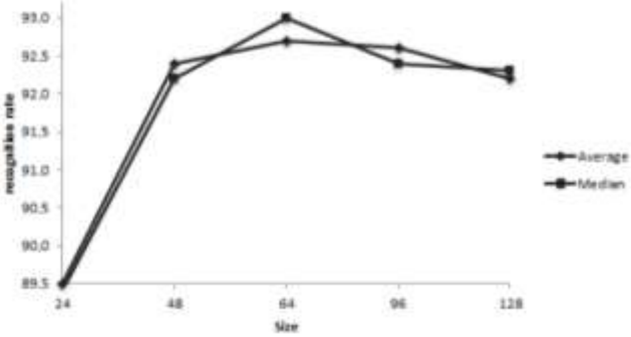
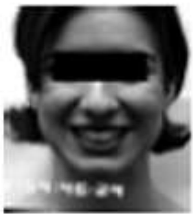
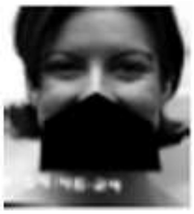


Figure 13. Results obtained as a function of the sizes of the optical streams used as input to the recognition CNN. The graph shows the median and average results obtained by averaging over 100 seeds for the sizes: 24×24, 48×48, 64×64, 96×96, and 128×128

In the previous section, the experimental protocol used to obtain the scores is explained in detail. The resulting scores are then compiled and displayed in Table 1, serving as an initial reference point for the evaluation of the proposed approach. This comparison is made with scores obtained using the same CNN architecture, but the training is conducted on non-occluded faces. The evaluation is performed in two scenarios: first, testing without any occlusion, and second, testing with various partial occlusions applied to the face. By conducting these tests, the proposed approach's effectiveness can be assessed in handling occluded facial data compared to the baseline performance on non-occluded faces.

Table 1. Comparison base with scores obtained without using the proposed approach. These results are obtained with the proposed CNN architecture by training on non-occluded faces and testing on the one hand, without occlusion and, on the other hand, in presence of different partial occlusions of the face

Original image	Eyes region occlusion	Mouth area occlusion
		
92.8%	73.9%	71.2%

4.3.1. Evaluation of the proposed method as a function of the cost function used In the evaluation process, the optical flows utilized are calculated between the initial (neutral) and final (apex) frames of each CK+ sequence. The results of the evaluation are depicted in Table 2, where the performance is analyzed based on different cost functions employed for backpropagation of the reconstruction AE architecture. The findings indicate that the endpoint is the most effective loss function, achieving a recognition rate of 87.2% for eye occlusion and 80.1% for mouth area occlusion. These results highlight the significance of the endpoint cost function in handling occluded facial data, demonstrating its superiority over other cost functions for this specific task.

Table 2. Results obtained as a function of the cost function used for backpropagation of the reconstruction autoencoder architecture

Cost function	Eyes region occlusion	Mouth area occlusion
MSE	86.2%	74.0%
Wing	86.1%	79.5%
EndPoint	87.2%	80.1%

In Table 3, the gains achieved through the utilization of different cost functions are presented. These gains are determined by calculating the difference between the results obtained using the proposed approach. By comparing the performance improvements with various cost functions, we can identify which approach yields the most significant gains in terms of recognition. This analysis helps in understanding the impact of different cost functions on the effectiveness of the proposed approach and guides in selecting the most suitable cost function.

Table 3. Gains obtained as a function of the cost function used. The gains are obtained by calculating the difference between the results obtained with the proposed approach

Cost function	Eyes region occlusion	Mouth area occlusion
MSE	+12.3%	+2.8%
Wing	+12.2%	+8.3%
EndPoint	+13.3%	+8.9%

4.3.2. Evaluation of the proposed method based on the skip connections added to the autoencoder reconstruction architecture

To enhance the quality of reconstructions, the approach draws inspiration from the Hourglass architecture, introducing connections between the encoder and decoder layers. These connections facilitate the retrieval of higher-level characteristics present in the encoder at the decoder level. By incorporating these connections, the decoder can access and utilize crucial information from earlier stages of the encoding process, leading to finer and more detailed reconstructions. Figure 14 illustrates the various types of connections studied, providing a visual representation of how information flows between the encoder and decoder modules, which contributes to the improved reconstruction performance.

Table 4 summarizes the set of results obtained by adding different connections on the reconstruction AE. The results in this table show that adding these connections can recover useful information for FER. These results also allow us to note that the best architecture is the one that contains residual connections between all hidden layers of the network.

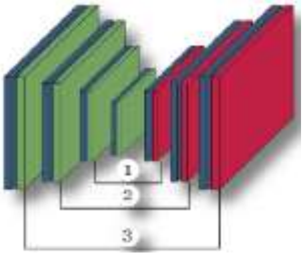


Figure 14. Illustration of different connections studied

Table 4. Recognition rates obtained as a function of different connections added on the reconstruction architecture

Connections	Eyes region occlusion	Mouth area occlusion
/	87.2%	80.1%
1	88.4%	81.1%
2	89.3%	81.8%
3	89.4%	81.0%
1+2	88.9%	82.6%
1+3	89.3%	82.5%
2+3	89.5%	82.6%
1+2+3	89.7%	83.2%

4.3.3. Evaluating the capacity for generalization

We now propose to evaluate the generalizability of our approach and the choice of these meta-parameters using another classifier. To do this, for the recognition stage, we use an support vector machine (SVM) classifier with radial basis function (RBF) kernel trained with the same protocol as above, integrating, for each iteration, the validation fold with the training folds. Optical flows are vectorized and used as SVM inputs.

Table 5 shows the results obtained with this new classifier. The first row of this table shows the results obtained without using reconstruction, i.e. training on unoccluded data and testing on occluded data. The next row shows the results obtained by testing on reconstructed data using our reconstruction approach.

We note that CNN extracts the features that are more robust to occlusions than the raw flow at the SVM input, particularly for the occlusion of the mouth. Indeed, occlusions have a greater impact in this experiment. Nevertheless, we note that reconstruction enables us to return to results comparable to those obtained with the CNN. The gain obtained thanks to reconstruction, noted below, is therefore more significant. These results demonstrate the generalizability of reconstruction to classifiers other than the CNN.

Table 5. Results obtained with an SVM classifier compared with results obtained with the CNN

	SVM		CNN	
Occluded region	Eyes region	Mouth area	Eyes region	Mouth area
Without reconstruction	74.9%	45.5%	73.9%	71.2%
With reconstruction	86.4%	78.1%	89.7%	83.2%
Gain	+11.5%	+32.6%	+15.8%	+12%

4.3.4. State-of-the-art comparison

The Table 6 shows a comparison between the results obtained by our approach with the parameters optimized in the previous sections and the results of other state-of-the-art approaches evaluated on CK+. We highlight the best results in this table by indicating, for each occlusion, the best result in bold and underlining the second best. Given the slight differences in results without occlusion, we add to the table the loss generated by the different occlusions, i.e., the difference between results without occlusion and results obtained by the different approaches. The proposed method is clearly competitive with the state-of-the-art. We also point out that we obtain higher results for occlusion of the mouth.

Table 6. Comparing the results of the proposed method with state-of-the-art results obtained from the CK+ database for eye and mouth occlusions

	Without occlusion	Eyes region occlusion	Mouth area occlusion
Huang *et al.* [15]	93.2%	93%/-0.2%	73.5%/-19.7%
Dapogny *et al.* [16]	93.4%	76%/-17.4%	67.1%/-26.3%
Our method	92.8%	89.7%/-3.1%	83.2%/-9.6%

The differences in results between these different solutions could be explained, among other things, by the choice of features exploited in each of these methods. In contrast to our approach, those of Huang *et al.* [15] and Dapogny *et al.* [16] are based on a static analysis of images and do not use temporal information, which could explain the slightly greater loss in results obtained. Huang *et al.* [15] propose an analysis based on temporal descriptors of shape and texture. The temporal dimension of these descriptors seems to allow additional information to be recovered despite occlusions. However, Huang *et al.* [15] do not rely on dense motion descriptors, which means that subtle facial deformations cannot be detected. However, when the epicenter of facial deformations is occluded, only subtle deformations may remain.

Our approach is based on an optical flow calculation, which enables us to retain this subtle information despite occlusions, which may partly explain the performance of our approach. What's more, our approach is the only one of the three to be based on reconstruction. Reconstruction could therefore also prove more effective than an analysis based on visible regions. Finally, we have selected the comparative approaches based on the experimental protocols closest to those proposed with our method.

5. CONCLUSION

In conclusion, our proposed approach offers a solution to address the occlusion problem in automatic FER by reconstructing hidden facial areas using the optical flow domain. The denoising AE architecture is well-suited for handling noisy data reconstruction and exploiting inter-personal similarity through motion. The primary objective is to recover facial expression-related information affected by occlusions. To enhance the method further, we plan to explore a GAN-based architecture with an adversary network focused on recognizing facial expressions. It is essential to note that the current approach is tailored to static occlusions observed between the first and last frames of each sequence. For a more comprehensive study, we intend to investigate the impact of dynamic occlusions, such as a hand passing in front of the face, on the computed optical flows, as these occlusions introduce additional movements unrelated to facial expressions.

REFERENCES

[1] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, Dec. 2014, doi: 10.1016/j.neucom.2014.05.008.

[2] S. Liu, Y. Zhang, and K. Liu, "Facial expression recognition under partial occlusion based on weber local descriptor histogram and decision fusion," in *Proceedings of the 33rd Chinese Control Conference, CCC 2014*, 2014, pp. 4664–4668, doi: 10.1109/ChiCC.2014.6895725.

[3] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, May 2019, doi: 10.1109/TIP.2018.2886767.

[4] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[5] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion," in *Procedings of the British Machine Vision Conference 2001*, 2001, pp. 23.1-23.10, doi: 10.5244/C.15.23.

[6] F. Bourel, C. C. Chibelushi, and A. A. Low, "Robust facial expression recognition using a state-based model of spatially-localised facial dynamics," in *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, 2002, pp. 113–118, doi: 10.1109/AFGR.2002.1004141.

[7] J. Y. R. Cornejo and H. Pedrini, "Emotion recognition from occluded facial expressions using weber local descriptor," *International Conference on Systems, Signals, and Image Processing*, vol. 2018-June, pp. 1–5, 2018, doi: 10.1109/IWSSIP.2018.8439631.

[8] Y. Lu, S. Wang, W. Zhao, and Y. Zhao, "WGAN-based robust occluded facial expression recognition," *IEEE Access*, vol. 7, pp. 93594–93610, 2019, doi: 10.1109/ACCESS.2019.2928125.

[9] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Y. preprint ArXiv:2001.06937, and U. 2020, "A review on generative adversarial networks: algorithms, theory, and applications," *IEEE transactions on knowledge and data engineering*, vol. 135, no. 4, pp. 3313–3332, 2021, doi: 10.1109/TKDE.2021.3130191.

[10] A. Kemmou, A. El Makrani, and I. El Azami, "An overview of the impact of partial occlusions on automatic facial expression recognition," in *Lecture Notes in Networks and Systems*, vol. 669 LNNS, Springer, 2023, pp. 451–462.

[11] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, Jul. 2008, doi: 10.1016/j.imavis.2007.11.004.

[12] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010, doi: 10.1109/JPROC.2010.2044470.

[13] S. F. Cotter, "Weighted voting of sparse representation classifiers for facial expression recognition," in *European Signal Processing Conference*, 2010, pp. 1164–1168.

[14] S. F. Cotter, "Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers," in *2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, Jan. 2011, pp. 437–442, doi: 10.1109/DSP-SPE.2011.5739254.

[15] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2181–2191, Dec. 2012, doi: 10.1016/j.patrec.2012.07.015.

[16] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *International Journal of Computer Vision*, vol. 126, no. 2–4, pp. 255–271, Apr. 2018, doi: 10.1007/s11263-017-1010-1.

[17] H. Towner and M. Slater, "Reconstruction and recognition of occluded facial expressions using PCA," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4738 LNCS, pp. 36–47, 2007, doi: 10.1007/978-3-540-74889-2_4.

[18] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh, and M. A. Hossain, "Adaptive facial point detection and emotion recognition for a humanoid robot," *Computer Vision and Image Understanding*, vol. 140, pp. 93–114, Nov. 2015, doi: 10.1016/j.cviu.2015.07.007.

[19] F. D. L. Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 362–369, doi: 10.1109/ICCV.2001.937541.

[20] I. T. Jolliffe, "Principal components in regression analysis," *Principal Component Analysis*, pp. 167–198, 2006, doi: 10.1007/0-387-22440-8_8.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[22] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 0, pp. 241–246, 2016, doi: 10.1109/ICDMW.2016.0041.

[23] Y. Zhang, R. Liu, S. Zhang, and M. Zhu, "Occlusion-robust face recognition using iterative stacked denoising autoencoder," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8228 LNCS, no. PART 3, pp. 352–359, 2013, doi: 10.1007/978-3-642-42051-1_44.

[24] Y. A. Chen, W. C. Chen, C. P. Wei, and Y. C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *Proceedings - International Conference on Image Processing, ICIP*, 2017, vol. 2017-Septe, pp. 1202–1206, doi: 10.1109/ICIP.2017.8296472.

[25] A. Dapogny, M. Cord, and P. Perez, "The missing data encoder: cross-channel image completion with hide-and-seek adversarial network," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020, pp. 10688–10695, doi: 10.1609/aaai.v34i07.6696.

[26] I. J. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, p. 9, 2014.

[27] Y. Li, S. Liu, J. Yang, and M. H. Yang, "Generative face completion," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 5892–5900, 2017, doi: 10.1109/CVPR.2017.624.

[28] V. Vielzeuf, C. Kervadec, S. Pateux, and F. Jurie, "The many variations of emotion," in *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, 2019, pp. 1–7, doi: 10.1109/FG.2019.8756560.

[29] Q. Wang, H. Fan, L. Zhu, and Y. Tang, "Deeply supervised face completion with multi-context generative adversarial network," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 400–404, Mar. 2019, doi: 10.1109/LSP.2018.2890205.

[30] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2857–2864, doi: 10.1109/CVPR.2011.5995710.

[31] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9912 LNCS, pp. 483–499, 2016, doi: 10.1007/978-3-319-46484-8_29.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, 2015, pp. 234–241.

[33] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 2235–2245, doi: 10.1109/CVPR.2018.00238.

[34] A. Dosovitskiy *et al.*, "FlowNet: learning optical flow with convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2758–2766, doi: 10.1109/ICCV.2015.316.

[35] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *arXiv preprint*, 2019.

BIOGRAPHIES OF AUTHORS

Abdelaali Kemmou received the Diploma of Master at the Faculty of Science, Ibn Tofail University, Kenitra. Morocco. In 2020 he joined the doctoral study center of Ibn Tofail University, Kenitra, Morocco. He is a member of the laboratory of research in computer science (L@RI). He is currently a Ph.D. researcher on the study and development of deep learning algorithms for big data. He can be contacted at email: kemabd@gmail.com.

Adil El Makrani is a Research Professor in Computer Science at the Faculty of Science, Ibn Tofail University, Kenitra. He received the Master in a computer science, computer graphics and imagery, and Ph.D. in Computer Science, Sidi Med Ben Abdellah University in 2009 and 2015, respectively. He affiliated to the Research in Informatics laboratory (L@RI). His research is currently focuses on artificial intelligence technologies, big data analytics, and their applications. He can be contacted at email: adil.elmakrani@uit.ac.ma.

Ikram El Azami currently works at the Department of informatics, Université Ibn Tofail. He does research in databases, machine learning, data mining and distributed computing. His most recent publication is "AraTrans the new transformer model to generate Arabic text". He can be contacted at email: ikram.elazami@uit.ac.ma.

Moulay Hafid Aabidi is Professor of Computing Science Research at the University Sultan Moulay Slimane, Higher School of Technology, Khenifra. His study focuses on the optimization of NP-hard issues with the metaheuristic approaches in artificial intelligence and big data for decision-making in diverse fields. Includes big data analytics, artificial intelligence, and information system. He can be contacted at email: myhafidaabidi@yahoo.fr.