Learning Multiscale Active Facial Patches for Expression Analysis

Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N. Metaxas, Senior Member, IEEE

Abstract—In this paper, we present a new idea to analyze facial expression by exploring some common and specific information among different expressions. Inspired by the observation that only a few facial parts are active in expression disclosure (e.g., around mouth, eye), we try to discover the common and specific patches which are important to discriminate all the expressions and only a particular expression, respectively. A two-stage multitask sparse learning (MTSL) framework is proposed to efficiently locate those discriminative patches. In the first stage MTSL, expression recognition tasks are combined to located common patches. Each of the tasks aims to find dominant patches for each expression. Secondly, two related tasks, facial expression recognition and face verification tasks, are coupled to learn specific facial patches for individual expression. The two-stage patch learning is performed on patches sampled by multiscale strategy. Extensive experiments validate the existence and significance of common and specific patches. Utilizing these learned patches, we achieve superior performances on expression recognition compared to the state-of-the-arts.

Index Terms—Cohn–Kanade, facial expression, GEMEP-FERA, group sparsity, MMI, multiscale facial patch selection, multitask learning.

I. INTRODUCTION

Facial expressions play significant roles in our daily communication, due to their abilities to reflect human emotions, and social interaction. In the past three decades, automatic facial expressions recognition has become an increasingly fascinating topic in the computer vision and pattern recognition communities for their extensive applications, such as human-computer interface, multimedia, and security [38], [54], [56]. However, as the basis of expression recognition, the exploration of the functional facial features is still an open problem.

Studies in psychology show that facial features of expressions are located around mouth, nose, and eyes, and their locations are essential for explaining and categorizing facial expressions. Through electrical muscle stimulation, References [1] and [18] found that most expressions are invoked by a small number of facial muscles around the mouth, nose and eyes [see Fig. 1(a)]. This indicates that most of the descriptive regions for each expression are located around certain face parts. Moreover, expressions can be generally categorized into six popular “basic” or “universal” expressions [26]: anger, disgust, fear, happiness, sadness, and surprise. (Of course, there are a lot of complex expressions which are not basic expressions). These expressions seems to be universal across different ethnicities and cultures [21].

Previous expression recognition methods can be generally categorized into two groups: AU-based methods and message and sign judgement methods. AU-based methods [20], [46], [47], [49], [51] recognize expressions by detecting AUs, which have more descriptive power, but these methods suffer from the difficulties of AU detection. Message and sign judgement methods [31], [44], [58] reveal the differences among expressions by facial appearance variations, which has been proved to be more reliable on single still images. However, these methods treat different facial parts equally or assign different weights to them empirically, thus lacking statistical support for the weight settings. This motivates us to fully make use of the prior knowledge from facial muscles and AU studies to extract the most discriminative regions, which can further assist expression analysis.

Inspired by the locations of AUs, we divide human face into nonoverlapping patches on different scale levels, and then conceptually group these patches into three categories: common facial patches, specific facial patches, and the rest. Common facial patches are active ones for all expressions. Specific facial patches are only active for one particular expression. Therefore, the most important facial patches are the common ones shared by all expressions; specific patches are only a few and only useful to discriminate a particular expression; the rest of the patches are of less help to expression recognition. The effective facial patches corresponding to different facial muscles may have different sizes, and the optimal size of patches is hard to be determined and it may vary among different areas of the face, we employ three different scale patches in this paper to cover effective facial parts with different sizes. That is, each face image is divided into $8 \times 8$, $6 \times 6$, and $4 \times 4$ patches. The patches are smaller when the
A two-stage multistage sparse learning (MTSL) framework is proposed to explore common and specific patches statistically on each scale level, respectively. In the first stage, the binary classification problem for each expression is treated as an individual task (see Fig. 4), then a MTSL model is built based on these related tasks to extract the common facial patches. In the second stage, the face verification task (see Fig. 5) is designed to be coupled with the previous classification task for one expression. In this way, another MTSL model can be constructed to find out the specific patches for this particular expression. Similarly, the specific patches for all the expressions can be figured out separately. After the common and specific patches on different scales are learned by the two-stage MTSL model, they are combined to boost the facial expression recognition accuracy.

For all the scales, the common and specific patches, found by extensive experiments on the Cohn–Kanade database [28] and the MMI database [52], not only confirm the psychology discoveries of the facial muscles and AUs, but also provide more accurate appearance locations. Moreover, these common and specific patches at the same scale can be used to boost the performance of expression recognition. The learned patches are shown to be effective across different databases, e.g., Cohn–Kanade database [28], GEMEP-FERA [33]. Only using relatively small number of patches (∼1/3 of the face), our method still outperforms other methods in expression recognition. Finally, the learned effective patches at different scales can be also combined to further improve the expression recognition performance.

Our contributions are as follows.
1) We provide a solid validation for an important psychology discovery, that only partial area of the face (corresponding to underlying facial muscles) are discriminative for expression recognition.
2) A two-stage MTSL framework is proposed to formulate the commonalities among expressions, and to find out the locations of common and specific patches for expressions.
3) Multiscale image division strategy is utilized to generate patches of different size for facial expression analysis. More convincing conclusion about facial parts (muscles) could be achieved, since they are of different sizes.
4) Extensive experiments with three different scales on three public databases demonstrate that these active patches are effective in recognizing expressions. The common and specific patches can be combined to improve the performances of state-of-the-art. Patches across different scales can also been fused to further boost the performance.

The rest of this paper is organized as follows. Section II reviews related work of facial expression analysis and MTSL. Section III presents our framework to learn common and specific patches based on MTSL. These effective patches on all different scales are learned in this section. The experimental results on three public databases are shown in Section IV. We conclude this paper in Section V.

II. RELATED WORK

A. Facial Expression Analysis

Most facial expression analysis methods generally follow the aforementioned two categories: AU-based and message and sign judgement methods. Although this paper belongs to the latter category, it is still necessary to give a completed review of related works on expression analysis.

1) AU-based facial expression analysis inspired by the well-known study on facial activity, facial action coding system (FACS) [19]. In this system, the subtle changes in facial appearance are encoded into 32 AUs with individual linguistic description. Since each basic expressions can be decomposed into several related AUs, the expression recognition problem can be transferred to AUs detection problem. Bartlett et al. [6] recognized six single upper face AUs, but no simultaneous AUs are considered in combination. Tian et al. [47] detected 16 AUs from face image sequences using lip tracking, template matching and neural networks. More works have been done on spontaneous facial expression data by automatic recognition of AUs [7], [8], [15], [16], [27], [53]. Some differences between spontaneous and deliberate facial behavior are also studied by [17]. Recently, AU detection and AU-based expression recognition methods make a lot of significant progresses. Tong et al. [49] explore the dynamic and semantic relationships of facial AUs to improve their recognition performance. A dynamic Bayesian network is built by Tong et al. [48] for better facial activity understanding. Senechal et al. [42] combine different types of features using SimpleMKL learning algorithm to extract geometric and appearance information simultaneously. Sandbach et al. [39], [40] exploit 3-D motion-based features between frames of 3-D facial geometry sequences for dynamic AU detection and further expression recognition. AU-based methods decompose the facial expressions into different individual muscle activities, and then infer the expression categories based on the AU detection results. These methods can have great representation power, but AU detection itself is quite difficult and it is still an open problem to the community.
2) Message and sign judgement facial expression analysis methods generally consist of the two main steps: facial representation and expression recognition.

a) Facial representation derives a set of features from original facial images to effectively represent all faces. Different features have been applied to either the whole-face or specific face regions to extract the facial appearance changes, such as Gabor [7], [24], [31], haar-like features [58], local binary patterns (LBPs) [35], [43]. Zafeiriou and Pitas [64] explored the graph structures with landmarks to represent the variations among different expressions. In [44], facial images are equally divided into small regions, and then LBP features are extracted from these empirically weighted sub-regions to represent the facial appearance. The LBP features are shown to be effective in expression recognition, so this paper will also utilize the LBP features with the same sub-region division strategy. Different from their work, we will focus on learning the effective sub-regions statistically.

b) Expression recognition aims to correctly categorize different facial representations. Support vector machine (SVM) [7], [44], [61] is the most popular and effective learning method in facial expression recognition. Reference [44] is the most similar work to ours, so it will be considered as the baseline. For fair comparison, this paper will also employ SVM as the classification algorithm.

Besides these works, there are also some works utilizing the geometric features [11], [36], [37], [69], such as the location of facial feature points (corners of the eyes, mouth, etc.). Some methods perform facial expression analysis based on 3-D face models [11], [41], [62]. More works on fusion of audio and visual information can be found in [65].

B. MTSL

Sparsity methods have attracted much attention in computer vision, multimedia and medical image communities, and have been employed in many applications, such as face recognition [59], background subtraction [25], image annotation and retrieval [67], and shape prior-based segmentation [68]. Many algorithm are proposed to solve these problems of sparsity priors, such as greedy methods [basis pursuit (BP) [13], matching pursuit [32], orthogonal matching pursuit (OMP) [50]], or L1 norm relaxation and convex optimization [9], [22], [29].

MTSL is an inductive transfer machine learning approach. It aims to learn a problem together with some related problems for better performance [2], [10]. MTSL is then designed in [3] for feature selection, through encouraging multiple predictors from different tasks to share similar parameter sparsity patterns. MTSL also obtained a rewarding performance on handwritten character recognition in [23]. Yuan et al. [63] developed a visual classification algorithm by learning the shared parts among different representation tasks. Recently, Chen et al. [12] provided a faster solution to MTSL problems.

Suppose there are $T$ related tasks, and $(x_i^t, y_i^t), i = 1, 2, \ldots, N_t$ is the training set of task $t$, where each sample is represented by $K$-dimensional features, $x_i^t \in R^K$, and $y_i^t \in \{-1, 1\}$ indexes $x_i^t$ is negative or positive, $w^t$ is a $K$-dimensional vector of representation coefficients for task $t$. All the $w^t$s are the rows of the matrix $W = [w_1^t, w_2^t, \ldots, w_T^t]$. MTSL aims to learn the shared sparse information among all the tasks. The formulation with $L_1/L_2$ mixed-norm regularization is as follows:

$$
\arg\min_W \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} J^t(w^t, x_i^t, y_i^t) + \lambda \sum_{k=1}^{K} \|w_k\|_2
$$

where $J^t(w^t, x_i^t, y_i^t)$ is the cost function of the $t$th task, $\lambda$ is a constant to balance the sparsity, and $\sum$ is the mathematical format for $L_1$ norm. The regularization term encourages most columns of matrix $W$ to be zero, and the remaining nonzero columns indicate the corresponding features are shared features across all the tasks.

III. PROPOSED WORK

In this section, we first introduce the multiscale facial appearance representation strategy, and then the learning procedures of common and specific patches at each scale level are illustrated. Finally, we design the classifier with these learned effective patches.
A. Multiscale Appearance Representation

Facial expressions are usually manifested by local facial appearance variations. However, it is not easy to automatically localize these local active areas on a facial image. A facial image is divided into \( p \) local patches, and then LBP features are used to represent the local appearance of the patch. These features have been proven to be a powerful descriptor in expression recognition [44] and face verification [57]. Since the facial parts corresponding to facial muscles do not have equal size, multiscale division strategy should be applied to facial images to generate facial patches with different sizes. These patches could offer a more complete coverage for the effective facial parts than the generated patches if only one single scale is applied. The facial patches should have reasonable size, which can not be too big to cover too much facial expressions, the performance could not be the best, because each expression also has its special properties besides the common properties. Here, we aim to explore some specific

B. Learning Common Patches Across Expressions

Discovering the common patches across all the expressions is actually equivalent to learning the shared discriminative patches for all the expressions. Since MTSL can learn common representations among multiple related tasks [3], our problem can be transferred into a MTSL problem. \( T \) related tasks are defined as \( T \) discriminative patch learners for \( T \) facial expressions, respectively (we set \( T = 6 \) for six basic expressions). Suppose each image has \( p \) patches, it can be represented by \( (p \times m) \)-dimensional LBP-based histogram features. Let \( K = p \times m \). However, (1) cannot directly model our problem. Different from the MTSL model described in (1), we focus on the selection of common patches instead of individual features. Since a group of consecutive features stand for one patch, and

\[
\text{Feature Matrix}: [x_1', x_2', \ldots, x_s'] \bullet (w')' \rightarrow [y_1', y_2', \ldots, y_s']
\]

C. Learning Specific Patches for Individual Expression

Although learned common patches can discriminate all facial expressions, the performance could not be the best, because each expression also has its special properties besides the common properties. Here, we aim to explore some specific

The number of common patches is not large, group sparsity prior can be assumed [66], [67]. Our problem is modeled as the following MTSL problem, in which the regularization term of (1) is modified to a patch level sparse constraint:

\[
\arg \min_{W} \sum_{i=1}^{T} \frac{1}{N_i} \sum_{i=1}^{N_i} J^t(w', x_i', y_i') + \lambda \sum_{j=1}^{p} \| \text{AG}_j \|_2^2.
\]
Algorithm 1 Algorithm for Learning Common Patches

1: Input: Training data \( \{(x_i^t, y_i^t), i = 1, \ldots, N\} \), define \( X^t = [x_1^t; \ldots; x_{N_t}^t], Y^t = [y_1^t; \ldots; y_{N_t}^t], V = [v_1; \ldots; v_p] \), \( t \) indicates the task index, and \( t = 1, \ldots, T, j \) is the group index, and \( j = 1, \ldots, p \).
2: Initialize: \( W_0 \) takes equal weights, \( V_0 = W_0 \) and \( a_0 = 1 \).
3: for \( s = 0 \ldots S \) do
4: \( w_{s+1}^{t} = V_s \cdot \eta \left[ \frac{1}{1 + \exp(-(Y_s^t)^T V_s W_s^2) \exp(-(Y_s^t)^T V_s W_s^2) - (Y_s^t)^T V_s W_s^2} \right] \)
5: if \( \| w_{G_s + 1}^t \|_2 \geq \lambda \eta \) then
6: \( W_{G_s + 1} = (1 - \frac{\lambda \eta}{\| w_{G_s + 1}^t \|_2}) W_{G_s + 1} \)
7: else
8: \( W_{G_s + 1} = 0 \)
9: end if
10: \( a_{s+1} = \frac{2}{1 + \exp(-\epsilon)} \), \( \delta_{s+1} = W_{s+1} - W_s \)
11: \( V_{s+1} = W_{s+1} + \frac{1}{a_s} a_{s+1} \delta_{s+1} \)
12: if \( \| \delta_{s+1} \|_2 \leq \epsilon \) then
13: break
14: end if
15: end for
16: Normalization: \( \eta = \frac{w^t}{\| w^t \|_2} \)
17: \( w_j^t = \sum w_j^t \), where \( W_j^t \) is the weight for patch \( j \), and \( \{w_j^t \in G_j \} \)
18: Output: order \( W_j^t \) decreasingly, and output the top \( k \) patches as the common patches for all expressions.

Face verification task

![Fig. 5. Design of face verification task. Image pairs from the same subject are considered as positive samples. Otherwise, as negative samples.](image)

Facial patches for each expression with the help of face verification, and then they are used to further boost the performance of common facial patches.

The motivation to employ the face verification task is that those special facial patches are important face regions, which are not only useful for recognizing the expression, but also very significant for identifying the subjects. Take an expression \( e \) for example. Recognition \( e \) task will prefer to select out those patches which are useful only to recognizing the expression.

When both common and specific patches are investigated, the MTSL model can be used to couple these two tasks and select out those shared important patches between these two tasks more robustly. The learned patches should embed some specific signatures of the face identity. This MTSL framework for specific patches is the same as the framework of learning the common patches except the different task design. The individual expression analysis task is organized in the same way as in Fig. 4. Fig. 5 illustrates how to organize the task of face verification. For face verification, we need to compare two images and label them as the same person or not, so we organize the training data of this task by the feature difference between two images. Assuming \( \{x_i^e; y_i^e\}_{i=1}^{N_e} \) is the training set, \( x_i^e \) is the feature difference between two images \( x_i \) and \( y_i \). \( y_i^e \in \{1, -1\} \) indicates whether the two images in \( i \)th image pair come from one subject or not. \( N_e \) is the number of image pairs. The superscript 2 means this task is the second task in the MTSL model. The procedure for solving this problem is the same with Algorithm 1. Because there are six expressions, six MTSL models are needed to be built to learn their specific patches, respectively.

The specific patches have overlap with the learned common patchs. Since the common patches will be used for all expressions, the overlapped patches are removed from the specific patches. The rest patches are considered as the final specific patches.

D. Classifier Design

With the extracted common and specific patch features based on the training data, classifiers are then built based on these features for testing data. MTSL model can directly give out classification results [63]. However, to fairly compare with [30] and [44], SVM is adopted to learn the expression classifiers and the one-against-all strategy is employed to decompose the six class problem into multiple binary classification problems. Each binary classification will output a confidence value of the test sample belonging to this class. The class label with the highest confidence will be the final classification result of this sample. The performances of common patches and the combination of common and specific patches are evaluated, respectively. For expression \( e \), denotes the common patches as \( P_c \), and the specific patches as \( [P_s^e]_{e=1}^{6} \). When both common and specific patches are investigated, the features from \( P_c \) and \( P_s^e \) are concatenated to represent facial images, and train the SVM classifiers. While only use the features of \( P_c \) when common patches are tested.

IV. EXPERIMENTS

We evaluate the learned common and specific patches for facial expression recognition. All methods are compared on three datasets, the Cohn–Kanade database [28], the MMI database [52] and the GEMEP-FERA [33], which are widely used for facial expression recognition algorithms. Our methods are denoted as CPL and CSPL, respectively (see Table I). To efficiently evaluate the performance of our proposed methods, they are compared with [44], which is the most recent comprehensive study on expression recognition with remarkable
Fig. 6. Results for six expressions in the coefficient matrix after MTSL for learning the common patches. $x$-axis corresponds to the feature index in the coefficient matrix, where features indexes are ordered consecutively as group by patches. $y$-axis is the weight values for features in each task after MTSL. The nonzeros parts are grouped, and matches across all tasks. (a) Anger. (b) Disgust. (c) Fear. (d) Happiness. (e) Sadness. (f) Surprise.

![Image of six basic expressions from the Cohn–Kanade database](image)

Fig. 7. Example of six basic expressions from the Cohn–Kanade database (anger, disgust, fear, happiness, sadness, and surprise).

TABLE I

<table>
<thead>
<tr>
<th>METHOD ABBREVIATIONS</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADL</td>
<td>only use patches selected by ADboost are used.</td>
</tr>
<tr>
<td>AFL</td>
<td>All patches of the whole Face are used.</td>
</tr>
<tr>
<td>MADL</td>
<td>only use Multi-scale patches selected by ADboost are used.</td>
</tr>
<tr>
<td>MAFL</td>
<td>Multi-scale All patches of the whole Face are used.</td>
</tr>
<tr>
<td>CPL</td>
<td>only use Common Patches. (our method)</td>
</tr>
<tr>
<td>CSPL</td>
<td>use Common and Specific Patches. (our method)</td>
</tr>
<tr>
<td>MCPL</td>
<td>only use Multi-scale Common Patches. (our method)</td>
</tr>
<tr>
<td>MCSPL</td>
<td>use Multi-scale Common and Specific Patches. (our method)</td>
</tr>
</tbody>
</table>

results. In [44], two methods are evaluated, denoted as ADL and AFL, respectively. ADL uses Adaboost to select important patches and then performs SVM on the extracted LBP features of these patches. AFL uses all the patches to train the classifier without feature selection. MCPL, MCSPL, MADL, and MAFL are the methods using multiscale patches for CPL, CSPL, ADL, and AFL, respectively. For fair comparison, all the methods are based on the same patch (sub-region) division strategy, same feature representation, and the same classification method (SVM). The only difference among the methods is the patches they use. All method abbreviations are listed in Table I. Ten folds cross-validation is employed for all methods.

A. Experiments on the Cohn–Kanade Database

The Cohn–Kanade database consists of 100 university students aged from 18 to 30 years old, of which 65% were female, 15% were African-American, and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were based on description of prototypic emotions. For our experiments, image sequences are selected out from 96 subjects, whose sequences could be labeled as one of the six basic emotions. For each sequence, we only use the three peak frames with the most expressions. The faces are detected automatically by Viola’s face detector [55], and then they are normalized to $96 \times 96$ as in [45] based on the location of the eyes. Fig. 7 shows some normalized samples with all expressions.

To better demonstrate the patch selection strategy and the physical meaning of the selected patches, we first apply our method to only one patch scale ($S_8$). The performance of the recognition can be further boosted by combining the patch selection across all different scales.

1) Analysis of Common Patches: As described in Section II-A, the proposed MTSL aims to select the shared patches instead of the shared features, so we apply the $L_1/L_2$ norm regularization on the patch level to obtain patch-based group sparsity. Fig. 6 reports the representation coefficient results for six expression tasks. We can see that the representation coefficients of features are sparse, and show the property of patch-based group sparsity. It is also clear to see the index correspondences for nonzero values across six expressions, which indicates the commonalities among them. So, this result demonstrates the effectiveness of our proposed algorithm in learning the shared common patches for expressions.

Before evaluating the recognition performance of the common patches, we want to inspect the performance when a different number of common patches is selected. Fig. 8 reports the results with a different number of the common patches. We can see that the recognition rate increases quickly with the first leading common patches, and when the number of the selected patches reaches around 20, it will get a recognition rate of 88.42%. If too many common patches are selected, the performance goes down slightly and fluctuates. It means that only some common patches are discriminative for all the expressions. When some patches with little importance are selected...
Fig. 8. Expression recognition rate with a different number of common patches. (a) Recognition result with selected common patches for the scale $S_8$. The patch number for the three faces images marked with selected common patches are 10, 20, and 40, respectively. (b) Results for $S_6$. Patch number are 11 and 24, respectively. (c) Results for $S_4$. Patch number are 5 and 10, respectively. All results show the most effective patches are around the mouth and the eyes, and using only one third of all the patches can achieve satisfied performance.

Fig. 9. Distribution of selected common patches on faces. The darker the red color is, the more times (shown as numbers) the patch has been selected as common patches in ten-fold experiments.

Fig. 10. Top three specific patches for six expressions after eliminating the shared patches on the Cohn–Kanade database.

Table IV reports the detailed recognition performance of the common patches on each expression, where the expressions of anger, disgust, fear, happiness, sadness, surprise are denoted as ag, dg, fa, hp, sd, and sp for simplicity. Promising recognition rates are obtained on all the expressions except anger. Anger is often misclassified as sadness. This is because these two expressions have similar appearance variations on the common patches. This problem can be alleviated by adding some specific patches, which will be discussed next.

![Table II: Confusion Matrix of AFL on Cohn–Kanade Database (Measured by Recognition Rate: %)]

<table>
<thead>
<tr>
<th></th>
<th>ag</th>
<th>dg</th>
<th>fa</th>
<th>hp</th>
<th>sd</th>
<th>sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>ag</td>
<td>66.67</td>
<td>7.5</td>
<td>0</td>
<td>0.83</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>dg</td>
<td>6.67</td>
<td>87.67</td>
<td>0.67</td>
<td>1.33</td>
<td>3.67</td>
<td>0</td>
</tr>
<tr>
<td>fa</td>
<td>3.73</td>
<td>1.43</td>
<td>77.54</td>
<td>10.40</td>
<td>6.90</td>
<td>0</td>
</tr>
<tr>
<td>hp</td>
<td>1.00</td>
<td>0.33</td>
<td>2.58</td>
<td>95.42</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>sd</td>
<td>10.60</td>
<td>1.25</td>
<td>2.87</td>
<td>0</td>
<td>84.54</td>
<td>0.74</td>
</tr>
<tr>
<td>sp</td>
<td>0</td>
<td>0</td>
<td>1.73</td>
<td>0</td>
<td>1.25</td>
<td>97.02</td>
</tr>
</tbody>
</table>

![Table III: Confusion Matrix of ADL on Cohn–Kanade Database (Measured by Recognition Rate: %)]

<table>
<thead>
<tr>
<th></th>
<th>ag</th>
<th>dg</th>
<th>fa</th>
<th>hp</th>
<th>sd</th>
<th>sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>ag</td>
<td>64.72</td>
<td>10.00</td>
<td>1.11</td>
<td>0</td>
<td>24.17</td>
<td>0</td>
</tr>
<tr>
<td>dg</td>
<td>5.33</td>
<td>89.33</td>
<td>2.00</td>
<td>1.33</td>
<td>2.00</td>
<td>0</td>
</tr>
<tr>
<td>fa</td>
<td>3.65</td>
<td>1.43</td>
<td>78.57</td>
<td>10.87</td>
<td>5.48</td>
<td>0</td>
</tr>
<tr>
<td>hp</td>
<td>1.00</td>
<td>0.33</td>
<td>3.24</td>
<td>94.76</td>
<td>0.67</td>
<td>0</td>
</tr>
<tr>
<td>sd</td>
<td>11.43</td>
<td>1.25</td>
<td>2.50</td>
<td>0</td>
<td>84.40</td>
<td>1.42</td>
</tr>
<tr>
<td>sp</td>
<td>0</td>
<td>0</td>
<td>1.73</td>
<td>0</td>
<td>0</td>
<td>98.27</td>
</tr>
</tbody>
</table>

![Table IV: Confusion Matrix of CPL on the Cohn–Kanade Database (Measured by Recognition Rate: %)]

<table>
<thead>
<tr>
<th></th>
<th>ag</th>
<th>dg</th>
<th>fa</th>
<th>hp</th>
<th>sd</th>
<th>sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>ag</td>
<td>65.56</td>
<td>8.33</td>
<td>0</td>
<td>0</td>
<td>25.28</td>
<td>0.83</td>
</tr>
<tr>
<td>dg</td>
<td>2.67</td>
<td>92.67</td>
<td>0.67</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>fa</td>
<td>0</td>
<td>1.98</td>
<td>78.97</td>
<td>13.25</td>
<td>5.79</td>
<td>0</td>
</tr>
<tr>
<td>hp</td>
<td>0.33</td>
<td>0.67</td>
<td>4.24</td>
<td>94.76</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sd</td>
<td>6.20</td>
<td>1.67</td>
<td>3.33</td>
<td>0</td>
<td>87.69</td>
<td>1.11</td>
</tr>
<tr>
<td>sp</td>
<td>0</td>
<td>0</td>
<td>1.25</td>
<td>0</td>
<td>0.48</td>
<td>98.27</td>
</tr>
</tbody>
</table>
2) Analysis of Specific Patches: Although a rewarding recognition result can be obtained by only using the common patches, the performance can be further improved by integrating some specific patches of each expression. Fig. 10 shows the top three learned specific patches for each expression based on the proposed multitask learning. We can see the locations of these patches are highly related to expression types. Take surprise for example. The selected specific patches show the characteristics of surprise expression, in which special appearance changes are distributed in opened mouth, on stared eye, and raised eyebrow. In CSPL, the common patches and the specific patches are integrated together, and the experimental results are reported in Table V. Compared to the results of CPL (Table IV), we can see that adding specific patches can further improve the performance of the common patches.

3) Comparisons With Other Methods: To further evaluate the proposed CPL and CSPL, we compare them to ADL and AFL developed in [44]. Table VI lists the F1 measure for every expression and the overall recognition rates of these four methods. AFL gets the recognition rate of 86.94%, which is much worse than our methods. It shows the importance of selecting discriminative patches. The confusion matrixes of methods, AFL and ADL, can also be found in Tables II and III. Although ADL also uses Adaboost to select the patches, it does not take the commonalities among all the expressions into account. ADL gets a recognition rate of 82.26% with the selected patches (highest rate with 20 ± 3 patches), while the recognition rates of CPL and CSPL are 88.42% and 89.89%, respectively. It demonstrates that the learned common and specific patches by our proposed two-stage MTSL can really improve the performance of expression recognition.

4) Analysis of Multiscale Patches: The selected patches are shown to be effective in expression recognition, even the number of them is limited [70]. However, no prior of the optimal patch size is given. Besides, patches with different scales may represent different information for recognition. So, we can employ the patch selection strategy on different patch sizes, and then combine the selected patches with different scales together to further improve the recognition performance. In Fig. 11, we show the selected common patches for all the scales for one of the ten-fold experiments. For each scale, only 1/3 patches are selected out. The results show that for different scales, the effective patches are around the mouth and the eyes. As shown in Table VI, the performance of these multiscale patches can achieve better performance (90.34%) than only using the common patches from a single scale S8 (88.42%). The F1 measures for each expression also validate the usefulness of multiscale strategy. The specific patches are selected out at different scale individually, following the way we already shown for the scale S8. Incorporating selected specific patches, we get the best performance (91.53%). Thus, the experiment results show that multiscale patches could contain more useful statistic information for recognition than one scale with a single division strategy. Our patches selection are effective in selecting the discriminative patches across all patch scales.

2) Analysis of Specific Patches: Although a rewarding recognition result can be obtained by only using the common patches, the performance can be further improved by integrating some specific patches of each expression. Fig. 10 shows the top three learned specific patches for each expression based on the proposed multitask learning. We can see the locations of these patches are highly related to expression types. Take surprise for example. The selected specific patches show the characteristics of surprise expression, in which special appearance changes are distributed in opened mouth, on stared eye, and raised eyebrow. In CSPL, the common patches and the specific patches are integrated together, and the experimental results are reported in Table V. Compared to the results of CPL (Table IV), we can see that adding specific patches can further improve the performance of the common patches.

3) Comparisons With Other Methods: To further evaluate the proposed CPL and CSPL, we compare them to ADL and AFL developed in [44]. Table VI lists the F1 measure for every expression and the overall recognition rates of these four methods. AFL gets the recognition rate of 86.94%, which is much worse than our methods. It shows the importance of selecting discriminative patches. The confusion matrixes of methods, AFL and ADL, can also be found in Tables II and III. Although ADL also uses Adaboost to select the patches, it does not take the commonalities among all the expressions into account. ADL gets a recognition rate of 82.26% with the selected patches (highest rate with 20 ± 3 patches), while the recognition rates of CPL and CSPL are 88.42% and 89.89%, respectively. It demonstrates that the learned common and specific patches by our proposed two-stage MTSL can really improve the performance of expression recognition.

4) Analysis of Multiscale Patches: The selected patches are shown to be effective in expression recognition, even the number of them is limited [70]. However, no prior of the optimal patch size is given. Besides, patches with different scales may represent different information for recognition. So, we can employ the patch selection strategy on different patch sizes, and then combine the selected patches with different scales together to further improve the recognition performance. In Fig. 11, we show the selected common patches for all the scales for one of the ten-fold experiments. For each scale, only 1/3 patches are selected out. The results show that for different scales, the effective patches are around the mouth and the eyes. As shown in Table VI, the performance of these multiscale patches can achieve better performance (90.34%) than only using the common patches from a single scale S8 (88.42%). The F1 measures for each expression also validate the usefulness of multiscale strategy. The specific patches are selected out at different scale individually, following the way we already shown for the scale S8. Incorporating selected specific patches, we get the best performance (91.53%). Thus, the experiment results show that multiscale patches could contain more useful statistic information for recognition than one scale with a single division strategy. Our patches selection are effective in selecting the discriminative patches across all patch scales.

B. Results on the MMI Database

The MMI database includes 30 students and research staff members aged from 19 to 62, of whom 44% are female, having either a European, Asian, or South American ethnic background. In this database, 213 sequences have been labeled with six basic expressions, in which 205 sequences are with frontal face. Different from [44], in which only the experimental data are collected from 99 selected sequences, we conduct our experiments on the data from all the 205 sequences. As in [44], the apex images are extracted from the sequences as the experimental data. Facial image are cropped based on locations of eyes, and resize it to 96 × 96, same as on Cohn–Kanade database.

MMI is a more challenging database than the Cohn–Kanade database. First, the subjects make expressions nonuniformly. Different people make the same expression in different ways. Second, some subjects wear accessories, such as glasses, headcloth, or moustache. Additionally, in some sequences, the apex frames are not with high expression intensity. All these factors will greatly degrade the recognition performance.

We first investigate the performance of the common patches with different patch number, and Fig. 12 shows the results. It can be seen that the results are similar to the results on the Cohn–Kanade database. About 20 common patches are discriminative for all the expressions, so we set the number of the common patches as 20 on this database too. Table IX lists the F1 measures for each expression and the overall recognition rates of CPL, CSPL, AFL, ADL, AFL, and their corresponding multiscale methods, respectively. Same as on the Cohn–Kanade data, CPL, and CSPL are superior to AFL and ADL. However, the performances of all four methods are much lower than that of the Cohn–Kanade database, because this database has several challenging factors as mentioned above. CSPL obtains much better performance than CPL. This is because each expression has a very big variance due to the diversity of the subjects in this database, but the common patches cannot describe these specific variations. Although a much better result of 86.7% is reported in [44], their experimental data are carefully chosen 99 sequences, while we perform the experiments on all the 205 sequences. Besides, they adopt sliding and multiscale windows to extract much more patches. We only divide the facial image into 64 patches, and we also obtain a recognition rate of 73.53% on more than double the size of the data than in [44]. Tables VII and VIII list the confusion matrix of CPL and CSPL, respectively.

<table>
<thead>
<tr>
<th>TABLE V</th>
<th>Confusion Matrix of CSPL on the Cohn–Kanade Database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ag</td>
</tr>
<tr>
<td>ag</td>
<td>71.3889</td>
</tr>
<tr>
<td>dg</td>
<td>2.67</td>
</tr>
<tr>
<td>fa</td>
<td>0</td>
</tr>
<tr>
<td>hp</td>
<td>0.33</td>
</tr>
<tr>
<td>sd</td>
<td>7.45</td>
</tr>
<tr>
<td>sp</td>
<td>0</td>
</tr>
</tbody>
</table>

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.
TABLE VI
RECOGNITION PERFORMANCES AND F1 MEASURES PER EXPRESSION FOR ALL COMPARED METHODS (I.E., AFL, ADL, CPL, CSPL, MAFL, MADL, MCPL, MCSPL) ON THE COHN–KANADE DATABASE

<table>
<thead>
<tr>
<th>Expressions</th>
<th>AFL</th>
<th>ADL</th>
<th>CPL</th>
<th>CSPL</th>
<th>MAFL</th>
<th>MADL</th>
<th>MCPL</th>
<th>MCSPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.6407</td>
<td>0.6281</td>
<td><strong>0.7144</strong></td>
<td><strong>0.7440</strong></td>
<td>0.6512</td>
<td>0.6325</td>
<td><strong>0.7350</strong></td>
<td><strong>0.7628</strong></td>
</tr>
<tr>
<td>Disgust</td>
<td>0.8782</td>
<td>0.8776</td>
<td><strong>0.8927</strong></td>
<td><strong>0.9134</strong></td>
<td>0.8875</td>
<td>0.8796</td>
<td><strong>0.9105</strong></td>
<td><strong>0.9411</strong></td>
</tr>
<tr>
<td>Fear</td>
<td>0.8235</td>
<td>0.8206</td>
<td><strong>0.8209</strong></td>
<td><strong>0.8432</strong></td>
<td>0.8369</td>
<td>0.8316</td>
<td><strong>0.8462</strong></td>
<td><strong>0.8619</strong></td>
</tr>
<tr>
<td>Happiness</td>
<td>0.9416</td>
<td>0.9381</td>
<td><strong>0.9305</strong></td>
<td><strong>0.9462</strong></td>
<td>0.9536</td>
<td>0.9467</td>
<td><strong>0.9512</strong></td>
<td><strong>0.9635</strong></td>
</tr>
<tr>
<td>Sadness</td>
<td>0.8204</td>
<td>0.8346</td>
<td><strong>0.8515</strong></td>
<td><strong>0.8619</strong></td>
<td>0.8269</td>
<td>0.8362</td>
<td><strong>0.8645</strong></td>
<td><strong>0.8829</strong></td>
</tr>
<tr>
<td>Surprise</td>
<td>0.9806</td>
<td>0.9791</td>
<td><strong>0.9827</strong></td>
<td><strong>0.9870</strong></td>
<td>0.9806</td>
<td>0.9821</td>
<td><strong>0.9842</strong></td>
<td><strong>0.9870</strong></td>
</tr>
<tr>
<td>Recognition Rate</td>
<td>0.8694</td>
<td>0.8226</td>
<td><strong>0.8842</strong></td>
<td><strong>0.8989</strong></td>
<td>0.8732</td>
<td>0.8324</td>
<td><strong>0.9034</strong></td>
<td><strong>0.9153</strong></td>
</tr>
</tbody>
</table>

TABLE VII
CONFUSION MATRIX OF CPL ON MMI DATABASE (MEASURED BY RECOGNITION RATE: %)

<table>
<thead>
<tr>
<th>(%)</th>
<th>ag</th>
<th>dg</th>
<th>fa</th>
<th>hp</th>
<th>sd</th>
<th>sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>ag</td>
<td>46.94</td>
<td>25.83</td>
<td>1.11</td>
<td>1.11</td>
<td>25.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dg</td>
<td>24.17</td>
<td>46.67</td>
<td>0</td>
<td>16.67</td>
<td>11.67</td>
<td>0.83</td>
</tr>
<tr>
<td>fa</td>
<td>2.22</td>
<td>7.78</td>
<td>49.44</td>
<td>13.33</td>
<td>22.22</td>
<td>5.00</td>
</tr>
<tr>
<td>hp</td>
<td>3.50</td>
<td>7.67</td>
<td>7.83</td>
<td><strong>70.67</strong></td>
<td>4.17</td>
<td>6.17</td>
</tr>
<tr>
<td>sd</td>
<td>22.50</td>
<td>7.50</td>
<td>7.78</td>
<td>11.39</td>
<td><strong>50.83</strong></td>
<td>0.00</td>
</tr>
<tr>
<td>sp</td>
<td>0.83</td>
<td>2.50</td>
<td>25.33</td>
<td>2.50</td>
<td>2.50</td>
<td><strong>66.33</strong></td>
</tr>
</tbody>
</table>

We further explore the common and specific patches selection for different patch sizes. The accuracy rate of the baseline method MAFL and MADL increases with more multiscale patches. The recognition rate of MCPL and MCSPL using more multiscale patches both achieve better performance than CPL and CSPL with single patch scale. The detail results are also shown in Table IX.

The experimental results indicate the location of learned common and specific patches, which confirms the previous knowledge about active facial parts in psychology. The rewarding performances of these patches in facial expression recognition provide a solid basis for patches selection and weight setting in similar applications. This paper opens the road for the researches of utilizing the prior knowledge of facial muscles in psychology, and further improves the performances of existing methods in computer vision.

C. Experiments on the GEMEP-FERA2011 Database

The Geneva multimodal emotion portrayals (GEMEP) is a collection of audio and video recordings [4], [5]. It consists of over 7000 audiovisual emotion portrayals, representing 18 emotions portrayed by ten actors trained by a professional director. The GEMEP-FERA2011 Database contains both AU sub-challenge and Emotion sub-challenge. This paper will focus on the Emotion sub-challenge, in which a total of 289 portrayals are selected (155 for training and 134 for testing). The training set included seven actors with 3 to 5 instances of each emotion per actor. The test set includes six actors, half of which were not present in the training...
For each video in training or testing, five frames are evenly extracted from the video. We use the landmark detection from [71] to localize the position of the eyes, based on which we align and crop out the face images. The landmark detection result and cropped images are shown in Fig. 13.

The emotion recognition challenge involves the classification of the following five emotions: anger, fear, joy, relief, and sadness. When determining the label of the test video, we first classify the extracted frames individually using a five-way forced strategy, then the emotion class which obtains the highest score will be the winner of the sequence.

Since the subjects in training part of database are quite few, it is quite difficult to learn the common patches and the specific patches on this database itself. In this experiment, we utilize the knowledge of the learned patches from the CK database. It also shows the capability of generalization across different databases. We follow the training/testing partition in the database, and the performances of our methods are compared with some related works in Table X. Our method MCSPL achieves much better recognition rate than the baseline work (i.e., [14], [33]). These results that the patches selected out by our algorithm are discriminative for expression recognition and are also robust across databases. Compared to the 1st prize of 2011 FERA competition (i.e., [42]), which gets 83.5% accuracy, our method MCSPL achieves 80.0%.

The detailed confusion matrix of the result is shown in Table XI. Senechal et al. [42] utilize a lot of features, such as LGPB histograms, 2.5 D Active Appearance model to combine appearance and geometry information. It also employs complex classification algorithm, such as, multikernels SVMs.
temporal filtering. However, our method just uses the LBP feature and naïve SVM. Considering these factors, our patch learning method is proved to be effective, and the results of our method are promising. More detailed comparisons of using different patches are given in Table XII.

V. CONCLUSION

In this paper, a new method to analyze facial expressions is proposed. Different from previous work, we aimed at exploring the commonalities among the expressions by discovering the common and specific patches. A two-stage sparse learning model is proposed to learn the locations of these patches based on the prior knowledge of facial muscles and AUs. A multiscale face division strategy is employed to obtain facial patches with different coverage area and eliminate the side effects from fixed patch size. The effectiveness of these patches are evaluated by facial expression recognition. Extensive experiments show that common patches can generally discriminate all the expressions, and the recognition performance can be further improved by integrating specific patches. More comprehensive patches can also be selected out to achieve better performance by using multiscale patch division strategy. The learned location information of these patches also confirms the location knowledge of facial muscles in psychology.

REFERENCES


Authors’ photographs and biographies not available at the time of publication.