HAL
open science

Facial Expressions Analysis Under Occlusions Based on Specificities of Facial Motion Propagation

Delphine Poux, Benjamin Allaert, José Mennesson, Nacim Ihaddadene, Ioan Marius Bilasco, Chaabane Djeraba

To cite this version:

Delphine Poux, Benjamin Allaert, José Mennesson, Nacim Ihaddadene, Ioan Marius Bilasco, et al.. Facial Expressions Analysis Under Occlusions Based on Specificities of Facial Motion Propagation. Multimedia Tools and Applications, In press. hal-02859856

HAL Id: hal-02859856

https://hal.science/hal-02859856v1

Submitted on 8 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Manuscript Click here to access/download;Manuscript;MTAP_Poux.pdf

Click here to view linked References

Noname manuscript No.
(will be inserted by the editor)

Facial Expressions Analysis Under Occlusions Based on Specificities of Facial Motion Propagation

Delphine Poux* · Benjamin Allaert · Jose Mennesson · Nacim Ihaddadene · Ioan Marius Bilasco · Chaabane Djeraba

Received: date / Accepted: date

Abstract Although much progress has been made in the facial expression analysis field, facial occlusions are still challenging. The main innovation brought by this contribution consists in exploiting the specificities of facial movement propagation for recognizing expressions in presence of important occlusions. The movement induced by an expression extends beyond the movement epicenter. Thus, the movement occurring in an occluded region propagates towards neighboring visible regions. In presence of occlusions, per expression, we compute the importance of each unoccluded facial region and we construct adapted facial frameworks that boost the performance of per expression binary classifier. The output of each expression-dependant binary classifier is then aggregated and fed into a fusion process that aims constructing, per occlusion, a unique model that recognizes all the facial expressions considered. The evaluations highlight the robustness of this approach in presence of significant facial occlusions.

Keywords Facial occlusions, motion propagation, facial framework, facial expressions

1 Introduction

Facial expression analysis is a field of research that has been extensively studied in recent years. Facial expressions give some clues about the emotional state of a

D. Poux (corresponding author), B. Allaert, I.M. Bilasco and C. Djeraba
Centre de Recherche en Informatique Signal et Automatique de Lille, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL -, F-59000 Lille, France
E-mail: {delphine.poux, marius.bilasco, chabane.djeraba}@univ-lille1.fr

J. Mennesson
IMT Lille Douai, Centre de Recherche en Informatique Signal et Automatique de Lille, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL -, F-59000 Lille, France
E-mail: jose.mennesson@univ-lille1.fr

N. Ihaddadene
ISEN Lille, Yncrea Hauts-de-France, France
E-mail: nacim.ihaddadene@yncrea.fr

person. A smile may indicate a happy person, a yawning indicates tiredness for example. This information is essential in many applications areas such as health, security or communication. Indeed, it is possible to automatically detect tiredness or anger for road safety, or to detect the level of pain of a patient for health applications.

The majority of the approaches dealing with facial expressions are generally trained on unoccluded faces and give very good results. However, these approaches perform poorly when deployed on un-controlled data (e.g., video surveillance system), where the face can be highly occluded. Two types of approaches have been proposed to address challenges in presence of occlusions. The first approach tend to reconstruct the occluded parts of the face and simulate an ideal analysis context. The second approach consists in characterizing the face despite the facial occlusion and let classifier identify the closest expression among the training data. In all cases, the facial expression analysis remains challenging when occlusions occur.

In this paper, we propose an innovative approach to overcome facial occlusions challenges. We assume that the facial movement induced by an expression is relatively close between individuals although the texture or facial geometry of each individual is highly different. The innovation brought by our contribution relies on the propagation properties of the facial movement. The movement induced by an expression spreads beyond the movement epicenter to neighbouring regions. Hence, if a region is occluded then it is possible to focus on the movement information that has been propagated to neighbouring regions. This paper is an extension of our previous work [21] where specific facial frameworks (i.e., specific sets of facial regions) are constructed per expression, according to the importance of each facial region to recognize the underlying expression in presence of specific occlusions. Only the most relevant regions are selected in order to be robust to both small and large occlusions. The previous work gets one output per expression and there is no unified process to recognize all facial expressions under one occlusion. In this paper, we propose to merge the per-expression facial frameworks into a unique model in order to recognize globally any given facial expression in presence of a specific occlusion.

As well as the majority of proposed approach to handle occlusions, we work on completely controlled dataset with frontal faces and simulated occlusions. Simulated occlusions allows the comparison of results with unoccluded data. Thus, we can quantify clearly the performance gap that is to various kind of occlusions. Because we are working on movement, we focus on the analysis of video sequences.

In Section 2, we highlight the main contributions of the paper and discuss approaches used to handle facial occlusions challenges. The construction of optimized facial frameworks per expression in presence of given occlusions is explained in Section 3. The merging of these facial frameworks is introduced in Section 4. In Section 5, we present the data used for learning and the experimental protocol used. Then, we present the performances obtained considering one expression at a time or all expressions simultaneously. In Section 6, we analyze the ability of the facial frameworks to recognize a given expression in presence of specific occlusions. In Section 7, we evaluate the generic expression recognition performance in presence of large occlusions and compare our approach to the other approaches from the literature. To conclude, we summarize the results and discuss perspectives in section 8.

2 Scope and background

This section starts with an overview of the state-of-the-art methods to recognize facial expressions. Later, we highlight the main objectives of our contribution and we provide a brief overview of existing approaches to handle facial occlusions.

2.1 Facial Expression Recognition

Facial expression can be studied statically or dynamically. Static approaches consider one image, usually associated to the expression apex, and focus on distinctive static features. Dynamic approaches encode changes in the evolution of the facial features over time starting from the onset of the expression until the apex.

Static facial expression recognition is mainly based on texture or geometric features extraction [5]. Texture features [23, 10] are based on the intensity of the pixels while geometrical ones [14] consider distances or deformations of the face, often based on facial landmarks. In all cases, these features encode relevant information about the facial expression and are used to train classifiers. Traditionally, KNN and SVM algorithms are used for this classification step [11]. Recently, deep learning based approaches have been proposed, essentially based on CNN architectures where features are learnt directly from the data [5].

Some works have studied the role of facial movement to recognize expressions and have shown that considering the whole activation sequence seem more informative than considering only the apex alone [2]. Existing descriptors have been then extended to support the temporal dimension. For instance, LBP [23] has been extended LBP-TOP [28] by encoding the changes overtime. Other works are based on the facial landmarks dynamics. Some approaches track these points to characterize the deformation of the face [24] in time. Recurrent neural networks such as LSTM have been proposed to keep track of temporal changes in deep learning approaches [25]. In order to characterize more precisely the movement, optical flow is particularly adapted and some features are derived from it. Recently, Allaert et al. [1] proposed a descriptor called Local Motion Patterns (LMP) based on optical flow. It characterizes the facial movement by retaining only the main directions related to facial expressions, while avoiding motion discontinuities. In order to characterize the movement within the face, LMPs are applied to small regions that are laid out on the face according to the facial muscles scheme. Hence, based on the relevance of the movement in the presence of facial expressions and the location of facial muscles, the face is segmented into 25 regions.

All these methods have proven their effectiveness in controlled situations. Nevertheless, these methods are still challenged under occlusions. Furthermore, Kotsia et al. [15] have shown that depending on the localisation of the occlusions the loss of performance differs greatly. For instance, occlusions of the mouth for example have greater impact than occlusions of the eyes regions. In the following section, we discuss how the current state-of-the art approaches deal with occlusions.

2.2 Background to overcome facial occlusions

Among the approaches proposed to handle occlusions, two categories can be distinguished: approaches that reconstruct the occluded parts of the face in order to retrieve an ideal analysis context, and approaches that characterize the face despite the occlusions.

Among the reconstruction approaches, the in-painting approach is the most commonly used [4,13,17,26]. In order to improve the reconstruction, recent works proposed to add some texture information from another unoccluded face. This unoccluded face is selected according to its similarities with the occluded one. Jampour et al. [13] use a guidance face to help the reconstruction of an occluded face. These solutions have proven their effectiveness for the task of face recognition because a similar face is chosen to reconstruct the occluded one. Nevertheless, based on the texture, this solution does not seem appropriate for facial expression recognition task where similar expressions do not necessarily imply similar faces. Recently, new approaches based on deep neural networks and more specifically on generative algorithms networks have been proposed [22,4,17,26]. These new approaches try to automatically reconstruct the hidden regions of the face thanks to a generative algorithm. However, these network architectures are large and parameter tuning process is complex. Besides, the large intra-face variation between individuals in the presence of facial expressions, the reconstruction of occluded regions remains relatively complex.

Regarding the approaches characterizing facial expressions despite the presence of occlusions, they can be divided in two categories : sparse representation approaches and sub-regions approaches. Sparse representation approaches recognize facial expression on an occluded face by representing a test image as a linear combination of unoccluded images from a dictionary [19,12]. This dictionary is composed of a set of unoccluded training images. Because the dictionary is composed of unoccluded data, occlusions cause errors in the linear combination. When these errors reach a threshold, they are implicitly considered as occlusions and are represented by an identity matrix which is isolated from the extracted facial features used by the classification process. These approaches have the advantage to implicitly localize occlusions. However, these approaches require large dictionaries covering variations for each expression in order to build accurate linear combinations and in order to have enough characteristics to discriminate between expressions. In the sub-regions approaches, the face is divided into different regions and each region is analyzed individually [7]. The results are then merged to recognize the expression. The advantage of these approaches is that they perfom well even in the absence of a large set of training data. However, the granularity of the subdivision of the face into local regions and its effect on performance remains an open question, particularly in the presence of important occlusions.

2.3 Contribution

Although much progress has been made in facial expression analysis field, facial occlusions are still challenging. Recent approaches proposed in the literature are not sufficient to properly characterize facial expressions in the presence of occlu-

sions. In addition, the large intra-face variety of individuals in presence of facial expressions increases the complexity of the learning process.

Considering the descriptors used to characterize facial expressions, majority of approaches are based on texture or geometry descriptors. However, in presence of an important facial occlusion, the information to characterize the facial expression is almost completely lost or has a high probability of being noisy due to estimation errors. Recent approaches have proven the effectiveness of optical flow in characterizing facial expressions [1]. Thanks to the physical properties of skin, descriptors based on movement seem adapted in the case of occlusion. Indeed, despite the fact that the epicenter of a movement is situated in an occluded part of the face, the movement related to the expression is still visible in the neighboring regions, as illustrated in Fig. 1 (see input data part of the image), where the motion induced by the smile has, as a secondary effect, the rise of the cheeks.

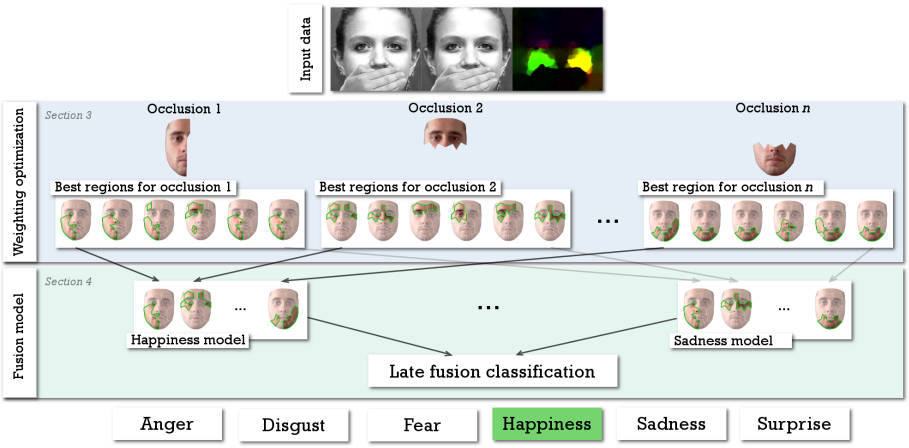


Fig. 1: Overview of the proposed approach.

Assuming that the movement within a face region spreads to neighbouring regions, we consider it appropriate to characterize facial expressions based on the evolution of movement through specific regions of the face. Inspired by the subregion approaches, we propose an innovative approach to overcome facial occlusions. Fig. 1 illustrates an overview of our approach which consists in recognizing facial expressions in presence of partial occlusions of the face. This approach is composed of two main steps. The first step consists in building optimized facial frameworks defining the facial regions contributing the most to the recognition of specific expressions in presence of a given occlusion. These facial frameworks are generated thanks to optimized weights computed for each facial region. These weights represent the contribution of each region to recognize a particular expression. The most important ones are selected in order to construct dedicated facial frameworks. The second step, illustrated in the lower part of Fig. 1, takes advantage of the obtained facial frameworks in order to train one binary model per expression. The results obtained with these binary models are then merged and a unique model per occlusion is trained in order to classify all expressions.

3 Weighting optimization algorithm

In this section, we investigate the best compromise between the minimum number of facial regions required to recognize facial expression and the performance obtained in different occlusions.

3.1 Weighting facial region scheme

The weighting algorithm consists in three steps. The first step generates various partial facial frameworks (a subset of facial regions), called configurations, including fewer regions than the initial facial framework. Inspired by [1], we consider a facial framework using 25 regions laid out following the facial muscle scheme. For each configuration C_j, the weighting algorithm evaluates the performance of the classification process using only the motion information contained in the regions R_j composing C_j. Then, the recognition rate obtained for a given configuration C_j serves to infer the contribution of each region R_j to the classification process.

3.1.1 Configurations

The choice of the retained configurations in the weighting algorithm is essential. Generating the whole set of configurations that covers all combinations of one to twenty-five regions is heavy and time consuming. Instead, in order to reinforce the motion propagation properties, we decided to consider only configurations containing pair-wise connected regions. As illustrated in Fig. 2-A, from the region R_{12}, the combinations $\{R_{12}\}$, $\{R_6, R_{12}\}$, $\{R_8, R_{12}\}$, $\{R_{14}, R_{12}\}$ and $\{R_{15}, R_{12}\}$ of size one and two are obtained. Indeed, the regions R_6, R_8, R_{14} and R_{15} are directly connected to the region R_{12}. Bigger combinations are obtained using the pair-wise connectivity of regions.

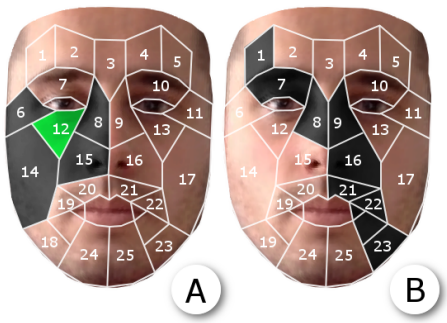


Fig. 2: Neighboring configurations.

We have chosen to explore configurations containing up to 8 regions as these configurations cover already horizontal, vertical and diagonal parts of the face as illustrated in Fig. 2-B. The configuration construction process guarantees that the

configurations cover several muscles of the face and enable us to study the correlation between them. Considering the configurations containing up to 8 regions, we had a total number of 21,294 different combinations.

We have chosen to explore configurations containing less than 8 regions as these configurations cover already horizontal, vertical and diagonal parts of the face as illustrated in Fig. 2-B. The configuration construction process guarantees that the configurations cover several muscles of the face and enable us to study the correlation between them.

3.1.2 Transferring weights to regions

The collected results obtained from all configurations are directly used to compute each region weight. At the beginning, each region receives a zero weight. Then, the classification rate obtained for each configuration C_j is used to compute a score according to the mean classification rate of all configurations normalized by the standard deviation. This score is calculated as follows :

$$\omega(C_j, emo) = exp((a(C_j, emo) - \mu_i)/\sigma_i)/exp(i) \tag{1}$$

where i is the number of regions of the configuration C_j, $j \in [1, 21294]$, $i = |C_j| \in [1, 8]$, and $a(C_j)$ refers to the accuracy obtained with the configuration C_j evaluated on the expression *emo*. μ_i and std_i are respectively the mean and the standard deviation of the results obtained with all the configurations containing i regions. Finally, $exp(i)$ which is the exponential of i, moderates the contribution of each configuration with regard to its size. Indeed, configurations covering larger portion of the face are expected to provide higher recognition rates.

The score obtained is then added to the current weights of each region R_j included in the configuration C_j. Finally, each region weight is normalized with regard to the number of apparition of the region in all the combinations. The obtained weights reflect the importance of each region for recognizing each expression.

Fig. 3 illustrates the heatmaps obtained on CK+ [20] dataset using the LMP descriptor which is a descriptor based on optical flow, an SVM classifier with RBF kernel and a 10-fold cross-validation protocol. Details about the specific parameters used for obtaining these heatmaps are provided in Section 6. We illustrate them here, in order to give a better understanding of the outcome of the weighting scheme. This figure reveals that for almost all expressions: the bottom of the face is activated, except for the anger, which mainly activates the eyebrows regions. Moreover, one can notice that these heatmaps are not symmetric. This asymmetry seems to be completely normal as some works have shown the asymmetry of facial expressions [9, 8].

The weight transferring process is represented in Fig. 4. The example shows the construction of the heatmap for the sadness expression. Fig. 4-A (bounded by the purple border) presents the heatmap obtained in absence of occlusions and in Fig. 4-B (bounded by the blue border) presents the heatmap obtained in presence of one occlusion.

As seen in the weighting heatmap obtained without any occlusion (i.e. considering entire set of configurations C), the most important regions for this expression are situated under the mouth. Considering the weight evaluation in presence of

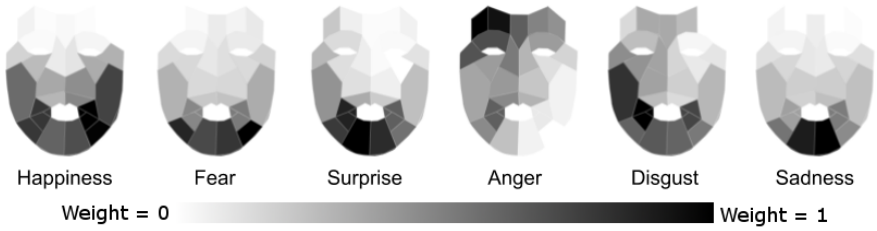


Fig. 3: Heatmaps of the importance of regions per facial expression computed using LMP descriptor [1] on CK+ dataset [20]. See Section 6 for details.

occlusions, the process is very similar to the unoccluded situation, but during the weight transferring part, we only consider configurations C that include only unoccluded regions (e.g., corresponds to the checked green configurations in Fig. 4-B). Thus, the importance of each visible region is computed independently from the occluded regions. Besides, occluded regions have a zero weight at the end of the process. This result is completely consistent because an occluded region gives no information about the facial expression.

As seen in the heatmap computed in presence of an occlusion impacting all the right part of the face, all configurations involving right regions are filtered out before transferring weights to regions. The resulting heatmap has zero weights for all regions on the right side of the face (blank areas) and the weights on the left side of the face are different with regard to the unoccluded heatmap notably for the cheek regions.

3.2 Optimizing facial framework for expressions recognition

The regions are sorted according to their weights in order to determine the ranking of the regions for each expression. This ranking is then used to generate models containing from one to twenty-five regions. Each model contains the n best regions for each facial expression. The obtained results reveals : a) the optimal facial frameworks for each facial expression; and b)the minimal number of regions required to recognize the expressions with performances similar to those obtained in absence of occlusion. These facial frameworks are illustrated in Fig.5 for the expression of happiness in presence of different occlusion patterns by selecting the 6 best regions.

4 Fusion of facial expression models

The weighing optimization algorithm allows the construction of one model per expression and per occlusion. Each model corresponds to a binary classifier and indicates if the input data corresponds to the underlying expression or not. In order to recognize an expression, regardless of the binary classifiers, we add a fusion step and, hence, construct a unified model for all expressions. The overview of the whole process are illustrated in Fig. 6.

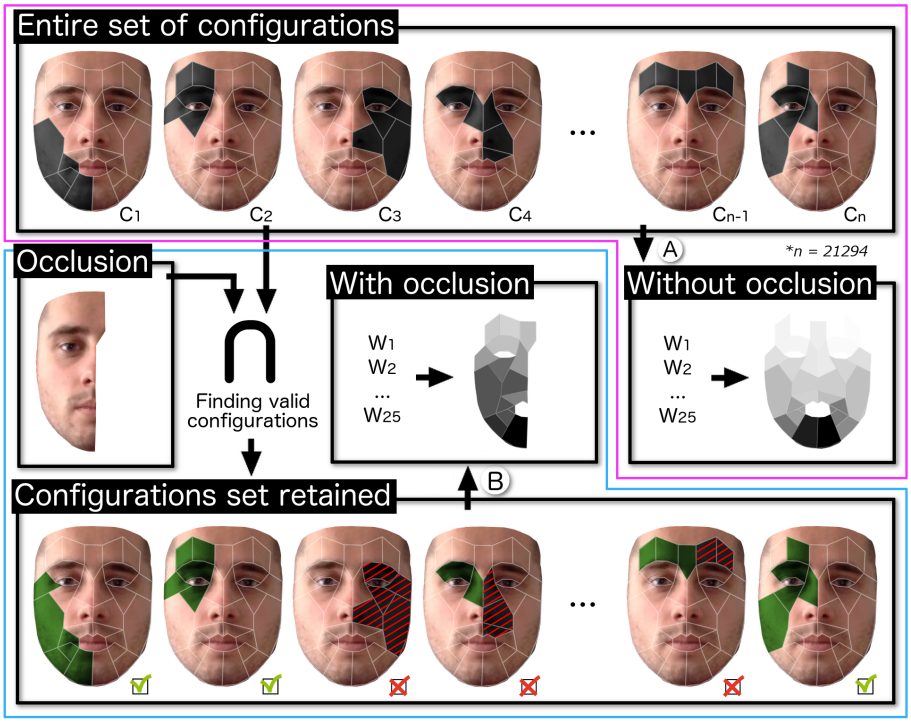


Fig. 4: Weights transfer considering facial occlusion (sadness expression).

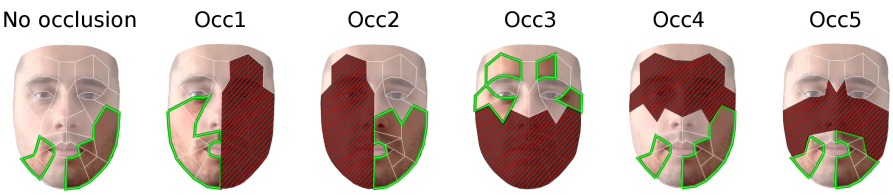


Fig. 5: Best facial frameworks for happiness expression under different occlusions.

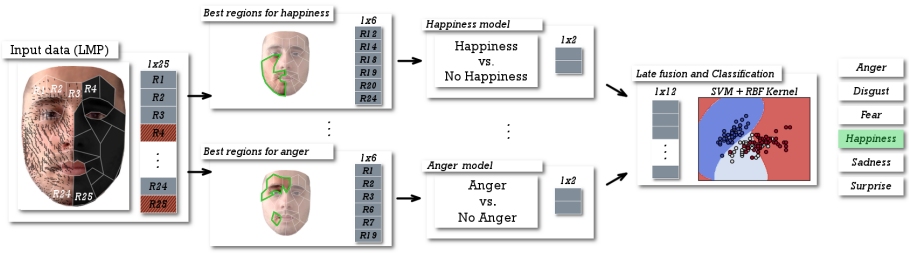


Fig. 6: Overview of the fusion process.

As we build a learning architecture using two layers, we proceed with two learning processes : one concerning the binary classifiers and one concerning the fusion layer. For each learning process, an adapted training set is prepared.

At first, six models are trained to recognize one expression against the others. In this case, we need a training set per expression (one expression against the others). For each model, the regions are first selected according to the x best regions that characterize an expression under a specific occlusion.

The constructed facial frameworks are then used to train, per expression, binary classifiers. The outputs of these models represent : a) the probability of an input sample to be classified as the underlying expression; and b) the probability of an input sample to belong to a different expression class.

A new training step is then performed with another training set which covers all expressions. In order to do this second training, raw data is fed to each binary classifier previously trained. The binary classifiers are not trained anymore and the models do not change. They are used only to compute, per expression, the probabilities that the input sample belongs or not to a specific expression class. These probabilities are concatenated into feature vectors that are fed into the fusion process.

5 Evaluation protocol

In this section, we present the protocol used to conduct our evaluations. First, we introduce the descriptors used to characterize facial movements. Then, we detail the dataset and the selected facial occlusions.

5.1 Facial motion characterization

To characterize the movement, we used the LMP descriptor proposed by Allaert et al. [1]. The LMP descriptor is based on optical flow which is particularly adapted to characterize the movement. Moreover, this descriptor has been created especially for facial expression recognition. Indeed, it takes into account the facial muscles scheme to filter discontinuities in optical flow. For experiments, we use the same segmentation used in Allaert et al. [1].

5.2 Dataset

The proposed approach is evaluated on the CK+ dataset as it is one of the most frequently used dataset in the literature to handle occlusions [6,7,12,18] and it contains video sequences which are adapted to study the movement. CK+ is a controlled dataset which contains 374 labelled video sequences. Each video sequence starts from the neutral face and ends with the apex of the expression.

In this dataset, images do not contain any occlusions, so, they have to be simulated. On one hand, occlusions are not totally realistic and there is a little gap between a real occlusion and a simulated one. But, on the other hand, we can totally control the experiments. By controlling the occlusion process, we can clearly quantify its impact on the overall performance. Besides, it offers also the possibility to construct precise benchmarks for comparison purpose.

5.3 Selected facial occlusions

Generally, the occluded regions are located at the level of the mouth and eyes, under different sizes. In order to simulate head pose variation, some approach hide half of the faces (right or left). Occlusion are often generated by the altering parts of the face by adding white, black or noisy pixels. Sometimes a blur effect can be applied instead. Some examples are presented in the left part of Fig. 7.

Not having a stable and widely accepted baseline to compare the performance of our approach on occluded faces, we choose to simulate important occlusions to challenge our approach, as illustrated in the right part of Fig. 7.

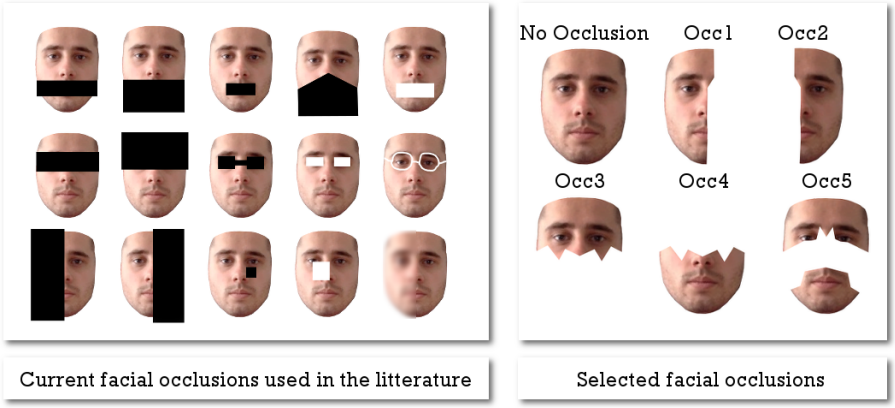


Fig. 7: Selected occlusions according to those used in the literature.

Inspired by the wide range of occlusions used in the literature, we choose a limited set of occlusions but which covers all the challenges. Indeed, the occlusion considered in our study present occlusions that impact larger facial area than those usually met in the literature.

The first two configurations (Occ1 and Occ2) present important occlusions on the left and right parts of the face. The third occlusion is inspired by the observations of Kotsia et al. [15] that underline the fact that the mouth has a great importance to recognize expression. Hence, in order to strongly challenge our approach, we define an occlusion configuration that impacts the mouth, the cheeks and the nose. Two other configurations consider important occlusions appearing on the upper part of the face and occlusions appearing in the middle part of the face.

6 Evaluation per expression

We propose a per expression evaluation in order to check if the constructed facial frameworks provide interesting results. We first evaluate the impact of the region selection when there is no occlusion. This first evaluation allows us to evaluate the accuracy of our weight calculation and, also, to find a minimal number of regions

required to recognize an expression. Then, we evaluate the efficiency of our per expression recognition method in presence of the selected occlusions.

6.1 Experimental protocol

In order to work individually with each expression and to build relevant model for per-expression recognition task, we generated several subsets of the CK+ dataset per expression. In each newly generated subset, all the sequences available for one expression are compared to a randomly stratified combination of all other expressions. For example the happiness subset contains two classes : happiness versus no-happiness. All the videos labeled happiness from the initial dataset are kept. For the no-happiness class, videos labeled with the five other expressions are randomly picked and a stratification scheme is employed in order to guarantee the same representativity of the other expressions as in the initial dataset. The defined distribution is described in Table 1.

Table 1: Per expression subsets size for the computation of per region weights.

	Happiness	Fear	Surprise	Anger	Disgust	Sadness	Total
Happiness subset	95	19	19	19	19	19	190
Fear subset	10	50	10	10	10	10	100
Surprise subset	16	16	80	16	16	16	160
Anger subset	7	7	7	35	7	7	70
Disgust subset	8	8	8	8	40	8	80
Sadness subset	13	13	13	13	13	65	130

For this evaluation, we have generated 25 configurations using one region, 46 configurations using two regions and so on until $12,827$ configurations using 8 regions. A total number of $21,294$ configurations are generated. The $21,294$ configurations are generated for each expression and all these models are sent to SVM classifiers. Weights are calculated for the twenty-five regions and for each expression. The regions are ranked according to the computed weights. The ranking is further used to generate twenty-five models by facial expression containing from one to twenty-five regions.

6.2 Impact of the selection process

Fig. 8 shows the results obtained for each facial expression using the sorted regions. These results show that this approach allows to be robust to really important occlusions. Indeed, facial expressions corresponding to surprise, happiness and disgust have quite optimal results with only one region. Sadness must have at least three regions to be recognized and anger needs at least six regions. This result is related to the complexity of the emotion. The anger and disgust expressions shares the same facial regions, which makes it hard to distinguish them with fewer regions. It is then necessary to take into consideration a larger number of regions.

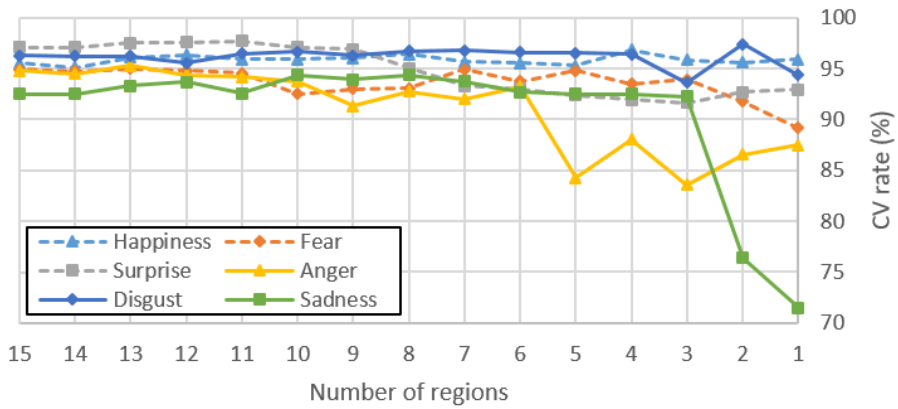


Fig. 8: Expression recognition rate according to the number of regions.

6.3 Efficiency of the approach on occluded faces

New weights are calculated for each occlusion and each facial expression to get specific models robust to the considered occlusion. For each expression and for each occlusion, the model which gives the best result is selected. We report results on the models containing the six best regions calculated for the occlusion on the facial expression. These 6-regions models are more stable with regard to some occlusions than the full facial framework using 25 regions. They reasonably limit the number of required unoccluded regions.

Fig. 9 shows the results obtained with and without intelligent facial frameworks per expression in presence of occlusions. The results obtained without intelligent facial frameworks are calculated with a model trained on unoccluded data using entire facial framework composed of the twenty-five regions. The optimized results are obtained with our approach by taking the best results considering the visible regions on one hand, and by taking the results given with the six best regions on the other. Finally, the black lines represent the results obtained in the case of an entire unoccluded dataset. According to these results, it is clear that our approach improves significantly the results even if only six regions are used. Indeed, considering the results provided by the best six regions calculated for each occlusion and for each expression gives results close to the best results obtained in unoccluded settings.

With regard to the obtained results, except for anger, the lower part of the face is really important for almost almost all expressions. Indeed, without optimization, the worst results are obtained with the mouth occlusion and the proposed approach significantly increases the performances in all situations. Concerning the anger expression, we can see that a lot of information is localized around the eyes. However, interesting results are obtained in spite of eyes occlusion. Finally, we can see that the occlusion of the nose and cheeks have an impact more or less important according to the expression. This observation shows that the effect of propagation of the movement give non negligible information.

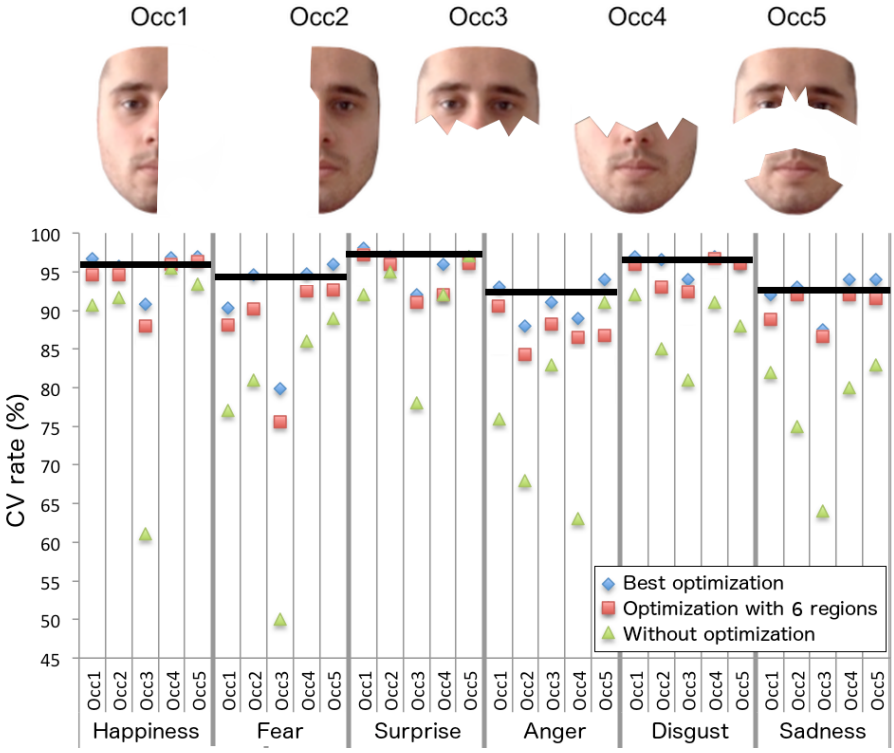


Fig. 9: Performance comparison with occlusion by expression on CK+.

7 Evaluation for all expressions

In this section, We present an evaluation of the entire process. We first evaluate the effectiveness of our approach to characterize the six universal expressions (happiness, anger, disgust, fear, sadness and surprise) under different facial occlusions. Then, a comparison with representative approaches from the literature is performed.

7.1 Experimental protocol

In our evaluation, we selected the six best regions for each expression for each occlusion calculated. The selected facial frameworks for each expression per occlusion are illustrated in Fig. 10.

In order to evaluate our approach, we had to split the dataset in two training subsets. One used for training the per expression models and the second one for training and evaluating the fusion model. We take 40% of the sequences to train the per expression model. The remaining 60% are then used to train and evaluate the fusion. Performances on each model are reported using a 10-folds cross-validation protocol. The detailed distributions of each expression is presented in Table 2.

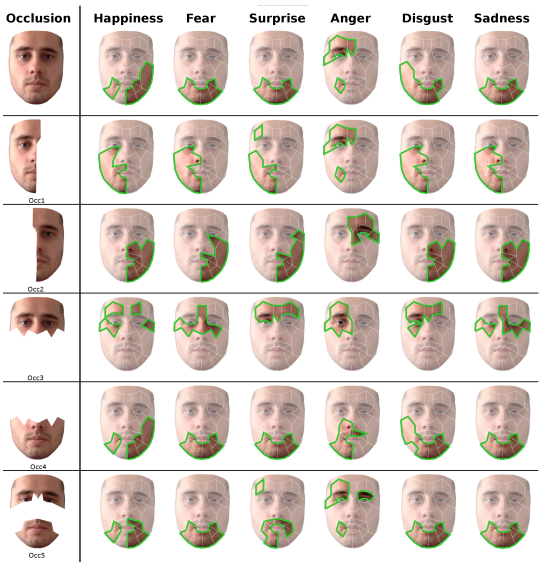


Fig. 10: The six best regions according to occlusions per expression.

‘

Table 2: Size of per expression subsets for the evaluation of the fusion step.

	Per expression models (40%)	Fusion model (60%)	Total number of data in CK+
Happiness	38	57	95
Fear	21	31	53
Surprise	33	49	83
Anger	14	22	37
Disgust	16	24	41
Sadness	26	39	65

For the first training sets, we need six different training sets : one per expression. In order to build these training sets, we take all sequences for the current expression. In order to have balanced distribution, a same number of data for the expression and for the others expressions are respected. Thus, we randomly pick 1/5 of the number of data of the expression for the five other expressions. By randomizing the initial dataset, we then created ten different sets of the training sets.

These results are calculated for the ten training sets and we report the mean result obtain for the ten runs. For each training set, an SVM classifier with RBF kernel are used with a 10-fold cross-validation protocol. Then, the average classification rates are reported.

7.2 Performances analysis

In order to evaluate the performance of our approach, we study three criteria. At first, we analyze the recognition performance of facial expressions in the presence

of different occlusions. Then, we compare our performance with other approaches proposed in the literature.

7.2.1 Performances analysis under occlusions

In this section, we analyze the performances of our approach to characterize the six universal facial expressions (happiness, sadness, disgust, fear, surprise and anger) under different occlusions. Table 3 shows the results obtained with our process with and without occlusion. The process without occlusion considers the six most important regions of the face to recognize each expression. The process with occlusion consider the calculated regions considering the several occlusions.

Table 3: Accuracy of our approach on CK+ dataset with and without occlusion.

No occlusion					
91.3%	73.4%	88.8%	89.0%	89.3%	90.7%

As observed in Table 3, we can conclude that the proposed approach is relatively robust in the presence of severe facial occlusions. As we can see, the results obtained with an occlusion of the bottom of the face drop significantly. It demonstrates the importance of the mouth regions about the expression and it is harder to compensate with the information found in the upper part of the face. To go further, the per expression results are presented in Fig. 11. This graph highlights the robustness of our method under the studied occlusions. As one could expect, the occlusion of the eyes regions has a strong impact on the expression of anger as these regions are really important. The confusion matrix of the upper part occlusion in Fig. 12 indicates that anger is mainly confused with sadness. This could be explained by the fact that these two expressions implies both pulling down of the lips. With these results, one can also notice that, the fear expression is particularly impacted by the bottom face occlusion. The confusion matrix of the lower part occlusion presented in Fig. 12 indicates that fear is mainly confused with surprise and sadness. Indeed, surprise and fear induce a raise of the eyebrows, while the inner corners of the eyebrows are pulled towards each other for both fear and sadness.

7.2.2 Performances comparison with others approaches

In this section, we compare the performance of our approach with the other approaches proposed in the literature on the CK+ database. Since there is no predefined baseline to compare the different approaches, we only analyze the occlusions that are closest to the other approaches. The results are represented in Fig. 13.

In Fig. 13, we compare our results with the results obtained by Kotsia et al. [15]. In this paper, they analyse the impact of facial occlusion on facial expression

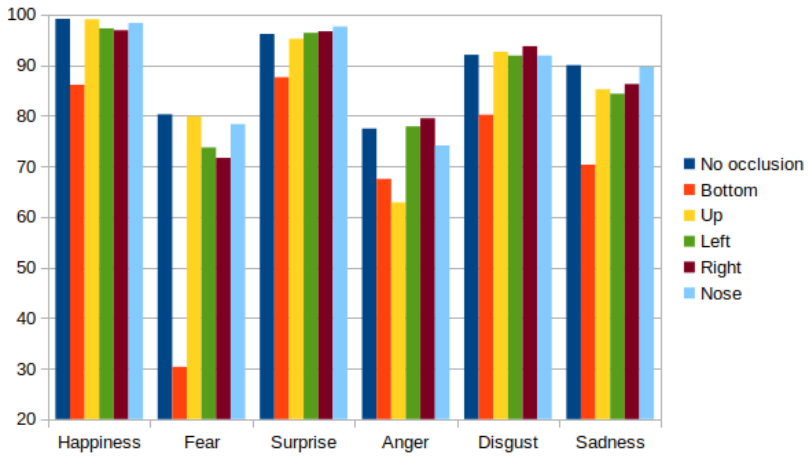


Fig. 11: Per expression results of our approach on CK+ dataset with and without occlusion.

Fig. 12: Confusion matrices obtained under occlusion of the lower and the upper part of the face.

recognition by analysing the results obtained by a human or by several descriptors. In Fig. 13 we have kept the results obtained respectively, with the DNMF algorithm [27] (column 1), with a descriptor based on the tracking of geometrical points of the face from the first frame to the last frame [16] (column 2) and, finally, with Gabor filters [3] (column 3). Gabor filters and DNMF algorithm are both texture-based features. We can notice from these results that, with exactly the same protocol, the results obtained by tracking geometric points, which takes into account the temporal aspect, seems to be more robust to occlusions. Dapogny et al. [7] proposed local descriptors for several regions of the face. These descriptors are fused to train the classifier. They also propose an auto-encoder trained to estimate confidence scores on the different regions. The obtained scores are used as weights for the

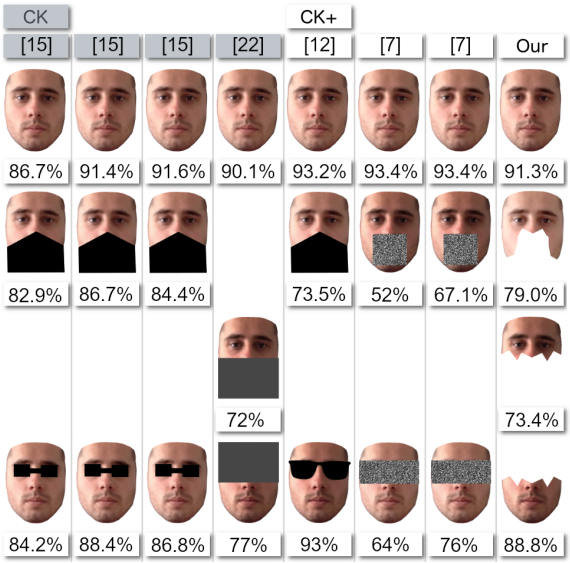


Fig. 13: Comparison of performances with others approaches in the literature.

different features during the fusion step. This weighting step is used to enhance the robustness to occlusions. The results obtained by using only the local features without the weights, and, the results with weighted features are shown in Fig. 13.

In view of the results, our approach gives very competitive performances. It is important to note that our occlusions are more severe than those used in other approaches except for [22] which may explain the difference with some approaches. To fairly compare the results obtained with other approaches, a less severe occlusion of the mouth is also presented in this comparison. This result as well as the one with the upper part of the face shows that our solution is really competitive with other approaches evaluated on the CK+ dataset. Moreover, the difference between the result obtained with and without the cheeks tends to show the importance of the cheeks and, thus, the importance of the propagation of the movement.

7.2.3 Analysis under realistic occlusions

Fig. 14 shows the qualitative results obtained on sequences captured in presence of real static occlusions. This figure illustrates, for each expression, the filtered optical flow, the selected regions for per expression recognition and the probabilities obtained for each model. The selected regions are outlined in green. As seen in this figure, the selected regions calculated with our method, focus the analysis on regions containing important and discriminant movements. The intermediate probabilities (reported below each thumbnail) are calculated in a cross dataset context. The per expression models and fusion models are trained on CK+ dataset. No data from the collected sequences is included in the training process.

Fig. 14: Qualitative results obtained in presence of real occlusions of the upper and lower part of the face.

8 Conclusion

In this paper, we design an approach that handle expression recognition in presence of occlusions. We propose, as a first step, a method to calculate a facial framework for each expression adapted to a considered occlusion. Based on the calculated facial frameworks, we propose then a fusion step in order to build an entire process which takes an input data and predict, at the end, the expression.

In order to do that, we pre-trained several models: one per expression in order to get the probabilities that the input data belongs to an expression class. These probabilities are then aggregated and they are used for training the fusion model.

The results obtained with this process are competitive with state-of-the-art methods, although we have considered larger occlusions. Nevertheless, it is still difficult to compare with other approaches especially due to reproductibility issues. One of our future work consists in building a benchmark regrouping a large set of occlusions and allowing the community to benefit from a stable evaluation framework.

References

1. Allaert, B., Bilasco, I., Djeraba, C.: Micro and macro facial expression recognition using advanced local motion patterns. IEEE Transactions on Affective Computing (2020)
2. Bassili, J.N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of personality and social psychology **37**(11), 2049 (1979)
3. Buciu, I., Kotsia, I., Pitas, I.: Facial expression analysis under partial occlusion. In: ICASSP, vol. 5, pp. v–453. IEEE (2005)
4. Chen, Y.A., Chen, W.C., Wei, C.P., Wang, Y.C.F.: Occlusion-aware face inpainting via generative adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1202–1206. IEEE (2017)
5. Corneanu, C.A., Simón, M.O., Cohn, J.F., Guerrero, S.E.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence **38**(8), 1548–1568 (2016)
6. Cornejo, J.Y.R., Pedrini, H.: Emotion recognition from occluded facial expressions using weber local descriptor. In: 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–5. IEEE (2018)
7. Dapogny, A., Bailly, K., Dubuisson, S.: Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. IJCV **126**(2-4), 255–271 (2018)
8. Dopson, W.G., Beckwith, B.E., Tucker, D.M., Bullard-Bates, P.C.: Asymmetry of facial expression in spontaneous emotion. Cortex **20**(2), 243–251 (1984)
9. Ekman, P.: Asymmetry in facial expression. Science **209**(4458), 833–834 (1980)
10. Ghimire, D., Lee, J.: Histogram of orientation gradient feature-based facial expression classification using bagging with extreme learning machine. Advanced Science Letters **17**(1), 156–161 (2012)
11. Huang Y.; Chen, F.L.S.W.X.: Facial expression recognition: A survey. Symmetry 2019 **11**(1189)
12. Huang, X., Zhao, G., Zheng, W., Pietikäinen, M.: Towards a dynamic expression recognition system under facial occlusion. Pattern Recognition Letters **33**(16), 2181–2191 (2012)
13. Jampour, M., Li, C., Yu, L.F., Zhou, K., Lin, S., Bischof, H.: Face inpainting based on high-level facial attributes. Computer vision and image understanding **161**, 29–41 (2017)
14. Kacem, A., Daoudi, M., Ben Amor, B., Carlos Alvarez-Paiva, J.: A novel space-time representation on the positive semidefinite cone for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3180–3189 (2017)
15. Kotsia, I., Buciu, I., Pitas, I.: An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing **26**(7), 1052–1067 (2008)

16. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE transactions on image processing **16**(1), 172–187 (2006)
17. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3911–3919 (2017)
18. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing (2018)
19. Liu, S.S., Zhang, Y., Liu, K.P.: Facial expression recognition under random block occlusion based on maximum likelihood estimation sparse representation. In: IJCNN, pp. 1285–1290. IEEE (2014)
20. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPRW, pp. 94–101. IEEE (2010)
21. Poux, D., Allaert, B., Mennesson, J., Ihaddadene, N., Bilasco, L.M., Dieraba, C.: Mastering occlusions by using intelligent facial frameworks based on the propagation of movement. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–6. IEEE (2018)
22. Ranzato, M., Susskind, J., Mnih, V., Hinton, G.: On deep generative models with applications to recognition (2011)
23. 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)
24. Tie, Y., Guan, L.: A deformable 3-d facial expression model for dynamic human emotional state recognition. IEEE Transactions on Circuits and Systems for Video Technology **23**(1), 142–157 (2012)
25. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., Rigoll, G.: Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing **31**(2), 153–163 (2013)
26. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514 (2018)
27. Zafeiriou, S., Tefas, A., Buciu, I., Pitas, I.: Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Transactions on Neural Networks **17**(3), 683–695 (2006)
28. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 915–928 (2007)

Click here to access/download;Author Biographies;author_biographies.pdf

Delphine Poux received his MS degree on Complex Model, Algorithms and Data in Computer Science from the University of Lille, France. She is currently a Ph.D. student at the Computer Science Laboratory in Lille (CRIStAL). Her research interests include computer vision and affective computing.

Benjamin Allaert received his MS degree on Image,Vision and Interaction and his Ph.D. on analysis of facial expressions in video flows in Computer Science from the University of Lille, France. He is currently a research engineer at the Computer Science Laboratory in Lille (CRIStAL). His research interests include computer vision and affective computing, and current focus of interest is the automatic analysis of human behavior.

José Mennesson is an Assistant Professor at IMT Lille Douai, France, since 2017. He received his MS degree on Calcul and Image and his Ph.D. on
Frequency methods for color image recognition in Computer Science from the University of La Rochelle. In 2013, he integrated the Computer Science Laboratory in Lille (CRIStAL, formerly LIFL). Since, he extended his research to human behavior understanding and image indexing.

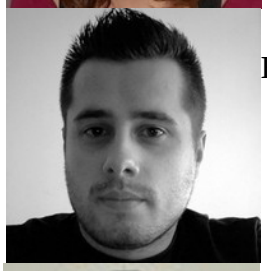
Nacim Ihaddadene is an Assistant Professor at the engineer school Isen, France. He received his MS degree on software architectures and his Ph.D. on
Extracting business process models from event logs in Computer Science from respectively the University of Nantes and the University of Lille. Then, he integrated the HEI ISA ISEN Group. Since, he extended his research to human behavior understanding.

Ioan Marius Bilasco is an Assistant Professor at the University of Lille, France, since 2009. He received his MS degree on multimedia adaptation and his Ph.D. on semantic adaptation of 3D data in Computer Science from the University Joseph Fourier in Grenoble. In 2008, he integrated the Computer Science Laboratory in Lille (CRIStAL, formerly LIFL) as an expert in metadata modeling activities. Since, he extended his research to facial expressions and human behavior analysis.

Chaabane Djeraba obtained a MS and Ph.D. degrees in Computer Science, from respectively the “Pierre Mendes France” University of Grenoble (France) and the “Claude Bernard” University of Lyon (France). He then became an Assistant and Associate Professor in Computer Science at the Polytechnic School of Nantes University, France. Since 2003, he has been a full Professor at the University of Lille. His current research interests cover the extraction of human behavior related information from videos, as well as multimedia indexing and mining.

Author Photographs

Delphine Poux

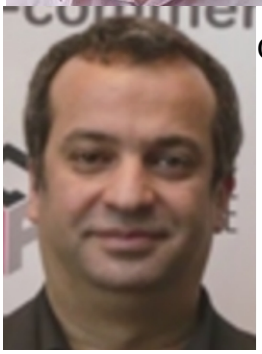


Benjamin Allaert

José Mennesson

Nacim Ihaddadene

Ioan Marius Bilasco



Chaabane Djeraba