Simultaneous Recognition of Facial Expression and Identity via Sparse Representation

Mohammad Reza Mohammadi  
Sharif University of Technology  
Tehran, Iran  
mrmohammadi@ee.sharif.edu

Emad Fatemizadeh  
Sharif University of Technology  
Tehran, Iran  
fatemizadeh@sharif.edu

Mohammad H. Mahoor  
University of Denver  
Colorado USA  
mmahoor@du.edu

Abstract

Automatic recognition of facial expression and facial identity from visual data are two challenging problems that are tied together. In the past decade, researchers have mostly tried to solve these two problems separately to come up with face identification systems that are expression-independent and facial expressions recognition systems that are person-independent. This paper presents a new framework using sparse representation for simultaneous recognition of facial expression and identity. Our framework is based on the assumption that any facial appearance is a sparse combination of identities and expressions (i.e., one identity and one expression). Our experimental results using the CK+ and MMI face datasets show that the proposed approach outperforms methods that conduct face identification and face recognition individually.

1. Introduction

Most of the proposed methods for facial image analysis deal with only one attribute of face at a time, either the identity or expression. These methods are compatible with the classical models of face perception in neuroscience that propose separate neural processing routes for facial expression recognition and face identification [2]. However, recent models of face perception suggest these routes are not independent of each other [3]. The interaction of facial expression and identity are more considered in the proposed algorithms in recent years. However, there are a limited number of works that attempt to recognize both facial expression and identity simultaneously.

In some studies such as [14] facial expression is recognized first, and then some selected features and classifiers are used for face identification. According to [14], using the information of facial expression improves the face identification rate. On the other hand, in some studies such as [19] the identity is predicted first, then its associated expression model is used to recognize the facial expression. These studies show that using the information of identity is helpful to recognize the facial expression and vice versa.

Giving a priority to either tasks (i.e., facial expression recognition or face identification) has two main drawbacks: 1) the preceded task is performed without using the other one and hence no improvement is occurred in the recognition rate of the first task; 2) the failures in the recognition of the first attribute may affect directly the result of the second one. Thus, by developing an algorithm to simultaneously conduct facial expression and identity recognition, we can overcome these drawbacks and improve the recognition rate of both attributes. Moreover, as was mentioned in [3], the position of the systems for recognition of facial expression and identity are separated in the brain but their routes are not independent. Thus, giving a priority to one of these tasks is not consistent with the biological systems. We have attention to biological systems because the performance of them in cognition is very better than the existing computer vision algorithms.

The N-mode SVD tensor decomposition on facial images is proposed in [30] to separate the influence of identity, pose, illumination, and expression. After decomposition of a facial image, different facial attributes are recognized by some classifiers. Similarly, Higher-Order SVD is proposed in [32] to learn expression and person subspaces. In [23], bilinear models are used to separate the parameters which control the expression and those which control the identity. Then, joint expression-invariant face recognition and identity-invariant expression recognition are efficiently achieved. Using principal components extracted from training data, facial emotion is represented as a linear combination of its identity and expression in [31]. Finally, nearest neighbor classifier employed for identity recognition, and support vector machines (SVM) used for expression recognition. Both works [23, 31] use 3D facial images in their experiments.

There are six basic emotions (i.e., Anger, Disgust, Fear, Happiness, Sadness, and Surprise) that are considered as
universal and their corresponding facial expressions are recognized across different cultures [11]. Therefore, it is expected that the changes associated to the facial appearance of each basic expression from the neutral expression be similar among individuals. Hence, it can be assumed that the facial image of a subject’s expression be the sum of the subjects neutral facial image and the corresponding emotion. Fig. 1 shows two samples of this linear decomposition of facial images to neutral and difference images. As expected and observed in Fig. 1(c), the corresponding images of facial images to neutral and difference images. As expected, the differences among similar basic expressions are denoted by $D_{k,l}$, where $k$ is the index of the subject, and $S$ is the number of subjects. With this definition, the difference image $(D_{k,l})$ is defined as follow:

$$D_{k,l} = E_{k,l} - N_{k,l}$$

where $N_{k,l}$ is the neutral image of the $k^{th}$ subject (it is possible that all neutral images of each subject be exactly identical). Neutral images of a subject $(i.e., N_{k,l})$ with fixed $k$ are very similar. On the other hand, as mentioned before, difference images corresponding to one of the basic expressions $(i.e., D_{k,l})$ with fixed $l$ are similar with some subject-dependent variations. Thus, face identification and facial expression recognition can perform accurately using $N$ and $D$, respectively. Nevertheless, in the test step only $E$ is available, and its decomposition to $N$ and $D$ is not a simple work. In this section, after a brief review of the SRC framework [33], we propose a novel method to simultaneous recognition of both identity and expression from $E$.

### 2.1. Sparse Representation-based Classification

In SRC, it is assumed that a test feature vector $x_i$ can be linearly represented based on the training feature vectors of the corresponding class:

$$x_i = s_{i,1}a_{i,1} + s_{i,2}a_{i,2} + \cdots + s_{i,n_i}a_{i,n_i} = A_is_i$$

where $a_{i,j}$ is the $j^{th}$ training feature vector of the $i^{th}$ class, $s_{i,j}$ is the corresponding coefficient of the $j^{th}$ feature, and $n_i$ is the number of training samples of the $i^{th}$ class. Nonetheless, the class label of the test feature vector is not known. For this reason, all training feature vectors are concatenated to make a matrix with the following structure:

$$A = [a_{1,1} a_{1,2} \ldots a_{1,n_1} \ldots a_{c,1} \ldots a_{c,n_c}] = [A_1 A_2 \ldots A_c]$$

where $c$ is the number of classes. In sparse notation, $A$ is called dictionary and its columns are called atoms. By solving (2), an ideal representation of $x_i$ over $A$ is $[0, \ldots, 0, s_{i,1}, \ldots, s_{i,n_i}, 0, \ldots, 0]$ where most of its elements are 0. Thus, representation of any test feature vector $x$ over $A$ may be a sparse vector $s$ that the index of its non-zero elements corresponds to the atoms of the correct class.
When the number of training samples is larger than the number of features, the system $x = As$ is called underdetermined. Thus, the number of solutions $s$ that satisfies $x = As$ can be more than one. Therefore, to choose one of the solutions, some extra constraints need to be added to the problem. Because the number of atoms for each class is a small fraction of all atoms, one constraint is the sparsity of the solution which is defined using L0 norm and is presented as follows:

$$\ell^0 : \hat{s}_0 = \arg\min \|s\|_0 \text{ subject to } As = x$$

(4)

After optimization of $\ell^0$, the class label is predicted based on the sparse solution. For instance, the reconstruction error of each class is used in [33] for classification:

$$\min_i r_i(x) = \|x - A\delta_i(\hat{s}_0)\|_2$$

(5)

where $r_i(x)$ is the reconstruction error for the $i^{th}$ class, $\hat{s}_0$ is the sparse representation of $x$ over $A$ obtained by any sparse solver, and $\delta_i(\hat{s}_0)$ is $\hat{s}_0$ whose elements correspond to other classes than $i^{th}$ are set to zero.

In real applications, it is possible that the test feature vector cannot be represented linearly based on the corresponding training feature vectors, especially when the number of training samples is small. In this condition, since L0 norm is very sensitive, the sparse solution may be quite different from the ideal solution (the location of nonzero elements differ). The noisy version of the test feature vector can be written as:

$$x = x_0 + e_0 = As_0 + Ie_0$$

(6)

where $x_0$ is the noiseless feature vector, $e_0$ is the corruption noise, $I$ is an identity matrix, and $A_{rob}$ is a robust version of $A$ to corruption and occlusion. If only a little portion of $x$ is corrupted, $e_0$ will be sparse. Thus, $w_0$, that is the concatenation of $s_0$ and $e_0$, will be sparse too. In summary, by concatenating $A$ and $I$, a new dictionary is built that is robust to corruption and occlusion. To achieve this advantage, we will add the identity matrix to the dictionaries of all experiments.

For face recognition, dictionary $A$ can be formed by $N_{k,l}$ vectors where its label is $k$. Similarly, for facial expression recognition, dictionary $A$ can be formed by $D_{k,l}$ vectors that its label is $l$. In the following subsection a new dictionary is introduced for simultaneous recognition of facial expression and identity from a facial image (i.e., $E$).

2.2. Proposed Simultaneous Recognition using SRC

For independent recognition of identity and expression, we propose to use two separate dictionaries as defined below:

$$A_{id} = [N_{1,1}, N_{1,2} \ldots N_{1,6}, N_{2,1} \ldots N_{2,6}]$$

(7)

$$A_{ex} = [D_{1,1}, D_{1,2} \ldots D_{1,6}, D_{2,1} \ldots D_{2,6}]$$

(8)

Dictionaries $A_{id}$ and $A_{ex}$ contain the neutral and difference images, respectively. Therefore, any neutral test feature vector can be represented over its dictionary by $N_{test} = A_{id}s_{id}$. Similarly, $D_{test} = A_{ex}s_{ex}$ is obtained for any difference image. Using the aforementioned fact that a given facial image, $E_{test}$, can be decomposed into a neutral and a difference image, the relation among these images can be written as:

$$E_{test} = N_{test} + D_{test} = A_{id}s_{id} + A_{ex}s_{ex}$$

(9)

where $A_{sim}$ is the designed dictionary for simultaneous recognition of facial expression and identity that is obtained by concatenating $A_{id}$ and $A_{ex}$. Since it is assumed that both $s_{id}$ and $s_{ex}$ are sparse, $s_{sim}$ is sparse too and can be found using an optimization problem such as $\ell^0$ in Eq. 4. An overview of this approach is illustrated in Fig. 2. Then, the sparse solution is used to recognize the facial expression and identity based on a joint reconstruction error:

$$\min_{k,l} r_{k,l}(E_{test}) = \|E_{test} - A_{sim}\delta_{k,l}(\hat{s}_{sim})\|_2$$

(10)

where $r_{k,l}(E_{test})$ is the joint reconstruction error for the $k^{th}$ subject and the $l^{th}$ expression, $\hat{s}_{sim}$ is the sparse representation of $x$ over $A_{sim}$ obtained by any sparse solver, and $\delta_{k,l}(\hat{s}_{sim})$ is $\hat{s}_{sim}$ whose elements correspond to other subjects than $k^{th}$ and other expressions than $l^{th}$ are set to zero.

To make the approach more consistent with biological systems (separate position but dependent routes for expression and identity recognition), we propose to separately recognize each attribute:

$$\min_k r_k(E_{test}) = \|E_{test} - A_{sim}\delta_k(\hat{s}_{sim})\|_2$$

(11)

$$\min_l r_l(E_{test}) = \|E_{test} - A_{sim}\delta_l(\hat{s}_{sim})\|_2$$

(12)

where in $\delta_k(\hat{s}_{sim})$ the coefficients correspond to the $k^{th}$ subject as well as all expressions are retained, and others are set to zero. A similar definition is applied to
Figure 2. Overview of the proposed approach. (a) a test facial image, (b) sparse representation vector of the test image on the dictionary that the coefficients correspond to the correct classes are marked with red (identity) and green (expression) colors, and (c) the symbolic dictionary with marked correct classes by color boxes.

\( \delta_{k,l}(\hat{s}_{sim}) \). We call the first scheme using Eq. (10) “joint” scheme and the second scheme using Eqs. (11) and (12) “separate” scheme.

The pseudo-code of the proposed method for simultaneous recognition of facial expression and identity via sparse representation is presented in Table 1.

### 3. Experimental Results

This section presents the experimental results of the proposed method for recognition of identity and six basic expressions using two public datasets, the Extended Cohn-Kanade (CK+) dataset [16], and the MMI dataset [24]. The CK+ dataset includes 593 video sequences recorded from 123 university students ranging from 18 to 30 years old. Each of the sequences contains images from neutral expression to peak expression (last frame) where these two frames are used as the neutral image (i.e., \( N_{k,l} \)) and the expressive image (i.e., \( E_{k,l} \)), respectively. In CK+, some subjects express only some of the six basic emotions (varies from 0 to 6). Therefore, 309 facial images from 106 subjects are in the final collection of the CK+ dataset where Surprise and Fear have the maximum (83) and the minimum (25) number of images, respectively.

The MMI dataset includes images of more than 20 students and research staff members of both genders (44% female), ranging in age from 19 to 62 years, having either a European, Asian, or South American ethnic background [28]. For each session, an image sequence is captured that have neutral faces at the beginning and the end. In our experiments, with this criterion that the image sequence labeling by an expert be identical to the ground-truth, 121 image sequences were selected from the MMI dataset. These sequences come from 28 subjects, with 1 to 6 emotions per subject where Surprise and Sadness have the maximum (25) and the minimum (12) number of sequences, respectively.

Table 1. Pseudo-code of the proposed method for simultaneous recognition of facial expression and identity via sparse representation

<table>
<thead>
<tr>
<th>Train: Dictionary Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inputs:</td>
</tr>
<tr>
<td>( A_{id} ): a matrix of neutral training samples in Eq. 7</td>
</tr>
<tr>
<td>( A_{ex} ): a matrix of difference training samples in Eq. 8</td>
</tr>
<tr>
<td>2. Dictionary building: ( A_{sim} = [A_{id} ~ A_{ex} ~ I] )</td>
</tr>
<tr>
<td>3. Output:</td>
</tr>
<tr>
<td>( A_{sim} ): final dictionary</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test: Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inputs:</td>
</tr>
<tr>
<td>( E ): a test sample</td>
</tr>
<tr>
<td>( A_{sim} ): dictionary</td>
</tr>
<tr>
<td>2. Solve the following ( \ell^0 ) problem:</td>
</tr>
<tr>
<td>( \hat{s}_{sim} = \arg \min | s |<em>0 \text{ subject to } A</em>{sim}s = E )</td>
</tr>
<tr>
<td>3. Recognition of identity and expression with one of the following schemes:</td>
</tr>
<tr>
<td>( k ): recognized identity</td>
</tr>
<tr>
<td>( l ): recognized expression</td>
</tr>
</tbody>
</table>

In each sequence, the first and the peak frames are chosen as the neutral and the expressive facial images, respectively. Some example images of the CK+ and the MMI datasets are shown in Fig. (3).

To register the facial images, the two outer corners of the eyes are transferred to two fix points by a rigid body trans-
form, and then the images are stretched such that the subnasale point locates to another fix point. By this method, the in-plane rotations of face can be removed, and the important locations in the facial images are located on each other. Indeed, we use only these three points because they are approximately fixed relative to each other in all expressions of one subject. If we use other points that displace in some expressions, the registration can reduce the effect of expression on the registered facial images. To extract the required points in the facial images, we use the automatic Constrained Local Models (CLM) algorithm [7]. Finally, a fixed rectangle around these 3 points is considered as the facial region. Fig. 4 demonstrates a sample facial image registered using this method. Afterwards, the cropped facial image is resized to $18 \times 15$ pixels and vectorized to a feature vector with 270 elements. Z-score normalization technique (i.e., zero-mean and unit-variance feature vector) is applied to make the system robust against illumination variations and different skin colors.

An important step in the proposed algorithm is finding the sparse solution of Eq. 4. It is shown in [1] that the optimization problem of $\ell^0$ in Eq. 4 is NP-Hard. Because of this, some algorithms such as Basis Pursuit (BP) [4], Matching Pursuit (MP) [18], and Smoothed L0 norm (SL0) [22] exists to approximate the solution of $\ell^0$. Due to the strong theoretical support and high performance reported in the studies of SRC [33, 34, 6], the BP method is used as the main sparse solver in this study (we will compare different sparse solvers in subsection 3.4) which minimizes the L1 norm problem instead of L0 norm:

$$\ell^1 : \hat{s}_1 = \arg\min_{s} \|s\|_1 \quad \text{subject to} \quad As = x \quad (13)$$

It is shown in [9] that if the solution vector is more sparse than a threshold, optimization results of $\ell^0$ and $\ell^1$ are the same.

In the two following subsections, the results of different approaches presented in this paper are reported using the CK+ dataset. Then, in subsection 3.3, the experimental results on the MMI dataset are presented and then compared with the results on the CK+ dataset.

### 3.1. Independent Recognition Results

For face recognition, all the neutral images in the CK+ collection are used to build the dictionary $A_{id}$ that is defined in Eq. 7. Using all the expressive images as test images, the achieved face recognition rate (FR) is 89.97%.

For facial expression recognition, the difference images are used with Leave-One-Subject-Out (LOSO) cross validation method. In LOSO, all images of one subject are hold out as test images and images of other subjects are used to build the dictionary. In other words, the atoms of dictionary $A_{ex}$ in Eq. 8 that correspond to the test subject are removed for each subject. This process is repeated for every subject, and the average recognition rate computed over all subjects. The facial expression recognition rate (FER) by this method is 90.61% on the CK+ dataset. The confusion matrix for this experiment is shown in Table 2 that demonstrates the similarity of some expressions such as Fear and Happy. The approach requires to have the neutral image of the test subject which can be considered as a limitation.

### 3.2. Simultaneous Recognition Results

In this experiment, the dictionary is built using all neutral images and difference images of other subjects than the test subject (i.e., LOSO evaluation for facial expression recognition) similar to Eq. 9. Then, any expressive images of the test subject are represented using the dictionary, and the labels (i.e., identity and expression) of each test image

| Table 2. Confusion matrix for facial expression recognition using difference images and SRC (%) on the CK+ dataset |
|---|---|---|---|---|---|---|
|  | An | Di | Fe | Ha | Sa | Su |
| An | 86.7 | 8.9 | 0 | 0 | 2.2 | 2.2 |
| Di | 3.4 | 96.6 | 0 | 0 | 0 | 0 |
| Fe | 8.3 | 0 | 54.2 | 20.8 | 0 | 16.7 |
| Ha | 0 | 1.4 | 0 | 98.6 | 0 | 0 |
| Sa | 10.7 | 3.6 | 0 | 3.6 | 71.4 | 10.7 |
| Su | 0 | 0 | 0 | 1.2 | 0 | 98.8 |

are predicted using the joint and separate schemes. Using the joint scheme, Eq. 10, the face recognition rate is increased to 97.41% (about 7.5% improvement), but the facial expression recognition rate is reduced to 86.41% (about 4% worse). On the other hand, using the separate scheme, Eqs. 11 and 12, face recognition rate is 97.41%, and facial expression recognition rate is 91.59%.

The recognition rates using different approaches are summarized in Table 3. As can be observed, best accuracies for both facial expressions and identity recognitions are obtained when the simultaneous dictionary and the separate recognition scheme are used. Comparing the results of the two schemes show that improvement in the facial expression recognition is significant (we will see in the next experiments that improvement in the expression recognition is always higher than the identity recognition). The main reason for the different improvements is the different contribution of two attributes in the combined feature vector. To quantify the contribution of each attribute, let’s define the energy of an image as the sum of the squared values of its gray levels. By this definition, the energy of the neutral images is on average about 30 times more than the energy of the difference images. For this reason, the separate scheme will improve the expression recognition more than the identity recognition.

Another interesting fact is that the facial expression recognition rate by the separate scheme is higher when using difference images, although the neutral image of the test subject is not used directly. One reason for this improvement is the existence of some neutral images of each subject in the dictionary while in the difference image only one neutral image is used, and some small errors in that neutral image can cause misclassification.

### 3.3. Results on the MMI Dataset

In this section, the recognition rates on the MMI dataset are reported. Table 4 illustrates the recognition rates with different approaches. As can be observed, the trend of results in this table is the same as Table 3 (i.e., for the CK+ dataset) with higher face recognition rate and lower facial expression recognition rate. The number of subjects and sequences in the MMI dataset is lower than the CK+ dataset that can bias the results (i.e., improve the recognition rate for identity and degrade for expression). Furthermore, facial expression recognition on the MMI dataset is more challenging than the CK+ dataset.

To increase the number of training samples for each expression, we collect the images of both CK+ and MMI datasets as a new collection with 430 sequences from 134 subjects where Surprise has the maximum (108) and both Fear and Sadness have the minimum (40) number of sequences. The recognition rates on this collection are reported in Table 5 where the final recognition rates of both expression and identity are more than the weighted average of two datasets. This improvement indicates the effect of the number of training samples.

### 3.4. Comparing sparse solvers

In this section, we compare the performance of state-of-the-art sparse solvers in estimating the sparse solution of $A_{\text{sim}} b_{\text{sim}} = x$. The comparison metrics are the facial expression and identity recognition rates (in the separate scheme), and the execution times of the algorithms. We compare the well-known BP [4], OMP [25], and SL0 [22] sparse solvers. As presented in [21], using non-negative (NN) sparse decomposition is useful in the SRC-based approaches, especially when the classes are opposite of each other (such as difference images of some expressions). Therefore, we also compare the NN-BP [10], NN-OMP [26], and CSLO [21] (the non-negativity only applies to the atoms correspond to $A_{\text{id}}$ and $A_{\text{ex}}$ because the coefficients corresponding to the identity matrix are related to pixel corruptions and can be negative). The results of this experiment are reported in Table 6.

### 3.5. Facial Expression Recognition for Unseen Subjects

In the last experiment, we evaluate the proposed approach for facial expression recognition of unseen subjects.
We assume that the neutral images of the test subject are removed from the simultaneous dictionary. Then, the facial images of the test subject are represented sparsely on the dictionary and the recognition is done. Although face recognition is meaningless in this case, we can decide that the test subject is unknown based on the sparsity concentration index (SCI) proposed in [33]. However, it is expected that the facial expression recognition can perform with some performance reduction. Indeed, the neutral portion of the test facial image can be approximated by the first part of the dictionary and the expression can be recognized by the second part.

Using the CSL0 as the sparse solver, the result of FER on the CK+ dataset is 87.38% (7.12% reduction in comparison when the neutral images exist in the dictionary). If we use the joint scheme instead of the separate scheme, the FER is 58.90% (more than 28% lower). The reductions in the joint scheme instead of the separate scheme, the FER when the neutral images exist in the dictionary). If we use the second part.

Table 6. Recognition rates (%) of the proposed approach in the separate scheme using different sparse solvers on the CK+ dataset

<table>
<thead>
<tr>
<th>Sparse solver</th>
<th>FR (%)</th>
<th>FER (%)</th>
<th>Average Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>97.41</td>
<td>91.59</td>
<td>1.491</td>
</tr>
<tr>
<td>NN BP</td>
<td>97.41</td>
<td>92.56</td>
<td>1.838</td>
</tr>
<tr>
<td>OMP ((k = 20))</td>
<td>87.70</td>
<td>89.97</td>
<td>\textbf{0.012}</td>
</tr>
<tr>
<td>NN OMP ((k = 20))</td>
<td>89.00</td>
<td>91.59</td>
<td>0.024</td>
</tr>
<tr>
<td>SL0</td>
<td>96.44</td>
<td>91.59</td>
<td>0.143</td>
</tr>
<tr>
<td>CSL0</td>
<td>96.76</td>
<td>\textbf{94.50}</td>
<td>0.239</td>
</tr>
</tbody>
</table>

Using the CSL0 as the sparse solver, the result of FER on the CK+ dataset is 87.38% (7.12% reduction in comparison when the neutral images exist in the dictionary). If we use the joint scheme instead of the separate scheme, the FER is 58.90% (more than 28% lower). The reductions in the second part.

Using the CSL0 as the sparse solver, the result of FER on the CK+ dataset is 87.38% (7.12% reduction in comparison when the neutral images exist in the dictionary). If we use the joint scheme instead of the separate scheme, the FER is 58.90% (more than 28% lower). The reductions in the joint scheme instead of the separate scheme, the FER when the neutral images exist in the dictionary). If we use the second part.

3.6. Comparison with the previous works

Since the CK+ dataset is developed for expression analysis, the face identification results on it are less reported in the literature. Therefore, we only compare our facial expression recognition results with the previous works in Table 7. As can be found, usually the recognition rate of the dynamic algorithms are higher than the static ones. The proposed algorithm in this study was static, and its recognition rate is comparable with the best static algorithms.

4. Conclusion

We presented a novel approach for simultaneous recognition of two facial attributes, facial expression and identity, using sparse representation. We discussed that often two recognition tasks are conducted separately, though the feature representations utilized for describing the facial appearances are the same. Based on the theoretical discussions reviewed in the paper, simultaneous recognition of facial expression and identity outperform conducting each recognition task individually. Our experimental results also support the proposed theory. In the future, we will work on developing and utilizing supervised dictionary learning for simultaneous recognition of multiple attributes.

References


