# Détection de Micro-expressions par Reconnaissance de Motif Local Temporel de Mouvements Faciaux

J. LI C. Soladie R. Séguier

Équipe FAST, IETR/CentraleSupélec

jingting.li@supelec.fr

## Résumé

*Les micro-expressions (MEs) sont porteuses d'informations non verbales spécifiques. Cependant, de part leur nature locale et brève, il est difficile de les détecter. Dans cet article, nous proposons une méthode de détection par reconnaissance d'un motif local et temporel de mouvement du visage. Dans notre système, afin d'améliorer la précision de détection, nous extrayons des motifs temporels sur une fenêtre de 300ms (durée moyenne d'une ME), issus de la projection dans l'espace ACP des déformations faciales. Ce motif temporel est caractéristique des MEs et identique quelque soit la ME. A l'aide d'un algorithme de classification classique (SVM), nous distinguons ainsi les MEs des autres mouvements faciaux. Pour finir, nous appliquons une fusion globale sur l'ensemble du visage pour éliminer les faux positifs et les vrais négatifs. Les expérimentions sont effectuées sur deux bases de données publiques. Les résultats de détection montrent que la méthode proposée surpasse la méthode de détection la plus populaire en termes de précision.*

## Mots Clef

Micro-expression, Détection, Motif temporel et local, Apprentissage automatique.

## Abstract

*Micro expressions (MEs) convey specific non-verbal information. However, as they occurs briefly on local facial region, ME detection is a difficult task. In our paper, a local temporal pattern (LTP) of facial movement is proposed for the ME detection. In order to increase the detection accuracy, our method extracts the LTP from a PCA space projection in the video by a sliding window of 300ms (mean duration of MEs). The LTPs represent the ME movement variation and they are identical for all MEs. ME frames are then recognized by a classical classification method (SVM). To eliminate the false positives and true negatives, we employ finally a global fusion on the entire facial region. Experiments are performed on two public databases, and the detection results show that our proposed method performs better than the most popular detection method in terms of accuracy.*

## Keywords

Micro-expression, Detection, Local temporal pattern, Machine learning.

# 1 Introduction

L'expression faciale est l'un des indicateurs externes les plus importants pour connaître l'émotion et le statut psychologique d'une personne [1]. Parmi les expressions faciales, les micro-expressions (MEs) [9] sont des expressions locales et brèves qui apparaissent involontairement, notamment dans le cas de forte pression émotionnelle. Leurs durées varient de 1/25 à 1/5 de seconde [9]. Leur caractère involontaire permet souvent d'affirmer qu'elles représentent des émotions véritables d'une personne [9]. La détection de MEs a des applications nombreuses notamment dans le domaine de la sécurité nationale [7], des soins médicaux [11], des études sur la psychologie politique [21] et la psychologie de l'éducation [4].

L'existence de MEs a d'abord été découverte par Haggard et Isaacs en 1966 [13] puis Ekman et Friesen [9] l'ont nommée en 1969. Plusieurs années plus tard, ils ont développé un outil pour former les personnes à la détection de micro-expressions (METT) [6]. Cependant, même pour un expert, le taux de reconnaissance des MEs est inférieur à 50% [12]. Pour coder les MEs, le système de codage d'actions faciales (FACS) [8] est souvent utilisé. Il a été créé pour analyser la relation entre la déformation du muscle facial et l'expression émotionnelle. Les unités d'action (AUs) sont les composantes faciales du FACS, qui représentent le mouvement musculaire local. L'étiquette de l'AU sur le visage permet d'identifier la ou les régions où la ME se produit. En conséquence, le système FACS peut aider à annoter l'apparence et la dynamique d'une ME dans une vidéo.

Depuis les années 2000, la recherche sur la détection et la reconnaissance automatique de micro-expressions (MEDR) s'est développée. Le nombre des articles reste

faible et les résultats ne sont pas encore très satisfaisants du fait de la nature même des MEs (micro) ainsi que du nombre limité de bases de données (BDDs) publiques de MEs. Cependant, il y a eu de plus en plus d'études émergentes ces dernières années (28 papiers en 2017 versus 12 en 2015 et 5 en 2013). Les études se focalisent principalement sur la reconnaissance des MEs, plus que sur la détection (plus de 2/3 des papiers traitent de la reconnaissance). Ces études supposent connues les images d'onset et d'offset des MEs. Pourtant, de part leur nature brève et de faible intensité, les MEs sont très difficiles à détecter. Les résultats de détection des méthodes proposées actuellement ne sont pas assez précis, et le taux de détection est reste faible (meilleur taux d'environ 70% [15]). Même lorsque les ME sont produites dans un environnement strictement contrôlé, les faux positifs sont très nombreux en raison des mouvements de la tête ou du clignement des yeux.

Ce papier explore un système automatique permettant de détecter des MEs. Un tel système doit être capable de :

— séparer les mouvements relatifs aux MEs des mouvements de la tête ou du clignement des yeux.
— détecter la région dans laquelle la ME se produit.

La principale contribution de cet article est de proposer une nouvelle méthode de détection de MEs utilisant un motif temporel local (local temporal pattern, LTP) issu d'une projection des déformations faciales dans un espace créé par ACP. L'originalité de l'approche est l'utilisation d'un motif temporel, correspondant aux onset et offset des MEs. Ce motif temporel est caractéristique des MEs et est identique pour toutes les MEs quelque soit la zone du visage où la ME se produit. Il correspond à l'activation ou au relâchement du muscle lors d'une ME. Dans notre système, ce motif temporel est appris sur les vidéos avec ou sans MEs et correspond à un intervalle de longueur de la durée moyenne d'une ME. L'extraction de caractéristiques temporelles sur un intervalle long est réalisée par le calcul de la distance entre la première image et la $n^{ieme}$ image de l'intervalle. Afin de conserver la variation la plus significative, une ACP est préalablement réalisée. La forme de la courbe dans Figure 1 représente la variation temporelle du début au sommet de la ME (onset à apex). La particularité de l'approche est la combinaison de traitements locaux et globaux. La détection des motifs temporels est réalisée de façon locale par ROI (Region of Interest - Région d'Intérêt), puis un système de règles sur l'ensemble du visage permet de séparer les MEs des autres mouvements faciaux. Par construction, notre méthode permet aussi de déterminer la position temporelle du début de la ME, c'est-à-dire les index des images où les motifs sont détectés. En outre, de part de nature de processus, nous pouvons localiser la région où la ME se produit.

L'article est organisé comme suit : la section II présente les travaux connexes sur la détection des MEs ; la section III décrit notre méthode en utilisant le motif temporel pour la détection des MEs ; la section IV présente les résultats expérimentaux ; la section V conclut le papier.

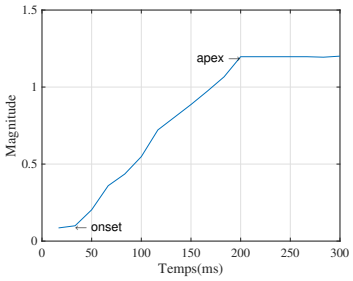



FIGURE 1 – Exemple de motif local temporel lors d'une ME dans la région du sourcil droit sur une durée de 300ms (Vidéo Sub01_EP01_5 de CASMEI).

## 2 Travaux connexes

Étant donné que les MEs sont des expressions faciales involontaires, la majorité des recherches sur la détection de ME est développée à partir de BDDs publiques spontanées. Cette section présente les travaux de recherche connexes en introduisant leur méthode principale et les caractéristiques utilisées. En outre, les avantages et les inconvénients de ces méthodes sont discutés.

L'idée essentielle actuelle de la plupart des méthodes est de calculer des différences entre valeurs de caractéristiques faciales, c'est-à-dire entre la première image et les autres images, dans une fenêtre temporelle donnée. Les caractéristiques utilisées sont diverses, par exemple LBP, HOG, flux optique, projection intégrale. Moilanen et coll. [19, 15] ont utilisé l'analyse de la distance chi-Square du motif binaire local (LBP). Les MEs ont ensuite été extraites par la comparaison à un seuil et la détection de crête. Yan et coll. [25] ont quantifié les MEs et repéré les images de sommet (apex) par trois algorithmes d'extraction de caractéristiques : Contrainte Modèle Local (CLM), LBP et flux optique. Liong et coll. [17] ont repéré l'image de sommet en employant une stratégie de recherche binaire. Patel et coll. [20] ont utilisé un flux optique, puis une intégration spatio-temporelle pour repérer l'image de sommet et aussi identifier la localisation de l'apparition (onset) et de la disparition (offset) de la ME. Davison et al. [5] ont utilisé des Histogrammes de gradient orienté 3D (HOG) comme mesure pour calculer la dissimilitude entre les images et extraire la ME en utilisant une ligne de base individualisée. Liong et coll. [16] ont présenté la contrainte optique comme caractéristique qui est plus efficace pour identifier les déformations subtiles du muscle facial. Wang et coll. [22] ont proposé une méthode utilisant l'amplitude de la différence maximale dans la direction principale du flux optique. Le calcul du flux optique étant coûteux, Lu et coll. [18] ont présenté une méthode avec un faible coût de calcul basé sur les différences entre les caractéristiques de la projection intégrale (IP) dans des images séquentielles. Le principal avantage de ces approches est de pouvoir effectuer les comparaisons entre images sur une fenêtre temporelle de la taille d'une ME. En revanche, c'est le mouve-

ment entre images qui est détecté, et pas spécifiquement le mouvement caractéristique des MEs. C'est pourquoi la capacité à distinguer les MEs des autres mouvements (tel que le clignement des yeux ou les mouvements de la tête) reste faible.

Récemment, de nouvelles approchent utilisant l'apprentissage automatique sont apparues. Elles ont notamment pour objectif de pouvoir différencier les mouvements des MEs des autres mouvements faciaux. Xia et coll. [23] ont utilisé AdaBoost pour estimer la probabilité initiale pour chaque image avec un modèle de marche aléatoire qui repère la ME en considérant la corrélation entre les images. Hong et coll. [14] ont proposé une approche basée sur des fenêtres coulissantes à échelles multiples. Les descripteurs LBP-TOP, HOG-TOP et HIGO (Histogramme de gradient orienté de l'image)-TOP ont été extraits et les MEs sont détectées par classification binaire. Borza et al. [2] ont utilisé la magnitude du mouvement comme caractéristique, puis l'algorithme Adaboost a été appliqué pour détecter les images de ME. L'apprentissage automatique permet effectivement d'éviter de détecter certains mouvements alors qu'il n'y a pas de ME dans la vidéo. Cependant, même si des descripteurs telles que LBP-TOP extraient des caractéristiques temporelles dans une petite fenêtre temporelle, la durée est trop courte pour représenter un modèle de mouvement temporel pour l'ensemble de la ME. Le classificateur se focalise principalement sur la détection d'un motif spatial mais pas sur la variation de motif temporel dans une fenêtre de durée de l'ordre d'une ME.

En outre, toutes les méthodes ci-dessus n'utilisent pas explicitement le fait que la ME est un mouvement facial local : les caractéristiques extraites de la région locale sont intégrées dans une caractéristique qui représente le mouvement du visage entier.

En conséquence, nous avons proposé une méthode de détection de MEs qui utilise directement un modèle temporel extrait de régions locales. Comparés aux méthodes existantes, nous prenons en compte beaucoup plus d'images pour avoir un motif temporel complet (et non une différence entre deux images) et local. Ce motif est analysé ensuite par un classifieur, afin de le distinguer d'autres mouvements faciaux. Pour ce qui est du coté spatial des MEs, nous n'apprenons pas de motif de déformation spatial mais nous utilisons une approche mixte : local / global. Cette méthode nous permet d'améliorer la capacité de distinguer la ME des autres mouvements, de trouver la localisation spatiale de la ME et la localisation temporelle du début de la ME.

# 3 Notre méthode : LTP-ML

La méthode proposée se compose de trois parties : un prétraitement permettant de détecter les régions d'intérêt (ROIs) du visage, l'extraction des LTPs (local temporal pattern) sur ces régions et finalement la détection à proprement parler des MEs. La figure 2 présente le processus global. Notre méthode est appelée LTP-ML pour local temporal pattern et machine learning.

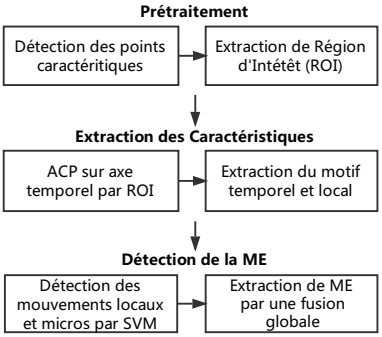



FIGURE 2 – Processus principal de notre méthode LTP.

## 3.1 Prétraitement : Détermination des ROIs

Le prétraitement se décompose en deux étapes. La première étape consiste à détecter des points caractéristiques du visage. Pour cela, nous utilisons l'outil de 'Intraface' [24] qui détecte automatiquement 49 points caractéristiques (PtCs) pour chaque image. La deuxième étape consiste à extraire des régions d'intérêt (ROI) où peuvent se produire les MEs. Les régions d'intérêt sont générées en forme de carré autour de certains PtCs. La longueur $a$ du côté du carré est déterminée par la distance $L$ entre les deux PtCs les plus proches entre l'œil gauche et l'œil droit (n° 23 et 26) : $a = (1/5) \times L$. La figure 3 illustre le résultat du prétraitement sur une image.

Les ROIs que nous avons sélectionnées correspondent aux AU où se produisent les MEs le plus fréquemment. Le tableau 1 donne le lien entre AU et ROI selon la distribution des AUs [8]. En raison de la rigidité du nez, cette région est choisie comme une référence pour éliminer des mouvements globaux de la tête (ROI 11,14), e.g. changement de pose. Les régions des yeux sont négligées à cause du clignement.

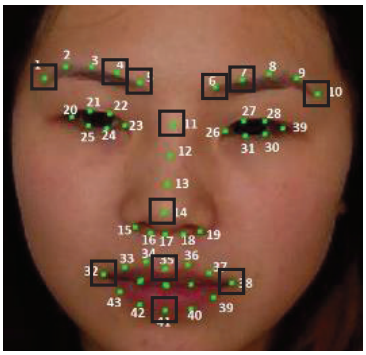


FIGURE 3 – Points caractéristiques et distribution des ROIs. 49 points caractéristiques sont détectés et 12 ROIs sont générées dépendant de la position des points choisis dans les régions des sourcils, du nez et du contour de la bouche (©Xiaolan Fu).

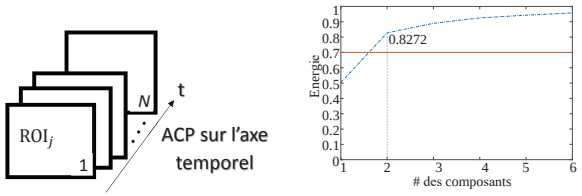
TABLE 1 – Lien entre AU et ROI

| Région du visage | Index de ROI | AU correspondante |
|---|---|---|
| Sourcil | 1, 4, 5, 6, 7, 10 | 1, 2, 4 |
| Nez | 11, 14 | 9 |
| Bouche | 32, 35, 38, 41 | 10, 12, 14, 15, 17, 25 |

## 3.2 Extraction des caractéristiques : LTP

L'objectif de cette partie est d'extraire des caractéristiques temporelles locales permettant de distinguer les MEs des autres mouvements faciaux. Pour cela, nous extrayons d'abord les principales déformations de texture (niveau de gris) au cours du temps pour chaque ROI. À cette fin, nous construisons une séquence temporelle locale par ROI et puis nous réalisons une analyse en composantes principales (ACP) sur chaque séquence. La figure 4(a) illustre ce traitement sur une des ROIs.

S'il y a $N$ images dans une vidéo, chaque image locale est de taille de $a^2$, la matrice $\mathbf{I}$ à traiter par ACP est de $N \times a^2$. Les premières composantes de l'ACP vont nous donner l'information du mouvement au cours du temps pour cette ROI. En conservant les deux premières composantes, comme montré dans la figure 4(b), plus de 70% énergie est conservée, et puis nous obtenons une matrice $\mathbf{P}$ de taille $N \times 2$.




(a) (b)

FIGURE 4 – Analyse d'ACP processus. Figure 4(a) affiche ACP sur une séquence temporelle locale. Une petite séquence vidéo pour la ROI$_j$ contient $N$ images, et est traitée par ACP sur l'axe temporel. Figure 4(b) montre un exemple d'analyse énergétique d'ACP. Les deux premiers composants peuvent conserver plus de 70% d'énergie.

Chaque point $P_i$ de cette matrice a deux composantes et représente une image locale. Il y a une relation entre la distance des points $P_i$ entre eux et l'amplitude du mouvement entre les deux images correspondantes : quand la distance augmente, l'amplitude du mouvement augmente.

Les variations de distances sont étudiées sur des fenêtres glissantes pour chaque ROI. La durée de la fenêtre est de 300ms afin de correspondre à la durée moyenne d'une ME. Les distances entre la première image de la fenêtre et les autres images sont calculées. Supposons qu'il y ait $K+1$ images dans la fenêtre (intervalle), l'ensemble des distances dans cet intervalle est de :

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

où $i \in [1, K]$, $j$ est l'index de la ROI, $n$ est l'index de l'image courante, $\Delta_j(n, n+i)$ signifie la distance Euclidienne entre le point $P_{n+i}$ de la $(i+1)^{eme}$ image de l'intervalle et le point $P_n$ de la première image dans l'intervalle. Ainsi, à chaque image correspond un ensemble de $K$ distances comme indiqué dans le tableau 2.

TABLE 2 – Ensemble de distances pour chaque image

| index de l'image | distances originales |
|---|---|
| 1ère image | $\Delta_j(1,2), \cdots, \Delta_j(1, K+1)$ |
| ... | ... |
| n ème image | $\Delta_j(n, n+1), \cdots, \Delta_j(n, n+K)$ |
| ... | ... |
| N ème image | $0, \cdots, 0$ |

La valeur de l'amplitude des mouvements n'est pas la même pour les MEs de vidéos différentes. Nous réalisons donc une normalisation de ces distances. Comme la période de l'onset est de l'ordre de 150ms (soit $K/2$), nous normalisons par rapport à la distance maximale sur cette période pour chaque ROI :

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

Le coefficient de la normalisation est $CN_j = 1/\Delta_j^{max}$ et la distance normalisée est :

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

La figure 1 affiche le motif temporel final qui va caractériser la ME. Dans ce motif spécifique pour la ME, la tendance croissante est d'environ $K/2$ images et monte à une valeur d'environ 1 (grâce à la normalisation). Cela correspond au mouvement de l'onset de la ME. La courbe dans $[K/2, K]$ reste stationnaire ou commence à décroître (apex et/ou début d'offset).

Pour finir, nous ajoutons la valeur des CNs dans la caractéristique afin de pouvoir éliminer les mouvements trop subtils. Si le nombre de ROIs est $J$, les caractéristiques d'une image sont donc une matrice de taille $J \times (K+1)$, comme présentée dans le tableau 3.

TABLE 3 – Caractéristique LTP d'une image

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

## 3.3 Détection de la ME

Dans cette partie, nous réalisons la détection des MEs en deux temps : reconnaissance locale du motif et fusion globale des ROIs.

**Reconnaissance du motif local et temporel.** Les LTP sont identiques quel que soit la ROI de la ME. En effet, les deux premières composantes de l'ACP conservent les principales variations de mouvement et des informations spatiales sur la ME. Grâce à la nature identique du LTP, les mouvements locaux peuvent être classés par apprentissage automatique indépendamment de la ROI.
Nous utilisons la méthode SVM qui est une classification supervisée. Pour étiqueter la BDD, nous avons trouvé empiriquement que les $K/3$ images qui sont avant l'onset de ME conservent le meilleur LTP motif. C'est raisonnable parce que la variation d'onset à apex que nous voudrons étudier est incluse dans les motifs des images avant l'onset. Nous annotons donc les images de chaque ROI séquence comme le schéma illustré dans Figure 5.


FIGURE 5 – Schéma de l'annotation d'une base de données. Le bleu représente une vidéo entière, la partie orange correspond à la ME. Les images dans l'intervalle [onset-$K/3$, onset] sont classées en étiquette 1 et les autres images sont étiquetés en 0 vis-à-vis du motif LTP.

L'index des ROIs pour l'entraînement est ensuite sélectionné par rapport à l'AU de la ME de la vidéo, comme illustré dans le tableau 4. Les indices d'AU viennent de la théorie d'Ekman [10]. Les résultats de la détection locale sont générés par LOSubOCV (Leave-One-Subject-Out Cross Validation).

TABLE 4 – Sélection des ROIs pour l'entraînement

| Condition sur les AUs | Index de ROI | Région du visage |
|---|---|---|
| Indice des AUs < 6 | 1,4,5,6,7,10 | Sourcil |
| Indice des AUs > 9 | 32,35,38,41 | Contour de la bouche |
| Autrement | Toutes les ROIs | Visage entier |

**Fusion globale.** Dans cette sous partie, nous éliminons des faux positifs (qui concernent les autres types de mouvements) et réduisons le nombre de vrais négatifs causés par notre processus de reconnaissance. Cela est réalisé en trois étapes : une qualification locale, une fusion spatiale et un processus de lissage.
Tout d'abord, nous utilisons deux seuils $T_{CN}$ et $T_{dist}$ pour supprimer les mouvements trop subtils qui pourraient correspondre à du bruit, bien qu'ayant le motif LTP. Si le CN de cette ROI est supérieur à $T_{CN}$ ou si la valeur de distance maximale dans l'ensemble est inférieure à $T_{dist}$, le résultat détecté est considéré comme un mouvement subtil qui peut être ignoré. De plus, nous limitons localement le nombre d'images avec une étiquette à 1 dans un intervalle de longueur $K$. En effet, par construction, la durée de le ME est d'environ K images et la condition optimale de détection de $K/3$ images avant l'onset de la ME. Avoir moins de $K/9$ motifs détectés ou plus de $K/2$ ne correspond pas à une ME.
Deuxièmement, nous effectuons une fusion spatiale des ROIs pour obtenir un résultat global sur le visage entier. Pour réduire le nombre de faux positifs, les mouvements ne correspondant pas à des ME sont éliminés par les règles spatiales suivantes.S'il y a plus de $J/2$ ROIs du visage entier ou plus d'une ROI du nez qui sont correspondent au motif LTP, ce mouvement est alors considéré comme un mouvement de tête global. De plus, le clignement des yeux conduit à un mouvement de toutes les zones autour des yeux. Ainsi, si toutes les ROIs des sourcils détectent un motif, notre système suppose qu'il y a un clignement des yeux et considère alors que ce n'est pas une ME.
Troisièmement, comme les résultats de l'algorithme de reconnaissance sont donnés par image, nous fusionnons des zones proches pour lesquelles un motif a été détecté.

# 4 Expérimentation

Dans cette section, la performance de LTP-ML est évaluée par la comparaison avec la méthode LBP-Chi-square-distance en utilisant deux BDDs publiques. Les résultats sont analysés par vidéo et par image. Afin d'analyser les performances de la fusion globale, les contributions de chaque étape sont illustrées.

## 4.1 Méthode de comparaison

La méthode LBP-Chi-square-distance (LBP-$\chi^2$) a été premièrement proposée par Moilanen et al. en 2014 [19]. C'est la méthode à laquelle nous allons nous comparer car c'est celle qui est la plus généralement utilisée pour comparer les résultats concernant la détection de MEs. Certaines méthodes [18, 15] évaluent leurs résultats à l'aide des métriques ROC et AUC. Dans notre cas, ces métriques ne sont pas adaptées car il n'y a pas de paramètre utile à ajuster. A noter que les résultats présentés dans le papier [19] considèrent que le clignement des yeux est un vrai positif, ce qui n'est pas le cas dans la vérité terrain des BDDs. Pour cette raison, nous avons ré-implémenté la méthode à partir de l'article et réussi à atteindre le même niveau de taux de détection.

## 4.2 BDDs et configuration

**Bases de données.** Les expérimentations sont faites sur deux BDDs de MEs spontanées : CASME I [26] et CASME II [25]. Les MEs dans ces deux BDDs sont étiquetées avec une vérité de terrain fiable, incluant la localisation temporelle d'onset, d'apext et d'ofset de la ME. Les paramètres essentiels sont répertoriés dans le tableau 5. Toutes les séquences dans CASME I et CASME II sont utilisées dans l'expérimentation.

**Configuration de l'expérimentation.** La configuration est faite d'après les descriptions des trois articles : [19], [15] et [14]. Pour la méthode LBP-$\chi^2$ : le visage est divisé en 36 blocs avec un chevauchement. Les taux de chevauchement dans la direction de X et Y sont 0.2 et 0.3 respectivement. Pour l'extraction de caractéristiques LBP du

bloc, un mappage uniforme est utilisé, le rayon r est mis à $r = 3$, et le nombre de points voisins p est fixé à $p = 8$. Les distances $\chi^2$ de chaque image sont calculées dans un intervalle $2 \times L + 1$. la valeur d'L de deux bases de données est affichée dans le tableau 5. Pour la vérité terrain de LBP-$\chi^2$, nous considérons l'intervalle de [début$-L/2$, fin$+L/2$] comme étant valeur 1.

Pour notre méthode LTP-ML, la taille $K$ de l'intervalle temporel est aussi affichée dans le tableau 5, cela correspond à 300ms d'après les FPS de chaque BDDs, qui est la durée moyenne d'une ME. L'apprentissage et la reconnaissance sont effectués à l'aide du logiciel Lib-SVM avec noyau linéaire [3]. Comme la distribution d'étiquettes dans les BDDs est très déséquilibrée, c'est-à-dire qu'il y a trop d'images non ME, nous effectuons un échantillonnement de 1 image sur 8. Les résultats sont obtenus par Leave-one-Subject-Out cross validation (LOSubOCV). Comme LTP-ML détecte le motif spécifique local du début de ME, la condition optimale est de détecter les motifs dans l'intervalle de [début$-K/3$, début] et dans l'intervalle de [sommet$-K/3$, sommet]. Nous introduisons donc un décalage de $K/3$ comparé à la vérité terrain de LBP-$\chi^2$, ce qui signifie que la vérité terrain est de [début$-K/3 - L/2$ , fin$-K/3 + L/2$].

TABLE 5 – Paramètres principaux des BDDs

| BDD | Sujets | MEs | FPS | L | K |
|---|---|---|---|---|---|
| CASMEI-A | 7 | 96 | 60 | 21 | 18 |
| CASMEI-B | 12 | 101 | 60 | 21 | 18 |
| CASMEII | 26 | 255 | 200 | 65 | 60 |

## 4.3 Comparaison des résultats avec LBP-$\chi^2$

Puisque le processus LBP-$\chi^2$ n'effectue pas de lissage après la détection de crête, les résultats obtenus par notre méthode LTP-ML sont d'abord comparés à LBP-$\chi^2$ sans processus de lissage. Pour évaluer la performance, le résultat de la détection est mesuré par vidéo et par image respectivement.

**Comparaison des résultats par vidéo.** Il est nécessaire de déterminer si la vidéo de test contient les ME séquences. Par conséquent, le résultat de la détection par vidéo est analysé. Dans ces deux BDDs, chaque vidéo à tester a un clip ME, et les pics détectés comme faux positifs sont déterminés par une fenêtre de recherche sans chevauchement (600ms). Tableau 6 illustre le résultat des méthodes LBP-$\chi^2$ et LTP-ML.

TABLE 6 – Résultat de détection par vidéo

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

CASME I-A a 96 vidéos, et notre méthode a détecté 80 MEs avec succès. Même si le nombre de faux positifs est un peu plus élevé que LBP-$\chi^2$, les mesures de précision sont plus élevées que LBP-$\chi^2$ en terme de TPR (Rappel), précision et F1-score, ce qui indique que LTP-ML détecte mieux les MEs que LBP-$\chi^2$. Par ailleurs, nous avons des résultats légèrement supérieurs sur CASME II, et des résultats équivalents pour CASME I-B. Les proportions d'images avec ME par rapport aux images sans ME sont de 0,19 pour CASME I et de 0,38 pour CASME II. De plus, la résolution faciale de CASME I-A est supérieure à celle de CASME I-B, la ROI contient plus de pixels mais apporte aussi plus de bruit. Selon la comparaison ci-dessus, LTP-ML est aussi bien que LBP-$\chi^2$ et la méthode semble mieux fonctionner dans les cas où il y a plus de mouvements du visage et plus de bruit (CASME-A).

**Résultat de la détection par image.** Les caractéristiques LTP, la classification locale et les deux premières étapes de la fusion globale sont effectuées pour chaque image. Ainsi, nous pouvons calculer une mesure de précision par image. Cette information est intéressante car ce sont les numéros d'image d'onset et d'offset qui sont utilisées dans les méthodes de reconnaissance de ME. Pour analyser les résultats, nous utilisons la matrice de confusion [TPR, FNR ; FPR, TNR]. Les résultats des deux méthodes de détection sont présentés dans le tableau 7. En raison de l'absence de lissage, le TPR n'est pas très élevé. Comparé au résultat de LBP-$\chi^2$, notre méthode surpasse la détection de ME car le TPR de notre méthode est supérieur à 20 % pour toutes les bases de données alors que le pourcentage de TPR de LBP-$\chi^2$ est toujours inférieur à 10 %. Comme certains mouvements faciaux ont un motif similaire à celui d'une ME, le résultat de détection non-ME de notre méthode est légèrement inférieur. Néanmoins, le FPR est maintenu relativement faible : 0,08 pour le CASME I-A, 0,05 pour CASME I-B et 0,09 pour CASME II.

TABLE 7 – Matrice de confusion pour la détection de MEs

| BDD | | LBP-$\chi^2$ | | 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 |

Afin de confirmer l'efficacité de notre méthode, nous avons choisi les trois métriques suivantes : ACC, précision et score F1. Ces métriques sont celles qui sont le plus souvent utilisées pour l'évaluation de la méthode d'apprentissage automatique. Les résultats sont listés dans le tableau 8. Tout d'abord, notons que la moyenne d'ACC est maintenue à 75 %. Les résultats obtenus rendent difficile la comparaison à la méthode LBP-$\chi^2$ en termes d'ACC et de précision. Par contre, notre méthode surpasse largement LBP-$\chi^2$ en termes de F1-score pour chaque BDD. Cela signifie que

notre méthode LTP-ML peut détecter plus d'images de ME tout en maintenant un FPR acceptable et une meilleure performance de classification. De plus, pour éliminer les faux négatifs, le lissage est appliqué sur les images détectées par LTP-ML. Le tableau 9 montre le résultat final de la détection LTP-ML avec le processus de lissage. Par rapport aux résultats sans lissage, les valeurs TPR et F1-score ont nettement augmenté. Le lissage permet de fusionner les images détectées de façon discrète et d'améliorer la capacité de la détection des MEs.

TABLE 8 – Résultat de détection par image

| BDD | Méthode | ACC | Précision | F1-score |
|---|---|---|---|---|
| CASME I-A | LBP-$\chi^2$ | **78.75%** | 38.35% | 8.78% |
| | LTP-ML | 77.90% | **43.24%** | **29.87%** |
| CASME I-B | LBP-$\chi^2$ | **82.92%** | **46.34%** | 12.05% |
| | LTP-ML | 82.61% | 45.67% | **29.64%** |
| CASME II | LBP-$\chi^2$ | 64.08% | **73.67%** | 15.66% |
| | LTP-ML | **65.07%** | 60.59% | **34.96%** |

TABLE 9 – Résultat de détection de LTP-ML avec le processus de lissage

| BDD | TPR | FPR | ACC | Précision | 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%** |

## 4.4 Analyse de la contribution de chaque étape de la fusion globale

Comme la fusion globale (FG) comprend trois étapes : qualification locale (QL), fusion spatiale (FS) et lissage (LG), les contributions de chaque étape valent la peine d'être étudiées pour évaluer leur impact respectif. A supposer que l'ensemble des résultats de reconnaissance des ROIs $\{X_{ROI_1}, \cdots, X_{ROI_j}, \cdots, X_{ROI_J}\}$ est $X_R$, le résultat global $X_G$ est obtenu par la condition suivante :

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

La figure 6 donne un exemple des contributions de chaque étape de fusion sur le résultat global en combinant toutes les ROIs choisies. La première couche de la figure est le résultat global $D_{G_{original}}$ obtenu directement par la reconnaissance locale (RL). La deuxième, la troisième et la quatrième couche sont les résultats globaux $D_{G_{QL}}$, $D_{G_{FS}}$ et $D_{G_{LG}}$ après QL, FS et LG respectivement sur l'ensemble de $X_{R_{original}}$. Et la cinquième couche est le résultat de $D_{G_{FG}}$, ce qui signifie le résultat final après les trois étapes de traitement.

Les résultats de la contribution de chaque étape pour la base CASME II sont présentés dans le tableau 10. La QL et la FS réduisent considérablement les FP mais aussi le nombre TP. Inversement, le processus de lissage augmente le nombre de TP et de FP. La combinaison de ces trois étapes permet d'atteindre un équilibre : l'ACC augmente à 0,71, le FPR diminue considérablement et le TPR est légèrement impacté.

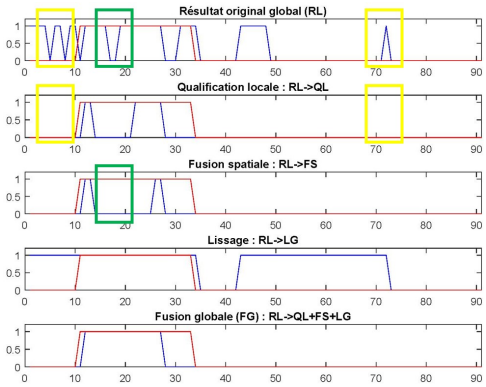



FIGURE 6 – Contribution de chaque étape de fusion globale. L'axe X correspond au numéro de l'image et l'axe Y à l'étiquette prédite. La courbe rouge représente la vérité terrain pour cette vidéo (CASME I_Sub08_EP12_2_1). Le premier bloc jaune dans la première couche représente des mouvements trop petits et le deuxième bloc jaune est un mouvement très court, ils sont éliminés par la qualification locale ; le bloc vert est un mouvement de la tête qui est éliminé par la fusion spatiale ; le résultat après trois étapes de la fusion globale correspond à la vérité terrain.

TABLE 10 – Analyse de chaque étape de FG sur la base CASME II.

| | TP | FP | TPR | FPR | ACC |
|---|---|---|---|---|---|
| RL | 15855 | 11141 | 0.65 | 0.28 | 0.69 |
| QL | 9492 | 5443 | 0.39 | 0.14 | 0.68 |
| FS | 6429 | 4257 | 0.26 | 0.11 | 0.65 |
| LG | 19778 | 18105 | 0.81 | 0.46 | 0.65 |
| FG | 13444 | 7624 | 0.55 | 0.19 | **0.71** |

# 5 Conclusion

La méthode LTP détecte les MEs en utilisant un motif local temporel du mouvement facial, qui est le même pour toutes les ROIs et tous les types de ME. Ce motif permet de distinguer les MEs des autres mouvements. Pour cela nous avons utilisé un algorithme d'apprentissage supervisé. Ce motif nous permet aussi d'identifier la localisation spatiale où la ME se produit. Ce motif est issu d'une ACP locale qui permet à la fois d'extraire l'information de mouvement principale mais aussi de faciliter la classification de SVM en réduisant la dimension des données.

Dans les travaux futurs, nous souhaitons continuer à travailler sur la diminution du nombre de faux positifs. Les relations entre les ROIs et l'influence de la taille de ROI vont être considérées. De plus, la performance d'apprentissage machine doit être améliorée pour renforcer la capacité

de distinguer les LTPs et les autres motifs de mouvement. Pour finir, il serait intéressant de faire des expérimentations sur d'autres BBDs et sur des vidéos longues. Il faut aussi comparer nos résultats avec d'autres méthodes. Cela reste une tâche difficile car chaque méthode propose ses propres métriques.

## Remerciement

Les auteurs remercient sincèrement le soutien financier du China Scholarship Council et ANR Reflet.

# Références

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

[2] D. Borza, R. Danescu, R. Itu, et al., High-speed video system for micro-expression detection and recognition, *Sensors*, vol. 17(12), 2913, 2017.

[3] C.-C. Chang, C.-J. Lin, Libsvm : a library for support vector machines, *ACM transactions on intelligent systems and technology (TIST)*, vol. 2(3), 27, 2011.

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

[5] A. K. Davison, M. H. Yap, C. Lansley, Micro-facial movement detection using individualised baselines and histogram-based descriptors, in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, p. 1864–1869, IEEE, 2015.

[6] P. Eckman, Emotions revealed, *St. Martin's Griffin, New York*, 2003.

[7] P. Ekman, Lie catching and microexpressions, *The philosophy of deception*, p. 118–133, 2009.

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

[9] P. Ekman, W. V. Friesen, Nonverbal leakage and clues to deception, *Psychiatry*, vol. 32(1), 88–106, 1969.

[10] P. Ekman, W. V. Friesen, Facial action coding system, 1977.

[11] J. Endres, A. Laidlaw, Micro-expression recognition training in medical students : a pilot study, *BMC medical education*, vol. 9(1), 47, 2009.

[12] M. Frank, M. Herbasz, K. Sinuk, et al., I see how you feel : Training laypeople and professionals to recognize fleeting emotions, in *The Annual Meeting of the International Communication Association. Sheraton New York, New York City*, 2009.

[13] E. A. Haggard, K. S. Isaacs, Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, in *Methods of research in psychotherapy*, p. 154–165, Springer, 1966.

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

[15] X. Li, X. Hong, A. Moilanen, et al., Towards reading hidden emotions : A comparative study of spontaneous micro-expression spotting and recognition methods, *IEEE Transactions on Affective Computing*, 2017.

[16] S.-T. Liong, J. See, R. C.-W. Phan, et al., Spontaneous subtle expression detection and recognition based on facial strain, *Signal Processing : Image Communication*, vol. 47, 170–182, 2016.

[17] S.-T. Liong, J. See, K. Wong, et al., Automatic apex frame spotting in micro-expression database, in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, p. 665–669, IEEE, 2015.

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

[19] A. Moilanen, G. Zhao, M. Pietikäinen, Spotting rapid facial movements from videos using appearance-based feature difference analysis, in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, p. 1722–1727, IEEE, 2014.

[20] D. Patel, G. Zhao, M. Pietikäinen, Spatiotemporal integration of optical flow vectors for micro-expression detection, in *International Conference on Advanced Concepts for Intelligent Vision Systems*, p. 369–380, Springer, 2015.

[21] P. A. Stewart, B. M. Waller, J. N. Schubert, Presidential speechmaking style : Emotional response to micro-expressions of facial affect, *Motivation and Emotion*, vol. 33(2), 125, 2009.

[22] S.-J. Wang, S. Wu, X. Fu, A main directional maximal difference analysis for spotting micro-expressions, in *Asian Conference on Computer Vision*, p. 449–461, Springer, 2016.

[23] Z. Xia, X. Feng, J. Peng, et al., Spontaneous micro-expression spotting via geometric deformation modeling, *Computer Vision and Image Understanding*, vol. 147, 87–94, 2016.

[24] X. Xiong, 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*, p. 532–539, 2013.

[25] W.-J. Yan, X. Li, S.-J. Wang, et al., Casme ii : An improved spontaneous micro-expression database and the baseline evaluation, *PloS one*, vol. 9(1), e86041, 2014.

[26] W.-J. Yan, Q. Wu, Y.-J. Liu, et al., 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)*, pp. 1–7, 2013.