Joint Representation and Estimator Learning for Facial Action Unit Intensity Estimation Supplementary Material

Yong Zhang[1], Baoyuan Wu[1]*, Weiming Dong[2], Zhifeng Li[1], Wei Liu[1], Bao-Gang Hu[2], and Qiang Ji[3]
[1]Tencent AI Lab, [2]National Laboratory of Pattern Recognition, CASIA, [3]Rensselaer Polytechnic Institute
{zha6.5ngyong201303,wubaoyuan1987}@gmail.com, weiming.dong@ia.ac.cn
michaelzfli@tencent.com, wl2223@columbia.edu, hubg@nlpr.ia.ac.cn, qji@ecse.rpi.edu

This supplementary material contains:

1. Derivation of the objective (Sec. 1)
2. Details of optimization (Sec. 2)
3. Convergence of the proposed algorithm (Sec. 3)
4. Influence of feature dimension (Sec. 4)
5. Visualization of the learned representation (Sec. 5)

Table 1. Notations in the formulation

Name	Description
	original space
$\mathbf{S}_u$	A set of segments without AU intensity annotation
$\mathbf{S}_u^m$	The m-th segment in $\mathbf{S}_u$
$\mathbf{X}_l$	A set of frames with AU intensity annotation
$\mathbf{Y}_l$	The intensities of the frames $\mathbf{X}_l$
	learned space
$\mathbf{B}$	The basic vectors of the learned space
$\mathbf{\Phi}_u$	The coefficients of frames of segments $\mathbf{S}_u$
$\mathbf{\Phi}_u^m$	The coefficients of frames of the m-th segment $\mathbf{S}_u^m$ in $\mathbf{S}_u$
$\mathbf{\Phi}_l$	The coefficients of the annotated frames $\mathbf{X}_l$
$\mathbf{w}$	The parameters of the intensity estimator
	auxiliary matrix
$\mathbf{\Gamma}^m$	A matrix with $\mathbf{\Gamma}^m_{i,i} = 1$, $\mathbf{\Gamma}^m_{i,i+1} = -1$ and other elements being 0. It is used to represent the label ranking.
$\mathbf{\Gamma}$	A combined matrix $\mathbf{\Gamma} = \text{diag}([\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, ..., \mathbf{\Gamma}^M])$
$\mathbf{C}^m$	The adjacent matrix of the m-th segment. $\mathbf{C}^m_{i,j} = 1$ if $\|i-j\| = 1$. Otherwise, $\mathbf{C}^m_{i,j} = 0$. It is used to represent the label and feature smoothness.
$\mathbf{D}^m$	A diagonal matrix. The i-th diagonal element is $\mathbf{D}^m_{i,i} = \sum_j \mathbf{C}^m_{i,j}$.
$\mathbf{L}^m$	A matrix. $\mathbf{L}^m = \mathbf{D}^m - \mathbf{C}^m$.
$\mathbf{L}$	A combined matrix $\mathbf{L} = \text{diag}([\mathbf{L}^1, \mathbf{L}^2, ..., \mathbf{L}^M])$

Table 2. Notations in the ADMM optimization

Name	Description
$\mathbf{C}_l$,$\mathbf{C}_u$, $\mathbf{Z}_1$,$\mathbf{Z}_2$,$\mathbf{Z}_3$	Introduced variables in the scaled form of the augmented Lagrangian function
$\mathbf{\Lambda}_l$, $\mathbf{\Lambda}_u$, $\mathbf{V}_1$,$\mathbf{V}_2$,$\mathbf{V}_3$	The multipliers in the scaled form of the augmented Lagrangian function

1. Derivation of the Objective

In Section 3.5 and Section 3.6 of the main paper, we describe the objective function and the algorithm for optimization, respectively. Here we present the details of the derivation of the objective and the details of the optimization. Involved notations are summarized in Table 1 and 2.

We consider an example, i.e.,

$$\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{AB}||_F^2 + \lambda||\mathbf{A}||_{2,1} \tag{1}$$

*Corresponding author.

The equivalent problem is

$$\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{AB}||_F^2 + \lambda||\mathbf{C}||_{2,1}.$$
$$\text{s.t. } \mathbf{A} = \mathbf{C}. \tag{2}$$

The augmented Lagrangian function is

$$L_\rho(\mathbf{A}, \tilde{\mathbf{\Lambda}}) = ||\mathbf{X} - \mathbf{AB}||_F^2 + \lambda||\mathbf{C}||_{2,1} + \frac{\rho}{2}||\mathbf{A} - \mathbf{C}||_F^2 + \text{tr}\big(\tilde{\mathbf{\Lambda}}^T(\mathbf{A} - \mathbf{C})\big). \tag{3}$$

Let $\mathbf{\Lambda} = \frac{\tilde{\mathbf{\Lambda}}}{\rho}$. We can get the scaled form of the augmented Lagrangian function, i.e.,

$$L_\rho(\mathbf{A}, \mathbf{\Lambda}) = ||\mathbf{X} - \mathbf{AB}||_F^2 + \lambda||\mathbf{C}||_{2,1} + \frac{\rho}{2}||\mathbf{A} - \mathbf{C} + \mathbf{\Lambda}||_F^2 - \frac{\rho}{2}||\mathbf{\Lambda}||_F^2. \tag{4}$$

Then, we derive the scaled form of our problem. The augmented Lagrangian function can be written as

$$\begin{aligned}
&L_{\rho_1,\rho_2,\rho_3}(\mathbf{\Phi}_l, \mathbf{\Phi}_u, \mathbf{B}, \mathbf{w}, \mathbf{C}_\cdot, \tilde{\mathbf{\Lambda}}_\cdot, \mathbf{Z}_\cdot, \tilde{\mathbf{V}}_\cdot) \\
= &\frac{1}{2}\left\| \begin{bmatrix}\mathbf{X}_l \\ \mathbf{S}_u\end{bmatrix} - \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}\mathbf{B} \right\|_F^2 + \lambda_1 \left\| \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix}^T \right\|_{2,1} \\
&+ \text{tr}\Big(\begin{bmatrix}\tilde{\mathbf{\Lambda}}_l \\ \tilde{\mathbf{\Lambda}}_u\end{bmatrix}^T \begin{bmatrix}\mathbf{\Phi}_l - \mathbf{C}_l \\ \mathbf{\Phi}_u - \mathbf{C}_u\end{bmatrix} \Big) + \frac{\rho_1}{2} \left\| \begin{bmatrix}\mathbf{\Phi}_l - \mathbf{C}_l \\ \mathbf{\Phi}_u - \mathbf{C}_u\end{bmatrix} \right\|_F^2 \\
&+ \mathbf{I}_-(\mathbf{Z}_0) + \text{tr}(\tilde{\mathbf{V}}_0^T(\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0)) + \frac{\rho_2}{2}||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0||^2 \\
&+ \mathbf{I}_+(\mathbf{Z}_1) + \text{tr}(\tilde{\mathbf{V}}_1^T(\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1)) + \frac{\rho_3}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1||^2 \\
&+ \mathbf{I}_+(\mathbf{Z}_2) + \text{tr}(\tilde{\mathbf{V}}_2^T(\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2)) + \frac{\rho_3}{2}||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2||^2 \\
&+ \frac{\lambda_0}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Y}_l||^2 + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u\mathbf{w} \\
&+ \lambda_3\text{tr}(\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u),
\end{aligned} \tag{5}$$

where $\tilde{\mathbf{\Lambda}}_l$, $\tilde{\mathbf{\Lambda}}_u$, $\tilde{\mathbf{V}}_0$, $\tilde{\mathbf{V}}_2$, and $\tilde{\mathbf{V}}_2$ are Lagrangian multipliers. $\mathbf{C}_l$, $\mathbf{C}_u$, $\mathbf{Z}_0$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ are introduced variables.

Let $\mathbf{\Lambda}_l = \frac{\tilde{\mathbf{\Lambda}}_l}{\rho_1}$, $\mathbf{\Lambda}_u = \frac{\tilde{\mathbf{\Lambda}}_u}{\rho_1}$, $\mathbf{V}_0 = \frac{\tilde{\mathbf{V}}_0}{\rho_2}$, $\mathbf{V}_1 = \frac{\tilde{\mathbf{V}}_1}{\rho_3}$, and $\mathbf{V}_2 = \frac{\tilde{\mathbf{V}}_2}{\rho_3}$. The scaled form of the Lagrangian function can be written as

$$
\begin{aligned}
& L_{\rho_1,\rho_2,\rho_3}(\mathbf{\Phi}_l, \mathbf{\Phi}_u, \mathbf{B}, \mathbf{w}, \mathbf{C}_\cdot, \mathbf{\Lambda}_\cdot, \mathbf{Z}_\cdot, \mathbf{V}_\cdot) \qquad (6)\\
= \; & \frac{1}{2}\left\| \begin{bmatrix}\mathbf{X}_l \\ \mathbf{S}_u\end{bmatrix} - \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}\mathbf{B}\right\|_F^2 + \lambda_1 \left\| \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix}^T \right\|_{2,1} \\
& + \frac{\rho_1}{2}\left\| \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix} - \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix} + \begin{bmatrix}\mathbf{\Lambda}_l \\ \mathbf{\Lambda}_u\end{bmatrix}\right\|_F^2 - \frac{\rho_1}{2}\left\| \begin{bmatrix}\mathbf{\Lambda}_l \\ \mathbf{\Lambda}_u\end{bmatrix}\right\|_F^2 \\
& + \mathbf{I}_-(\mathbf{Z}_0) + \frac{\rho_2}{2}||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0||^2 - \frac{\rho_2}{2}||\mathbf{V}_0||^2 \\
& + \mathbf{I}_+(\mathbf{Z}_1) + \frac{\rho_3}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1 + \mathbf{V}_1||^2 - \frac{\rho_3}{2}||\mathbf{V}_1||^2 \\
& + \mathbf{I}_+(\mathbf{Z}_2) + \frac{\rho_3}{2}||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2||^2 - \frac{\rho_3}{2}||\mathbf{V}_2||^2 \\
& + \frac{\lambda_0}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Y}_l||^2 + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u\mathbf{w} + \lambda_3\mathrm{tr}(\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u).
\end{aligned}
$$

2. Details of Optimization

In Eq.(6), $\mathbf{\Phi}_u\mathbf{w}$ and $\mathbf{\Phi}_l\mathbf{w}$ cannot be treated as new variables because $\mathbf{\Phi}_u$ and $\mathbf{\Phi}_l$ are also coupled with $\mathbf{B}$ and $\mathrm{tr}(\mathbf{\Phi}_u\mathbf{L}^T\mathbf{\Phi}_u)$ involves only $\mathbf{\Phi}_u$. Hence, we optimize each variable alternatively based on ADMM as follows. PCA [2] is used to initialize $\mathbf{B}$, $\mathbf{\Phi}_l$, and $\mathbf{\Phi}_u$. $\mathbf{C}_l = \mathbf{\Phi}_l$ and $\mathbf{C}_u = \mathbf{\Phi}_u$, while other variables are randomly initialized. Note that the currently updated variable will be used to update other variables.

Optimizing $\mathbf{B}$ The subproblem with respect to $\mathbf{B}$ is

$$
\min_{\mathbf{B}\in\mathcal{B}} \; \frac{1}{2}\left\| \begin{bmatrix}\mathbf{X}_l \\ \mathbf{S}_u\end{bmatrix} - \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix}\mathbf{B}\right\|_F^2. \qquad (7)
$$

Taking the derivative of $\mathbf{B}$ and setting it to 0, we have

$$
\tilde{\mathbf{B}} = [\mathbf{\Phi}_l^T\mathbf{\Phi}_l + \mathbf{\Phi}_u^T\mathbf{\Phi}_u]^{-1}[\mathbf{\Phi}_l^T\mathbf{X}_l + \mathbf{\Phi}_u^T\mathbf{X}_u]. \qquad (8)
$$

To project it into $\mathcal{B} = \{\mathbf{b} : ||\mathbf{b}||_2 = 1\}$, we normalize each row of $\tilde{\mathbf{B}}$, i.e., $\mathbf{B}_{i\cdot} = \frac{\tilde{\mathbf{B}}_{i\cdot}}{||\tilde{\mathbf{B}}_{i\cdot}||_2}$.

Optimizing $\mathbf{\Phi}_l$ and $\mathbf{\Phi}_l$ The subproblem with respect to $\mathbf{\Phi}_l$ is

$$
\begin{aligned}
\min_{\mathbf{\Phi}_l} \; & \frac{1}{2}||\mathbf{X}_l - \mathbf{\Phi}_l\mathbf{B}||^2 + \frac{\lambda_0}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Y}_l||^2 \qquad (9)\\
& + \frac{\rho_1}{2}||\mathbf{\Phi}_l - \mathbf{C}_l + \mathbf{\Lambda}_l||^2 + \frac{\rho_3}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1 + \mathbf{V}_1||^2.
\end{aligned}
$$

Taking the derivative of $\mathbf{\Phi}_l$ and setting it to 0, we have the closed-form solution

$$
\begin{aligned}
\mathbf{\Phi}_l = & [\mathbf{X}_l\mathbf{B}^T + \lambda_0\mathbf{Y}_l\mathbf{w}^T + \rho_1(\mathbf{C}_l - \mathbf{\Lambda}_l) \qquad (10)\\
& + \rho_3(\mathbf{Z}_1 - \mathbf{V}_1)\mathbf{w}^T][\mathbf{B}\mathbf{B}^T + (\lambda_0 + \rho_3)\mathbf{w}\mathbf{w}^T + \rho_1\mathbf{I}]^{-1},
\end{aligned}
$$

where $\mathbf{I}$ is an identity matrix. The subproblem with respective to $\mathbf{\Phi}_u$ is

$$
\begin{aligned}
\min_{\mathbf{\Phi}_u} \; & \frac{1}{2}||\mathbf{X}_u - \mathbf{\Phi}_u\mathbf{B}||_F^2 + \frac{\rho_1}{2}||\mathbf{\Phi}_u - \mathbf{C}_u + \mathbf{\Lambda}_u||_F^2 \\
& + \frac{\rho_2}{2}||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0||^2 + \frac{\rho_3}{2}||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2||^2 \\
& + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u^T\mathbf{L}\mathbf{\Phi}_u\mathbf{w} + \lambda_3\mathrm{tr}(\mathbf{\Phi}_u^T\mathbf{L}\mathbf{\Phi}_u). \qquad (11)
\end{aligned}
$$

The gradient of $\mathbf{\Phi}_u$ is

$$
\begin{aligned}
\nabla_u = \; & \mathbf{\Phi}_u[\mathbf{B}\mathbf{B}^T + \rho_1\mathbf{I} + \rho_3\mathbf{w}\mathbf{w}^T] + \lambda_3(\mathbf{L} + \mathbf{L}^T)\mathbf{\Phi}_u \\
& + [\lambda_2(\mathbf{L} + \mathbf{L}^T) + \rho_2\mathbf{\Gamma}^T\mathbf{\Gamma}]\mathbf{\Phi}_u\mathbf{w}\mathbf{w}^T - \mathbf{X}_u\mathbf{B}^T \\
& - \rho_1(\mathbf{C}_u - \mathbf{\Lambda}_u) - \rho_3(\mathbf{Z}_2 - \mathbf{V}_2)\mathbf{w}^T \\
& - \rho_2\mathbf{\Gamma}^T(\mathbf{Z}_0 - \mathbf{V}_0)\mathbf{w}^T. \qquad (12)
\end{aligned}
$$

Though it has a closed-form solution, the computation is inefficient since it involves the inverse of a large matrix. Instead we use a gradient-based method to update $\mathbf{\Phi}_u$, i.e.,

$$
\mathbf{\Phi}_u \leftarrow \mathbf{\Phi}_u - \alpha\nabla_u, \qquad (13)
$$

where the step size α is obtained by exact line search. α is computed as $\alpha = \frac{t_1}{t_2}$, where

$$
\begin{aligned}
t_1 = & \; \mathrm{tr}((\mathbf{\Phi}_u\mathbf{B} - \mathbf{X}_u)^T\nabla_u\mathbf{B}) + \rho_1\mathrm{tr}((\mathbf{\Phi}_u - \mathbf{C}_u + \mathbf{\Lambda}_u)^T\nabla_u) \\
& + [\rho_2(\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0)^T\mathbf{\Gamma} + \rho_3(\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2)^T]\nabla_u\mathbf{w} \\
& + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u^T(\mathbf{L} + \mathbf{L}^T)\nabla_u\mathbf{w} + \lambda_3\mathrm{tr}(\nabla_u^T(\mathbf{L} + \mathbf{L}^T)\mathbf{\Phi}_u), \\
t2 = & \; tr(\mathbf{B}^T\nabla_u^T\nabla_u\mathbf{B}) + \mathrm{tr}(\nabla_u^T(\rho_1\mathbf{I} + 2\lambda_3\mathbf{L})\nabla_u) \\
& + \mathbf{w}^T\nabla_u^T(\rho_3\mathbf{I} + 2\lambda_2\mathbf{L} + \rho_2\mathbf{\Gamma}^T\mathbf{\Gamma})\nabla_u\mathbf{w}. \qquad (14)
\end{aligned}
$$

Optimizing $\mathbf{w}$ The subproblem with respect to $\mathbf{w}$ is

$$
\begin{aligned}
\min_{\mathbf{w}} \; & \frac{\lambda_0}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Y}_l||^2 + \lambda_2\mathbf{w}^T\mathbf{\Phi}_u^T\mathbf{L}\mathbf{\Phi}_u\mathbf{w} \\
& + \frac{\rho_2}{2}||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0||^2 + \frac{\rho_3}{2}||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1 + \mathbf{V}_1||^2 \\
& + \frac{\rho_3}{2}||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2||^2. \qquad (15)
\end{aligned}
$$

Taking the derivative of $\mathbf{w}$ and setting it to 0, we can obtain the closed-form solution

$$
\begin{aligned}
\mathbf{w} = & \big[(\lambda_0 + \rho_3)\mathbf{\Phi}_l^T\mathbf{\Phi}_l + \mathbf{\Phi}_u^T[\lambda_2(\mathbf{L} + \mathbf{L}^T) + \rho_2\mathbf{\Gamma}^T\mathbf{\Gamma} \\
& + \rho_3\mathbf{I}]\mathbf{\Phi}_u\big]^{-1}\big[\lambda_0\mathbf{\Phi}_l^T\mathbf{Y}_l + \rho_2\mathbf{\Phi}_u^T\mathbf{\Gamma}^T(\mathbf{Z}_0 - \mathbf{V}_0) \\
& + \rho_3\mathbf{\Phi}_l^T(\mathbf{Z}_1 - \mathbf{V}_1) + \rho_3\mathbf{\Phi}_u^T(\mathbf{Z}_2 - \mathbf{V}_2)\big]. \qquad (16)
\end{aligned}
$$

Optimizing $\mathbf{C}_l$ and $\mathbf{C}_u$ The subproblem with respect to $\mathbf{C}_l$ and $\mathbf{C}_u$ is

$$
\min_{\mathbf{C}_l,\mathbf{C}_u} \lambda_1\left\| \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix}^T \right\|_{2,1} + \frac{\rho_1}{2}\left\| \begin{bmatrix}\mathbf{\Phi}_l \\ \mathbf{\Phi}_u\end{bmatrix} - \begin{bmatrix}\mathbf{C}_l \\ \mathbf{C}_u\end{bmatrix} + \begin{bmatrix}\mathbf{\Lambda}_l \\ \mathbf{\Lambda}_u\end{bmatrix}\right\|_F^2.
$$

Let $\mathbf{C} = [\mathbf{C}_l; \mathbf{C}_u]$, $\mathbf{\Phi} = [\mathbf{\Phi}_l; \mathbf{\Phi}_u]$, and $\mathbf{\Lambda} = [\mathbf{\Lambda}_l; \mathbf{\Lambda}_u]$. The problem can be decomposed into small problems, i.e.,

$$\mathbf{C}_{\cdot i} = \arg\min_{\mathbf{C}_{\cdot i}} \lambda_1 ||\mathbf{C}_{\cdot i}||_2 + \frac{\rho_1}{2} ||\mathbf{\Phi}_{\cdot i} - \mathbf{C}_{\cdot i} + \mathbf{\Lambda}_{\cdot i}||_F^2,$$

where $\mathbf{C}_{\cdot i}$ is the i-th column of $\mathbf{C}$, $\mathbf{\Phi}_{\cdot i}$ is the i-th column of $\mathbf{\Phi}$, and $\mathbf{\Lambda}_{\cdot i}$ is the i-th column of $\mathbf{\Lambda}$. The solution is

$$\mathbf{C}_{\cdot i} = S_{\lambda_1/\rho_1}(\mathbf{\Phi}_{\cdot i} + \mathbf{\Lambda}_{\cdot i}), \tag{17}$$

where $S_k(\mathbf{a}) = [1 - \frac{k}{||\mathbf{a}||_2}]_+ \odot \mathbf{a}$ and $S_k(0) = 0$. $[\cdot]_+ = \max(\cdot, 0)$. $\odot$ represents pairwise product.

Optimizing $\mathbf{Z}_0$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ The subproblems with respect to $\mathbf{Z}_0$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are

$$\min_{\mathbf{Z}_0} \mathbf{I}_-(\mathbf{Z}_0) + \frac{\rho_2}{2} ||\mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0 + \mathbf{V}_0||_F^2 - \frac{\rho_2}{2} ||\mathbf{V}_0||_F^2,$$
$$\min_{\mathbf{Z}_1} \mathbf{I}_+(\mathbf{Z}_1) + \frac{\rho_3}{2} ||\mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1 + \mathbf{V}_1||_F^2 - \frac{\rho_3}{2} ||\mathbf{V}_1||_F^2,$$
$$\min_{\mathbf{Z}_2} \mathbf{I}_+(\mathbf{Z}_2) + \frac{\rho_3}{2} ||\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2 + \mathbf{V}_2||_F^2 - \frac{\rho_3}{2} ||\mathbf{V}_2||_F^2.$$

The solutions are

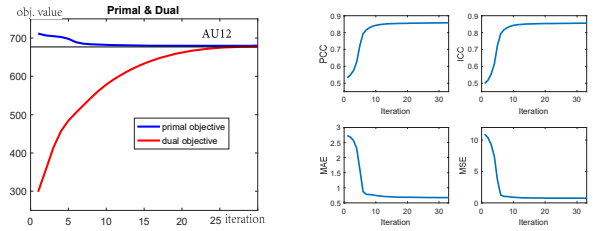
$$\mathbf{Z}_0 = \min\{0, \mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} + \mathbf{V}_0\}, \tag{18}$$
$$\mathbf{Z}_1 = \max\{0, \mathbf{\Phi}_l\mathbf{w} + \mathbf{V}_1\}, \tag{19}$$
$$\mathbf{Z}_2 = \max\{0, \mathbf{\Phi}_u\mathbf{w} + \mathbf{V}_2\}. \tag{20}$$

Optimizing $\mathbf{\Lambda}_l$, $\mathbf{\Lambda}_u$, $\mathbf{V}_0$, $\mathbf{V}_1$, and $\mathbf{V}_2$ The Lagrange multipliers can be updated as

$$\mathbf{\Lambda}_l \leftarrow \mathbf{\Lambda}_l + \mathbf{\Phi}_l - \mathbf{C}_l, \tag{21}$$
$$\mathbf{\Lambda}_u \leftarrow \mathbf{\Lambda}_u + \mathbf{\Phi}_u - \mathbf{C}_u, \tag{22}$$
$$\mathbf{V}_0 \leftarrow \mathbf{V}_0 + \mathbf{\Gamma}\mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_0, \tag{23}$$
$$\mathbf{V}_1 \leftarrow \mathbf{V}_1 + \mathbf{\Phi}_l\mathbf{w} - \mathbf{Z}_1, \tag{24}$$
$$\mathbf{V}_2 \leftarrow \mathbf{V}_2 + \mathbf{\Phi}_u\mathbf{w} - \mathbf{Z}_2. \tag{25}$$

3. Convergence of the Proposed Algorithm

In Section 4.1, we briefly present the convergence of the algorithm. Here, we present the testing performance in all evaluation criteria as the iteration proceeds. Fig. 1(a) shows the learning curve of AU12 on FERA 2015 under the scenario that 6% of training frames are annotated. Fig. 1(b) shows corresponding performance on the testing set at each iteration. As shown in Fig. 1, the decrease of the primal objective and the improvement of the performance happen within in the first 10 steps. Then, the objective changes slowly and converges within 30 iterations.



(a) (b)

Figure 1. Convergence of the algorithm. (a) The learning curve of AU12 on FERA 2015. (b) The performance of AU12 on the testing set at different iterations.

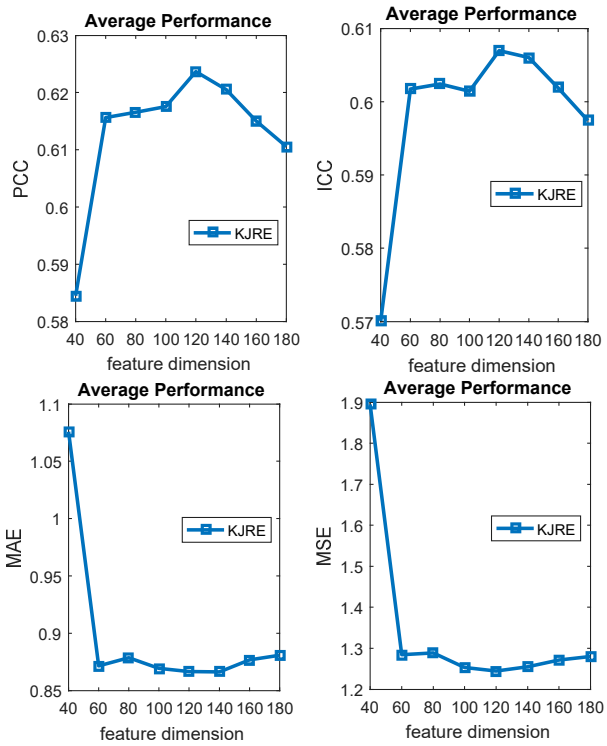


Figure 2. Influence of the feature dimension.

4. Influence of Feature Dimension

In the main paper, the feature dimension K is treated as a hyperparameter which can be selected though cross-validation. Here we show the average performance under different K values. The experiment is performed on FERA 2015 under the annotation rate of 6%. The results are shown in Fig. 2. When the feature dimension is less than 60, the performance is poor. The performance starts to decrease when the feature dimension is larger than 120.

5. Visualization of Learned Representation

The comparison between the learned and original representations of 3,000 test samples of AU12 in BP4D is shown in Fig. 3. The feature dimension is 219D for the original representation and 120D for the learned representation. We

use t-sne [1] to project high dimensional samples into a 2D space. Samples with the same or close intensities are clustered better in the learned space than in the original space.

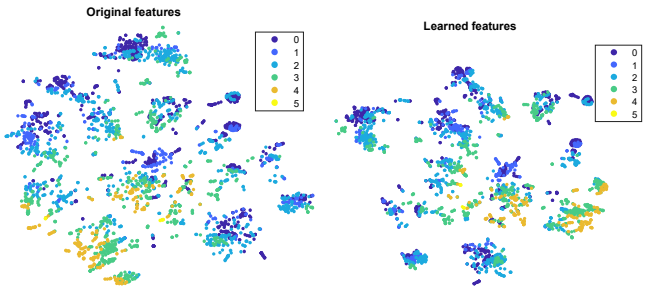


Figure 3. Comparison between the learned representation and the original representation. Left: original representation. Right: learned representation.

References

[1] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014. 4

[2] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 2