Optical Engineering

SPIEDigitalLibrary.org/oe

Person-independent facial expression analysis by fusing multiscale cell features

Lubing Zhou
Han Wang

Person-independent facial expression analysis by fusing multiscale cell features

Lubing Zhou
Han Wang
Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Avenue
639798 Singapore
E-mail: zhou0145@e.ntu.edu.sg

Abstract. Automatic facial expression recognition is an interesting and challenging task. To achieve satisfactory accuracy, deriving a robust facial representation is especially important. A novel appearance-based feature, the multiscale cell local intensity increasing patterns (MC-LIIP), to represent facial images and conduct person-independent facial expression analysis is presented. The LIIP uses a decimal number to encode the texture or intensity distribution around each pixel via pixel-to-pixel intensity comparison. To boost noise resistance, MC-LIIP carries out comparison computation on the average values of scalable cells instead of individual pixels. The facial descriptor fuses region-based histograms of MC-LIIP features from various scales, so as to encode not only textural microstructures but also the macrostructures of facial images. Finally, a support vector machine classifier is applied for expression recognition. Experimental results on the CK+ and Karolinska directed emotional faces databases show the superiority of the proposed method. [DOI: 10.1117/1.OE.52.3.037201]

Subject terms: image processing; image representation; facial expression recognition; support vector machine; local features.

Paper 121395 received Sep. 26, 2012; revised manuscript received Feb. 1, 2013; accepted for publication Feb. 4, 2013; published online Mar. 4, 2013.

1 Introduction

The goal of facial expression recognition is to determine the emotional words saying on the face, e.g., anger, disgust, fear, joy, sadness, surprise or neutral. These emotions are generalized across human beings, regardless of the countries and races. They provide a powerful, natural and immediate indication of the inner thoughts and intentions of persons. Automatic emotion detection is an important and challenging task. It has practical significance in human computer interaction, lie detection, data-driven animation and video indexing.[1] Although much progress has been made,[1–6] facial expression recognition with a high accuracy is still a tough problem due to the complexity and variety of facial expressions.

After the face is located, facial expression analysis generally involves two components: facial feature representation and facial expression recognition. The former can be further divided into two types: geometric feature-based methods and appearance-feature based methods. On the other hand, depending on whether the temporal information is used, the latter component can also be divided into frame-based and sequence-based methods.[1]

Facial feature representation is a crucial step for successful facial expression recognition, and it constitutes the main topic of this paper. A good facial descriptor should be able to minimize the intra-expression variations while maximize the inter-expression variations. Geometric features describe either the shape of facial components (eyes, nose, eyebrows, and mouth) or locations of facial landmarks (eye corners, brows, lip, and, etc.), and a feature vector is formed to represent the face geometry. The paper[7] tracked a couple of facial features and adopted a constrained local models to represent the expression faces. Another representative is the famous Facial Action Coding System,[8] which uses so-called Action Units to encode the movements of muscles. Appearance-based features extract the intensity distribution or textures of the whole face or specific face regions. Gabor wavelets have been widely adopted to extract the facial texture frequencies due to the favourable performance.[4,9,10] The drawback is the tremendous time and memory costs to convolve facial images with a bank of Gabor filters. Another notable appearance-based feature is the local binary patterns (LBP)[11] and its variants.[12] The LBP, which was originally introduced for texture analysis, encodes pixels in an image by a decimal number via thresholding the 3×3 neighbourhood of the central pixel. Since the first time that LBP was successfully used as a facial representation for the purpose of face recognition,[13] many works have been done to extend its scope in terms of feature variants or combining with different learning methods. The work[5] presented a comprehensive study of facial expression recognition based on LBP features. The LBP is simple, efficient and invariant to monotonic illumination changes, because it is based on pixel-wise comparisons. Also due to the simple pixel-to-pixel comparisons, the LBP operator is sensitive to white noise.[6] The paper[6] proposed a novel descriptor to identify expressions based on a local directional pattern (LDP) feature which computes eight directional edge responses and encodes their strength rankings with an 8-bit string. Similarity-normalized shape (SPTS) and canonical appearance (CAPP) features were combined to gain promising recognition rates in the work.[14]

Based on the extracted features, facial expression recognition are commonly accomplished by applying the state-of-the-art pattern recognition techniques including template

0091-3286/2013/$25.00

matching, linear programming, subspace methods like eigenfaces and fisherfaces, Support vector machines (SVM), neural network (NN), hidden Markov models (HMM), Adaboost, and others. In expression recognition, the input can be either a single frame or a sequence of frames. The recognition methods are classified into frame-based and sequence-based categories, respectively. Intuitively, frame-based expression recognition uses a static image as input or treats each frame of a sequence independently. Geometric features can be located and extracted to form a specific expression-specified model, or appearance features may be applied to extract the textures in the frame. Frame-based method is relatively simple and direct, and can recognize the expression label for each individual image. Currently, most of the proposed methods belong to frame-based type.[4–6,10,14] On the other hand, sequence-based approaches use temporal information in the sequence by tracking some specific facial features over all frames. Using the dynamic facial motion can help to produce more accurate and robust expression recognition.[15] And the key challenge is the tracking the face and its features,[2] which affects the final recognition accuracy. Yeasin et al.[16] used the horizontal and vertical components of the flow as features, discrete HMMs were trained to recognize the temporal signatures associated with each type of expressions. Aleksic and Katsaggelos[17] proposed facial animation parameters as features and used multistream HMMs for recognition. In the work,[15] two spatio-temporal local features were proposed: the volume local binary patterns (VLBP and LBP from three orthogonal planes (LBP-TOP), and SVM classifier was utilized. Gabor filter and genetic algorithm (GA) were combined for dynamic analysis of facial expression from video sequences in the paper.[18]

In this work, we present a novel facial descriptor based on multiscale cell local intensity increasing patterns (MC-LIIP) for person-independent facial expression analysis. The LIIP is a texture feature which was first adopted for open/closed eye state recognition via boosting algorithm.[19] The MC-LIIP extends LIIP operator from pixel level to scalable cell level in order to enhance the noise resistance. It can represent the textural micro-structures and macro-structures of images in various scales. SVM classifier is applied to conduct the recognition task. The rest of the paper is organized as follows: in Sec. 2 a short review of related local features is given; Sec. 3 introduces the proposed multiscale approach for facial expression recognition; then Sec. 4 presents the experiment evaluation of the proposed method, and finally the conclusion is drawn.

2 LBP and LDP

The LBP has gained great success for face recognition and facial expression recognition.[5,13] LBP operator assigns pixels in an image with 8-bit string by thresholding the intensity of a 3×3 neighborhood with central pixel value, and then transforms the string to a decimal label, as expressed in Eq. (1).

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=0}^{7} s(i_p - i_c)2^p, \qquad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{1}$$

where i_p is the intensity of p'th neighbor and i_c is intensity of central pixel. Totally, there are $2^8 = 256$ patterns. Uniform LBP (ULBP)[11] is one most frequently used LBP variant that reduces the pattern number to 59. Uniform patterns are the patterns that include at most two 1–0 or 0–1 transitions in the circular 8-bit string. These patterns present majority of the prominent micro-structures such as edges, spots and corners. Without further statement, LBP refers to ULBP in following context. Facial representation is often formed by concatenating regional histograms of these local features under certain face partition. The LBP is computationally efficient and robust to monotonic illumination changes. But one weakness is that the pattern is easily affected by random noise or small intensity perturbations.[6]

The LDP feature[6] was proposed to enhance the resistance to moderate white noise or nonmonotonic illumination changes. The LDP convolves a local intensity patch with eight Kirsch edge masks, and sets the three directions with highest responses to 1, and others to 0. In this way, it produces ${}^8C_3 = 56$ patterns in total. The authors reported better facial expression recognition results than the LBP method[5] by using different frame sequences of the Cohn-Kanade (CK) database.[20] The drawback of LDP is the increased computational complexity from calculating and ranking the eight edge responses.

3 Multiscale Feature Fusion Approach

In this section, we propose the use of a multiscale feature fusion approach for facial expression recognition. The texture feature is first introduced and further extended to represent facial images, and then the emotions are recognized by applying a SVM classifier.

3.1 *Local Intensity Increasing Patterns*

The LBP features can represent the micro-level texture information of edges, spots and corners using the information of intensity changes. On the other hand, the quantized magnitude and orientation of local gradient has also been proved to be a very discriminative feature, and it was defined as the famous SIFT feature.[21] The motivation of LIIP is to develop a local feature that retains the simplicity and efficiency of texture feature LBP, and meanwhile contains sufficient gradient information.

Similar to LBP and LDP, LIIP encodes pixels in an image with decimal labels to express the intensity context. Figure 1 presents the generation of LIIP code. Given a 3×3 intensity patch, the eight surrounding pixels are thresholded to 1 or 0 via intensity comparison with the central pixel. A pixel is encoded as 1 if its intensity is greater than the central pixel, otherwise it is encoded as 0. Code 1 implies the intensity change in corresponding direction is increasing from

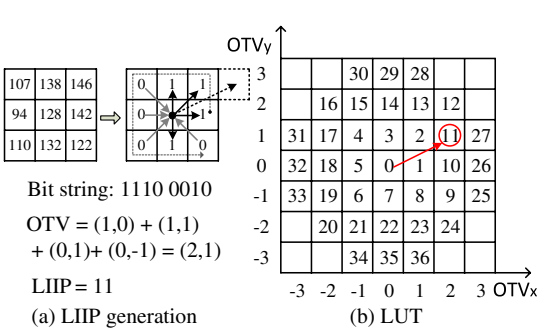

Fig. 1 LIIP pattern generation.

central pixel toward that surrounding pixel, and code 0 means the opposite increasing trend. The increasing trends with respect to the eight surrounding pixels are represented by eight trend vectors (TVs) as indicated by the solid arrows. Note the origin locates at the central pixel. An individual TV can only depict the intensity change trend in one direction of the local patch. To describe the overall trend, the overall trend vector (OTV) is defined by summing the eight individual TVs. From Fig. 1(a), it is easy to see that the sum of 1-coded TVs (black solid arrows) equals to the sum of 0-coded TVs (grey solid arrows) for symmetrically distributed surrounding pixels. So equivalently, the OTV is obtained by summing all the 1-coded TVs. The resulting OTV points from low-intensity area toward high-intensity area. It is capable of indicating the intensity increasing direction and even rough magnitude. Finally, the OTV is transformed to a decimal pattern label using a predefined two-dimensional look-up table (LUT) as displayed in Fig. 1(b). For instance, the OTV of the bit string 1110 0010 in Fig. 1(a) is (2, 1), and the label is 11. Overall, there are 37 patterns for 3×3 patch.

The LIIP has several excellent properties as a texture feature. First, it inherits the advantages from traditional LBP: implementation simplicity and runtime efficiency. In addition, LIIP is invariant to monotonic illumination changes as the operator is based on pixel-wise intensity comparisons. Second, the number of patterns 37 is favourable for the commonly adopted region-histogram based facial representation. ULBP and LDP contain 59 and 56 patterns, respectively. Less pattern number leads to fewer informationless zero histogram bins when the region size is moderate or smaller. Third, LIIP is able to encode the uniform patterns like ULBP. The peripheral patterns in the LUT of Fig. 1(b) fall into the uniform patterns that carry more texture information, while these patterns that are close to the origin mostly fall into nonuniform patterns that are more textureless. Figure 2 gives four examples: the left two patterns are uniform patterns and they represent more informative micro-patterns such as edge and corner, while the other two patterns locate nearby the origin and belong to nonuniform patterns. Finally, LIIP provides rough implication of gradient information. The larger pattern values in the LUT correspond to the local patches with more biased intensity distribution or higher gradient toward the OTV direction, and smaller values correspond to such patches that have no dominant gradient toward any direction.

3.2 *Multiscale Cell LIIP*

The motivation of MC-LIIP is to address the following drawbacks of the LIIP, LBP and LDP operators for facial representation. In LBP and LIIP, the bit string or pattern label is generated from pixel-wise comparison that is easily affected by noise.[6] LDP handles the issue by replacing the pixel comparison with ranking of eight directional edge responses. In addition, the above three features calculated from the local 3×3 neighborhood can only capture textural micro-structures, and is incapable of providing information of larger scale macro-structures which is more robust and may be the dominant features.

In MC-LIIP, the comparison computation is applied to the mean values of small cells instead of individual pixels. At scale s, the cell size is $s \times s$ [Fig. 3(a)–3(d)]. Giving an original image, a set of cell-based average images are first obtained by calculating the mean values of nonoverlapping $s \times s$ cells. Note that the original image can be considered as the average image at scale 1 [Fig. 3(e)] because the cell size is 1×1, and the size of average image at scale s is $1/s$ of the original image [Fig. 3(f)–3(h)]. Next, these multiscale cell-based average images are transformed to LIIP label images. That is why we call it multiscale cell LIIP. For simplicity, the MC-LIIP feature of scale s is notated as MC-LIIP$_s$. In fact, MC-LIIP$_s$ operator is composed of nine $s \times s$ cells in the original image as shown in Fig. 3(a)–3(d). Figure 3 presents the four MC-LIIP$_s$ operators of $s = 1, 2, 3, 4$, the corresponding nonoverlapping cell-based average images and the MC-LIIP$_s$ images. It is observed, the forehead and the cheeks below eyes in the face are encoded to textureless MC-LIIP patterns with label 0 (darkest), because these areas have been deteriorated due to strong highlights, It needs to mention that the cell-based averaging can be processed efficiently by direct cell summing for small s, or integral image[22] for big s. Thus, the extraction of MC-LIIP$_s$ is still very fast.

Figure 4 shows the MC-LIIP$^{1-4}$ images of two emotional facial images. For the purpose of illustration convenience, the average images are calculated from overlapping cells instead of nonoverlapping cells. It is observed that, the MC-LIIP$_s$ images with small s presents small-scale, local, more sensitive and micro patterns of face structures, which is profitable to describe the local detailed features of the face. In contrast, MC-LIIP$_s$ with large s shows better noise resistance, and presents relatively larger-scale, regional and macro patterns of face structures, but the local details are dropped. So the features from different scales can provide complementary information to each other. The rationale behind the proposed facial descriptor is to fuse the MC-LIIP$_s$ features at various scales, so as to get a more complete facial representation. From the MC-LIIP images in Fig. 4, we can observe the obvious differences in the

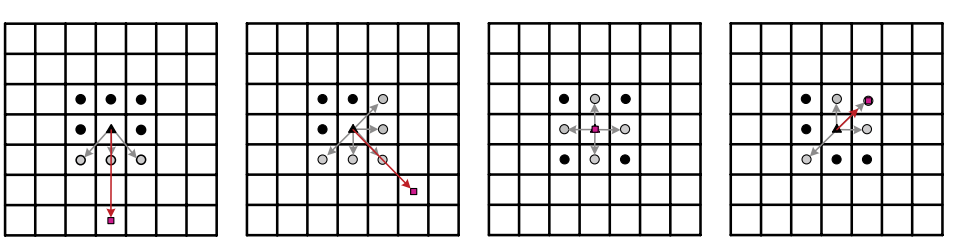

Fig. 2 Four LIIP patterns, bright circle: higher intensity than the center, black circle: lower intensity than the center, red arrow: OTV.

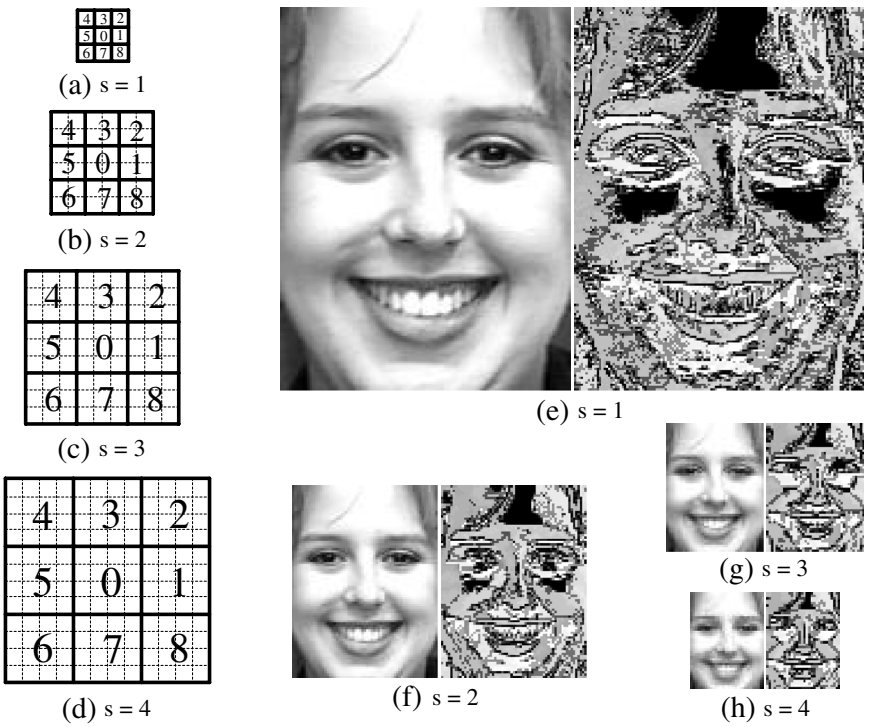

Fig. 3 MC-LIIP label images at different scales, (a), (e): $s = 1$; (b), (f): $s = 2$; (c), (g): $s = 3$; and (d), (h): $s = 4$.

forehead, eyes and mouth areas between the joy and surprise images.

3.3 *Feature Fusion and Classifier*

In above section, the MC-LIIP$_s$ images are gained at several scales. To distinguish different expressions, a rational feature representation of facial images is a vital factor. Directly using the raw pixels in a pattern image can fully retain the spatial localization information of the local image structures, but it is usually susceptible to noises and within-class variations because it is too local to being reliable. Contrarily, a global representation like a single histogram is less influenced, however, it discards the whole spatial distribution information of patterns. Often, a compromise is reached between global and local features. For small images like eyes (20×40), histograms of patterns are extracted from regions that are exhaustively explored by a sliding window at various locations and scales.[19] Comparatively, larger facial images commonly adopt a grid partition and local patterns histograms are extracted from the divisions.[5,6,13,15] Our proposed facial descriptor also utilizes grid-based regional histograms.

The feature fusion procedure of the facial representation is illustrated in Fig. 5. Given a normalized facial image, non-overlapping cell-based MC-LIIP$_s$ images are first computed at L scales as shown in Fig. 3 ($s = 1, \ldots, L$). Then, MC-LIIP$_s$ images are divided into M_s regions $R_0, R_1, \ldots, R_{M_s}$, from which the pattern histograms are extracted and individually normalized to unit length. The jth histogram bin of the i'th region at scale s before normalization is expressed as

$$H_s(i,j) = \sum_{(x,y)\in R_i} 1[\text{MC-LIIP}_s(x,y) = j], \tag{2}$$

where $1(*)$ is the zero/one indicator function. Finally, all the histogram bins from different regions of different scales are concatenated together to a single feature vector. Such an united facial descriptor is more comprehensive as it contains texture features from multiple scales. In addition, the descriptor conveys textural and statistical information in several levels: pixel and cell level for MC-LIIP operator, regional level for block histograms, and global level for

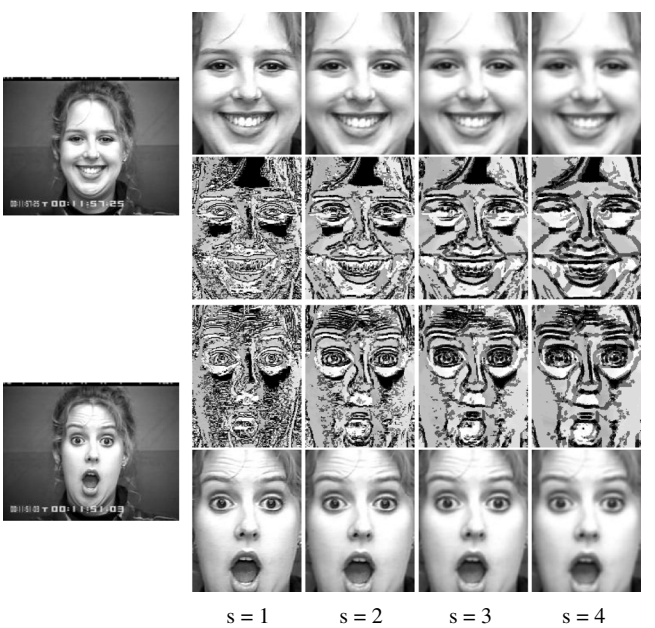

Fig. 4 MC-LIIP illustration of two emotional images, rows 1 and 4 are the overlapping cell-based average images, rows 2 and 3 are the corresponding MC-LIIP images, and scale s of each column is given.

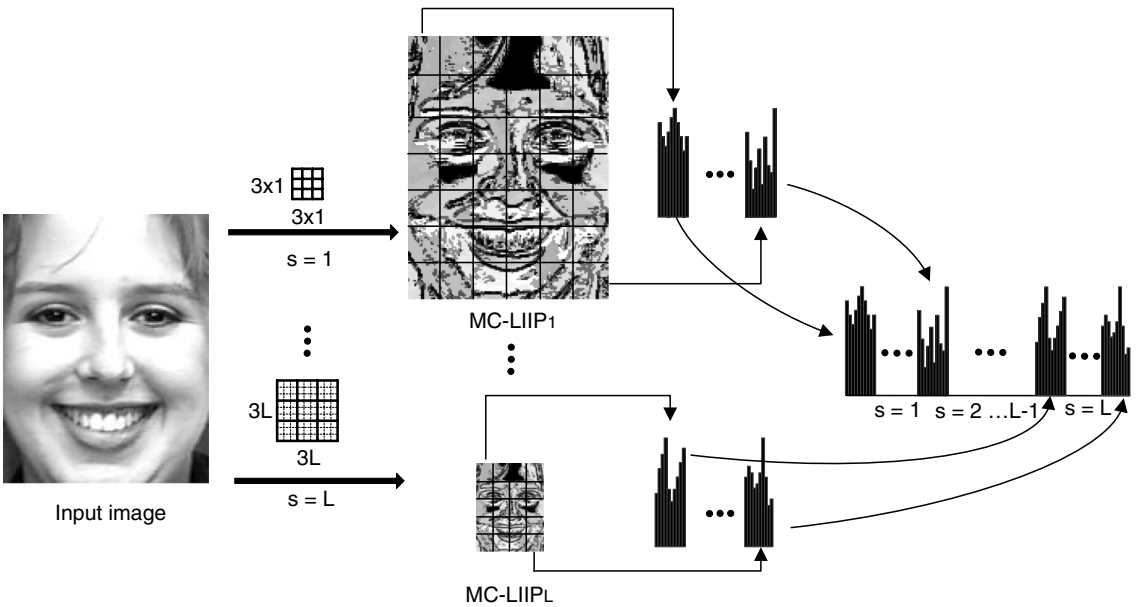

Fig. 5 Fusion procedure of MC-LIIP$_s$ features.

the feature vector. Feature size of the descriptor is $37\sum_{s=1}^{L} M_s$.

Multiclass support vector machine (SVM) has been proved to be the best classifier for facial expression classification.[4–6] In two-class case, SVM maps data onto a higher dimensional space, where a linear hyperplane with maximal margin is found between the two classes. Then a new sample x is classified by the decision function:

$$f(x) = \text{sign}\left[\sum_i \alpha_i y_i K(x_i, x) + b\right], \tag{3}$$

where (x_i, y_i) are support vectors, α_i are Lagrange multipliers, b is the hyperplane parameter, and K is a kernel function. More details of SVM can be found in the paper.[23] The only classifier used in this work is a SVM with linear kernel, because it is fast and our aim is not to evaluate different classifiers. Original SVM only makes binary decision, and multiclass SVM is achieved using one-against-rest scheme. As suggested in paper,[23] grid-search during 10-fold cross-validation is applied to optimize SVM parameters. The setting that produces the best cross-validation recognition rate is picked. This work uses the SVM implementation in the public OpenCV library (http://sourceforge.net/projects/opencvlibrary/).

4 Experiments

Most papers attempt to evaluate the proposed methods in recognizing the six universal emotions: anger, disgust, fear, joy, sadness and surprise. And often, the 6-class classification is extended to 7-class case by bringing in a neutral expression. In this work, both 6-class and 7-class expression recognition are considered. The algorithm evaluation and comparative experiments are mainly conducted on the extended Cohn-Kanade (CK+) database.[14] We investigate the recognition performance of MC-LIIP features at different or fused scales. In addition, the recognition accuracies of our approach and several other approaches on the Karolinska directed emotional faces (KDEF) database are also provided.

4.1 *Experimental Data and Setup*

The most widely used test-bed for facial expression analysis is the Cohn-Kanade (CK)[20] database, which consists of 486 sequences of 97 subjects. But the problem of CK dataset is that it is hard to evaluate newly proposed algorithm against reported results of existing works since the standard emotion labels are not provided. Commonly, authors pick and manually label different sequences from the database. The CK+ database[14] is an augmented distribution of CK set, with an additional 22% sequences and 27% subjects. More important, CK+ provides standard emotion labels of 309 sequences of 106 subjects for the six universal emotions. Thus, we choose CK+ as the main evaluation dataset.

CK+ database is composed of 123 subjects aged from 18 to 50 years, with 69% female, 81% Euro-American, 13% Afro-American, and 6% other groups. Image sequences start from a neutral emotion and gradually end at a peak prototypic emotion. They are digitized into 490×640 or 480×640 pixel arrays. In the experiments, the first neutral frame and last three peak frames were picked from the 309 labelled sequences, resulting in 1,236 images, including 135 anger, 177 disgust, 75 fear, 207 joy, 84 sadness, 249 surprise and 309 neutral. Following the works,[5,10] facial images were cropped from original image and normalized to 110×150 pixels by restricting a fixed distance between the two eyes. No further geometric or illumination normalization was performed in our experiments. Note that MC-LIIP possesses fair stability to illumination changes.

Person-independent repeated 10-fold cross-validation was adopted to ensure the generalization performance to new subjects. That is, we first divided all the 106 subjects

Table 1 Performance evaluation of different image partitions (MC-LIIP$_1$).

Partitions	Feature size	6-Class (%)	7-Class (%)
5×6	1110	91.9 ± 5.1	89.0 ± 3.9
6×7	1554	91.9 ± 4.5	89.2 ± 4.1
7×8	2072	93.6 ± 4.4	89.9 ± 4.3
8×9	2664	93.7 ± 4.0	91.8 ± 3.7
9×11	3663	93.1 ± 3.9	90.7 ± 3.8

Table 2 Recognition results of different MC-LIIP$_s$ methods.

Features	Feature size	6-Class (%)	7-Class (%)
MC-LIIP$_1$	2664	93.7 ± 4.0	91.8 ± 3.7
MC-LIIP$_2$	2072	94.7 ± 3.7	91.7 ± 3.8
MC-LIIP$_3$	1110	94.1 ± 3.9	90.7 ± 3.8
MC-LIIP$_4$	1110	93.7 ± 3.3	88.4 ± 3.5
MC-LIIP$_{12}$	4736	94.1 ± 4.0	92.1 ± 3.7
MC-LIIP$_{123}$	5846	94.6 ± 3.8	92.8 ± 3.7
MC-LIIP$_{1234}$	6956	95.3 ± 3.6	93.2 ± 3.4

randomly into 10 groups with roughly equal number of subjects, and then nine groups were used to train the classifier, while the rest one was used to test the classifier. Each group had only one chance of being the test data in the 10 rounds, and a mean recognition rate was gained. The above process was repeated 10 times to achieve an average accuracy by redoing the random data grouping each time.

4.2 *Determination of Image Partitions*

MC-LIIP$_s$ images are divided into different nonoverlapping blocks, from which the MC-LIIP$_s$ histograms are extracted. The recognition performance can be boosted by adjusting the block partition. Table 1 lists the recognition rates of several partitions for MC-LIIP$_1$ image. It is observed that either too few or too many regions will degrade the accuracy. The former case is due to that much spatial location information of local textures is missing, while the latter causes more zero bins in the regional histograms, which is too local to be robust. From the table, a 8×9 partition is optimal for MC-LIIP$_1$. Similarly, the optimized partitions for MC-LIIP$_2$, MC-LIIP$_3$ and MC-LIIP$_4$ images are 7×8, 5×6 and 5×6.

4.3 *Evaluation of* MC-LIIP$_s$ *Features*

Experiments were carried out to evaluate MC-LIIP$_s$ feature at individual scale as well as feature fusion of several scales. Table 2 records the results. Note that notation MC-LIIP$_{123}$ means the descriptor is a fusion of features from scales $s=1, 2, 3$. The best results come from the MC-LIIP$_{1234}$, feature fusion of the four scales. It achieves accuracies of 95.3% and 93.2% for 6-class and 7-class expression recognition, respectively. Among the four features of individual scale, MC-LIIP$_2$ performs the best, because it maintains a better balance between local details and noise resistance.

The confusion matrices (CM) of 6-class and 7-class recognition with MC-LIIP$_{1234}$ and linear SVM are shown in Tables 3 and 4. Comparing with the reported results of "LBP + SVM (RBF)" method,[5] the recognition performance for every expression is increased or at least comparable, except sadness. Also as reported in their work, we observe that high recognition rates can be achieved for disgust, joy, surprise, and neutral, while most confusions come from fear and sadness. Interestingly, happiness (joy) is the simplest emotion, while sadness is the hardest emotion.

4.4 *Comparison with Other Methods*

So far the majority of published methods are frame-based type, and many others belong to sequence-based type. For the evaluation database, previous methods mostly utilized the CK/CK+ database.[1,3,5,10,15,18] The data is suitable for both frame-based and sequence-based types. Hence, algorithm comparisons with other state-of-the-art approaches are mainly based on this dataset. These methods adopted different facial descriptors based on discriminative feature representation, and applied a SVM classifier unless otherwise stated. Table 5 compares our method with those methods in terms of person number (PN), sequence number (SN), class number (CN), whether person-independent(PI), dynamic or static, and with different metrics. It provides the overall recognition results with the CK/CK+ database. It should be clarified that the PN and SN vary among different methods. The reason is that the emotion labels were lacked in first distribution of CK database, and authors personally selected and labelled the sequences for algorithm evaluation before year 2010. In the new distribution CK+, the 7 expressions of 309 sequences of 109 subjects are clearly labelled. Our method is evaluated with all these 309 sequences, hence the experimental setup is replicable. In Table 5, it is observed that the proposed method exhibits superior recognition rates than other frame-based

Table 3 CM of 6-class facial expression recognition (%).

	Anger	Disgust	Fear	Joy	Sadness	Surprise
	(%)	(%)	(%)	(%)	(%)	(%)
Anger	**93.2**	1.7	0	0	4.3	0.8
Disgust	0.9	**99.1**	0	0	0	0
Fear	1.3	0	**87.2**	4.1	0	7.4
Joy	0	0	0	**100**	0	0
Sadness	15.2	1.1	2.4	0	**78.1**	3.2
Surprise	0.1	0.4	0.2	0.1	1	**98.2**

Table 4 CM of 7-class facial expression recognition (%).

	Anger	Disgust	Fear	Joy	Sad	Surprise	Neutral
	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Anger	**84.9**	1.2	0	0	0.2	0	13.7
Disgust	0	**96.4**	0	0	0	0	3.6
Fear	0.4	0	**78.6**	4.1	0	4.4	12.5
Joy	0	0	0	**100**	0	0	0
Sadness	5.6	0	0	0	**68.7**	0.6	25.1
Surprise	0	0	0.1	0	0	**96.6**	3.3
Neutral	0.8	0	0.1	0	0.7	0.4	**98.0**

methods, and most of the dynamic methods. For 6-class problem, the performance of our method is slightly inferior to the spatial-temporal LBP approach,[15] which introduces an extra temporal plane to the original LBP. Considering it is a marginal difference and the experimental setups such as PN and SN are different, the performance of the proposed method is still reasonable and satisfactory.

Additionally, we pay extra attention to the LBP and LDP approaches since they are similar to ours in the algorithm. The proposed MC-LIIP outperforms the other two. Notably, LDP-based method[6] reported comparable results to ours, but person-independent recognition was not performed in their work. In addition, we have conducted the experiments with LBP-based and LDP-based methods using the similar setup as our method indicated in Table 5. Separately, the 6(7)-class recognition rates of the LBP-based and LDP-based approaches are: 89.4% (91.1%) and 90.4% (92.2%). Considering processing speed, the average times to extract LBP, LDP, MC-LIIP$_1$ and MC-LIIP$_{1234}$ for one frame on a 3.2 GHz i5 CPU are 0.574 ms, 2.594 ms, 0.518 ms and 0.890 ms.

4.5 *Experiments on KDEF Database*

We further conducted experiments on the KDEF database,[24] which is a large-scale and more balanced facial expression database. The used data contains 980 images sized 562×762 from 70 subjects including 35 males. Each subject has 14 images, with 2 images for each of the seven emotions: anger, disgust, fear, joy, sadness, surprise and neutral. Compared to CK+, the data is less biased among different expressions, genders and subjects. Same as in CK+ database, 10-fold cross-validation was carried out for person-independent emotion detection, and the hit rates for the 6 (7)-class recognition were: anger—82.5% (81.5%), disgust—86.4% (86.4%), fear—69.7% (66.0%), joy—97.8% (97.8%),

Table 5 Comparison with different approaches on CK/CK+ database. (loso: leave-one-subject-out).

Method	PN	SN	CN	PI	Dynamic	Metric	Accuracy (%)
Gabor (Neural Network)[10]	97	375	6	N	N	—	92.2
Gabor[4]	90	313	7	Y	N	loso	88.0
LBP[5]	96	320	7 (6)	Y	N	10-fold	88.1 (91.5)
LDP[6]	96	408	7 (6)	N	N	7-fold	92.8 (94.9)
SPTS + CAPP[14]	—	327	8 (+contempt)	Y	N	loso	88.3
Optical flow + HMM[16]	97	—	6	Y	Y	5-fold	90.9
Multi-stream HMM[17]	90	284	6	Y	Y	—	93.66
Spatio-temporal LBP[15]	97	374	6	Y	Y	10-fold	96.26
Gabor + Genetic Algorithm[18]	115	321	6	Y	Y	loso	92.97
ours	106	309	7 (6)	Y	N	10-fold	93.2 (95.3)

Table 6 Comparison with different approaches on KDEF database.

Method	6-class accuracy (%)	7-class accuracy (%)
LBP[5]	81.6 ± 4.1	82.0 ± 3.8
LDP[6]	81.7 ± 4.0	81.7 ± 4.2
SPTS + CAPP[14]	—	82.8
ours	84.4 ± 3.9	84.5 ± 3.7

sadness—79.0% (77.2%), surprise—90.9% (90.7%), neutral—(92.2)%. This dataset is applicable to only the frame-based methods, since the dataset does not contains continuous sequence of frames. Our main motivation to conduct this experiment is to compare our method with the LBP and LDP methods as they are closely related, and provide a reference results with this database for future papers. We believe the KDEF dataset, which was originally developed for psychological and medical research purpose, is very valuable but has not received sufficient attention from facial expression analysis researchers. Table 6 records the recognition results. It is observed that the accuracies differ slightly for 6-class and 7-class expression because the data is less biased between classes.

5 Conclusion

This paper presents a new facial representation based on MC-LIIP for person-independent facial expression recognition. The MC-LIIP is a texture feature that combines the strengths of LBP, uniform patterns and gradient information, and indicates intensity distribution trend of local images. Furthermore, scalable MC-LIIP is capable of representing both local textural micro-patterns and macro-patterns. The MC-LIIP histogram based facial descriptor contains textural and statistical information in pixel, regional and global levels. Thus, the descriptor demonstrates good discriminative ability for expression recognition by incorporating with a SVM classifier. The proposed method can achieve reasonably higher accuracy with the CK+ and KDEF databases.

References

1. Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds., pp. 487–519, Springer Verlag, New York (2011).
2. M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Trans. PAMI* **22**(12), 1424–1445 (2000).
3. B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recogn.* **36**(1), 259–275 (2003).
4. M. S. Bartlett et al., "Recognizing facial expression: machine learning and application to spontaneous behavior," in *IEEE Conf. Comput. Vision Pattern Recogn.*, Vol. 2, pp. 568–573 (2005).
5. C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image Vision Comput.* **27**(6), 803–816 (2009).
6. T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI Journal* **32**(5), 784–794 (2010).
7. S. W. Chew et al., "Person-independent facial expression detection using constrained local models," in *Int. Conf. Automatic Face and Gesture Recognition*, pp. 915–920, IEEE Computer Society, Santa Barbara, California (2011).
8. P. Ekman and W. V. Friesen, *Facial action coding system: a technique for the measurement of facial movement*, Consulting Psychologists Press, Palo Alto, CA (1978).
9. J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech, Signal Process.* **36**(7), 1169–1179 (1988).
10. Y. Tian, "Evaluation of face resolution for expression analysis," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 82–89, IEEE Computer Society, Washington DC (2004).
11. T. Ojala, M. Pietikäinen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI* **24**(7), 971–987 (2002).
12. M. P. Inen et al., *Computer vision using local binary patterns*, Vol. 10, Springer Verlag, London (2011).
13. T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV*, pp. 469–481 (2004).
14. P. Lucey et al., "The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 94–101, IEEE Computer Society, San Francisco, Califonia (2010).
15. G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI* **29**(6), 915–928 (2007).
16. M. Yeasin, B. Bullot, and R. Sharma, "From facial expression to level of interest: a spatio-temporal approach," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 922–927, IEEE Computer Society, Washington DC (2004).
17. S. P. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multi-stream HMMs," *IEEE Trans. Inform. Forensics Sec.* **1**(1), 3–11 (2006).
18. S. Shojaeilangari, S. Yun, and T. E. Khwang, "Person independent facial expression analysis using Gabor features and genetic algorithm," in *IEEE Conf. on Information, Communications and Signal Processing*, pp. 1–5, IEEE Computer Society, Singapore (2011).
19. L. B. Zhou and H. Wang, "Open/closed eye recognition by local binary increasing intensity patterns," in *IEEE Conf. on Robotics, Automation and Mechatronics (RAM)*, pp. 7–11, IEEE Computer Society, Qingdao, China (2011).
20. T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 46–53, IEEE Computer Society, Grenoble, France (2000).
21. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
22. P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vision* **57**(2), 137–154 (2004).
23. C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (2003).
24. D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces—KDEF," CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institute, ISBN 91-630-7164-9 (1998).

Lubing Zhou received his BEng from Beihang University, Beijing, in 2008. From 2009 to current, he is pursuing his PhD in the School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore. His research interests include image processing, computer vision and machine learning.

Han Wang is in the School of EEE, Nanyang Technological University since 1992 and is currently an associate professor. He received his bachelor's degree in computer science from Northeast Heavy Machinery Institute in China and PhD from the University of Leeds in the United Kingdom, respectively. His research interests include computer vision and robotics. He has done significant research work in his research areas and published over 120 top quality international conference and journal papers. He has been invited as a member of the editorial advisory board of the *Open Electrical and Electronic Engineering Journal*. He is a senior member of IEEE.