Improving Facial Expression Analysis using Histograms of Log-Transformed Nonnegative Sparse Representation with a Spatial Pyramid Structure

Ping Liu Shizhong Han Yan Tong
Department of Computer Science and Engineering
University of South Carolina, Columbia, SC 29208
{liu264, han38, tongy}@email.sc.edu

Abstract— **Facial activity is the most direct signal for perceiving emotional states in people. Emotion analysis from facial displays has been attracted an increasing attention because of its wide applications from human-centered computing to neuropsychiatry. Recently, image representation based on sparse coding has shown promising results in facial expression recognition.**

In this paper, we introduce a novel image representation for facial expression analysis. Specifically, we propose to use the histograms of nonnegative sparse coded image features to represent a facial image. In order to capture fine appearance variations caused by facial expression, logarithmic transformation is further employed on each nonnegative sparse coded feature. In addition, the proposed Histograms of Log-Transformed Nonnegative Sparse Coding (HLNNSC) features are calculated and organized in a pyramid-like structure such that the spatial relationships among the features are captured and utilized to enhance the performance of facial expression recognition. Extensive experiments on the Cohn-Kanade database show that the proposed approach yields a significant improvement in facial expression recognition and outperforms the other sparse coding based baseline approaches. Furthermore, experimental results on the GEMEP-FERA2011 dataset demonstrate that the proposed approach is promising for recognition under less controlled and thus more challenging environment.

I. Introduction

Facial activity is the most powerful and natural means of emotion expression and perception. Emotion analysis from facial displays has attracted an increasing attention because of its wide applications such as human behavior analysis, human-centered computing, neuropsychiatry, and entertainment. Extensive efforts have been devoted to facial expression recognition from visual images and video data and great progress has been made over the years on automatic facial expression recognition ([27], [41], [32]). However, recognition performance suffers dramatically in real-world conditions with unconstrained pose and illumination, low resolution, and spontaneous facial displays as demonstrated in the most recent Facial Expression Recognition and Analysis (FERA2011) challenge [32].

Most existing approaches for facial expression analysis utilize various human hand-designed features extracted from videos/images including Local Binary Patterns (LBP) ([29], [32]), Histograms of Oriented Gradients (HOG) ([12], [6]), Haar wavelet [34], Gabor coefficients ([44], [43], [31], [2], [34]), and Scale-invariant feature transform (SIFT) descriptors [12]. These features are all designed deliberately with specific expert knowledge. Although these features can achieve promising performance when they have the desirable discrimination power in a specific application, their generalization between different applications is still questionable. To find the best feature set for a specific application, performance evaluation is conducted by enumerating all candidate feature sets on a testing dataset. It is not surprisingly one kind of these features need to be replaced by another one when a different testing dataset is used.

Unlike the human tuned feature representations, sparse coding technique, proposed by Olshausen and Field [26], is an unsupervised feature extraction method aiming to find an over-complete feature representation for the input data. Since the sparse coding representation is over-complete, it can capture a wide range of variations that are not targeted to a specific application. When it is employed in a specific application, a few sparse coding basis vectors that are related to the application will be selected according to the training data such that each sample data is represented by a linear combination of these selected basis vectors. Sparse representation has attracted increasing attention and shown promising results in the application of facial expression recognition ([40], [23], [38], [19], [22], [47]).

Most recently, Nonnegative Sparse Coding (NNSC) technique [11] has been developed to integrate the advantages of the sparse coding [26] and Nonnegative Matrix Factorization (NMF) [18]. By using NNSC, each image is represented by only "additions" of a few basic patterns. This nonnegative representation is consistent with human vision system, where the firing rate of the simple cell in the primary visual cortex is nonnegative [11]. More importantly, it is more natural to represent a face that is a combination of facial components (e.g., eyes, eyebrows, nose, and lip) with different shapes and appearance. There are a few early attempts ([3], [46], [39]) managing to take advantage of nonnegative representation for facial expression analysis.

In this work, we propose a novel sparse-coding based image representation for facial expression analysis. Inspired by the Bag-of-Features (BoF), we intend to utilize the statistics of the sparse coded image features to capture the facial appearance variations caused by facial expressions. Our proposed approach has four primary contributions.

First, we employ the NNSC technique to learn an over-

complete dictionary from a large number of local patches extracted from facial images without expression labels. In this work, we employ the Labeled Faces in the Wild (LFW) database [15] for constructing the NNSC dictionary. The LFW database has a wide range of variations in demographics, camera view points, and more importantly, spontaneous facial expressions as shown in Fig. 1. We believe that, the dictionary learned from LFW database is more comprehensive than from any publicly available facial expression database and is especially suitable to characterize spontaneous facial expression under uncontrolled environment.

Second, given a new input image, thousands of image patches can be extracted, each of which is represented by an NNSC coded image feature. A histogram, in this work, is calculated to represent the statistics of the NNSC coded image features for each facial image.

Third, most of the elements of an NNSC coded image feature, however, have very small values (around 10^{-4}). Our data analysis on facial images shows that over 50% nonzero elements are less than 0.1, while the full range is $(0, 6]$. As a result, the fine appearance variations cannot be characterized if the histograms are computed from the original NNSC coded features. In order to handle this issue, logarithmic transformation is further employed on each NNSC coded feature to enhance their discriminative ability.

Finally, motivated by the success of spatial pyramid matching in image classification [35], the proposed Histograms of Log-Transformed NNSC (HLNNSC) features are organized in a pyramid-like structure such that the spatial relationships among the features are captured and utilized to enhance the performance of facial expression recognition.

An overview of the proposed Spatial Pyramid structured HLNNSC (SP-HLNNSC) based approach is illustrated in Fig. 1. In order to evaluate the proposed approach, extensive experiments have been conducted on two well-known facial expression datasets. The results on the Cohn-Kanade database show that the proposed method yields a significant improvement and outperforms the other sparse coding based baseline approaches. Furthermore, the results on the GEMEP-FERA2011 dataset demonstrate that the proposed approach achieves a promising recognition performance under less controlled and thus more challenging environment.

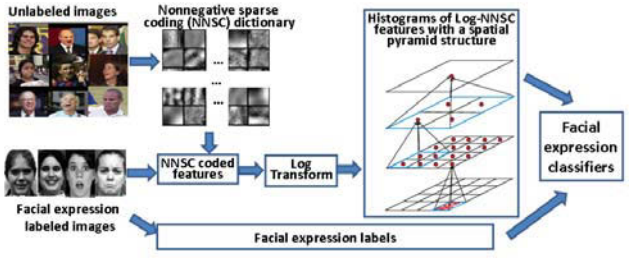

Fig. 1. An overview of facial expression recognition using histograms of log-transformed NNSC coded features with a pyramid-like structure.

II. Prior Work

Generally speaking, most of existing work on facial expression recognition are based on a two-stage training procedure. First, a set of features are extracted from the original input images/videos to characterize expression related facial appearance or geometry changes; and then a target facial expression is recognized by different recognition engines from the extracted features. According to the features employed, these approaches can be grouped into two classes: *holistic methods* ([2], [29], [34], [32], [36], [45]) extracting features from the whole face and *local methods* ([13], [44], [28], [31], [5], [9], [43], [17]) extracting features by detecting and tracking a small set of predefined feature points. Local approaches require to deliberately design special purposed features for each facial expression, respectively; and are sensitive to tracking errors. Holistic approaches, on the contrary, often employ general-purposed image features such as Gabor wavelet coefficients ([2], [34]), Haar wavelet ([34], [36]), and histograms of LBP features ([45], [32], [29]), where how these features are selected and utilized is determined by the recognition engine.

Besides these human tuned feature representations, sparse coding based feature representations ([40], [23], [38], [19], [22], [47], [3], [46], [39]) have been recently employed in facial expression analysis. Among them, sparse representations with nonnegative constraints ([3], [46], [39]) have shown promising results. Zhi et al [46] developed a Graph-Preserving Sparse NMF (GPSNMF) method. With the locality-preserving constraints, the GPSNMF achieves a better discriminant capability and is effective to handle the partial occlusions in the facial images. Zafeiriou and Petrou [39] proposed a Projected Gradient Kernel NMF (PGKNMF) method by combining the NMF with arbitrary positive definite kernels and obtained reasonable performance in the application of facial expression recognition. Bociu and Pitas [3] developed a Discriminant NMF (DNMF) method, where the sparse coding based image decomposition is performed in a supervised way by employing the class information in the cost function.

Our proposed approach differs from aforementioned sparse coding based methods in two major aspects. First, instead of using the sparse coded image features directly, we exploit the statistics of the sparse coded features. Second, the sparse coded features are organized into a pyramid-like structure to incorporate their spatial correlations.

III. Methodology

A. Feature Representation based on NNSC

In this work, the NNSC technique [10] is adopted to build an over-complete and discriminative feature representation, i.e., a dictionary $\mathbf{D} = [D_1, D_2, ...D_S]$ with S basis vectors, for input data. The input data can be whole images or image patches extracted from the region of interest. Specifically, we learn basic vectors from local patches extracted from facial images because the local patches are more appropriate to describe the facial images, which consist of multiple facial components such as eyes, nose, lip and eyebrows, and can capture a wide range of facial shape/appearance variations.

Given a learned dictionary $\mathbf{D}$, the information contained in each image patch P_i of a set of K patches $\mathbf{P} =$

$[P_1, P_2, \cdots, P_K]$ can be encoded by a linear combination of a few basis vectors with a sparse coefficient vector Z_i. Different from the traditional sparse coding technique [26], where the image patch P_i is represented as the addition and subtraction of basis vectors in the dictionary, only additive of basis vectors are permitted using NNSC. In other words, each element in Z_i is nonnegative. This nonnegative constraint is proved to be more compatible with the intuition of "come together to form a whole" and is especially suitable to represent a part-based object such as the face [18].

The NNSC based feature extraction in our work consists of two major steps: *dictionary construction*, also called *basis vectors learning* and *sparse coefficients estimation*. First, the *basis vectors learning* is performed by minimizing the reconstruction errors between the image patches $\mathbf{P}$ and the reconstructed data with an additional nonnegative constraint. The objective function for optimization is as follows:

$$\hat{\mathbf{Z}} = \underset{\mathbf{Z}}{\operatorname{argmin}} \|\mathbf{P} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda \|\mathbf{Z}\|_1, \quad (1)$$

where $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_K]$ is the set of optimized sparse coefficient vectors for all patches $\mathbf{P}$ such that each image patch in P_i can be reconstructed as $\hat{P}_i = \mathbf{D}Z_i$; and λ is a penalty parameter used to control the sparsity of $\mathbf{Z}$.

The optimization process is divided to two subproblems. One is to find the optimized sparse coefficient vectors $\mathbf{Z}$ given the current estimation of the dictionary $\mathbf{D}^t$ at the t^{th} iteration, which solution can be found as follows [10]:

$$\hat{\mathbf{Z}}^{t+1} = \mathbf{Z}^t[(\mathbf{D}^t)^T\mathbf{P}]./[(\mathbf{D}^t)^T\mathbf{D}^t + \lambda] \quad (2)$$

The other one is to find the best basis vectors to construct $\mathbf{D}$, while sparse coefficient vectors $\mathbf{Z}^t$ is fixed. Because of the nonnegative constraints, this subproblem is more complicated than that in the traditional sparse coding learning. Following the work by [10], we use projected gradient descent method to solve this subproblem as follows:

1. $\mathbf{D}'$= $\mathbf{D}^t - \lambda(\mathbf{D}^t\mathbf{Z}^t - \mathbf{P})(\mathbf{Z}^t)^T$.
2. Set the negative values in $\mathbf{D}'$ to zero.
3. Rescale each column of $\mathbf{D}'$ to a vector with unit L_2 norm, and then set $\mathbf{D}^{t+1} = \mathbf{D}'$.

Therefore, we can update $\mathbf{D}^t$ and $\mathbf{Z}^t$ alternatively with the other one fixed. After the dictionary $\mathbf{D}$ is constructed, we can estimate the sparse coefficient vector for a new input image patch. Following the same optimization process described in Eq.(2), we can find the best Z_i, which is satisfied with the nonnegative constraint and gives the minimum reconstruction error, with a fixed $\mathbf{D}$.

B. HLNNSC Coded Features

As discussed above, an NNSC coefficient vector Z_i is employed to represent a local image patch. However, facial expression analysis is usually performed on the whole face region such that the expression-related facial shape deformation and facial appearance variation can be fully captured. In this work, we randomly extract N local patches from each face region. By doing so, the information of each face can be sufficiently and concisely characterized by these local patches. As a result, each image I_j can be represented by $\tilde{Z}_j = [Z_{<1,j>}, \cdots, Z_{<N,j>}]$, where $Z_{<i,j>}$ is the NNSC coefficient vector corresponding to the i^{th} image patch P_i extracted from I_j.

We should note that the dimension of $\tilde{Z}_j$ $(N * S)$ is very large: usually above 1 million. In order to reduce the dimension of $\tilde{Z}_j$, we develop a new feature representation, i.e., the histograms of NNSC coefficients (HNNSC). Specifically, the HNNSC feature H_j for the image I_j is defined as $H_j = [H_{<1,j>}, \cdots, H_{<m,j>}, \cdots, H_{<S,j>}]$, where $H_{<m,j>}$ is estimated as

$$H_{<m,j>} = hist(Z_{<1,m,j>}, \cdots, Z_{<i,m,j>}, \cdots, Z_{<N,m,j>}) \quad (3)$$

where $Z_{<i,m,j>}$ is the m^{th} element of $Z_{<i,j>}$. Hence, $H_{<m,j>}$ is the histogram of the m^{th} element of the NNSC coefficient vectors corresponding to N image patches of I_j.

The HNNSC intends to model the distribution of the NNSC coded features on the whole image. However, most of the elements of an NNSC coded feature have very small values (around 10^{-4}). As an example, we use 100 facial images with different subjects and different expressions and extract 2000 image patches from each image. For each image patch, an NNSC coefficient vector with 4000 coefficients is calculated. Thus, we have $4000 * 2000 * 100$ NNSC coefficients in total, from which a statistic analysis is performed. Fig. 2(a) shows the distribution of all nonzero NNSC coefficients on these 100 images. As illustrated in Fig. 2(a), over 50% nonzero NNSC coefficients are less than 0.1, while the full range is $(0, 6]$. Therefore, the histogram computed from this distribution will be extremely unbalanced such that the sample data are concentrated in the bin with the smallest value. Such a histogram is not able to characterize the subtle facial appearance variations.

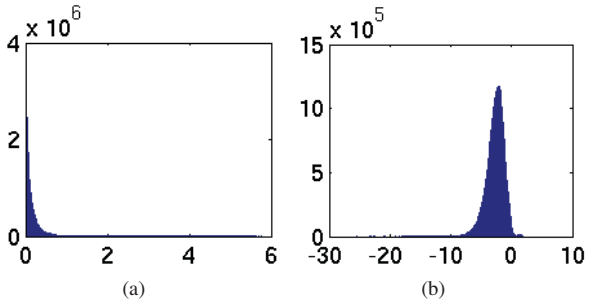

Fig. 2. (a) The distribution of nonzero NNSC coefficients, most of which are within the range of [0,0.1]. (b) The distribution of log-transformed nonzero NNSC coefficients.

In image processing, we know that logarithmic transformation is useful to highlight the details in the dark region. Motivated by this, we apply the logarithmic transformation on each nonzero NNSC coefficient to enhance its discriminative ability. As shown in Fig. 2(b), the distribution of the log-transformed nonzero NNSC coefficients now becomes a bell-shape. Over 50% transformed coefficients (corresponding to $(0, 0.1]$ before taking transformation) have been stretched in a range of [-10,-2].

C. HLNNSC with a Spatial Pyramid Structure

HLNNSC actually belongs to the category of *bag-of-feature* image representations, where the spatial information of the local patches are eliminated. However, the layout of different facial components is crucial to represent a face. The spatial relationships among the image patches encode important information of face deformation and facial appearance variations caused by facial expression changes.

Recently, a *Spatial Pyramid Matching* (SPM) approach, utilizing a spatial pyramid constructed feature representation, has been demonstrated to be effective on image classification problems [35]. Motivated by this, we extend our HLNNSC features to having a spatial pyramid structure.

Specifically, a three-layer spatial pyramid is constructed as illustrated in Fig. 3. At the l^{th} layer of the pyramid, the image is divided into $2^{2-l} \times 2^{2-l}$ cells. In this work, $l = 0$ corresponds to the lowest layer in the pyramid; while $l = 2$ corresponds to the highest layer. For each cell, an HLNNSC feature is calculated from all image patches whose centers are within the cell. The final Spatial Pyramid structured HLNNSC (SP-HLNNSC) feature is obtained by concatenating all HLNNSC features from all cells across three layers. By doing so, the lower level of the pyramid captures the locality information of the image patches and is sensitive to subtle facial expression changes. The higher level, on the contrary, reserves the advantages of the *bag-of-feature* methods and is not sensitive to the misalignment errors in face localization.

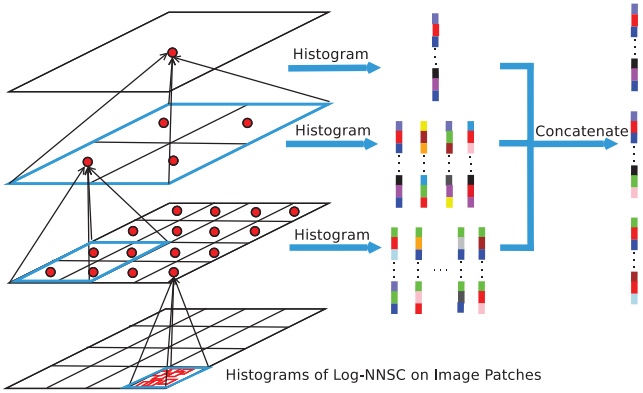

Fig. 3. An illustration of computing SP-HLNNSC features.

IV. Experiment

A. Image Databases and Experimental Setup

In order to demonstrate the effectiveness of the proposed SP-HLNNSC based facial expression recognition method, we have performed extensive validation studies on two well-known facial expression databases. The first database is the Cohn and Kanade's DFAT-504 database [16], which consists of more than 100 subjects and has been widely used for evaluating facial expression recognition system. The results on this database will be used to demonstrate the generalization capability of the proposed framework on a large population. The second database is the GEMEP-FERA2011 dataset [32], which has been used as the benchmark dataset for the FERA2011 challenge [32]. The facial expressions displayed in this database are more natural compared to the Cohn-Kanade database with considerable head movements involved. The results on the second database intend to demonstrate the robustness under less controlled environment and to facilitate a performance comparison with the state-of-the-art facial expression recognition techniques. In this work, the recognition performance is evaluated quantitatively in terms of classification rate ($\frac{\text{number of correctly recognized samples}}{\text{total number of samples}}$).

B. Weakly Supervised Dictionary Construction

Rather than learning the NNSC dictionary directly from the facial expression databases, i.e., the Cohn-Kanade database and the GEMEP-FERA2011 dataset, we employ the LFW database [15] for constructing the dictionary. This is because these facial expression databases have a small size in terms of subjects and expression categories. As a result, a new dictionary needs to be learned given a new testing set in order to capture the facial appearance variations in the unseen subjects and/or unseen expression categories. In practice, the testing images cannot be obtained in advance. Consequently, the generalization capability is often compromised.

Inspired by the recent advances in weakly-supervised learning ([42], [7], [21]), where unlabeled but relevant data has been used for improving the performance of a classifier, we would like to take advantage of the huge unlabeled data, i.e., facial images without expression labels. In particular, we use the LFW database to learn the NNSC dictionary. The LFW database has nearly 14000 images from over 5000 subjects and contains significant variations in background, demographics, face view angles, expressions, and illuminations as well as partial occlusions.

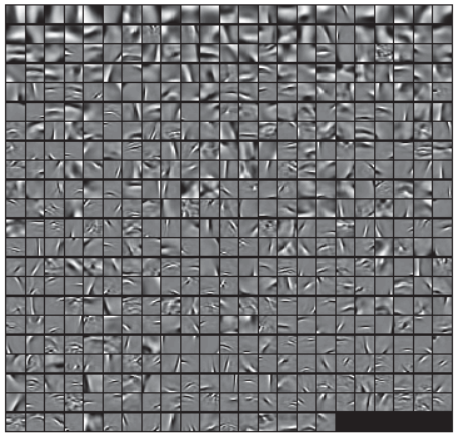

Fig. 4. Visualization of learned basis vectors.

For preprocessing purpose, each image in the LFW database has been funneled by image congealing [14] such that the face regions across different images have been aligned to remove the scale and positional variance. After congealing, the face regions are then cropped to 130×130. In this work, we randomly extract 300 image patches (14×14 square) from each cropped facial image in the LFW database. Hence, we have more than 4 million image patches for dictionary learning. We adapt SPArse Modeling Software (*SPAMS*) [24] to learn NNSC dictionary from these extracted

image patches. Some examples of learned basis vectors are visualized in Fig. 4. We can see that the learned basis vectors indeed capture basic patterns of the input images. For example, if we look closer, we can find that some of the basis vectors in the first row describe the major facial components such as nose. This learned NNSC dictionary has been employed in all experiments discussed below in spite of the facial expression database used for testing.

C. Performance Evaluation on Cohn-Kanade Database

We first evaluate the proposed SP-HLNNSC based method on the Cohn-Kanade database [16]. The images in Cohn-Kanade database were collected from around 100 subjects under a controlled environment without obvious head movements. The expression categories covered in Cohn-Kanade include neutral, anger, disgust, fear, happiness, sadness, and surprise. Since each subject has only a few expressions activated and labeled, we have 204 image sequences with expression labels in total. For each image sequence, there is only one expression label provided, which corresponds to the last frame (the peak frame). Therefore, we select the last three frames for training/testing purpose from each image sequence. In addition, we also select one image without any expression for each subject and use it as a neutral expression. By doing so, we build an experimental dataset named *CK-DB* with a total of 613 images. The one-versus-all classification strategy is employed for this multiclass classification problem. When we recognize an expression, the positive samples are selected as the images with the expression occurring; and the other images are used as the negative samples.

For each image in the CK-DB, the face region is cropped and normalized to 150×130 based on the eye positions provided in the database. For each cropped facial image, we randomly extract 2000 image patches, from which the SP-HLNNSC feature is calculated using the NNSC dictionary learned on the LFW database.

In order to evaluate the performance for generalization to novel subjects, the CK-DB is divided into 8 subsets, where the subjects in any two subsets are non-overlapped. For each run, we use one subset for testing and the remaining 7 subsets for training the classifiers. We will perform such 8 runs by enumerating the subset used for testing; and the recognition performance is computed as the average of the 8 runs. In this experiment, the classification rate for each expression is computed based on a per-image detection since we have expression labels for each image in the *CK-DB*.

In this experiment, we intend to evaluate the effectiveness of using the proposed SP-HLNNSC based method. Specifically, we would like to compare the recognition performance of the proposed method with a set of baseline methods including a) histograms of sparse coded features without nonnegative constraints (HSC) computed from the whole image, b) HNNSC features computed from the whole image, and c) a spatial pyramid of HNNSC features (SP-HNNSC).

For each image in the CK-DB, we calculate different types of features including the proposed SP-HLNNSC and the features listed above. For all methods in comparison, we use AdaBoost classifiers for classification given a type of features. For a fair comparison, we ensure all the methods are compared under the same condition: using the same data for training/testing.

TABLE I
PERFORMANCE COMPARISON ON THE CK-DB.

Expression	HSC	HNNSC	SP-HNNSC	SP-HLNNSC (Proposed)
Neutral	0.87	0.86	0.93	0.94
Anger	0.78	0.78	0.89	0.85
Disgust	0.77	0.88	0.95	0.94
Fear	0.75	0.88	0.93	0.94
Happiness	0.88	0.93	0.97	0.98
Sadness	0.81	0.73	0.77	0.95
Surprise	0.9	0.94	0.9	0.99
AVG	0.82	0.86	0.91	0.94

In Table I, we compare the recognition performance using the proposed method and the baseline methods for the 7 emotion categories in terms of classification rate. From Table I, we can find that the proposed SP-HLNNSC based method achieves the best recognition performance among all the methods in comparison in terms of the average classification rate of the 7 emotion categories (0.94). Specifically, by using the nonnegative constraints (HNNSC), the average classification rate was improved by 0.04 compared to the HSC based method. The pyramid-like structure (SP-HNNSC) further improved the average classification rate by 0.05 compared to the HNNSC based method. Finally, the proposed SP-HLNNSC based method outperformed the SP-HNNSC based method by 0.03. This clearly demonstrates that the log-transformed sparse features are more effective to describe the subtle facial appearance changes.

Furthermore, we also compare the proposed SP-HLNNSC based method with other state-of-the-art NNSC based methods that were evaluated on the Cohn-Kanade database including PGKNMF method by Zafeiriou and Petrou [39] and GPSNMF method by Zhi et al [46]. In this work, we use the experimental results reported in their papers. The PGKNMF method [39] and the GPSNMF method [46] were evaluated on the Cohn-Kanade database using a subset of 13 subjects and 30 subjects, respectively. In contrast, we use all 82 subjects that have at least one expression labeled.

Table II shows the performance comparison in terms of average classification rate for 6 emotion categories (anger, disgust, fear, happiness, sadness, and surprise). The proposed SP-HLNNSC based approach outperformed the PGKNMF method significantly by 0.1. Although the average classification rate of the proposed method is almost the same as that of GPSNMF method, the GPSNMF method was evaluated in a relative easier experimental setup, where the same image sequence has been divided into training and testing sets.

D. Performance Evaluation on GEMEP-FERA2011 Dataset

It is more desirable to recognize facial expressions when the subjects express their expressions naturally with free head motions. This holds true in the database of GEMEP-FERA2011 [32], where subject sometimes undergoes large

TABLE II

PERFORMANCE COMPARISON ON THE COHN-KANADE DATABASE IN TERMS OF AVERAGE CLASSIFICATION RATE FOR 6 EMOTIONS.

PGKNMF [39]	GPSNMF [46]	SP-HLNNSC (Proposed)
0.842	0.943	0.941

TABLE III

RECOGNITION PERFORMANCE ON GEMEP-FERA2011 DATASET [32].

Expression	Person-Independent	Person-Specific	Overall
Anger	0.643	0.846	0.741
Fear	0.333	1.00	0.600
Joy	1.00	1.00	1.00
Relief	0.688	0.900	0.769
Sadness	0.667	1.00	0.800
AVG	0.666	0.949	0.782

head movements. In GEMEP-FERA2011 database, videos are collected from 10 subjects and are divided into two groups: one is the training set and the other one is the testing set. Each video has a single expression label available to the users and thus every frame in the same video shares the same expression label. There are 5 expression categories in the database, i.e., anger, fear, joy, relief and sadness. The training set consists of 7 subjects with 155 videos. The testing set consists of 6 subjects with 134 videos, which cover the same expression categories as the training set. The expression labels for the testing set are blind to the users for a fair comparison. Half of the subjects in the testing set are not present in the training set.

For GEMEP-FERA2011 database, we use an eye detector to detect the positions of the eyes [33], based on which each frame of the training/testing videos is preprocessed following the same way as used for the CK-DB. Then, the SP-HLNNSC feature for each frame of the videos is calculated from 2000 random sampled image patches and used for facial expression recognition.

Similar to the experiments on the CK-DB, we use a one-versus-all classification strategy and train an AdaBoost classifier for each expression category. For training purpose, the positive samples of an expression are selected as all the frames of the videos that have the target expression; while all the frames in the remaining videos are used as the negative samples. During the testing process, we first estimate the expression label for each frame of a testing video. Then, a majority voting is utilized to obtain the expression label for each testing video.

The per-video recognition performance of the proposed SP-HLNNSC method is reported in Table III in terms of classification rate. In Table III, the performance of *person-specific* test, *person-independent* test, and *overall* test are reported. In the *person-independent* test, the test subjects are not present in the training set; while in the *person-dependent* test, the test subjects appear in the training set also. The *overall* test uses all testing videos.

We further compare the performance of the proposed method with other published results in FERA2011 Challenge in Table IV. We can see that our proposed method is ranked

TABLE IV

PERFORMANCE COMPARISON ON GEMEP-FERA2011 DATABASE IN TERMS OF AVERAGE CLASSIFICATION RATE [32].

TEAM	Person-Independent	Person-Specific	Overall
UC Riverside [37]	0.752	0.962	0.838
UIUC-UMC [30]	0.655	1.00	0.798
Proposed method	0.666	0.949	0.782
KIT [8]	0.658	0.944	0.773
UCSD-CERT [20]	0.714	0.837	0.761
UCLIC [25]	0.609	0.837	0.700
U. Montreal [6]	0.579	0.870	0.700
Queensland Univ. of Tech. [4]	0.624	0.554	0.600
Baseline [32]	0.440	0.730	0.560
MIT-Cambridge [1]	0.448	0.433	0.440

at the 3^{rd} place for *person-independent*, *person-specific*, and *overall* tests. Note that the result of the person-specific test is much higher than that of the person-independent test. It is because the GEMEP-FERA2011 dataset contains only 10 subjects, among which only 7 subjects appear in the training set. This observation implies that the expression recognition for a registered user is more reliable and accurate and it remains challenging to generalize a trained system to unseen users.

Although the performance of the proposed method is lower than [37] and [30], our method is much more easier and direct to implement. Both [37] and [30] need to empirically choose and tune specific human designed features such as SIFT, LBP, and LPQ employed individually or in a combination, while our work automatically learns and designs the features in an unsupervised way. In addition, the proposed method does not rely on motion information as [30], and thus is particularly suitable for expression recognition from a single image.

V. CONCLUSION AND FUTURE WORK

Facial expression recognition is challenging because of subtle and complex facial deformations, and changes in view point and illumination. It is of extreme importance to develop image features that are capable of capturing subtle facial appearance changes caused by facial expressions. In this paper, we propose a novel image representation by exploiting the statistics of sparse coded image features. Specifically, we first learn the NNSC dictionary from a large number of facial images without expression labels. Then, the histograms of NNSC coded image features are employed to represent image patches extracted from facial images. Logarithmic transformation is further applied on each NNSC coded feature to enhance its discriminative ability. In order to characterize the spatial relationships among the features, a pyramid-like structure formed by the proposed HLNNSC features is employed.

Extensive experiments on two well-known facial expression databases (Cohn-Kanade database and GEMEP-FERA2011 dataset) demonstrate that our proposed SP-HLNNSC feature outperforms the other sparse coding based image features in comparison. Furthermore, the proposed method also shows promise for expression recognition under

less controlled environment with significant head movements. Since the proposed SP-HLNNSC feature does not limit to the application of facial expression recognition, we plan to extend this work to other face-related classification problems such as biometrics and recognition of micro-expressions.

VI. Acknowledgments

This work was supported by National Science Foundation under CAREER Award IIS-1149787.

References

[1] T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. E. Kaliouby, P. Robinson, and R. Picard. Real-time inference of mental states from facial expressions and upper body gestures. In *FG*, pages 909–914, 2011.

[2] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *CVPR*, volume 2, pages 568–573, 2005.

[3] I. Bociu and I. Pitas. A new sparse image representation algorithm applied to facial expression recognition. In *Proc. IEEE Int'l Workshop on Machine Learning for Signal Processing*, pages 539–548, 2004.

[4] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FG*, pages 915–920, 2011.

[5] J. F. Cohn, L. I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *IEEE SMC*, volume 1, pages 610–616, 2004.

[6] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based HoG features. In *FG*, March 2011.

[7] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *CVPR*, pages 1–8, 2007.

[8] T. Gehrig and H. Ekenel. A common framework for real-time emotion recognition and facial action unit detection. In *CVPR Workshops*, pages 1–6, 2011.

[9] H. Gu and Q. Ji. Facial event classification with task oriented dynamic Bayesian network. In *CVPR*, volume 2, pages 870–875, 2004.

[10] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

[11] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Machine Learning Research*, 5:1457–1469, 2004.

[12] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang. Multi-view facial expression recognition. In *FG*, pages 1–6, 2008.

[13] C. Huang and Y. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. *J. Visual Communication and Image Representation*, 8(3):278–290, 1997.

[14] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[16] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *FG*, pages 46–53, 2000.

[17] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Trans. on PAMI*, 32(11):1940–1954, Nov. 2010.

[18] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] Y. Lin, M. Song, D. Quynh, Y. He, and C. Chen. Sparse coding for flexible, robust 3d facial-expression synthesis. *Computer Graphics and Applications*, 32(2):76–88, 2012.

[20] G. Littlewort, J. Whitehill, T. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett. The motion in emotion – A CERT based approach to the FERA emotion challenge. In *FG*, pages 897–902, 2011.

[21] J. Liu, K. Yu, Y. Zhang, and Y. Huang. Training conditional random fields using transfer learning for gesture recognition. In *ICDM*, pages 314–323, 2010.

[22] W. Liu, C. Song, and Y. Wang. Facial expression recognition based on discriminative dictionary learning. In *ICPR*, 2012.

[23] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *FG*, pages 336–342. IEEE, 2011.

[24] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Machine Learning Research*, 11:19–60, 2010.

[25] H. Meng, B. Romera-Paredes, and N. Berthouze. Emotion recognition by two view SVM-2K classifier on dynamic facial expression features. In *FG*, pages 854–859, 2011.

[26] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[27] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang. Human computing and machine understanding of human behavior: A survey. In T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, editors, *Artificial Intelligence for Human Computing*, Lecture Notes in Artificial Intelligence. Springer Verlag, London, 2007.

[28] M. Pantic, L. J. M. Rothkrantz, and H. Koppelaar. Automation of non-verbal communication of facial expressions. In *Proc. of Conf. Euromedia*, pages 86–93, 1998.

[29] T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Combining LGBP histograms with AAM coefficients in the multi-kernel SVM framework to detect facial action units. In *FG*, March 2011.

[30] U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. Huang, X. Lv, and T. Han. Emotion recognition from an ensemble of features. In *FG*, pages 872–877, 2011.

[31] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *FG*, pages 229–234, May 2002.

[32] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analyis of the first facial expression recognition challenge. *IEEE Trans. on SMC-Part B: Cybernetics*, 42(4):966–979, 2012.

[33] P. Wang and Q. Ji. Multi-view face and eye detection using discriminant features. *CVIU*, 105(2):99–111, February 2007.

[34] J. Whitehill, M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Towards practical smile detection. *IEEE Trans. on PAMI*, 31(11):2106–2111, Nov. 2009.

[35] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009.

[36] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *CVPR*, pages 1–6, June 2007.

[37] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *FG*, pages 866–871, 2011.

[38] Z. Ying, Z. Wang, and M. Huang. Facial expression recognition based on fusion of sparse representation. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, volume 6216/2010 of *Lecture Notes in Computer Science*.

[39] S. Zafeiriou and M. Petrou. Nonlinear non-negative component analysis algorithms. *IEEE Trans. on IP*, 19(4):1050–1066, 2010.

[40] S. Zafeiriou and M. Petrou. Sparse representations for facial expressions recognition via L1 optimization. In *CVPR Workshops*, pages 32–39, 2010.

[41] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on PAMI*, 31(1):39–58, Jan. 2009.

[42] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted HMMs for unusual event detection. In *CVPR*, volume 1, pages 611–618, 2005.

[43] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on PAMI*, 27(5):699–714, May 2005.

[44] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *FG*, pages 454–459, 1998.

[45] G. Zhao and M. Pietiäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI*, 29(6):915–928, June 2007.

[46] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Trans. on SMC-Part B: Cybernetics*, (99):1–15, 2010.

[47] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.