Micro-Expression Spotting using the Riesz Pyramid

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

Micro-Expression Spotting using the Riesz Pyramid

Carlos Arango Duque Olivier Alata Rémi Emonet Anne-Claire Legrand Hubert Konik
Univ Lyon, UJM-Saint-Etienne, CNRS, IOGS, Laboratoire Hubert Curien UMR 5516
F-42023, Saint-Etienne, France
carlos.arango.duque@univ-st-etienne.fr

Abstract

Facial micro-expressions (MEs) are fast and involuntary facial expressions which reveal people hidden emotions. ME spotting refers to the process of finding the temporal locations of rapid facial movements from a video sequence. However, detecting these events is difficult due to their short durations and low intensities. Also, a distinction must be made between MEs and eye-related movements (blinking, eye-gaze change, etc). Taking inspiration from video magnification techniques, we design a workflow for automatically spotting MEs based on the Riesz pyramid. In addition, we propose a filtering and masking scheme that segment motions of interest without producing undesired artifacts or delays. Furthermore, the system is able to differentiate between MEs and eye movements. Experiments are carried out on two databases containing videos of spontaneous micro-expressions. Finally, we show that our method is able to outperform other methods from the state of the art in this challenging task.

1. Introduction

Although the study of automatic emotion recognition based on facial micro-expressions (MEs) have gained momentum in the last couple of years, much of this work has focused on classifying emotions [3, 5]. However, most of these papers use temporally and manually segmented videos (that are known to contain MEs).

Considering that a lot of real-life applications require to detect when an event takes place, spotting MEs becomes a primary step for a fully automated facial expression recognition (AFER) system. However, this is a challenging task because MEs are quick and subtle facial movements of low-spatial amplitude which are very difficult to detect by the human eye.

Some authors propose to spot MEs by analyzing the difference between appearance-based features of sequential frames and searching the frame in which the peak facial change occurs [7, 21]. In addition, [8] provides not only temporal information but also spatial information about the movements in the face. In [11], the authors propose to calculate the spatio-temporal strain (the deformation incurred on the facial skin during non-rigid motion) using dense optical flow.

Another possible approach would be to use *Eulerian* motion amplification techniques [17, 19] in order to enhance the spotting process. These techniques have already been used in the past for ME recognition [9, 10]. Although it has been used as a pre-processing step, it has been shown to degrade the spotting accuracy [5]. Our careful examination showed that intermediate representations produced by the latter methods can be used as proxies for motion.

In this paper, we propose a method which is able to spot MEs in a video by analyzing the phase variations between frames obtained from the Riesz Pyramid. The proposed method does not require training or pre-labeling of the videos. This paper is organized as follows. Sec. 2 serves as an introduction to the theoretical material to understand the Riesz Pyramid outputs, its quaternionic representation and filtering. Sec. 3 describes our proposed methodology and contributions. Sec. 4 describes our experiments, results and discussion. Finally, Sec. 5 presents our conclusions.

2. Background

Eulerian motion magnification is a family of techniques that amplifies subtle motion in videos. They are inspired by the *Eulerian* perspective, in reference to fluid dynamics where the properties of a voxel of fluid, such as pressure and velocity, evolve over time. The first of these techniques [19] exaggerates motion by amplifying temporal color changes at fixed positions. However, this method can significantly amplify noise when the magnification factor is increased. [17] proposed a method to amplify the phase variation over time within each image subband using the steerable pyramid (an over-complete transform that decomposes an image according to spatial scale, orientation, and position). However, the main disadvantage of this method comes from the complex steerable pyramids which are very overcomplete and costly to construct. Later, a new method based on

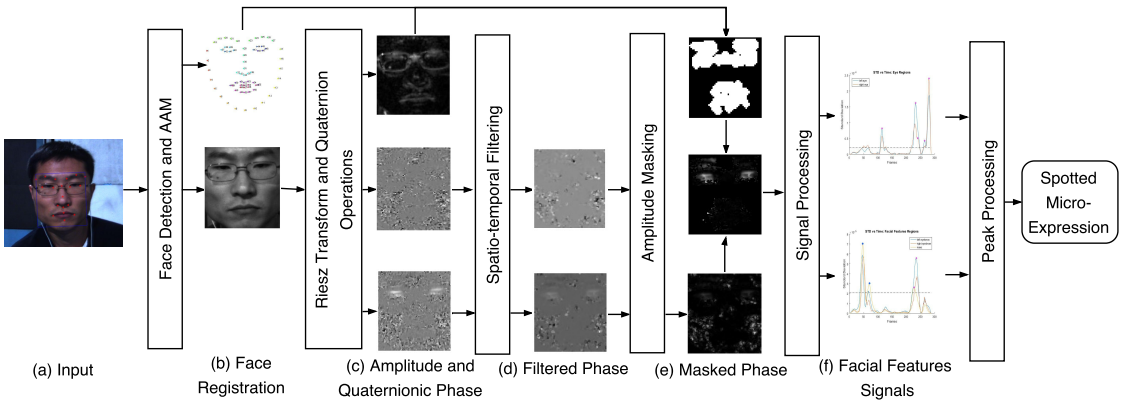


Figure 1: ME spotting framework. From the input video (a) the face is detected and cropped. Facial landmarks locations are also extracted (b). Each frame in the cropped video sequence is processed with the Riesz Pyramid to obtain the local amplitude and the quaternionic phase (c). We apply an improved spatio-temporal filter to the quaternionic phase (d). Then we use local amplitude and facial landmarks to mask relevant areas in the quaternionic phase (e). The phase is processed into a series of 1-D signals (f) and we detect and classify the resulting peaks to spot the MEs.

the Riesz pyramid was proposed which produced motion-magnified videos of comparable quality to the previous one, but the videos can be processed in one quarter of the time, making it more suitable for real-time or online processing applications [16, 18].

Although, video magnification is a powerful tool to magnify subtle motions in videos, it does not precisely indicate the moment when these motions take place. However, the filtered quaternionic phase difference obtained during the Riesz magnification approach [18] seems to be a good proxy for motion, thus it could potentially be used for temporally segmenting subtle motions. In this section, we introduce the Riesz pyramid, its quaternionic representation and quaternionic filtering.

2.1. Riesz Monogenic Signal

The Riesz pyramid is constructed by first breaking the input image into non-oriented subbands using an efficient, invertible replacement for the Laplacian pyramid, and then taking an approximate Riesz transform of each band. The key insight into why this representation can be used for motion analysis is that the Riesz transform is a steerable Hilbert transformer and allows us to compute a quadrature pair that is 90 degrees out of phase with respect to the dominant orientation at every pixel. This allows us to phase-shift and translate image features only in the direction of the dominant orientation at every pixel.

Following [14], in two dimensions, the Riesz transform is a pair of filters with transfer functions

$$-i\frac{\omega_x}{\|\overrightarrow{\omega}\|}, -i\frac{\omega_y}{\|\overrightarrow{\omega}\|} \tag{1}$$

with $\overrightarrow{\omega} = [\omega_x, \omega_y]$ being the signal dimensions in the frequency domain. If we filter a given image subband I using Eq. 1, the result is the pair of filter responses, $(R_1; R_2)$. The input I and Riesz transform $(R_1; R_2)$ together form a triple (the monogenic signal) that can be converted to spherical coordinates to yield the local amplitude A, local orientation θ and local phase ϕ using the equations

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

2.2. Quaternion Representation of Riesz Pyramid

The Riesz pyramid coefficient triplet $(I; R_1; R_2)$ can be represented as a quaternion $\mathbf{r}$ with the original subband I being the real part and the two Riesz transform components $(R_1; R_2)$ being the imaginary i and j components of the quaternion.

$$\mathbf{r} = I + iR_1 + jR_2 \tag{3}$$

The previous equation can be rewritten using (2) as:

$$\mathbf{r} = A\cos(\phi) + iA\sin(\phi)\cos(\theta) + jA\sin(\phi)\sin(\theta) \tag{4}$$

However, the decomposition proposed by (4) is not unique. That means that both (A, ϕ, θ) and $(A, -\phi, \theta + \pi)$ are possible solutions. This can be solved if we consider

$$\phi\cos(\theta), \phi\sin(\theta) \tag{5}$$

which are invariant to this sign ambiguity. If the Riesz pyramid coefficients are viewed as a quaternion, then Eq. 5 is the quaternion logarithm of the normalized coefficient[1]. Thus,

[1] An extended review of the quaternionic representation including complex exponentiation and logarithms can be found in [18].

the local amplitude A and quaternionic phase defined in Eq. 5 are computed:

$$A = \|\mathbf{r}\| \tag{6}$$

$$i\phi\cos(\theta) + j\phi\sin(\theta) = \log(\mathbf{r}/\|\mathbf{r}\|) \tag{7}$$

In previous *Eulerian* motion amplification papers, motions of interest were isolated and denoised with temporal filters. However, the quaternionic phase cannot be naively filtered since it is a wrapped quantity. Therefore, a technique based on [4] is used to filter a sequence of unit quaternions by first unwrapping the quaternionic phases in time and then using a linear time invariant (LTI) filter [18] in order to isolate motions of interest in the quaternionic phase.

Furthermore, the signal to noise ratio (SNR) of the phase signal can be increased by spatially denoising each frame with an amplitude-weighted spatial blur using a Gaussian kernel with standard deviation ρ on the i and j components of the temporally filtered signal. In the following section, we show how the phase signal can be used to spot MEs.

3. Proposed Framework

Our proposed algorithm goes as follows: First we detect and track stable facial landmarks through the video. Secondly, we use the Riesz Pyramid to calculate the amplitude and quaternionic phase of the stabilized face images and we implement a proper spatio-temporal filtering scheme which can enhance motions of interest without producing delays or undesired artifacts. Thirdly, we isolate areas of potential subtle motions based on amplitude and facial landmarks. Finally, we measure the dissimilarities of quaternionic phases over time and transform it into a series of 1-D signals, which are used to estimate the apex frame of the MEs. A graphic representation of our framework can be seen in Fig. 1.

3.1. Face Detection and Registration

We start by detecting the face in the first frame using the cascade object detector of Viola and Jones [15]. Then, we use the active appearance model (AAM) proposed by Tzimiropoulos and Pantic [13] to detect a set of facial landmarks. Next, we select certain facial landmarks which won't move during facial expressions (we selected the inner corner of the eyes and the lower point of the nose between the nostrils). We track these points using the Kanade-Lucas-Tomasi (KLT) algorithm [12] and use them to realign the face over time. Finally, we use these landmarks to crop the aligned facial region of interest for each frame in the video (See Fig. 1(b)).

3.2. Riesz Transform and Filtering

For the sequence of cropped images obtained in 3.1 of N frames, we perform the process described in Sec. 2.1 and Sec. 2.2 for each frame $n \in [1, \ldots, N]$. However, depending on the spatial resolution of the cropped faces and the speed of the image acquisition system, the local phase computed from certain frequency subbands will contribute more to the construction of a certain motion compared to the ones from other levels. In other words, not all the levels of the pyramid are able to provide useful information about the subtle motion. Thus, we must test the quaternionic phase from different subbands (pyramid level) and select the level which better represents the subtle motion we want to detect.

We then obtain both local amplitude A_n and quaternionic phase $(\phi_n\cos(\theta), \phi_n\sin(\theta))$ from the selected subband. We apply the process described in Sec. 2.2 to obtain the filtered quaternionic phase ϕ (Sec. 2.2). However, since we are aiming to detect any significant quaternionic phase shifts between frames and to compensate for the cumulative sum made done in the quaternionic phase unwrapping [18], we calculate the difference of two consecutive filtered quaternionic phases:

$$\Delta\phi_n\mathbf{u} = \phi_n\,\mathbf{u} - \phi_{n-1}\,\mathbf{u} \tag{8}$$

where $\mathbf{u} = i\cos\theta + j\sin\theta$.

We also must consider what kind of temporal filter we must implement for our application. The previous works in *Eulerian* motion magnification (See Sec. 2.2) have given their users freedom to choose any temporal filtering method available. However, since we require to pinpoint the exact moment when a subtle motion is detected we cannot use traditional causal filters which delay the signal response (for example, Fig. 2c shows the delay caused by the filtering scheme used in [16]). Therefore, we propose to use a digital non-causal zero-phase finite impulse response (FIR) filter.

$$\Phi_n\mathbf{u} = b_0\Delta\phi_n\mathbf{u} + \sum_{k=1}^{p} b_k(\Delta\phi_{n+k}\mathbf{u} + \Delta\phi_{n-k}\mathbf{u}) \tag{9}$$

where p is an even number and b_k is a coefficient of a FIR filter of length $2p+1$. One limitation of this method is that non-causal filters requires to use the previous and following p frames from the current frame (therefore for online applications there must be a delay of at least p frames).

Another element to consider is that *Eulerian* amplification methods are tailored for a particular task. These methods focus on amplifying subtle periodical movements (such as human breathing, the vibration of an engine, the oscillations of a guitar string, etc) by temporally band-passing some potential movements. However, these methods do not consider subtle non periodical movements (such as blinking or facial MEs). The latter type of motion, when band-passed, creates some large oscillations near the beginning and the end of the subtle motion (Fig. 2d) as stated by Gibbs phenomenon. Therefore, we decided to use low-pass filtering for this type of signals (Fig. 2e).

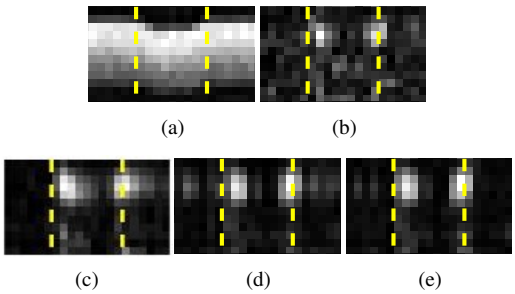


Figure 2: A comparison of different filter responses for subtle motion detection. (a) is a slice in time of an input image sequence with subtle non periodical motion (The yellow dashed lines indicate when the subtle movement starts and ends). (b) is the calculated quaternionic phase shift $\Delta\phi_n$ of (a). We reduce the noise in (b) using three different filtering schemes: (c) an IIR Butterworth causal filter which delays the signal; (d) a FIR non-causal band-pass filter which does not delay the signal but it creates some artifacts before and after the motion has taken place (Gibbs phenomenon); (e) a FIR non-causal low-pass filter (our proposal).

3.3. Masking regions of interest

In order to optimize the spotting process, we decided to mask facial regions of interest (ROIs) in which, according to the Facial Action Coding System (FACS) [2], MEs might appear. For that purpose, we create a mask M_1 using the facial landmarks localized in Sec. 3.1.

Another thing to consider before motion spotting is the problem of image noise. Assuming the general case of two static images corrupted with some level of additive Gaussian noise, their quaternionic phase difference would be non-zero ($|\Phi_n| > 0$) even after the phase SNR is improved by ways of spatial and temporal filtering (Sec. 2.2). We have observed that the Φ_n values could have a high variance in areas where local amplitude A has a relative low value regardless of the presence of motion. Considering that the motion phase-signal in regions of low amplitude is not meaningful [17], we decided to isolate these areas using an adaptive threshold of validation computed from the local amplitude. However, since the scale of local amplitude might vary from subject to subject (for example, in videos with subjects wearing glasses the local amplitude in the border of the glass frames was very high compared to the rest of the face), we need to normalize the local amplitude before we can threshold it.

$$M_2 = \begin{cases} 1 & \text{if} \quad \beta \leq \frac{A_n}{A_q} \\ 0 & \text{if} \quad \beta > \frac{A_n}{A_q} \end{cases} \tag{10}$$

where A_n is the calculated local amplitude of the image at frame n, A_q is the 95-percentile of the empirical distribution of the amplitudes along the video and β is a threshold

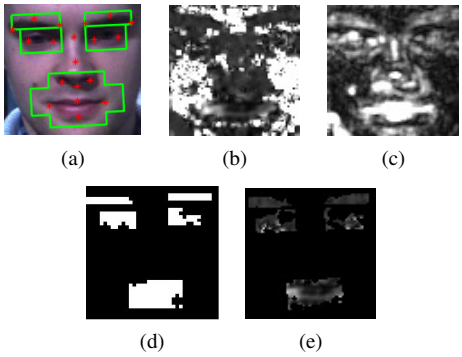


Figure 3: Masking facial regions of interest. (a) from the input video the face has been cropped, facial landmarks have been located and some ROIs have been delimited (green rectangles). (b) is the filtered quaternionic phase shift Φ_n and (c) is its local amplitude. (d) is the mask created using the amplitude and the facial ROIs. Finally, (e) is the result of masking (b) with (d).

selected by the user (See Sec. 4.2). By masking the low amplitudes areas we have effectively selected the regions in which MEs can be detected. We combine both masks ($M = M_1 \& M_2$) and refine the result further using morphological opening. Finally, we mask the quaternionic phase ($\Phi_n \mathbf{u}$) with M (as seen in Fig. 3e).

3.4. Micro-Expression Spotting

3.4.1 Preprocessing

Our first step is to minimize the effect of potential macro-movements from the spotting process. For this, we will consider any head pose change or translation that might occur during the video as rigid motion. With this in mind, we subtract the average of the masked quaternionic phase:

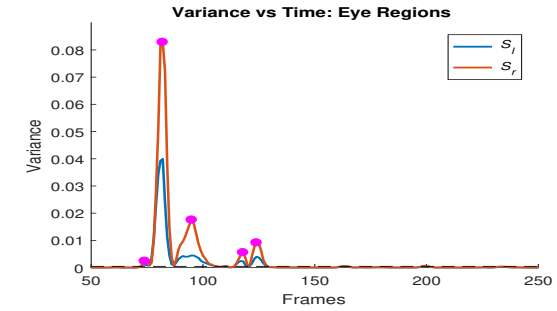
$$\Phi'_{n,s}\mathbf{u'}_s = \Phi_{n,s}\mathbf{u}_s - \frac{1}{card(M)} \sum_{r \in M} \Phi_{n,r}\mathbf{u}_r \tag{11}$$

where s is a masked pixel, $\mathbf{u'}_s = i\cos\theta'_s + j\sin\theta'_s$ and $card(M)$ is the cardinality of the mask M. All the elements of a rigid motion such as translation follow the same orientation and magnitude, thus Eq. 11 reduces the effect of this kind of movements for the spotting step while the elements in non-rigid movements such as facial expressions follow different orientations and magnitudes and are not heavily affected by this step. Next, we get rid of orientation and calculate the euclidean norm of the phase thus:

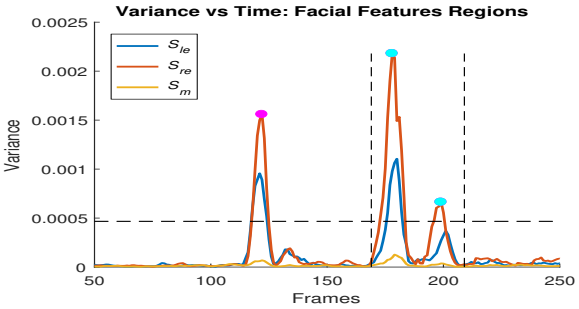
$$|\Phi'_n| = \sqrt{(\Phi'_n \sin\theta')^2 + (\Phi'_n \cos\theta')^2} \tag{12}$$

3.4.2 Feature Signal Generation

The next step would be to calculate the phase variations over time and spot any subtle movements as MEs. However, by



(a) eye signals peak detection



(b) facial features signals peak detection

Figure 4: Micro-expression spotting process. The vertical dashed lines represent the time period between the true onset and true offset. First, we detect peaks on the signal that go over a threshold (the horizontal dashed line). For (a) these peaks (magenta dots) represent eye blinks or movements while in (b) represent ME candidates. Finally, we select the pair of peaks in (b) that does not coincide in time with the ones in (a) as a true MEs (cyan dots)

this logic, eye blinks and eye gaze changes could be wrongfully considered as MEs. Instead of just ignoring the information given by the eye areas, we can use it to help our system discard the possible false-positives. Thus, we divide our masked data in five different areas: two eye areas (left and right eye), and three facial features areas (left and right eyebrow and mouth area). For each area and each frame, we calculate the variance:

$$S(n) = \frac{1}{L}\sum_{l=0}^{L-1} |\Phi'_{n,l} - \mu_n|^2 \tag{13}$$

where S is a 1-D signal which peaks represent subtle changes in the video, μ_n is the mean value of $|\Phi'_n|$ in a given area, l is the index of the pixels in a selected area and L is the total number of pixels in that area.

3.4.3 Peak Analysis

From the previous step, we obtain 5 signals: S_l and S_r which correspond to the left and right eye areas and S_{le}, S_{re} and S_m which correspond to the left eyebrow, right eyebrow and mouth areas respectively. Next, we compute the median and maximum values of the calculated signals:

$$max_S = \max_{\forall n \in N} (S_{le}, S_{re}, S_m) \tag{14}$$

$$med_S = \underset{\forall n \in N}{\text{median}} (S_{le}, S_{re}, S_m) \tag{15}$$

and we use them to create a series of adaptive thresholds:

$$T_E = \frac{max_S}{2} \tag{16}$$

$$T_F = med_S + (max_S - med_S) \times \alpha \tag{17}$$

The next step is to localize any peak or local maxima in the signals that surpasses the previously computed thresholds (using T_E as threshold for S_l and S_r and T_F as threshold for S_{le}, S_{re}, S_m). For each signal S, we obtain a matrix:

$$\mathbf{P} = \begin{bmatrix} x_1 & n_1 \\ & \vdots \\ x_k & n_k \end{bmatrix} \tag{18}$$

where k is the total number of detected peaks, x_i and n_i are the magnitude and time (frame number) of a detected peak.

We perform a procedure to refine $\mathbf{P}$ by choosing the tallest peak and discard all peaks which are closer than a minimal peak-to-peak separation (ψ frames). Then, the procedure is repeated for the tallest remaining peak and iterates until it runs out of peaks to consider.

The next step is to discard redundant information by combining different peak matrices into one without duplicates (see Algorithm 1). We obtain $\mathbf{P}_E = \text{FUSEPEAKS}(\mathbf{P}_l, \mathbf{P}_r)$ for the eyes areas and $\mathbf{P}_F = \text{FUSEPEAKS}(\mathbf{P}_{le}, \mathbf{P}_{re}, \mathbf{P}_m)$ for the facial features areas. Finally, we discard as an eye blink any peak from $\mathbf{P}_F$ that has a corresponding peak in $\mathbf{P}_E$ (see Algorithm 2).

One thing to take into consideration is the nature of MEs. During most MEs, the face goes from a neutral state (onset) to a moment when the ME is at its peak (apex) and then, after a short period of time, it goes back to a neutral state (offset). However, there are some micro-expressions that have fast onset phase but very slow offset phase (some even remain in apex for seconds) which some methods would fail to detect. That means that a ME is comprised of either one or two subtle motions. Therefore, our method has been adapted to detect either one or two peaks per ME (a pair of peaks would be identified as one ME as seen in Fig. 4b) depending on the case. One advantage of this approach is that, if an eye movement happens during the onset phase or the offset phase, our method might discard only one peak and the ME would still be spotted.

Algorithm 1: FusePeak: Peak fusion algorithm

Input : A set of input peak matrices $\{\mathbf{P}_1, \ldots, \mathbf{P}_m\}$
Output: Fused peak matrix $\mathbf{P}_u$

Concatenate the elements of the input matrices: $\mathbf{P}_f = \begin{bmatrix} \mathbf{P}_1 \\ \vdots \\ \mathbf{P}_m \end{bmatrix}$ where k_f = number of rows in $\mathbf{P}_f$
Sort $\mathbf{P}_f$ using $n_{f,i}, i = 1, \ldots, k_f$
$s = 1$ and $\mathbf{P}_{u,s} = [x_{u,s}, n_{u,s}] = \mathbf{P}_{f,1} = [x_{f,1}, n_{f,1}]$
for $i \leftarrow 2$ **to** k_f **do**
  if $n_{u,s} = n_{f,i}$ **then**
    if $x_{u,s} \leq x_{f,i}$ **then**
      $x_{u,s} = x_{f,i}$
    end
  else
    $s = s + 1$
    $\mathbf{P}_{u,s} = [x_{u,s}, n_{u,s}] = \mathbf{P}_{f,i} = [x_{f,i}, n_{f,i}]$
  end
end

Algorithm 2: ME spotting

Input : Peak matrices $\mathbf{P}_E$, $\mathbf{P}_F$
Output: ME peak matrix $\mathbf{P}_m$

k_F = number of rows in $\mathbf{P}_F$;
k_E = number of rows in $\mathbf{P}_E$;
$s = 1$;
for $i \leftarrow 1$ **to** k_F **do**
  $flag = 0$
  for $j \leftarrow 1$ **to** k_E **do**
    if $n_{F,i} = n_{E,j}$ *AND* $x_{F,i} \leq x_{E,j}$ **then**
      $flag = 1$
    end
  end
  if $flag = 0$ **then**
    $\mathbf{P}_{m,s} = [x_{m,s}, n_{m,s}] = \mathbf{P}_{F,i} = [x_{F,i}, n_{F,i}]$
    $s = s + 1$
  end
end

4. Experimental Results and Discussions

In this section, we describe the experimental procedures. Firstly, we briefly introduce the selected datasets that we use to test our methodology. Secondly, we talk about the parameters selection and evaluation scheme for our proposed method. Thirdly, we present our results and compare them with the state of the art. Fourthly, we study the impact of the different parameters on our system. Finally, we discuss the relevance of our results and what challenges we faced in the implementation of our method.

4.1. Datasets

For our experimentation, we selected two spontaneoulsy elicited ME databases. The first one, the Spontaneous Micro-Expression Database (SMIC) [6] 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 pixels). Ground truth annotation provides the frame numbers indicating the onset and offset frames (the moment when a ME starts and ends respectively). The MEs are labeled into three emotion classes: positive, surprise and negative emotions. For our experimentation we decided to use only the SMIC-HS dataset.

The second database, the improved Chinese Academy of Sciences Micro-expression (CASME II) [20] 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 pixels. Ground truth annotations not only provides the frame numbers indicating the onset and offset but also the apex frames (the moment when the ME change is at its highest intensity). The MEs are labeled into five emotion classes: happiness, surprise, disgust, repression and others.

4.2. Evaluation Procedure

For both datasets, we can calculate up to 4 levels of the Riesz pyramid. The first two levels (which have the information of the high frequency subbands) seem to carry an important amount of undesired noise and the third level of the pyramid seems more sensible to detect MEs compared to the fourth level (See Sec. 4.4). Thus, we choose to use only the third level of the pyramid.

Based on [22], we know that the duration of MEs is in the order of 100 ms. Thus, we design a FIR non-causal low-pass temporal filter with cutoff frequency of 10 Hz, corresponding to a filter of order 18 for the SMIC-HS database and 36 for the CASME II database. We used a Gaussian Kernel K_ρ with standard deviation $\rho = 2$ for spatial filtering for both cases. One thing to consider in the parameter selection is the impact of the duration of MEs in the spotting process. Since MEs can last from 170 to 500 ms [22], we selected a conservative lower bound for the peak-to-peak minimal separation ψ by taking half of the minimal expected duration (that would be 85 milliseconds which will correspond to 9 frames for SMIC-HS and 18 frames for CASME II). For the β parameter, we perform a leave-one-subject-out cross validation which shows that the best values for β are close to 0.1 for the SMIC-HS dataset and between 0.2 and 0.22 for the CASME II dataset. We report the aggregate results of the evaluation measure presented in Sec. 4.3.

After the peak detection step, all the spotted peak frames are compared with ground truth labels to tell whether they are true or false positive spots. With a certain threshold level, if one spotted peak is located within the frame range (between the onset and offset) of a labeled ME video, the spotted sequence will be considered as one true positive ME. Otherwise, we will count a penalty equal to the duration of the labeled ME (offset−onset+1 frames) as false positive. We define the true positive rate (TPR) as the percentage of frames of correctly spotted MEs divided by the total number of ground truth ME frames in the database. The false positive rate (FPR) is calculated as the percentage of incorrectly spotted frames divided by the total number of non-ME frames from all the image sequences. We evaluate the performance of our ME spotting method using receiver operating characteristics (ROC) curves with TPR as the y axis and FPR as the x axis.

4.3. Results

We performed the spotting experiment on CASME II and the high speed camera dataset SMIC-HS. The spotting results on each dataset are presented in Fig. 5. The ROC curve is drawn by varying the parameter α in Eq. 17 (from 0 to 1 with a step size of 0.05). We evaluate the accuracy of our method by calculating the area under the ROC curve (AUC) for each dataset. The AUC percentage for the SMIC-HS database is $\mathbf{88.61\%}$, and for CASME II it is $\mathbf{90.93\%}$. However, most of the operating points in the ROC curve are not reasonable due to high number of false positives. Instead, some examples can be given with a reasonable ratio between true and false positives. For SMIC-HS dataset a spotting accuracy of 75.20% was achieved with only 10.89% FPR using $\alpha = 0.4$, and for CASME II dataset a spotting accuracy of 83.16% was achieved with only 10.52% of FPR using $\alpha = 0.15$.

We wanted to compare our results with the work made by [5] since we use a similar method of evaluation (ROC curves). However, this might become challenging since their method and parameters are different from ours. In order to compare results we changed some parameters that we used in Sec. 4.2. Specifically, the spotting range was changed to $[\text{ONSET} - (L-1)/4, \text{OFFSET} + (L-1)/4]$ and the false positive penalty was changed to L frames, being L a time window of about 0.32 seconds according to their work ($L = 33$ for SMIC-HS and $L = 65$ for CASME II). As it can be observed in table 1, the results of our method outperforms the results reported in [5].

4.4. Parameter Analysis

To evaluate the impact of the different parameters on our system, we test our proposed framework while varying its parameter values. The parameters we decided to evaluate are: the pyramid level (from 1 to 4), amplitude masking threshold β (from 0 to 0.9) and minimal peak-to-peak separation ψ (from 5 to 25 frames for the SMIC-HS dataset and from 5 to 40 frames in the CASME II dataset). We perform our spotting experiments similarly as in Sec. 4.2 for both datasets and for each set of parameters we obtain the AUC. We decided to represent the result for each dataset as four result surfaces (one for each Riesz pyramid level) depicting the AUC as a function of ψ and β (As seen in Fig. 6b and 6d). Exhaustive results for all the levels of the pyramid, are provided in the supplementary materials.

We show the impact of the Riesz pyramid level by calculating the mean AUC as a function of β (The results are shown in Fig. 6a and 6c). We observe that we obtain the best performance values using the third level of the pyramid for the SMIC-HS dataset and the third and fourth levels for the CASME II dataset. However, a further inspection of the result surfaces in CASME II show that the results in the third level of the pyramid are slightly better. We further inspect the result surface of the third pyramid level for both datasets (Fig. 6b and 6d). We obtain stable results when β ranges between 0.05 and 0.25 in the SMIC-HS dataset and between 0.05 to 0.3 in the CASME II dataset. An even closer inspection shows that we obtain the best results when β varies between 0.07 and 0.13 and ψ varies between 5 to 15 frames (at a 100 fps it corresponds to 50 to 150 milliseconds) for the SMIC-HS dataset (Fig. 6b) and when β varies between 0.12 and 0.15 and between 0.18 and 0.22 and ψ is bigger than 10 frames (at a 200 fps it corresponds to 50 milliseconds) for CASME II dataset (Fig. 6d).

One thing to take in consideration is that, by further augmenting the value of ψ, we might unknowingly discard MEs

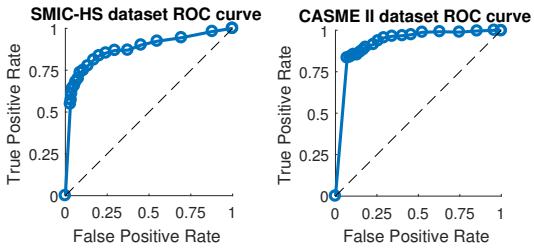
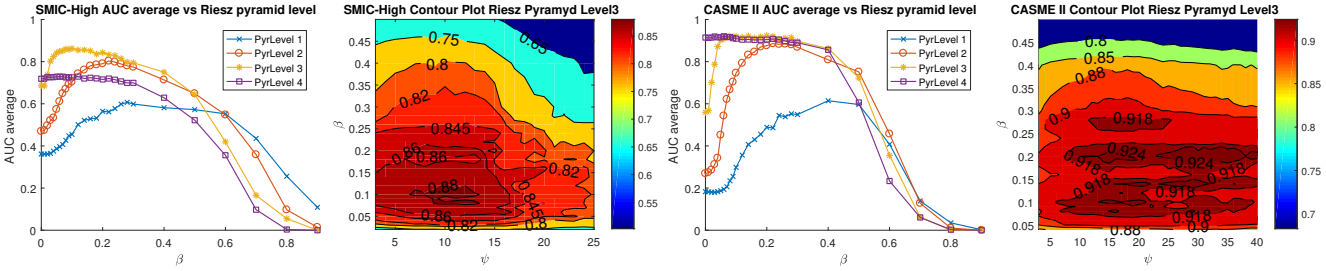


Figure 5: ROC curves for ME spotting on SMIC-HS and CASME II datasets

Database	SMIC-HS	CASME II
LBP [5]	83.32%	92.98%
HOOF [5]	69.41%	64.99%
Ours	**89.80%**	**95.13%**

Table 1: AUC values of the ME spotting experiments using different methods on SMIC-HS and CASME II datasets using the measure metrics from [5].



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

Figure 6: Parameter Evaluation results in SMIC-HS and CASME II datasets

that occur within a small window of time. Considering that some videos in the SMIC-HS dataset contain multiple MEs but all videos in the CASME II dataset contain only one ME, it would explain why augmenting the value of ψ affects the results for SMIC-HS but not the results for CASME II.

After analyzing our results, we estimate that, since both of the evaluated datasets have the same spatial resolution (640×480 pixels), we could use the same level of the pyramid for both and we could fairly establish a robust value range for β (between 0.05 and 0.25). Furthermore, this range value for β is coherent with our values obtained by our cross validation scheme. We can also estimate that a robust value for ψ for both datasets should correspond to at least 50 milliseconds. Moreover, we can conclude that the level of the pyramid and β are the parameters which have more impact in the performance of our system.

4.5. Discussion

Although both databases are similar, the SMIC-HS contain longer video-clips making the spotting task more difficult and more prone to false positive. We suspect this might be the reason why a higher AUC is achieved on CASME II database. Furthermore, the subjects in the CASME II dataset were at a closer distance to the camera during video recording, thus the captured faces had a bigger resolution which result in a shift of the ME motion to low frequencies. This might explain why we could obtain good results using the fourth level of the Riesz pyramid in the CASME II dataset but not in the SMIC-HS dataset. A scale normalization of the captured faces could allow us to fix the pyramid level selection, regardless of the database.

Upon detailed examination of the spotting results, we found that a large portion of false negatives were caused by our algorithm dismissing true MEs as eye movements. This might happened because our system does not differentiate eye blinks from eye-gaze changes, thus discarding MEs that happen simultaneously with eye-gaze changes.

One of the main challenges of comparing our ME spotting method with the state of the art comes from the fact that there is not a single standard performance metric. For example, [7] evaluate their work using mean absolute error and standard error. [1] also uses ROC curves to evaluate their work but fails to disclose how they compute their false positives. Although our evaluation method is very similar from the one by [5], the false positive penalties were different. That is the reason why the initial AUCs obtained in the beginning of Sec. 4.3 are different from the ones in our comparative Table 1, because the first tests had higher false positive penalties. This happens because these penalties were based on an assumed duration of MEs. However, since micro expressions have different duration times (as discussed in Sec. 3.4), a different approach should be considered for computing false positives.

5. Conclusions

We presented a facial micro-expression spotting method based on the Riesz pyramid. Our method adapts the quaternionic representation of the Riesz monogenic signal by proposing a new filtering scheme. We are also able to mask regions of interest where subtle motion might take place in order to reduce the effect of noise using facial landmarks and the image amplitude. Additionally, we propose a methodology that separates real micro-expressions from subtle eye movements decreasing the quantity of possible false positives. Experiments on two different databases show that our method surpasses other methods from the state of the art. Moreover, the results of our parameter analysis suggest that the method is robust to changes in parameters. Furthermore, the quaternionic representation of phase and orientation from the Riesz monogenic signal could potentially be exploited in the future for a more general subtle motion detection scheme and for facial micro-expression recognition.

Acknowledgements

This work has been supported by a research grant from the Foundation of Jean Monnet University (UJM).

References

[1] 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.

[2] P. Ekman and E. L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, Apr. 2005.

[3] D. H. Kim, W. J. Baddar, and Y. M. Ro. Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 382–386, New York, NY, USA, 2016. ACM.

[4] J. Lee and S. Y. Shin. General construction of time-domain filters for orientation data. *IEEE Transactions on Visualization and Computer Graphics*, 8(2):119–128, Apr. 2002.

[5] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen. Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017.

[6] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Apr. 2013.

[7] S. T. Liong, J. See, K. Wong, A. C. L. Ngo, Y. H. Oh, and R. Phan. Automatic apex frame spotting in micro-expression database. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 665–669, Nov. 2015.

[8] A. Moilanen, G. Zhao, and M. Pietikainen. Spotting Rapid Facial Movements from Videos Using Appearance-Based Feature Difference Analysis. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, pages 1722–1727, Aug. 2014.

[9] S. Park and D. Kim. Subtle facial expression recognition using motion magnification. *Pattern Recognition Letters*, 30(7):708–716, May 2009.

[10] S. Y. Park, S. H. Lee, and Y. M. Ro. Subtle Facial Expression Recognition Using Adaptive Magnification of Discriminative Facial Motion. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 911–914, New York, NY, USA, 2015. ACM.

[11] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 51–56, Mar. 2011.

[12] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical report, International Journal of Computer Vision, 1991.

[13] G. Tzimiropoulos and M. Pantic. Optimization Problems for Fast AAM Fitting in-the-Wild. In *2013 IEEE International Conference on Computer Vision (ICCV)*, pages 593–600, Dec. 2013.

[14] M. Unser, D. Sage, and D. V. D. Ville. Multiresolution Monogenic Signal Analysis Using the Riesz-Laplace Wavelet Transform. *IEEE Transactions on Image Processing*, 18(11):2402–2418, Nov. 2009.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, volume 1, pages I–511–I–518 vol.1, 2001.

[16] N. Wadhwa, M. Rubinstein, F. Durand, and W. Freeman. Riesz pyramids for fast phase-based video magnification. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, May 2014.

[17] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Phase-based Video Motion Processing. *ACM Trans. Graph.*, 32(4):80:1–80:10, July 2013.

[18] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Quaternionic representation of the riesz pyramid for video magnification. Technical report, 2014.

[19] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian Video Magnification for Revealing Subtle Changes in the World. *ACM Trans. Graph.*, 31(4):65:1–65:8, July 2012.

[20] 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*, 9(1), Jan. 2014.

[21] W.-J. Yan, S.-J. Wang, Y.-H. Chen, G. Zhao, and X. Fu. Quantifying Micro-expressions with Constraint Local Model and Local Binary Pattern. In L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops*, number 8925 in Lecture Notes in Computer Science, pages 296–305. Springer International Publishing, Sept. 2014. DOI: 10.1007/978-3-319-16178-5_20.

[22] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu. How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions. *Journal of Nonverbal Behavior*, 37(4):217–230, July 2013.