LTP-ML: Micro-Expression Detection by Recognition of Local Temporal Pattern of Facial Movements

Conference Paper · May 2018

DOI: 10.1109/FG.2018.00100

Jingting Li
Chinese Academy of Sciences
23 PUBLICATIONS **183** CITATIONS

Catherine Soladié
École Supérieure d'Electricité
47 PUBLICATIONS **187** CITATIONS

Renaud Seguier
École Supérieure d'Electricité
127 PUBLICATIONS **866** CITATIONS

LTP-ML : Micro-Expression Detection by Recognition of Local temporal Pattern of Facial Movements

Jingting LI, Catherine SOLADIE, Renaud SEGUIER
FAST Research Team
CENTRALESUPELEC/IETR
Rennes, France
Email: {jingting.li, catherine.soladie, renaud.seguier}@centralesupelec.fr

***Abstract*—The Micro-expressions (MEs) carry specific non-verbal information, for example the facial movement caused by pain. However, as a consequence of their local and short nature, it is difficult to detect MEs. This paper presents a novel detection method by recognizing a local and temporal pattern (LTP) of facial movement. In our system, with the purpose of improving the detection accuracy, temporal local features are generated from the video in a sliding window of 300ms (mean duration of a ME). These features are extracted from a projection in PCA space and form a specific pattern during ME which is the same for all MEs. Using a classical classification algorithm (SVM), MEs are then distinguished from other facial movements. Finally, a global fusion analysis is applied on the whole face to eliminate false positives. Experiments are performed on two databases: CASME I and CASME II. The detection results show that the proposed method outperforms the most popular detection method in terms of F1-score according to the analysis of multiple metrics.**

***Keywords*-Micro-expression; Detection; Local temporal pattern; Machine learning**

I. Introduction

The facial expression is one of the most important external indicators to reveal the emotion and the psychological status of a person [1]. Among the facial expressions, the micro-expressions (MEs) [2] are local and brief expressions which appear involuntarily, particularly in the case of strong pressure. The duration varies from 1/25 to 1/5 of a second [2]. The involuntary nature makes it possible to affirm that it represents the real emotions of a person [3]. The detection of micro-expressions has many applications particularly in the field of national security [3], medical care [4], and studies on political psychology [5] and educational psychology [6].

To code the micro-expressions, the Facial Action Coding System (FACS) [7] is widely used. FACS was created to analyze the relationship between facial muscle deformation and emotional expression. The action units (AUs) are the facial components of the FACS, which represent local muscle movement. The AU index identifies the region(s) where the ME occurs. Thus, the FACS system can help annotate the appearance and dynamics of an ME in a video.

Since the 2000s, research on the automatic detection and recognition of micro-expression (MEDR) has developed.

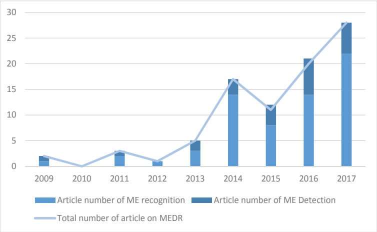


Figure 1. MEDR research trend. The curve shows that the number of articles on MEDR is increasing by year, mainly in the area of ME recognition (bottom column), ME detection research has not yet attracted sufficient attention (column at the top).

Figure 1 shows the trend of number of the MEDR research articles. The number remains low and the results are not yet very satisfactory because of the ME nature and also the limited number of public ME database. However, there have been more and more emerging studies in recent years. Compared with ME detection, there are already numerous studies on ME recognition. In addition, the recognition rate is getting higher, e.g. in Davison [8], the highest recognition accuracy reached 76.60%. However, most of ME recognitions are performed between the onset and offset frame. In the meantime, due to their short duration and low intensity, micro-expressions are very difficult to detect. The detection results of current proposed methods are not precise enough, the detection rate is around 70% [9]. Even through the ME samples are produced in a strictly controlled environment, there are many false positives due to head movement or eye blinking. What's more, the metrics used to analyze the result in different papers are divers. The accuracies are studied per frame, per interval or per video, while the metric could be TPR, ROC, ACC and other measures. It is difficult to define one metric rather than another. This paper explores an automatic system for detecting MEs which could:

- separate motions related to MEs from head movement or eye blinking.
- detect the region where the ME occurs.

The principal contribution of this article is to propose a novel ME detection method using a local temporal pattern (LTP) extracted from a projection in PCA space. The origi-

978-1-5386-2335-0/18/$31.00 ©2018 European Union

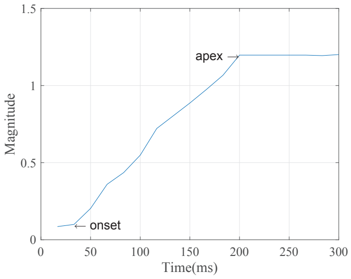


Figure 2. Example of local temporal pattern during an ME in the right eyebrow region over a period of 300ms. (Video: Sub01_EP01_5 of CASMEI)

nality of the approach is the utilization of a temporal pattern, corresponding to the onset and offset of MEs. Specifically, the pattern is same for all kinds of MEs. Therefore, the detection is not limited to certain ME. This temporal pattern is extracted from videos with or without MEs. Meanwhile, the pattern corresponds to a length interval of the average duration of an ME. The extraction of temporal characteristics over a long interval is performed by calculating the distance between the first frame and the n^{th} frame of the interval. In order to conserve the most significant variation, principal component analysis (PCA) is first performed. The shape of the curve in figure 2 represents the temporal variation from the onset to the apex of the ME.

The second contribution concerns locating the region where ME occurs. The particularity of the approach is the combination of local and global treatments. The detection of temporal patterns is done locally by ROI (Region of Interest), a fusion system on the entire face then separates MEs from other facial movements.

Finally, depending on the construction, the time position of the onset of the ME can be determined, i.e. the indexes of the frames where the patterns are detected.

The article is organized as follows: section II introduces the related work on the detection of MEs; section III describes our method by employing the temporal pattern for ME detection; section IV presents the experimental results; section V concludes the paper.

II. Related works

Since the micro-expression is an involuntary facial expression, the majority of ME detection research is developed based on spontaneous public databases. This section presents the related research works by introducing their principal method and utilized features. In addition, the advantages and drawbacks of these method are discussed.

The main idea of most methods for ME detection is to calculate the difference between their own features, which means the differences between the first frame and the other frames in a time window. Meanwhile, the utilized features are diverse, including LBP, HOG, optical flow, integral projection.

Moilanen et al. [10], [9] used linear binary pattern (LBP) feature difference analysis to spot ME. MEs were then extracted by thresholding and peak detection. Yan et al. [11] quantified ME and spotted the apex frames by three feature extraction algorithms: Constraint Local Model (CLM), LBP and optical flow. Liong et al. [12] developed the method and spotted the apex frame by employing a binary search strategy. Patel et al. [13] utilized optical flow and then a spatiotemporal integration to spot apex frame and identify the location of onset and offset. Davison et al. [14] applied 3D-HOG as the feature distance measure to calculate the dissimilarity between frames and detected the ME using an individualized baseline. Liong et al. [15] spotted apex frame by employing optical strain which is more effective in identifying the subtle deformable facial muscle. Li et al. [16] integrated a deep multi-task learning method for the facial landmarks location to help detect the ME with HOOF (histograms of oriented optical flow). Wang et al. [17] proposed the main directional maximal difference analysis of optical flow to spot the ME. Lu et al. [18] presented a method with a low computation cost based on differences in the Integral Projection (IP) of sequential frames for ME detection.The main advantage of these approaches is to be able to make comparisons between frames over a time window of the size of an ME. However, they detect the movement between frames, and not specifically the ME movement. This is why the ability to distinguish MEs from other movements (such as blinking or head movements) remains weak.

Nowadays, methods utilizing machine learning are emerging. Xia et al. [19] utilized Adaboost model to estimate the initial probability for each frame and then a random walk model to spot the ME by considering the correlation between frames. Hong et al. [20] proposed a multi-scale sliding window based approach. LBP-TOP, HOG-TOP and HIGO-TOP were extracted as feature and MEs were detected by binary classification. Borza et al. [21] used the movement magnitude across frames by simple absolute image differences, then Adaboost algorithm was applied to detect ME frames. The machine learning process makes it possible to avoid detecting certain movements when there is no ME in the video. However, the classifier focuses primarily on detecting a spatial pattern. Even though, features like LBP-TOP extracts temporal characteristics in a small temporal window, the duration is too short to represent a global temporal movement pattern for ME. The temporal pattern variation in a ME duration has yet to attract sufficient attention. In addition, all of the above methods of machine learning do not explicitly use the fact that the ME is a local facial movement, and the extracted features from local region are integrated into a feature which represents the movement for entire face.

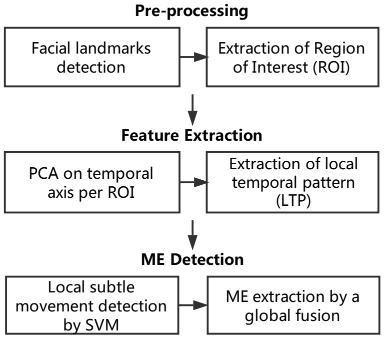

Pre-processing
Facial landmarks detection
Extraction of Region of Interest (ROI)
Feature Extraction
PCA on temporal axis per ROI
Extraction of local temporal pattern (LTP)
ME Detection
Local subtle movement detection by SVM
ME extraction by a global fusion

Figure 3. Overview of our LTP-ML method

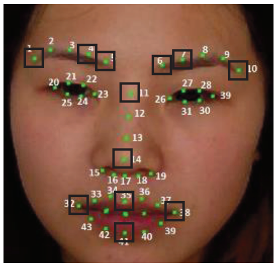


Figure 4. Facial landmarks and ROIs distribution. 49 landmarks are detected and ROIs are generated depending on the position of chosen landmarks. 12 ROIs are generated in the region of eyebrows, nose and mouth contour.(©Xiaolan Fu)

Table I
CHOSEN ROIS AND RELATED AU INDEX

Facial region	ROI index	Related AU
Eyebrows	1, 4, 5, 6, 7, 10	1, 2, 4
Nose	11, 14	9
Mouth	32, 35, 38, 41	10, 12, 14, 15, 17, 25

Therefore, our method is proposed to detect ME using directly a temporal pattern extracted from local region. Several frames are taken into account to obtain a real temporal and local pattern (LTP), and then analyzed by a classifier. Even though the spatial pattern is not studied, the detected facial motions are distinguished by a fusion analysis from local to global. This method helps to improve the ability to distinguish ME from other movements. In addition, it is capable of finding the ME spatial location and the temporal index of the onset of ME.

III. OUR METHOD - LOCAL TEMPORAL PATTERN - MACHINE LEARNING

The proposed method consists of three parts: a pre-processing to precisely detect facial landmarks and extract the region of interets (ROIs), then the extraction of local temporal pattern (LTP) on these ROIs and eventually the detection of MEs. Since the LTPs are classified by machine learning, our method will be presented as LTP-ML in brief. Figure 3 displays the overall process.

A. *Pre-processing: local ROI detection*

As the ME is a local facial movement, a pre-processing is performed on the face to determine local ROIs. The process contains two stages. The first step is detecting 49 landmarks (LMs) on human face for each image using the tool 'Intraface' [22]. The second step is to extract ROIs where the MEs could possibly occur. These ROIs are generated in the form of a square around the chosen landmarks. The length of the side of the square a is determined by the distance L between the LM #23 and #26, i.e. the distance between the left and right inner corners of eyes: $a = (1/5) \times L$. The small length avoids the overlap of adjacent ROIs. Figure 4 illustrates the result of preprocessing on an image.

The ROIs include the regions of the two eyebrows (ROI 1,4,5,6,7,10) and the contour of the mouth (ROI 32,35,38,41). These ROIs contain most evident muscle movement of ME compared with ROI #2, 3, 8, 9 of eyebrow and other ROIs of mouth region. The eye area is neglected due to blinking. Because of the rigidity of the nose, two samples on this region are chosen as a reference to eliminate overall movements of the head (ROI 11,14). Table I gives the link between AU and ROI location. Our selected ROIs connects to the AUs where occur ME most frequently.

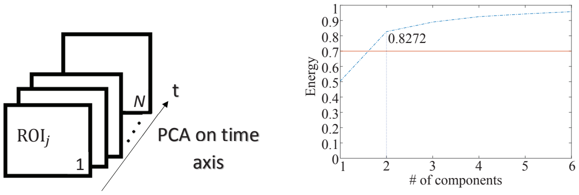
B. *Feature extraction: Local Temporal Pattern extraction*

The objective of this part is to extract LTPs allowing MEs to be distinguished from other facial movements. Therefore, the main texture deformations are extracted at gray level over time for each ROI. First, local sequences that only include one ROI are conducted. PCA [23] is then performed on each ROI sequence to conserve the principal variation at this region. Figure 5(a) illustrates this processing on one of the ROI sequences.

Let N be the number of frames in a video, and a^2 be the size of the ROI, the size of matrix I processed by PCA is $a^2 \times N$. We note $\bar{I} \in M_{a^2,N}(\mathbb{R})$ the mean value matrix of each pixel in chosen ROI and $\Phi \in M_{2,N}(\mathbb{R})$ the projection matrix in the PCA space reduced to the first 2 dimensions. As the PCA energy analysis example shown in Figure 5(b), these first two components can normally conserve more than 70% energy. The matrix P in size of $2\times N$, i.e. the projection of I in the PCA space with first two dimensions, is obtained by following formula:

$$\begin{bmatrix} P_1(x) & \cdots & P_N(x) \\ P_1(y) & \cdots & P_N(y) \end{bmatrix} = \Phi\times(\begin{bmatrix} I_1(1) & \cdots & I_N(1) \\ & \ddots & \\ I_1(a^2) & \cdots & I_N(a^2) \end{bmatrix} - \bar{I})$$

Each point P_n represents the most significant regional motion for one frame, facial changes can be analyzed on the time axis. A relation between the distance of the points and the movement magnitude can be found: while the distance gets larger, the magnitude increases.



(a) (b)

Figure 5. PCA process analysis. Figure 5(a) shows PCA per ROI among all the video. A local sequence video of ROI$_j$ with N frames is processed by the PCA on the time axis. Figure 5(b) shows a PCA energy analysis example. The first 2 components can conserve more than 70% energy. (Sub01_EP01_5_ROI1 of CASMEI)

Table II
DISTANCE DATASETS PER FRAME FOR ONE ROI VIDEO SEQUENCE

Frame index	Original distances
1st frame	$\Delta_j(1,2), \cdots, \Delta_j(1,K+1)$
$\cdots$	$\cdots$
n th frame	$\Delta_j(n,n+1), \cdots, \Delta_j(n,n+K)$
$\cdots$	$\cdots$
N th frame	$0, \cdots, 0$

The variation of the distances is then studied on the sliding windows of each ROI. The duration taken is 300ms to correspond to the average duration of a ME. The distance between the first frame and the other frames in the interval is calculated. Suppose there are $K+1$ frames in the interval, the set of distances in this interval is:

$$\{\Delta_j(n,n+1), \cdots, \Delta_j(n,n+i) \cdots, \Delta_j(n,n+K)\}$$

Where $i \in [1,K], j$ is the ROI index, n is the index of the first frame in the interval, $\Delta_j(n,n+i)$ represents the euclidean distance between the point P_{n+i} of the $(i+1)^{th}$ frame in interval and the point P_n of the first frame n in interval.

Therefore, each frame has a dataset which consists of K distances as listed in Table II.

As the movement magnitudes are not same in different videos, these above distance sets need to be normalized. In average, the curve reaches the top in 150ms ($K/2$) for the ME. Thus, the normalization is applied depending on the maximum distance in this period for each ROI:

$$\Delta_{j_{max}} = max_{n=1\cdots N,\ i=1\cdots K/2}(\Delta_j(n,n+i))$$

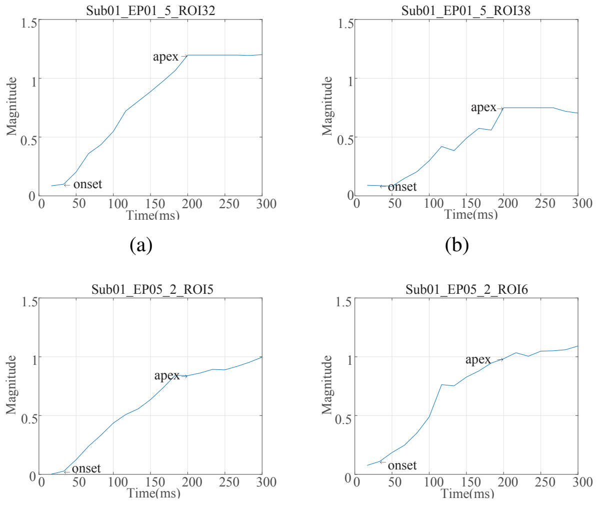
Then, the coefficient of normalization (CN) is computed : $CN_j = 1/\Delta_{j_{max}}$ and the normalized distance is:

$$d_j(n,i) = \frac{\Delta_j(n,n+i)}{\Delta_{j_{max}}} = \Delta_j(n,n+i) \times CN_j$$

Finally, CN is added to into the feature in order to eliminate the movements which are too subtle. Suppose there are J chosen ROIs, the features are connected in parallel. Table III lists the constructed features of frame n, which is a matrix of $J \times (K+1)$.

Table III
FEATURE CONSTRUCTION OF ONE FRAME WITH ALL CHOSEN ROIS

ROI 1	CN_1	$d_1(n,n+1), \cdots, d_1(n,n+K)$
$\cdots$	$\cdots$	$\cdots$
ROI j	CN_j	$d_j(n,n+1), \cdots, d_j(n,n+K)$
$\cdots$	$\cdots$	$\cdots$
ROI J	CN_J	$d_J(n,n+1), \cdots, d_J(n,n+K)$



(a) (b) (c) (d)

Figure 6. Temporal patterns of two different videos. Figures 6(a) and 6(b) illustrate the ME movements at ROI 32 and ROI 38 (the right and left corners of the mouth) in the video Sub01_EP01_5 of CASME I. Emotion of this video is labeled as joy, which is often expressed by mouth corner displacement. The curves in Figures 6(c) and 6(d) present the movements of the stress emotion at ROI 5 and ROI 6 (the inside corners of the right and left eyebrows) in the video Sub01_EP01_5 of CASME I.The pattern of curves in these four images are similar even through the ROIs and videos are different.

C. ME detection

In this subsection, the ME detection is processed by two steps : a local pattern classification and a global merge.

1) Local classification: Figure 6 shows that the LTPs are identical regardless of the ROI and the ME type. In fact, because the first two components of PCA retain the main variations, the pattern is only influenced by the movement in this local region. Thanks to the identical nature of the LTP for ME, the local movements can be classified by machine learning.

A supervised classification SVM is employed. To label the database, the $K/3$ frames before ME onset are found to retain the best pattern, where $K+1$ is the length of interval as mentioned in the above section and $K/3$ is an empirical value. Hence, the frames are annotated as shown in Figure 7.

The ROI index for training is then selected based on the AU info for each video, as illustrated in Table IV. Although, in the stage of recognition, the features are constructed by entire chosen ROIs from all the videos. The results of local detection are generated by LOSubOCV (Leave-One-Subject-

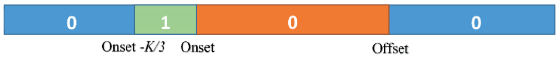


Figure 7. Frame label annotation diagram. The blue rectangle represents an entire video, and the orange part means the ME sequence. Frames in the range [onset- $K/3$, onset] are classified in label 1 and the rest frames are labeled in 0.

Table IV
ROI SELECTION FOR TRAINING STAGE OF MACHINE LEARNING

AU condition	ROI index	Facial region
All given AU index lower than 6	1,4,5,6,7,10	Eyebrows
All given AU index larger than 9	32,35,38,41	Mouth contour
Otherwise	all chosen ROIs	Entire face

Out Cross Validation).

2) Global fusion: In this part, the false positives concerning other movements and true negatives caused by our recognition process are reduced by these three steps: a local qualification, a spatial fusion and a merge process.

First of all, two thresholds T_{CN} and T_{dist} are set to eliminate the movements which are too subtle and may be generated by noise, even through they have LTP pattern. Furthermore, the frame number with label to 1 is limited locally in an interval of length K. In fact, the duration of ME is normally around K frames, and the optimal condition is to detect all $K/3$ frames before the ME onset. Having less than $K/9$ or more than $K/2$ patterns detected does not correspond to an ME.

Secondly, in the part of spatial fusion, all ROIs which have detected movement in one frame are integrated by certain strategies to obtain a global detection result for entire face. To reduce the number of false positive, the unrelated motions are eliminated by the following rule. If there are more than $J/2$ ROIs of entier face or more than one nose region that have been detected with some movement, this motion is then considered as a global head movement. What's more, the eye blinking leads to all the muscles around eye. Thus, if all the ROIs of eyebrows detect movement, the ME detection system supposes there is an eye blinking, and treat the current frame as non-ME.

Thirdly, as the results are given per frame, the detected ME frames distribute discretely, which brings many true negatives, the nearby zones where have detected LTP are merged.

IV. Experiments and Results

In this section, the performance of LTP-ML is evaluated by the comparison with the LBP-Chi-square-distance method using two public databases. The results are analyzed by video and by frames. In order to analyze the performance of the global fusion, the contributions of each step are illustrated. In addition, we compute the accuracy of detection by emotion.

A. Method for comparison

The LBP-Chi-square-distance method (LBP-χ^2) was first proposed by Moilanen et al. in 2014 [10]. We use this method to compare with, because it is the one that is most commonly used in other articles to compare with their own ME detection results. However, the majority methods [18], [9] evaluate their results using ROC and AUC metrics, in our method, these metrics are not suitable since there is no valuable parameter to adjust. Moreover, the results presented in [10] consider that the eye blinking is a true positive, which is not the case of the ground truth in the databases. As a result, we re-implement the method from the article and succeed to achieve the same level of detection rate.

B. Database and experiment configuration

1) Database: Experiments are performed on two spontaneous ME databases: CASME I [24] and CASME II [11]. The MEs in these two databases are labeled with reliable ground truth, including the temporal location of the onset, apex and offset of the ME. CASME I has two sections : section A and section B because of two lighting conditions and also two different resolutions. The essential parameters of these two databases are listed in Table V. All ME sequences in CASME I and CASME II are used in the experiment.

2) Experiment configuration: The configuration of LBP-χ^2 method is based on the descriptions of the three articles: [10], [25] and [20]. The face is divided into 36 blocks with an overlap. The overlap rates in the direction of X and Y are 0.2 and 0.3 respectively. A uniform mapping is applied for the LBP feature extraction from the block, the radius r is set to $r = 3$, and the number of neighboring points p is set to $p = 8$. The χ^2 distances of the current frame are computed in an $2 \times L + 1$ interval. L value for two databases are listed in Table V. The ground truth of LBP-χ^2 is put in the range of [onset$-L/2$, offset$+L/2$].

For our LTP-ML method, the size of the time interval K corresponds to 300ms according to the fps of each database as shown in Table V, which is the average duration of a ME. Training and recognition are performed using the software Lib-SVM with linear kernel [26]. Since the dataset is very unbalanced, these non-ME frames are sampled by 1 out of 8 for SVM training stage. The parameter of cost c for SVM training is set to 5 and the weight w for each class are set to 1 and 2.5 respectively. All frames are considered in the testing stage. The results are obtained by LOSubOCV.

Since LTP-ML detects the special pattern of the onset of local facial movement, the optimal condition is to detect patterns in the interval of [onset$-K/3$, onset] and in the meantime in [apex$-K/3$, apex]. Thus, the ground truth of LTP-ML is defined by adding a $K/3$ shift to that of LBP-χ^2, i.e. [onset$-K/3 - L/2$, offset$-K/3 + L/2$].

Table V
Principal parameters and experiment configuration for CASME I and CASME II

Database	Subject	ME sequence	FPS	L	K
CASME I-A	7	96	60	21	18
CASME I-B	12	101	60	21	18
CASME II	26	255	200	65	60

Table VI
ME detection results per video on CASME I and CASME II by LBP-χ^2 and LTP-ML. The results are evaluated in term of TP, FP, TPR, Precision and F1-score, in which the best results are highlighted in bold.

Database	Method	TP	FP	TPR	Precision	F1-score
CASME I-A	LBP-χ^2	53	**91**	0.55	0.37	44.16%
	LTP-ML	**80**	111	**0.83**	**0.42**	**55.75%**
CASME I-B	LBP-χ^2	76	106	0.75	0.42	53.71%
	LTP-ML	**77**	**103**	**0.76**	**0.43**	**54.80%**
CASME II	LBP-χ^2	221	**134**	0.87	**0.62**	72.46%
	LTP-ML	**229**	148	**0.90**	0.61	**72.47%**

Table VII
ME detection confusion matrix per frame on CASME I and CASME II. The best results are highlighted in bold.

Database		LBP-χ^2		LTP-ML	
		ME	Non-ME	ME	Non-ME
CASME I-A	ME	0.05	0.95	**0.23**	0.77
	Non-ME	0.02	**0.98**	0.08	0.92
CASME I-B	ME	0.07	0.93	**0.22**	0.78
	Non-ME	0.02	**0.98**	0.05	0.95
CASME II	ME	0.09	0.91	**0.24**	0.76
	Non-ME	0.02	**0.98**	0.09	0.91

Table VIII
Detection result per frame on CASME I and CASME II by LBP-χ^2 and LTP-ML. The results are evaluated in term of ACC, Precision and F1-score, in which the best results are highlighted in bold.

Database	Method	ACC	Precision	F1-score
CASME I-A	LBP-χ^2	**78.75%**	38.35%	8.78%
	LTP-ML	77.90%	**43.24%**	**29.87%**
CASME I-B	LBP-χ^2	**82.92%**	**46.34%**	12.05%
	LTP-ML	82.61%	45.67%	**29.64%**
CASME II	LBP-χ^2	64.08%	**73.67%**	15.66%
	LTP-ML	**65.07%**	60.59%	**34.96%**

C. Result comparison with LBP-χ^2

Since the original LBP-χ^2 process do not perform merge after the peak detection, the results obtained by our LTP-ML method is firstly compared with LBP-χ^2 without merge process. To evaluate the performance, the detection result is measured per video and per frame respectively.

1) Result comparison per video: It is necessary to determine whether the test video contains the ME clips. Hence, the detection result per video is analyzed. In these two databases, each video has one ME clip, and the false positive detected peaks are determined by a non-overlap search window (600ms). Table VI illustrates the result of LBP-χ^2 and LTP-ML methods.

CASME I-A has 96 videos, and our method has detected successfully 80 ME sequences. Even though the number of false positive is little higher than LBP-χ^2, the accuracy measurements are higher than LBP-χ^2 in TPR(Recall), precision and F1-score, which indicates the LTP-ML performs better than LBP-χ^2. In the meantime, according to the chosen measurement, the ME detection on CASME I-B and on CASME II performs slightly better than that of LBP-χ^2.

The proportions of ME frames are 0.19 for CASME I and 0.38 for CASME II. Moreover, the facial resolution of CASME I-A is higher than that of CASME I-B, the ROI contains more pixels but also brings more noise. Depending on the above comparison, our LTP-ML method performs as well as LBP-χ^2, and it performs better in the case of more facial movements and more noise in CASME I-A.

2) Result comparison per frame: The LTP features are extracted from each frame. In addition, local classification and the first two steps of global fusion are performed on frame. The accuracy measurement per frame can demonstrate the ME-NonME recognition ability of our method. The confusion matrix in our article is constructed as [TPR, FNR; FPR, TNR], and the matrix for the two detection methods are shown in Table VII. Due to the absence of merging the detected frames into temporal interval, the TPR is not very high. Compared to the result of LBP-χ^2, our method outperforms LBP-χ^2 for ME detection because the TPR of our method is higher than 20% while the percentage of TPR of LBP-χ^2 is lower than 10%. In the meantime, since there are some facial motions have similar pattern as LTP for ME, the non-ME detection result of our method is slightly lower. Although, the FPR is maintained relatively low, 0.08 for the CASME I-A, 0.05 for CASME I-B and 0.09 for CASME II.

In order to confirm the effectiveness of our method, three metrics i.e. ACC, precision and F1-score are chosen because they are most commonly used for the evaluation of machine learning method. The results are listed in Table VIII. First of all, the average ACC of our method can be maintained at 75%. However, the obtained results make it difficult to compare with LBP-χ^2 in term of ACC and precision. In addition, our method outperforms LBP-χ^2 in term of F1-score for each database. As a result, our LTP-ML method can detect more ME frames while maintaining an acceptable FRP and a better classification performance. Furthermore, to eliminate the false negative, the frames detected by LTP-ML are then merged. Table IX shows the final LTP-ML detection result with merge process. Compared with the results without merge, the TPR and F1-score value have increased after the merge process.

D. Distribution analysis of each step in global fusion

As described in section III, the global fusion is divide into three steps: local qualification, spatial fusion and merge process. The accuracy measurement for each step should be

Table IX

LTP-ML DETECTION RESULT WITH MERGE PROCESS ON CASME I AND CASME II. THE RESULTS ARE EVALUATED IN TERM OF TPR, FPR, ACC, PRECISION AND F1-SCORE, IN WHICH THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Database	TPR	FPR	ACC	Precision	F1-score
CASME I-A	**0.37**	0.13	76.35%	41.70%	**39.14%**
CASME I-B	**0.34**	0.09	80.95%	42.19%	**37.92%**
CASME II	**0.55**	0.19	**71.00%**	**63.81%**	**59.09%**

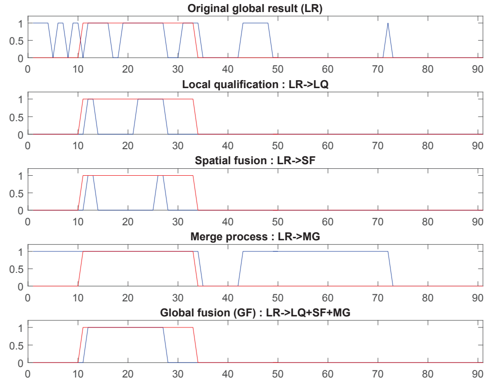


Figure 8. Example of global result obtained by each step in global fusion. The X axis is the frames index, the Y axis is the predicted label, the red curve represents the ground truth for this video and the blue curve is our global detection result.(CASME I_Sub08_EP12_2_1)

studied to improve the efficiency of decreasing the FP and FN. The process is analyzed on two levels, one is detection result per ROI and the other one is global result per frame by combining all ROIs. Supposing the dataset of results per ROI X_R is defined as $\{X_{ROI_1}, \cdots, X_{ROI_j}, \cdots, X_{ROI_J}\}$, the global result can be obtained by the following conditions:

$$X_G = \begin{cases} 1 & \exists X_{ROI_j} = 1, j \in [1, J] \\ 0 & \text{otherwise} \end{cases}$$

Figure 8 illustrates an example of the contribution of each step to the global result. The first layer in the figure shows the global result $D_{G_{original}}$ obtained directly by the local recognition (LR). The second, third and fourth layers give the global results $D_{G_{LQ}}$, $D_{G_{SF}}$ and $D_{G_{MG}}$ after local qualification (LQ), spatial fusion (SF) and merge process (MG) respectively over the dataset of $X_{R_{original}}$. And the fifth layer is the result of $D_{G_{GF}}$, in other words, the final result after three global fusion steps (GF).

As shown in the second and third layer of figure, the local qualification and spatial fusion eliminate the detected peaks which do not fit the selection criteria, meanwhile there is a risk that some TP frames are deleted. In the other side, the merge process puts the neighboring detected frames into an interval who has appropriate length. The figure at fifth layer illustrates the result with three steps of global fusion. Merged intervals that meet the requirement of LQ and SF are conserved.

Table X

ANALYSIS OF EACH STEP GF ON CASME II. THE RESULTS ARE EVALUATED IN TERM OF TP, FP, TPR, FPR, ACC. THE BEST RESULT OF ACC IS HIGHLIGHTED IN BOLD.

	TP	FP	TPR	FPR	ACC
LR	15855	11141	0.65	0.28	0.69
LQ	9492	5443	0.39	0.14	0.68
SF	6429	4257	0.26	0.11	0.65
MG	19778	18105	0.81	0.46	0.65
GF	13444	7624	0.55	0.19	**0.71**

Table XI

DETECTION RESULTS PER EMOTION IN TERM OF ACC ON CASME I

Emotion	CASME I-A		CASME I-B	
	NbVideo	ACC	NbVideo	ACC
Tense	48	0.77	23	0.81
Happiness	4	0.79	5	0.71
Repression	30	0.79	10	0.79
Disgust	4	0.80	42	0.81
Surprise	7	0.82	14	0.84
Contempt	4	0.68	6	0.86
Fear	1	0.55	1	0.72

The contribution is then evaluated for CASME II by accuracy metrics, as listed in Table X. The local qualification and spatial fusion can largely reduce the FP frames while the TP number is also decreased. Conversely, the merge process increase the number of both TP and FP. The combination of these three steps can reach to a balance, the ACC increases to 0.71 while the FPR is acceptable and the TPR is slightly impacted..

E. Detection per emotion

The databases are labeled by emotion type and AU information. The positions of chosen ROIs are related to the labeled AU. Hence, the detection performance per emotion is worth to analyze. Table XI lists the measurement per frame for CASME I. Despite of fear emotion with only one video sequence, there is nothing significant found by the analysis of ACC measure. While the tense and surprise emotion are mostly linked to the AU1, AU2 and AU4 of eyebrow, the emotion of happiness and repression always lead to the mouth movement. Even though the different emotion links to different AU, the type of emotion does not influence the detection result. Moreover, the result confirms that our proposed LTP is identical regardless of the ROI and the ME type.

V. CONCLUSION AND FUTURE WORKS

The LTP-ML method detects MEs using a local temporal pattern of facial movement, which is the same pattern for all the ROIs and alls the ME types. MEs can be distinguished from other movements by this pattern. A supervised learning algorithm is utilized to achieve this goal. In addition, this pattern allows us to identify spatial location where the ME occurs. Moreover, PCA is performed to facilitate the classification through the SVM by reducing the data dimension.

In future work, even though our method outperforms LBP-χ^2, there is plenty work to be done in this reducing false positives. In addition, more facial regions connected to ME movement should be considered, and the machine learning performance need to be improved to enhance the ability of distinguishing LTP and other motion pattern. Meanwhile, experiments should be performed on other databases, and the results should be compared with other methods, which is a difficult task because each method offers its own measures.

Acknowledgment

The authors would like to thank the financial support of China scholarship Council and ANR Reflet.

References

[1] R. L. Birdwhistell, "Communication without words," *Ekistics*, pp. 439–444, 1968.

[2] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, p. 88106, 1969.

[3] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, p. 118133, 2009.

[4] J. Endres and A. Laidlaw, "Micro-expression recognition training in medical students: a pilot study," *BMC medical education*, vol. 9, no. 1, p. 47, 2009.

[5] P. A. Stewart, B. M. Waller, and J. N. Schubert, "Presidential speechmaking style: Emotional response to micro-expressions of facial affect," *Motivation and Emotion*, vol. 33, no. 2, p. 125, 2009.

[6] M.-H. Chiu, H. L. Liaw, Y.-R. Yu, and C.-C. Chou, "Facial micro-expression states as an indicator for conceptual change in students understanding of air pressure and boiling points," *British Journal of Educational Technology*.

[7] P. Ekman and W. Friesen, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto: Consulting Psychologists*, 1978.

[8] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *arXiv preprint arXiv:1708.07549*, 2017.

[9] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, 2017.

[10] A. Moilanen, G. Zhao, and M. Pietikinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, p. 17221727.

[11] 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, p. e86041, 2014.

[12] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE, 2015, p. 665669.

[13] D. Patel, G. Zhao, and M. Pietikinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, p. 369380.

[14] A. K. Davison, C. Lansley, C. C. Ng, K. Tan, and M. H. Yap, "Objective micro-facial movement detection using facs-based regions and baseline evaluation," *arXiv preprint arXiv:1612.05038*, 2016.

[15] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *Signal Processing: Image Communication*, vol. 47, p. 170182, 2016.

[16] X. Li, J. Yu, and S. Zhan, "Spontaneous facial micro-expression detection based on deep learning," in *Signal Processing (ICSP), 2016 IEEE 13th International Conference on*. IEEE, 2016, p. 11301134.

[17] S.-J. Wang, S. Wu, and X. Fu, "A main directional maximal difference analysis for spotting micro-expressions," in *Asian Conference on Computer Vision*. Springer, 2016, p. 449461.

[18] H. Lu, K. Kpalma, and J. Ronsin, "Micro-expression detection using integral projections," 2017.

[19] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Computer Vision and Image Understanding*, vol. 147, p. 8794, 2016.

[20] X. Hong, T.-K. Tran, and G. Zhao, "Micro-expression spotting: A benchmark," *arXiv preprint arXiv:1710.02820*, 2017.

[21] D. Borza, R. Danescu, R. Itu, and A. Darabant, "High-speed video system for micro-expression detection and recognition," *Sensors*, vol. 17, no. 12, p. 2913, 2017.

[22] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, p. 532539.

[23] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, p. 433459, 2010.

[24] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–7.

[25] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikinen, "Reading hidden emotions: spontaneous micro-expression spotting and recognition," *arXiv preprint arXiv:1511.00423*, 2015.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.