Improved Spatiotemporal Local Monogenic Binary Pattern for Emotion Recognition in The Wild

Xiaohua Huang  
Center for Machine Vision Research  
Department of Computer Science and Engineering  
University of Oulu, Finland  
huang.xiaohua@ee.oulu.fi

Qiuhai He  
Center for Machine Vision Research  
Department of Computer Science and Engineering  
University of Oulu, Finland  
qiuhai.he@ee.oulu.fi

Xiaopeng Hong  
Center for Machine Vision Research  
Department of Computer Science and Engineering  
University of Oulu, Finland  
xhong@ee.oulu.fi

Guoying Zhao  
Center for Machine Vision Research  
Department of Computer Science and Engineering  
University of Oulu, Finland  
gyzhao@ee.oulu.fi

Matti Pietikäinen  
Center for Machine Vision Research  
Department of Computer Science and Engineering  
University of Oulu, Finland  
mkp@ee.oulu.fi

ABSTRACT

Local binary pattern from three orthogonal planes (LBP-TOP) has been widely used in emotion recognition in the wild. However, it suffers from illumination and pose changes. This paper mainly focuses on the robustness of LBP-TOP to unconstrained environment. Recent proposed method, spatiotemporal local monogenic binary pattern (STL MBP) [14], was verified to work promisingly in different illumination conditions. Thus this paper proposes an improved spatiotemporal feature descriptor based on STLMBP. The improved descriptor uses not only magnitude and orientation, but also the phase information, which provide complementary information. In detail, the magnitude, orientation and phase images are obtained by using an effective monogenic filter, and multiple feature vectors are finally fused by multiple kernel learning. STLMBP and the proposed method are evaluated in the Acted Facial Expression in the Wild as part of the 2014 Emotion Recognition in the Wild Challenge. They achieve competitive results, with an accuracy gain of 6.35% and 7.65% above the challenge baseline (LBP-TOP) over video.

Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications; I.5.4 [Pattern Recognition]: Applications

Keywords

Spatiotemporal feature; Monogenic signal analysis; Multiple Kernel Learning

1. INTRODUCTION

Facial expression recognition has become a valuable measurement for determining emotional state of a human being based on the facial images. The most common target is to recognize seven universal facial expressions, i.e. happiness, sadness, surprise, neutral, anger, fear and disgust, which are proposed by [9]. Early stage research mostly focused on the posed facial expression databases such as Cohn-Kanade (CK) database [19] and MMI database [25]. Unfortunately, these databases are collected under laboratory controlled environments. Such lab controlled data poorly represents the environment and conditions faced in the world situation. Though some attempts have been made to investigate the influence of sole environmental condition such as different illuminations or occlusions [13, 16], there is still no use facial expression recognition system in the wild, which contains numerous challenging environmental conditions. It is thus valuable to pursue the recognition under realistic image acquisition conditions.

The first Emotion Recognition in the Wild (EmotiW 2013) challenge [6] has recently provided a platform to create and verify their facial expression recognition systems. This challenge was based on an extended form of the Acted Facial Expression in the Wild (AFEW) database [7], in which short video clips have been annotated for different emotions. In this challenge, local binary pattern from three orthogonal planes [32] (LBP-TOP), which achieved promising performance in texture classification and posed facial expression, was implemented as the baseline. As far as we know, there are numerous facial expression recognition systems achieving more promising results than the baseline on this database [15, 18, 23]. Among them, the recent popular method, deep neural network [15], was the first time to be used in emotion recognition and won this competition. In their work, they...
combined multiple deep neural networks for video and audio features. On the other hand, some machine learning methods, e.g. multiple kernel learning (MKL) [23] and partial least square regression (PLSR) [18], were employed to distinguish different facial expressions. Some attempts were made to use other appearance-based features, e.g. Local Gabor binary pattern from three orthogonal planes (LGBP-TOP) [2], optical flow and Gabor [17]. In addition, the geometric features were only used in [17]. However, it is observed that the combination of optical flow, Gabor and geometric features work poorer than the baseline. One reason may be that optical flow and geometric features are sensitive to pose change and mis-alignment errors.

Now the second Emotion Recognition in the wild (EmotiW 2014) provides additional videos (567 videos samples for training, 371 for validation and 407 for test) compared with EmotiW 2013 (380, 396 and 312 for training, validation and test sets, respectively). The illumination variations and pose change are two critical problems to facial expression recognition. It is necessary to find out a robust feature representation for resisting them. In general, feature representation methods are categorized into appearance-based features and geometric-based ones. Certain studies [10, 28] demonstrate that appearance-based features are more robust to illumination variation and mis-alignment error than geometric features. In fact, the face detection, landmark localization and alignment procedure may fail due to the challenging conditions. It causes geometric feature not stable. Recalled from their work in EmotiW 2013, it is observed that appearance-based features are widely applied. The main reason is that they are robust to illumination variations and pose change.

Additionally, it is known that LBP-TOP [32] obtained promising results on facial expression recognition under controlled environment, while in the wild it poorly worked. One main reason is that the pixel-level information contains too much noise caused by illumination variation or pose change. There is still some space to increase the robustness of LBP-TOP in the wild. Thus we propose a novel facial expression recognition system based on face preprocessing procedure and appearance-based features, which is shown in Figure 1. In face preprocessing procedure, landmark detection is employed based on [33] and IntraFace. Each frame is aligned to the first frame. In facial feature extraction, we propose a two-layer architecture to describe the video sequences, which is based on spatiotemporal local monogenic binary pattern (STLMBP) [14]. In the first layer, an effective monogenic filter is presented to capture the lower-level features including magnitude, orientation and phase information. In the second layer, these lower-level features are further encoded by the respective methods, e.g., the difference of sign (DoS) [21] or phase-quadrant encoding method (PQDC) [4]. The MKL method [22] is used to fuse multiple feature vector as output of the second layer. The main contributions of this work are to improve the method [14] by introducing the phase information and use the proposed method for the complex AFEW database, which are shown to promisingly outperform the challenge baseline results in EmotiW 2014.

The paper is organized as follows. Section 2 gives a brief overview of existing related work, followed by Section 3 that describes the proposed method in detail. Next, Section 4 summarizes the results, after which a conclusion is presented in Section 5.

2. RELATED WORK

This paper mainly focused on facial features for facial expressions. Various approaches have been proposed for facial feature representations [10, 28]. Two representative ones are local binary pattern (LBP) [21] and Gabor [20][30]. It is found that spatiotemporal LBP [32] and Gabor [2] have been widely used in EmotiW 2013, especially that LBP is also used as baseline algorithm in EmotiW 2014.

Some researchers have recently attempted to extract robust features on the lower-level still images, e.g., Gabor images, rather than on the pixel-level [26, 27, 29, 30, 31]. In [31], the Gabor magnitude and LBP operator were used as the first and second layers, respectively. The first layer enhanced facial features, while the second layer used local structural information and histogram of sub-regions to describe the facial features. Some temporal extension of Gabor-based LBP are established in [1], which combined motion and appearance features extraction with a prior Gabor wavelet filtering. However, in [27], they suggested that the Gabor filter suffers from three limitations: (1) it is not optimal if the broad spectral information with maximal spatial localization need to be sought, (2) its maximum bandwidth is restricted to about one octave, and (3) the computational cost and the storage space are expensive due to many scales and orientations. An alternative method, called monogenic signal analysis [11], can avoid these limitations, because it is itself a compact representation of features with little information loss and does not use steerable filters to create multi-orientation features. The advantages of monogenic filters motivated some researchers to employ it as the first layer [27, 29]. They demonstrated that the monogenic binary pattern outperformed local Gabor binary pattern in face recognition.

Recent studies show that orientation information not only reduces the effect of specularity but also increases the stability to illumination variation [24]. The orientation has been developed as complement magnitude in monogenic binary pattern [27, 29]. Their works have been extended to temporal domain by [14]. In [14], they decomposed the orientation information into real and imaginary parts. They also demonstrated that orientation worked well in different illumination conditions. However, their works missed another important information, i.e., phase information, which can...
be output by monogenic filter. Further improvement to this work can be found in [26, 30]. They showed that the use of phase can provide complementary information for Gabor magnitude.

3. PROPOSED METHODOLOGY

3.1 Monogenic representation

The monogenic signal [11] generalizes the analytic signal to 2-D by the introduction of a Riesz filter $f(x,y)$, in which Fourier domain representation is $[F_1(u,v), F_2(u,v)]$. It can capture the phase, orientation and magnitude of an image in a rotation invariant way. The spatial visualization of Riesz transform is illustrated in Figure 3.

Given an image $I(x,y)$, the monogenic signal is represented as the convolution of 2-D signal with two Riesz filtered components $f(x,y) = g_x(\omega) * I(x,y) + [h_x(x,y), h_2(x,y)]$, where "∗" is convolution operator, $h_x(x,y)$ is the spatial domain representation of $F_1(u,v)$, and $g_x(\omega)$ is a Log-Gabor function on the linear frequency scale $s$. Here the Log-Gabor function has a transfer function of the form [12] as follows:

$$g_x(\omega) = F^{-1}(G(\omega)) = F^{-1}(\exp(-\log(\omega/\omega_c)^2)/2(\log(k/\omega_c)^2)),$$

where $\omega_c \propto s$ is the filter’s center frequency and $s$ is number of wavelet scale of which a lower value will reveal more fine scale features while a larger value will highlight coarse features.

Therefore, the monogenic magnitude, phase and orientation of the signal can be formulated by:

$$A = \sqrt{h^2 + h_x^2 + h_y^2},$$

$$\phi = \arctan(\sqrt{h_x^2 + h_y^2}/h),$$

$$\theta = \arctan(h_x/h_y),$$

where $h = g_x(\cdot) * I(\cdot)$, $h_x = h(\cdot) * h_1(\cdot)$, $h_y = h(\cdot) * h_2(\cdot)$, and the value ranges of phase and orientation are in $[0 2\pi]$ and $[0 \pi]$, respectively.

3.2 Multiscale extension of monogenic

Since the log-Gabor filters are band-pass filters, multi-scale monogenic representation is thus required to fully describe a signal. The direct concatenation of multi-scale representation for the sequences may cause high dimensionality and expensive computational time. Alternative, the sum of multi-scale representation can solve these limitations. Monogenic magnitude, phase and orientation (see in Eqs. 2, 3 and 4) can be re-written as:

$$\tilde{A} = \sqrt{\tilde{h}^2 + \tilde{h}_x^2 + \tilde{h}_y^2},$$

$$\tilde{\phi} = \arctan(\sqrt{\tilde{h}_x^2 + \tilde{h}_y^2}/\tilde{h}),$$

$$\tilde{\theta} = \arctan(\tilde{h}_x/\tilde{h}_y),$$

where $\tilde{h} = \sum_{s=1}^{S} g_s(\cdot) * f(\cdot)$, $\tilde{h}_x = \sum_{s=1}^{S} g_s(\cdot) * f(\cdot) * h_1(\cdot)$, $\tilde{h}_y = \sum_{s=1}^{S} g_s(\cdot) * f(\cdot) * h_2(\cdot)$, and $S$ is the number of scale in multi-scale log-Gabor filters.

3.3 Improved spatiotemporal local monogenic binary pattern

It is noted that each facial image can be represented by magnitude, phase and orientation of its monogenic signal. The magnitude and orientation have been utilized in [14], while the phase information is not still used. It is found that the phase information is useful to face recognition [26]. Motivated by their work, we attempt to exploit the information of phase for improving STLMBP. Next we will describe the formulation of spatiotemporal features.

**Magnitude:** In the monogenic representation of the image, the magnitude measures local structure energy, and the LBP operator can then be directly used to encode the variation of local energy. The histogram descriptor can thus be formulated as:

$$H_{mag}^{P,R} = \sum_{i=0}^{P-1} L((\tilde{A}_{x,p,y,p} - \tilde{A}_{x,c,y,c}) \geq 0) 2^p,$$

where, $L(\tilde{A}_{x,p,y,p} - \tilde{A}_{x,c,y,c}) = \begin{cases} 1, & \tilde{A}_{x,p,y,p} - \tilde{A}_{x,c,y,c} \geq 0 \\ 0, & \tilde{A}_{x,p,y,p} - \tilde{A}_{x,c,y,c} < 0 \end{cases}$, and $\tilde{A}_{x,c,y,c}$ is the monogenic magnitude value at the position $(x_c,y_c)$. $\tilde{A}_{x,p,y,p}$ is the magnitude value of $P$ equally spaced pixels on a circle of radius $R$ at this position.

**Orientation:** The orientation information $\tilde{\theta}$ is decomposed into two subcomponents including real and imaginary pictures. The encoding measurement for real and imaginary parts can be defined as follows:

$$B(\tilde{h}_i) = \begin{cases} 1, & \tilde{h}_i > 0 \\ 0, & \tilde{h}_i \leq 0 \end{cases},$$

where, $i = x$ for real picture, and $i = y$ for imaginary picture. For a brief description, $\tilde{h}_i$ is denoted as $\tilde{h}_i$.

Then an alternative binary-compare function, local bit exclusive or operator (LXP), is exploited to calculate the correlation of the center and its neighbors. The histogram of real/imaginary can be formulated as follows:

$$H_{ori}^{P,R} = \sum_{p=0}^{P-1} (B(\tilde{f}_{x,p,y,p} \oplus B(\tilde{f}_{x,c,y,c})) 2^p,$$

where $B(\tilde{f}_{x,p,y,p})$ is the PQDC value of $P$ equally spaced pixels on a circle of radius $R$ at the position $(x_c,y_c)$, and $\oplus$ denotes the bit exclusive or operator. Furthermore, the uniform pattern in LBP, which contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string

Figure 3: Spatial domain representation of (a) $F_1(u,v)$ and (b) $F_2(u,v)$ of Riesz transform.
is considered circular, is used to preserve a simple rotation-invariant property and reduce the length of the feature vector.

**Phase:** Given the phase angle $\widetilde{\phi}(x, y)$ and its angles of neighbors $\widetilde{\phi}(x_p, y_p)$, it needs to exploit the operations of two phase angles according to the formulation of LBP. However, it is not reasonable to compare the difference of phase angles and use the unit function. Furthermore, the threshold method for the difference of phase angles is complicated and unstable since the threshold is empirically set. Instead, we rather quantify the phase angle by applying the quantification function:

$$q = \text{mod}\left(\left\lfloor \frac{\widetilde{\phi}(x, y)}{\varphi} + 0.5 \right\rfloor, \varphi \right),$$  \hspace{1cm} (11)

where \text{mod} is the modulo operation, and $\varphi$ is the number of dominant orientation.

According to Eqn. 11, it is easy to get the dominant phase bins $q_{x,y}$ and $q_{x_p,y_p}$ of $(x, y)$ and $(x_p, y_p)$, respectively. Then their operation is calculated as follows:

$$Q_p = \begin{cases} 0, & q_{x,y} = q_{x_p,y_p} \\ 1, & q_{x,y} \neq q_{x_p,y_p}. \end{cases}$$  \hspace{1cm} (12)

The histogram of phase can be formulated as follows:

$$H_{\text{phase}}^{P,R} = \sum_{p=0}^{P-1} Q_p 2^p.$$  \hspace{1cm} (13)

**Extension to temporal domain:** It is known that a video sequence can provide much more information than the static image. This induces another problem how to extend these methods to video sequences. Here we take the procedure producing the histogram of magnitude as an example. The procedure of all informations is illustrated in Figure 2. As shown in Figure 2, the magnitude images $\{A_n\}_{n=1}^N$ extracted from $N$ frames are orderly put into one cuboid. Furthermore, the features are extracted from three orthogonal planes (XY, XT and YT planes) for describing appearance and motions. For histogram of each plane, the uniform pattern is employed. Finally all histograms from three planes are concatenated into one feature vector. For other information, the similar procedure is implemented. One difference is the corresponding operators should be used.

In the implementation, the facial structure information is obtained by block-based method, which divided the facial image into several blocks. For each facial image, we divided all facial images into non-overlapping spatial $8 \times 8$ blocks.

### 3.4 Feature fusion and classification

Given multiple feature representation for each video, it is hoped to combine them in a way that increases the discriminative power of the classifier. Two common methods for fusing multiple feature representations are feature-level, where a single classifier is trained using all features as input, and decision-level, where a classifier is trained for each feature separately and a decision rule combines the classifier outputs. We followed a feature fusion approach based on Multiple Kernel Learning (MKL) [8]. In our implementation, we used the UFO-MKL method as given in [22].

### 4. EXPERIMENTAL ANALYSIS

#### 4.1 Data description and preprocess

The database provided by EmotiW 2014 [7] consists of short video clips extracted from popular movies. Each clip contains an actor expressing one of seven basic facial expressions: neutral (NE), happiness (HA), sadness (SA), disgust (DI), fear (FE), anger (AN), and surprise (SUR). The data is divided into three sets: \textit{Train}, \textit{Val} and \textit{Test}, respectively containing 567, 371 and 407 video clips. \textit{Train} and \textit{Val} sets with facial expression labels are available for researchers, which are used to adjust the parameter and verify the facial expression model, while \textit{Test} set contains unseen video clips.
Figure 4: The recognition rate on validation set with different scale.

Figure 5: The recognition rate on validation set with different number of dominant phase.

clips. The goal of this dataset is to address the challenges in recognizing emotions in real-world conditions.

We simply use the aligned face images provided by EmotiW 2014 organizers. In their database, the face is localized using mixture of parts framework of [33] and tracking is performed using IntraFace Library. The fiducial points generated by IntraFace are used for aligning the face. The size of all face images is normalized into $128 \times 128$ pixels.

4.2 Experiment parameter

The proposed method has two important parameters, i.e. the number of scale and the number of dominant phase in Eqn. 11. We attempt to find out the best parameters for them on Val set.

Firstly, to evaluate the influence of scale in Section 3.2, we use STLMBP for obtaining the best parameter, since we expect less influence of another parameter. We tune the parameter $s$ in range $[1, 6]$ and obtain difference results as shown in Figure 4. It is found that too many scales are not useful on predicting Val set. For Test set, we select the best parameter on validation set, i.e. $s = 3$ in Eqns. 5, 6 and 7.

Secondly, we only consider the phase information with three scales. Lib-SVM with Gaussian kernel [3] is used to classify the emotion on Val set. The number of dominant phase $\varphi$ is tested in $2, 4, 8$ and $16$. We obtain the comparative results with different number as shown in Figure 5. It is found that when the number of dominant phase is $4$, the feature descriptor achieves best. For Test set, we select $\varphi = 4$ for phase information in Eqn 11.

4.3 Result comparison

We demonstrate all of our results on the Val and Test sets, where the test results are evaluation feedbacks from EmotiW 2014 organizers) in Table 1. Firstly, we compare the performance of spatiotemporal features separately based on magnitude, real part, imaginary part and phase on Val set. Here we name them as STMBP, STRBP, STIBP and STPBP, respectively. It is found that all features except magnitude achieve better performance than the baseline on Val set (which adopted LBP-TOP features [32] for video and kernel SVM for classification) [5]. Furthermore, we also show the performance of STLMBP and improved STLMBP on Val and Test sets. STLMBP outperforms the baseline at the recognition rate of 6.35%. Improved STLMBP performs better than the STLMBP and baseline on both Val and Test sets.

The confusion matrix of final test results are shown in Table 2 and 3. We can see that “Happiness” and “Anger” are easy to be distinguished from other expressions, but it is still hard to do well on some difficult emotion classes such as “Disgust”. One main reason is that the motion from “Happiness” and “Anger” is more obvious than other classes.
Table 3: Confusion matrix (%) on test data by using improved STLMBP

<table>
<thead>
<tr>
<th>Expression</th>
<th>AN</th>
<th>DI</th>
<th>FE</th>
<th>HA</th>
<th>NE</th>
<th>SA</th>
<th>SUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN</td>
<td>63.79</td>
<td>6.90</td>
<td>6.90</td>
<td>5.17</td>
<td>5.17</td>
<td>6.90</td>
<td>5.17</td>
</tr>
<tr>
<td>DI</td>
<td>11.54</td>
<td>15.38</td>
<td>7.69</td>
<td>30.77</td>
<td>7.69</td>
<td>19.23</td>
<td>7.69</td>
</tr>
<tr>
<td>FE</td>
<td>30.43</td>
<td>0</td>
<td>17.39</td>
<td>8.70</td>
<td>23.91</td>
<td>10.87</td>
<td>8.70</td>
</tr>
<tr>
<td>HA</td>
<td>7.41</td>
<td>0</td>
<td>3.70</td>
<td>70.37</td>
<td>6.17</td>
<td>11.11</td>
<td>1.23</td>
</tr>
<tr>
<td>NE</td>
<td>10.26</td>
<td>4.27</td>
<td>5.13</td>
<td>16.24</td>
<td>38.46</td>
<td>15.38</td>
<td>10.26</td>
</tr>
<tr>
<td>SA</td>
<td>15.09</td>
<td>1.89</td>
<td>3.77</td>
<td>26.42</td>
<td>18.87</td>
<td>16.98</td>
<td>16.98</td>
</tr>
<tr>
<td>SUR</td>
<td>7.69</td>
<td>3.85</td>
<td>11.54</td>
<td>15.38</td>
<td>15.38</td>
<td>11.54</td>
<td>34.62</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper, we propose a new spatiotemporal feature descriptor for video-based emotion recognition in the real-world environment. For each video clip, magnitude, orientation and phase information are contained by using monogenic filter, and then they are encoded to represent the appearance-based features. In classification, multiple kernel learning is applied to fuse multiple feature vector and classify the video clip. The method is evaluated on EmotiW 2014 and shows promising results on unseen test data. In the future, we will try to explore the discriminative information from multiple feature vector to further improve the recognition performance.

6. ACKNOWLEDGMENTS

This work was sponsored by the Academy of Finland, Infotech Oulu and Nokia Foundation.

7. REFERENCES


