https://doi.org/10.1016/j.jvcir.2018.05.024

Histogram-based Local Descriptors for Facial Expression Recognition (FER): A comprehensive Study

Cigdem Turan*, Kin-Man Lam

Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

*Corresponding author.

E-mail addresses: cigdem.turan@connect.polyu.hk (C. Turan), enkmlam@polyu.edu.hk (K.-M. Lam)

Postal address: DE503, The Hong Kong Polytechnic University, 11 Yuk Choi Rd, Hung Hom, Hong Kong

ABSTRACT

This paper aims to present histogram-based local descriptors applied to Facial Expression Recognition (FER) from static images, and provide a systematic review and analysis of them. First, we describe the main steps in encoding binary patterns in a local patch, which are required in every histogram-based local descriptor. Then, we list the existing local descriptors, while analysing their strengths and weaknesses. Finally, we present the experimental results of all these descriptors on commonly used facial expression databases, with varying resolution, noise, occlusion, and number of sub-regions, as well as comparing them with the results obtained by the state-of-the-art deep learning methods. This paper aims to bring together different studies of the visual features for FER by evaluating their performances under the same experimental setup, and critically reviewing various classifiers making use of the local descriptors.

1. Introduction

Facial expressions, which are an important aspect of non-verbal communication, have been extensively studied in different fields, such as psychology [7; 8]. Ekman and Friesen [7] identified six facial expressions (i.e. anger, disgust, fear, joy, sadness, and surprise) as prototypical expressions that are universal among humans regardless of their age, race and gender. Early research on automatic FER focused on those six emotions [14; 15; 16]. In recent years, FER has shown its importance in human-computer interaction (HCI), such as assistive driving [19], embodied agents [22], and in applications such as diagnosis [27; 28] and computer games [29]. Thus, the demand has been increasing for an effective FER technology that can determine one's emotional state, based on face images, regardless of one's age, gender or race. Although much progress has been made on recognizing facial expressions, it is still a difficult task, due to the complexity and variability of facial expressions, and an effective facial-representation method is a vital step to improving the recognition rate in FER.

Facial feature representations proposed in the literature can now be divided into three categories: geometrical, appearance-based, and deep features. Geometrical features [32] take advantage of shape and location information of facial components and salient points, i.e. the eyes, lips, nose tip, etc. FER with Action Unit (AU) recognition is a geometrical feature-based approach, which has achieved more attention recently with the advancement in deep neural-network structures [35; 36]. However, geometrical features still require an accurate and reliable reconstruction and tracking of the facial landmarks. Therefore, it is difficult to achieve in real-life situations. Furthermore, AU-based facial expression recognition may require training data whose Action Units are already labelled by experts, which is a labor-intensive and time-consuming process. Recent studies have shown that appearance-based methods can achieve similar or better performance than AU recognition-based methods [10].

Appearance-based features are based on texture information related to the expressions on a face, e.g. wrinkles, skin changes, etc., which can be applied to the whole face or specific facial regions. Appearance-based features do not require the accurate reconstruction of all the facial landmarks, but only eye-pupil points, since the eyes are usually used to align the faces for further facial representation, i.e. feature extraction. Furthermore, appearance-based features only need the emotion labels of the samples for the training process. These advantages make appearance-based methods more favorable in comparison to geometrical features. One of the first attempts of FER, based on texture classification, is to use Local Binary Pattern (LBP), which was proposed by Ojala et al. [24]. LBP is one of the most widely used descriptors, due to its computational simplicity, discriminative power, and insensitivity to monotonic grayscale changes.

The successful application of LBP on the FER problems has inspired further studies for local descriptors. These studies focus on enhancing the coding techniques, e.g. different neighbourhood sizes, processing of input images, e.g. linear filtering, transformations, etc., to emphasize the expression-specific information. Numerous variants of LBP have been proposed for the problems, such as face recognition [43; 44], facial expression recognition [46], texture classification [47], spatiotemporal feature representation [50], and medical image analysis [52]. Some comprehensive studies of LBP variants can be found in [53; 54; 55].

Recently, local binary feature learning methods have been proposed for efficient and data-adaptive face representation, because LBP and other hand-crafted features require strong prior knowledge of the problem in order to engineer them by hand [56; 57; 58]. The objective behind the feature learning methods is to learn a feature mapping using raw pixels to project each local pixel difference (PDV) into a low-dimensional binary vector that can efficiently represent the face data. Therefore, a codebook constructed using the learned binary codes can be used to obtain a histogram feature for each image [58]. To the best of our knowledge, local binary feature learning methods have not been applied to the FER problem, but only to age estimation [62] and face recognition [63; 64; 65]. As LBP has been successfully applied to the tasks for facial image analysis, it is worthwhile evaluating the recently proposed local binary feature learning methods on FER.

Recently, deep neural networks have been studied widely for many pattern-recognition tasks, such as human pose estimation [66], face recognition [68], gender recognition [69], image recognition [71; 72], which require learning from a large amount of data. The increasing popularity and the success of deep features are also rooted in the FER problems [1; 3; 4; 12; 17; 20; 25]. Although the increase in recognition rate for FER is undeniable, the debate between hand-crafted features and deep features is still active. Benitez-Garcia et al. [23] proposed a local descriptor, i.e. a handcrafted feature, which can achieve a higher recognition rate than any deep neural-network structure until now. This suggests that the domain-specific knowledge and the handcrafted features are still effective and favourable for visual classification. In this paper, we present a comprehensive study of appearance-based facial features, i.e. handcrafted features, and then we compare their best results with those methods based on recently proposed local binary feature learning methods and deep features for FER.

The steps of a basic FER framework with the use of appearance-based features can be listed as follows: 1) detecting and aligning the face images, 2) dividing each face image into several overlapping or non-overlapping regions, 3) extracting local features from these regions based on the local descriptors, 4) concatenating the respective local features to form a single feature vector, followed by unsupervised or supervised dimensionality reduction, 5) training a classifier based on the feature vectors from training samples, and 6) predicting the class label of a new query based on the trained classifier. The classification results depend on almost every step listed above. However, most of the recent studies have focused only on developing more robust local features [23; 31; 73; 75]. A robust feature should be highly discriminative, easily computed, of low dimensionality, insensitive to noise, such as illumination changes, and have low intra-class variations.

It is difficult to balance these properties for a local descriptor. For example, LBP is computationally simple and discriminative, but sensitive to random noise. Similarly, although Gabor-based local descriptors have shown their achievements, especially in face recognition [38; 76; 77; 78], the features suffer from the expensive computational requirement and high dimensionality. Thus, developing a robust local descriptor is still an open issue for many fields of image representation and classification, such as texture representation [75; 79; 80; 81] and face representation [44; 75].

In the field of computer vision, popular local descriptors are often applied to different problems or applications. For instance, although LBP was originally devised for texture classification, it has been applied to face recognition [82], image retrieval [83], facial expression recognition [16], etc. However, it might not always be true for a new descriptor. Facial expression recognition is a problem different from face recognition or other types of recognition. “A good face-recognition local descriptor” should represent discriminative identity information about face images, while “a good facial-expression local descriptor” should discard the subject’s identity information and highlight the expression-specific information of a face. Therefore, it is important to be attentive to the nature of a problem in choosing an appropriate local descriptor.

The local descriptors, proposed in the literature, often benchmark their results against previously reported ones. However, the reliability of the benchmarking may not be high, due to the following reasons:

- A few benchmark databases were used, and the descriptors were evaluated with different databases.

- Each of the databases may have a different set of expression categories.
- Different image preprocessing techniques, e.g. face alignment, illumination, different normalization, etc., are used in experiments.
- The evaluation procedures/testing protocols, e.g. the choice of the classifier, the cross-validation scheme used, etc., are different.
- The overall experiments cannot be reproduced because not all the experimental setup is known.

In the literature, there have been several attempts to compare the performances of LBP-like descriptors using the same experimental settings. One of the most recent experimental studies on the LBP-like descriptors was conducted by Liu et al. [84], which evaluated thirty-two LBP variants for texture classification. However, there are still many other texture descriptors for facial expression recognition, which should be compared.

Kristensen et al. [55] presented an overview of "binary flavored features" for FER. Although a set of commonly used terms was defined so as to encourage consistency in terminology and to explain the current challenges, the depth of the survey in terms of performance comparison is limited. Another aim of this paper is to fill this gap by providing a comprehensive performance analysis on those recent local descriptors used for FER.

In this paper, we compare the performances of 27 local descriptors on four popular databases with the same experimental setup, including the use of two classifiers, different image resolutions, and different numbers of sub-regions. In addition to their accuracy, other important aspects, such as face resolutions for best performances, are also studied. Moreover, we compare the results achieved by handcrafted features, e.g. histogram-based local features, with the results obtained by the "Compact Binary Face Descriptor (CBFD) [57]" and the state-of-the-art deep features. We also evaluate the robustness of the respective local descriptors in the scenario of a cross-dataset facial expression recognition problem. In our evaluation, we found that the best overall performances are obtained by Local Phase Quantization (LPQ) and Local Gabor Binary Pattern Histogram Sequence (LGBPHS), with consistency across most of the databases used in our experiments.

The rest of the paper is organized as follows: Section 2 introduces a taxonomy for histogram-based local descriptors and highlights the representative examples of the specific steps. In Section 3, the experimental setup is first described, then comprehensive experimental results are presented. Section 4 concludes the paper.

2. Construction of the histogram-based local descriptors

Histogram-based local descriptors compute local statistical information at key points, and describe the features in a region using a histogram representation. Almost all the local statistical feature (LSF) methods, as described in [43], have two main parts: statistical histogram feature extraction and statistical feature combination. Unlike [43] which divides the statistical histogram feature extraction further into three steps, we divide it into five steps in this paper, in order to describe different local descriptors in more detail. In the rest of this section, each step is explained while the corresponding representative descriptors are highlighted with their strengths and weaknesses.

2.1. Local variation coding

Histogram-based local-feature descriptors represent the centre pixel of a local region as a decimal number, according to its values compared to its neighbouring pixels. Regardless of the input image, local variation coding is a general method used to encode the pattern features in a local patch. For each local patch, with a given neighbourhood, a typical local variation coding has five steps, including linear filtering, quantization, binarization, encoding and binary to decimal conversion. In the following sub-sections, these five steps will be explained in detail.

2.1.1. Linear filtering

The first step of local variation coding is to convolve a patch with a predefined set of linear filters. The most commonly used linear filters in histogram-based local descriptors are Kirsch [33; 34], Prewitt [5; 6; 33], Sobel [5; 6; 13; 33; 40], and Derivative-Gaussian [33].

From the computational point of view, Sobel operators are more efficient than the Kirsch operators, as less pixels and multiplications are involved. These linear filters operate on a local patch with a 3×3 mask, and custom linear filters, which consider higher-order derivatives, have also been proposed. For example, Local Arc Pattern (LAP) [21] and Local Monotonic Pattern (LMP) [45] encode the first and the second-order derivatives of a local patch in different orientations, using a set of custom filters. Although LAP and LMP can represent a bigger micro pattern with multiple radii, they use intensity values, as LBP, and are therefore sensitive to non-monotonic changes. Local Transitional Pattern (LTrP) [59] and Local Monotonic Pattern (LMP) [45] encode the transition of intensity change in different directions over a local patch. Local Derivative Pattern (LDP) [85] encodes the second and higher-order derivatives of a local patch. Although the higher-order derivatives can represent local variations with more details, the dimensionality of the resulting feature vector will become higher, as well as the computational cost.

2.1.2. Quantization

The second step of the local variation coding is the quantization of the linear-filter responses. The most common way of quantization used in the different descriptors is the unit step function. The local descriptors, such as LBP, Median Binary Pattern (MBP) [51], etc., quantize their filter responses using the unit-step function. However, this will generate inconsistent binary codes in uniform and near-uniform face regions, because the filter responses may vary slightly around the threshold value, usually zero. Local Ternary Pattern (LTeP) [51], Median Ternary Pattern (MTP) [51], Gradient Directional Pattern (GDP) [5; 6], and Gradient Local Ternary Pattern (GLTeP) [10; 13] add an extra level of thresholding, which facilitates the generation of more consistent codes for local patterns in smooth facial regions, as well as highly textured regions.

Quantization of the filter responses does not necessarily result in binary values. A common way of non-binary quantization is the *k*-bin method. Histogram of Oriented Gradients (HOG) [86] and Pyramids of Histogram of Oriented Gradients (PHOG) [70] are two examples, which quantize the gradient angles to *k* intervals, and then count the gradient magnitudes of those pixels whose gradient orientations are within a specific interval. Another method of non-binary quantization of the filter responses, such as the angle or phase information, is to use the quadrant information [37; 43; 76], i.e. the 2-D Cartesian coordinate system, where four quadrants are defined by the *x*- and *y*-axes.

2.1.3. Binarization

After quantization, the filter responses of some descriptors, such as LBP [24], GDP [5; 6], have already been in binary form, i.e. 0 and 1. However, the other descriptors need a binarization process. The filter responses can be binarized in two ways:

Binarization by splitting into different levels of binary codes: One example of this method is LTeP [51], which has three levels after thresholding. A common way of encoding these three-level responses is to split the responses into two binary codes: "1" and "0" form one binary code, while "0" and "-1" form the other one. Therefore, two histograms are formed, and this results in a higher dimensional feature vector.

Binarization by logical operators: This method can be utilized in two different circumstances: when the quantized values are in binary form [45; 59], or not in binary form [37]. The common logical operators are "AND" [45] and "XOR" [18; 59]. These two logical operators have their unique advantages in information encoding. "AND" encodes the alikeness/sameness of the values, while "XOR" encodes the opposition between the values.

2.1.4. Encoding

The bits in a binary codeword correspond to the binarized responses of the different abovementioned filters. A basic way of creating a codeword is to use all the resultant binary codes to form a string. In the case of 3×3 neighborhood, i.e. 8 neighbours, each code string will be 8-bit long, which forms a decimal value between 0 and 255. LBP, LTeP, MBP, MTP and GDP utilize this basic code. Local Directional Pattern (LDiP) [26] computes the eight directional edge responses, by using the Kirsch masks. However, as the response values are not equally important in all the directions, LDiP encodes the *k* most prominent directions, i.e. a customized codeword. LDiP can provide more stable codes, in the presence of gray-level distortion, such as noise and non-monotonic illumination changes. High-frequency regions in a face carry more information about texture information, such as the human eye regions. Therefore, to achieve a more competent face representation, textural regions with high contrast/frequency should influence the LDiP code more. However, LDiP considers both low and high-frequency regions equally. To incorporate this importance into the LDiP codes, an extension of LDiP, named Local Directional Pattern Variance (LDiPv) [30], was proposed, which introduces the variance of the codes as weights in constructing the histograms. However, both LDiP and LDiPv consider the filter responses in absolute value, which lose the important direction information, e.g. different transitions in a region. Furthermore, they are sensitive to rotation variations, because a fixed start position has to be defined for encoding a binary string, and they are profoundly dependent on the number of the most prominent directions considered. Local Directional Number Pattern (LDN) [33] also encodes the principal directions, i.e. the most positive and negative directions, so a more discriminative representation of directions can be achieved. Local Directional Texture Pattern (LDTP) [34] also encodes the principal directions, which discards the insignificant details that may vary on the samples belonging to the same class. However, different from the other descriptors, LDTP encodes both the principal directions and the intensity information (the intensity difference of the two principal directions). Therefore, LDTP is robust against both rotation and illumination changes.

Recently in the fields of texture classification, image retrieval, and facial feature representation, an extensive amount of customized coding schemes has been proposed [44; 75; 80; 81]. All these coding schemes aim at producing robust features, which are important for the image-classification problem.

2.1.5. Binary to decimal conversion

The last step of local variation coding is to convert a binary codeword into a decimal value, which represents the local pattern of the pixel under consideration. After computing the feature values for all the pixels in a patch, the statistics of these numbers, in the form of a histogram, can be used to represent the patch.

2.1.6. Local Binary Pattern and other local variation coding schemes

LBP, as a local variation coding method, has four steps as discussed previously: linear filtering, quantization with the unit step function, encoding the binary codeword, and binary to decimal conversion. LBP has also been extended to use different neighbourhood sizes, as well as uniform LBP codewords, i.e. those codewords have no more than two transitions from 1 to 0 or 0 to 1. A codeword is non-uniform if it has more than two transitions. This idea was inspired by the fact that the uniform codewords occur more frequently than those non-uniform codewords in images.

LBP encodes the relationship between the central pixel and its neighbours. Some local descriptors extract high-order local information. A high-order descriptor can capture more detailed discriminative information. Other local descriptors also encode different distinctive spatial relationships in a local region. More information about LBP variants can be found in [84].

2.2. Local feature representation

LBP and other histogram-based local descriptors encode the distribution of local variation codes within a region. A frequency-based or weighted-vote-based histogram constructed for a whole face image will lose the spatial information about the patterns encoded by a local descriptor. To represent the facial features more effectively, face images are divided into a number of overlapping or non-overlapping small sub-regions. Local features extracted from the sub-regions can achieve better recognition rates than those using holistic features, such as Eigenfaces and Fisherfaces [87].

Different regions in a face carry different amount of information about an expression. To eliminate the excessive and non-informative features for face or expression recognition, weighted histogram representation has been adopted. In this representation, weights are often set according to the discriminability of the regions [16], e.g. a small weight near the image’s borders, and a higher weight around the eye and mouth regions.

Another local-feature representation uses only those regions that carry salient information about facial expressions. Benitez-Garcia et al. [23] developed an algorithm to detect salient regions based on fiducial points for feature extraction. In [15], we observed that the features extracted from the eye and mouth regions can achieve higher recognition rates than the features extracted from the sub-regions divided from a whole face.

2.3. Inputs to local variation coding

Most of the early descriptors extract local features from intensity information, using a local variation coding method. However, the intensity information is sensitive to noise and illumination variations. Therefore, other types of input have been considered for local variation coding. Since gradients are more stable than intensity under the presence of illumination variations, several descriptors utilize gradient information to encode local variations. For example, GDP encodes gradient angles, while GLTeP encodes gradient magnitudes.

After the successful applications of LBP, several descriptors, which are based on Gabor filtering with a predefined number of scales and orientations, have been proposed. Examples of these descriptors include Local Gabor Binary Pattern Histogram Sequence (LGBPHS) [38], Local Gabor Directional

Pattern (LGDiP) [39], and Local Gabor Transitional Pattern (LGTrP) [42]. These descriptors often encode the magnitude information of the transform, i.e. the Gabor Magnitude Image, because the magnitude information is robust to misalignment. Gabor features are robust to image variations in terms of illumination and noise, but extracting the features is computationally expensive and the resulting feature vector has a high dimensionality.

Binary Pattern of Phase Congruency (BPPC) [2] applies wavelet transform to the logarithmic Gabor features, followed by computing the phase congruency (PC). PC is a dimensionless quantity, and can be considered as the gradient where high energy values of PC occur on edges, corners, etc. Monogenic signal analysis [88], which is a 2-D generalization of the 1-D analytic signal, is an alternative method to Gabor filtering. Monogenic signal analysis can estimate the multi-resolution amplitude, orientation, and phase components of a signal, which represent the signal energetic, structural, and geometric information, respectively. One advantage of monogenic signal analysis over Gabor transformations is that it has a lower time and space complexity.

In 2010, two local descriptors, which use monogenic signal analysis, were proposed for texture classification [89] and face recognition [90], where only the monogenic phase information and both the amplitude and orientation information, respectively, are encoded. Several other this kind of local descriptors exist in the literature [44; 91; 92]. Monogenic signal analysis has also been used for spatiotemporal facial expression recognition, with the local descriptor named “Spatiotemporal Local Monogenic Binary Patterns (STLMBP)” [93]. However, to the best of our knowledge, Monogenic Binary Coding (MBC) [43] is the only descriptor that applied monogenic signal analysis to static facial expression images [61]. MBC encodes the amplitude (MBC_A), phase (MBC_P), and orientation (MBC_O) information separately.

Local Phase Quantization (LPQ) [48] is a local descriptor, which extracts features from the discrete Fourier transform (DFT) over an image. LPQ is robust against blur and low resolution because it quantizes the phase information in local neighbourhoods. However, LPQ requires the point spread function (PSF) to be positive and valued in the low-frequency domain. Local Frequency Descriptor (LFD) [37], which also extracts information from DFT, encodes both the magnitude and phase information using LBP and Local XNOR Pattern (LXNORP). LFD does not require PSF to be positive, and can carry more information than LPQ, but the dimension of the feature vector is doubled.

Weber Local Descriptor (WLD) [69, 70] was inspired by the Weber's Law, which states that the significance of a change in the stimuli depends on the initial value of the stimuli. WLD, which computes the differential excitation and the orientation of an image, forms a joint histogram for the differential excitation and the orientation. WLD has been applied to several problems successfully, including facial expression recognition [74]. However, WLD discards the orientation information of the differential excitation and neighbouring pixel pairs.

Recently, Jang et al. [18] proposed an extension of WLD, named Improved Weber Binary Coding (IWBC), to solve the drawbacks of WLD. IWBC generates two images, which are called the Novel Weber Magnitude Image and the Novel Weber Orientation Image, which are then encoded using Local XOR Pattern and LBP, respectively. Although IWBC can represent a face more accurately than WLD by including the orientation information about the neighbouring pixels, it suffers from the problem of high dimensionality. To the best of our knowledge, IWBC has never been applied to the FER problem. Since IWBC has been shown to outperform WLD on the face recognition problem, so we include IWBC in our experiments to evaluate its performance as a local descriptor for FER.

3. Experiments

In this section, a number of histogram-based local descriptors are evaluated for facial-expression recognition, with the same experiment settings. We will first describe the experimental setup, including the benchmark databases, pre-processing, feature extraction, and classification schemes, and then analyse the experimental results.

3.1. Experimental setup

3.1.1. Databases and the corresponding numbers of expression classes

The performances of the local descriptors are compared on commonly-used, acted databases, as well as spontaneous databases. The facial-expression databases used in our experiments are BAUM-2 [94], CK+ [95], JAFFE [96], and TFEID [97].

The CK+ database, which is one of the acted facial-expression databases mostly used, contains a total of 593 posed sequences across 123 subjects. 327 of the sequences were labelled with one of the seven discrete expressions — anger, contempt, disgust, fear, happiness, sadness, and surprise. The last three frames of each sequence and their landmarks provided are used for experiments. JAFFE and TFEID are two acted face databases with six prototypical expressions and the neutral expression, which contain 213 images from 10 Japanese females and 268 images from 40 Taiwanese subjects, respectively. The BAUM-2 database consists of expression videos extracted from movies. The expressions in the videos are in the close-to-real-life conditions, i.e. with pose, age, and illumination variations. In our experiments, the image dataset, namely BAUM-2i, consisting of images with peak expressions extracted from the videos in BAUM-2 are considered. There are 1,057 face images from 250 subjects, which have seven discrete expressions and the neutral expression in BAUM-2i.

A list of the descriptors, and the corresponding feature dimensions, used in our experiments.

	Abbreviation	**Descriptor Name**	**Dimension**
1	BPPC [2]	Binary Pattern of Phase Congruency	1062
2	GDP [5; 6]	Gradient Directional Pattern	256
3	GDP2 [9]	Gradient Direction Pattern	8
4	GLTeP [10; 13]	Gradient Local Ternary Pattern	512
5	IWBC [18]	Improved Weber Binary Coding	2048
6	LAP [21]	Local Arc Pattern	272
7	LBP [24]	Local Binary Pattern	59
8	LDiP [26]	Local Directional Pattern	56
9	LDiPv [30]	Local Directional Pattern Variance	56
10	LDN [33]	Local Directional Number Pattern	56
11	LDTP [34]	Local Directional Texture Pattern	72
12	LFD [37]	Local Frequency Descriptor	512
13	LGBPHS [38]	Local Gabor Binary Pattern Histogram Sequence	256
14	LGDiP [39]	Local Gabor Directional Pattern	280 *
15	LGIP [40]	Local Gradient Increasing Pattern	37
16	LGP [41]	Local Gradient Pattern	7
17	LGTrP [42]	Local Gabor Transitional Pattern	256
18	LMP [45]	Local Monotonic Pattern	256
19	LPQ [48; 49]	Local Phase Quantization	256
20	LTeP [51]	Local Ternary Pattern	512
21	LTrP [59; 60]	Local Transitional Pattern	256
22	MBC [43; 61]	Monogenic Binary Coding	3072 *
23	MBP [51]	Median Binary Pattern	256
24	MRELBP [67]	Median Robust Extended Local Binary Pattern	800
25	MTP [51]	Median Ternary Pattern	512
26	PHOG [70]	Pyramid of Histogram of Oriented Gradients	168 *
27	WLD [73; 74]	Weber Local Descriptor	32 *

* the feature dimension used in our experiments

The abovementioned databases have their own characteristics in terms of where the expression images were taken, the expression classes, the race, age and gender of the participants, etc.

3.1.2. Descriptors

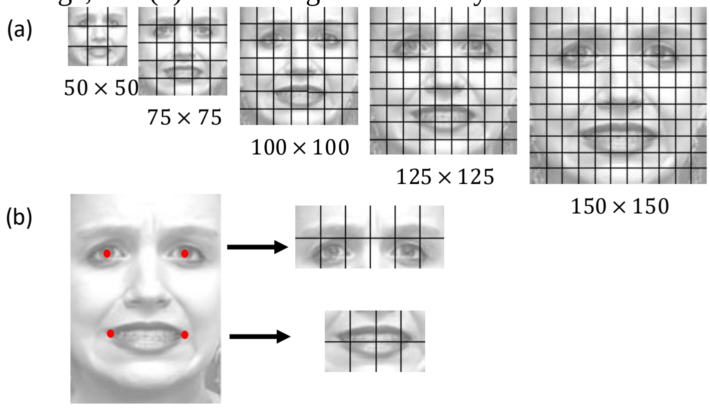
From the list of descriptors in [55], we select those descriptors based on spatial features, because this paper considers FER on static images only. The descriptors described in this paper and used in FER are also included in the experiments. Because of numerous LBP variants, only the basic LBP variants and MRELBP [67], which achieved the best performances in a recent comparative study for texture classifications [84], are chosen for our comparative analysis. The other local descriptors, which are not based on LBP, but inspired by LBP, for facial expression recognition are also included. Most of the descriptors presented and evaluated in this paper belong to the sixth category defined in [84], which is called "other methods inspired by LBP". A feature learning method, named "Compact Binary Face Descriptor (CBFD) [57]", is also used in our experiments to evaluate its performance on FER, in comparison to other state-of-the-art methods.

Table 1. A list of selected descriptors for our experiments and a comparison of the types of input data used and the local variation coding methods.

Descriptors	Input for local variation coding	Local variation coding
IWBC	Weber magnitude, Weber orientation	Local Xor Pattern (LXP) and Local Binary Pattern (LBP)
LAP	Intensity	first- and second-order derivatives using a set of custom filters
LBP	Intensity	-
LGBPHS	Gabor image	Local Binary Pattern (LBP)
LGIP	Intensity	Horizantal and vertical responses of sobel masks
LMP	Intensity	Local And Pattern with sign information of two level intensity differences
LPQ	Phase from Fourier Transform	Quantization
LTeP	Intensity	Two-level LBP
MBC_P	Phase, orientation or amplitude from Riesz transform	Local Nand (not and) Pattern
WLD	Differential excitations and orientations	Quantization

The descriptors (represented by their abbreviations), evaluated in our experiments, are listed in **Error! Reference source not found.**. To conduct a more detailed performance analysis, the best ten descriptors, along with the corresponding input information used and coding methods, are listed in Table 2. It is worth noting that MRELBP, IWBC, and CBFD are applied for the first time to the FER problem.

Figure 1. Examples of the sub-regions used in our experiments: (a) regular sub-regions in an image, and (b) the sub-regions for the eye window and mouth window.



3.1.3. Pre-processing and feature extraction

For the first set of experiments, face images from the different databases are scaled to different resolutions, including 50×50, 75×75, 100×100, 125×125, and 150×150. Then, features are extracted from the images with different numbers of sub-regions.

Table 2. The recognition rates for different resolutions, different numbers of sub-regions, on the CK+ database. "-" means that the corresponding results are unavailable because the dimensionality of the feature vectors are too high for experiments.

	Database	CK+ – LOSO – 6-class											
	Resolution	50x50	75x75		100x100		125x125			150x150			
	# of sub-regions	3x3	3x3	5x5	5x5	7x7	5x5	7x7	9x9	5x5	7x7	9x9	11x11
1	BPPC [2]	85.33	85.76	90.40	89.75	**90.83**	90.40	90.40	89.86	87.06	90.40	89.21	89.86
2	GDP [5; 6]	74.54	75.73	83.82	86.30	86.62	86.30	85.98	**86.84**	85.65	85.76	86.08	86.08
3	GDP2 [9; 10; 11]	57.71	57.39	83.06	81.98	90.51	82.20	89.97	92.45	83.06	89.43	94.28	**94.82**
4	GLTP [13]	82.85	85.98	91.15	92.66	92.66	92.34	91.69	91.05	92.13	**93.64**	91.59	92.99
5	IWBC [18]	88.67	90.72	91.69	90.51	93.42	91.15	93.10	**94.82**	90.83	92.13	93.31	92.99
6	LAP [21]	83.17	80.26	89.75	89.21	91.69	90.29	91.26	93.42	91.05	91.05	93.42	**94.17**
7	LBP [24]	84.68	84.03	92.23	91.48	91.69	91.91	93.53	93.85	91.80	93.20	93.74	**95.25**
8	LDiP [26]	68.72	71.52	86.73	85.44	89.00	86.08	89.54	89.54	84.68	89.64	89.32	**89.75**
9	LDiPv [30; 31]	68.93	71.20	83.17	82.85	86.95	83.50	87.70	89.00	85.11	85.98	88.67	**89.21**
10	LDN [33]	80.91	82.96	88.46	88.24	90.29	90.40	91.15	92.66	90.83	90.40	**92.66**	91.91
11	LDTP [34]	82.74	80.69	85.87	85.65	90.08	86.19	89.75	93.10	83.60	87.06	**93.53**	89.75
12	LFD [37]	86.62	82.09	**90.61**	88.78	90.51	87.38	89.43	89.21	86.62	88.57	88.78	87.49
13	LGBPHS [38]	86.19	87.27	92.02	92.88	92.99	91.26	90.72	91.48	90.29	89.75	91.48	**95.25**
14	LGDiP [39]	71.09	69.15	75.19	80.15	79.72	77.35	78.86	79.07	80.04	**83.39**	80.80	78.64
15	LGIP [40]	83.28	84.14	93.20	91.59	92.88	91.69	92.66	93.96	91.69	92.34	93.31	**95.15**
16	LGP [41]	50.70	51.13	79.50	77.13	87.38	76.27	85.33	92.45	76.27	86.41	93.10	**93.31**
17	LGTrP [42]	48.76	50.16	62.46	64.51	68.72	65.26	64.40	66.67	62.03	**69.26**	64.40	68.82
18	LMP [45]	86.30	87.38	90.83	92.23	92.34	92.99	95.04	**95.25**	91.59	94.50	93.85	93.96
19	LPQ [48; 49]	90.08	92.45	93.96	**94.39**	93.31	93.31	94.28	94.17	92.77	93.74	93.74	93.64
20	LTeP [51]	88.35	89.10	91.80	92.45	93.31	92.99	93.96	**95.69**	92.56	94.50	95.04	94.93
21	LTrP [59; 60]	74.76	75.73	85.65	85.44	88.13	84.36	87.70	88.24	87.38	**89.54**	89.43	87.27
22	MBC_A [43; 61]	**92.56**	89.54	89.97	90.51	88.35	89.43	89.43	-	90.08	89.32	-	-
23	MBC_P [43; 61]	88.89	89.32	92.88	**94.28**	90.51	91.80	93.42	-	92.56	92.45	-	-
24	MBC_O [43; 61]	88.89	87.81	92.02	91.80	91.37	90.94	**92.56**	-	**92.56**	92.02	-	-
25	MBP [51]	83.71	82.85	90.08	90.61	90.94	91.05	91.48	93.53	90.40	91.69	93.42	**94.07**
26	MRELBP [67]	87.70	88.13	92.13	90.29	92.45	90.72	92.02	**93.53**	91.05	92.88	92.88	93.31
27	MTP [51]	90.72	87.92	90.51	90.08	89.97	89.64	89.97	**92.77**	89.43	89.00	91.59	90.94
28	PHOG [70]	87.59	89.54	89.54	90.29	89.32	89.00	**91.80**	90.72	89.21	90.83	90.51	90.40
29	WLD [73; 74]	81.45	79.61	91.37	90.94	92.23	90.83	93.10	**95.90**	92.23	93.31	95.47	95.47

In the second set of experiments, face images from the different databases are all scaled to the size of 126×189 pixels, with a distance of 64 pixels between the two eyes. To locate the eye and mouth windows, the facial landmarks, i.e. the eye and mouth corners, are used. If facial landmarks are not provided for a database, the required facial-feature points are marked manually. The eye window and the mouth window are further divided into 12 and 8 sub-regions, respectively. Figure 1 shows examples of selected sub-regions in both the first and the second set of experiments.

Table 3. The recognition rates for different resolutions and different numbers of sub-regions, on the BAUM-2i database. "-" means that the corresponding results are unavailable because the dimensionality of the feature vectors are too high for experiments.

	Database	BAUM-2i – 10-fold – 6-class											
	Resolution	50x50	75x75		100x100		125x125			150x150			
	# of sub-regions	3x3	3x3	5x5	5x5	7x7	5x5	7x7	9x9	5x5	7x7	9x9	11x11
1	BPPC [2]	51.18	52.48	53.90	56.26	58.98	54.37	**59.10**	57.45	56.15	55.67	57.21	54.85
2	GDP [5; 6]	40.78	45.39	50.71	46.10	**52.84**	45.98	50.83	51.89	46.22	50.24	51.89	49.76
3	GDP2 [9; 10; 11]	23.88	26.12	29.91	27.90	37.35	28.01	40.54	48.94	28.84	40.66	48.11	**53.31**
4	GLTP [13]	50.00	54.14	57.57	55.44	59.57	53.90	59.10	58.27	54.73	59.22	57.57	**60.05**
5	IWBC [18]	55.67	57.33	**59.22**	58.39	58.16	56.97	57.92	57.57	57.09	57.21	56.86	56.26
6	LAP [21]	46.22	44.80	53.66	48.70	51.77	48.35	50.24	56.03	49.05	50.24	56.03	**56.38**
7	LBP [24]	48.35	48.23	54.26	54.02	56.62	53.07	56.86	58.63	53.31	54.61	56.62	**59.46**
8	LDiP [26]	27.30	30.61	43.62	45.39	53.07	48.70	52.25	53.66	45.15	52.60	**56.03**	**56.03**
9	LDiPv [30; 31]	24.11	28.25	40.19	36.05	49.88	40.54	48.94	52.25	40.78	48.82	52.84	**53.90**
10	LDN [33]	37.23	34.16	48.11	47.52	51.06	47.28	54.14	55.67	47.16	54.61	55.91	**60.99**
11	LDTP [34]	30.26	34.04	48.11	43.38	48.82	42.79	46.81	49.41	44.33	47.64	46.81	**51.06**
12	LFD [37]	46.69	44.56	53.90	50.59	57.21	51.77	56.74	57.57	51.06	54.73	56.50	**57.92**
13	LGBPHS [38]	49.41	50.00	56.62	57.57	59.46	59.22	60.28	60.76	59.81	61.11	**62.41**	57.92
14	LGDiP [39]	30.97	32.62	39.60	42.67	44.09	41.25	46.10	**47.52**	39.83	42.55	44.44	43.85
15	LGIP [40]	30.50	30.97	49.88	47.04	54.61	49.17	54.02	56.74	48.11	52.96	55.44	**58.39**
16	LGP [41]	23.40	21.75	23.52	26.36	31.32	25.41	32.98	42.43	25.30	30.02	41.84	**46.22**
17	LGTrP [42]	24.47	23.05	31.44	30.38	31.09	32.03	34.04	35.11	31.68	36.29	**40.07**	36.29
18	LMP [45]	49.76	50.35	55.91	56.86	58.75	56.50	58.16	**60.64**	52.36	55.32	58.27	60.17
19	LPQ [48; 49]	56.38	56.03	61.35	61.35	60.28	59.46	**61.47**	61.23	57.68	59.57	60.28	**61.47**
20	LTeP [51]	52.36	50.59	55.32	52.96	58.87	52.96	59.46	59.22	54.02	59.57	**60.28**	**60.28**
21	LTrP [59; 60]	35.34	38.89	42.79	46.22	51.06	45.15	49.17	52.01	46.57	50.47	51.77	**53.78**
22	MBC_A [43; 61]	56.62	56.62	**59.81**	57.57	59.34	56.38	58.63	55.08	56.03	55.67	55.20	-
23	MBC_P [43; 61]	56.03	54.96	59.93	59.81	61.58	59.57	61.47	61.94	59.46	60.87	**62.06**	-
24	MBC_O [43; 61]	57.68	55.79	61.35	**61.82**	60.99	60.05	60.17	61.23	58.27	60.99	60.40	-
25	MBP [3; 4; 51]	43.62	47.04	54.37	54.37	54.49	53.55	55.32	**59.46**	52.60	53.43	55.08	57.33
26	MRELBP [67]	46.34	48.70	55.56	56.86	57.68	57.80	57.92	**59.34**	57.45	57.57	58.98	**59.34**
27	MTP [51]	43.97	41.96	51.54	50.24	**54.02**	47.87	53.78	51.65	42.91	51.65	52.13	52.60
28	PHOG [70]	47.52	50.35	51.42	53.43	53.90	51.77	54.26	52.96	54.14	54.02	**54.61**	53.43
29	WLD [73; 74]	30.26	24.82	51.77	46.10	57.80	47.52	55.44	56.15	46.93	54.73	55.79	**58.75**

3.1.4. Dimensionality reduction and classification

In the first two sets of experiments, the local descriptors listed in **Error! Reference source not found.** were first extracted. Then, the subspace-learning method, Soft Locality Preserving Projection (SLPM) [98], is applied for manifold learning and dimensionality reduction. SLPM is a graph-based subspace-learning method, which uses the *k*-neighborhood information and the class information. The key feature of SLPM is that it aims to control the level of spread of the different classes, because the spread of the classes in the underlying manifold is closely connected to the generalizability of the learned subspace. In our experiments, we employ SLPM for dimensionality reduction and for increasing the discriminative ability of the extracted features. Finally, the nearest neighbour (NN) classifier is used for classification. The third set of experiments were conducted, with the best setting for each of the databases, using the Support Vector Machine (SVM) classifier, with the linear kernel. The results are then compared to those based on the nearest neighbour classifier. Two different cross-validation schemes are adopted in our experiments: Leave-One-Subject-Out (LOSO) to encourage the reproducibility of the experiments, and 10-fold cross-validation, which is used when there are sufficient number of images for each subject in the database, i.e. BAUM-2i. Furthermore, both the 10-fold and LOSO cross-validation schemes are used for comparison on the JAFFE and TFEID databases.

3.2. Experimental results

Table 4. The comparison of recognition rates obtained by the selected local descriptors on the BAUM-2i database (the best of sub-regions) using 10-fold cross validation. 6-class: AN, DI, FE, HA, SA, and SU. 7-class: AN, CO, DI, FE, HA, SA, and SU. 8-class: AN, CO, DI, FE, HA, NE, SA, and SU.

	BAUM-2i		
	6-class	7-class	8-class
IWBC	59.22	55.53	52.53
LAP	56.38	54.97	49.00
LBP	59.46	58.32	52.44
LGBPHS	**62.41**	57.99	54.15
LGIP	58.39	54.75	49.86
LMP	60.64	57.54	52.53
LPQ	61.47	**58.99**	**54.73**
LTeP	60.28	57.21	52.63
MBC_P	62.06	58.10	54.25
WLD	58.75	54.41	50.53

In this section, the experiment results on the four facial-expression databases (BAUM-2i, CK+, JAFFE, TFEID) under different experimental settings are presented and discussed. The experiments are designed to measure the performances of the respective descriptors, for face images at different resolutions and divided into different sub-regions, and with different classifiers.

3.2.1. Performance analysis for varying resolution and number of sub-regions

All the face images are first aligned based on the positions of the two eye pupils, and cropped to the different resolutions. For each resolution, face images are divided into different numbers of sub-regions, say $l \times l$, where l varies from 3 to 11.

Table 3 and Table 4 present the results on CK+ and BAUM-2i for all the descriptors. As observed from the results shown in Tables 3 and 4, in general, the classification performances improve when the image resolution and the number of sub-regions increase. Therefore, higher resolution and more sub-regions lead to better classification performances. However, with more sub-regions, the feature dimension will become very high. In other words, the better performance is at the expenses of higher computational requirements.

For more detailed performance analysis, the best ten descriptors, which have achieved promising results, were chosen to repeat the first set of experiments on the four databases separately with different numbers of expression classes, as well as the two different classification schemes. Table 5 shows the best classification rates on BAUM-2i with different numbers of expression classes. In Tables 6 to 8, the columns named “best of sub-regions” show the best classification rates for the number of sub-regions being used. We only show the best results, otherwise there are too many data to be shown.

3.2.2. Performance analysis of the eye and mouth regions

The second set of experiments was conducted with the features extracted from the eye and mouth windows of face images. The CK+, JAFFE and TFEID databases are used to test the performances of the respective features extracted from the eye and the mouth regions. The BAUM-2i database is not used because it consists of images in the wild. Labelling the facial landmarks is a complicated task. In Tables 6 to 8, the two columns under “eye and mouth windows” show the classification accuracies of the selected features, using the LOSO and 10-fold cross-validation schemes.

As observed from the tables, using the features extracted from the eye and the mouth windows achieves lower classification accuracies than that using features extracted from the sub-regions of whole face images. However, for the results based on sub-regions, we show the best classification accuracies achieved for the five different resolutions and the five different numbers of sub-regions. Furthermore, each descriptor achieves the best performance on a different resolution and a different number of sub-regions. Experiment results show that there are not a particular resolution and a particular number of

Table 5. The recognition rates of selected local descriptors on the CK+ database, with 6 classes (AN, DI, FE, HA, SA, and SU) and 7 classes (AN, CO, DI, FE, HA, SA, and SU), using LOSO.

	CK+			
	Eye and mouth windows		Best of sub-regions	
	6-class	7-class	6-class	7-class
IWBC	94.61	93.68	94.82	94.50
LAP	91.37	91.44	94.17	92.86
LBP	93.31	92.56	95.25	93.99
LGBPHS	92.23	90.72	95.25	93.99
LGIP	91.26	92.35	95.15	94.50
LMP	**94.71**	94.19	95.25	**94.90**
LPQ	94.61	**94.90**	94.39	94.19
LTeP	93.53	93.17	95.69	94.80
MBC_P	91.69	89.40	94.28	92.46
WLD	93.31	91.44	**95.90**	94.80

Table 6. The recognition rates of the selected best local descriptors on the JAFFE database.

	JAFFE			
	Eye and mouth windows		Best of sub-regions	
	LOSO	10-fold	LOSO	10-fold
IWBC	58.69	88.73	65.73	90.61
LAP	**68.08**	90.61	68.54	94.84
LBP	61.50	86.38	65.73	93.43
LGBPHS	63.38	**93.90**	**71.83**	93.90
LGIP	62.91	87.32	66.20	93.90
LMP	60.09	85.92	67.14	93.43
LPQ	67.61	92.02	69.95	93.43
LTeP	61.03	89.20	62.44	94.37
MBC_P	63.38	92.96	66.67	93.90
WLD	63.38	86.85	69.01	**96.24**
CBFD	66.20	89.67	-	-

Table 7. The comparison of recognition rates obtained by the selected local descriptors on the TFEID database.

	TFEID			
	Eye and mouth windows		Best of sub-regions	
	LOSO	10-fold	LOSO	10-fold
IWBC	89.55	90.67	92.91	91.79
LAP	91.04	91.04	94.40	**95.15**
LBP	91.79	92.54	93.66	94.78
LGBPHS	**94.40**	91.04	**95.15**	93.66
LGIP	89.18	86.19	94.78	93.28
LMP	91.42	92.16	94.03	94.03
LPQ	**94.40**	**93.28**	94.03	94.40
LTeP	90.30	92.16	94.40	**95.15**
MBC_P	94.30	91.79	94.40	93.66
WLD	92.16	91.42	94.78	94.40
CBFD	93.66	92.16	-	-

sub-regions that can work the best for all the descriptors.

3.2.3. Performance analysis of the classifiers

Table 9 presents the experiment results obtained with the NN and the SVM classifiers. We can observe that both LGBPHS and LPQ achieve similar performances in the use of NN and SVM. However, the

NN classifier can achieve equal or higher performance than the SVM classifier if a supervised dimensionality reduction method is employed. In our experiments, we utilize SLPM for dimensionality reduction.

Table 8. The comparison of the recognition rates obtained with features extracted from the eye and mouth regions by the nearest neighbor classifier (NN) and SVM classifier using LOSO.

	CK+		JAFFE		TFEID	
	SLPM + NN	SVM	SLPM + NN	SVM	SLPM + NN	SVM
LGBPHS	92.23	91.91	63.38	61.50	94.40	94.40
LPQ	94.61	**94.93**	**67.61**	67.14	94.40	94.40

3.2.4. Performance analysis of cross-dataset facial expression recognition

In real-life applications, query samples are often different from the training samples in terms of uncontrolled variations such as illumination. Therefore, it is important for a local descriptor to have a good generalization power, and the descriptor can still achieve a good performance when the training and test sets are from different databases. In this paper, we also conduct experiments to test the robustness and accuracy of the best selected descriptors in the scenario of cross-dataset FER.

Table 10 shows the experiment results when the training and the testing sets are two different datasets, which have different acquisition conditions. As you can observe in Table 10, the recognition rates for the 6 basic emotions decrease significantly, because cross-dataset FER is a challenging task. Although no local descriptor can perform consistently better than the others, MBC_P achieves the highest recognition rates when the model is trained using JAFFE while tested on TFEID, and vice versa. MBC_P uses monogenic signal analysis to estimate the phase component of the images, which represents the images' geometric information. Since the JAFFE database consists of images of Japanese women, while TFEID consists of images of Taiwanese men and women, we can observe that the phase information of the monogenic signal analysis is insensitive to cross-cultural face representation for FER.

Table 9. The comparison of the recognition rates of the ten selected descriptors on cross-dataset facial expression recognition, with features extracted from the eye and mouth windows.

Trained on	CK+		JAFFE		TFEID	
Tested on	JAFFE	TFEID	CK+	TFEID	CK+	JAFFE
IWBC	25.00	33.77	34.52	42.98	38.30	30.98
LAP	29.89	32.46	24.16	42.98	**45.85**	26.63
LBP	21.74	34.65	29.02	44.74	35.81	28.26
LGBPHS	18.48	33.33	37.22	60.09	39.48	44.02
LGIP	**30.43**	31.14	31.18	41.23	42.61	31.52
LMP	29.35	32.46	**37.32**	48.25	37.32	23.37
LPQ	19.57	**38.16**	32.58	50.44	42.07	35.33
LTeP	19.57	35.96	25.03	25.88	26.86	35.87
MBC_P	25.54	31.14	37.00	**63.60**	38.83	**47.28**
WLD	15.76	35.09	27.18	32.89	42.61	24.46

3.2.5. Comparison with deep features

Recently, convolutional deep neural networks have been applied to FER [1; 3; 4; 12; 17; 20; 25]. Table 11 presents the performances of deep learning methods applied on the CK+ database. 3DCNN-DAP [1] adapts a deformable parts learning component to detect discriminative facial action parts for spatiotemporal FER, where a Boosted Deep Belief Network [3] (BDBN) is used to learn and select the expression-related facial features to develop a strong classifier in a unified loopy framework iteratively. Iterative learning of the BDBN framework strengthens the discriminative capabilities of the features. STM-ExpLet [4] learns a spatiotemporal manifold (STM) from low-level features from each expression video clip, followed by learning a universal manifold model that statistically unify all the STMs. With this method, expression videos are also aligned. Different from these methods, DTAGN [12] trains two models, with temporal geometry features and temporal appearance features, respectively, from image

Table 10. The comparison of recognition rates of deep learning methods and the best recognition rate obtained with handcrafted features

Method	Feature Type	Accuracy (%)
3DCNN-DAP [1]	Deep features	92.4
BDBN [3]	Deep features	96.7
STM-ExpLet [4]	Deep features	94.2
DTAGN [12]	Deep features	97.3
Inception [17]	Deep features	93.2
PPDN [20]	Deep features	97.3
LFC + FFD [23]	Handcrafted features	97.9
FN2EN [25]	Deep features	**98.6**
LPQ-SLPM-NN	Handcrafted features	**95.9**

sequences, and these two features are complementary to each other. In [17], a network, which consists of two convolutional layers with max pooling and four inception layers, was proposed. The network was evaluated for its generalizability by experiments, with cross-database classification. To boost the generalizability of learning, [20] presented a peak-piloted deep network (PPDN), which uses the samples with high-intensity expressions to supervise the samples with low-intensity expression that are hard to classify. Until now, FN2EN [25], which uses a two-stage training algorithm, achieved the best performance on the CK+ dataset. FN2EN, in the first stage, trains the convolutional layers, whose outputs from the last pooling layer are used to supervise the expression net in the second stage.

As observed in Table 11, LPQ with NN outperforms several deep learning methods. LFC + FFD [23] is also a histogram-based feature extraction method, which achieves higher classification accuracy than all the listed methods, except FN2EN [25]. To the best of our knowledge, FN2EN achieves the highest classification accuracy on the CK+ database. However, expensive computational cost is a drawback of most of the methods based on deep convolutional neural networks. Furthermore, the CK+ database consists of images taken under controlled environments, i.e. posed expressions and the number of expression samples are limited in the CK+ database. These factors direct us the need of a large-scale facial-expression database in the wild. There have been several attempts to collect facial-expression images in the wild [99; 100; 101]. [102] and [103] are two recently published databases, which contain large-scale face images with varying expressions. These databases will be very useful for FER based on deep learning.

4. Conclusion

This paper provides a systematic review and analysis of current histogram-based local feature descriptors, which have been applied for facial-expression recognition. The weaknesses and strengths of the existing descriptors, as well as their underlying connections, have also been discussed and analysed. Then, a comprehensive evaluation of the performances of different descriptors for facial-expression recognition is conducted and presented. In total, 27 local descriptors have been applied on four facial-expression databases, under the same experimental settings. The robustness of the respective local descriptors is tested under different conditions, such as varying image resolutions and number of sub-regions, and the classifiers. Moreover, a brief performance comparison with seven recent deep features and two handcrafted features has been conducted.

Several remarks from the experiment results are listed as follows:

- The databases have different characteristics, which affect the choice of the ideal descriptor for a particular database. Even the number of expression classes can also affect the performances of the descriptors.
- The results show a trade-off between the number of sub-regions and the overall classification accuracy. The use of the eye and the mouth windows decreases the number of sub-regions and the dimensionality of the resulting feature vectors, with a slight loss in terms of accuracy.

- The resolution of face images and the number of sub-regions are the two most important factors that affect the overall classification accuracies.
- The highest classification accuracies are obtained mostly by LGBPHS and LPQ. This shows that Gabor wavelets and phase information are important features for representing expression-specific information. However, we should keep in mind that Gabor features suffer from high computational cost.
- According to the comprehensive analysis shown in this paper, the best local descriptors for FER, by considering the feature length, computational cost, and the classification accuracy simultaneously, is LPQ.
- Deep neural-network-based methods indeed can achieve excellent classification accuracies on FER. However, these methods also suffer from time and space complexities as LGBPHS.

In conclusion, our comprehensive experiment results show that the trade-off between the computational cost and the classification accuracy still exists today.

5. Acknowledgements

The work described in this paper was supported by a research grant from The Hong Kong Polytechnic University (project code: G-YBKF).

References

[1] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, Deeply learning deformable facial action parts model for dynamic expression analysis, Asian Conference on Computer Vision, Springer, 2014, pp. 143-157.

[2] S. Shojaeilangari, W.-Y. Yau, J. Li, and E.-K. Teoh, Feature extraction through binary pattern of phase congruency for facial expression recognition, Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on, IEEE, 2012, pp. 166-170.

[3] P. Liu, S. Han, Z. Meng, and Y. Tong, Facial expression recognition via a boosted deep belief network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805-1812.

[4] M. Liu, S. Shan, R. Wang, and X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1749-1756.

[5] F. Ahmed, Gradient directional pattern: a robust feature descriptor for facial expression recognition. Electronics letters 48 (2012) 1203-1204.

[6] W. Chu, Facial expression recognition based on local binary pattern and gradient directional pattern, Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, IEEE, 2013, pp. 1458-1462.

[7] P. Ekman, and W.V. Friesen, Constants across cultures in the face and emotion. Journal of personality and social psychology 17 (1971) 124.

[8] P. Ekman, Darwin, deception, and facial expression. Annals of the New York Academy of Sciences 1000 (2003) 205-221.
[9] M.S. Islam, Gender Classification using Gradient Direction Pattern. Science International 25 (2013).
[10] M. Valstar, and M. Pantic, Fully automatic facial action unit detection and temporal analysis, Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, IEEE, 2006, pp. 149-149.
[11] R. Walecki, V. Pavlovic, B. Schuller, and M. Pantic, Deep Structured Learning for Facial Action Unit Intensity Estimation. arXiv preprint arXiv:1704.04481 (2017).
[12] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, Deep temporal appearance-geometry network for facial expression recognition. arXiv preprint arXiv:1503.01532 (2015).
[13] F. Ahmed, and E. Hossain, Automated facial expression recognition using gradient-based ternary texture patterns. Chinese Journal of Engineering 2013 (2013).
[14] C. Turan, K.-M. Lam, and X. He, Facial expression recognition with emotion-based feature fusion, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific, IEEE, 2015, pp. 1-6.
[15] C. Turan, and K.-M. Lam, Region-based feature fusion for facial-expression recognition, Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 5966-5970.
[16] C. Shan, S. Gong, and P.W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27 (2009) 803-816.
[17] A. Mollahosseini, D. Chan, and M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1-10.
[18] B.-Q. Yang, T. Zhang, C.-C. Gu, K.-J. Wu, and X.-P. Guan, A novel face recognition method based on iwld and iwbc. Multimedia Tools and Applications 75 (2016) 6979.
[19] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, Drowsy driver detection through facial movement analysis. Human–computer interaction (2007) 6-18.
[20] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, Peak-piloted deep network for facial expression recognition, European Conference on Computer Vision, Springer, 2016, pp. 425-442.
[21] M.S. Islam, and S. Auwatanamongkol, Facial Expression Recognition using Local Arc Pattern. Trends in Applied Sciences Research 9 (2014) 113.
[22] T. Fong, I. Nourbakhsh, and K. Dautenhahn, A survey of socially interactive robots. Robotics and autonomous systems 42 (2003) 143-166.
[23] G. Benitez-Garcia, T. Nakamura, and M. Kaneko, Facial Expression Recognition Based on Local Fourier Coefficients and Facial Fourier Descriptors. Journal of Signal and Information Processing 8 (2017) 132.
[24] T. Ojala, M. Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence 24 (2002) 971-987.
[25] H. Ding, S.K. Zhou, and R. Chellappa, Facenet2expnet: Regularizing a deep face recognition net for expression recognition, Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 118-126.
[26] T. Jabid, M.H. Kabir, and O. Chae, Local directional pattern (LDP)–A robust image descriptor for object recognition, Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, IEEE, 2010, pp. 482-487.
[27] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, and F. De la Torre, Detecting depression from facial actions and vocal prosody, Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, IEEE, 2009, pp. 1-7.
[28] S. Jaiswal, M.F. Valstar, A. Gillott, and D. Daley, Automatic detection of ADHD and ASD from expressive behaviour in RGBD data, Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 762-769.
[29] S.J. Kirsh, and J.R. Mounts, Violent video game play impacts facial emotion recognition. Aggressive behavior 33 (2007) 353-358.

[30] M.H. Kabir, T. Jabid, and O. Chae, A local directional pattern variance (LDPv) based face descriptor for human facial expression recognition, Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, IEEE, 2010, pp. 526-532.

[31] K.-C. Fan, and T.-Y. Hung, A novel local pattern descriptor—local vector pattern in high-order derivative space for face recognition. IEEE transactions on image processing 23 (2014) 2877-2891.

[32] B. Martinez, M.F. Valstar, B. Jiang, and M. Pantic, Automatic analysis of facial actions: A survey. IEEE Transactions on Affective Computing (2017).

[33] A.R. Rivera, J.R. Castillo, and O.O. Chae, Local directional number pattern for face analysis: Face and expression recognition. IEEE transactions on image processing 22 (2013) 1740-1752.

[34] A.R. Rivera, J.R. Castillo, and O. Chae, Local directional texture pattern image descriptor. Pattern Recognition Letters 51 (2015) 94-100.

[35] S. Jaiswal, and M. Valstar, Deep learning the dynamic appearance and shape of facial action units, Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, 2016, pp. 1-8.

[36] Z. Tősér, L.A. Jeni, A. Lőrincz, and J.F. Cohn, Deep learning for facial action unit detection under large head poses, Computer Vision–ECCV 2016 Workshops, Springer, 2016, pp. 359-371.

[37] Z. Lei, T. Ahonen, M. Pietikäinen, and S.Z. Li, Local frequency descriptor for low-resolution face recognition, Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 161-166.

[38] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition, Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE, 2005, pp. 786-791.

[39] S.Z. Ishraque, A.H. Banna, and O. Chae, Local Gabor directional pattern for facial expression recognition, Computer and Information Technology (ICCIT), 2012 15th International Conference on, IEEE, 2012, pp. 164-167.

[40] Z. Lubing, and W. Han, Local gradient increasing pattern for facial expression recognition, Image Processing (ICIP), 2012 19th IEEE International Conference on, IEEE, 2012, pp. 2601-2604.

[41] M.S. Islam, Local gradient pattern-A novel feature representation for facial expression recognition. Journal of AI and Data Mining 2 (2014) 33-38.

[42] T. Ahsan, T. Jabid, and U.-P. Chong, Facial expression recognition using local transitional pattern on Gabor filtered facial images. IETE Technical Review 30 (2013) 47-52.

[43] M. Yang, L. Zhang, S.C.-K. Shiu, and D. Zhang, Monogenic binary coding: An efficient local feature extraction approach to face recognition. IEEE Transactions on Information Forensics and Security 7 (2012) 1738-1751.

[44] J. Li, N. Sang, and C. Gao, Face recognition with Riesz binary pattern. Digital Signal Processing 51 (2016) 196-201.

[45] T. Mohammad, and M.L. Ali, Robust facial expression recognition based on local monotonic pattern (LMP), Computer and Information Technology (ICCIT), 2011 14th International Conference on, IEEE, 2011, pp. 572-576.

[46] R.A. Khan, A. Meyer, H. Konik, and S. Bouakaz, Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognition Letters 34 (2013) 1159-1168.

[47] T. Song, H. Li, F. Meng, Q. Wu, and J. Cai, LETRIST: Locally Encoded Transform Feature Histogram for Rotation-Invariant Texture Classification. IEEE Transactions on Circuits and Systems for Video Technology (2017).

[48] V. Ojansivu, and J. Heikkilä, Blur insensitive texture classification using local phase quantization, International conference on image and signal processing, Springer, 2008, pp. 236-243.

[49] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, Emotion recognition using PHOG and LPQ features, Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 878-883.

[50] G. Zhao, and M. Pietikäinen, Dynamic texture recognition using volume local binary patterns, Dynamical Vision, Springer, 2007, pp. 165-177.

[51] F. Bashar, A. Khan, F. Ahmed, and M.H. Kabir, Robust facial expression recognition based on median ternary pattern (MTP), Electrical Information and Communication Technology (EICT), 2013 International Conference on, IEEE, 2014, pp. 1-5.
[52] L. Nanni, A. Lumini, and S. Brahnam, Local binary patterns variants as texture descriptors for medical image analysis. Artificial intelligence in medicine 49 (2010) 117-125.
[53] N.P. Doshi, and G. Schaefer, A comprehensive benchmark of local binary pattern algorithms for texture retrieval, Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, 2012, pp. 2760-2763.
[54] L. Nanni, A. Lumini, and S. Brahnam, Survey on LBP based texture descriptors for image classification. Expert Systems with Applications 39 (2012) 3634-3641.
[55] R.L. Kristensen, Z.-H. Tan, Z. Ma, and J. Guo, Binary pattern flavored feature extractors for Facial Expression Recognition: An overview, Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, IEEE, 2015, pp. 1131-1137.
[56] V. Balntas, L. Tang, and K. Mikolajczyk, Binary online learned descriptors. IEEE transactions on pattern analysis and machine intelligence 40 (2018) 555-567.
[57] J. Lu, V.E. Liong, X. Zhou, and J. Zhou, Learning compact binary face descriptor for face recognition. IEEE transactions on pattern analysis and machine intelligence 37 (2015) 2041-2056.
[58] Y. Duan, J. Lu, J. Feng, and J. Zhou, Learning rotation-invariant local binary descriptor. IEEE Transactions on Image Processing 26 (2017) 3636-3651.
[59] T. Jabid, and O. Chae, Local Transitional Pattern: A Robust Facial Image Descriptor for Automatic Facial Expression Recognition, Proc. International Conference on Computer Convergence Technology, Seoul, Korea, 2011, pp. 333-44.
[60] T. Jabid, and O. Chae, Facial Expression Recognition Based on Local Transitional Pattern. International Information Institute (Tokyo). Information 15 (2012) 2007.
[61] X.X. Xia, Z.L. Ying, and W.J. Chu, Facial Expression Recognition Based on Monogenic Binary Coding, Applied Mechanics and Materials, Trans Tech Publ, 2014, pp. 437-440.
[62] J. Lu, V.E. Liong, and J. Zhou, Cost-sensitive local binary feature learning for facial age estimation. IEEE Transactions on Image Processing 24 (2015) 5356-5368.
[63] Y. Duan, J. Lu, J. Feng, and J. Zhou, Context-aware local binary feature learning for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
[64] J. Lu, V. Erin Liong, and J. Zhou, Simultaneous local binary feature learning and encoding for face recognition, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3721-3729.
[65] J. Lu, V.E. Liong, and J. Zhou, Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. IEEE transactions on pattern analysis and machine intelligence (2017).
[66] A. Toshev, and C. Szegedy, Deeppose: Human pose estimation via deep neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653-1660.
[67] L. Liu, S. Lao, P.W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, Median robust extended local binary pattern for texture classification. IEEE Transactions on Image Processing 25 (2016) 1368-1381.
[68] F. Schroff, D. Kalenichenko, and J. Philbin, Facenet: A unified embedding for face recognition and clustering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823.
[69] J. Mansanet, A. Albiol, and R. Paredes, Local deep neural networks for gender recognition. Pattern Recognition Letters 70 (2016) 80-86.
[70] A. Bosch, A. Zisserman, and X. Munoz, Representing shape with a spatial pyramid kernel, Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, 2007, pp. 401-408.
[71] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[72] K. Simonyan, and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
[73] S. Li, D. Gong, and Y. Yuan, Face recognition using Weber local descriptors. Neurocomputing 122 (2013) 272-283.
[74] S. Liu, Y. Zhang, and K. Liu, Facial expression recognition under partial occlusion based on Weber Local Descriptor histogram and decision fusion, Control Conference (CCC), 2014 33rd Chinese, IEEE, 2014, pp. 4664-4668.
[75] S. Chakraborty, S. Singh, and P. Chakraborty, Local Gradient Hexa Pattern: A Descriptor for Face Recognition and Retrieval. IEEE Transactions on Circuits and Systems for Video Technology (2016).
[76] B. Zhang, S. Shan, X. Chen, and W. Gao, Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. IEEE Transactions on Image Processing 16 (2007) 57-68.
[77] W.-P. Choi, S.-H. Tse, K.-W. Wong, and K.-M. Lam, Simplified Gabor wavelets for human face recognition. Pattern Recognition 41 (2008) 1186-1199.
[78] K.-H. Pong, and K.-M. Lam, Multi-resolution feature fusion for face recognition. Pattern Recognition 47 (2014) 556-567.
[79] S. Du, Y. Yan, and Y. Ma, Local spiking pattern and its application to rotation-and illumination-invariant texture classification. Optik-International Journal for Light and Electron Optics 127 (2016) 6583-6589.
[80] M. Verma, and B. Raman, Local tri-directional patterns: A new texture feature descriptor for image retrieval. Digital Signal Processing 51 (2016) 62-72.
[81] S. Fadaei, R. Amirfattahi, and M.R. Ahmadzadeh, Local derivative radial patterns: A new texture descriptor for content-based image retrieval. Signal Processing 137 (2017) 274-286.
[82] T. Ahonen, A. Hadid, and M. Pietikainen, Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence 28 (2006) 2037-2041.
[83] V. Takala, T. Ahonen, and M. Pietikäinen, Block-based methods for image retrieval using local binary patterns. Image analysis (2005) 13-181.
[84] L. Liu, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, Local binary features for texture classification: Taxonomy and experimental study. Pattern Recognition 62 (2017) 135-160.
[85] B. Zhang, Y. Gao, S. Zhao, and J. Liu, Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. IEEE transactions on image processing 19 (2010) 533-544.
[86] N. Dalal, and B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE, 2005, pp. 886-893.
[87] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on pattern analysis and machine intelligence 19 (1997) 711-720.
[88] M. Felsberg, and G. Sommer, The monogenic signal. IEEE Transactions on Signal Processing 49 (2001) 3136-3144.
[89] L. Zhang, L. Zhang, Z. Guo, and D. Zhang, Monogenic-LBP: A new approach for rotation invariant texture classification, Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE, 2010, pp. 2677-2680.
[90] M. Yang, L. Zhang, L. Zhang, and D. Zhang, Monogenic binary pattern (MBP): A novel feature extraction and representation model for face recognition, Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, pp. 2680-2683.
[91] Y.-H. Oh, A.C. Le Ngo, J. See, S.-T. Liong, R.C.-W. Phan, and H.-C. Ling, Monogenic Riesz wavelet representation for micro-expression recognition, Digital Signal Processing (DSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 1237-1241.
[92] Z. Zeng, L. Song, Q. Zheng, and Y. Chi, A new image retrieval model based on monogenic signal representation. Journal of Visual Communication and Image Representation 33 (2015) 85-93.
[93] X. Huang, G. Zhao, W. Zheng, and M. Pietikainen, Spatiotemporal local monogenic binary patterns for facial expression recognition. IEEE Signal Processing Letters 19 (2012) 243-246.

[94] C.E. Erdem, C. Turan, and Z. Aydin, BAUM-2: a multilingual audio-visual affective face database. Multimedia Tools and Applications 74 (2015) 7429-7459.

[95] T. Kanade, J.F. Cohn, and Y. Tian, Comprehensive database for facial expression analysis, Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, IEEE, 2000, pp. 46-53.

[96] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, Coding facial expressions with gabor wavelets, Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, IEEE, 1998, pp. 200-205.

[97] L.-F. Chen, and Y.-S. Yen, Taiwanese facial expression image database. Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan (2007).

[98] C. Turan, K.-M. Lam, and X. He, Soft Locality Preserving Map (SLPM) for Facial Expression Recognition. arXiv preprint arXiv:1801.03754 (2018).

[99] D. McDuff, R. El Kaliouby, and R.W. Picard, Crowdsourcing facial responses to online videos, Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 512-518.

[100] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, and A. Batliner, The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. Affective computing and intelligent interaction (2007) 488-500.

[101] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark, Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE, 2011, pp. 2106-2112.

[102] X. Peng, Z. Xia, L. Li, and X. Feng, Towards facial expression recognition in the wild: a new database and deep recognition system, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 93-99.

[103] A. Mollahosseini, B. Hasani, and M.H. Mahoor, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. arXiv preprint arXiv:1708.03985 (2017).