INTRINSIC TWO-DIMENSIONAL LOCAL STRUCTURES FOR MICRO-EXPRESSION RECOGNITION

Yee-Hui Oh[†] *Anh Cat Le Ngo*[†] *Raphael C.-W. Phan*[†] *John See*[†] *Huo-Chong Ling*[⋆]

[†] Multimedia University, 63100 Cyberjaya, Malaysia.
[⋆] Curtin University, 98009 Miri, Malaysia.

ABSTRACT

An elapsed facial emotion involves changes of facial contour due to the motions (such as contraction or stretch) of facial muscles located at the eyes, nose, lips and etc. Thus, the important information such as corners of facial contours that are located in various regions of the face are crucial to the recognition of facial expressions, and even more apparent for micro-expressions. In this paper, we propose the first known notion of employing intrinsic two-dimensional (i2D) local structures to represent these features for micro-expression recognition. To retrieve i2D local structures such as phase and orientation, higher order Riesz transforms are employed by means of monogenic curvature tensors. Experiments performed on micro-expression datasets show the effectiveness of i2D local structures in recognizing micro-expressions.

Index Terms— Emotion, micro-expressions, i2D, higher order Riesz transform

1. INTRODUCTION

Micro-expressions are facial expressions involving only minute and brief facial motions that last typically from 1/25 to 1/5 of a second [1]. Just like normal facial expressions, micro-expressions include six universal expressions, namely happy, sad, fear, surprise, anger and disgust. Micro-expressions are involuntary and usually reveal the genuine emotion state of a person [2]. Thus, recognizing micro-expressions is beneficial as we can interpret and identify the true feeling of someone in order to avoid conflict, danger or being deceived. However, experiments conducted by Frank et al. [3] have quantified the difficulty of recognizing micro-expressions with naked eyes. His results revealed that the recognition rates for five subtle expressions were just 32% and 47% for untrained and trained human experts respectively.

More recently, attention has increasingly been paid to micro-expressions recognition. However, its performance is not as good as normal facial expressions recognition; the latter achieving up to over 90% accuracy. The reasons could be the following: (1) lack of well established micro-expressions databases due to difficulty in inducing, capturing and identifying subtle emotions and (2) difficulty in recognizing micro-expressions due to the short duration and low-intensity facial motions of elapsed expressions. And yet such capability is beneficial for diverse application settings where revealing hidden or suppressed emotions is vital; such is the case for public safety (e.g. truth concealment or suspicious intent) and even in the corporate world where the integrity of high-ranking officials needs to be evaluated.

To our best knowledge, there are only two publicly available spontaneous micro-expressions databases: CASME II [4] and SMIC [5]. In [4], the baseline was established on CASMEII by employing 5×5 block-based Local Binary Pattern - Three Orthogonal Planes (LBP-TOP) [6] and Support Vector Machine (SVM) for feature extraction and classification respectively, with Leave-One-Video-Out cross-validation (LOVOCV). The same techniques for feature extraction and classification were adopted by [5] to produce a baseline for SMIC. In the case of this work, temporal interpolation model (TIM) was additionally applied as a preprocessing technique to fix the frame length for each video. There are few other related works that employ different feature extraction techniques for micro-expressions: optical strain [7, 8] and variants of LBP-TOP [9, 10].

More recently, monogenic signal theory [11] was proposed to extract the local structures of images for emotion recognition. In the work of [12], only local magnitude and local real and imaginary parts of orientation were utilized to perform normal facial expression recognition. In their later work [13], they included the local phase, orientation (i.e., the real and imaginary parts) and magnitude as features for in-the-wild emotion recognition. In [14], the same concept of extracting local signal structures such as amplitude, phase and orientation as features was first employed for micro-expression recognition. These previous works utilized the *first order* Riesz transform. However, first order Riesz transform can only be used to analyze intrinsic one-dimensional (i1D) local structures such as lines and edges [15]. Since facial expressions comprise complex contours such as corners, intrinsic two-dimensional (i2D) local structures would be a better and more suitable feature representation. Therefore, in this paper we propose a new feature representation technique which adopts i2D local structures from multiple scales, for the first time in micro-expression recognition. The i2D lo-

cal structures such as phase and orientation are retrieved by employing monogenic curvature tensors based on *second* and *third order* Riesz transforms. The i2D phase and orientation from all scales are then encoded by the proposed encoding schemes, and described by a dynamic texture descriptor, which forms the final feature vector for SVM classification. In our experiments, we show that i2D local structures outperform i1D local structures in micro-expression recognition. Concisely, the main contributions of this paper are: (1) the first known work that employs i2D local structures for micro-expression recognition; (2) new formulations of encoding schemes for phase and orientation information; and (3) the comparison between i1D and i2D local structures as feature representations for micro-expression recognition.

The paper is organized as follows: Section 2 gives a brief introduction on higher order Riesz transform, followed by Section 3 that describes the proposed method in detail. Section 4 then presents the experimental results and analysis while our conclusion is drawn in Section 5.

2. HIGHER ORDER RIESZ TRANSFORM

The intrinsic dimension of a signal expresses the number of degrees of freedom required to describe local structures [11]. An image can be categorized into three possible intrinsic dimensionalities: i0D, i1D and i2D; e.g., constant areas are of i0D, straight lines and edges are of i1D while more complex patterns such as corners and junctions are of i2D [16].

The monogenic signal which is built on first order Riesz transform, is an isotropic 2D extension of the traditional 1D analytic signal. It can analyze i1D local structures such as straight lines and edges effectively in a rotation invariant manner. In the case of a 2D image, $\mathbf{X}$, the first order Riesz kernel in the spatial domain is

$$(R_x(\mathbf{X}), R_y(\mathbf{X})) = \left(\frac{x}{2\pi|\mathbf{X}|^3}, \frac{y}{2\pi|\mathbf{X}|^3}\right), \mathbf{X} = (x, y) \in \mathbb{R}^2 \tag{1}$$

where R_x and R_y represent the Riesz transform operator corresponding to x and y directions. After Fourier transform, the transfer function of the kernel is

$$(H_u(\mathbf{u}), H_v(\mathbf{u})) = \left(i\frac{u}{|\mathbf{u}|}, i\frac{v}{|\mathbf{u}|}\right), \mathbf{u} = (u, v) \in \mathbb{R}^2 \tag{2}$$

and the monogenic signal of an image $f(\mathbf{X})$ is defined as the combination of $f(\mathbf{X})$ and its Riesz transform

$$f_M(\mathbf{X}) = \Big(f(\mathbf{X}), R_x\{f\}(\mathbf{X}), R_y\{f\}(\mathbf{X})\Big) \tag{3}$$

Concisely, the i1D local phase ϕ and i1D local orientation θ can be retrieved by

$$\phi_{i1D} = atan2\left(\frac{\sqrt{R_x^2 + R_y^2}}{f}\right), \quad \theta_{i1D} = atan\left(\frac{R_y}{R_x}\right) \tag{4}$$

To retrieve i2D local structures such as corners and junctions on an image, higher order Riesz transform has to be employed. The Laplacian of Poisson filter (LOP) proposed by Fleischmann [15] is employed as it can derive the second and third order filter kernels easily. Higher order filter kernels are produced by multiplying the Riesz transform with itself and with the kernel filter according to the convolution theorem in the frequency domain.

The second order Riesz transform in the spatial domain has three components:

$$R_{xx} = \mathcal{F}^{-1}\{\mathcal{F}\{R_x\{R_x\} * LOP\}(\mathbf{u})\} \tag{5}$$

$$R_{xy} = \mathcal{F}^{-1}\{\mathcal{F}\{R_x\{R_y\} * LOP\}(\mathbf{u})\} \tag{6}$$

$$R_{yy} = \mathcal{F}^{-1}\{\mathcal{F}\{R_y\{R_y\} * LOP\}(\mathbf{u})\} \tag{7}$$

while the third order Riesz transform in spatial domain has four components:

$$R_{xxx} = \mathcal{F}^{-1}\{\mathcal{F}\{R_x\{R_x\{R_x\}\} * LOP\}(\mathbf{u})\} \tag{8}$$

$$R_{xxy} = \mathcal{F}^{-1}\{\mathcal{F}\{R_x\{R_x\{R_y\}\} * LOP\}(\mathbf{u})\} \tag{9}$$

$$R_{xyy} = \mathcal{F}^{-1}\{\mathcal{F}\{R_x\{R_y\{R_y\}\} * LOP\}(\mathbf{u})\} \tag{10}$$

$$R_{yyy} = \mathcal{F}^{-1}\{\mathcal{F}\{R_y\{R_y\{R_y\}\} * LOP\}(\mathbf{u})\} \tag{11}$$

where the $*$ is the convolution operator.

To retrieve the i2D local structures for the phase and orientation, monogenic curvature tensors are employed. For our needs, we make use of both the even and odd part of the monogenic curvature tensor [15], which are defined as

$$T_{even} = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{xy} & R_{yy} \end{bmatrix} \tag{12}$$

$$T_{odd} = \begin{bmatrix} R_{xxx} + e_{12}R_{xxy} & e_{12}R_{xxy} - R_{xyy} \\ e_{12}R_{xxy} - R_{xyy} & R_{xyy} + e_{12}R_{yyy} \end{bmatrix} \tag{13}$$

where e_{12} is a bivector basis of Clifford algebra [17]. By means of Clifford algebra, the odd tensor can be split into two parts $T_{odd} = T_{oddx} + e_{12}T_{oddy}$. From T_{odd} and T_{even}, the i2D local phase ϕ and orientation θ can be derived by

$$\phi_{i2D} = atan2\left(\frac{|det(T_{odd})|}{det(T_{even})}\right) \tag{14}$$

$$\theta_{i2D} = atan\left(\frac{det(T_{oddx})}{det(T_{oddy})}\right) \tag{15}$$

where $det(\cdot)$ is the determinant of the tensor matrix.

3. PROPOSED METHOD

Our proposed approach is illustrated in Fig. 1, and more details of each step are given in the following subsections.

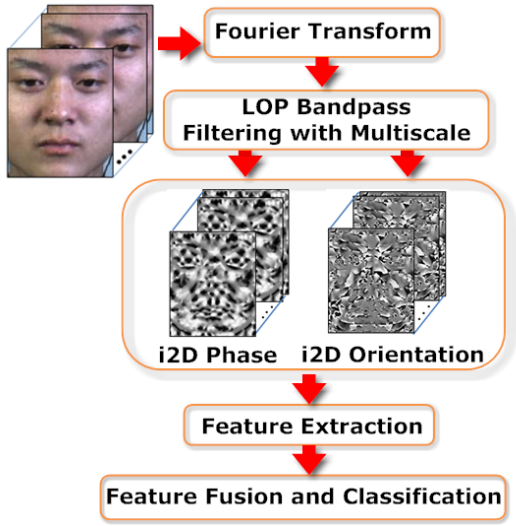

Fig. 1. Flowchart of the proposed algorithm

3.1. Signal Preprocessing

In the real world, a signal e.g. an image, is usually formed by multiple frequencies. In monogenic representation, local signal properties of image structures e.g. amplitude, phase and orientation are extracted from a narrow sub-band of the whole image spectrum. Thus, a bandpass filter is required to represent an image with band limited signals. In our work, we adopt the Laplacian of Poisson (LOP) filter of [15] as the bandpass filter to construct higher order Riesz transform. For first order Riesz transform, we employ the standard Poisson filter as it is not possible to derive the first order Riesz transform from the LOP filter [15]. The transfer functions of the LOP and Poisson filters in the Fourier domain are as follows:

$$\mathcal{F}\{LOP\}(\mathbf{u}) = -4\pi^2|\mathbf{u}^2|exp(-2\pi|\mathbf{u}|s), \mathbf{u} \in \mathbb{R}^2 \quad (16)$$

$$\mathcal{F}\{P\}(\mathbf{u}) = exp(-2\pi|\mathbf{u}|s), \mathbf{u} \in \mathbb{R}^2 \quad (17)$$

where $ss > 0$ is the scale space parameter. In our experiment, s is set to 4, 8 and 16 to form a multiscale bandpass signal. The multiscale bandpass signals are then convoluted with the second and third order Riesz transforms to build the even and odd monogenic curvature tensors. The i2D phase and orientation are retrieved from the even and odd monogenic curvature tensors based on Eqs. (14) and (15) respectively, while the i1D phase and orientation are computed from Eq. (4).

3.2. Feature Extraction

In this work, we introduce new encoding schemes for both the i1D and i2D signal structures. Instead of encoding the input into binary code by comparing the difference using the unit step function (which was found to be not reasonable for phase and orientation angles [13]), a simple quantification formula is used to quantize the phase and orientation into a number of discrete levels. These quantified phase and orientation are then described by the state-of-the-art dynamic texture descriptor LBP-TOP to include essential temporal and appearance information found in sequences [6]. To extract the local texture pattern, a block-based approach to LBP-TOP is implemented, where input images are first partitioned into 5×5 non-overlapping blocks (following e.g. [14]), then histograms are obtained from each block volume and concatenated to form the final feature histogram.

3.2.1. Encoding Local Orientation

The i1D and i2D orientation structures are encoded by the following quantification formula:

$$j = sign(\theta(x,y))mod([\frac{\theta(x,y)}{\frac{\pi/2}{\vartheta}}], \vartheta) \quad (18)$$

where $\theta(x,y)$ denotes the orientations and ϑ denotes the quantification levels. As the orientation spans from $\frac{-\pi}{2}$ to $\frac{\pi}{2}$, the sign of the orientation has to be taken into consideration for quantifying. The correlation between the orientation of the center pixel and the neighboring pixels is computed as:

$$J_p = \begin{cases} 0, & j_{x_c,y_c} = j_{x_p,y_p} \\ 1, & j_{x_c,y_c} \neq j_{x_p,y_p} \end{cases} \quad (19)$$

where j_{x_c,y_c} and j_{x_p,y_p} denote the orientation of the center point and neighboring point respectively. Finally, the orientation histogram is computed as follows:

$$H_\theta = \sum_{p=0}^{P-1} J_p 2^p \quad (20)$$

3.2.2. Encoding Local Phase

The i1D and i2D phase structures are first encoded by:

$$p = mod([\frac{\phi(x,y)}{\frac{2\pi}{\varphi}}], \varphi) \quad (21)$$

a quantification function, where $\phi(x,y)$ denotes the phase angle, and φ is the number of quantification levels. The binary encoded dominant phase Q_P can easily obtained by:

$$Q_p = \begin{cases} 0, & q_{x_c,y_c} = q_{x_p,y_p} \\ 1, & q_{x_c,y_c} \neq q_{x_p,y_p} \end{cases} \quad (22)$$

where q_{x_c,y_c} and q_{x_p,y_p} denote the quantified phase value of the center point and neighboring point respectively. Finally, the phase histogram is computed as follows:

$$H_\phi = \sum_{p=0}^{P-1} Q_p 2^p \quad (23)$$

3.3. Feature Fusion and Classification

The multiple feature representations, H_c, where $c = \{\theta, \phi\}$ across all s scales are fused by direct concatenation. A stan-

dard multi-class linear-kernel SVM classifier is then used to perform micro-expression classification of image sequences.

4. EXPERIMENTS

4.1. Database Description and Preprocessing

The proposed algorithm is evaluated on two publicly available spontaneous micro-expression databases: CASME II [4] and SMIC [5]. The videos from these databases have been recorded under constrained lab conditions. CASME II consists of 247 videos elicited from 26 subjects, containing five micro-expression classes: Happiness (HAP), Disgust (DIS), Repression (REP), Surprise (SUR), and Others (OTH). The videos were recorded with a high speed camera with frame rate of 200 *fps* and a spatial resolution of 280×340 pixels. SMIC contains 164 micro-expression samples from 16 participants. The videos were captured and recorded with 100 *fps* at a resolution of 640×480 pixels. This database contains three classes of micro-expressions: positive (POS), negative (NEG) and Surprise (SUR). POS labels are mainly happy expressions while NEG labels include sad, fear and disgust expressions. All video frames in these databases have been well-registered, and the face regions have been properly aligned and cropped. Due to the different resolution of videos in the database, the input frames have been resized to the average resolution of 340×280 pixels.

4.2. Experiment Settings

In the encoding scheme, the quantification levels of local phase φ and local orientation ϑ are set to 16 and 8 respectively, as these are empirically found to perform the best in the datasets. SVM with linear kernel (c = 10000) is applied for classification. To avoid over-fitting our data which is of small sample size, linear kernels are chosen. As the database comprised multiple subjects, the Leave-One-Subject-Out cross-validation (LOSOCV) setting is applied in our evaluation, whereby for each of the k folds (for k subjects), the image sequences of one subject are used as testing samples while the remaining image sequences are used as training samples; finally, the average score across k folds is taken.

4.3. Results and Discussions

We present our experimental results in Table 1 for both CASME II and SMIC databases. The proposed approach using i2D local structures clearly outperforms the i1D local structures and the baseline methods of [4, 5] for both datasets. This strengthens the fact that micro-expressions involve fine variations of complex facial contours which include corners and junctions. These patterns are well described by i2D local structures as compared to gradient-like i1D local structures. The confusion matrices of our proposed method shown in Tables 2 and 3 further shed light as to which expressions are responding well to features encoded by i2D local structures. Interestingly, not all emotions improved with the use of I2D features; but specifically, the HAP and OTH (from CASME II), and POS and SUR (from SMIC) have shown improvement.

Table 1. The recognition performance based on F1-score (F1), precision (P) and recall (R) on the CASME II and SMIC

Local	CASME II [4]			SMIC [5]		
Structures	F1	P	R	F1	P	R
Baseline	0.35	0.36	0.34	0.43	0.43	0.44
i1D	0.32	0.37	0.28	0.34	0.33	0.36
i2D	**0.41**	**0.46**	**0.37**	**0.44**	**0.44**	**0.45**

Table 2. Confusion matrix for CASMEII by using i2D local structures

Expression	HAP	DIS	REP	SUR	OTH
HAP	**0.47**	0.03	0.06	0.03	0.41
DIS	0.11	0.32	0.00	0.00	0.57
REP	0.30	0.07	0.26	0.00	0.37
SUR	0.40	0.12	0.00	0.16	0.32
OTH	0.13	0.12	0.02	0.06	**0.67**

Table 3. Confusion matrix for SMIC by using i2D local structures

Expression	NEG	POS	SUR
NEG	0.33	0.39	0.29
POS	0.37	**0.55**	0.08
SUR	0.44	0.09	**0.47**

5. CONCLUSION

We present a novel idea of exploiting i2D local structures as the features for micro-expressions recognition, in contrast to previous works that utilized i1D local structures. In order to retrieve i2D local structures particularly the phase and orientation information, we adopt the even and odd parts of monogenic curvature tensors based on second-order and third-order Riesz transforms. The i2D local phase and i2D local orientation from all scales are then encoded by our proposed encoding schemes and represented by an LBP-TOP descriptor. Our experiments on two publicly available spontaneous micro-expression datasets demonstrate the effectiveness of i2D local structures over i1D local structures for micro-expression recognition.

6. ACKNOWLEDGEMENT

This research was supported by TM (Telekom Malaysia) under the projects UbeAware and 2beAware.

7. REFERENCES

[1] Paul Ekman and Wallace V Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[2] Paul Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*, WW Norton & Company, 2009.

[3] Mark Frank, Malgorzata Herbasz, Kang Sinuk, Amy Marie Keller, and Courtney Nolan, "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.

[4] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, pp. e86041, 2014.

[5] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[6] Guoying Zhao and Matti Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[7] Sze-Teng Liong, John See, Raphael C-W. Phan, Anh Cat Le Ngo, Yee-Hui Oh, and KokSheik Wong, "Subtle expression recognition using optical strain weighted features," in *Computer Vision - ACCV 2014 Workshops, Revised Selected Papers, Part II*, 2014, pp. 644–657.

[8] Sze-Teng Liong, Raphael C-W. Phan, John See, Yee-Hui. Oh, and KokSheik Wong, "Optical strain based recognition of subtle emotions," in *International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2014*, 2014, pp. 180–184.

[9] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh, "LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition," in *Computer Vision–ACCV 2014*, pp. 525–537. Springer, 2015.

[10] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition," *PloS one*, vol. 10, no. 5, pp. e0124674–e0124674, 2015.

[11] Michael Felsberg and Gerald Sommer, "The monogenic signal," *Signal Processing*, vol. 49(12), pp. 3136–3144, 2001.

[12] Xiaohua Huang, Guoying Zhao, Wenming Zheng, and Matti Pietikainen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *Signal Processing Letters*, vol. 19(5), pp. 243–246, 2012.

[13] Xiaohua Huang, Qiuhai He, Xiaopeng Hong, Guoying Zhao, and Matti Pietikainen, "Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 514–520.

[14] Yee-Hui Oh, Anh Cat Le Ngo, John See, Sze-Teng Liong, Raphael C-W Phan, and Huo-Chong Ling, "Monogenic riesz wavelet representation for micro-expression recognition," in *Digital Signal Processing (DSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1237–1241.

[15] Oliver Fleischmann, *2D signal analysis by generalized Hilbert transforms*, Thesis, University of Kiel, 2008.

[16] Lennart Wietzke, Oliver Fleischmann, and Gerald Sommer, "2D image analysis by generalized hilbert transforms in conformal space," in *Computer Vision–ECCV 2008*, pp. 638–649. Springer, 2008.

[17] David Hestenes and Garret Sobczyk, *Clifford Algebra to Geometric Calculus: A unified language for mathematics and physics*, vol. 5, Springer Science & Business Media, 2012.