Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 7206041, 12 pages
https://doi.org/10.1155/2017/7206041

*Research Article*

# Facial Expression Recognition from Video Sequences Based on Spatial-Temporal Motion Local Binary Pattern and Gabor Multiorientation Fusion Histogram

**Lei Zhao,[1] Zengcai Wang,[1,2] and Guoxin Zhang[1]**

[1] *Vehicle Engineering Research Institute, School of Mechanical Engineering, Shandong University, Jinan 250061, China*
[2] *Key Laboratory of High Efficiency and Clean Mechanical Manufacture, Ministry of Education, School of Mechanical Engineering, Shandong University, Jinan 250061, China*

Correspondence should be addressed to Zengcai Wang; wangzc@sdu.edu.cn

Received 14 November 2016; Revised 3 January 2017; Accepted 26 January 2017; Published 19 February 2017

Academic Editor: Alessandro Gasparetto



This paper proposes novel framework for facial expressions analysis using dynamic and static information in video sequences. First, based on incremental formulation, discriminative deformable face alignment method is adapted to locate facial points to correct in-plane head rotation and break up facial region from background. Then, spatial-temporal motion local binary pattern (LBP) feature is extracted and integrated with Gabor multiorientation fusion histogram to give descriptors, which reflect static and dynamic texture information of facial expressions. Finally, a one-versus-one strategy based multiclass support vector machine (SVM) classifier is applied to classify facial expressions. Experiments on Cohn-Kanade (CK) + facial expression dataset illustrate that integrated framework outperforms methods using single descriptors. Compared with other state-of-the-art methods on CK+, MMI, and Oulu-CASIA VIS datasets, our proposed framework performs better.

## 1. Introduction

Content and intent of human communication can be clarified by facial expressions synchronizing dialogues [1, 2]. Ekman and Friesen first proposed that six basic facial statuses reflect different emotion content [3]. Basic emotions are summarized into six facial statuses including happiness, anger, sadness, disgust, surprise, and fear and are often analyzed in facial expression recognition research. Attaining automatic recognition and understanding facial emotion will promote development of various fields, such as computer technology, security technology, and medicine technology in applications, such as human-computer interaction, driver fatigue recognition, pain and depression recognition, and deception detection and stress monitoring. Owing to practical value and theoretical significance of medical and cognitive scientists, facial expression recognition became of great interest in research [4, 5]. Recently, this field made much progress because of advancements in related research subareas, such as face detection [6], tracking and recognition [7], and new developments in machine learning areas, such as supervised learning [8, 9] and feature extraction. However, accurate recognition of facial expressions is still a challenging problem because of its dynamic, complex, and subtle facial expression changes and different head poses [10, 11].

Most recent facial expression recognition methods focused on means of analyzing frame of neutral expression or frame of peak phase of any of six basic expressions in video sequence (i.e., static-based facial expression recognition) [12]. Static-based method analyzes facial expressions and disregards some important dynamic expression features. Consequently, this method performs poorly in some practical applications. Dynamic-based method extracts time and motion information from image sequence of facial expressions. Movement of facial landmarks and changes in facial texture are dynamic features containing useful information

representing underlying emotional status. Therefore, dynamic information should be extracted from over entire video sequence during facial expression recognition.

This paper presents video-based method for facial expressions analysis using dynamic and static textures features of facial region. First, constrained local model (CLM) based on incremental formulation is used to align and track face landmark. Then, dynamic and static features are extracted and fused to enhance recognition accuracy: spatial-temporal motion LBP concatenating LBP histograms on XT and YT planes of whole video sequence to estimate changes of facial texture during the whole process and Gabor multiorientation fusion histogram in peak expression frame to represent status of facial texture when facial expression occur. Finally, classification is accomplished using multiclass SVM using one-versus-one strategy.

This paper mainly provides the following contributions: (a) integration of spatial-temporal motion LBP with Gabor multiorientation fusion histogram, (b) analysis on contributions of three LBP histograms on three orthogonal planes to accuracy of face expression recognition, and (c) spatial-temporal motion LBP facial dynamic features.

The remainder of this paper is structured as follows. Section 2 presents background of facial expression recognition methods and discusses contribution of current approaches. Section 3 shows detailed description of proposed approaches. Section 4 details experiments employed for evaluation of proposed framework performance. Finally, Section 5 concludes the paper.

## 2. Background

Facial expression analysis systems are designed to classify image or video sequences into one basic emotion of six previously mentioned expressions. Facial expressions change face shape and texture (i.e., deformation of eyebrows, mouth, eyes, and skin texture) through movement of facial muscles. These variations can be extracted and adopted to classify given facial image or video sequences. Though methods for facial expression recognition may include different aspects, recent research can be divided into two main streams: facial action unit based (AU-based) techniques and content-based techniques, both of which are detailed in Sections 2.1 and 2.2, respectively.

*2.1. AU-Based Expression Recognition.* In AU-based approaches, Facial Action Coding System (FACS) is extensively used tool for describing facial movements [25]. FACS bridges facial expression changes and motions of facial muscles producing them. This system defines nine different AUs in upper part of the face, eighteen AUs in lower part of the face, and five AUs which belong to neither upper nor lower part of the face. Moreover, this system defines some action units, such as eleven for describing head position, nine for states of eyes, and fourteen for other actions. AUs, which are the smallest identifiable facial movements [26], are used by many expression recognition systems [27–30].

In [31], several approaches are compared to classify AU-based facial expressions; these approaches include principal component analysis (PCA), linear discriminate analysis (LDA), independent component analysis, optical flow, and Gabor filters. Results showed that methods using local spatial features perform excellently in expression recognition. However, PCA and LDA technologies destroy capacity to discern local features.

The system mentioned in [26] tracks a series of facial feature points and extracts temporal and spatial geometric features from motions of feature points. Geometric features are less susceptible to physical changes, such as variation of illumination and differences among human faces. However, in some AUs, failure is observed in methods based on geometric features, such as AU15 (mouth pressure) and AU14 (dimple). The appearance of these AUs can change facial texture but not facial feature points.

Based on dynamic Bayesian network (DBN) for spontaneous AU intensity recognition, unified probabilistic model is proposed in [32]. The model contains two steps, that is, AU intensity observation extraction and DBN inference. Gabor feature, histogram of oriented gradients (HOG) feature, and SVM classifier-based framework are employed to extract AU intensity observation. Then, DBN is used to systematically model dependencies among AU, multi-AU intensity levels, and temporal relationships. DBN model combines with image observation to recognize AU intensity through DBN probabilistic inference. Results show that the proposed method can improve AU intensity recognition accuracy.

In [33] work, hidden Markov model (HMM) is used to model AU in time evolution process. Classification is accomplished by maximizing extracted facial features and probability of HMM. Work in [34] combines SVM with HMM to recognize AU, and recognition accuracy on each AU is higher than result yielded when using SVM for time evolution of feature. Although both methods consider time characteristics of AUs, interaction model among AUs is not built.

*2.2. Content-Based Expression Recognition.* Non-AU methods are content-based and use local or global facial features to analyze expressions. Content-based techniques for facial expression include static-based method [35, 36] and dynamic-based method [16, 18, 20, 22, 37].

In [37], Kanade-Lucas-Tomas tracker is used to track grid points on face to obtain point displacements. These points are used to train SVM classifier. However, this method only extracts geometric features through locating and tracking points on face. In [16], system integrating pyramid histogram of gradients on three orthogonal planes (PHOG-TOP) with optical flow is adapted to classify the emotions. Two kinds of features respectively represent movement of facial landmarks and variation of facial texture. Results show that integrated framework outperforms methods using single feature operator.

The work in [18] adopts method using texture operator called LBP instead of geometric features. Important features are selected by using boosting algorithm before training SVM

classifier. In [22], volume LBP (VLBP) and LBP histograms on three orthogonal planes (LBPTOP) are used to reveal dynamic texture of facial expressions in video sequence. Then, based on multiclass SVM using one-versus-one strategy, a classifier is applied. Experiments conducted on Cohn-Kanade (CK) dataset demonstrate that LBPTOP outperforms former methods.

To obtain higher recognition accuracy, combined geometric and texture features are adopted. The work in [20] fuses geometric feature and Gabor feature of local facial region to enhance the facial expression recognition accuracy. Dynamic information of video sequence is extracted for features to construct parametric space with 300 dimensions. Compared with other methods, system performs excellently in expression recognition when combining geometric and texture features.

*2.3. Discussion and Contributions.* AU is middle-level interpretation of facial muscle motion; such actions associate high-level semantics with low-level features of facial expressions and identify meaningful human facial movements [27]. However, drawbacks of the AU-based expression recognition include affiliation of errors to whole recognition process and chances of reduction in accuracy when middle-level classification step is added in the system. Static-based expression recognition technology only reflects expression state at specific time points. Dynamic features extracted from video sequence reveal facial expression changes and are more accurate and robust for expression recognition [16, 20, 22].

In this paper, proposed framework focuses on content-based and features of fusion-based facial expression recognition technique, and it identifies expression in video more accurately. Most feature extraction methods of recent work are content-based and are limited to single form of dynamic or static features, while in this paper we propose a method to recognize facial expression using the spatial-temporal motion LBP from the whole video sequence and Gabor multiorientation fusion histogram from the peak expression frame. Moreover, we first study contributions of LBP histograms from different orthogonal planes to evaluate accuracy of expression recognition and to concatenate LBP histograms on XT and YT planes to establish spatial-temporal motion LBP features based on significance of expression recognition. Thus, in this paper, framework outperforms those from previous studies while considering dynamic and static texture features of facial expressions.

## 3. Methodology

Proposed approach includes location and tracking of facial points (preprocessing), spatial-temporal motion LBP and Gabor multiorientation fusion histogram extraction (feature extraction), and expression classification. Figure 1 illustrates this method. Framework and its components are detailed in the following sections.

*3.1. Preprocessing.* When image sequence is introduced to facial expression recognition systems, facial regions should be detected and cropped as preprocessing step. Real-time Viola–Jones face detector [38], commonly used in different areas, such as face detection, recognition, and expression analysis, is adopted to crop face region coarsely in this paper. Viola–Jones face detector consists of cascade of classifiers employing Haar feature trained by AdaBoost. Haar feature is based on integral image filters, computed simply and fast at any location and scale.

To segment face regions more accurately, fiducial points should be located and tracked in video sequences. Based on incremental formulation [39], CLM model is employed to detect and track face points in our framework. Original CLM framework is patch-based method, and faces can be represented by series of image patches cropped from location of fiducial points. Patch-expert of each point is trained by linear SVM and positive and negative patches and used as detector to update point location. Instead of dealing with updating shape model, CLM model based on incremental formulation focuses on updating function, which can map facial texture to facial shape, and constructing robust and discriminatively deformable model that outperforms state-of-the-art CLM alignment frameworks [39].

Locations of detected points on contour of eyes are used to compute the angle between line of two inner corners and image level edge in each frame to crop the facial region from background. Angle can be rotated to horizontal axis to correct any in-plane head rotation. Points of two outer eyes corners and nose tip are used to crop and scale images into $104 \times 128$ rectangular region containing all local areas related to facial expressions. After preprocessing, center in horizontal direction of cropped image is $x$ coordinate of center of two eyes, while $y$ coordinate of nose tip is found in lower third in vertical direction of cropped image.

*3.2. Spatial-Temporal Motion LBP Feature.* LBP is well-known operator used to describe local texture characteristics of image and is first proposed by Ojala et al. [40]. Operator labels image pixels by thresholding $n \times n$ neighborhood of each pixel with value of center and considering results as binary numbers. Then, histogram of all labels in image can be adopted as texture descriptor. LBP is widely used in many aspects, such as face recognition, texture classification, texture segmentation, and facial expression recognition, because of gray scale and rotation invariance advantages.

Recently, some extended LBP operators are introduced to describe dynamic texture and outperform other texture operators. LBPTOP is spatial-temporal operator extended from LBP and is proposed in the work of [22]. As shown in Figure 2, LBP codes are extracted from three orthogonal planes (i.e., XY, XT, and YT) of video sequences. For all pixels, statistical histograms of three different planes can be, respectively, denoted as LBP-XY, LBP-XT, and LBP - YT and are obtained and concatenated into one histogram. Histogram is calculated as follows:

$$H_{i,j} = \sum_{x,y,t} I\left\{f_j(x,y,t) = i\right\}, \quad i = 0, \ldots, n_j - 1, \ \ j = 0, 1, 2, \tag{1}$$

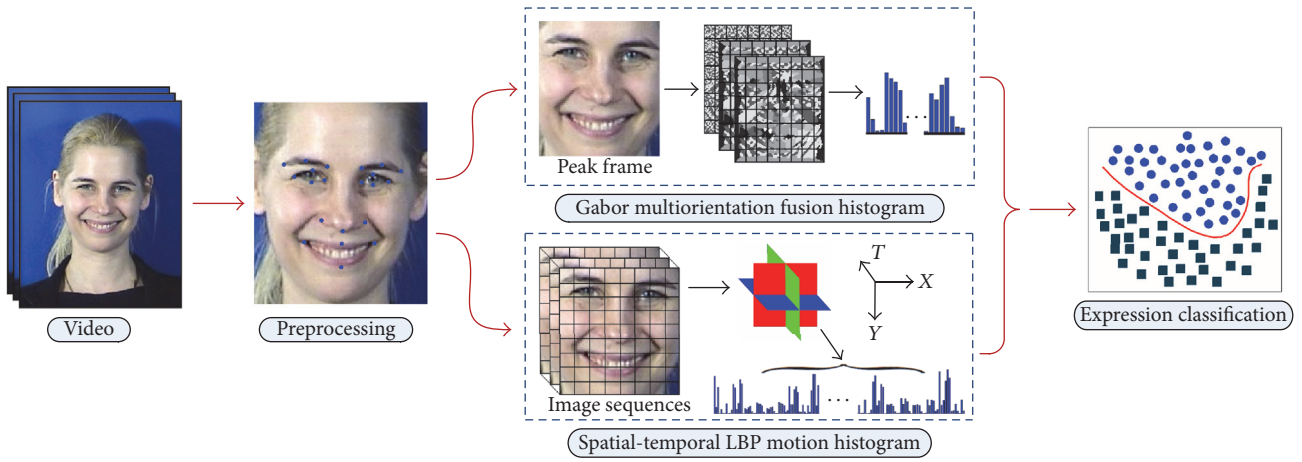


Figure 1: Proposed framework for facial recognition.

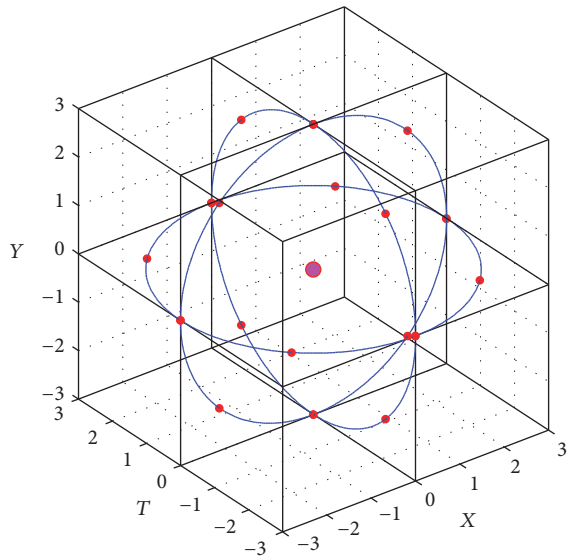


Figure 2: Example of LBP code using radii three and eight neighboring points on three planes.

where $n_j$ is numbers of different labels produced by LBP operator in $j$th plane ($j$ = 0: XY, 1: XT and 2: YT) and $f_j(x, y, t)$ represents LBP code of center pixel $(x, y, t)$ in $j$th plane.

As shown in formula (1) and Figure 2, dynamic texture in video sequence can be extracted by incorporating LBP histogram on XY plane (LBP-XY) reflecting information of spatial domain and LBP histograms on XT and YT planes representing spatial-temporal motion information in horizontal and vertical directions, respectively.

All LBPTOP components are LBP histograms extracted from three orthogonal planes and have equal contribution in facial expression recognition [22]. However, not all LBPTOP components are significant. LBP-XY includes characteristics of different subjects, resulting in difficulties in accurate expression recognition. LBP-XT and LBP-YT contain most information on spatial-temporal texture changes of facial expression. Consequently, as shown in Figure 3, LBP-XY is abandoned, and LBP-XT and LBP-XT are merged to construct spatial-temporal motion LBP feature, which is applied in facial expression recognition in our study. In this paper, LBP code is used and includes radii three and eight neighboring points on each plane; moreover, image of each frame is equally divided into $8 \times 9$ rectangular subblocks with overlap ratio of 70%.

*3.3. Gabor Multiorientation Fusion Histogram.* Gabor wavelet is well-known descriptor representing texture information of an image. Gabor filter can be calculated by Gaussian kernel multiplied by sinusoidal plane. Gabor feature is highly capable of describing facial textures used in different research, such as identity recognition, feature location, and facial expression recognition. However, Gabor performs weakly in characterizing global features, and feature data are redundant. Figure 4 shows extracted features based on Gabor multiorientation fusion histogram proposed in [41]; these features are used in reducing feature dimension and improving recognition accuracy.

First, Gabor filter of five scales and eight orientations is used to transform images. The following represents multiscale and multiorientation features of pixel point $z(x, y)$ in images:

$$\{G_{u,v}(z) : u \in (0, \ldots, 7),\ v \in (0, \ldots, 4)\}, \tag{2}$$

where $u$ and $v$ denote orientations and scales and $G_{u,v}(z)$ is norm of Gabor features.

Transformation of each original image produces 40 corresponding images with different scales and orientations. Dimension of obtained feature is 40 times as high as that of original facial image. Therefore, on the same scale, features of eight orientations are fused according to a certain rule. Fusion rules are as follows:

$$T_v(z) = \arg\max_u \{\|G_{u,v}(z)\|\}, \quad u \in (0, \ldots, 7),\ v \in (0, \ldots, 4), \tag{3}$$

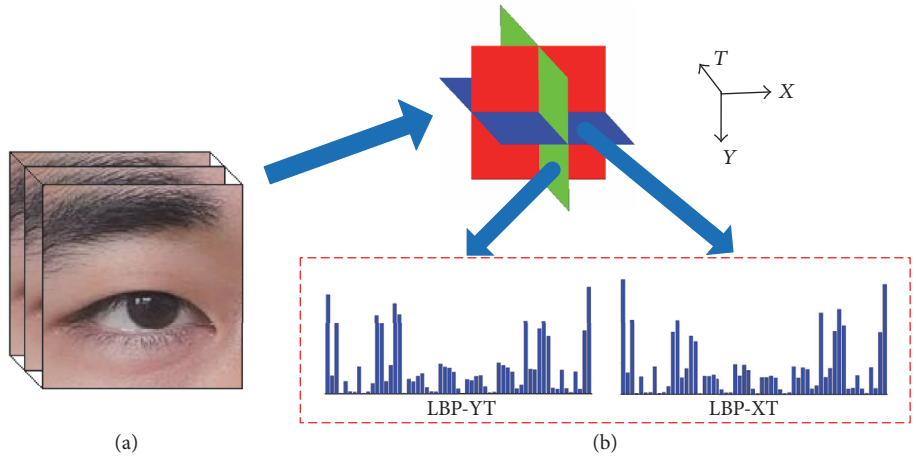


Figure 3: Features in each block sequence. (a) Block sequence and (b) LBP histograms from XT and YT planes.

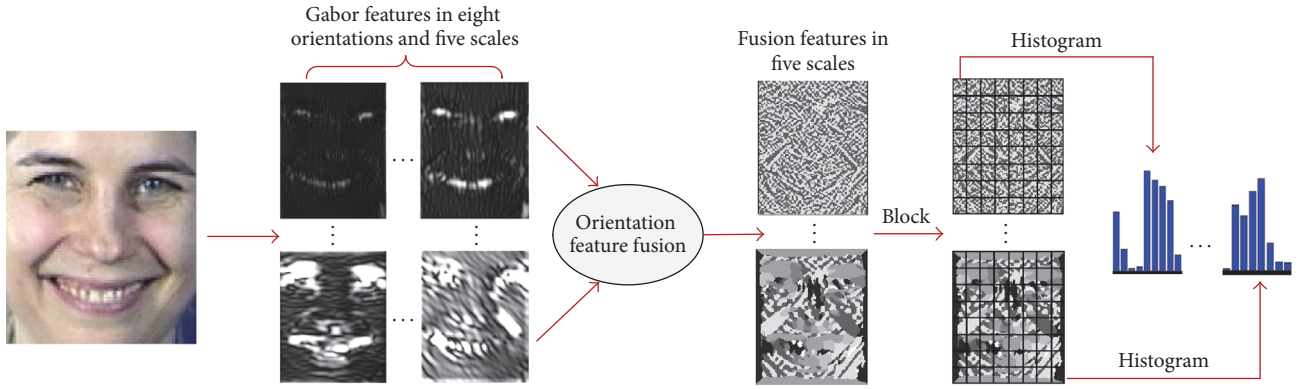


Figure 4: Feature extraction procedure based on Gabor multiorientation fusion histogram.

where $T_v(z) \in [1, 8]$ is encoding value of pixel point $z(x, y)$ and $G_{u,v}(z)$ is Gabor features.

Here, orientation of local area is evaluated by orientation code of maximum value of Gabor features of eight orientations. Each encoding value $T_v(z)$ represents one local orientation. Finally, each expression image is transformed into five scales of images, and each scale fuses eight orientations. Fusion images are shown in Figure 5. As can be seen from figure, fused images retain the most obvious change information of Gabor subfiltering. Therefore, Gabor fusion feature shows high discrimination for local texture changes.

Histogram can describe global features of images. However, many structural details maybe lost when histogram of entire image is directly calculated. Thus, each fusion image is divided into $8 \times 8$ nonoverlapping and equal rectangular subblocks. Accordingly, histogram distributions of all subblocks are calculated according to formula (4) and combined to construct Gabor multiorientation fusion histogram of whole image.

$$h_{v,r,i} = \sum_{z} I\left(R_{v,r}(z) = i\right), \quad i = 0, \ldots, 8, \tag{4}$$

where $R_{v,r}(z)$ is encoding value in point $z(x, y)$ of $i$th subblock.

As obtained by coding five scales on multiple directions of Gabor feature, Gabor multiorientation fusion histogram not only inherits Gabor wavelet advantage capturing corresponding spatial frequency (scale), spatial location, and orientation selectivity but also reduces redundancy of feature data and computation complexity.

*3.4. Multifeature Fusion.* Features extracted in Sections 3.2 and 3.3 are scaled to $[-1, 1]$ and integrated into one vector using the following equation:

$$F = \alpha P + (1 - \alpha) Q, \quad 0 \le \alpha \le 1, \tag{5}$$

where $P$ is spatial-temporal motion LBP histogram and $Q$ is Gabor multiorientation fusion histogram. $F$ is fusion vector of $P$ and $Q$ and is feature vector for final prediction. Parameter $\alpha$ usually depends on performance of each feature. For all experiments in this paper, we set value of $\alpha$ to 0.5 while considering experiments results shown in Tables 2 and 3.

*3.5. Classifier.* Many classifiers are used for facial expression recognition; these classifiers include J48 based on ID3, nearest neighbor based on fuzzy-rough sets, random forest [20], SVM [22], and HMM [29]. In this paper, SVM is selected

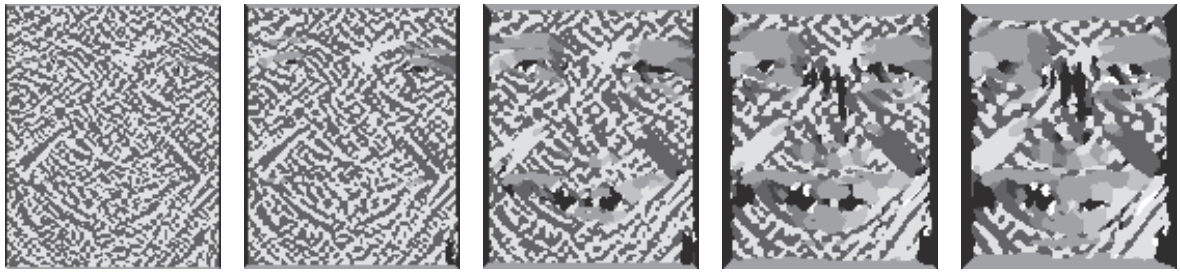

Figure 5: Fusion images on five scales.

in our framework because of the following properties: (1) classifier based on statistical learning theory, (2) high generalization performance, and (3) ability to deal with feature having high dimensions.

To achieve more effective classification of facial expressions, SVM classifier with RBF kernel is adapted. Given a training set of instance-label pairs $(x_i, y_i)$, the SVM requires the solution of the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i \left( w^T \phi(x_i) + b \right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \end{aligned} \tag{6}$$

where $x_i$ are training samples; $l$ is the number of training set; $y_i$ are the labels of $x_i$; $C$ is punishment coefficient that prevents overfitting; $\xi_i$ is relaxation variable; and $w$ and $b$ are the parameters of the classifier. Here training vectors $x_i$ are mapped into a higher dimensional space by using function $\phi(x_i)$. SVM finds a linear separating hyperplane with the maximal margin in the higher dimensional space. $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel function of SVM. The RBF kernel function is shown as follows:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \gamma > 0, \tag{7}$$

where $\gamma$ is parameter of RBF kernel. In this paper, grid-search strategy and tenfold cross validation [42] are used to select and estimate punishment coefficient $C$ and kernel parameters $\gamma$ of SVM. SVM is binary classifier used only in two sets of feature spaces of objects. Facial expression recognition is multiclass problem. Thus, multiclassifier, which consists of binary SVM classifiers, is constructed to classify facial expressions [42]. Two strategies are commonly used: one-versus-one strategy and one-versus-all strategy. One-versus-one strategy is applied in our method because of its robustness in classification and simplicity of application. In this strategy, SVM is trained for classifying any two kinds of facial expression. Thus, six expressions require training 15 binary SVM classifiers (fear versus anger, happiness versus disgust, and surprise versus sadness, among others). Final result is expression classification with highest number of votes. However, occasionally, class with most votes is not the only result. In this case, nearest neighbor classifier is applied to these classes to obtain one final result.

Table 1: Number of subjects for each expression class in three datasets.

| Expression | CK+ | MMI | Oulu (each condition) |
|---|---|---|---|
| Anger | 45 | 32 | 80 |
| Disgust | 59 | 28 | 80 |
| Fear | 25 | 28 | 80 |
| Happiness | 69 | 42 | 80 |
| Sadness | 28 | 32 | 80 |
| Surprise | 83 | 41 | 80 |
| Total | 309 | 203 | 480 |

## 4. Experiments

*4.1. Facial Expression Datasets.* CK+ dataset [14] is extended from CK dataset [43], which was built in 2000; it is available and most widely used dataset for evaluating performance of facial expression recognition systems. This dataset consists of 593 video sequences, including seven basic facial expressions performed by 123 participants. Ages of subjects range from 18 to 30 years; 35% are male; 13% are Afro-American; eighty-one percent are Euro-American; and six percent are people of other race. In video sequences, image of each frame is 640 × 490 or 640 × 480 pixels. Most frames are gray images with eight-bit precision for grayscale values. Video sequences contain frames from neutral phase to peak phase of facial expressions, and video rate is 30 frames per second. In our study, 309 image sequences containing 97 subjects with six expressions (anger, disgust, fear, happiness, sadness, and surprise) are selected for our experiments. Each subject has one to six expressions. Top row of Figure 6 shows peak frames of six expressions from six subjects in CK+ dataset. Table 1 shows number of subjects for each expression class in CK+ dataset in our experiment.

MMI dataset [44] is applied to evaluate performance of expression recognition methods; this dataset contains 203 video sequences, including different head poses and subtle expressions. These expression sequences were performed by 19 participants with ages ranging from 19 to 62 years. Subjects are Asian, European, and South American. Male participants account for 40% of subjects. Different from CK+ dataset, MMI dataset contains six basic facial expressions comprising neutral, onset, apex, and offset frames. In video sequences, image of each frame is 720 × 576 pixels with RGB full color. Original frames are processed into images with eight-bit

Table 2: Classification results of using spatial-temporal motion LBP on CK+ dataset.

| Expression | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Average |
|---|---|---|---|---|---|---|---|
| Anger | 42 | **0** | **0** | **0** | **3** | **0** | **93.3** |
| Disgust | **0** | 59 | **0** | **0** | **0** | **0** | **100** |
| Fear | **1** | **0** | 20 | **2** | **2** | **0** | **80.0** |
| Happiness | **0** | **0** | **0** | 69 | **0** | **0** | **100** |
| Sadness | **4** | **0** | **2** | **0** | 21 | **1** | **75.0** |
| Surprise | **0** | **0** | **2** | **1** | **0** | 80 | **96.4** |
| Overall | | | | | | | **94.1** |

Table 3: Classification results of using Gabor multiorientation fusion histogram on CK+ dataset.

| Expression | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Average |
|---|---|---|---|---|---|---|---|
| Anger | 40 | **2** | **0** | **0** | **3** | **0** | **88.9** |
| Disgust | **2** | 56 | **0** | **1** | **0** | **0** | **94.9** |
| Fear | **1** | **0** | 17 | **3** | **4** | **0** | **68.0** |
| Happiness | **0** | **0** | **0** | 69 | **0** | **0** | **100** |
| Sadness | **4** | **0** | **0** | **0** | 24 | **0** | **85.7** |
| Surprise | **0** | **1** | **0** | **0** | **0** | 82 | **98.8** |
| Overall | | | | | | | **93.2** |


Figure 6: Sample images in CK+ dataset and MMI dataset.

precision for grayscale values for our study and extracted frames from neutral to apex phases of facial expressions. Bottom row of Figure 6 shows peak frames of six expressions from six subjects in MMI dataset. Table 1 shows number of each expression class in MMI dataset in our experiment. As seen from Table 1, MMI dataset includes fewer subjects than CK+ dataset.

Oulu-CASIA VIS facial expression database [24] includes six basic expressions from 80 subjects between 23 to 58 years old. 26.2% of the subjects are females. All the image sequences were taken under the visible light condition. Oulu-CASIA VIS contains three subsets. All expressions in the three subsets were respectively captured in three different illumination conditions: normal, weak, and dark. The images of normal illumination condition are captured by using good normal lighting. Weak illumination means that only computer display is on and each subject sits in front of the display. Dark illumination means almost darkness. Video sequences contain frames from neutral phase to peak phase of facial expressions, and video rate is 25 frames per second. The face regions of six expressions in Oulu-CASIA VIS dataset are shown in Figure 7 (top row: normal, middle row: weak, and bottom row: dark). The resolution of the images in Oulu-CASIA VIS dataset (320 × 240) is lower than that of the images in CK+ and MMI datasets. The number of video sequences is 480 for each illumination condition. Table 1

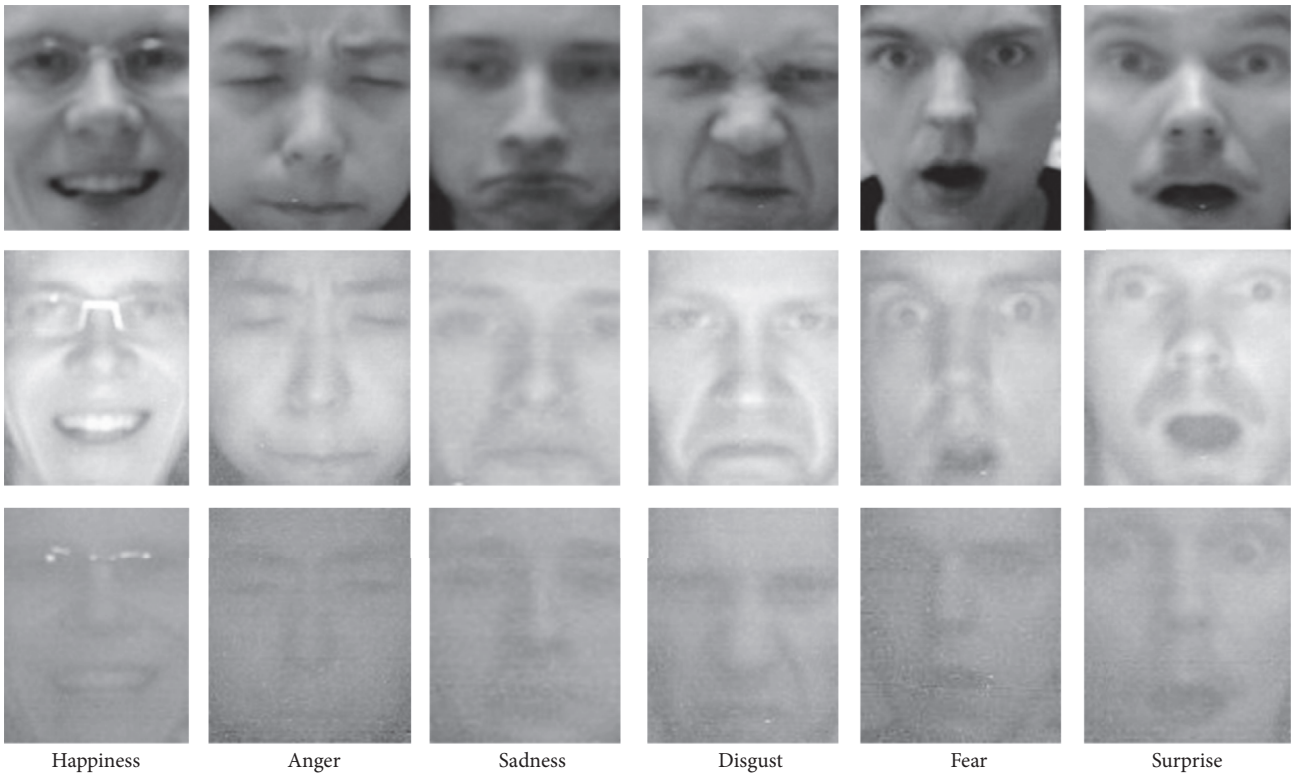


Figure 7: Sample images in Oulu-CASIA VIS dataset.

shows number of each expression class in Oulu-CASIA VIS dataset in our experiment.

*4.2. Experimental Results on CK+ Dataset.* Three sets of experiments are conducted to evaluate performance of our method on CK+ dataset because head motions of subjects in CK+ dataset are smaller than that in other two datasets. Then, in this paper, performance of framework proposed is compared with seven state-of-the-art methods. Leave-one-out cross validation strategy is applied in these sets of experiments. The punishment coefficient $C$ and kernel parameters $\gamma$ of SVM for CK+ dataset are 1,048,576 and $3.72 \times 10^{-9}$, respectively.

In first set of experiments, contributions of three LBP histograms on three planes (three components of LBPTOP) are investigated separately to determine accuracy of face recognition. Figure 8 presents recognition rates of classifying six basic expressions using LBP histograms from three individual planes (i.e., LBP-XY, LBP-XT, and LBP-YT) and their combination (i.e., LBPTOP). The figure indicates that using LBPTOP yields the highest recognition rate (94.5%); LBP-YT describes better variation in texture along vertical direction (93.5%), and lowest performance is attributed to LBP-XY containing appearance characteristics of different subjects (82.2%). Then we investigate the performance of spatial-temporal motion LBP feature which joins the LBP histograms on XT and YT planes together and compare with LBPTOP. Based on results shown in Figure 9, recognition rates are almost similar. Furthermore, spatial-temporal motion LBP feature has higher recognition rate than LBPTOP in terms of anger and disgust. Results shown in Figures 8 and 9 mean the following: (1) variation of texture on XT and YT plane is more significant than XY plane in expression sequence; (2) compared with LBPTOP, dimension of spatial-temporal motion LBP feature is reduced by 1/3, but recognition accuracy is not decreased.

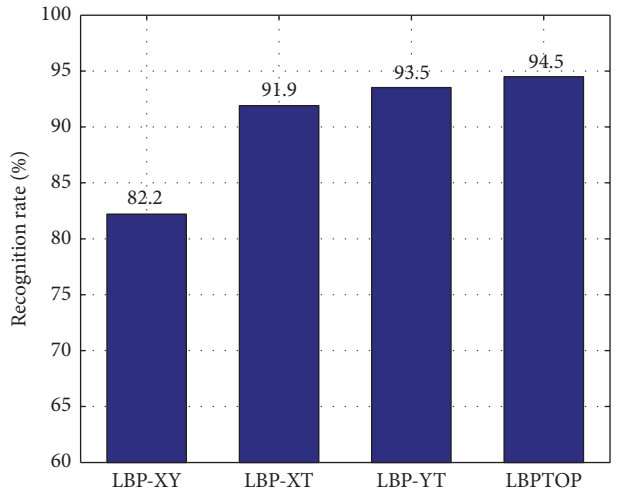


Figure 8: Recognition rates of methods using LBP histograms from three planes and LBPTOP.

Table 4: Classification results of using fusion features on CK+ dataset.

| Expression | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Average |
|---|---|---|---|---|---|---|---|
| Anger | 42 | **0** | **0** | **0** | **3** | **0** | **93.3** |
| Disgust | **0** | 59 | **0** | **0** | **0** | **0** | **100** |
| Fear | **0** | **0** | 22 | **0** | **2** | **1** | **88.0** |
| Happiness | **0** | **0** | **0** | 69 | **0** | **0** | **100** |
| Sadness | **3** | **0** | **1** | **0** | 24 | **0** | **85.7** |
| Surprise | **0** | **0** | **2** | **1** | **0** | 80 | **96.4** |
| Overall | | | | | | | **95.8** |

Table 5: Comparison results of using fusion feature and LBPTOP on CK+ dataset.

| Expression | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Average |
|---|---|---|---|---|---|---|---|
| Anger | 93.3 (+2.2) | **0** | **0** | **0** | **6.7** | **0** | |
| Disgust | **0** | 100 (0.0) | **0** | **0** | **0** | **0** | |
| Fear | **0** | **0** | 88.0 (0.0) | **0** | **8** | **4** | |
| Happiness | **0** | **0** | **0** | 100 (0.0) | **0** | **0** | |
| Sadness | **10.7** | **0** | **3.6** | **0** | 86.7 (+8.1) | **0** | |
| Surprise | **0** | **0** | **2.4** | **1.2** | **0** | 96.4 (0.0) | |
| Overall | | | | | | | 95.8 (+1.3) |

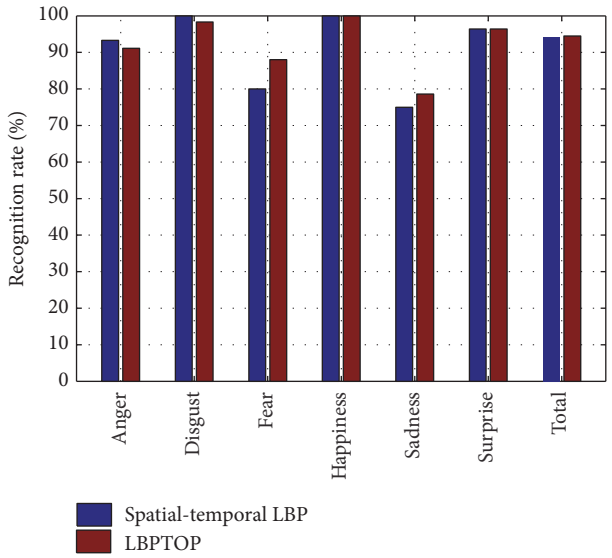


Figure 9: Recognition rates of methods using LBPTOP and spatial-temporal motion LBP.

Second set of experiments evaluates effectiveness of combining spatial-temporal motion LBP with Gabor multiorientation fusion histogram. Tables 2–4 show results obtained by using spatial-temporal motion LBP, Gabor multiorientation fusion histogram, and proposed framework with multiclass SVM classifier. Results include recognition rate of each expression and overall classifications accuracy. Each black digital represents number of correctly classified subjects of each expression category, and each dark and bolder digital image represents number of misclassified subjects of each expression category. Results show that when using combination of spatial-temporal motion LBP and Gabor multiorientation fusion histogram framework performs better than using individual features.

Third set of experiments compares performance between proposed methods and approach using LBPTOP [22]. Results in Table 5 show recognition rate of method using combination of spatial-temporal motion LBP with Gabor multiorientation fusion histogram (i.e., proposed framework), where figures in parentheses represent difference in accuracy in using LBPTOP and features proposed in this paper; positive figure shows better accuracy of proposed features. Based on Table 5, proposed framework performs better in two expressions and same in others. Furthermore, overall recognition rate is slightly better when using proposed method. Therefore, method fusing spatial-temporal motion LBP with Gabor multiorientation fusion histogram outperforms LBPTOP.

In this set experiment, we compared our method with systems proposed by Eskil and Benli [13], Lucey et al. [14], Chew et al. [15], Fan and Tjahjadi [16], and Jain et al. [17] on CK+ dataset. The datasets used in these papers are the same and identical to the dataset adopted in our study. Therefore, results presented in these papers can be directly compared with our method. Methods proposed by Eskil and Benli [13], Lucey et al. [14], Chew et al. [15], and Fan and Tjahjadi [16] used leave-one-subject-out cross validation method, and technology proposed by Jain et al. [17] used tenfold cross validation. Based on results shown in Table 6, proposed framework shows average accuracy of 95.8% for all six basic facial expressions, performing better than methods developed by mentioned authors in this paragraph.

*4.3. Experimental Results on MMI Dataset.* Seven methods in [18–20] are used to perform quantitative comparisons with

Table 6: Comparison of results in proposed framework and seven state-of-the-art methods on CK+ dataset.

| Research | Methodology | Accuracy |
|---|---|---|
| Eskil and Benli [13] | High-polygon wireframe mode | 85.0 |
| Lucey et al. [14] | AAM shape | 68.9 |
| | AAM appearance | 84.5 |
| | Combined | 88.7 |
| Chew et al. [15] | Uni-hyperplane classification | 89.4 |
| Fan and Tjahjadi [16] | PHOGTOP and optical flow | 90.9 |
| Jain et al. [17] | Temporal modeling of shapes | 91.9 |
| Proposed | | 95.8 |

Table 7: Comparison result of proposed framework and seven systems on MMI dataset.

| Research | Methodology | Accuracy (%) |
|---|---|---|
| Shan et al. [18] | LBP | 47.78 |
| Wang et al. [19] | AdaBoost | 47.8 |
| | HMM | 51.5 |
| | ITBN | 59.7 |
| Fang et al. [20] | Active shape model | 62.38 |
| | Active appearance model | 64.35 |
| | Gabor + Geometry + SVM | 70.67 |
| Proposed | | 71.92 |

our framework on MMI dataset. The punishment coefficient $C$ and kernel parameters $\gamma$ of SVM for MMI dataset are 2,097,152 and $9.31 \times 10^{-10}$, respectively. Datasets used in these papers are identical to the dataset adapted in our study. In this experiment, proposed framework performs tenfold cross validation. Shan et al. [18] and Fang et al. [20] developed methods by conducting experiments using the same strategy we employed. Methods proposed by Wang et al. [19] performed twentyfold cross validation, and methods using active shape model [20] and active appearance model [20] performed twofold cross validation; results are shown in Table 7. Table 7 indicates that performance of proposed method on MMI dataset is inferior to CK+ dataset because of subtler expressions, notable changes in head rotation angle, and fewer training data. However, proposed framework still performs better than the other seven state-of-the-art methods.

*4.4. Experimental Results on Oulu-CASIA VIS Dataset.* In this section, we evaluated our framework on Oulu-CASIA VIS dataset. In these set experiments, proposed framework performs tenfold cross validation. The punishment coefficient $C$ and kernel parameters $\gamma$ of SVM for this dataset are 1,048,507 and $9.33 \times 10^{-10}$, respectively. First, our method is tested on three subsets of Oulu-CASIA VIS, respectively. The experimental results are shown in Figure 10. As shown in Figure 10, the recognition rate obtained for the normal illumination condition is highest, and the recognition rate obtained for the dark illumination condition is worst. Evidently, the facial images captured in the weak or dark illumination are often missing much useful texture information

Table 8: Comparison result of proposed framework and four methods on Oulu-CASIA VIS dataset.

| Research | Methodology | Accuracy (%) |
|---|---|---|
| Scovanner et al. [21] | 3D SIFT | 55.83 |
| Zhao and Pietikäinen [22] | LBPTOP | 68.13 |
| Kläser et al. [23] | HOG3D | 70.63 |
| Zhao et al. [24] | AdaLBP | 73.54 |
| Proposed | | 74.37 |

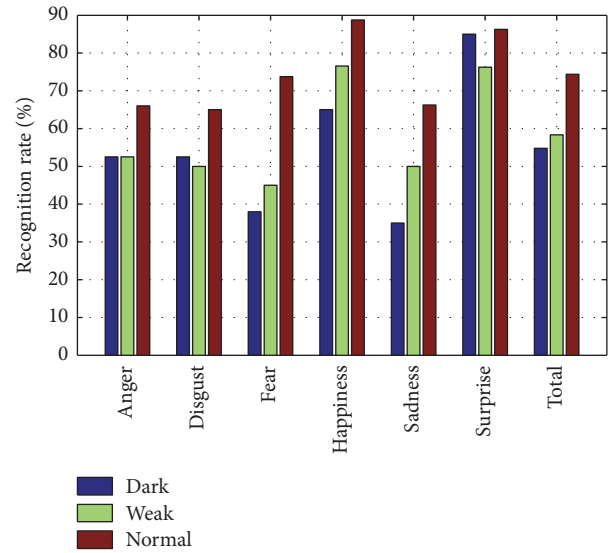


Figure 10: Recognition results of proposed methods on three subsets of Oulu-CASIA VIS dataset.

for facial expression recognition. Then, four methods in [21–24] are used to perform quantitative comparisons with our framework on the subset obtained in normal illumination conditions. The experimental results are shown in Table 8. Table 8 indicates that performance of proposed method on Oulu-CASIA VIS dataset is inferior to CK+ dataset but superior to MMI dataset. Proposed framework still outperforms the other four state-of-the-art methods.

*4.5. Computational Complexity Evaluation.* In this section, computational complexity of our method is analyzed in CK+, MMI, and Oulu-CASIA VIS dataset. Processing time is approximated to be time needed for preprocessing, extracting spatial-temporal motion LBP and Gabor multiorientation fusion histogram, and classifying each image frame of video sequence. Runtime (measured using computer system clock) is estimated by using mixed-language programming based on MATLAB and C++ on an Intel (R) Core (TM) i7-6700 CPU at 3.40 GHz with 8 GB RAM running on Windows 10 operating system. Average processing time per image is under 280, 286, and 249 ms for CK+, MMI, and Oulu-CASIA VIS datasets, respectively. Then, we compare computational time of our method with system using LBPTOP [22]. Comparison results in Table 9 show that computational time of our method is lower than system using LBPTOP in three datasets. Results show that our method is more efficient and effective than

Table 9: Runtime of our method and LBPTOP in CK+, MMI, and Oulu-CASIA VIS datasets.

| Research | CK+ | MMI | Oulu |
|---|---|---|---|
| LBPTOP | 349 ms | 357 ms | 317 ms |
| proposed | 280 ms | 286 ms | 249 ms |

LBPTOP-based technology for facial expression recognition in CK+, MMI, and Oulu-CASIA VIS datasets.

## 5. Conclusion

This paper presents novel facial expression recognition framework integrating spatial-temporal motion LBP with Gabor multiorientation fusion histogram. Framework comprises preprocessing (face recognition and points location and tracking), dynamic and static feature extraction, and feature classification, outperforming seven state-of-the-art methods on CK+ dataset, seven other methods on MMI dataset, and four methods on Oulu-CASIA VIS dataset. Expressions of disgust and happiness are easier to classify than other expressions, demonstrating better performance of proposed method. However, recognition rates are lower for fear (88.0%) and sadness (86.7%) on CK+ dataset compared with others, because most anger expressions are misclassified as sadness and vice-versa. A lamination of proposed method is that the occlusions, head rotations, and illumination can reduce accuracy of expression recognition; these phenomena will be solved in our future work. A framework integrating other useful information (pupil, head, and body movements) with the features extracted in this paper could perform better. This issue will also be studied in our future work. The proposed method cannot achieve real-time facial expression recognition. We will attempt to establish an efficient dimensionality reduction method to reduce the computational complexity of our framework in the future. Furthermore, proposed system will be improved and applied in study of human drowsiness expression analysis in our further work.

## Competing Interests

The authors declare no conflict of interests.

## Acknowledgments

This work was supported by the Open Foundation of State Key Laboratory of Automotive Simulation and Control (China, Grant no. 20161105).

## References

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

[2] J. A. Russell, J. M. Fernandez-Dols, and G. Mandler, *The Psychology of Facial Expression*, Cambridge University Press, Cambridge, UK, 1997.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[4] B. Golomb and T. Sejnowski, "Benefits of machine understanding of facial expressions," in *NSF Report—Facial Expression Understanding*, pp. 55–71, 1997.

[5] M. Pantic, "Face for ambient interface," in *Ambient Intelligence in Everyday Life*, vol. 3864 of *Lecture Notes on Artificial Intelligence*, pp. 32–66, Springer, Berlin, Germany, 2006.

[6] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[7] H. Fang and N. Costen, "From Rank-N to Rank-1 face recognition based on motion similarity," in *Proceedings of the 20th British Machine Vision Conference (BMVC '09)*, pp. 1–11, London, UK, September 2009.

[8] C. Cornelis, M. D. Cock, and A. M. Radzikowska, "Vaguely quantified rough sets," in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing: 11th International Conference, RSFDGrC 2007, Toronto, Canada, May 14–16, 2007. Proceedings*, vol. 4482 of *Lecture Notes in Computer Science*, pp. 87–94, Springer, Berlin, Germany, 2007.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] M. Pantic and L. Ü. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[11] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[12] C. P. Sumathi, T. Santhanam, and M. Mahadevi, "Automatic facial expression analysis: a survey," *International Journal of Computer Science & Engineering Survey*, vol. 3, no. 6, pp. 47–59, 2012.

[13] M. T. Eskil and K. S. Benli, "Facial expression recognition based on anatomy," *Computer Vision and Image Understanding*, vol. 119, pp. 1–14, 2014.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPRW '10)*, pp. 94–101, San Francisco, Calif, USA, June 2010.

[15] S. W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. F. Conn, "Improved facial expression recognition via uni-hyperplane classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2554–2561, IEEE, June 2012.

[16] X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognition*, vol. 48, no. 11, pp. 3407–3416, 2015.

[17] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops '11)*, pp. 1642–1649, November 2011.

[18] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A Comprehensive Study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[19] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3422–3429, IEEE, Portland, Ore, USA, June 2013.

[20] H. Fang, N. Mac Parthaláin, A. J. Aubrey et al., "Facial expression recognition in dynamic sequences: an integrated approach," *Pattern Recognition*, vol. 47, no. 3, pp. 1271–1281, 2014.

[21] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia (MM '07)*, pp. 357–360, ACM, Augsburg, Germany, September 2007.

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

[23] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proceedings of the 19th British Machine Vision Conference (BMVC '08)*, pp. 275–281, Leeds, UK, September 2008.

[24] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[25] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*, A Human Face, Salt Lake City, Utah, USA, 2002.

[26] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 1, pp. 28–43, 2012.

[27] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 699–714, 2005.

[28] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 2, pp. 433–449, 2006.

[29] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.

[30] Y. Zhang, Q. Ji, Z. Zhu, and B. Yi, "Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 10, pp. 1383–1396, 2008.

[31] G. Donate, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[32] Y. Li, S. M. Mavadati, M. H. Mahoor, Y. Zhao, and Q. Ji, "Measuring the intensity of spontaneous facial action units with dynamic Bayesian network," *Pattern Recognition*, vol. 48, no. 11, pp. 3417–3427, 2015.

[33] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131–146, 2000.

[34] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden Markov models for modeling facial action temporal dynamics," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4796, pp. 118–127, 2007.

[35] T. Danisman, I. M. Bilasco, J. Martinet, and C. Djeraba, "Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron," *Signal Processing*, vol. 93, no. 6, pp. 1547–1556, 2013.

[36] W.-L. Chao, J.-J. Ding, and J.-Z. Liu, "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection," *Signal Processing*, vol. 117, pp. 1–10, 2015.

[37] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[38] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[39] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 1859–1866, IEEE, Columbus, Ohio, USA, June 2014.

[40] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[41] S.-S. Liu, Y.-T. Tian, and C. Wan, "Facial expression recognition method based on Gabor multi-orientation features fusion and block histogram," *Zidonghua Xuebao/Acta Automatica Sinica*, vol. 37, no. 12, pp. 1455–1463, 2011.

[42] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[43] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG '00)*, pp. 46–53, Grenoble, France, March 2000.

[44] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 317–321, Amsterdam, Netherlands, July 2005.