HIERARCHICAL SPARSE AND COLLABORATIVE LOW-RANK REPRESENTATION FOR EMOTION RECOGNITION

Xiang Xiang, Minh Dao, Gregory D. Hager, Trac D. Tran

Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
\{xxiang, minh.dao, ghager1, trac\}@jhu.edu

ABSTRACT
In this paper, we design a Collaborative-Hierarchical Sparse and Low-Rank (C-HiSLR) model that is natural for recognizing human emotion in visual data. Previous attempts require explicit expression components, which are often unavailable and difficult to recover. Instead, our model exploits the low-rank property to subtract neutral faces from expressive facial frames as well as performs sparse representation on the expression components with group sparsity enforced. For the CK+ dataset, C-HiSLR on raw expressive faces performs as competitive as the Sparse Representation based Classification (SRC) applied on manually prepared emotions. Our C-HiSLR performs even better than SRC in terms of true positive rate.

Index Terms— Low-rank, group sparsity, multichannel

1. INTRODUCTION
In this paper, the problem of interest is to recognize the emotion given a video of a human face and emotion category [1]. As shown in Fig.1, an expressive face can be separated into a dominant neutral face and a sparse expression component, which we term emotion and is usually encoded in a sparse noise term e. We investigate if we can sparsely represent the emotion over a dictionary of emotions [2] rather than expressive faces [3], which may confuse a similar expression with a similar identity [2]. Firstly, how to get rid of the neutral face? Surely we can prepare an expression with a neutral face explicitly provided as suggested in [2]. Differently, we treat an emotion as an action and assume neutral faces stay the same. If we stack vectors of neutral faces as a matrix, it should be low-rank (ideally with rank 1). Similarly, over time sparse vectors of emotions form a sparse matrix. Secondly, how to recover the low-rank and sparse components? In [4], the (low-rank) Principal Component Pursuit (PCP) [5] is performed explicitly. While theoretically the recovery is exact under conditions [5], it is of approximate nature in practice. Finally, since we only care about the sparse component, can we avoid such an approximate explicit PCP step? This drives us to exploit Sparse representation and Low-Rank property jointly in one model named SLR (Sec. 3.1).

Different from image-based methods [2, 4], we treat an emotion video as a multichannel signal. If we just use a single channel such as one frame to represent an emotion, much information is lost since all frames collaboratively represent an emotion. Therefore, we prefer using all or most of them. Should we treat them separately or simultaneously? The former just needs to recover the sparse coefficient vector for each frame. The latter gives a spatial-temporal representation, while it requires the recovery of a sparse coefficient matrix, which should often exhibit a specific structure. Should we enforce a class-wise sparsity separately or enforce a group sparsity collaboratively? [4] models the class-wise sparsity separately for the recognition of a neutral face’s identity and an expression image’s emotion once they have been separated. Alternatively, we can exploit the low-rankness as well as structured sparsity by inter-channel observation. Since class decisions may be inconsistent, we prefer a collaborative model [6] with group sparsity enforced [7]. This motivates us to introduce the group sparsity as a root-level sparsity to the SLR model embedded with a leaf-level atom-wise sparsity. The reason of keeping both levels is that signals over frames share class-wise yet not necessarily atom-wise sparsity patterns [8]. Therefore, we term this model Collaborative-Hierarchical Sparse and Low-Rank (C-HiSLR) model.

In the remainder of this paper, we review sparse and low-rank representation literature in Sec. 2, elaborate our model in Sec. 3, discuss the optimization in Sec. 4, empirically validate the model in Sec. 5, and draw a conclusion in Sec. 6.

2. RELATED WORKS
When observing a random signal y for recognition, we hope to send the classifier a discriminative compact representation x, which satisfies Ax = y and is yet computed by pursuing the best reconstruction. When A is under-complete, a closed-form approximate solution can be obtained by Least-Squares:
\[ x^* = \arg \min_x \|y - Ax\|^2 \approx (A^TA)^{-1}A^Ty. \]

When \( A \) is over-complete, we add a Tikhonov regularizer [9]:
\[ x^* = \arg \min_x \|y - Ax\|^2 + \lambda_e \|x\|^2 = \arg \min_x \|y - Ax\|^2 \approx (A^TA + \lambda_e I)^{-1}A^Ty \]
where \( y = [y, 0]^T, \ A = [A, \sqrt{\lambda_e}I]^T \)
is always under-complete. But \( x^* \) is not necessarily compact yet generally dense. Alternatively, we can seek a sparse usage of \( A \). Sparse Representation based Classification [10] (SRC) expresses a test sample \( y \) as a linear combination of prepared fixed training emotions. The matrix \( Y \) is over-complete and needs to be sparsely used. SRC still performs robustly well for denoising and coding tasks such as well-aligned noisy face identifications.

3. REPRESENTATION MODELS

In this section, we explain how to model \( X \) using \( Y \) and training data \( D \), which contains \( K \in \mathbb{Z}^+ \) types of emotions. We would like to classify a test video as one of the \( K \) classes.

3.1. SLR: Joint Sparse Representation and Low-Rankness

First of all, we need an explicit representation \( Y \) of an expressive face. The matrix \( Y \in \mathbb{R}^{d \times \tau} \) can be an arrangement of \( d \)-dimensional feature vectors \( y \in \mathbb{R}^d \) such as Gabor features [23] or concatenated image raw intensities [10] of the \( \tau \) frames: \( Y = [Y_1|...|Y_{\tau}]_{d \times \tau} \). We emphasize our model’s power by simply using the raw pixel intensities.

Now, we seek an implicit latent representation \( X \in \mathbb{R}^{n \times \tau} \) of an input test face’s emotion \( Y_e \in \mathbb{R}^{d \times \tau} \) as a sparse linear combination of prepared fixed training emotions \( D \in \mathbb{R}^{d \times n} \):
\[ Y_e = DX. \]

Since an expressive face \( y = y_e + y_n \) is a superposition of an emotion \( y_e \in \mathbb{R}^d \) and a neutral face \( y_n \in \mathbb{R}^d \), we have
\[ Y = Y_e + L, \]
where \( L \in \mathbb{R}^{d \times \tau} \) is ideally \( \tau \)-times repetition of the column vector of a neutral face \( y_n \in \mathbb{R}^d \). Presumably \( L = [y_n|...|y_n]_{d \times \tau} \). As shown in Fig. 2, \( X \) subjects to
\[ Y = DX + L, \]
where the dictionary matrix \( D_{d \times n} \) is an arrangement of all sub-matrices \( D_{[y]}, \ j = 1, ..., \lfloor d/n \rfloor \). Only for training, we have \( \lfloor d/n \rfloor \) training emotions with neutral faces subtracted. The above constraint of \( X \) characterizes an affine transformation from the latent representation \( X \) to the observation \( Y \). If we write \( X \) and \( Y \) in homogeneous forms [24], then we have
\[ Y_{d \times \tau} \begin{bmatrix} 1 & 1_{1 \times \tau} \end{bmatrix} = D_{d \times n} \left( y_n \right)_{d \times 1} \begin{bmatrix} 1 & 1_{1 \times \tau} \end{bmatrix} X_{n \times \tau} \begin{bmatrix} 1 & 1_{1 \times \tau} \end{bmatrix}. \]

In the ideal case with \( rank(L) = 1 \), if the neutral face \( y_n \) is pre-obtained [2, 4], it is trivial to solve for \( X \). Normally, \( y_n \) is unknown and \( L \) is not with rank 1 due to noises. As \( X \) is supposed to be sparse and \( rank(L) \) is expected to be as small as possible (maybe even 1), intuitively our objective is to
\[ \min_{X,L} \text{sparsity}(X) + \lambda_L \cdot \text{rank}(L), \]
where \( \text{rank}(L) \) can be seen as the sparsity of the vector formed by the singular values of \( L \). Here \( \lambda_L \) is a non-negative weighting parameter we need to tune [25]. When \( \lambda_L = 0 \), the optimization problem reduces to that in SRC. With both terms relaxed to be \( \ell_1 \) norm, we alternatively solve
\[ \min_{X,L} \|X\|_1 + \lambda_L \|L\|_*, \]
where \( \|\cdot\|_1 \) is the entry-wise \( \ell_1 \) matrix norm, whereas \( \|\cdot\|_* \) is the Schatten \( \ell_1 \) matrix norm (nuclear norm, trace norm) which can be seen as applying \( \ell_1 \) norm to the vector of singular values. Now, the proposed joint SLR model is expressed as
\[ \min_{X,L} \|X\|_1 + \lambda_L \|L\|_* \quad \text{s.t.} \quad Y = DX + L \]
Both SLR and C-HiSLR models can be seen as solving a problem of multi-channel Lasso (Least Absolute Shrinkage and Selection Operator). For a single-channel signal, Group Lasso [27] has explored the group structure for Lasso yet does not enforce sparsity within a group, while Sparse Group Lasso [28] yields an atom-wise sparsity as well as a group sparsity. Then, [8] extends Sparse Group Lasso to multichannel, resulting in a Collaborative-Hierarchical Lasso (C-HiLasso) model. For our problem, we do need \( L \), which induces a Collaborative-Hierarchical Sparse and Low-Rank (C-HiSLR) model:

\[
\min_{X,L} \|X\|_1 + \lambda L \|L\|_s + \lambda_g \sum_{G \in \mathcal{G}} \|X[G]\|_F
\]

\[
st. \quad Y = DX + L
\]

where \( X[G] \) is the sub-matrix formed by all the rows indexed by the elements in group \( G \subseteq \{1, \ldots, n\} \). As shown in Fig. 3, given a group \( G \) of indices, the sub-dictionary of columns indexed by \( G \) is denoted as \( D[G] \). \( \mathcal{G} = \{G_1, \ldots, G_K\} \) is a non-overlapping partition of \( \{1, \ldots, n\} \). Here \( \|\cdot\|_F \) denotes the Frobenius norm, which is the entry-wise \( \ell_2 \) norm as well as the Schatten \( \ell_2 \) matrix norm and can be seen as a group’s magnitude. \( \lambda_g \) is a non-negative weighting parameter for the group regularizer, which is generalized from an \( \ell_1 \) regularizer (consider \( \mathcal{G} = \{\{1\}, \{2\}, \ldots, \{n\}\} \) for singleton groups) [8]. When \( \lambda_g = 0, \) C-HiSLR degenerates into SLR. When \( \lambda_L = 0, \) we get back to collaborative Sparse Group Lasso.

3.3. Classification

Following SRC, for each class \( c \in \{1, 2, \ldots, K\} \), let \( D[G_c] \) denote the sub-matrix of \( D \) which consists of all the columns of \( D \) that correspond to emotion class \( c \) and similarly for \( X[G_c] \). We classify \( Y \) by assigning it to the class with minimal residual as \( c^* = \arg \min_c r_c(Y) := \|Y - D[G_c]X[G_c] - L\|_F \).

4. OPTIMIZATION

Both SLR and C-HiSLR models can be seen as solving

\[
\min_{X,L} f(X) + \lambda L \|L\|_s \quad s.t. \quad Y = DX + L
\]

To follow a standard iterative ADMM procedure [26], we write down the augmented Lagrangian function for (3) as

\[
\mathcal{L}(X,L,A) = f(X) + \lambda L \|L\|_s + \langle A, Y - DX - L \rangle + \frac{\beta}{2} \|Y - DX - L\|_F^2,
\]

where \( A \) is the matrix of multipliers, \( \langle \cdot, \cdot \rangle \) is inner product, and \( \beta \) is a positive weighting parameter for the penalty (augmentation). A single update at the \( k \)-th iteration includes

\[
L_{k+1} = \arg \min_L \lambda L \|L\|_s + \frac{\beta}{2} \|Y - DX_k - L + \frac{1}{\beta} A_k\|_F^2
\]

\[
X_{k+1} = \arg \min_X f(X) + \frac{\beta}{2} \|Y - DX - L_{k+1} + \frac{1}{\beta} A_k\|_F^2
\]

\[
A_{k+1} = A_k + \beta(Y - DX_{k+1} - L_{k+1}).
\]

The sub-step of solving (5) has a closed-form solution:

\[
L_{k+1} = \mathcal{D}(Y - DX_k + \frac{1}{\beta} A_k),
\]

where \( \mathcal{D} \) is the shrinkage thresholding operator. In SLR where \( f(X) = \|X\|_1 \), (6) is a Lasso problem, which we solve by using an existing fast solver [29]. When \( f(X) \) follows (2) of C-HiSLR, computing \( X_{k+1} \) needs an approximation based on the Taylor expansion at \( X_k \). We refer the reader to [8] for the convergence analysis and recovery guarantee.

5. EXPERIMENTAL RESULTS

All experiments are conducted on the CK+ dataset [31] which consists of 321 emotion sequences with labels (angry, contempt\(^1\), disgust, fear, happiness, sadness, surprise)\(^2\) and is randomly divided into a training set (10 sequences per category) and a testing set (5 sequences per category). For SRC, we assume that the information of neutral face is provided. We subtract the first frame (a neutral face) from the last frame per sequence for both training and testing. Thus, each emotion is represented as an image. However, for SLR and C-HiSLR, we assume no prior knowledge of the neutral face. We form a dictionary by subtracting the first frame from the last \( \tau_{trn} \) frames per sequence and form a testing unit using the last \( (\tau_{tst} - 1) \) frames together with the first frame, which is not explicitly known as a neutral face. Thus, each emotion is represented as a video. Here, we set \( \tau_{trn} = 4 \) or \( 8, \tau_{tst} = 8, \lambda_L = 10 \) and \( \lambda_G = 4.5 \). Fig. 4 visualizes the recovery results given by C-HiSLR. Facial images are cropped using the Viola-Jones detector [32] and resized to \( 64 \times 64 \). As shown in Fig. 5, imperfect alignment may affect the performance.

Firstly, SRC achieves a total recognition rate of \( 0.80 \), against \( 0.80 \) for eigenface with nearest subclassifier and \( 0.72 \) for eigenface with nearest neighbor classifier. This verifies that emotion is sparsely representable by training data and SRC can be an alternative to subspace based methods. Secondly, Table 1-3 present the confusion matrix (\( \tau_{trn} = 4 \)) and Table 4 summarizes the true positive rate (i.e., sensitivity). We have anticipated that SLR (0.70) performs worse than SRC (0.80) since SRC is equipped with neutral faces. However, C-HiSLR’s result (0.80) is comparable with SRC’s. C-HiSLR performs even better in terms of sensitivity, which verifies that the group sparsity indeed boosts the performance.

\(^1\)Contempt is discarded in [2, 4] due to its confusion with other classes.

\(^2\)Please visit http://www.cs.jhu.edu/~xxliang/slr/ for the cropped face data and programs of C-HiSLR, SLR, SRC and Eigenface.
6. CONCLUSION
We design the C-HiSLR representation model for emotion recognition, unlike [2] requiring neutral faces as inputs and [4] generating labels of identity and emotion as mutual by-products with extra efforts. Our contribution is two-fold. First, we do not recover emotion explicitly. Instead, we treat frames simultaneously and implicitly subtract the low-rank neutral face. Second, we preserve the label consistency by enforcing atom-wise as well as group sparsity. For the CK+ dataset, C-HiSLR’s performance on raw data is comparable with SRC given neutral faces, which verifies that emotion is automatically separable from expressive faces as well as sparsely representable. Future works will include handling misalignment [33] and incorporating dictionary learning [12].

ACKNOWLEDGMENTS
This work is supported by US National Science Foundation under Grants CCF-1117545 and CCF-1422995, Army Research Office under Grant 60219-MA, and Office of Naval Research under Grant N00014-12-1-0765. The first author is grateful for the fellowship from China Scholarship Council.

| Model  | An | Co | Di | Fe | Ha | Sa | Su |
|-------|----|----|----|----|----|----|----|
| SRC   | 0.71 | 0.60 | 0.93 | 0.25 | 0.96 | 0.24 | 0.98 |
| SLR   | 0.51 | 0.63 | 0.74 | 0.51 | 0.85 | 0.70 | 0.94 |
| C-HiSLR | 0.77 | 0.84 | 0.93 | 0.53 | 0.93 | 0.65 | 0.95 |

Table 4. Comparison of sensitivity. The bold and italics denote the highest and lowest respectively. Difference within 0.05 is treated as comparable. C-HiSLR performs the best.
7. REFERENCES

[1] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE T-PAMI, vol. 31, no. 1, pp. 39–58, 2009.

[2] Stefanos Zafeiriou and Maria Petrou, “Sparse representations for facial expressions recognition via II optimization,” in IEEE CVPR Workshop, 2010.

[3] Raymond R. R. A. Tropp, Anna C. Gilbert, “Signal recovery from random measurements via Orthogonal Matching Pursuit,” IEEE Trans. Inf. Theory, vol. 53, no. 12, pp. 4655–4666, 2007.

[4] Emmanuelle J. Candes, Justin Romberg, and Terence Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489–509, 2006.

[5] Edoardo Amaldi and Viggo Kann, “On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems,” Theoretical Computer Science, vol. 209, pp. 237–260, 1998.

[6] Wikipedia, “Tikhonov regularization,” http://en.wikipedia.org/wiki/Tikhonov_regularization.

[7] Wikipedia, “Homogeneous coordinates,” http://en.wikipedia.org/wiki/Homogeneous_coordinates.

[8] MathWorks, “Matlab Computer Vision System Toolbox,” http://www.mathworks.com/products/computer-vision/

[9] Gregory D. Hager and Peter N. Belhumeur, “Efficient region tracking with parametric models of geometry and illumination,” IEEE T-PAMI, vol. 20, no. 10, pp. 1025–1039, 1998.

[10] Emmanuel J. Candes and Terence Tao, “Decoding by linear programming,” IEEE Trans. Inf. Theory, vol. 51, no. 12, pp. 4203–4215, 2005.

[11] Guangcan Liu, Zhouchen Liu, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, “Robust recovery of subspace structures by low-rank representation,” IEEE T-PAMI, vol. 35, no. 1, pp. 171–185, 2013.

[12] Ehsan Elhamifar and Rene Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” IEEE T-PAMI, vol. 35, no. 11, pp. 2765–2781, 2013.

[13] Meng Yang and Lei Zhang, “Gabor feature based sparse representation for face recognition with gabor occlusion dictionary,” in ECCV, 2010.

[14] Raymond R. R. A. Tropp, Anna C. Gilbert, “Signal recovery from random measurements via Orthogonal Matching Pursuit,” IEEE Trans. Inf. Theory, vol. 53, no. 12, pp. 4655–4666, 2007.

[15] Emmanuelle J. Candes, Justin Romberg, and Terence Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489–509, 2006.

[16] Edoardo Amaldi and Viggo Kann, “On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems,” Theoretical Computer Science, vol. 209, pp. 237–260, 1998.

[17] Emmanuelle J. Candes, Justin Romberg, and Terence Tao, “Stable signal recovery from incomplete and inaccurate measurements,” Comm. Pure Appl. Math., vol. 59, pp. 1207–1223, 2006.

[18] Wikipedia, “Tikhonov regularization,” http://en.wikipedia.org/wiki/Tikhonov_regularization.

[19] MathWorks, “Matlab Computer Vision System Toolbox,” http://www.mathworks.com/products/computer-vision/.

[20] Wikipedia, “Homogeneous coordinates,” http://en.wikipedia.org/wiki/Homogeneous_coordinates.

[21] MathWorks, “Matlab Computer Vision System Toolbox,” http://www.mathworks.com/products/computer-vision/.

[22] Gregory D. Hager and Peter N. Belhumeur, “Efficient region tracking with parametric models of geometry and illumination,” IEEE T-PAMI, vol. 20, no. 10, pp. 1025–1039, 1998.