Pattern Recognition Letters 135 (2020) 382–389

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Mean oriented Riesz features for micro expression classification☆

Carlos Arango Duque, Olivier Alata*, Rémi Emonet, Hubert Konik, Anne-Claire Legrand

Lab. Hubert Curien, CNRS UMR 5516, UJM, IOGS, Univ. Lyon, 42023 Saint Etienne, France

ARTICLE INFO

Article history:
Received 14 October 2019
Revised 26 April 2020
Accepted 5 May 2020
Available online 8 May 2020

MSC:
68U10
68T10

Keywords:
Micro-expressions
Mean oriented Riez features
Recognition
Classification
Subtle motion analysis
Riesz pyramid
Monogenic signal
SMIC database
CASME2 database

ABSTRACT

Micro-expressions are brief and subtle facial expressions that go on and off the face in a fraction of a second. This kind of facial expressions usually occurs in high stake situations and is considered to reflect a human's real intent. There has been some interest in micro-expression analysis, however, a great majority of the methods are based on classically established computer vision methods such as local binary patterns, histogram of gradients and optical flow. A novel methodology for micro-expression recognition using the Riesz pyramid, a multi-scale steerable Hilbert transform is presented. In fact, an image sequence is transformed with this tool, then the image phase variations are extracted and filtered as proxies for motion. Furthermore, the dominant orientation constancy from the Riesz transform is exploited to average the micro-expression sequence into an image pair. Based on that, the Mean Oriented Riesz Feature description is introduced. Finally the performance of our methods are tested in two spontaneous micro-expressions databases and compared to state-of-the-art methods.

本文贡献
1.使用里斯金字塔(Riesz pyramid)方法，多尺度可操纵的希尔伯特变换
2.图片序列使用此变换，可以提取图像帧的变化并过滤用于标识动作。
3.利用Riesz变换的显性取向常数将微表情序列平均成图像对
4.基于此，引入均值导向的Riesz特征
5.试验结果：在两个自发微表情数据集上测试，获得sota方法

1. Introduction

Micro-expressions (MEs) are brief and subtle facial expressions that last a fraction of a second which are considered to reflect a human's hidden emotions [6]. Analyzing them has become a challenging problem in computer vision with different state-of-the-art approaches based on well-known computer vision methods such as local binary patterns (LBP), histogram of gradients (HOG) and optical flow (OF). However, MEs are composed of subtle motions which might be difficult to process with classical approaches. One possible solution is to analyze them by comparing the phase variations between images. Fleet and Jepson [8], Gautama and Van Hulle [9] initially proposed that spatio-temporally band-passed video provides a good approximation to the motion field.

However, its true potential became evident with motion magnification a method in which phase variations of subtle motions are amplified [33]. Specifically, the Riesz pyramid-based representation for video magnification [32] has shown to be a simple, adaptable and fast-processing method that can work in almost real-time. In addition to allowing the motion to be exaggerated, the intermediate representations produced by these methods can be used to directly analyze subtle motion. Indeed, it has been already used for subtle motion analysis [5] and ME spotting [4].

This paper proposes a framework based on this tool to extract multi-scale oriented phase variation features using the Riesz pyramid in order to model and classify MEs. There have been some other authors who have proposed to extract phase variations using a Riesz wavelet or transform as part of their ME classification scheme [20,25,26]. However, to our knowledge, there is no other method that uses the multi-scale phase variations as the main feature of their solution.

This paper is divided as follows: Section 2 presents a brief recapitulation of the state of the art in ME recognition. Section 3 introduces the reader to the monogenic signal components and the Riesz pyramid. Section 4 presents a novel way to extract motion features using the orientation and phase from the monogenic signal to model MEs. Section 5 analyzes the results of our experiments and compares them with state-of-the-art methods. Finally, in Section 6, our conclusions are presented.

☆ Editor: Prof. S. Sarkar

* Corresponding author.
E-mail address: olivier.alata@univ-st-etienne.fr (O. Alata).

https://doi.org/10.1016/j.patrec.2020.05.008
0167-8655/

2. Related work

The ME recognition frameworks can be divided into different feature representation families. The first one is composed of *LBP-based methods*. They use the intensity information of the image with the intention of describing facial features that appear temporally during any kind of facial expression. A great numbers of descriptors are based on local binary patterns (LBP) introduced in [39]. A 3D extension called LBP from Three Orthogonal Planes (LBP-TOP) takes a stack of consecutive frames as 3D volume and compute LBP over three orthogonal planes and has been used in several ME recognition frameworks [17]. Due to the popularity of LBP-TOP, a plethora of variations have emerged for feature extraction. For instance, some authors propose to take the average plane from each stack first, and then compute the LBP on the three average planes (MOP-LBP) [35]. Another variation called Spatio-Temporal LBP with Improved Integral Projection (STLBP-IIP) preserves the shape property of MEs and then enhances discrimination of the features for ME recognition [12,13]. Another proposal called Spatio-temporal Completed Local Quantized Pattern (STCLQP) extracts sign, magnitude and orientation information while creating a compact and discriminative codebook [14].

The second family is composed of *OF-based methods*. They use the distribution of the apparent velocities of objects in an image for motion representation. Some authors propose to derive the OF vectors to calculate the optical strain (OS) or non-rigid deformation for the analysis of MEs [21]. Inspired in the success of histogram features in the object recognition community, [3] proposed the Histogram of Oriented OF (HOOF) descriptor to model the distribution of OF during a video sequence. Another approach called the fuzzy histogram of oriented optical flow orientations (FHOFO) collects the motion directions into angular bins based on the fuzzy membership function [11]. Some similar descriptors like Main Directional Mean Optical flow (MDMO) [24] and Facial Dynamic Maps (FDM) [37] extract OF motion vectors from selected facial regions. Other authors have proposed to calculate which facial regions have high probabilities of movement (RHPM) and use them to filter the OF [1]. Another approach called Bi-Weighted oriented OF (BI-WOOF) is a variation of HOOF that uses optical strain as a weighting coefficient [23].

The third family is composed of *Deep learning methods*. These methods normally combine feature learning and classification in the same pipeline. In [15], the spatial features of micro-expressions at different expression-states are encoded using a CNN and then the temporal characteristics of the different expression-states of the micro-expressions are encoded using long short-term memory (LSTM) recurrent neural networks. In [18], a VGGNet trained for face recognition is finely tuned and adapted for ME recognition. Other authors propose to extract the ME features using a pre-trained network (ImageNET) and an evolutionary feature selection scheme to remove the irrelevant deep features [28]. In [16], a CNN is trained using both gray-scale images and optical flow as input (3D-FCNN).

There are certain proposals that do not really fit the previous families of methods. For instance, [19] uses spatio-temporal Gabor filters (ST-Gabor) to extract features at different scales and orientations by convolving a bank of oriented bandpass filters to an image sequence. Some approaches also consider combining appearance-based features and OF. For instance, [20] proposes to mix the BI-WOOF and the phase components from the monogenic signal obtained by a Riesz transform. Furthermore, [22] uses optical strain as weights for LBP-TOP (OSW-LBP-TOP).

In this paper, a complete framework for ME recognition based on the Riesz pyramid representation, a fast multi-scale approximation of the Riesz transform is presented. A new feature called Mean Oriented Riesz Feature (MORF), using the multi-scale oriented phase of the Riesz pyramid is introduced. Let us now recall the basis of the Riesz pyramid and its components.

3. Background

3.1. The monogenic signal and the Riesz transform

In signal analysis, a real valued 1-D signal can be represented as a complex valued signal. From this representation some useful information can be extracted such as the local amplitude and local phase. The analytical representation is composed of the original function and its Hilbert transform. In the case of 2-D signals (images), [7] proposed an isotropic generalization called the monogenic signal. The monogenic signal is a triple comprised of the original image and a quadrature pair produced by the Riesz transform (a 2-D steerable generalization of the Hilbert transform). This quadrature pair is 90 degrees phase-shifted with respect to the dominant orientation at every pixel [32], thus we can extract the local amplitude and the local phase variations in the direction of the dominant orientation from the monogenic signal. Let $I(\boldsymbol{x})$ be a 2D gray scale image of a spatial variable $\boldsymbol{x} = (x, y)^\intercal$, and let $F(\boldsymbol{\omega})$ be its frequency-domain representation found using the 2D Fourier transform, where $\boldsymbol{\omega} = (\omega_1, \omega_2)^\intercal$ is a two-dimensional frequency [2]. The two odd parts F_{R_1} and F_{R_2} of the monogenic signal are:

$$F_{R_l}(\boldsymbol{\omega}) = \begin{cases} i\frac{\omega_l}{\|\boldsymbol{\omega}\|}F(\boldsymbol{\omega}), & \boldsymbol{\omega} \neq 0 \\ 0, & \boldsymbol{\omega} = 0 \end{cases} \tag{1}$$

where $l = 1$ or 2 and $R_1(\boldsymbol{x})$ and $R_2(\boldsymbol{x})$ correspond to the image domain representation of F_{R_1} and F_{R_2} respectively.

3.2. The Riesz pyramid

The Riesz pyramid decomposes the image into multiple sub-bands, each of which corresponds to a different spatial scale, and then applies the Riesz transform of each sub-band (Fig. 1b). An ideal version of the Riesz pyramid could be built in the frequency domain using octave (or sub-octave) filters similar to the ones proposed in [33] and the frequency domain Riesz transform (Eq. (1)). However, it requires the use of costly Fourier transforms to be built. In order to make the Riesz pyramid faster, some adaptations need to be made. Firstly, instead of using the Fourier transform, the image is decomposed into non-oriented sub-bands using an invertible image pyramid such as the Laplacian pyramid. Secondly, an approximate Riesz transform can be defined by two finite difference filters, which is significantly more efficient to compute. Since most of the energy from the previously processed sub-bands are concentrated in a frequency band around $\|\boldsymbol{\omega}\| = \frac{\pi}{2}$, the Riesz transform can be approximated with the three tap finite difference filters $[0.5, 0, -0.5]$ and $[0.5, 0, -0.5]^\intercal$ [32].

3.3. Riesz pyramid coefficients

If a given image subband I is filtered using this method, the result is the pair of filter responses, $(R_1; R_2)$. The input I and the Riesz transform $(R_1; R_2)$ together form a triple (the monogenic signal) that can be converted to spherical coordinates (applying the method from [32]) to yield the local amplitude A, local orientation θ and local phase ϕ from the Riesz coefficients.

$$\begin{aligned} I &= A\cos(\phi) \\ R_1 &= A\sin(\phi)\cos(\theta) \\ R_2 &= A\sin(\phi)\sin(\theta) \end{aligned} \tag{2}$$

While Eq. (2) can be solved, both (A, ϕ, θ) and $(A, -\phi, \theta + \pi)$ are possible solutions. This predicament can be fixed by considering

$$\phi\cos(\theta), \phi\sin(\theta) \tag{3}$$

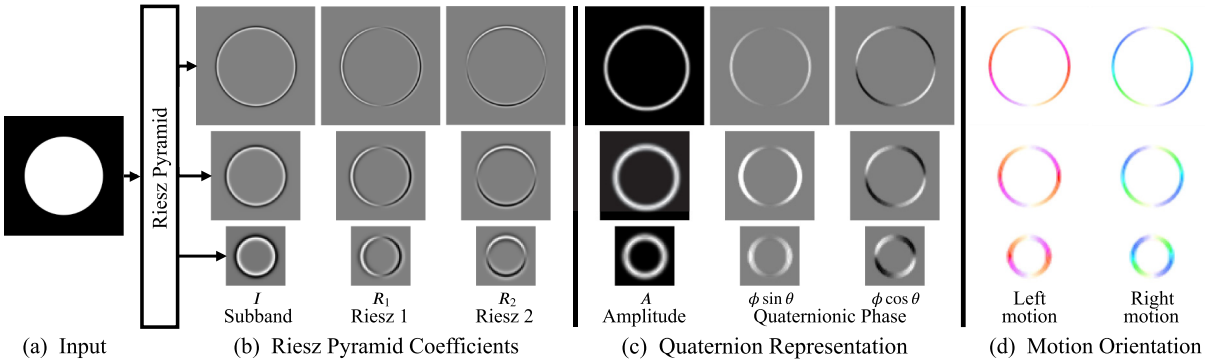


Fig. 1. Different representations of the Riesz pyramid. The input is a circle with a sharp edge (a). In (b), the input is decomposed into multiple spatial sub-bands using an invertible transform, and an approximate Riesz transform is taken of each band to form the Riesz pyramid. In (c), the Riesz pyramid coefficients are transformed into a quaternion. Then, for each subband the amplitude and the quaternionic phase can be extracted. In (d), we show the quaternionic phase difference between two consecutive frames for the input image translated one pixel to the left (Left motion) and one pixel to the right (Right motion).

which are invariant to this sign ambiguity. If the methods of [34] in which the Riesz pyramid coefficients are represented as a quaternion are applied:

$$\mathbf{r} = I + iR_1 + jR_2 \tag{4}$$

then, the previous equation can be rewritten using Eq. (2) as:

$$\mathbf{r} = A\cos(\phi) + iA\sin(\phi)\cos(\theta) + jA\sin(\phi)\sin(\theta) \tag{5}$$

Thus, the local amplitude A and the quaternionic phase ($\phi\cos(\theta)$, $\phi\sin(\theta)$) are computed as:

$$\begin{aligned} A &= \|\mathbf{r}\| \\ i\phi\cos(\theta) + j\phi\sin(\theta) &= \log(\mathbf{r}/\|\mathbf{r}\|) \end{aligned} \tag{6}$$

Furthermore, the quaternionic phase can be denoised by applying a temporal quaternionic filtering scheme. A complete explanation of the quaternionic operation used to extract the Riesz coefficients and filtering can be found in [34]. Fig. 1c shows the local amplitude and filtered quaternionic phase extracted from different levels of the Riesz pyramid applied to an input image translated one pixel to the left.

3.4. Local orientation of the quaternionic phase

One of the advantages of the monogenic signal is that, in the same way as the analytical signal, it preserves the split of identity. This means that, the local phase is invariant to changes of the local orientation, and the local orientation is invariant to changes of the local structure (which means that we can split them). If we can recover the correct local direction from the local orientation, we have an ideal split of identity with respect to energetic, geometric, and structural information of the signal. However, there are a few problems in estimating the correct local direction. The first one is that the estimation of local orientation is unstable if the local phase ϕ is close to 0. The second problem is that it's not possible to find an absolute estimation for the local direction but rather the relative estimation.

The main question becomes whether the orientation component of the quaternionic phase can be used to differentiate between opposing motions. Fig. 1d presents a circle with a sharp edge which is translated one pixel in any given direction. The resulting filtered quaternionic phase with pseudo-colors where the image saturation represents the phase ϕ component and the color hue represents the orientation θ component is presented. The areas of low amplitude are masked using the technique presented in [5] for better visualization. The image is translated one pixel to the left (left motion) and to the right (right motion).In that way motion in different directions can be represented (as evidenced by the different hues from the pseudo color image representations from the edges of Fig. 1d). The orientation of the quaternionic phase is not the same as that of the translation but rather it is perpendicular to the orientation of the edge. This means that the oriented quaternionic phase is affected by the aperture problem. Nevertheless, when comparing two opposing motions (left vs right), their respective oriented quaternionic phases are also opposite. This becomes important for ME recognition, where different MEs represent motions in different directions. For example, when analysing the eyebrow movement during an ME of surprise, the eyebrows rise. On the other hand during an ME of anger, the eyebrows are contracted (lowered).

4. Mean oriented Riesz features

This paper now proposes a descriptor to extract the oriented phase elements from the monogenic signal called **Mean Oriented Riesz Features (MORF)**. While Section 4.1 introduces the concept of the mean oriented Riesz image pair, Section 4.2 describes the implementation of our proposed descriptor.

4.1. Mean oriented Riesz image pair

In Section 3.4 it is shown how using a relative quaternion phase estimation motions from different directions can be differentiated. However, only the motion between two consecutive frames is analyzed. Considering the MEs are captured as video sequences of several frames, it is also necessary to analyse the temporal evolution of these motions before proposing an ME modelling scheme. It can be considered unnecessary to analyse the whole ME sequence but rather a shorter sequence from ME onset to ME apex (when the face goes from a neutral state to a state of peak expressiveness) because the spatial displacement of the facial muscles is more evident compared to the sequence that goes from ME apex to ME offset (the face goes from peak expressiveness to a neutral state). It is also necessary to deal with the high variance of the quaternionic phase ($\phi\cos(\theta)$, $\phi\sin(\theta)$) in areas of low local amplitude A [4]. proposes to crop a series of local ROIs and to mask them using the local amplitude from the Riesz pyramid in order to isolate areas of potential noise. Although this approach was effective for ME spotting, it ignores certain facial areas of low amplitude which might have some interesting information while an ME is taking place (such as the cheek areas).

Taking the aforementioned considerations into account, we propose to model the temporal evolution of the ME in two single images called the **mean oriented Riesz (MOR) image pair**. The filtered quaternionic phase of an ME sequence is simply calculated

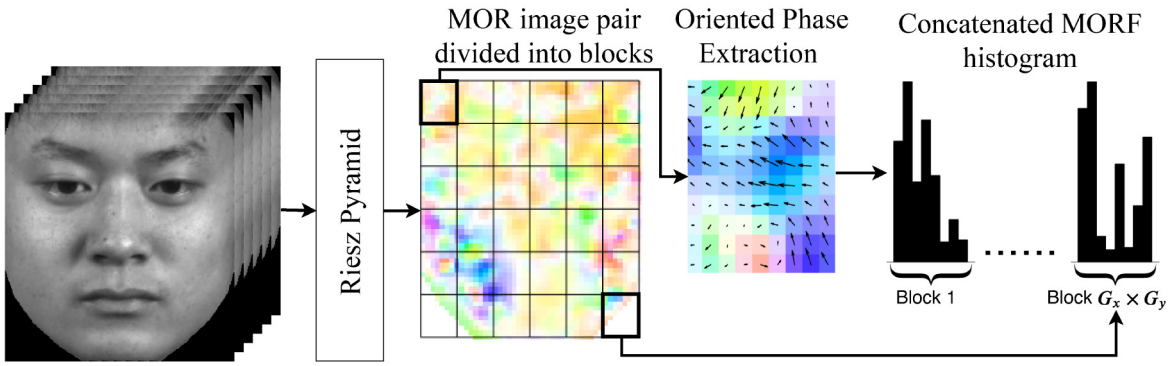


Fig. 2. Extraction of the MORF descriptor.

from onset to apex and then, for each pixel, the results are averaged through the time axis :

$$\overline{\phi\lambda(\theta)} = \frac{1}{f_a - f_o + 1} \sum_{t=f_o}^{f_a} \phi_t \lambda(\theta_t) \tag{7}$$

with f_o and f_a, the frame the onset begins and the frame of the apex respectively, and λ either cos or sin. The main intuition is that by temporally averaging the filtered quaternionic phase, the real motion of each pixel is modelled in a single orientation and magnitude while reducing the effect of wrongfully detected motion due to noise.

4.2. MORF extraction

To begin with the face is detected in the first frame [31], then, an active appearance model (AAM) [30] is used to detect a set of facial landmarks. Next, certain facial landmarks which will not move during facial expressions are selected (the inner corners of the eyes and the lower point of the nose between the nostrils are selected). These points are tracked using the KLT algorithm [29] and a cropped face image sequence is obtained (the area outside the face border is masked - see Fig. 2). The Riesz pyramid is applied to obtain the quaternionic phase (Section 3.3) and it is filtered with the method described in [4]. Then, Eq. (7) is then used to obtain the MOR image pair (see center of Fig. 2).

The face is divided into a grid of equally sized non-overlapping rectangle areas. As can be seen in Fig. 2, each pixel in the image pair represents a motion vector with a magnitude and angle. They can be extracted from the oriented phase by:

$$\overline{\phi_R} = \sqrt[2]{(\overline{\phi_R \cos(\theta_R)})^2 + (\overline{\phi_R \sin(\theta_R)})^2} \tag{8}$$

$$\overline{\theta_R} = \arctan\left(\frac{\overline{\phi_R \sin(\theta_R)}}{\overline{\phi_R \cos(\theta_R)}}\right) \tag{9}$$

where $\overline{\phi_R}$ is a matrix containing the phase of every pixel, $\overline{\theta_R}$ is a matrix containing the dominant orientation of every pixel and R corresponds to the level of the Riesz pyramid, the oriented phase is being extracted from. we are extracting the oriented phase from. The next step is to create the histogram of oriented phase for each one of the rectangular blocks. For each pixel, a bin is selected based on the orientation θ and a weighted vote is cast based on the value of the phase ϕ. The final histogram is the concatenation of all the histograms (Fig. 2).

The MORF descriptor depends on three parameters: G which determines the grid division of ($[G_x, G_y]$) ROIs, O which determines the number of orientations bins of the descriptor and R which determines the level of the Riesz pyramid going to be extracted. Thus $\text{MORF}_{G,O,R}$ produces a feature vector of $G_x \times G_y \times O$ length.

5. Experimental results

5.1. Datasets

For our experimentation, two spontaneously elicited ME databases are selected. First, the SMIC database [17] consists of 164 spontaneous facial MEs image sequences from 16 subjects. The full version of SMIC contains three datasets: the SMIC-HS dataset recorded by a high speed camera at 100 fps; the SMIC-VIS dataset recorded by a color camera at 25 fps; and the SMIC-NIR dataset recorded by a near infrared camera at 25 fps (all with a spatial resolution of 640 × 480). Ground truth annotations provide the frame numbers indicating the onset and offset frames. The MEs are labeled into three emotion classes: positive, surprise and negative emotions. For our experimentation it is decided to use only the SMIC-HS dataset.

Secondly, the CASME II [38] database consists of 247 spontaneous facial MEs image sequences from 26 subjects. They were recorded using a high speed camera at 200 fps and spatial resolution of 640 × 480. Ground truth annotations not only provide the frame numbers indicating the onset and offset but also the apex frames (the moment when the ME is at its highest intensity). The MEs are labeled into five classes: happiness, surprise, disgust, repression and others.

5.2. Parameter analysis

To evaluate the impact of the different parameters on our system, our proposed framework is tested while its parameter values are varied. It is decided to evaluate the following parameters: the pyramid level (from 2nd to 4th level[1]), the grid division ($G_x = [4, \ldots, 10]$ and $G_y = [6, \ldots, 12]$) and orientation binning ($O = [4, \ldots, 10]$). For classification, a multiclass-SVM (one-vs-all) using a polynomial kernel of degree three is trained. The hyperparameters are tuned by an exhaustive grid search process. The results are tested by a Leave-One-Subject-Out (LOSO) cross-validation.

The effect of the level of the Riesz pyramid is shown (Fig. 3). In both the SMIC-HS dataset and the CASME II dataset extracting the phase values from the 2nd level of the pyramid results in a higher median accuracy. Furthermore, the effect of the angle division for the MORF descriptor is shown (Fig. 4). In both cases, the best angle division for the level that yields the best results (2nd) is between 6 and 7 divisions. The effect of the grid division for the MORF descriptor is also shown (Fig. 5). In the SMIC-HS dataset, the best grid division is between 7 and 10 for G_x and 6 to 8 for G_y. In the CASME II dataset, the best grid division is between 7 and 9 for

[1] since the first level has the information of the highest frequency sub-band and seems to carry an important amount of undesired noise.

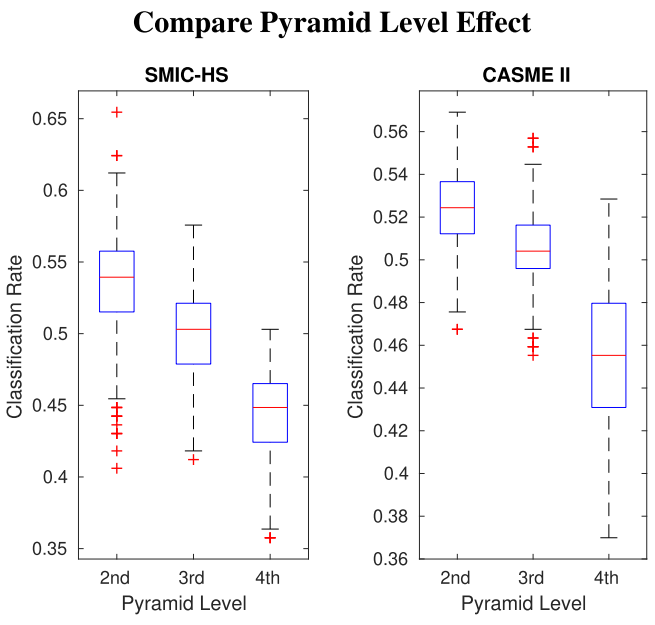


Fig. 3. Riesz pyramid level evaluation.

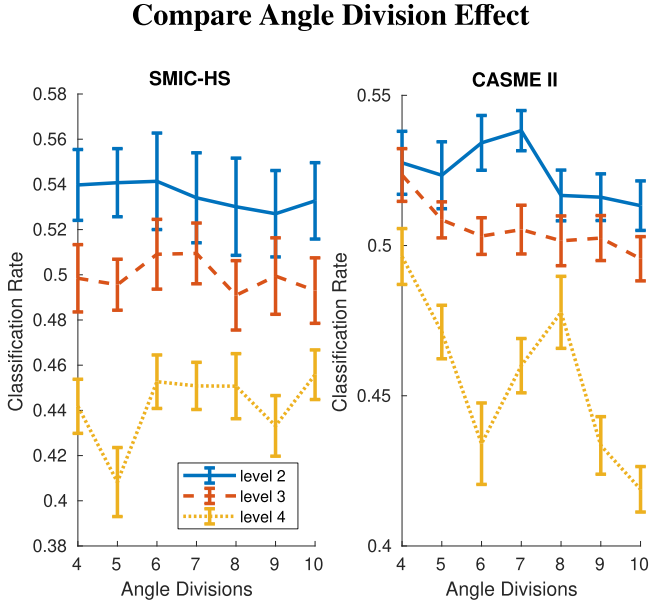
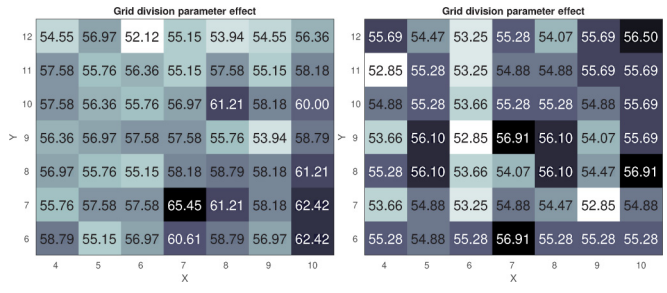


Fig. 4. Orientation binning parameter evaluation.



(a) SMIC-HS (b) CASME II

Fig. 5. Grid division parameter evaluation.

G_x and 8 to 10 for G_y (However there are some peaks when G_x is 10).

5.3. MORF variations

Some variations to the MORF descriptor are tested by combining different data and methodologies. Firstly, the results of two or more MORF histograms are merged from different levels of the Riesz pyramid (**F-MORF**). The idea is to use the oriented phase calculated from different sub-bands to potentially complement the information for modelling an ME. Secondly, it is decided to use the amplification process of [33] in which the subtle phase changes are multiplied by a scalar without amplifying the noise (**A-MORF**). This is done by multipling the quaternionic filtered phase ($\phi\cos(\theta)$, $\phi\sin(\theta)$) by a magnification factor (α), then, after performing a quaternion exponentiation on it, the amplified quaternioninc phase is extracted:

$$\sin(\alpha\phi)\cos(\theta), \sin(\alpha\phi)\sin(\theta) \tag{10}$$

Finally, this representation is used to calculate the MOR image pair and extract the MORF descriptor. The pyramid levels can also be both merged and amplified to obtain **AF-MORF**. The recognition performance of the proposed method is measured using both recognition accuracy and F-measure. The results are shown in Tables 1 and 2.

For SMIC-HS better results are obtained using MORF and CASME II using FA-MORF. This discrepancy comes from the differences of the datasets. The subjects in CASME II were at a closer distance to the camera during video recording compared to SMIC, thus the captured faces had a bigger resolution which result in a shift of the ME motion to low frequencies. This might explain why better results can be obtained using the 2nd level of the Riesz pyramid in the SMIC-HS dataset but in the CASME II dataset similar results are obtained both in the 2nd and 3rd levels. Consequently merging these two levels yields better results for CASME II but not for SMIC-HS. Furthermore the CASME II videos were captured with a camera twice as fast as those in SMIC-HS. This means that the phase differences between frames in the CASME II database are smaller and can potentially be improved by amplification which might explain why A-MORF and FA-MORF perform better in the CASME II dataset.

5.4. State-of-the-art comparison

Our classification results are compared with some representative methods from the state of the art[2] in Table 3. For the LBP-based methods, it can be seen how they have improved from the baseline proposed by Li et al. [17]. Each method extracts spatio-temporal information by creating a different code-book based on the intensity difference between a pixel and its 3D neighborhood. One reason why these methods tend to do better in the dataset CASME II is that, as previously mentioned, its images have a bigger resolution which means extracting better textured information of the MEs. For the OF-based methods, each method extracts motion information from OF. The best results come from Bi-WOOF [21] and OF Maps [1] for calculating the motion between onset and apex frames (instead of calculating the motion between consecutive frames). Furthermore, OF Maps extract the coherent movement on the face in different locations and use it to filter residual motion vectors (noise). For the deep learning methods, each method either trains or tunes a pre-trained convolutional neural network (CNN) for extracting features and classification. Although these methods are becoming more widely used in classification

[2] For a more thorough comparison, we refer the reader to state-of-the-art surveys presented in [27] and [10].

Table 1
ME classification for SMIC HS in terms of accuracy and F-measure.

SMIC-HS							
Feature	Pyramid Level	Non Amplified		A-MORF			
				$\alpha = 5$		$\alpha = 10$	
		Acc	F1-mea	Acc	F1-mea	Acc	F1-mea
MORF	2	**65.45%**	**0.6466**	58.79%	0.5764	53.33%	0.5266
	3	57.58%	0.5733	61.21%	0.6059	58.18%	0.5822
	4	50.30%	0.5043	50.30%	0.5054	50.91%	0.5158
F-MORF	2&3	58.79%	0.5937	60.00%	0.5999	59.39%	0.5895
	3&4	54.55%	0.5507	56.36%	0.5695	58.18%	0.5798
	2&3&4	58.79%	0.5913	58.79%	0.5794	59.39%	0.5868

Table 2
ME classification for CASME II in terms of accuracy and F-measure.

CASME II							
Feature	Pyramid Level	Non Amplified		A-MORF			
				$\alpha = 5$		$\alpha = 10$	
		Acc	F1-mea	Acc	F1-mea	Acc	F1-mea
MORF	2	56.91%	0.5878	58.94%	0.6045	58.54%	0.5983
	3	55.69%	0.5545	57.72%	0.5779	58.94%	0.5762
	4	52.85%	0.5110	53.66%	0.5183	54.88%	0.5355
F-MORF	2&3	58.54%	0.5923	**62.20%**	**0.6304**	62.20%	0.6171
	3&4	59.35%	0.5871	56.10%	0.5549	56.50%	0.5576
	2&3&4	59.76%	0.6012	58.54%	0.5850	58.13%	0.5827

Table 3
Comparison of micro-expression recognition performance in terms of accuracy and F-measure for feature-extraction state-of-the-art methods.

Micro-Expression Classification Methods						
Family	Method		Accuracy		F-measure	
	Features	Paper	SMIC HS	CASME II	SMIC HS	CASME II
LBP based	LBP-TOP	[17]	48.78%	-	-	-
	LBP-MOP	[35]	50.61%	45.75%	-	-
	STLBP-IP	[13]	57.93%	59.51%	0.58	0.57[3]
	STCQLP	[14]	58.39%	64.02%	0.6381	0.5836
	Di-STLBP-IP	[12]	63.41%	64.78%	-	-
OF based	OS	[21]	53.56%	-	-	-
	FDM	[37]	54.88%	41.96%	0.538	0.4053
	HFOFO	[11]	51.83%	56.64%	0.5243	0.5248
	Bi-WOOF	[21]	62.20%	58.85%	0.62	0.61
	OF Maps	[1]	-	**65.35%**	-	-
Deep learning	Imagenet	[28]	53.60%	47.30%	-	-
	3D-FCNN	[16]	55.49%	59.11%	-	-
	CNN + LSTM	[15]	-	60.98%	-	-
	VGGNet	[18]	-	63.30%	-	-
Others	Monogenic + LBP-TOP	[26]	-	-	0.44	0.41
	Riesz Wavelet + LBP-TOP	[25]	-	-	-	0.43
	OSW-LBP-TOP	[22]	53.66%	42.00%	0.54	0.38
	ST-Gabor	[19]	54.47%	55.28%	-	-
	Bi-WOOF + Riesz Phase	[20]	**68.29%**	62.55%	**0.67**	**0.65**
Our method	MORF		**65.45%**	56.91%	**0.6466**	0.5878
	F-MORF		58.79%	59.76%	0.5937	0.6012
	FA-MORF		61.21%	**62.20%**	0.6059	**0.6304**

problems, they also struggle to obtain good results when dealing with small datasets. And although they obtain good results with the CASME II, they avoid using a smaller dataset such as SMIC-HS.

Our proposed approach does not yield the best possible results. It is worth noting that, while other authors have been able to obtain better results starting from a baseline method (like LBP-TOP in the case of the LBP-based methods and HOOF in the case of OF-based methods) and propose an improved method by changing the way they extract features and how to code them (quantizing the values, developing a weighted histogram, pre-selecting regions of interest and/or frames to process, etc.), our method uses a more basic featuring extracting method. It can be imagined that applying a more sophisticated method to extract information from Riesz phase variations might result in better results. All things considered, our method is still able to surpass several descriptors in both datasets.

Our method outperforms other Riesz based methods like [25,26] by approximately 20%. However, Liong and Wong [20]

remains as the best Riesz phase-based method. It's worth noting that the performance of the previous version of this method, Bi-WOOF [21], is greatly improved by adding Riesz phase difference between onset and apex frames [20]. This implies that our results could be improved by complementing our phase-based features with other motion and texture features.

One cause of error might come from the limitations of the local orientation of the monogenic signal. The local orientation θ represents the dominant direction in the image at any given point. This representation comes from the formulation of the monogenic signal which assumes images as intrinsically one-dimensional signals. This means that the monogenic signal is useful for modelling image features such as edges and lines that have variation in one direction only, but cannot model image features such as corners that have variation in two directions (intrinsically 2D signals) [36].[3]

6. Conclusion

A facial micro-expression recognition method based on the quaternionic oriented phase representation of the multi-scale monogenic signals is proposed. Phase variations are quickly extracted from a video using an approximate Riesz transform called the Riesz pyramid. The temporal evolution of a micro-expression is modeled as an image pair that contains the mean oriented phase component of the monogenic signal which aims to reduce the effects of image noise. Furthermore, this model is extended into an easily-adaptable and low-dimensional feature descriptor which can also contain the amplification of the oriented phase or concatenate the multi-scale oriented phase representation. Our method achieves an accuracy of 65.45% and an F-score of 0.6466 for the SMIC-HS dataset and an accuracy of 62.20% and an F-score of 0.6304 for the CASME II dataset. These are the best results for methods focused on Riesz transform based features. It is also competitive against methods based on widely researched features such as LBP and OF and even against deep learning methods. Furthermore, it obtains the best results in the SMIC-HS for methods focused in one single type of feature.

Our Riesz pyramid-based method has shown itself to be a powerful tool for ME recognition. Adopting a more sophisticated feature extraction and codification, along with complementing the Riesz phase variations with motion or texture information, could be used to create an improved ME analysis technique in the future. In addition, since there is already a method that uses a similar basis for ME spotting [4], both methods can be merged for an integrated Riesz phase-based spotting and recognition framework.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. Allaert, I.M. Bilasco, C. Djeraba, Consistent optical flow maps for full and micro facial expression recognition, in: Proc. VISIGRAPP (5: VISAPP) INSTICC, SciTePress, 2017, pp. 235–242.
[2] C. P. Bridge, Introduction to the monogenic signal, 2017. abs/1703.09199.
[3] R. Chaudhry, A. Ravich, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: IEEE Conf. on Computer Vision and Pattern Recognition CVPR, 2009.
[4] C.A. Duque, O. Alata, R. Emonet, A.C. Legrand, H. Konik, Micro-expression spotting using the Riesz pyramid, in: 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018, pp. 66–74.
[5] C.A. Duque, O. Alata, R. Emonet, A.C. Legrand, H. Konik, Subtle motion analysis and spotting using the Riesz pyramid, in: Proc. VISIGRAPP (5: VISAPP), INSTICC, SciTePress, 2018, pp. 446–454.
[6] P. Ekman, E.L. Rosenberg, Smiles when Lying, in: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, 2005, pp. 201–216.
[7] M. Felsberg, G. Sommer, The monogenic signal, IEEE Trans. Signal Process. 49 (2001) 3136–3144.
[8] D.J. Fleet, A.D. Jepson, Computation of component image velocity from local phase information, Int. J. Comput. Vis. 5 (1990) 77–104, doi:10.1007/BF00056772.
[9] T. Gautama, M. Van Hulle, A phase-based approach to the estimation of the optical flow field using spatial filtering, IEEE Trans. Neural Netw. 13 (2002) 1127–1136, doi:10.1109/TNN.2002.1031944.
[10] K. Goh, C. Ng, L. Lim, U. Sheikh, Micro-expression recognition: an updated review of current trends, challenges and solutions, Vis. Comput. (2018), doi:10.1007/s00371-018-1607-6.
[11] S.L. Happy, A. Routray, Fuzzy histogram of optical flow orientations for micro–expression recognition, IEEE Trans. Affect. Comput. (2018) 394–406.
[12] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikainen, Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection, 2016. https://arxiv.org/abs/1608.02255.
[13] X. Huang, S. Wang, G. Zhao, M. Piteikinen, Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection, in: 2015 IEEE Int. Conf. on Computer Vision Workshop (ICCVW), 2015, pp. 1–9.
[14] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikinen, Spontaneous facial micro–expression analysis using spatiotemporal completed local quantized patterns, Neurocomput. 175 (2016) 564–578.
[15] D.H. Kim, W.J. Baddar, Y.M. Ro, Micro-expression recognition with expression-state constrained spatio-temporal feature representations, in: Proceedings of the 2016 ACM on Multimedia Conf., 2016, pp. 382–386.
[16] J. Li, Y. Wang, J. See, W. Liu, Micro-expression recognition based on 3D flow convolutional neural network, Pattern Anal. Appl. (2018).
[17] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikainen, A spontaneous micro-expression database: inducement collection and baseline, in: 2013 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6.
[18] Y. Li, X. Huang, G. Zhao, Can micro-expression be recognized based on single apex frame? in: 2018 25th IEEE Int. Conf. on Image Processing (ICIP), 2018, pp. 3094–3098.
[19] C. Lin, F. Long, J. Huang, J. Li, Micro-expression recognition based on spatiotemporal gabor filters, in: 2018 8th Int. Conf. on Information Science and Technology (ICIST), 2018, pp. 487–491.
[20] S. Liong, K. Wong, Micro-expression recognition using apex frame with phase information, in: 2017 Asia-Pacific Signal and Information Processing Assoc. Annual Summit and Conf. (APSIPA ASC), 2017, pp. 534–537.
[21] S.T. Liong, R.C.W. Phan, J. See, Y.H. Oh, K. Wong, Optical strain based recognition of subtle emotions, in: 2014 Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS), 2014, pp. 180–184.
[22] S.T. Liong, J. See, R.C.W. Phan, A.C.L. Ngo, Y.H. Oh, K. Wong, Subtle expression recognition using optical strain weighted features, in: C.V. Jawahar, S. Shan (Eds.), Computer Vision - ACCV 2014 Workshops, LNCS, Springer Int. Publishing, 2014, pp. 644–657.
[23] S.T. Liong, J. See, K. Wong, R.C.W. Phan, Less is more: micro-expression recognition from video using apex frame, Signal Process. Image Commun. 62 (2018) 82–92.
[24] Y.J. Liu, J.K. Zhang, W.J. Yan, S.J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Trans. Affect. Comput. 7 (2016) 299–310.
[25] Y.H. Oh, A.C. Le Ngo, R.C.W. Phari, J. See, H.C. Ling, Intrinsic two-dimensional local structures for micro-expression recognition, in: 2016 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2016, pp. 1851–1855, doi:10.1109/ICASSP.2016.7471997. ISSN: 2379-190X
[26] Y.H. Oh, A.C.L. Ngo, J. See, S.T. Liong, R.C.W. Phan, H.C. Ling, Monogenic Riesz wavelet representation for micro-expression recognition, in: 2015 IEEE Int. Conf. on Digital Signal Processing (DSP), 2015, pp. 1237–1241.
[27] Y.H. Oh, J. See, A.C. Le Ngo, R.C.W. Phan, V.M. Baskaran, A survey of automatic facial micro-expression analysis: databases methods and challenges, Front. Psychol. 9 (2018).
[28] D. Patel, X. Hong, G. Zhao, Selective deep features for micro-expression recognition, in: 2016 23rd Int. Conf. on Pattern Recognition (ICPR), 2016, pp. 2258–2263.
[29] C. Tomasi, T. Kanade, Detection and Tracking of Point Features, 1991. Technical Report.
[30] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, in: 2013 IEEE Int. Conf. on Computer Vision (ICCV), 2013, pp. 593–600.
[31] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition 2001. CVPR 2001, 2001.
[32] N. Wadhwa, M. Rubinstein, F. Durand, W. Freeman, Riesz pyramids for fast phase-based video magnification, in: 2014 IEEE Int. Conf. on Computational Photography (ICCP), 2014, pp. 1–10.
[33] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, Phase-based video motion processing, ACM Trans. Graph. 32 (2013) 80:1–80:10.

[3] Values extracted from [27].

[34] N. Wadhwa, M. Rubinstein, F. Durand, W.T. Freeman, Quaternionic Representation of the Riesz Pyramid for Video Magnification, 2014. Technical Report.
[35] Y. Wang, J. See, R.C.W. Phan, Y.H. Oh, Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition, PLoS ONE 10 (2015) 1–20.
[36] L. Wietzke, G. Sommer, O. Fleischmann, The geometry of 2d image signals, in: 2009 IEEE Conf. on Computer Vision and Pattern Recognition, 2009, pp. 1690–1697.
[37] F. Xu, J. Zhang, J.Z. Wang, Microexpression identification and categorization using a facial dynamics map, IEEE Trans. Affect. Comput. PP (2017) 254–267.
[38] W.J. Yan, X. Li, S.J. Wang, G. Zhao, Y.J. Liu, Y.H. Chen, X. Fu, CASME II: an improved spontaneous micro-expression database and the baseline evaluation, PLoS ONE 9 (2014).
[39] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach.Intell. 29 (2007) 915–928.