Joint Sparse and Low-Rank Representation for Emotion Recognition

Xiang Xiang, Fabian Prada, Hao Jiang

Abstract— **We present an approach to emotion recognition applying a join sparse and low rank decomposition. Previous sparse-based approaches to this problem require of explicitly provided neutral faces to assist the recognition task. Our model trade the neutral face requirement by a sound prior: neutral face can be capture as the low rank component of frame sequence with moderate variation. Satisfactory recognition rates were attained on 7 different emotions.**

I. Introduction

In this work we explore the task of classifying the emotional state of a subject provided an image or video sequence of its face. In order to get a solid formulation of the emotion recognition problem, we must first look at a more fundamental definition: Action Units[1]. An AU is a single gesture: raising eyebrows, opening mouth, closing eyes, etc. Usually AU's do not occur independently one of each other, for instance, frowning brow and stretching the mouth tend to come altogether. Based on these interdependence, we can define an emotion as a collection of AU's that may ocurr simultaneously or alternately.

Emotion recognition is currently one of the main challenges for effective Human Machine Interaction. The innate ability to identify emotional states on other people plays a crucial role on human communication: people care not only on the context of the message but also on its context. Human communication have been transformed by the technological developments of the last decades, devices capable of registering and transmitting video and voice on real time. On the next years, we will probably observe how these technologies incorporate emotion recognition to provide a more accurate feedback from the machine itself.

Fig. 1. An example of an expressive faces separability. Left: the last frame of fear face sequence, with slightly opened mouth, nearly closed eyes, tightened eyebrows, facial muscles and wrinkles. Middle: the neutral face with closed mouth, open eyes, flat eyebrows and facial muscles. Right: the expression with highlighted mouth, eyes and eyebrows.

Zeng et. al. [21] provide a motivational work to emotion recognition in visual and audio signals. In the specific context of facial emotion recognition Bettadapura [23] presents a relevant survey describing the different problem formulations, approaches and databases on the area. Approaches to facial emotion recognition using sparse representation were previously proposed by Zafeiriou et. al. [22] and Taheri [9] et. al. Zafeiriou et. al. proposes a sparse representation classification approach to emotion recognition. Training and testing samples in their work are obtained by explicit subtraction of expression and neutral faces, and posterior dimensionality reduction. The work of Taheri looks for join recognition of emotion and identity. In Taheri's work neutral face of the testing sample is also explicitly available in a neutral face dictionary. Both works strongly assist the recognition task using the neutral face of the testing subject. Our aim is to weak or disregard this assumption.

In Sections II and III we present the formulation of our model for video sequences and single frames. Section IV is devoted to describe the experiments that support the use of sparse representation for emotion recognition. Section V present a discussion on the results. Conclusions and future directions of work are described in Section VI and VII.

The authors are with the Departments of Computer Science and Applied Mathematics, The Johns Hopkins University, Baltimore, MD 21218. Email: {xxiang,fpradan1,hjiang13}@jhu.edu.

II. Joint Sparse and Low-rank On Video

A. Problem and Model Formulation

The problem we are trying to study is classifying facial emotions in videos. More specifically, we work in the following setup. Suppose we are given a video of a human face in which the face changes from neutral to showing certain emotion. We want to classify that facial emotion. We will consider 7 different emotions: *Angry, Contempt, Disgust, Fear, Happiness, Sadness* and *Surprise.*

We use Y to denote our input video, where the i-th column of Y is a vector that represents the grey-scale image at the i-th frame of the video. Since a face image can be viewed as a superposition of a neutral face component and an expression component, the video matrix Y can be decomposed into a low-rank matrix L whose columns represent neutral faces and a matrix Z which contains information of the expression. We also assume that we have a dictionary matrix A of facial expressions. And therefore, the matrix Z can be written as a sparse representation by $Z = AX$ where X is a sparse matrix. To summarize, an input video Y can be written as

$$Y = AX + L.$$

Our goal is to classify the expression using this decomposition. To this end, we propose the following optimization

framework.

$$\begin{aligned} &\min ||X||_1 + \alpha||L||_* \\ &\text{subject to } Y = AX + L. \end{aligned} \tag{1}$$

The rationale behind using ℓ_1 and nuclear norms is to try to recover the sparse matrix X and the low-rank matrix L respectively.

It is important to notice that the novelty of our approach is on computing the sparse representation and low rank term in a single optimization problem. Previous sparse representation models proceed in two steps: first, they solve for a low rank plus sparse decomposition (e.g, RPCA), then, the sparse component is classified from a dictionary (e.g., SRC). By solving a common optimization problem, our model forces the term represented by the dictionary and the low rank component to complement each other. As we will describe in the next section, classification does not require solving another optimization problem but a direct metric comparison.

For robustness, a sparse error term may be added to the formulation in (1) in order to correct for sparse variation on the testing emotion that may not be expressed by the dictionary. Thus, a robust formulation our model is provided by,

$$\begin{aligned} &\min ||X||_1 + \alpha||L||_* + \gamma||E||_1 \\ &\text{subject to } Y = AX + L + E. \end{aligned} \tag{2}$$

B. Algorithm

Suppose our dictionary contains K types of facial expressions, we want to classify a new test video Y as one of these K classes. Given Y, we first decompose it into the sparse-representation component AX, low-rank component L, and error term E by solving (2). For each class i, let $A^{(i)}$ denote the submatrix of A which consists of all the columns of A that correspond to expression class i. Similarly, let $X^{(i)}$ be the submatrix of X which consists of all the rows of X that correspond to expression class i. We then classify Y by assigning it to the expression class that minimizes the residual as follows:

$$r_i(Y) := ||Y - A^{(i)}X^{(i)} - L - E||_F \tag{3}$$

Algorithm 1 summarizes the complete classification procedure. We will elaborate on how we solve (2) in the next section.

Algorithm 1 Video-Based JS&LR Algorithm

1: **Input**: a dictionary matrix A, a test video Y.
2: Normalize the columns of A and Y to have unit ℓ_2 norm.
3: Solve the optimization problem (2).
4: Compute the residuals $r_i(Y)$ defined in (3)
5: **Output**: class of $Y = \text{argmin}_i\, r_i(Y)$.

C. Optimization Solution

We solve the convex optimization problem (2) using the alternating direction method of multipliers (ADMM). Following a standard ADMM procedure, we can write down the augmented Lagrangian function as

$$\begin{aligned} L(X, L, E, \Lambda) = &||X||_1 + \alpha||L||_* + \gamma||E||_1 \\ &+ \langle \Lambda, Y - AX - L - E\rangle + \frac{\beta}{2}||Y - AX - L - E||_F^2 \end{aligned} \tag{4}$$

where Λ is the matrix of multipliers and β is the positive penalty parameter. An update at the k-th iteration is:

$$L_{k+1} = \underset{L}{\text{argmin}}\, \alpha||L||_* + \frac{\beta}{2\alpha}||Y - AX_k - L - E + \frac{1}{\beta}\Lambda_k||_F^2 \tag{5}$$

$$X_{k+1} = \underset{X}{\text{argmin}}\, ||X||_1 + \frac{\beta}{2}||Y - AX - L_{k+1} - E + \frac{1}{\beta}\Lambda_k||_F^2 \tag{6}$$

$$E_{k+1} = \underset{E}{\text{argmin}}\, ||E||_1 + \frac{\beta}{2\gamma}||Y - AX - L_{k+1} - E + \frac{1}{\beta}\Lambda_k||_F^2 \tag{7}$$

$$\Lambda_{k+1} = \Lambda_k + \beta(Y - AX_{k+1} - L_{k+1} - E_{k+1}). \tag{8}$$

Steps (5) and (7) has closed-form solutions,

$$L_{k+1} = D(\frac{\alpha}{\beta}, (Y - AX_k - E + \frac{1}{\beta}\Lambda_k)) \tag{9}$$

$$E_{k+1} = S(\frac{\gamma}{\beta}, (Y - AX_k - L + \frac{1}{\beta}\Lambda_k)) \tag{10}$$

where D and S are shrinkage thresholding operator.

Step (6) is the well-known LASSO problem. This is was solved using an iterative thresholding algorithm[19].

III. Joint Sparse and Low-rank On Single Frame

In the previous section we describe our model in the case of a video sequence with moderate facial expression variation. In this section we explore a model for a more basic instance of the emotion recognition problem: only a single image provided.

In our model for video sequences, we attempt to attain the neutral and expression separation by capturing the neutral component as a the low rank of the sequence. When just a single frame is provided we will use the neutral faces of the training set to get an approximation of the unavailable neutral face of the testing image. Our assumption is that the testing image can be decomposed in a component that is sparsely represented by the emotion dictionary and a component that is low rank respect to a neutral faces matrix.

Given a neutral faces matrix M_c (each column is a neutral face in vector form), an emotions dictionary A and a testing image y, the optimization problem we solve for this context is given by,

$$\min ||x||_1 + \alpha||M||_* \quad (11)$$
subject to:
$$y = Ax + m.$$
$$M = [M_c|m]$$

Since we enforce m to be the component of the test image that (1) minimizes nuclear norm of the extended matrix of neutral faces and (2) is complemented by a sparse span of elements in the emotion dictionary, we expect m to be a reasonable approximation of the neutral face. As we will see in the experimental section this situation is satisfied in some of our results.

Robustness is also gained by introducing a sparse term to compensate for expression or facial detail that may not be captured by the dictionary. The complete model is provided by,

$$\min ||x||_1 + \alpha||M||_* + \gamma||e||_1 \quad (12)$$
subject to:
$$y = Ax + m + e.$$
$$M = [M_c|m]$$

A. Algorithm

Classification is computed in analogous way as to our video-based model. The residual for class i is now provided by:

$$r_i(Y) := ||y - A^{(i)}x^{(i)} - m - e||_F \quad (13)$$

For completeness, Algorithm 2 summarizes the entire classification procedure.

Algorithm 2 Frame-Based JS&LR Algorithm

1: **Input**: a emotion dictionary matrix A, a neutral faces matrix M_c, a test image y.
2: Normalize the columns of A, M_c and y to have unit ℓ_2 norm.
3: Solve the optimization problem (12).
4: Compute the residuals $r_i(Y)$ defined in (13)
5: **Output**: class of y is $\text{argmin}_i\, r_i(Y)$.

B. Optimization Solution

Solution to (12) is also computed using an ADMM procedure. The augmented Lagrangian formulation is now,

$$L(x, m, e, M, \Lambda^1, \Lambda^2) = ||x||_1 + \alpha||M||_* + \gamma||e||_1 \quad (14)$$
$$+ \langle \Lambda^1, y - Ax - m - e\rangle + \frac{\beta_1}{2}||y - Ax - m - e||_F^2$$
$$+ \langle \Lambda^2, M - [M_c|m]\rangle + \frac{\beta_2}{2}||M - [M_c|m]||_F^2$$

This problem is iteratively solved by computing,

$$M_{k+1} = \underset{M}{\text{argmin}}\, \alpha||M||_* + \frac{\beta_2}{2\alpha}||M - [M_c|m_k] + \frac{1}{\beta_2}\Lambda_k^2||_F^2 \quad (15)$$

$$x_1 = \underset{X}{\text{argmin}}\, ||x||_1 + \frac{\beta_1}{2}||y - Ax - m_k - e_k + \frac{1}{\beta}\Lambda_k^1||_F^2 \quad (16)$$

$$e_{k+1} = \underset{E}{\text{argmin}}\, ||e||_1 + \frac{\beta}{2\gamma}||y - Ax_{k+1} - m_k - e + \frac{1}{\beta}\Lambda_k^1||_F^2 \quad (17)$$

$$m_{k+1} = \underset{m}{\text{argmin}}\, \frac{\beta_1}{2}||y - Ax - m_k - e_k + \frac{1}{\beta}\Lambda_k^1||_F^2$$
$$+ \frac{\beta_2}{2\alpha}||M - [M_c|m_k] + \frac{1}{\beta_2}\Lambda_k^2||_F^2 \quad (18)$$

$$\Lambda_{k+1}^1 = \Lambda_k^1 + \beta_1(y - Ax_{k+1} - m_{k+1} - e_{k+1}). \quad (19)$$

$$\Lambda_{k+1}^2 = \Lambda_k^2 + \beta_2(M - [M_c|m]). \quad (20)$$

Step 16 is the only that does not admit a close computation. We use an iterative Lasso Solver[19] for its computation.

IV. Experiments

In this section we present a set of experiments we conducted to validate our models. In all our experiments the expression features captured by the dictionary correspond to the simple subtraction between an image of a person showing an expression and an image of the same person in neutral state (see the right picture in Figure 1). Our results confirm good recognition rates for certain emotions using this basic dictionary.

The data used in these experiments is provided by the Extended Cohn-Kanade (CK+) [1] database. We obtain our data samples from video sequences in the database whose label is already provided and are composed by at least 8 frames. This corresponds to 45,18,59,25, 69,28, and 83 samples (videos or images according to the experiment context) of the respective expression (in alphabetic order). Some of these samples may correspond to a same person under different emotions.

Due to the few samples for some of the emotion classes, and to keep a balance on the number of training examples, we choose 10 training and 5 testing samples per category to compute recognition rates. We repeat these experiments 5 times and take their average to reduce variance. These are the values presented in the results below.

The Augmented Lagrangian Method (ALM) we use on solving the optimization problems is motivated from the Lagrangian multipliers method and penalty methods. We call β the penalty coefficient. This coefficient can change over iterations while we prefer fixing it as a constant in practice. It is guaranteed that X converges to the optimal provided that Λ_k is a bounded sequence and β_k is always larger than the optimal [2]. However, increasing β_k increases the ill-conditioness or difficulty of minimizing the Lagrangian [2]. In practice, we alleviate the difficulty

	Ang.	Cont.	Dis.	Fear	Hap.	Sad.	Sur.
Ang.	**0.74**	0.01	0.08	0.01	0.02	0	0.14
Cont.	0.09	**0.56**	0.03	0	0.19	0.03	0.11
Dis.	0.01	0	**0.95**	0.01	0.02	0	0.01
Fear	0.07	0	0.11	**0.24**	0.26	0.01	0.31
Hap.	0	0	0	0.01	**0.96**	0	0.03
Sad.	0.25	0	0.13	0.01	0.02	**0.25**	0.34
Sur.	0	0.01	0	0	0.01	0	**0.98**

TABLE I
CONFUSION MATRIX FOR THE EMOTION RECOGNITION EXPERIMENT USING SPARSE REPRESENTATION CLASSIFICATION (SRC). GLOBAL RECOGNITION RATE IS 79%.

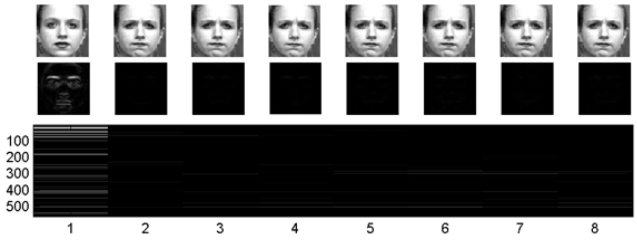

Fig. 2. *Tail samples.* First row show the frames in the testing sequence. Second row shows the value of AX provided by the solution to the optimization problem 2. The graph below shows the distribution of entries of coefficient matrix X. Results obtained using $\alpha = 10$ and $\gamma = 1$

by fixing β to be a reasonably large constant. We choose $\beta = 20/\text{ Avg}(|Y_{ij}|)$. The choose of β also affects the convergence rate, which in theory is linear if β_k is always larger than the optimal value and is even super linear if β_k goes to infinity.

A. Validation of Sparse Representation

Our models are build on the assumption that an emotion can be sparsely represented using a simple expression-minus-neutral dictionary. In this experiment we want to show that this is a correct assumption.

The first and last frame of each CK+ video sequence correspond to neutral and emotional state respectively . Thus, extraction of the emotion feature for our training and testing samples is done by simply subtracting the first frame to the last frame of their respective sequences.

The validation of our assumption exactly follows the sparse representation based classification (SRC) [3] approach. Firts, we compute a sparse vector to represent the testing sample from our dictionary elements. Then, we utilize nearest neighbor (minimum distance) as classifier. On Table I we present the confussion matrix of emotion recognition using this approach. The overall recognition rate is 79%. Similar recognition rates were obtained for Eigenface methods on emotion recognition for this setup: 72% on Eigenface with Nearest Neihbour classifier and 80% on Eigenface with Nearest Subspace classifier.

B. Join Sparse and Low Rank in Video

To test the video-based model we define the dictionary by subtracting the initial (neutral) frame of each training sample to some of its lasts frames (in our experiments we took the last 8 frames as those that capture the emotion). This allowed us to attain larger variation on each emotion, than by just adding a single dictionary element per trainning sample.

We evaluate two different kind of testing samples: *tail samples* and *equally spaced samples*.

Tail samples are frame sets conformed by the initial frame and some of the last frames of the video sequence. *Tail samples* are mostly dominated by frames that exhibit a completely formed emotion, as observed in Figure 2.

	Ang.	Cont.	Dis.	Fear	Hap.	Sad.	Sur.
Ang.	**0.84**	0	0.12	0	0	0	0.04
Cont.	0.04	**0.56**	0	0	0.16	0.12	0.12
Dis.	0.12	0	**0.84**	0	0.04	0	0
Fear	0.12	0	0	**0.56**	0.2	0.12	0
Hap.	0	0	0	0	**1**	0	0
Sad.	0.28	0	0.12	0	0	**0.52**	0.08
Sur.	0.04	0.04	0	0	0	0	**0.92**

TABLE II
CONFUSION MATRIX FOR *Tail sample* EXPERIMENT. RESULTS OBTAINED USING $\alpha = 10$ AND $\gamma = 1$. GLOBAL RECOGNITION RATE IS 74%.

Equally spaced samples are frame sets taken at approximately equally time separation along the video. As observed from Figure 3, there is no dominating expression on these frame sets.

C. Join Sparse and Low Rank in Single Frames

For the single frame model, our dictionary is formed by taking one element per training sequence : the subtraction of the last frame (expression) from the first frame (neutral) of the sequence. We choose 10 neutral images at random from the entire database and used them as the columns of matrix M_c.

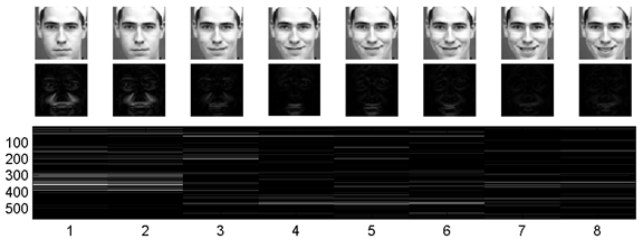

Fig. 3. *Equally spaced samples.*Results obtained using $\alpha = 15$ and $\gamma = 1$.

	Ang.	Cont.	Dis.	Fear	Hap.	Sad.	Sur.
Ang.	**0.68**	0.04	0.04	0	0	0.12	0.12
Cont.	0.04	**0.44**	0.04	0	0.12	0.32	0.04
Dis.	0	0	**0.84**	0.04	0	0.08	0.04
Fear	0.04	0	0	**0.44**	0.2	0.24	0.08
Hap.	0.04	0	0	0	**0.96**	0	0
Sad.	0.08	0	0.04	0.12	0	**0.76**	0
Sur.	0.08	0	0	0	0	0	**0.92**

TABLE III
CONFUSION MATRIX FOR *Equally spaced sample* EXPERIMENT. RESULTS OBTAINED USING $\alpha = 15$ AND $\gamma = 1$. GLOBAL RECOGNITION RATE IS 72%.

	Ang.	Cont.	Dis.	Fear	Hap.	Sad.	Sur.
Ang.	**0.36**	0	0.12	0.08	0.08	0.2	0.16
Cont.	0.08	**0.24**	0	0.16	0.16	0.2	0.16
Dis.	0.08	0	**0.52**	0.04	0	0.12	0.24
Fear	0.08	0.04	0	**0.44**	0.2	0.04	0.2
Hap.	0	0.04	0	0.04	**0.88**	0	0.0.04
Sad.	0.24	0	0.12	0.24	0	**0.28**	0.12
Sur.	0	0.08	0	0	0	0.04	**0.88**

TABLE IV
CONFUSION MATRIX FOR *Single sample* EXPERIMENT. RESULTS OBTAINED USING $\alpha = 30$ AND $\gamma = 0.5$. GLOBAL RECOGNITION RATE IS 51%.

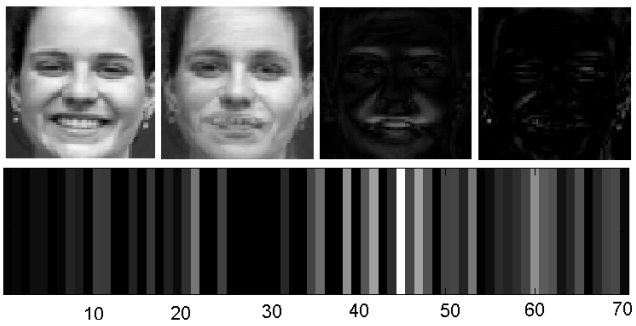

Fig. 4. Our model for a single frame(left) provide a decomposition on a neutral component(middle left), a sparsely represenatable emotion(middle right), and a sparse error (right). This sparse error compensate for features that are not capture by the emotion dictionary or the neutral face set (see the earrings in the picture on the right). Results obtained using $\alpha = 30$ and $\gamma = 0.5$

Fig. 5. The first row shows a decomposition of an expressive using the Single Frame model with $\alpha = 100$. In this case, the neutral face looks like an average of the neutral faces in matrix M_c. The second row shows a decomposition for $\alpha = 30$. For this second value,the neutral face corresponds to the same identity of the person in the testing image.

V. Analysis

As observed from Table I our validation experiment support the effectiveness of emotion recognition using sparse representation. Similar recognition rates (in the order of 70% to 80%) were also obtained for our experiment of emotion recognition on video sequences (see Tables II and III). The *Equally spaced samples* experiment confirm the capability of our model of recognizing emotion without requiring explicitness of the neutral face of the testing person.

Tail samples and *Equally spaced samples* provide two interesting testing scenarios. As observed from Figure 2, the *Tail samples* configuration is dominated by the expressive frames, thus the low rank term pushes the initial (neutral) frame to be transformed into an expressive image. Then, coefficient X is mostly dedicated to explain the residual between expression and neutral for this first frame. Since this residual can be sparsely expressed by the emotions dictionary, satisfactorious recognition was indeed obtained.

The *Equally spaced samples* test is a more challenging setup. As observed from Figure 3 there is no dominant term, so all the frames in the sequence may be pushed to a common frame (that may not correspond to a neutral face). In contrast to *Tail samples*, the X coefficient for *Equally spaced samples* is more dense since it requires represent the residual for each of the frames in the sequence. Despite the apparent more intricate scenario (and the simplicity of our dictionary) the recognition rate (72%) in this case remain close to the previous experiments.

Results provided in Tables II and III, were obtained from a few iterations of our ADMM solver implementation. As showed in Figures 2 and 3, by running a few iteration we may be capturing the most relevant descent direction on the testing input, which turns sufficient to get satisfactory recognition. Instead, running our solver implementation to converge introduce undesired artefacts as shown in Figure 6. There, we observe that rank and sparsity decrease, but columns of the low rank matrix are far from being neutral faces, and the sparse reconstruction do not correspond to a expression. A more precise tuning of the sparse error term parameter γ may attenuate these errors.

Recognition rates for the *Single frame* instance were much below than for video sequences (in the range of 50

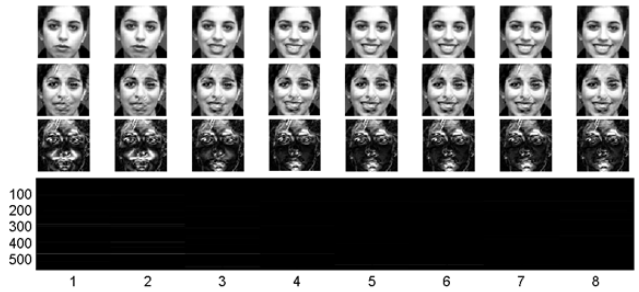

Fig. 6. Satisfactory recognition rates were obtained using some few itearions of ADMM (30 in our tests). For large number of iterations (1000 in this example) recognition rate drops by undesired artifacts introduced to our neutral plus expression decomposition.

to 60%). Figure 4, shows an example of a decomposition of an expressive face in a neutral face approximation, a sparse representable emotion and a sparse error term. Extracting a neutral face is indeed a challenging problem. In our model, the results for a given input image may variate widely according to the selected matrix of neutral face images (M_c) and the low rank parameter α (see Figure 5).

VI. Conclusion

In this worked we presented a method that attained emotion recognition in the $70\% - 80\%$ range (still not competitive the $80\% - 90\%$ state of the art methods) for equally spaced video sequences on 7 different expressions. The novelty of our approach is on using join sparse representation and low rank to disregard the explicitness of the neutral face of the testing sample. In contrasts to previous works in the area, no feature processing, dictionary learning, or join sparsity terms were involved on getting the presented results. Higher recognition rates might be expected by exploiting each of these ideas.

We observed a large variation on the differents emotion categories: while happiness and surprise are well classified, fear is specially hard to recognize. The feature space and the classification strategy we are currently using seems insufficient to get a good separation of some of these categories. Parameter tuning and checking convergence issues are two areas that require further work.

VII. Further Ideas

The current progress verifies the hypothesis that an expression is sparsely representable as well as separable from an expressive face. We will explore more in the following aspects to improve the perforamnce, reproduce other methods' performance reported in the literature and make a fair comparison under the same experimental setting.

A. Model design

Under Donoho & Xu's Restricted Nullspace Property (RUP, 2001), basis pursuit l-1 relaxation is exact for all S-sparse vectors. This sufficient and necessary property is even more strict than Candes & Tao's Restricted Isometry Property (RIP, 2005), which is only a sufficient condition for the exact recovery. When designing the objective, we can always decompose it into a loss function and a regularizer, in the machine learning perspective. The variety mostly lies in the choose of the regularizer. Here comes the Restricted Strong Convexity (RSC, 2009) [4], which guarantees the decomposability of an objective function and is a generalization of RUP.

B. Action unit and features

Since face expression is basically an action over time, we are curious about a reasonable format of a training or testing data unit. In literature, people employ action units [5] to characterize facial expressions. A whole sequence starting from a neutral face is not always necessary. In terms of discriminality, researchers have examined automatic detection of action units [6]. This direction will involve too many side materials and we prefer focusing on those related to sparse representation such as [7], which implies feature representation instead of raw pixel intensities, as expected.

C. Discovery of low-rank component

In our current model, we assume there exists a single low-rank component L. However, it is entirely possible that L is a combination of several low-rank components or subspaces. By introducing low-rank representation [8], we re-write the constraint to be $Y = AX + BL$, where A is a dictionary that linear spans the expression data space, and B is another dictionary that linearly spans the neutral data space. Both dictionaries are dynamically updated iteration by iteration. The initial dictionaries can be set as a concatenation of all training samples. By setting B as an identity matrix, the model falls back to our current model. This idea is highly related to Taheri et. al. [9].

D. Dictionary learning

The above section mentions online updating of dictionaries. As a first attempt, we prefer a pre-trained dictionary by K-SVD [10], which is motivated from minimizing residues or maximizing likelihood in the Bayesian perspective. It includes a sparse coding stage to generate a codebook and a codebook update stage to refine the codebook, which are similar with the computing the cancroids for each cluster and re-form the cluster in K-means, respectively. K-SVD provides a way to train a dictionary without providing a guarantee that such a dictionary will induce good enough discriminality of the recovered X. This part of work is still quite experimental, while the theoretical analysis of K-SVD is there and quite involved [11]. For the next step, we hope to move to online learning of dictionary since we have a short video as a data unit. We may refer to [12] in the context of action recognition.

E. Double sparse representation

[9] first aims to decompose a expressive face image into two components a neutral face and an expression. Since the model is for videos, Y matrix is decomposable into a low-rank matrix Y_n and a sparse matrix Y_e by Robust PCA. However, we may argue about the ambiguity of components

because a component can be sparse as well as low-rank. And then, [9] applies sparse representations for Y_n over all identity samples in the same identity class and for Y_e, respectively. This type of double sparse representation in a single objective is also termed Morphological Component Analysis (MCA) [13].

F. Group sparsity

Group sparsity [14] has played a key role in applications of sparse representation and robust PCA such as background subtraction problems [15]. We have tried enforcing row sparsity for X, for supports for frames in one testing unit should be almost the same. This modification results in more complicated optimization in ADMM. We refer to [16] regarding how to linearize one step or sub-problem in ADMM.

G. Classifier: label assignment

The current implementation adopts the 1-NN classifier, which is problematic when two or more classes have an equal number of votes.

References

[1] P Lucey, J Cohn, T Kanade, J Saragih, Z Ambadar, and I Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE CVPR*, 2010.

[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientic, 2002.

[3] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, "Robust face recognition via sparse representation," vol. 31, no. 2, 2009.

[4] S. Negahban, Ravikumar, M. J. Wainwright, and Yu, "Estimation rates of (near) low-rank matrices with noise and high-dimensional scaling," in *arXiv pre-print (Annals of Statistics)*, 2009, vol. 0912.5100.

[5] P. Ekman and W. Friesen, *Facial action coding system: A technique for the measurement of facial movement*, Consulting Psychologists Press, 1978.

[6] Tomas Simon, Minh Hoai Nguyen, Fernando De La Torre, and Jeffrey F. Cohn, "Action unit detection with segment-based svms," in *CVPR*, 2010.

[7] Taheri, "Structure-preserving sparse decomposition for facial expression analysis," 2013.

[8] Liu, "Robust recovery of subspace structures by low-rank representation," vol. 35, no. 1, 2012.

[9] S. Taheri, V. Patel, , and R. Chellappa, "Component-based recognition of faces and facial expressions," vol. 4, no. 4, 2013.

[10] Michal Aharon, "K-svd...," vol. 54, no. 11, 2006.

[11] Ron, "Analysis k-svd...," vol. 61, no. 3, 2013.

[12] Zhao, "Online detection of unusual events in videos via dynamic sparse coding," in *IEEE CVPR*, 2011.

[13] Starck, "Image decomposition via the combination of sparse representations and a variationl approach," vol. 14, no. 10, 2005.

[14] J Huang and T Zhang, "The benefit of group sparsity," 2010.

[15] Cui, "Background subtracion using low rank and group sparsity constraints," in *ECCV*, 2012.

[16] Lin, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *NIPS*, 2011.

[17] Allen Yang, Arvind Ganesh, Zihan Zhou, Andrew Wagner, Shankar Sastry Victor Shia, and Yi Ma, "Fast l-1 minimization algorithms," `http://www.eecs.berkeley.edu/~yang/software/l1benchmark/`, 2008.

[18] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE T-PAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[19] J Yang, "Alternating direction algorithms for l1-problems in compressive sensing," in *arXiv*, 2009, vol. 0912.1185.

[20] T Kanade, J F Cohn, and Y Tian, "Comprehensive database for facial expression analysis," in *IEEE FG*, 2000.

[21] G. Roisman Z. Zeng, M. Pantic and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," vol. 31, no. 1, 2009.

[22] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization," 2010.

[23] Vinay Bettadapura, "Face expression recognition and analysis: The state of the art," .