Micro and macro facial expression recognition using advanced Local Motion Patterns

Benjamin Allaert, Ioan Marius Bilasco, Chaabane Djeraba

Micro and macro facial expression recognition using advanced Local Motion Patterns

Benjamin Allaert*, Ioan Marius Bilasco*, and Chaabane Djeraba*
*Centre de Recherche en Informatique Signal et Automatique de Lille, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL -, F-59000 Lille, France

Abstract—In this paper, we develop a new method that recognizes facial expressions, on the basis of an innovative Local Motion Patterns (LMP) feature. The LMP feature analyzes locally the motion distribution in order to separate consistent mouvement patterns from noise. Indeed, facial motion extracted from the face is generally noisy and without specific processing, it can hardly cope with expression recognition requirements especially for micro-expressions. Direction and magnitude statistical profiles are jointly analyzed in order to filter out noise. This work presents three main contributions. The first one is the analysis of the face skin temporal elasticity and face deformations during expression. The second one is a unified approach for both macro and micro expression recognition leading the way to supporting a wide range of expression intensities. The third one is the step forward towards in-the-wild expression recognition, dealing with challenges such as various intensity and various expression activation patterns, illumination variations and small head pose variations. Our method outperforms state-of-the-art methods for micro expression recognition and positions itself among top-ranked state-of-the-art methods for macro expression recognition.

Index Terms—Macro expression, Micro expression, Optical flow, Facial expression, Local motion patterns.

1 Introduction

Facial expression recognition has attracted great interest over the past decade in wide application areas such as human machine interaction, behavior analysis, video communication, e-learning, well-being, e-health and marketing. For example, during visio-conferences between several participants, facial expression analysis strengthens dialogue and social interaction between all participants (i.e keep the viewers attention). In e-health applications, facial expressions recognition helps to better understand patient minds and pain, without any intrusive sensors.

Facial expressions are fundamentally covering both micro and macro expressions [1]. It is a very important issue, because by essence, both micro and macro expressions as well as intermediate expressions are present during human interactions. Addressing expression recognition problem with a unified approach, regardless of the expression intensity, is one important requirement related to in-the-wild expression recognition. In this work, we focus on the micro and macro expression recognition as they represent extreme intensities. Proposing a unified approach coping at once with very large intensity variations leads up the way to the coverage of the full range of facial expression intensities.

The difference between macro and micro expression depends essentially of the duration and the intensity of expression, as illustrated in Figure 1.

Macro expressions are voluntary facial expressions, and cover large face area. The underlying facial movements and texture deformations can be clearly discriminated from noise. The typical duration of macro expression is between 0.5 and 4 s [1]. On the opposite, micro expressions are involuntary facial expressions. Often, they convey hidden emotions that determine true human feelings and state-of-mind. Micro expressions tend to be subtle manifestations of a concealed emotion under a masked expression. Micro expressions are characterized by rapid facial movements and cover restricted facial area. The typical duration of micro expressions is between 65 ms and 500 ms [4]. In terms of facial muscles movements and texture changes, micro expressions are characterized by low intensities [5].

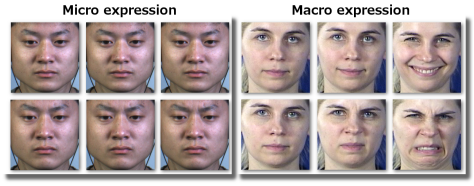

Fig. 1. Difference of motion intensity between micro and macro expression - happiness (line 1), disgust (line 2), from CASME II [2] and MMI [3], respectively micro and macro expression datasets.

We propose an innovative motion descriptor called Local Motion Patterns (LMP), with three main contributions. First, it takes into account mechanical facial skin deformation properties (local coherency and local propagation). Second, it empowers the construction of a unified method for micro expressions (disgust, happiness, repression, surprise) and macro expressions (anger, disgust, fear, happiness, sadness, surprise) recognition. When extracting motion information from the face, the unified method deals with inconsistencies and noise caused by face characteristics (skin smoothness, skin reflect and elasticity). Generally, related works on facial expression recognition have been proposed to deal separately with macro and micro expressions. The advantage

Manuscript received may 30, 2018; revised 2018. Corresponding author: B. Allaert (email: benjamin.allaert@univ-lille.fr).

of having a unified method characterizing both macro and micro expressions consists in its ability to cope with a large panel of facial expression intensities. Hence, a unified method narrows the gap towards in-the-wild settings with regard to intensity variations. Third, on the basis of local facial motion intensity and propagation, the method is the step forward and potentially suitable for in-the-wild expression recognition showing robustness to : motion noise, illumination changes (near infrared and natural illumination), small head pose variation and various activation pattern.

Our face expression recognition method is validated on representative datasets of facial expression recognition community for both micro (CASME II, SMIC) and macro expression (CK+, Oulu-CASIA, MMI, AFEW) recognition. Our method is better than the state-of-the-art methods for micro expression recognition, and is competitive with state-of-the-art macro expression recognition methods.

In section 2, we discuss works related to expression recognition. We introduce facial expression features, and recent expression recognition approaches. In section 3, we present our Local Motion Patterns (LMP) feature that considers local facial motion coherency (see "Feature extraction" part in Figure 2). In this section, we show how LMP deals with facial skin smoothness, reflection and elasticity. In section 4, we explore several strategies of encoding the facial motion for macro and micro expression recognition (see "Expression recognition" part in Figure 2). Experimental results, presented in section 5, outline the generalization capacity of our method for micro and macro expression recognition. Conclusions, summing up the main contributions, and perspectives are discussed in section 6.

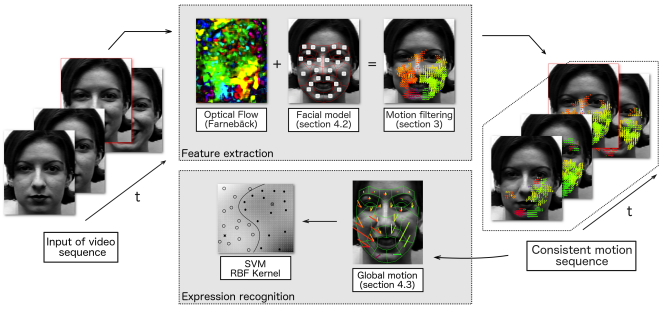

Fig. 2. Overview of our expression recognition method.

2 Related Work

This section presents the most significant macro and micro facial expression recognition approaches that have been proposed in the literature. Facial expression recognition approaches are based on features and face segmentation models. The facial segmentation model defines the regions of the faces from where information is extracted. The information is composed of features encoding changes in terms of texture and motion. We start the section by discussing features of macro and micro expression recognition, followed by face segmentation models. Finally, we focus on the combination of features and face segmentation models for macro and micro expression recognition.

2.1 Macro expression recognition

Important motions induced by face skin muscles characterize macro expressions. Furthermore, with regard to facial deformation, several types of techniques based on appearance and/or geometry are used to encode the changes.

Features, such as LBP [6] or HOG [7] obtained good results in the analysis of macro facial deformations. A similar comment applies to convolutional neural network (CNN) approaches [8]–[10] that learn a spatial feature representation on the apex frames. By relying on the spatial feature only, LBP, HOG and static CNN approaches do not exploit facial expression dynamics, which can limit the performance in presence of subtle expressions.

Psychological experiments by Bassili [11] showed that facial expressions can be recognized more accurately in a sequence of images. Therefore, a dynamic extension of LBP called Local Binary Pattern on Three Orthogonal Plans (LBP-TOP) is proposed in [12]. Considering the latest developments in dynamic texture domains, the optical flow regains interest from the community becoming one of the most widely used and recognized solution [13]. Optical flow estimates in a natural way the local dynamics and temporal texture characteristics. In recent deep learning approaches [14], [15], a recurrent neural network (RNN) was used with a conventional CNN to encode face dynamics and showed better performances compared to CNN only.

Most geometric approaches use Active Appearance Model (AAM) or variations like Active Shape Model (ASM), to track a dense set of facial points [16]. The location of these landmarks is then used in different ways to help extracting the shape- or motion-related facial features.

Hybrid approaches combine geometric and appearance features. As suggested in [17], combining them provides additional information to the recognition process. Jaiswal et al. [18] use a combination of Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM), which jointly learn shape, appearance and dynamics. They show that the combination of dynamic CNN features and BLSTM excels at modeling the temporal information. Several deep learning methods [15], [16] used a temporal geometric feature in order to reduce the effect of the identity on the learned features.

2.2 Micro expression recognition

Expression recognition approaches presented above, designed for macro expression, are not adapted to micro expression challenges (very short duration, low motion amplitude and limited texture changes). Liu et al. [19] apply directly macro expression approaches to micro expressions and show that detecting subtle changes by applying traditional macro expression approaches is a difficult task. Indeed, partial and low-intensity facial movements in micro expressions differ from ordinary expressions and it is difficult to split between true facial motion and noise due to head movement or motion discontinuities. The same conclusion has been drawn when using deep learning [20].

According to [21], micro expressions are much more difficult to detect without temporal information. Thus, researchers use spatio-temporal features for micro expression

analysis. Wang et al. [22] propose an extension of LBP-TOP based on the three intersecting lines crossing over the center point of the three histograms. They provide more compact and lightweight representation by minimizing the redundancy in LBP-TOP. Huang et al. [23] propose a new spatio-temporal LBP on an improved integral projection combining the benefit of texture and shape.

Although most micro expression recognition studies have considered LBP-TOP, several authors investigate alternative methods. Li et al. [21] employ temporal interpolation and motion magnification to counteract the low intensity of micro expressions. They show that longer interpolated sequences do not lead to improved performances, because the movement tends to be diluted. Interpolating micro expression segments using only 10 frames seems sufficient. Recently, Liu et al. [19] design a feature for micro expression recognition based on a robust optical flow method, and extract Main Directional Mean Optical-flow (MDMO). They showed that the magnitude is more discriminant than the direction when working with micro expressions. Furthermore, several deep-learning methods have been proposed to deal with micro expressions [14], [20], [24], but for now they all present lower performances than handcrafted approaches.

In this context, systems based on dynamic textures provide better performances. They allow detecting subtle changes occurring on the face while texture-based or geometry-based approaches fail in this case.

2.3 Face segmentation models

The face segmentation model, based on geometric information, defines the most appropriate layout for extracting face features in order to recognize expression. Assuming that face regions are well aligned; histogram-like features are often computed from equal-sized face grids [25]. However, apparent misalignment can be observed, primarily caused by face deformations induced by the expression itself. In most cases the geometric features are used to ensure that facial regions (eyes, eyebrows, lips corners) are well aligned.

Appearance features extracted from active face regions improve the performance of expression recognition. Therefore, some approaches define the regions with respect to specific facial locations (i.e. eyes, lips corners) using geometrical characteristics of the face [26].

Recent studies use facial landmarks to define facial regions. They increase robustness to facial deformation analysis during expression. Jiang et al. [27] define a mesh over the whole face with ASM, and extract features from the regions enclosed by the mesh. Sadeghi et al. [28] use a fixed geometric model for geometric normalization of facial images. The face image is divided into small sub-regions and then LBP histograms are calculated in each one for accurately describing the texture.

Face segmentation models based on salient patches, blocks, meshes or weighted masks have been explored overtime in combination with various features. Despite the use of similar features in macro and micro expression recognition, it is still difficult to find a unified facial segmentation model for analyzing macro and micro expressions together.

2.4 Synthesis

Micro expressions are quite different from macro expressions in terms of facial motion amplitudes and texture changes, which make them more difficult to characterize. Results from significant state-of-the-art approaches are illustrated in Table 1 and show the striking difference between macro and micro expression recognition performances.

Table 1 illustrates the established trends: appearance (static approaches), geometry and motion (dynamic texture and temporal approaches) in both macro and micro expression recognition fields. The main focus of the table resides in the difference in terms of performances between micro and macro expression recognition when the same underlying features and face segmentation models are used. Macro and micro expression recognition approaches are not directly comparable due to the fact that the underlying data is very different. However, we present them together in order to show that methods working well in one situation do not provide equivalent performances in the other. In order to allow an intra-category ranking, all macro expression approaches, cited in Table 1, use SVM as a final classifier and 10 fold cross-validation protocol. All cited micro expression approaches use SVM as a final classifier and leave-one-subject-out (LOSO) cross validation protocol.

TABLE 1
State-of-the-art methods for macro and micro expressions (* data augmentation).

Based on	Macro expression (CK+)		Micro expression (CASME II)	
App.	LBP [29] Block-based	90.05%	LBP [21] Block-based	55.87%
	PHOG [7] Salient region	95.30%	HIGO [21] Block-based	67.21% magnified
	CNN [8] Whole face	96.76% *	CNN [20] Whole face	47.30% *
Geom.	Gabor Jet [30] Facial points	95.17%	/	/
	DTGN [16] Facial points	92.35% *	/	/
Motion	LBP-TOP [12] Block-based	**96.26%**	DiSTLBP-IIP [23] Block-based	64.78%
	Optical flow [31] Facial meshes	93.17%	MDMO [19] Facial meshes	**67.37%**
	CNN + LSTM [14] Whole face	**98.62% ***	CNN + LSTM [24] Whole face	60.98% *

As shown in Table 1, well-known static methods like LBP have limited potential for micro expression recognition. The difference would be attributable to the fact that it cannot discriminate very low intensity motions [21]. LBP-TOP has shown promising performance for facial expression recognition. Therefore, many researchers have actively focused on LBP-TOP for micro expression recognition.

Geometric approaches deliver good results for macro expressions, but fail in detecting subtle motions in presence of micro expressions. Subtle motions require measuring skin surface changes. Algorithms tracking landmarks do not deliver the necessary accuracy for micro expressions.

Dynamic texture approaches are best suited to low facial motion amplitudes [23]. Specifically, methods based on optical flow appear to be promising for micro expression analysis [19]. Moreover, the optical flow approach proposed in [31] obtains competitive results in both macro and micro expression analysis. However, the optical flow approaches are often criticized for being heavily impacted by the presence of motion discontinuities and illumination changes. Recent optical flow algorithms (i.e. [32]) evolved to better

deal with noise. The majority of these algorithms is based on complex filtering and smooth motion propagation to reduce the discontinuity of local motion, improving the quality of optical flow. Still, in presence of high and low intensity of motion, the smoothing effect tends to induce false motions. Another technique consists in artificially amplifying the motion. This technique is being used increasingly and successfully in micro expression recognition [21]. The main disadvantage is the requirement of high intensity facial deformation. Such deformations alter significantly the facial morphology, especially in the presence of macro expression.

Concerning deep learning approaches, we underline important contrasts. On one hand, deep learning approaches provide good results for macro expression recognition (see * lines in Table 1). Deep learning approaches are based on auto-encoded features optimized for specific datasets. For example, Breuer and Kimmel [14] employ Ekman's facial action coding system (FACS) in order to boost the performances of their approach. On the other hand, deep learning results are clearly lower than handcrafted approaches in micro expressions recognition (Table 1).

Transposing efficiently features and face segmentation models from macro expression recognition to micro expression recognition is not yet achieved with regard to the current state-of-the-art. The selected representative works employ the same underlying feature in micro and macro expression recognition, however they need to change the facial segmentation model and the overall approach in order to maximize performances in both situations. Table 1 shows that it is still difficult to find a common methodology to analyze both macro and micro expressions accurately. However, for both, dynamic approaches seems promising.

Starting from these observations, we propose an innovative motion descriptor called Local Motion Patterns (LMP) that preserves the real facial motion and filters the motion discontinuity for both low and high intensities. Inspired by recent advances in the use of motion-based approaches for macro and micro expression recognition, we explore the use of magnitude and direction constraints in order to extract the relevant motion on the face. Considering the smoothing of motion in recent optical flow approaches, simple optical flow combined with magnitude constraint is appropriate for reducing the noise induced by illumination changes and small head movements. In the next section, we propose to filter optical flow information based on consistent local motion propagation to keep only the pertinent motion. Then, in section 4, we explore the construction of a unified facial segmentation model that generates discriminating features used to recognize effectively six macro expressions (anger, disgust, fear, happiness, sadness, surprise) and four micro expressions (disgust, happiness, repression, surprise).

3 Local Motion Patterns

The facial characteristics (skin smoothness, skin reflect and elasticity) induce inconsistencies when extracting motion information from the face. In our method, instead of explicitly computing the global motion field, the motion is computed in specific facial areas, defined in relation with the facial action coding system in order to keep only the pertinent motion on the face. The pertinent motion is extracted from regions where the movement intensity reflects natural facial movements. We consider natural facial movement to be uniform and locally continuous over neighboring regions.

We propose a new feature named Local Motion Patterns (LMP) that retrieves the coherent motion around epicenter ϵ(x,y) when considering natural motion propagation to neighboring regions. Each region, called Local Motion Region (LMR), is characterized by a histogram of optical flow $H_{LMR_{x,y}}$ of B bins. There are two types of LMR involved: Central Motion Region (CMR), and Neighboring Motion Region (NMR).

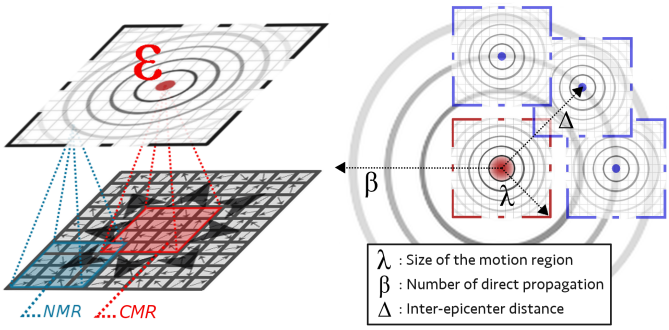

Fig. 3. Overview of local motion patterns (LMP) extraction.

LMP construction is illustrated in Figure 3. Eight NMR are generated around the CMR. All these regions are at distance Δ from the CMR. The bigger is the distance between two regions, the lower is the coherence in the overlapping area. λ is the size of the area under investigation around the epicenter. Finally, β characterizes the number of direct propagations from the epicenter that are carried out in order to retrieve all the coherent motions.

3.1 Local coherency of central motion region

In order to measure the consistency of the motion in terms of intensity and directions of LMP, we analyze the direction distribution in its CMR for several layers of magnitude. The motion on the face spreads progressively due to skin elasticity. We assume a regular progression of magnitude in specific directions.

We propose a method to compute the main direction in specific regions by analyzing jointly different layers of magnitude. This technique brings out main directions that are difficult to observe and reduces the motion noise.

The direction distribution of LMR is divided into q histograms, one per magnitude layer. The high intensity motion is more easily detected than low intensity motion. Each layer of magnitude is defined as following:

$$MH_{LMR_{x,y}}(n,m) = \{(bin_i, mag_i) \in H_{LMR_{x,y}} \mid mag_i \in [n,m]\}. \quad (1)$$

where n and m represent the magnitudes ranges and $i = 1,2,...,B$ is the index of bin. Each $MH_{LMR_{x,y}}$ is normalized, and directions representing less than 10%, are filtered out (set to zero). Then, magnitude layers are segmented into three parts $P_1 \in [0\%, 33\%]$, $P_2 \in]33\%, 66\%]$ and $P_3 \in]66\%, 100\%]$, represented by three cumulative histograms $ML_{LMR_{x,y}}(m1,m2)$ that are computed as follows:

$$ML_{LMR_{x,y}}(m1,m2) = \{(bin_i, card(\{(n,m) \mid \exists (bin_i, mag_i) \in MH_{LMR_{x,y}}(n,m) \mid mag_i \in [m1,m2]\}))\}. \quad (2)$$

The directional and magnified histogram $DMH_{LMR_{x,y}}$ is obtained by applying different weights to each part ω_1, ω_2 and ω_3 of the corresponding bins, as follows:

$$\begin{aligned} DMH_{LMR_{x,y}} &= ML_{LMR_{x,y}}(m1, m2) * \omega_1 \\ &+ ML_{LMR_{x,y}}(m2, m3) * \omega_2 \\ &+ ML_{LMR_{x,y}}(m3, m4) * \omega_3. \end{aligned} \quad (3)$$

in order to reinforce the local consistency of magnitude within each direction, we are applying 10-scale factors between layers ($\omega 1 = 1$, $\omega 2 = 10$ and $\omega 3 = 100$). We assume that the higher is the result, the higher is the pertinence of motion.

The motion filtering process is illustrated in Figure 4. Figure 4-A represents the histogram magnitude layers $MH_{LMR_{x,y}}$. Parameter n is varying between 0 and 10, by 0.2 magnitude steps. The parameter m is fixed to 10 in order to keep overlapping of magnitudes. The successive magnitude layers clearly distinguish the main direction. Next, the three magnitude layers $ML_{LMR_{x,y}}$ are represented in Figure 4-B, where each $ML_{LMR_{x,y}}$ corresponds to a row, and the number in each cell represents the number of magnitude occurrences for each bin. Finally, directional and magnified histogram $DMH_{LMR_{x,y}}$ is illustrated in Figure 4-C. The values associated with the black circles in Figure 4-C are computed by applying 10-scale factors to successive layers in 4-B and summing the results.

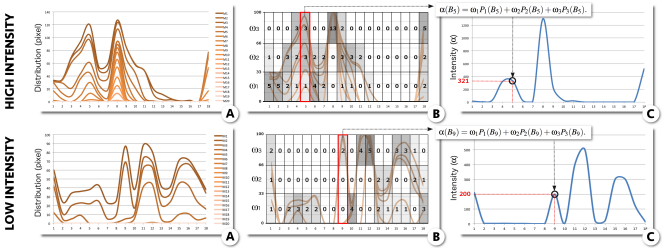

Fig. 4. The process of consistent local motion characterization in local motion region. (A) Magnitude histograms for different ranges, (B) Cumulative overlapping histograms, (C) Filtering motion.

Before analyzing the neighborhood and confirming the coherency of LMR, at least one main direction in the distribution shall be obtained after applying a fixed threshold α to the directional and magnified histogram ($DMH_{LMR_{x,y}}$). The threshold value reinforces the co-occurences of various intensities within the same direction bin. If no direction is found, LMR and LMP are locally incoherent. This means that the local intensity of motion does not exhibit the expected progressive behavior in any direction.

Afterwards, it must be ensured that the main orientation directions into $DMH_{LMR_{x,y}}$ are consistent. In fact, the local distribution in LMR can be consistent in terms of intensity, but it is possible to have a large number of bins with high values. This step ensures that the local motions spread coherently in the local neighborhood.

In order to ensure consistent distribution in terms of orientation, the density of k main directions is analyzed. Each main selected direction must satisfy several criteria. First criterion ensures that the main direction covers a limited number of bins (1 to s), where s is the threshold for the number of bin spans accepted. Indeed, if we analyze a small region in a face, a coherent facial motion is rarely spreading over more than 60°and the variance of movement is progressive. Otherwise, if one main direction is spreading over 60°, LMR stops analyzing the neighboring regions. Indeed, main directions spreading over 60°undermine the accurate identification of consistent motion by causing the propagation of false and misleading information. This criterion is defined by the following two equations. The first one characterizes the extent of main directions and the second filters out orientations spreading over s consecutive bins:

$$\begin{aligned} C(DMH_{LMR_{x,y}}) = \{E = [a..b] \mid \forall_i \in [a..b] \mid DMH_{LMR_{x,y}}(i) > \alpha \\ \wedge \nexists j \in \{a-1, b+1\} \mid DMH_{LMR_{x,y}}(j) > \alpha\}. \end{aligned} \quad (4)$$

$$C^{'}(DMH_{LMR_{x,y}}, s) = \{E \in C(DMH_{LMR_{x,y}}) \mid card(E) < s\}. \quad (5)$$

where $[a..b]$ represents the limits that the standard deviation of directions must meet and α is the threshold value of the intensity. Then, for each selected direction, we keep only the directions spreading over at most s consecutive bins.

In order to reinforce the fact that there is a gradual change in orientation, it is important that each main motion generates smooth transitions in terms of directions between neighbors. A maximum tolerance of Φ is supported as defined in the following:

$$\begin{aligned} C^{''}(DMH_{LMR_{x,y}}) = \{E = [a..b] \in C^{'}(DMH_{LMR_{x,y}}, s) \\ \mid \forall_{i,j} \in E, \|i - j\| \leq 1 \\ \mid \|DMH_{LMR_{x,y}}(i) - DMH_{LMR_{x,y}}(j)\| < \Phi\}. \end{aligned} \quad (6)$$

Finally, the filtered directional and magnified histogram $FDMH_{LMR_{x,y}}$ corresponds to k main directions in $DMH_{LMR_{x,y}}$. $FDMH_{LMR_{x,y}}$ is constructed as follows:

$$\begin{aligned} FDMH_{LMR_{x,y}} = \{(b_i, m_i) \in DMH_{LMR_{x,y}} \mid \exists E = [a..b] \\ \in C^{''}(DMH_{LMR_{x,y}}) \wedge b_i \in E\}. \end{aligned} \quad (7)$$

Despite CMR is considered coherent, LMP validation and computation have not yet been completed. Indeed, if we consider that natural facial movement is uniform during facial expressions, then the local facial motion should spread over at least one neighboring region.

3.2 Neighborhood propagation

When LMP is locally coherent in CMR, the approach verifies the motion expansion on neighboring motion regions (NMR). In some cases, physical rules (e.g. skin elasticity) ensure that local motion spreads to neighboring regions until motion exhaustion. Motion is subject to changes that may affect direction and magnitude in any location. However, intensity of moving facial region tends to remain constant during facial expression. Therefore, a pertinent motion observed and computed in CMR appears, eventually with lower or upper intensity, in at least one neighboring region.

Before analyzing the motion propagation, the local coherency of each NMR is analyzed with the same method discussed above for CMR. As for CMR, it must be ensured that the local distribution is consistent in terms of intensity and orientation. As an outcome of the process each locally consistent NMR_i is characterized by $FDMH_{LMR_{x_i,y_i}}$. However, it is important to check that the local distribution is similar to some extent with the previous adjacent

neighbor. Bhattacharyya coefficient is used to measure the overlap between two neighboring LMR as follows:

$$C'''(FDMH_{LMR_{x1,y1}}, FDMH'_{LMR_{x2,y2}}) = \sum_{i=1}^{B} \sqrt{FDMH_{LMR_{x1,y1}}(i) FDMH'_{LMR_{x2,y2}}(i)}. \quad (8)$$

where $FDMH_{LMR_{x1,y1}}$ and $FDMH'_{LMR_{x2,y2}}$ are the local distributions and B is the number of bins. LMR is considered consistent with his neighbor, if the coefficient is lower than the fixed threshold ρ.

The motion propagation into LMP after one iteration is given in Figure 5. If the local motion is inconsistent, NMR are represented in gray. If NMR are not coherent, three situations can be distinguished: a) the motion in NMR is locally inconsistent in terms of intensity; b) the motion in NMR is locally inconsistent in terms of orientation, and c) the distribution similarity between two regions is inconsistent.

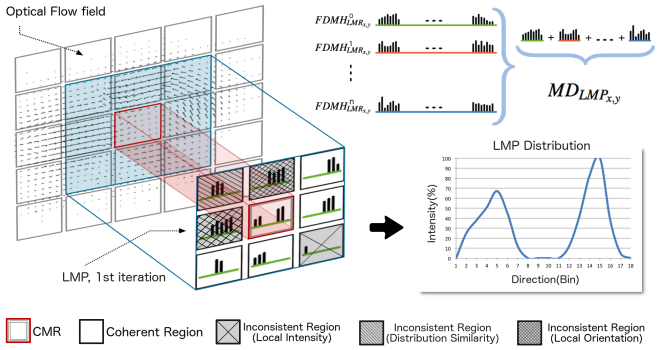

Fig. 5. LMP distribution, computed from the propagation in neighborhood of central motion region.

As long as at least one newly created NMR is inter-region coherent with its neighbor, recursively, for each subsequent NMR, the motion analysis is reconducted. The recursive process ends when the number of propagations β is reached.

Finally, each distribution ($FDMH_{LMR_{x,y}}$) corresponding to NMR that have direct or indirect connections to original CMR is cumulated into the LMP distribution. If the motion propagation between all NMR is inconsistent, the motion propagation is no more explored. This means that there are no more pertinent motions to collect into LMP. The final motion distribution of LMP is computed as follows:

$$MD_{LMP_{x,y}} = \{\sum_{i=0}^{n} FDMH_{LMR_{x,y}} \mid FDMH_{LMR_{x,y}} \in LMP_{x,y}\}. \quad (9)$$

where n is the number of consistent regions.

$MD_{LMP_{x,y}}$ is defined as a histogram over B bins, which contains, for each bin the sum of main direction intensities collected from coherent NMR and CMR. Then we are able to extract the coherent motions from a specific location on the face.

In summary, the proposed LMP feature collects pertinent motions and filters out the noise based on three criteria: convergence of motion intensity in the same direction, local coherency of direction distribution and coherent motion propagation. Each criterion can be configurable independently of the others, which makes it fully adaptable to many uses and contexts such as action recognition, facial expression recognition, tracking and other. To prove the effectiveness of our LMP, we analyze in the next section, the use of LMP for micro and macro facial expression analysis.

4 Expression Recognition

The choice of the facial segmentation model impacts greatly the performances. Various epicenters can be considered for coherent motion extraction. So, we study the impact of epicenters on the perceived motion while applying LMP. We show that the intensity of expression (macro or micro) plays a key role in locating LMP epicenter and, in the meantime, it impacts the way the consistent motion on the face is encoded. Then, we explore the integration of the coherent optical flow into facial model formulation, and discuss several strategies for considering discriminant local regions of the face.

4.1 Impact of LMP location

For macro expressions, motion propagation covers large facial area. If one CMR (Central Motion Region) is randomly placed in an area around the motion epicenter, then the motion consistency is most of the time observed. However, for micro expressions, the motion propagation covers restricted facial area. Motions are less intense, so motion propagation is discontinued. Figure 6 shows local motion distribution extracted in various points around left lip corner (blue, red and green dots). The original flow field and the local motion distribution extracted from a happiness sequence around the different locations are shown in the first three columns. The fourth column shows the distribution overlap.

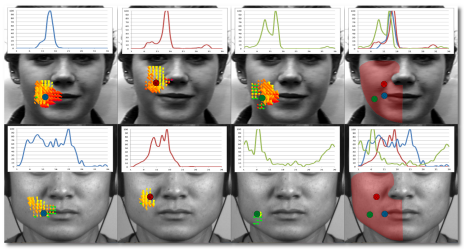

Fig. 6. Consistent motions from happiness sequence computed from different locations in the same region.

For macro expressions (first line), the location of each LMR is different, still the distributions present large overlaps (column 4). For micro expressions (second line), the distributions corresponding to the three columns are different. The experimentation can be reproduced in other facial regions with similar outcomes with regard to micro versus macro expressions. It is hence important to determine best discriminant facial regions for encoding coherent motion in the context of generic expression recognition process.

4.2 Best discriminant facial region

Macro and micro expression motions are very different in terms of intensity and propagation. It is therefore important to detect pertinent motions that generate features able to discriminate effectively some of the most common macro

expressions (happiness, sadness, fear, disgust, surprise and anger) and micro expressions (happiness, disgust, surprise, repression). In order to identify optimal LMP epicenters locations, we have considered samples for CK+ and CASME2 datasets.

To identify the locations within the face where motion often occurs, we first align frames based on eyes location, and we compute the optical flow of each frame of the sequence. This step eliminates in-plane head rotation and addresses individual differences in face shape. Then, each frame is segmented in $20x30$ blocks. LMP is extracted from each block, with LMP epicenter situated at the center of each block. Then, the consistent motion vector is computed in each LMP. Next, each relevant optical flow extracted from each frame is merged into a single binary motion mask. The consistent motion mask as well as motion information are extracted from video sequences of the same expression class. Finally, each consistent motion mask is normalized and merged into a heat map of motion for the underlying expression. The six consistent motion masks for the basic macro expressions are illustrated in the first line of Figure 7.

The extracted mask indicates that pertinent motions are located below the eyes, in the forehead, around the nose and mouth, as illustrated in Figure 7. Some motions are located in the same place during elicitation for several expressions, but they are distinguishable by their intensity, direction and density. For example, anger and sadness motions are similar as they appear around the mouth and the eyebrows. However, when a person is angry, motion is convergent (e.g mouth upwards and eyebrows downwards), and motion is divergent when a person is sad.

The same strategy for finding the best discriminant regions was used on CASME II dataset for micro expressions (happiness, disgust, surprise and repression). As illustrated in the second line of Figure 7, the pertinent motions are located near the eyebrows and the lips corner. Compared with macro expression motion maps, propagation distances are highly reduced for micro expressions.

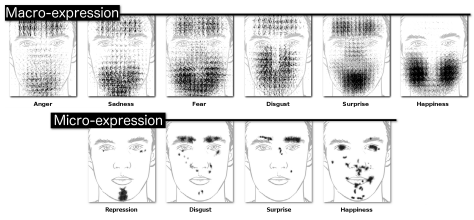

Fig. 7. Pertinent motions elicitating 6 macro expressions from CK+ dataset (top) and 4 micro expressions from CASME II dataset (bottom).

At this stage, the main facial regions of motion are accurately identified. We now construct a vector that encodes the relationships between facial region of motion and expressions. We use the facial landmarks to define regions that increase deformation robustness during expression. Similarly to Jiang et al. [27], the landmarks are used to define a mesh over the whole face, and a feature vector can be extracted from the regions enclosed by the mesh. Landmarks and geometrical features of the face are used to compute the set of points that defines a mesh over the whole face (forehead, cheek). Finally, the best discriminant landmarks are selected corresponding to active face regions, and specific points are computed to set out the mesh boundaries.

The partitioning into facial regions of interest (ROIs) is illustrated in Figure 8. The partitioning is based on the facial motion observed in the consistency maps constructed from both macro and micro expressions. The locations of these ROIs are uniquely determined by landmarks points for both micro and macro expressions. For example, the location of feature point P_Q is the average of two landmarks, P_{10} and P_{55}. The distance between eyebrows and forehead feature points (P_A,P_B,...,P_F) corresponds to the size of the nose $Distance_{P_{27},P_{33}}/4$. This allows maintaining the same distance for optimal adaptation to the size of the face. Note that, in order to deal precisely with the lip corners motion, regions 19 and 22 overlap regions 18 and 23, respectively.

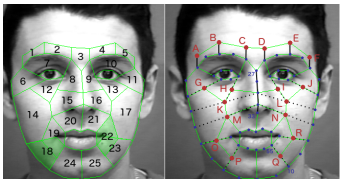

Fig. 8. Facial partition in interest regions based on facial muscles.

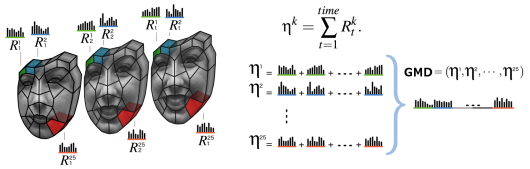

Fig. 9. Building the feature vector from the facial motion mask.

4.3 Facial motion descriptor

The facial motion mask is defined by the 25 ROIs presented above. In each frame f_t, we consider the filtered distribution motion inside each ROI R_t^k, where t is the frame index and $k = 1, 2, ..., 25$ is the ROI index. Inside each R_t^k, LMP is applied and $MD_{LMP_{x,y}}$ is computed as defined in equation 9. R_t^k motion distributions are summed into η^k, which corresponds to local facial motion in region k for the entire sequence.

$$\eta^k = \sum_{t=1}^{time} R_t^k. \tag{10}$$

Finally, histograms η^k are concatenated into one-row vector $GMD = (\eta^1, \eta^2, ..., \eta^{25})$, which is considered as the feature vector for the macro and micro expression. The feature vector size is equal to the number of ROI multiplied by the number of bins. An example is illustrated in Figure 9, where all motion distributions $MD_{LMP_{x,y}}$ corresponding to R_t^1, R_t^2 ... R_i^{25} with $t \in [1, time]$ are summed up in η^1, η^2 ... η^{25} respectively. η^1, η^2 ... η^{25} are then concatenated, and define the global motion distribution GMD.

4.4 Facial expression recognition framework

The framework, presented in Figure 2, is suitable to micro and macro expressions. First a preprocessing step is considered in order to extract landmarks and compute the 25

ROIs. Then Farnebck algorithm [33] is used to compute fast dense optical flow. It ensures that motion is not affected by smoothing and the computation time is low. LMP features are extracted from each ROI. Next, relevant motion in each facial region is cumulated over time. Each facial region is represented by a histogram based on the orientation and the intensity of motion. The concatenation of the histograms extracted from the various regions defines the feature vector used for classifying each video sequence. In order to evaluate the benefit brought by mixing motion and geometric information, for some of the experiments introduced in the next section, we enrich the feature vector with the shape characteristics of each ROI.

5 Evaluation

We highlight the performances obtained by our method on widely used datasets for micro expression recognition, namely CASME II [2] and SMIC [34], and widely used datasets for macro expression recognition, namely CK+ [35], Oulu-CASIA [36], MMI [3] and AFEW [37]. Experiments on these datasets cover aspects of in-the-wild recognition such as: head movement, illumination, visible and infrared contexts.

After introducing the datasets, we compare our performances with some major methods in the literature. We use LIBSVM [38] with RBF kernel and 10 fold cross-validation protocol for macro expressions and leave-one-subject-out (LOSO) for micro expressions [1].

5.1 Datasets

CASME II (micro expression dataset) contains 247 spontaneous micro expressions from 26 subjects, categorized into five classes: happiness (33 samples), disgust (60 samples), surprise (25 samples), repression (27 samples) and others (102 samples). The micro expressions are recorded at 200 fps in well-controlled laboratory environment.

SMIC (micro expression dataset) is divided into three sets : (i) HS dataset is recorded by high-speed camera at 100 fps and includes 164 sequences from 16 subjects, (ii) VIS dataset is recorded by standard color camera at 25 fps; and (iii) NIR dataset is recorded by near infrared camera at 25 fps. The high-speed (HS) camera was used to capture and record the whole data, while VIS and NIR cameras were only used for recording the last eight subjects (77 sequences). The three datasets include micro expression sequences from onset to offset. Each sequence is labeled with one of the following emotion classes: positive, surprise and negative.

CK+ (macro expression dataset) contains 593 acted facial expression sequences from 123 participants, with seven basic expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). In this dataset, the expression sequences start at neutral state and finishes at apex state. Expression recognition is completed in excellent conditions, because the deformations induced by ambient noise, facial alignment and intra-face occlusions are not significant with regard to the deformations directly related to the expression. However, the temporal activation pattern is variable and spreads from 4 frames to 66 frames with a mean sequence length of 17.8 ± 7.42 frames.

Oulu-CASIA (macro expression dataset) includes 480 sequences of 80 subjects taken under three different lighting conditions: strong, weak and dark illuminations. They are labeled with one of the six basic emotion labels (happiness, sadness, anger, disgust, surprise, and fear). Each sequence begins with neutral facial expression and ends with apex. Expressions are simultaneously captured in visible light and near infrared. Varying lighting conditions influence the recognition process.

MMI (macro expression dataset) includes 213 sequences from 30 subjects. The subjects were instructed to perform six expressions (happiness, sadness, anger, disgust, surprise, and fear). Subjects are free of their head movements and expressions show similarities with in-the-wild settings. Compared with CK+ and Oulu-CASIA, due to more important head pose variations of subjects, MMI is more challenging.

AFEW (macro expression dataset) contains sequences extracted from movies and is divided into three sets: Train (773 samples), Validation (383 samples) and Test (653 samples). In this experiment, we used the VReco sub-challenge data which consists in classifying a sample audio-video clip into one of the seven categories: anger, disgust, fear, happiness, neutral, sadness and surprise. Compared to other selected macro expression dataset, AFEW is the most challenging one, as it presents close to real world situations.

5.2 Micro expression

In this section, we show the experiment results on CASME II and SMIC micro expression datasets, followed by discussion and analysis of the results.

Experiments on CASME II Table 2 shows a comparison of our results with regard to the major state-of-the-art micro expression methods. In our method, the optical flow is calculated from two consecutive frames without any magnification nor temporal interpolation. For these experiments, we select only the activation part (e.g. onset to apex) from each sequence.

TABLE 2
Performances on CASME II dataset using LOSO *(* data augmentation)*
.

Method	Interpolat.	Magnifi.	Acc(%)
Baseline [2]	✗	✗	63.41%
LBP-SIP [22]	✗	✗	67.21%
Deep feat. (CNN) [20]	✗	✗	47.30%
STLBP-IIP [23]	✗	✗	62.75%
DiSTLBP-IPP [23]	✗	✗	64.78%
HIGO [21]	✓	✓	67.21%
CNN + LSTM [24]	✗	✗	60.98% *
CNN + AUs + LSTM [14]	✗	✗	59.47% *
LMP	✗	✗	**70.20%**
LMP	✗	✓	**68.43%**

In view of the results obtained in Table 2, our method outperforms the other state-of-the-art methods, including handcrafted and deep learning methods (see $*$ lines), in all cases. Looking closely, some authors summarize videos in fewer frames [21]. Indeed, the time lapse between two frames in CASME II is very small as the dataset is recorded with a high-speed camera (at 200 fps). The short time

1. Detailed informations about the data used for the experiments and the code for extracting LMP features are available here : https://gitlab.univ-lille.fr/marius.bilasco/lmp_for_review

lapse combined with the low expression intensity makes the distinction between the noise and the true facial motion very difficult. In [21] a magnification process, which consists of interpolating the frequency, in order to intensify the facial motion is used. These techniques perform well in presence of low intensity motion, but produce severe facial deformations in presence of high intensity motions or head pose variations. Although magnification shows interest in presence of descriptors such as LBP, this technique tends to reduce the performance of optical flow-based approaches. This is mainly due to the fact that the acquisition noise (low lighting change) is also intensified and does not facilitate the measurement of the optical flow. Even-though deep learning methods [14], [24] employ data augmentation, their performances are lower than those of handcrafted methods. The performances obtained on the CASME II dataset show the ability of our method to deal with micro expressions recognition in situations where no illumination changes appear. In the next paragraph, we evaluate our method on micro expressions in presence of various illumination settings.

Experiments on SMIC Table 3 compares the performances of the proposed method with those of major state-of-the-art methods on SMIC dataset under three different acquisition conditions: sequences recorded by high-speed camera at 100 fps (HS), sequences recorded by normal color camera at 25 fps (VIS) and sequences recorded by a near infrared camera both at 25 fps (NIR).

TABLE 3
Performances on SMIC dataset using LOSO *(* data augmentation)* .

Method	Magnifi.	SMIC-HS	SMIC-VIS	SMIC-NIR
LBP-TOP [34]	✗	48.78%	52.11%	38.03%
Deep feat. (CNN) [20]	✗	53.60% *	56.30% *	N/A
Facial Dynamics Map [39]	✗	54.88%	59.15%	57.75%
HIGO [21]	✗	65.24%	76.06%	59.15%
HIGO [21]	✓	68.29%	81.69%	67.61%
LMP	✗	**67.68%**	**86.11%**	**80.56%**
LMP	✓	**67.42%**	**83.12%**	**78.45%**

Our method outperforms the state-of-the-art methods, including handcrafted and deep learning methods, in all cases when no magnification is applied. We obtain comparable performances for the SMIC-HS subset when magnification is applied. Indeed, Li et al. [21] show that artificially amplifying the motion tends to improve the results for micro expression recognition. However, as our aim is to offer a unified micro and macro expression recognition solution, interpolating the video frequency cannot be appropriately generalized on macro expressions. The results obtained on SMIC dataset show good performances for micro expressions recognition with regard to near infrared and natural illumination settings. Our method based on optical flow seems to fit much better near infrared condition compared to other dynamic methods.

5.3 Macro expression

We study the performance of our method to recognize macro expressions on CK+, Oulu-CASIA and MMI datasets dealing respectively with variations in temporal activation sequences, illumination variations and small head movements. We are also considering AFEW dataset containing in-the-wild data in order to study the behaviour of our method in such settings without using any complex pre-processing.

Experiments on CK+ Table 4 compares the performance of the proposed method with major state-of-the-art methods on CK+ dataset. We use the most representative subset of CK+ dataset that contains 327 sequences and 7 expressions to evaluate the performances of our method.

TABLE 4
Performances on CK+ dataset using 10-fold cross validation protocol on 327 sequences *(* data augmentation)* .

Method	Acc(%)
Dis-ExpLet [40]	95.10%
RBM-based model [41]	95.66%
PHRNN-MCSNN [15]	98.50% *
DTAGN (joint) [16]	97.25% *
LMP	**96.94%**
LMP + Geom. feat.	**97.25%**

Compared to handcrafted approaches [40], [41], our method based only on optical flow obtains competitive results (96.94%). Despite the noise contained in the original optical flows, the variation in sequence length and expression activation patterns, the joint analysis of magnitudes and orientations keeps only the pertinent motion.

Inspired by improvements obtained by hybrid approaches, we combine motion features with geometric features by exploiting the shape of facial ROIs for the apex frame. Combination of geometric and LMP features improves slightly the results (97.25%). Results of recent deep learning approaches [15], [16] obtained on CK+ are comparable with the best results that we obtained using a handcrafted approach. Handcrafted approaches consider only the initial data and hence are more sustainable as limited quantity of data are required for training. The performances achieved using only the initial data are well positioned with regard to the augmented settings. This proves the discriminant power of the LMP features.

The facial segmentation model plays an important role in characterizing globally the local facial movement. The segmentation model can be subject to landmark detection errors. In order to quantify the effect of landmark detection and epicenter computation errors, we conduct a series of experiments where landmarks are randomly affected by small to large errors. Three landmarks noise levels were applied. The landmarks location were randomly shifted by $\pm$ 0.5%, $\pm$ 5% and $\pm$ 10% in relation to the size of the face. The results obtained are respectively 96.02%, 95.71% and 94.18%. Although performance tends to decrease as noise becomes more and more important, performance remains relatively stable.

Experiments on Oulu-CASIA Table 5 compares the performance of our method with major state-of-the-art methods on Oulu-CASIA dataset under normal illumination and near infrared settings. The majority of approaches, evaluated on Oulu-CASIA dataset, takes into account only the data under normal illumination conditions (VL). Performances on near infrared (NI) sequences are reported in [36].

Under various illumination settings, our method achieves better results than handcrafted approaches [12], [40], [42] and is competitive with regard to recent deep learning approaches [9], [15], [16]. The performances obtained using LMP in the near infrared domain outper-

TABLE 5
Performances on Oulu-CASIA dataset using 10-fold cross validation protocol on 480 sequences *(* data augmentation)* .

Method	VL-Acc(%)	NI-Acc(%)
LBP-TOP [12]	68.13%	-
AdaLBP [36]	73.54%	72.09%
LBP-TOP + Gabor [42]	74.37%	-
Dis-ExpLet [40]	79.00%	-
DTAGN (joint) [16]	81.46% *	-
PHRNN-MSCNN [15]	86.25% *	-
FN2EN [9]	87.71% *	-
LMP	**75.13%**	**81.88%**
LMP + Geom.feat.	**84.58%**	**81.49%**

form those of [36] (81.88%). According to the results, the combination of motion and geometric features clearly improves the performances (84.58%) in the VL setting and our method obtains competitive performances. Under NI settings, LMP features perform the best due to robustness to poor landmarks detection which impacts negatively the solution combining LMP and geometry features.

Experiments on MMI Table 6 compares the performance of recent state-of-the-art methods on MMI dataset. We have selected only the activation sequence (e.g. neutral to apex) for 205 sequences. The combination of motion and geometric features improves the performances (78.26%) and out stands other handcrafted approaches [12], [40], [42], [43]. Compared to deep learning approaches our approach performs better than [10], [16] and obtains competitive results with [12], [40], [43]. Compared to deep learning approaches our approach performs better than [10], [16] and obtains competitive results with [15].

TABLE 6
Performances on MMI dataset using 10-fold cross-validation protocol *(* data augmentation)* .

Method	Acc(%)
LBP-TOP [12]	59.51%
LBP-TOP + Gabor [42]	71.92%
CSPL [43]	73.53%
Dis-ExpLet [40]	77.60%
DTAGN (joint) [16]	70.24% *
PHRNN-MSCNN [15]	81.18% *
LMP	**74.40%**
LMP + Geom. feat.	**78.26%**

Experiments on AFEW Table 7 compares the performances of recent state-of-the-art methods on VReco subchallenge of the AFEW dataset. To deal with head pose variations, we use the same affine registration proposed in the baseline [44]. In view of the performance obtained by all the approaches, it can be seen that the existing solutions are not very robust on data acquired in-the-wild. Although LMPs do not give the best performance compared to deep learning approaches, they are better than LBP-TOP used in the baseline. In this context, deep learning based approaches tend to give better performances because they have the ability to fit better to data specificities (head pose variations, illumination changes, important head movements.)

TABLE 7
Performances on AFEW dataset *(* data augmentation)* .

Method	Acc(%)
LBP-TOP [44]	41.07%
LSTM [45]	58.81% *
CNN-RNN [46]	59.02% *
LMP	**49.16%**

In the next section, we synthesize the results and we highlight the capacity of the proposed facial expression recognition framework and underlying LMP feature to deal in a unified manner with the various challenges brought by micro and macro expressions.

5.4 Micro and macro expression evaluation synthesis

Table 8 summarizes the most relevant results with representative state-of-the-art methods on micro and macro expressions using unaltered versions of the datasets and the same evaluation protocols (10-fold cross validation for macro-expression and LOSO for micro-expression).

Results show that the proposed method has the singularity of dealing in a unified manner with both micro and macro expressions challenges. The method outperforms micro expression state-of-the-art methods. Overall, on average, our method performs 4.94% better than the best handcrafted approaches for each dataset and 14.18% better than learning-based approaches. Furthermore, we obtain very competitive results for macro expression recognition, whether it is under varying illumination condition or in presence of small head pose variations. Our method performs on average 5.17% better than the best handcrafted approaches for each dataset. Learning-based approaches using data augmentation perform on average 2.19% better than our approach when used in situation where no face normalization is required, as it is the case for CK+, Oulu-Casia and MMI datasets. For the specific case of AFEW, where important head pose variations occurs, the performances of our method are relatively low with regard to learning methods, but still high with regard to handcrafted methods. We did not added any complex pre-processing stages required by datasets as AFEW in order to keep a uniform framework for all cases.

Although recent deep learning methods achieve better performances for macro expression recognition, it is important to emphasize the relevance of a unified method that can characterize efficiently both micro and macro expressions. Proposing an unified approach capable to deal with very large intensity variations leads up the way to the coverage of the full range of facial expression intensities.

The parameters used to assess LMP performances on each dataset are given in Table 9. LMP settings vary slightly depending on the dataset, underlining the generalization capacity of the unified approach to deal with macro and micro expression specificities. Most of the time, the variations are due to the acquisition conditions (distance to the camera, resolution, frame rates).

Results obtained for micro and macro expression prove the efficiency and the robustness of our contribution, which stands as a good candidate for challenging contexts (e.g. variations in head movements, illumination, activation patterns and intensities).

6 Conclusion

The main contributions of our paper are articulated around three axes. The first one is an innovative Local Motion Patterns (LMP) feature that measures temporal physical phenomena related to skin elasticity of facial expression. The second one is a unified recognition approach of both

TABLE 8
Performance synthesis on all datasets *(* data augmentation)* .

	Method	Micro expression				Macro expression				
		CASME II	SMIC			CK+	CASIA		MMI	AFEW
			HS	VIS	NIR	7 classes	VL	NI		
Handcrafted	LBP-TOP [12]	-	-	-	-	- %	68.13%	-	59.51 %	-
	LBP-TOP + Gabor [42]	-	-	-	-	-	74.37%	-	71.92%	-
	AdaLBP [36]	-	-	-	-	-	73.54%	72.09%	-	-
	Dis-ExpLet [40]	-	-	-	-	95.10%	79.00%	-	77.60%	-
	HIGO + magnification [21]	67.21%	68.29%	81.69%	67.61%	-	-	-	-	-
	* LBP-TOP [44]	-	-	-	-	-	-	-	-	41.07%
	LMP	**70.20%**	**67.68%**	**86.11%**	**80.56%**	**97.25%**	**84.58%**	**81.46%**	**78.26%**	**49.16%**
Deep Learning	* CNN + LSTM [24]	60.98%	-	-	-	-	-	-	-	-
	* Deep feat. (CNN) [20]	47.30%	53.60%	56.30%	-	-	-	-	-	-
	* CNN + AUs + LSTM [14]	59.47%	-	-	-	-	-	-	-	-
	* PHRNN-MSCNN [15]	-	-	-	-	98.50%	86.25%	-	81.18%	-
	* FN2EN [9]	-	-	-	-	-	87.71%	-	-	-
	* CNN - RNN [46]	-	-	-	-	-	-	-	-	59.02%

TABLE 9
Parameter settings used for assessing the best results.

Datasets		λ	Δ	ρ	E	M	V	β	bin
Micro	CASME II	4	0.5	0.75	100	4	5	6	9
	SMIC-HS	3	0.5	0.75	100	3	5	6	9
	SMIC-VIS	5	0.5	0.75	100	4	5	3	9
	SMIC-NIR	4	0.5	0.75	100	3	5	3	12
Macro	CK+	3	0.5	1	100	4	5	3	12
	CASIA-VL	4	0.5	1	100	5	5	3	6
	CASIA-NI	5	0.5	0.75	100	5	5	6	9
	MMI	3	0.5	1	100	4	5	6	12
	AFEW	3	0.5	1	100	4	5	3	12

macro and micro expressions. The spatio-temporal features, extracted from videos, encode motion propagation into local motion regions situated near expression epicenters. As motion is inherent to any facial expressions our method is naturally suitable to deal with all expressions that cause facial skin deformation. The third one is related to the exponential potentiality and suitability of our method to meet in-the-wild requirements. We obtain good performances in various illumination (near infrared and natural) conditions for both micro and macro expression recognition.

The method outperforms micro expression state-of-the-art methods on CASME II (70.20%) and SMIC-VIS (86.11%). Furthermore, we obtain competitive results for macro expression recognition (97.25% for CK+, 84.58% for Oulu-CASIA and 78.26% for MMI). However, on data acquired under natural conditions, as in AFEW dataset, further efforts are still needed (49.19%). The important global head motions overcome the local motion characterizing facial expressions. Specific pre-processing steps as those illustrated in [47] are required in order to address challenges brought by large head movements.

Although our contribution unifies the micro and macro expression domains, other challenges such as dynamic background, occlusion, non-frontal poses, important head movements are still to be addressed. For example, let us consider the challenge of expression recognition in presence of important head movements. Although dynamic texture approaches perform well when analyzing facial expression in near frontal view, recognition of dynamic textures in presence of head movements remains a challenging problem. Indeed, dynamic textures must be well segmented in space and time. However, we believe that the registration based on facial components or shape are not adapted to dynamic approaches. Such registrations cause facial deformations and induce noisy motion [47]. We believe that suitable relationship between motion representation and registration is the key for expression recognition in presence of head movements.

REFERENCES

[1] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[2] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, 2014.

[3] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, 2005, pp. 5–pp.

[4] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[5] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, 2014.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[7] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Human vision inspired framework for facial expressions recognition," in *ICIP*, 2012, pp. 2593–2596.

[8] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[9] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *FG*. IEEE, 2017, pp. 118–126.

[10] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *WACV*. IEEE, 2016, pp. 1–10.

[11] J. N. Bassili, "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*, vol. 37, no. 11, pp. 2049–58, 1979.

[12] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[13] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: a survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015.

[14] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," in *CVPR Honolulu - June 21-26*, 2017.

[15] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015, pp. 2983–2991.

[17] I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition*, vol. 41, no. 3, pp. 833–851, 2008.
[18] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *WACV*, 2016, pp. 1–8.
[19] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.
[20] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *ICPR*. IEEE, 2016, pp. 2258–2263.
[21] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Reading hidden emotions: spontaneous micro-expression spotting and recognition," in *CVPR*, 2015, pp. 217–230.
[22] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *ACCV*, 2014, pp. 525–537.
[23] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikainen, "Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection," *CVPR*, 2016.
[24] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *Transactions on Affective Computing - issue 99*, 2017.
[25] X. Fan and T. Tjahjadi, "A dynamic framework based on local zernike moment and motion history image for facial expression recognition," *Pattern Recognition*, vol. 64, pp. 399–406, 2017.
[26] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *Affective Computing*, vol. 6, no. 1, pp. 1–12, 2015.
[27] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *ICPR*, 2014, pp. 1776–1781.
[28] H. Sadeghi, A.-A. Raie, and M.-R. Mohammadi, "Facial expression recognition using geometric normalization and appearance representation," in *MVIP*. IEEE, 2013, pp. 159–163.
[29] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
[30] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.
[31] B. Allaert, I. M. Bilasco, and C. Djeraba, "Consistent optical flow maps for full and micro facial expression recognition." in *VISAPP*, 2017, pp. 235–242.
[32] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Epicflow: Edge-preserving interpolation of correspondences for optical flow," in *CVPR*, 2015, pp. 1164–1172.
[33] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *SCIA*. Springer, 2003, pp. 363–370.
[34] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *FG*. IEEE, 2013.
[35] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPRW*. IEEE, 2010, pp. 94–101.
[36] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.
[37] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, July 2012.
[38] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM TIST*, vol. 2, no. 3, p. 27, 2011.
[39] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.
[40] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets via universal manifold model for dynamic facial expression recognition," *Transactions on Image Processing*, vol. 25, no. 12, pp. 5920–5932, 2016.
[41] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal rbm-based model for facial expression recognition," *Pattern Recognition*, vol. 49, pp. 152–161, 2016.
[42] L. Zhao, Z. Wang, and G. Zhang, "Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram," *Mathematical Problems in Engineering*, 2017.
[43] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *CVPR*. IEEE, 2012, pp. 2562–2569.
[44] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM international conference on multimodal interaction*. ACM, 2017, pp. 524–528.
[45] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 569–576.
[46] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.
[47] B. Allaert, J. Mennesson, I. M. Bilasco, and C. Djeraba, "Impact of the face registration techniques on facial expressions recognition," *Signal Processing: Image Communication*, vol. 61, pp. 44–53, 2018.

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.

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.