sciendo

Int. J. Appl. Math. Comput. Sci., 2018, Vol. 28, No. 2, 399–409
DOI: 10.2478/amcs-2018-0030

amcs

FACIAL EXPRESSION RECOGNITION UNDER DIFFICULT CONDITIONS: A COMPREHENSIVE STUDY ON EDGE DIRECTIONAL TEXTURE PATTERNS

FAISAL AHMED [a,*], MD. HASANUL KABIR [b]

[a]Department of Computer Science
University of Calgary, 2500 University Drive NW, Calgary, AB, Canada
e-mail: faahmed@ucalgary.ca

[b]Department of Computer Science and Engineering
Islamic University of Technology, Board Bazar, Gazipur 1704, Bangladesh

In recent years, research in automated facial expression recognition has attained significant attention for its potential applicability in human–computer interaction, surveillance systems, animation, and consumer electronics. However, recognition in uncontrolled environments under the presence of illumination and pose variations, low-resolution video, occlusion, and random noise is still a challenging research problem. In this paper, we investigate recognition of facial expression in difficult conditions by means of an effective facial feature descriptor, namely the directional ternary pattern (DTP). Given a face image, the DTP operator describes the facial feature by quantizing the eight-directional edge response values, capturing essential texture properties, such as presence of edges, corners, points, lines, etc. We also present an enhancement of the basic DTP encoding method, namely the compressed DTP (cDTP) that can describe the local texture more effectively with fewer features. The recognition performances of the proposed DTP and cDTP descriptors are evaluated using the Cohn–Kanade (CK) and the Japanese female facial expression (JAFFE) database. In our experiments, we simulate difficult conditions using original database images with lighting variations, low-resolution images obtained by down-sampling the original, and images corrupted with Gaussian noise. In all cases, the proposed method outperforms some of the well-known face feature descriptors.

Keywords: directional ternary pattern, compressed DTP, facial feature descriptor, texture encoding, support vector machine.

1. Introduction

In recent years, with the fast-paced growth of computing technologies, demands for personalized and customizable consumer products and applications are also increasing rapidly. In this context, automated facial expression recognition (FER) is an interesting research direction, since vision-based FER provides an intuitive solution for sensing the emotional states of consumers, thus enabling products or applications to dynamically respond to the situation (Uddin *et al.*, 2009). Thus, a more user-friendly and human-like interaction can be achieved in an unobtrusive manner. Automated facial expression recognition can potentially be used in biometrics, human–computer interaction, social robotics, data-driven animation, and consumer products (Ahmed and Kabir, 2012a; Jabid *et al.*, 2010).

A generic FER system comprises two basic components: (a) facial feature descriptor and (b) classifier. The facial feature descriptor describes the characteristics of a given facial image through features, which is used by the classifier to recognize the corresponding expression class. The success of an FER system critically depends on the discriminating capability of the underlying face feature descriptor (Ahmed and Kabir, 2012a; Ahmed, 2012). Even using the best classifier will result in poor recognition performance, if provided with features having low discriminating ability or inadequate information (Tenne, 2017; Sniezynski, 2015). An effective and discriminating feature descriptor can be characterized as having high inter-class and low intra-class variations

*Corresponding author

(Rivera *et al*., 2013). However, obtaining such feature descriptor for facial expression recognition is difficult due to several reasons, which are (a) differences in pose, alignment, and lighting condition, (b) presence of occlusion and noise, and (c) wearing glasses, ornaments, etc. Hence, recent research efforts mainly focus on designing effective and discriminative feature extraction methods in uncontrolled environments, which constitutes a difficult and challenging task (Tan and Triggs, 2007).

In this paper, we investigate facial feature description by means of a robust local texture pattern, namely the directional ternary pattern (DTP) for FER under uncontrolled and difficult conditions. The proposed method presents an intuitive texture encoding approach that quantizes edge responses obtained from all eight directions of a local neighborhood and thus represents the local texture by means of an encoded pattern. Unlike the other existing texture operators, the proposed DTP operator can differentiate between smooth and high-textured facial regions, which allows the classifier to discard features obtained from non-informative smooth regions, if necessary. We also introduce an enhancement of the basic DTP encoding scheme, namely the compressed DTP (cDTP), which utilizes the symmetric property of Robinson compass masks to quantize edge responses from opposite directions with a single level. Thus, the cDTP operator can effectively compress the original DTP feature vector length, without compromising any loss of texture information. To evaluate the performance of the DTP and cDTP based face feature descriptors under difficult conditions, we designed a set of experiments that includes: (a) original expression images with lighting variations, (b) low resolution images obtained by down-sampling the originals, and (c) images corrupted with Gaussian noise. The objective is to assess the effectiveness and robustness of directional texture patterns in challenging real-world scenarios, with respect to other state-of-the-art expression recognition methods. Experiments with two large databases, namely the Cohn–Kanade (CK) (Kanade *et al*., 2000) and Japanese female facial expression (JAFFE) (Lyons *et al*., 1999) databases demonstrate that, the proposed descriptor is more robust in extracting facial features under difficult conditions and achieves superior recognition performance compared to some state-of-the-art FER methods.

2. Related work

Based on the types of the extracted features, various FER methods found in the literature can roughly be categorized into two classes: (i) geometric feature-based methods and (ii) appearance-based methods (Shan *et al*., 2009). Early studies on facial expression recognition were mostly based on extracting geometric relations among different facial components (Jabid *et al*., 2010). One of the pioneer works by Ekman and Friesen has proposed the facial action coding system (FACS) (1978) which represented the physical behavior of some specific facial muscles through action units. This method of characterizing facial expressions through physical changes in facial muscles or points was later investigated by several researchers. Zhang (1999) utilized the geometric positions of a set of fiducial points, manually selected for each expression image. A similar type of features was adopted by Guo and Dyer (2003), who used linear programming for simultaneous feature selection coupled with classifier training. More recent studies on tracked action unit data conducted by Valstar *et al.* (2005) showed promising recognition performance for geometric methods. Compared with existing appearance-based methods, their method achieved equal or better results in the experiments. However, most geometric FER methods largely depend on accurate detection and tracking of facial components, which is a challenging and difficult task in a dynamic and changing environment. Therefore, applications of geometric FER in real world are quite limited (Jabid *et al*., 2010; Ahmed and Kabir, 2012a).

While geometric methods focuse on specific facial components, appearance-based methods extract a holistic representation of the facial appearance by applying a filter or filter banks on the whole or some specific regions of the face image. Various appearance-based methods can be found in the literature. Among these methods, principal component analysis (PCA) (Padgett and Cottrell, 1997), 2D PCA (Yang *et al*., 2004), independent component analysis (ICA) (Fa and Shin, 2006), Gabor wavelets (Lyons *et al*., 1999), and the more recent enhanced ICA (EICA) (Uddin *et al*., 2009) are the most common ones. In a comprehensive study of existing facial action recognition methods, Donato *et al.* (1999) investigated PCA, ICA, Gabor wavelets, local principal components (PCs), and local feature analysis (LFA). In their study, Gabor wavelets and ICA achieved the highest recognition performance. However, according to Jabid *et al.* (2010), the performance of PCA and ICA deteriorates in dynamic conditions, while the facial appearance representation based on Gabor wavelets is computation and memory expensive (Kabir *et al*., 2012).

In recent years, facial appearance descriptors based on the local binary pattern (LBP) (Shan *et al*., 2009; Ojala *et al*., 2002) and its variants (Ahmed and Kabir, 2012a; Guo *et al*., 2010) have attained significant attention. This is due to the basic advantages provided by the local texture operators, which are charaterized by relatively low computational complexity and better robustness to lighting and pose changes (Zhao and Pietikainen, 2009; Jabid *et al*., 2012). The local binary pattern operator encodes the texture information of a small local neighborhood into an eight-bit binary pattern, which acts as a template for detecting micro-level texture details.

However, it is sensitive to large illumination variations and random noise. Tan and Triggs (2007) addressed this issue by introducing an extra level in their local ternary pattern encoding. Zhao *et al.* (2008) proposed to use Sobel masks prior to applying the LBP operator in order to facilitate more robust texture encoding. The descriptor based on the local directional pattern (LDP) (Jabid *et al.*, 2010) introduced a different texture encoding approach that exploits eight-directional edge responses instead of gray levels. Although LDP yields better recognition performance than LBP, it tends to produce inconsistent texture response in smooth facial regions (Ahmed and Kabir, 2012a; Ahmed, 2012).

3. Proposed method

There are several steps involved in the proposed facial expression recognition method. First, a face detection algorithm is applied to detect faces in an image scene and the detected faces are then cropped from the original images. The Viola–Jones face detection method (Viola and Jones, 2004) is used for this purpose. After that, Robinson masks are applied to obtain the eight-directional edge response values for each pixel in the face image. These edge response values are then quantized using the proposed directional ternary pattern (DTP) operator and a feature descriptor for the corresponding face image is constructed based on the DTP codes. Finally, a support vector machine is used for the classification task. Figure 1 illustrates the components of the proposed method.

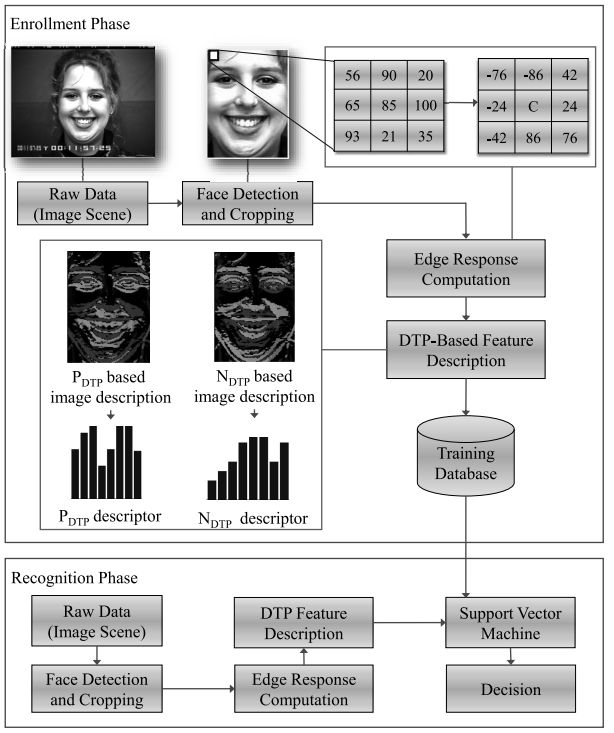

Fig. 1. Overview of the proposed methodology.

3.1. Directional ternary pattern (DTP).

3.1.1. DTP encoding scheme. As previous research (Chen *et al.*, 2000; Ling *et al.*, 2007) demonstrated, edge responses are largely insensitive to illumination variations. Consequently, an encoding scheme that exploits the edge responses in different directions can retain more information of the local region, since it holds the relations among pixels implicitly which is not available in the intensity space (Rivera *et al.*, 2013). Hence, the proposed directional ternary pattern (DTP) operator encodes the local texture by assigning a three-valued code to each pixel based on the edge response values in different directions about the center pixel. Here, eight directional edge response values are computed by Robinson masks centered on a pixel oriented in eight different directions as shown in Fig. 2.

By applying these eight masks, we obtain eight edge response values, each of which represents the edge significance in its corresponding direction. The presence of an edge or a corner will produce high edge-response values in their respective directions. On the other hand, uniform or smooth regions will produce edge response values of similar or same magnitudes in different directions. Therefore, unlike the LDP operator (Jabid *et al.*, 2010) that always sets the most prominent k directions to 1 and others to 0 for forming a binary code regardless of the local region being uniform or not, the DTP operator employs a ternary coding scheme that differentiates between the smooth and high textured regions using a threshold. After computing the average μ of all eight directional edge response values, those responses within a $\pm t$ margin about the mean μ are quantized to 0, those above $\mu + t$ and those below $\mu - t$ are quantized to $+1$ and -1, respectively, i.e.,

$$S_{DTP}(r_i) = \begin{cases} 1, & r_i > \mu + t, \\ 0, & \mu - t \le r_i \le \mu + t, \\ -1, & r_i < \mu - t. \end{cases} \quad (1)$$

Here, r_i is the edge response value in the i-th direction, μ is the average edge response value, and t is a threshold value. We illustrate the basic DTP encoding scheme in Fig. 3. Here, the positive and negative edge response values indicate the gradient direction of light and dark areas of the neighborhood, which reveals important information regarding the underlying structure of the locality (Rivera *et al.*, 2013). This distinction between light and dark responses is implicitly captured in the DTP code by the $+1$ and -1 discrimination levels, respectively, which enables DTP to differentiate between blocks with light and dark areas swapped, by

generating different DTP code for each individual case. Unlike DTP, other gradient-based methods (Jabid *et al.*, 2010; Zhao *et al.*, 2008) ignore the sign information of the edge response values, and thus, they fail to differentiate between certain facial regions (such as top and bottom edges of eyebrows), which have different intensity transitions (light to dark and dark to light, respectively). Figure 4 shows an example of the local binary pattern (LBP) and the proposed directional ternary pattern (DTP) encoding scheme applied on a small image patch under the presence of Gaussian white noise. It can be observed that, introducing the noise results the LBP code to change from its original pattern, while the DTP was able to retain the original relationships of the pixels in the local neighborhood. We argue that there are two contributing factors to the robustness of the proposed DTP encoding scheme: first, utilization of the eight-directional edge response values instead of raw gray scale values to represent the local neighborhood with a ternary pattern, which provides robustness under the presence of nonmonotonic changes. Secondly, utilization of the threshold t improves the overall stability by differentiating between smooth and high-textured regions.

In order to reduce the length of the feature vector, each DTP code is further split into its corresponding positive (which represents the direction of light area) and negative parts (which represents the direction of dark area), and treated as two separate binary patterns, namely P_{DTP} and N_{DTP}. Thus, the number of features reduces from $3^8(= 6561)$ to $2 \times 2^8(= 512)$. Here, P_{DTP} and N_{DTP} take the following form:

$$P_{DTP} = \sum_{i=0}^{7} S_P(S_{DTP}(r_i))2^i, \tag{2a}$$

$$S_P(v) = \begin{cases} 1, & v = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2b}$$

$$N_{DTP} = \sum_{i=0}^{7} S_N(S_{DTP}(r_i))2^i, \tag{3a}$$

$$S_N(v) = \begin{cases} 1, & v = -1, \\ 0, & \text{otherwise}, \end{cases} \tag{3b}$$

3.1.2. Adaptive threshold selection. The resultant texture patterns from the DTP operator is dependent on the selection of an effective threshold that can differentiate between the smooth and high-textured regions. In our earlier works (Ahmed and Kabir, 2012a; 2012b), a global threshold was found to be sufficient in achieving stable recognition performance. In this paper, we also present an adaptive local threshold selection method which derives the t value based on the edge response statistics of the local neighborhood. The proposed adaptive threshold can

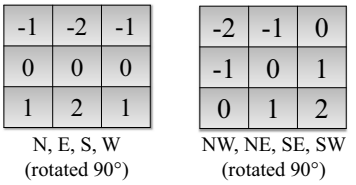

Fig. 2. Robinson eight directional edge response masks. Different orientations are obtained by rotating these masks by 90°. Here, N, S, E, and W correspond to North, South, East, and West, respectively.

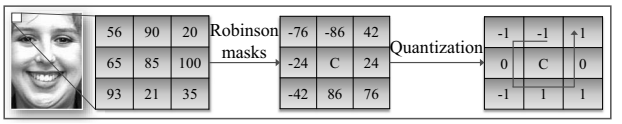

Fig. 3. DTP encoding scheme for $t = 40$. Here, the obtained DTP code for C is $(-1)(-1)0(-1)1101$.

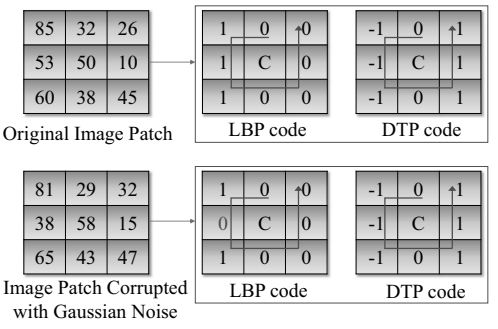

Fig. 4. Example of DTP (t = 40) vs. LBP encoding under the presence of Gaussian white noise. It can be observed that, while the LBP code has been affected by the noise, DTP was able to retain the same texture pattern.

be defined as

$$t = \sigma \times \alpha. \tag{4}$$

Here, σ represents the standard deviation of the eight directional edge response magnitudes and α is a scaling factor. This approach thus adjusts the threshold value based on the illumination condition as well as the presence of texture information in the local neighborhood, resulting in more robust DTP codes.

3.1.3. DTP face descriptor. Applying the DTP operator on all the pixels of an image, we get two encoded images, one for the P_{DTP} code and the other for the N_{DTP} code. First, histograms are computed from these two encoded images using

$$H_{P_{DTP}}(i) = \sum_{x=1}^{M} \sum_{y=1}^{N} f(P_{DTP}(x, y), i), \tag{5}$$

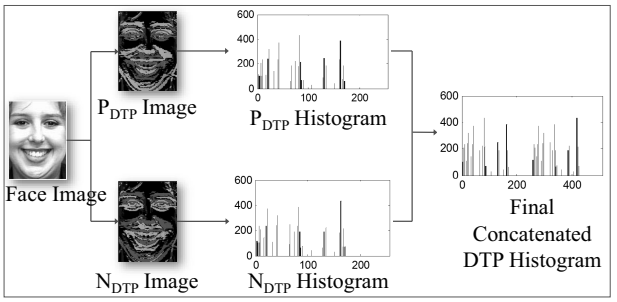

Fig. 5. DTP histogram generation process.

$$H_{N_{DTP}}(i) = \sum_{x=1}^{M}\sum_{y=1}^{N} f(N_{DTP}(x,y),i), \tag{6}$$

$$f(a,i) = \begin{cases} 1, & a = i, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Here, i is the positive or negative DTP code value computed on the (x, y) pixel of an $M \times N$ encoded image. Histograms computed from the P_{DTP} and N_{DTP} encoded images are then concatenated spatially to produce the DTP histogram, which represents the occurrence information of the P_{DTP} and N_{DTP} binary patterns. The process is shown in Fig. 5.

Histograms generated from the whole encoded image merely express the occurrence frequencies of the generated micro-patterns. It contains no locality information of those patterns. However, the presence of location information and spatial relationships provides a better facial feature representation and describes the image content more accurately (Jabid *et al.*, 2010; Ahonen *et al.*, 2006; Gundimada and Asari, 2009). Therefore, the DTP histogram is modified to an extended histogram in order to incorporate some degree of location information. First, each image is partitioned into a number of regions and individual DTP histograms are generated from each of those regions. Finally, the histograms of all the regions are concatenated to obtain an extended DTP histogram. For the facial expression recognition process, this histogram collection is used as the face feature vector. The process is illustrated in Fig. 6.

3.2. Compressed DTP: A fast DTP encoding scheme.

The strength of the DTP features lies in the quantization of stable edge response values and the discrimination between smooth and high-textured facial regions. To compute the eight directional edge responses, the DTP operator uses the Robinson compass masks. The reason is that Robinson masks are easier to implement than the other compass masks (such as Kirsch masks), since they rely only on the coefficients of 0, 1, and 2, and are symmetrical about their directional axis (Umbaugh, 2011). Therefore, only the responses on four of the masks are required to be computed. The responses from the other four masks can be obtained by simply negating the responses obtained from the first four (Umbaugh, 2011).

Assume that R_N and R_S are the edge response values from any two opposite directions, such as North and South, respectively, computed on a 3×3 local neighborhood. Due to the symmetric property of the Robinson masks, the magnitude of the edge responses R_N and R_S will always be the same, only the sign will be the opposite (for magnitudes greater than 0). Hence, the mean μ of the eight-directional edge response values around the center point of a local neighborhood will always be 0. As a result, while quantizing the edge response values with a threshold t, the edge responses from any two opposite directions (R_N and R_S) will always satisfy one of the following three instances:

1. If $(\mu - t) \leq R_N \leq (\mu + t)$, then $(\mu - t) \leq R_S \leq (\mu + t)$;
2. If $R_N > (\mu + t)$, then $R_S < (\mu - t)$;
3. If $R_N < (\mu - t)$, then $R_S > (\mu + t)$.

From Fig. 7 it can be observed that, if R_N lies within the threshold region $\mu \pm t$, R_S will also lie within the same region. Otherwise, edge responses from opposite two directions will lie at the opposite two sides of the threshold region. Therefore, instead of quantizing both of the edge response values from two opposite directions, we can quantize a single one without any loss of information (due to their symmetric property). Hence, we propose a variant of the basic DTP encoding scheme, namely the compressed DTP (cDTP) encoding, where each time two opposite edge response values are selected and a single discrimination level (0, 1, or 2, based on the cases listed above) is used to label both. Thus, we need only a 4-digit base-3 number to encode the eight-directional edge response values around the center of a local neighborhood, which effectively reduces the number of possible DTP patterns from $3^8 (= 6561)$ to $3^4 (= 81)$. The encoding scheme can be represented formally as follow:

$$cDTP = \sum_{i=0}^{3} S_{cDTP}(R_i) \times 3^i, \tag{8}$$

$$S_{cDTP}(R_i) = \begin{cases} 0, & \mu - t \leq R_i \leq \mu + t, \\ 1, & R_i > \mu + t, \\ 2, & R_i < \mu - t. \end{cases} \tag{9}$$

Here, R_0, R_1, R_2, and R_3 are the edge response values from the North, East, North-East, and North-West directions, respectively. The difference between the basic DTP and the compressed DTP encoding scheme is that, in the cDTP encoding, we use one of the three discrimination levels (0, 1, and 2) for labeling edge response values from two opposite directions simultaneously (which is then

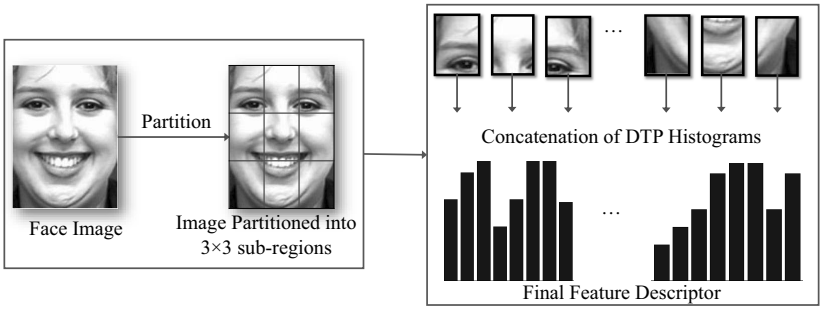

Fig. 6. DTP feature vector construction process.

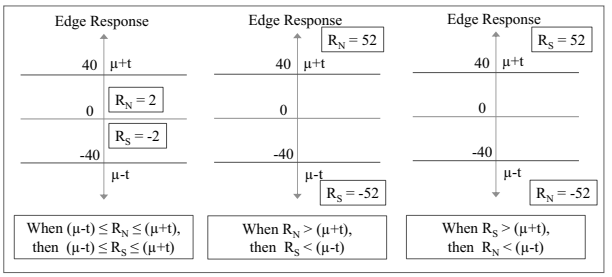

Fig. 7. Illustration of the possible three conditions for different values of R_N and R_S ($t = 40$).

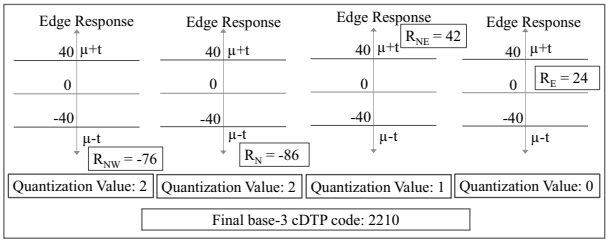

Fig. 8. cDTP encoding scheme ($t = 40$) for the 3×3 neighborhood shown in Fig. 3. Here, the final cDTP code is obtained by concatenating the quantization results of edge responses from North-West (R_{NW}), North (R_N), North-East (R_{NE}), and East (R_E), which is 2210.

concatenated to form a base-3 number), while in the basic DTP encoding, each edge response is labeled individually. The compressed DTP encoding is illustrated in Fig. 8.

Applying the compressed DTP operator to all the pixels of an image will result in an encoded cDTP image, where the value of each pixel will range between 0 and 80. The distribution information of these compressed DTP micro-patterns are then represented as a spatial histogram, namely the cDTP histogram. In order to incorporate location information of the cDTP micro-patterns, the cDTP histogram is modified to an extended histogram using the same method as the DTP. The extended cDTP histogram is then used as the facial feature vector for the classifier training and testing.

4. Experiments and results

4.1. Experimental setup and dataset description. The performance of the proposed method was evaluated based on its ability to recognize a set of prototypic emotional expressions, which includes anger, disgust, fear, joy, sadness, and surprise. By introducing additional neutral face expression images, this 6-class recognition problem can further be extended to a 7-class problem. The experiments were carried out on two well-known databases, namely the Cohn–Kanade (CK) database (Kanade *et al*., 2000) and the Japanese female facial expression (JAFFE) database (Lyons *et al*., 1999).

In the CK database, a sample set of 100 students, aging from 18 to 30 during image acquisition, was included. A majority of the subjects (65%) were female; 15% of the samples were African American, and 3% were Asian or of Latin descent. Each of the students displayed facial expressions starting from nonexpressiveness to one of the aforementioned six prototypic emotional expressions in the image acquisition process. These image sequences were then digitized into 640×480 or 640×690 pixel resolutions. In our setup, a set of 1224 facial image sequences were selected from 96 subjects and each of the images were given a label describing the subject's facial expression. The dataset containing the 6-classes of expressions was then extended by 408 images of neutral facial images to obtain the 7-class expression dataset.

The JAFFE database comprises facial expression images of 10 Japanese female subjects. All the images were digitized into a resolution of 256×256 pixels. The images were obtained from a frontal pose, and to ensure the exposure of all the expressive regions of the face, the subjects' hair was tied back. During image acquisition, tungsten lights were used to create an even illumination. Instead of revealing the actual names, the subjects are referred with their initials: KA, KL, KM, KR, MK, NA, NM, TM, UY, and YM. In our setup, the 6-class expression dataset consists of a total of 283 images, while the 7-class expression set includes additional 50 neutral

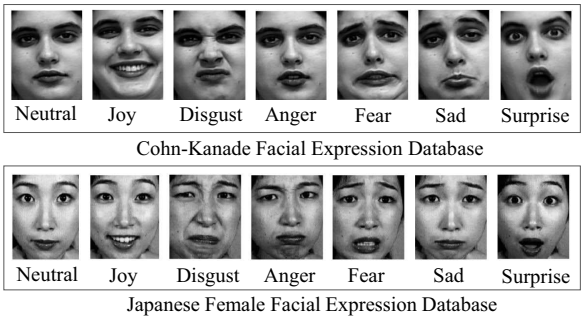

Fig. 9. Facial images from the CK and JAFFE databases.

expression images. Figure 9 shows the sample prototypic expression images from the CK and JAFFE databases.

The selected facial images were cropped from the original ones based on a bounding box detected using the Viola–Jones face detection method (Viola and Jones, 2004) and then normalized to 150×110 pixels. No attempt was made to remove illumination changes, since one of the objectives of our experiments is to demonstrate the effectiveness of DTP under lighting variations. We carried out a ten-fold cross-validation to compute the classification rate. In ten-fold cross-validation, ten subsets comprising equal numbers of instances are formed by partitioning the whole dataset randomly. The classifier is first trained on the nine subsets and then the remaining set is used for testing. This process is repeated 10 times and then the average classification rate is computed. A support vector machine (SVM) equipped with a radial basis function (RBF) kernel was used as the classifier (Yao *et al.*, 2014). A grid-search cross-validation was carried out for each facial descriptor in order to select appropriate hyper parameter values (C and γ), as suggested by Hsu and Lin (2002). Hence, for each descriptor, an optimal parameter setting is adopted. For the proposed DTP and cDTP operators, the optimal values of C and γ were found to be 1.0 and 0.105, respectively.

The performance of the proposed method can be influenced by adjusting two parameters: the threshold selection method (global or local adaptive) and the number of regions into which the expression images are to be partitioned. In literature, the commonly-used number of regions are 3×3, 5×5, 7×7, 7×6, and 9×8. In our experiments, we consider three different cases where the images were partitioned into 3×3, 5×5, and 7×6 regions. In order to determine the optimal global threshold t, we first fixed the number of regions to 3×3, and then searched for a t value that achieves the best recognition performance for the CK 6-class dataset. In our experiments, the highest classification rate was achieved for $t = 40$, hence the value of t was set to 40 for the global thresholding technique in all other experiments. For the proposed local adaptive thresholding, the scaling factor α was empirically set to 0.8.

4.2. Experimental results for the CK dataset. We have compared the performance of the proposed method with three well-known local pattern operators, namely local binary pattern (LBP) (Shan *et al.*, 2009), local ternary pattern (LTP) (Tan and Triggs, 2007), and local directional pattern (LDP) (Jabid *et al.*, 2010). Tables 1 and 2 show the recognition rates of these local pattern-based feature descriptors against the 6-class and the 7-class expression datasets, respectively. In both cases, DTP and cDTP exhibit superior performance in recognizing expression images. For the 6-class dataset, DTP and cDTP (with local thresholding) attain the highest recognition rates of 97.5% and 97.6%, respectively. On the other hand, for the 7-class dataset, the highest recognition rates of DTP and cDTP (both with local thresholding) are 96.0% and 95.3%, respectively. Here, inclusion of neutral expression images results in a decrease in the accuracy. For both the 6-class and the 7-class recognition problem, the highest classification rate is obtained for images partitioned into 7×6 regions.

Table 1. Confusion matrix for the CK 6-class recognition using the DTP feature descriptor (for images partitioned into 7×6 regions). Rows represent true classes and columns represent classification rates (%).

	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sad (%)	Surprise (%)
Anger	**97.7**	0	0	0	0	2.3
Disgust	0	**97.9**	0	1.6	0	0.5
Fear	0	0.4	**96.0**	0	0	3.6
Joy	0.6	0.5	0	**98.9**	0	0
Sad	0	0	0	0	**100**	0
Surprise	0	0	0	3.4	0	**96.6**

Table 2. Confusion matrix for the CK 7-class recognition using the DTP feature descriptor (for images partitioned into 7×6 regions).

	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sad (%)	Surprise (%)	Neutral (%)
Anger	**98.5**	0	0	0	0	0	1.5
Disgust	0	**98.0**	0	0	0	2.0	0
Fear	0	0.4	**98.5**	0	0	1.1	0
Joy	0.5	0.5	0	**93.2**	0	5.8	0
Sad	0	0	0	0	**93.4**	0	6.6
Surprise	0	0	0	1.9	0	**98.1**	0
Neutral	5.9	0	0	0	3.0	0	**91.1**

It can be observed that dividing an image into a higher number of regions results in a higher classification rate, since the feature descriptor then contains more location and spatial information of the local patterns. Nevertheless, the feature vector length is also higher in such cases, which affects the computational efficiency.

In addition, dividing an image into too many regions will result in a decrease in the recognition performance, since then the texture histograms of those very small local regions will fail to represent any significant texture information. In our case, the recognition rate decreased for images partitioned into more than $9 \times 8 (= 72)$ regions. Hence, selection of the number of regions is a trade-off between computational efficiency and classification rate.

The confusion matrix of recognition using the original DTP descriptor for the CK 6-class and the CK 7-class datasets are shown in Tables 3 and 4, respectively. These two tables provide a better picture of the recognition performance of the DTP descriptor for individual expression types. It can be observed that, for the 6-class recognition problem, all the expression types can be recognized with a high accuracy. On the other hand, for the 7-class recognition, while anger, disgust, fear, and surprise can be recognized with high accuracy, the recognition rates of joy, sadness, and neutral expressions are lower than the average. Here, the neutral expression images are confused with anger and sad and vice versa, which results in a decrease in the average recognition performance. In addition, the class labels in the CK database represent what the participants were asked to perform, rather than what were performed originally (Kanade *et al.*, 2000). This could also potentially lead to a mislabeling of some training data, causing cross-class errors.

Table 3. Recognition rate (%) for the CK 6-class dataset.

Feature descriptor	Number of regions		
	3×3	5×5	7×6
LBP	79.3	89.7	90.1
LTP	91.3	92.3	94.6
LDP	80.2	91.9	93.7
Original DTP	94.5	97.1	97.5
Original cDTP	94.2	96.8	97.3
DTP with adaptive threshold	**94.7**	**97.2**	97.3
cDTP with adaptive threshold	94.6	96.5	**97.6**

Table 4. Recognition rate (%) for the CK 7-class dataset.

Feature descriptor	Number of regions		
	3×3	5×5	7×6
LBP	73.8	80.9	83.3
LTP	85.3	88.5	88.9
LDP	75.7	86.3	88.4
Original DTP	90.3	**93.9**	95.8
Original cDTP	89.9	93.5	95.2
DTP with adaptive threshold	**90.7**	93.2	**96.0**
cDTP with adaptive threshold	88.7	93.7	95.3

4.3. Experimental results for the JAFFE dataset. For the JAFFE 6-class dataset, the global threshold-based DTP and cDTP feature descriptors achieve the highest classification rates of 92.5% and 91.2%, respectively. For the 7-class dataset, the highest recognition rates of 88.9% and 88.2% are obtained for the local thresholding-based DTP and cDTP descriptors, respectively. Tables 5 and 6 show the comparison of the recognition performances of the different feature descriptors for the JAFFE 6-class and the 7-class datasets, respectively. It can be observed that DTP and cDTP achieve the best two recognition rates for both datasets. Here, too the highest classification rate is obtained for images partitioned into 7×6 regions. The recognition performance in the JAFFE database are relatively lower than the CK database. The reason is that in the JAFFE database, some of the expression images are labeled with incorrect class labels or expressed incorrectly by the subjects (Jabid *et al.*, 2010).

Table 5. Recognition rate (%) for the JAFFE 6-class dataset.

Feature descriptor	Number of regions		
	3×3	5×5	7×6
LBP	84.1	87.6	90.5
LTP	84.3	87.9	90.9
LDP	83.2	88.9	90.7
Original DTP	87.5	**90.1**	**92.5**
Original cDTP	87.1	89.8	91.2
DTP with adaptive threshold	**87.7**	89.9	92.3
cDTP with adaptive threshold	86.9	89.5	91.0

Table 6. Recognition rate (%) for the JAFFE 7-class dataset.

Feature descriptor	Number of regions		
	3×3	5×5	7×6
LBP	81.5	82.3	85.3
LTP	84.6	85.0	86.7
LDP	83.3	85.3	85.9
Original DTP	85.3	**86.9**	88.7
Original cDTP	85.9	86.5	88.0
DTP with adaptive threshold	85.4	86.8	**88.9**
cDTP with adaptive threshold	**86.1**	86.3	88.2

4.4. Experimental results for low-resolution images. Automated facial expression analysis is useful in smart meeting, surveillance, and many other applications (Jabid *et al.*, 2010), where often only low-resolution video data are available. Since geometric-feature based methods like detection of facial action units or fiducial points are difficult to accommodate in these scenarios, appearance-based methods seem to be a better solution. Therefore, the performance of the proposed method is also evaluated on low-resolution images. Experiments were conducted on images from the CK 6-class expression dataset. We considered three different image resolutions: 75×55, 48×36, and 37×27. Sample expression images are shown in Fig. 10.

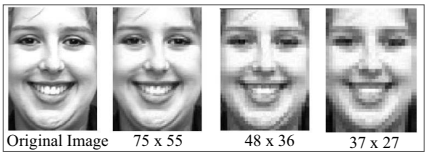

Fig. 10. Sample low resolution images used in the experiments.

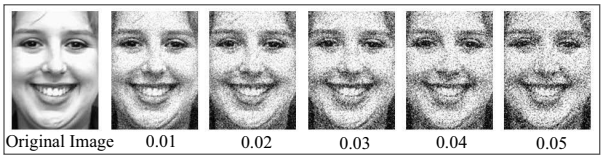

Fig. 11. Sample expression images contaminated with Gaussian white noise with zero mean and different variances.

The original images were down sampled to obtain these low-resolution images. All the images were partitioned into 7×6 regions while applying the texture operators. In this experiment also, the performances of the DTP and cDTP feature descriptors are compared with LBP, LTP, and LDP. Table 7 shows the recognition rates of different descriptors for low resolution expression images. From the recognition accuracy, it is evident that facial feature representation based on DTP and cDTP is more robust than other existing local texture patterns over a useful range of low resolutions.

4.5. Experimental results for noisy images. To investigate the robustness of the proposed DTP descriptor under the presence of noise, further experiments were conducted on the images from the CK 6-class expression dataset. In the experimental setup, the images in the testing set were contaminated with Gaussian white noise of different variances (as shown in Fig. 11), while the training samples were kept unchanged. All the images were partitioned into 7×6 regions during feature vector generation. Table 8 shows the corresponding recognition rates of LBP, LTP, LDP, and DTP against images corrupted with Gaussian white noise with zero mean and different variances (0.01, 0.02, 0.03, 0.04, and 0.05). It can be observed that in all cases DTP and cDTP achieve significantly higher recognition rates than the other texture operators. The superiority of DTP encoding is due to the utilization of stable edge responses and its discriminating capability of smooth and high textured areas from different face regions.

4.6. Discussion. The experimental results validate that the DTP and cDTP-based feature representation perform consistently better than some widely-used appearance-based face descriptors, even under the presence of illumination variations, random noise and

Table 7. Recognition rates (%) for low-resolution images from the CK 6-class dataset.

Feature descriptor	Image resolution		
	75x55	48x36	37x27
LBP	88.9	83.5	79.7
LTP	89.7	85.9	83.3
LDP	90.7	89.1	84.4
Original DTP	**93.9**	92.2	89.1
Original cDTP	93.5	91.9	88.7
DTP (adaptive threshold)	93.7	**92.4**	**89.2**
cDTP (adaptive threshold)	93.1	91.8	88.9

Table 8. Recognition rate (%) on images from the CK 6-class dataset corrupted with Gaussian white noise with zero mean and different variances.

Feature descriptor	Noise variance				
	0.01	0.02	0.03	0.04	0.05
LBP	73.7	66.7	64.5	62.3	61.9
LTP	77.1	70.4	67.3	65.0	62.3
LDP	70.9	61.3	55.1	52.4	48.2
Original DTP	87.7	**79.7**	73.4	69.6	66.5
Original cDTP	87.5	79.3	73.1	68.8	65.3
DTP (adaptive)	**87.8**	79.5	**73.6**	**69.8**	**66.6**
cDTP (adaptive)	87.6	79.2	73.3	68.7	65.2

in low resolution images. In a controlled environment with no illumination normalization, the proposed texture descriptors can attain an average recognition rate of 95.05% for the CK 6-class and JAFFE 6-class datasets and an average recognition rate of 92.45% for the CK 7-class and JAFFE 7-class datasets, while the existing texture operators can attain an average of 92.75% and 87.8% at most, respectively. On the other hand, under the presence of random noise and low resolution images, the performance of the proposed method is significantly higher than existing approaches. A summary of the recognition performances of the proposed method against other existing face descriptors can be found in Table 9. On the other hand, Table 10 shows the results of statistical two-sample t-tests performed on the recognition rates (%) obtained for DTP, compared with other facial descriptors. Here, for the t-tests performed on the (DTP vs. LBP) and (DTP vs. LDP) pairs, the null hypotheses was rejected due to low p-values of 0.0074 and 0.0321, respectively, which indicates the recognition rates obtained for the DTP descriptor is statistically significant with respect to LBP and LDP. However, the p-value obtained for the (DTP vs. LTP) was found to be higher than the threshold 0.05. Although computation of the eight directional edge response values makes the proposed texture descriptors more expensive than pure gray-level based methods like LBP, utilization of the symmetric property of the Robinson compass masks enables the computation of DTP features faster than other edge

Table 9. A summary of performances of different face descriptors under various conditions.

Feature descriptor	Average recognition rates (%) for different datasets			
	All 6-class datasets	All 7-class datasets	Low-resolution	Gaussian noise
LBP	86.90	81.18	84.03	65.82
LTP	90.21	86.50	86.30	68.42
LDP	88.10	84.15	88.06	57.58
Original DTP	**93.20**	90.15	91.73	75.38
Original cDTP	92.73	89.83	91.37	74.80
DTP (adaptive threshold)	93.18	**90.16**	**91.76**	**75.46**
cDTP (adaptive threshold)	92.68	89.71	91.27	74.80

Table 10. Two-sample t-test results for DTP vs. other methods.

Descriptors	p-value	Reject null hypotheses?
DTP vs. LBP	0.0074	Yes
DTP vs. LTP	0.1255	No
DTP vs. LDP	0.0321	Yes

response-based texture patterns, such as LDP (Jabid *et al*., 2010) and LDPv (Kabir *et al*., 2012).

5. Conclusion

This paper describes a local facial feature descriptor based on DTP for expression recognition. The DTP operator integrates the local edge responses for texture encoding, and also discriminates between smooth and non-smooth areas with three different levels. We also present a variant of DTP encoding, namely the compressed DTP, which can effectively reduce the feature vector length without any significant loss of information, resulting in almost similar recognition performance with a very low computational cost. In our experiments, both DTP and cDTP achieve superior performance than some widely-used feature descriptors. The effectiveness of the proposed method is due to its texture discriminating capability, robustness under illumination variations and the presence of noise the compared the existing representations. Therefore, it can also be used for face recognition and gender classification systems in consumer products.

References

Ahmed, F. (2012). Gradient directional pattern: A robust feature descriptor for facial expression recognition, *IET Electronics Letters* **48**(19): 1203–1204.

Ahmed, F. and Kabir, M.H. (2012a). Directional ternary pattern (DTP) for facial expression recognition, *IEEE International Conference on Consumer Electronics, Berlin, Germany*, pp. 265–266.

Ahmed, F. and Kabir, M.H. (2012b). Facial feature representation with directional ternary pattern (DTP): Application to gender classification, *Proceedings of the IEEE International Conference on Information Reuse and Integration, Las Vegas, NV, USA*, pp. 159–164.

Ahonen, T., Hadid, A. and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(12): 2037–2041.

Chen, H., Belhumeur, P. and Jacobs, D. (2000). In search of illumination invariants, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA*, Vol. 1, pp. 254–261.

Donato, G., Bartlett, M.S., Hagar, J.C., Ekman, P. and Sejnowski, T.J. (1999). Classifying facial actions, *IEEE Transactions on Pattern Analysis and Machince Intelligence* **21**(10): 974–989.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA.

Fa, C.C. and Shin, F.Y. (2006). Recognizing facial action units using independent component analysis and support vector machine, *Pattern Recognition* **39**(9): 1795–1798.

Gundimada, S. and Asari, V.K. (2009). Facial recognition using multisensor images based on localized kernel eigen spaces, *IEEE Transactions on Image Processing* **18**(6): 1314–1325.

Guo, G.D. and Dyer, C.R. (2003). Simultaneous feature selection and classifier training via linear programming: A case study for face expression recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, USA*, pp. 346–352.

Guo, Z., Zhang, L. and Zhang, D. (2010). Rotation invariant texture classification using LBP variance (LBPV) with global matching, *Pattern Recognition* **43**(3): 706–719.

Hsu, C.W. and Lin, C.J. (2002). A comparison on methods for multiclass support vector machines, *IEEE Transactions on Neural Networks* **13**(2): 415–425.

Jabid, T., Kabir, M.H. and Chae, O. (2010). Robust facial expression recognition based on local directional pattern, *ETRI Journal* **32**(5): 784–794.

Jabid, T., Kabir, M.H. and Chae, O. (2012). Local directional pattern (LDP) for face recognition, *International Journal of Innovative Computing, Information and Control* **8**(4): 2423–2437.

Kabir, H., Jabid, T. and Chae, O. (2012). Local directional pattern variance (LDPV): A robust feature descriptor for facial expression recognition, *International Arab Journal of Information Technology* **9**(4): 382–391.

Kanade, T., Cohn, J. and Tian, Y. (2000). Comprehensive database for facial expression analysis, *Proceedings of the IEEE International Conference on Automated Face and Gesture Recognition, Grenoble, France*, pp. 46–53.

Ling, H., Soatto, S., Ramanathan, N. and Jacobs, D. (2007). A study of face recognition as people age, *Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil*, pp. 1–8.

Lyons, M.J., Budynek, J. and Akamatsu, S. (1999). Automatic classification of single facial images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(12): 1357–1362.

Ojala, T., Pietikainen, M. and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(7): 971–987.

Padgett, C. and Cottrell, G. (1997). Representing face images for emotion classification, *in* M. Mozer *et al.* (Eds.), *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, Cambridge, MA, pp. 894–900.

Rivera, A.R., Castillo, J.R. and Chae, O. (2013). Local directional number pattern for face analysis: Face and expression recognition, *IEEE Transactions on Image Processing* **22**(5): 1740–1752.

Shan, C., Gong, S. and McOwan, P.W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study, *Image and Vision Computing* **27**(6): 803–816.

Sniezynski, B. (2015). A strategy learning model for autonomous agents based on classification, *International Journal of Applied Mathematics and Computer Science* **25**(3): 471–482, DOI: 10.1515/amcs-2015-0035.

Tan, X. and Triggs, B. (2007). Enhanced local texture feature sets for face recognition under difficult lighting conditions, *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures, Rio de Janeiro, Brazil*, pp. 168–182.

Tenne, Y. (2017). Machine-learning in optimization of expensive black-box functions, *International Journal of Applied Mathematics and Computer Science* **27**(1): 105–118, DOI: 10.1515/amcs-2017-0008.

Uddin, M.Z., Lee, J.J. and Kim, T.S. (2009). An enhanced independent component-based human expression recognition from video, *IEEE Transactions on Consumer Electronics* **55**(4): 2216–2224.

Umbaugh, S.E. (2011). *Digital Image Processing and Analysis*, CRC Press, Boca Raton, FL.

Valstar, M., Patras, I. and Pantic, M. (2005). Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data, *Proceedings of the IEEE CVPR Workshop, San Diego, CA, USA*, Vol. 3, pp. 76–84.

Viola, P. and Jones, M. (2004). Robust real-time face detection, *International Journal of Computer Vision* **57**(2): 137–154.

Yang, J., Zhang, D., Frangi, A. and Yang, J. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(1): 131–137.

Yao, B., Hu, P., Zhang, M. and Jin, M. (2014). A support vector machine with the tabu search algorithm for freeway incident detection, *International Journal of Applied Mathematics and Computer Science* **24**(2): 397–404, DOI: 10.2478/amcs-2014-0030.

Zhang, Z. (1999). Feature-based facial expression recognition: Sensitivity analysis and experiment with a multi-layer perceptron, *International Journal of Pattern Recognition and Artificial Intelligence* **13**(6): 893–911.

Zhao, G. and Pietikainen, M. (2009). Boosted multi-resolution spatiotemporal descriptors for facial expression recognition, *Pattern Recognition Letters* **30**(12): 1117–1127.

Zhao, S., Gao, Y. and Zhang, B. (2008). Sobel-LBP, *Proceedings of the IEEE International Conference on Image Processing, San Diego, CA, USA*, pp. 2144–2147.

Faisal Ahmed received his MSc degree in computer science from the University of Calgary, Canada, and his BSc in computer science and information technology from the Islamic University of Technology, Bangladesh. He is currently working as a data scientist in FarmersEdge, Canada. In the past, he has served as an instructor at the Islamic University of Technology and as a research assistant at the University of Calgary. His research interests include computer vision, machine learning, and data mining.

Md. Hasanul Kabir received his PhD in computer engineering from Kyung Hee University, South Korea, and his BSc in computer science and information technology from the Islamic University of Technology, Dhaka, Bangladesh. He is currently working as a professor at the Department of Computer Science and Engineering, Islamic University of Technology. His research interests include texture feature extraction, motion estimation, computer vision, medical image processing, and pattern recognition.

Received: 25 February 2017
Revised: 5 September 2017
Re-revised: 22 October 2017
Accepted: 2 November 2017