Multi-View Exclusive Unsupervised Dimension Reduction for Video-Based Facial Expression Recognition

Liping Xie[†‡], **Dacheng Tao**[‡], **Haikun Wei**[†]
[†]Key Laboratory of Measurement and Control of CSE, Ministry of Education,
School of Automation, Southeast University, Nanjing 210096, China
[‡]QCIS and FEIT, University of Technology Sydney, Australia
lpxie2007@gmail.com, dacheng.tao@uts.edu.au, hkwei@seu.edu.cn

Abstract

Video-based facial expression recognition (FER) has recently received increased attention as a result of its widespread application. Many kinds of features have been proposed to represent different properties of facial expressions in videos. However the dimensionality of these features is usually high. In addition, due to the complexity of the information available in video sequences, using only one type of feature is often inadequate. How to effectively reduce the dimensionality and combine multi-view features thus becomes a challenging problem. In this paper, motivated by the recent success in exclusive feature selection, we first introduce exclusive group LASSO (EG-LASSO) to unsupervised dimension reduction (UDR). This leads to the proposed exclusive UDR (EUDR) framework, which allows arbitrary sparse structures on the feature space. To properly combine multiple kinds of features, we further extend EUDR to multi-view EUDR (MEUDR), where the structured sparsity is enforced at both intra- and inter-view levels. In addition, combination weights are learned for all views to allow them to contribute differently to the final consensus presentation. A reliable solution is then obtained. Experiments on two challenging video-based FER datasets demonstrate the effectiveness of the proposed method.

1 Introduction

Automatic facial expression recognition (FER) plays an important role in pattern recognition and computer vision. Its extensive applications include: human-computer interaction systems, psychology, and security. Over the past decade, there has been extensive productive and fruitful study of static images [Gu *et al.*, 2012]. In reality, however, facial expression activity is dynamic, and its variability can be described as the onset, the apex and the offset [Xie *et al.*, 2014]. Many experiments, including those conducted in psychology [Ambadar *et al.*, 2005], have demonstrated that the utilization of the temporal information located in facial expression activity enhances recognition performance. Temporal information is thus an essential component of a successful FER system, and video-based FER has attracted particular attention [Zhao and Pietikainen, 2007; Xie *et al.*, 2014].

Numerous methods for video-based FER have been proposed, but they are all limited in that only a single type of feature is utilized. This may lead to unsatisfactory recognition results because the facial expression information in videos is very complex. Combining the different kinds of features has significant potential to improve the performance of video-based FER since each feature characterizes a different view. For example, spatio-temporal features based on the scale invariant feature transform (SIFT) [Gu *et al.*, 2012] and histograms of oriented gradients (HOG) [Klaser *et al.*, 2008] descriptors effectively characterize the respective shape and appearance of an expression, while the features based on the motion boundary histograms (MBH) [Wang *et al.*, 2013a] descriptor are especially good at capturing the motion information, and the widely used texture feature LBP-TOP [Zhao and Pietikainen, 2007] is particularly insensitive to monotonic gray-scale changes. In this paper, we treat each feature representation as a particular view for charactering facial expression in videos.

It has been demonstrated empirically in [Wang *et al.*, 2013a] that combining multiple features tends to achieve better recognition accuracy in action recognition, simply as a result of concatenating different features. However, the simple concatenation strategy is not physically meaningful because each view has a specific statistical property, and it often leads to over-fitting due to the high dimensionality [Luo *et al.*, 2016; Tao *et al.*, 2009] of the spatio-temporal features. For example, the dimensionality of the LBP-TOP features is about 3000 if 4×4 blocks are adopted for each frame, and the codebook size of the bag-of-features representations for HOG, HOF and MBH can also be several thousands. To properly combine the different views and also reduce feature dimensionality, we develop a novel multi-view dimension reduction (MVDR) algorithm, which aims to find a low dimensional representation for heterogeneous high dimensional data. MVDR can be performed in a supervised, semi-supervised, or unsupervised manner; most of the current works are unsupervised due to the high labeling cost in many real-world applications. We also focus on the unsupervised setting in this paper.

Inspired by the recently proposed exclusive feature selection [Kong *et al.*, 2014], we first propose a novel sparse framework for unsupervised dimensionality reduction (UDR) termed exclusive UDR (EUDR), which learns a low-dimensional and sufficiently informative pattern for the original feature [Tao *et al.*, 2007]. The exclusive group LASSO [Kong *et al.*, 2014] (EG-LASSO) is employed as a regularization term on the corresponding projection matrix. The main advantage of EUDR is that it allows arbitrary group structures being exploited on the feature space. To also deal with the multi-view features in video-based FER, we extend EUDR to multi-view exclusive UDR (MEUDR), which simultaneously combines multi-view features and reduces the dimensionality. In MEUDR, the structure sparsity brought by the EG-LASSO regularization is achieved at both intra- and inter-view levels. Therefore, the complementary nature of different views tends to be better exploited than it is in existing multi-view unsupervised dimensionality reduction (MUDR) methods. In addition, combination weights are learned for each view to allow various views to contribute differently to the final representation. Thus the model is robust to noisy views and complementarity exploration can be further enhanced.

To validate the effectiveness of the proposed MEUDR for FER, we conduct experiments on two challenging video-based FER datasets. The superiority of our method is obvious by comparing it with several competitive baselines and recently proposed MUDR approaches.

2 MEUDR: Multi-view Exclusive Unsupervised Dimension Reduction

Notations: For a matrix X, we use $X_{(i,:)}$ and $X_{(:,j)}$ to signify its i-th row and j-th column vector respectively. X_{ij} denotes the (i,j)-element of matrix X. I_r denotes an $r \times r$ identity matrix, $\|P\|_F$ is the Frobenius norm of matrix P, $\|p\|_1$ denotes the l_1-norm of vector p, and B^T is the transpose of B.

2.1 Problem formulation

Suppose we are given the original feature matrix $X \in \mathbb{R}^{n \times d}$, where n is the total number of samples, d is the dimensionality of high-dimensional data points. A basic formulation for UDR is to minimize the reconstruction errors of the original data represented by the matrix X, i.e.,

$$\begin{gathered}\arg \min_{\{B,P\}} \frac{1}{n}\|X - BP^T\|_F^2 + \gamma\|P\|_F^2, \\ \text{s.t. } B^T B = I_r.\end{gathered} \tag{1}$$

where $P = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_r] \in \mathbb{R}^{d \times r}$ is the projection matrix that maps the original high-dimensional data to a low-dimensional subspace, and $B \in \mathbb{R}^{n \times r}, r < d$ is the low-dimensional representation in the subspace. Although this formulation efficiently finds a low-dimensional representation, all input variables are encouraged to contribute to each dimension of the final representation. To remove junk dimensions and discern important ones, we propose to adaptively select variables for constructing each projection vector $\mathbf{p}_i \in \mathbb{R}^d$. This leads to the following general formulation of the sparse unsupervised dimension reduction (SUDR):

$$\begin{gathered}\arg \min_{\{B,P\}} \frac{1}{n}\|X - BP^T\|_F^2 + \gamma\sum_{i=1}^{r}\Omega(\mathbf{p}_i), \\ \text{s.t. } B^T B = I_r.\end{gathered} \tag{2}$$

where Ω is any convex sparse-promoting penalty. In this paper, we adopt the recently proposed exclusive group LASSO (EG-LASSO) [Kong *et al.*, 2014] as the regularizer for $\mathbf{p}_i$. EG-LASSO first adopts a general "group" setting to obtain arbitrary group structure of $\mathbf{p}_i$, following which l_1/l_2 norm penalty is used to achieve sparsity. In this way, this regularizer brings out sparsity with arbitrary structure on feature space. Therefore we obtain the following exclusive UDR (EUDR) problem:

$$\begin{gathered}\arg \min_{\{B,P\}} \frac{1}{n}\|X - BP^T\|_F^2 + \gamma\sum_{i=1}^{r}\Omega_{E_g}^{\mathcal{G}}(\mathbf{p}_i), \\ \text{s.t. } B^T B = I_r.\end{gathered} \tag{3}$$

where $\Omega_{E_g}^{\mathcal{G}}(\mathbf{p}_i) = \sum_{g \in \mathcal{G}} \|\mathbf{p}_{i,\mathcal{G}_g}\|_1^2$ is the EG-LASSO term, and $\mathcal{G}$ is the group set. However, this framework is limited in that only problem of UDR with single view data is available. To deal with the multi-view features in video-based FER, we extend it to multi-view dimensionality reduction (MVDR) by mapping the features of different views $X^{(v)}$ into a common subspace. Therefore we have the following multi-view exclusive unsupervised dimension reduction (MEUDR) formulation:

$$\begin{gathered}\min_{B,\{P^{(v)}\},\boldsymbol{\theta}} \frac{1}{n}\sum_{v=1}^{V}\theta_v\|X^{(v)} - B(P^{(v)})^T\|_F^2 \\ + \gamma\sum_{v=1}^{V}\sum_{i=1}^{r}\Omega_{E_g}^{\mathcal{G}}(\mathbf{p}_i^{(v)}) + \frac{\eta}{2}\|\boldsymbol{\theta}\|_2^2, \\ \text{s.t. } B^T B = I_r; \sum_{v=1}^{V}\theta_v = 1, \theta_v \geq 0.\end{gathered} \tag{4}$$

where $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_V]$ is a vector of view combination coefficients [Nie *et al.*, 2014] to explore the complementary information of different views, and also prevent the model from being contaminated by noisy views. To further exploit the feature relationships between different views, we reformulate Eq.(4) as:

$$\begin{gathered}\min_{B,\{P^{(v)}\},\boldsymbol{\theta}} \frac{1}{n}\sum_{v=1}^{V}\theta_v\|X^{(v)} - B(P^{(v)})^T\|_F^2 \\ + \gamma\sum_{i=1}^{r}\Omega_{E_g}^{\mathcal{G}}(\mathbf{p}_i^{cat}) + \frac{\eta}{2}\|\boldsymbol{\theta}\|_2^2, \\ \text{s.t. } B^T B = I_r; \sum_{v=1}^{V}\theta_v = 1, \theta_v \geq 0.\end{gathered} \tag{5}$$

Here, $\mathbf{p}_i^{cat} = [\mathbf{p}_i^{(1)}; \mathbf{p}_i^{(2)}; ...; \mathbf{p}_i^{(V)}]$ is a concatenation of projection vectors, and the group set $\mathcal{G}$ consists of two parts: we first regard the features in the same view as in a group to bring sparsity at intra-view level; then we put the highly correlated features of different views in a group to bring sparsity at inter-view level, where the feature correlation matrix is calculated as $R = (R_{ij}) \in \mathbb{R}^{n\times n}$, and R_{ij} denotes the correlations between i-th and j-th features, i.e., $R_{ij} = \frac{|\sum_t X_{it}X_{jt}|}{\sqrt{\sum_t X_{it}^2}\sqrt{\sum_t X_{jt}^2}}$. Benefiting from this two-part group setting, both the feature relationships within each view and the complementary nature between different views are well exploited.

2.2 Optimization

The global solution to the optimization problem in Eq.(5) is difficult to achieve since it is not joint convex with respect to the set of variables $(B, \{P^{(v)}\}, \boldsymbol{\theta})$. We therefore present an alternative iterative algorithm to solve the problem by converting the original problem into three sub-problems, in which only one variable is updated. First, we reformulate Eq.(5) as follows:

$$\begin{aligned} \min F(B,P,\boldsymbol{\theta}) &= \frac{1}{n}\|\hat{X} - BP^T\Theta_d\|_F^2 \\ &\quad + \gamma\sum_{i=1}^{r}\sum_{g\in\mathcal{G}}\|\boldsymbol{p}_{i,\mathcal{G}_g}^{cat}\|_1^2 + \frac{\eta}{2}\|\boldsymbol{\theta}\|_2^2, \qquad (6) \\ \text{s.t. } B^TB &= I_r; \sum_{v=1}\theta_v = 1, \theta_v \geq 0, v = 1, ..., V. \end{aligned}$$

where $\hat{X} = [\sqrt{\theta_1}X^{(1)}, ..., \sqrt{\theta_V}X^{(V)}] \in \mathbb{R}^{n\times d}$ is the concatenated feature matrix with the dimension $d = \sum_{v=1}^V d_v$, $P = [P^{(1)}; ...; P^{(V)}] \in \mathbb{R}^{d\times r}$ is the concatenated mapping matrices, and $\Theta_d \in \mathbb{R}^{d\times d}$ is a block diagonal matrix in which the v-th block is a diagonal matrix $\sqrt{\theta_v}I_v$. Then we update B, P and $\boldsymbol{\theta}$ alternatively until the termination criterion is achieved.

When P and $\boldsymbol{\theta}$ are fixed, the problem Eq.(6) becomes:

$$\min F(B) = \|\hat{X} - BP^T\Theta_d\|_F^2, \text{ s.t. } B^TB = I_r. \qquad (7)$$

According to [Han *et al.*, 2012], we first apply SVD to $E = \hat{X}\Theta_d^TP$, i.e., $E = USV$, and then obtain the solution for B as $B = UV$.

Next, when B and $\boldsymbol{\theta}$ are fixed, the problem Eq.(6) w.r.t. P is given as follows:

$$\min F(P) = \frac{1}{n}\|\hat{X} - (\sum_{i=1}^{r}\boldsymbol{b}_i\boldsymbol{p}_i^T)\Theta_d\|_F^2 + \gamma\sum_{i=1}^{r}\sum_{g\in\mathcal{G}}\|\boldsymbol{p}_{i,\mathcal{G}_g}^{cat}\|_1^2. \qquad (8)$$

where $\boldsymbol{b}_i = B_{(:,i)}$ is the i-th column of B. This problem can be solved by alternating for each $\boldsymbol{p}_i$ until convergence. The above problem w.r.t. $\boldsymbol{p}_i$ is given by:

$$\min F(\boldsymbol{p}_i) = \frac{1}{n}\|\tilde{X} - \boldsymbol{b}_i\boldsymbol{p}_i^T\Theta_d\|_F^2 + \gamma\sum_{g\in\mathcal{G}}\|\boldsymbol{p}_{i,\mathcal{G}_g}^{cat}\|_1^2. \qquad (9)$$

where $\tilde{X} = \hat{X} - (\sum_{k\neq i}\boldsymbol{b}_k\boldsymbol{p}_k^T)\Theta_d$. It is easy to verify that the optimal solution of Eq.(9) is exactly the solution of the following problem:

$$\min G(\boldsymbol{p}_i) = \frac{1}{n}\|\tilde{X} - \boldsymbol{b}_i\boldsymbol{p}_i^T\Theta_d\|_F^2 + \gamma\boldsymbol{p}_i^TD^i\boldsymbol{p}_i. \qquad (10)$$

where D^i is a diagonal matrix with the entry $D_{jj}^i = (\sum_g \frac{(I_{\mathcal{G}_g})_j\|\boldsymbol{p}_{i,\mathcal{G}_g}^{cat}\|_1}{|p_{ji}|})$, where $I_{\mathcal{G}_g} \in \{0,1\}^{d\times 1}$ is a vector of the group index indicator, and $|p_{ji}|$ is replaced by $\sqrt{p_{ji}^2 + \epsilon}$ when $p_{ji} = 0$ [Kong *et al.*, 2014]. Taking the derivative of the objective with respect to $\boldsymbol{p}_i$ and setting it to zero, we have:

$$-\frac{2}{n}\Theta_d\tilde{X}^T\boldsymbol{b}_i + \frac{2}{n}(\boldsymbol{b}_i^T\boldsymbol{b}_i)(\Theta_d\Theta_d^T)\boldsymbol{p}_i + 2\gamma D^i\boldsymbol{p}_i = 0. \qquad (11)$$

Then we obtain:

$$\boldsymbol{p}_i = \frac{1}{n}(\frac{1}{n}(\boldsymbol{b}_i^T\boldsymbol{b}_i)(\Theta_d\Theta_d^T) + \gamma D^i)^{-1}\Theta_d\tilde{X}^T\boldsymbol{b}_i. \qquad (12)$$

Note that D^i is dependent on P, and thus is also an unknown variable. Fortunately, we can prove that the solution for $\boldsymbol{p}_i$ can be obtained by repeating the following two steps until convergence:

- Calculate the diagonal matrix $D_{\tau+1}^i$ using $\boldsymbol{p}_i^\tau$;
- Update $\boldsymbol{p}_i^{\tau+1} = \frac{1}{n}(\frac{1}{n}(\boldsymbol{b}_i^T\boldsymbol{b}_i)(\Theta_d\Theta_d^T) + \gamma D_{\tau+1}^i)^{-1}\Theta_d\tilde{X}^T\boldsymbol{b}_i$.

This iteration can also be incorporated into the alternation of different $\boldsymbol{p}_i, i = 1, ..., r$.

Lastly, for fixed P and B, we can rewrite the original problem Eq.(6) as:

$$\begin{aligned} \min F(\boldsymbol{\theta}) &= \boldsymbol{\theta}^T\boldsymbol{q} + \frac{\eta}{2}\|\boldsymbol{\theta}\|_2^2, \\ \text{s.t. } \sum_{v=1}\theta_v &= 1, \theta_v \geq 0, v = 1, ..., V. \end{aligned} \qquad (13)$$

where $\boldsymbol{q} = \frac{1}{n}[q_1, ..., q_V]^T$ with each $q_v = \|X^{(v)} - B(P^{(v)})^T\|_F^2$. We adopt the coordinate descent algorithm [Huang *et al.*, 2015] to solve Eq.(13). Therefore, in each iteration of the descent procedure, only two elements θ_i and θ_j are selected to be updated; the others are fixed. By using the Lagrangian of Eq.(13) and considering the sum to one constraint, we obtain the following updating rule:

$$\begin{cases} \theta_i^* = \dfrac{\eta(\theta_i + \theta_j) + (q_j - q_i)}{2\eta}, \\ \theta_j^* = \theta_i + \theta_j - \theta_i^* \end{cases} \qquad (14)$$

The obtained θ_i^* or θ_j^* may violate the constraint $\theta_i \geq 0$. Thus we set $\theta_i^* = 0$ if $\eta(\theta_i + \theta_j) + (q_j - q_i) < 0$, and similarly for θ_j^*.

We summarize the main procedure of the optimization in Algorithm 1. The stopping criterion for terminating the algorithm is the difference of the objective value between two consecutive steps. That is, if $|O_{t+1} - O_t|/|O_t| < \epsilon$, then the iteration stops, where O_t is the objective value at the i-th iteration step. Since the sub-problems of Eq. (7), (8) and (13) are convex w.r.t. B, P, and $\boldsymbol{\theta}$ respectively, the algorithm is guaranteed to converge according to the following analysis.

2.3 Convergence analysis

In this section, we discuss the convergence of the proposed MEUDR algorithm. Let the initialized value of the objective Eq.(6) be $F(B^t, P^t, \boldsymbol{\theta}^t)$. Since Eq.(7) is convex and

Algorithm 1 The efficient iterative algorithm for solving Eq.(6)

```
Input: A matrix of the concatenated features $X = [X^{(1)}, ..., X^{(V)}] \in \mathbb{R}^{n \times d}$.
Output: A low-dimensional consensus pattern matrix $B \in \mathbb{R}^{n \times r}$, and a set of mapping matrices $P = [P^{(1)}; ...; P^{(V)}] \in \mathbb{R}^{d \times r}$.
Set $t = 0$. Initialize $P^0$ as random matrix, and $\theta_v^0 = \frac{1}{V}, v = 1, ..., V$.
While not converge do
Calculate $B^{t+1} = UV$, where $U$ and $V$ are obtained by SVD on $E = \hat{X}(\Theta_d^t)^T P^t$.
Set $\tau = 0$.
Repeat
    For $i = 1, ...r$
        Calculate the diagonal matrix $D_{\tau+1}^i$ based on $\boldsymbol{b}_i^{t+1}$ and $\boldsymbol{p}_{i,\tau}^t$;
        Update $\boldsymbol{p}_{i,\tau+1}^t = \frac{1}{n}(\frac{1}{n}((\boldsymbol{b}_i^{t+1})^T \boldsymbol{b}_i^{t+1})(\Theta_d^t(\Theta_d^t)^T) + \gamma D_{\tau+1}^i)^{-1} \Theta_d^t \tilde{X}^T \boldsymbol{b}_i^{t+1}$.
    End for
    $\tau = \tau + 1$.
Until converge
Update $P^{t+1} = [\boldsymbol{p}_{1,\tau+1}^t, ..., \boldsymbol{p}_{r,\tau+1}^t]$.
Calculate $\boldsymbol{\theta}^{t+1}$ by using the update rule Eq.(14).
$t = t + 1$.
End while
```

can be solved analytically, we have $F(B^{t+1}, P^t, \boldsymbol{\theta}^t) \leq F(B^t, P^t, \boldsymbol{\theta}^t)$. The problem in Eq. (8) is solved alternatively for each $\boldsymbol{p}_i$ with all the other $\boldsymbol{p}_k, k \neq i$ fixed. Let $F(\boldsymbol{p}_i^t)$ and $G(\boldsymbol{p}_i^t)$ be the objectives of Eq.(9) and (10) respectively at the t-th alternating step. Similar to [Kong *et al.*, 2014], we can prove that by using the updating rules for calculating $\boldsymbol{p}_i$, the following inequalities hold:

$$G(\boldsymbol{p}_{i,\tau+1}^t) < G(\boldsymbol{p}_{i,\tau}^t). \tag{15}$$

and

$$(F(\boldsymbol{p}_{i,\tau+1}^t) - F(\boldsymbol{p}_{i,\tau}^t)) < (G(\boldsymbol{p}_{i,\tau+1}^t) - G(\boldsymbol{p}_{i,\tau}^t)). \tag{16}$$

Therefore, we have $F(\boldsymbol{p}_{i,\tau+1}^t) \leq F(\boldsymbol{p}_{i,\tau}^t)$. This indicates that $F(\boldsymbol{p}_{i,\tau+1}^t, \{\boldsymbol{p}_{k,\tau}^t\}_{k \neq i}) \leq F(\{\boldsymbol{p}_{i,\tau}^t\}_{i=1}^r) = F(P^t)$. Because $F(P^{t+1}) = F(\{\boldsymbol{p}_{i,\tau+1}^t\}_{i=1}^r)$, we have $F(B^{t+1}, P^{t+1}, \boldsymbol{\theta}^t) \leq F(B^{t+1}, P^t, \boldsymbol{\theta}^t)$. Lastly, because Eq.(13) is a convex problem, we have $F(B^{t+1}, P^{t+1}, \boldsymbol{\theta}^{t+1}) \leq F(B^{t+1}, P^{t+1}, \boldsymbol{\theta}^t)$. This completes the proof.

2.4 Induction for out-of-sample data

In this section, we extend MEUDR to the out-of-sample data.

We use $x = [x^{(1)}, ..., x^{(V)}] \in \mathbb{R}^d$ to signify the new multi-view data point. And we expect to find its low-dimensional representation $b \in \mathbb{R}^r$. Since the mapping matrices P and view combination coefficients $\theta = [\theta_1, \theta_2, ..., \theta_V]$ have been derived based on the ever-known data, they can be used directly for the following solution. Based on the same strategy, which we have used in MEUDR, we expect b to be the solution to the following problem:

$$\begin{aligned} \min F(b) &= \|\hat{x} - bP^T \Theta_d\|_F^2, \\ \text{s.t. } B^T B &= I_r. \end{aligned} \tag{17}$$

where $\hat{x} = [\sqrt{\theta_1} x^{(1)}, ..., \sqrt{\theta_V} x^{(V)}]$. After applying SVD to $E = \hat{x}\Theta_d^T P$, i.e., $E = USV$, we obtain the solution for b as $b = UV$.

It can be seen from the above analysis that the consensus representation for an out-of-sample data can be obtained easily.

2.5 Related work

Multi-view learning [Wang *et al.*, 2013b; Luo *et al.*, 2015; Xu *et al.*, 2015; Cai *et al.*, 2014] has received much attention recently. The multiple views we refer to here are the multiple distinct feature sets used to describe a given sample, which are different from the multiple viewpoints in the traditional multi-view FER [Moore and Bowden, 2011]. We can roughly group the multi-view learning algorithms into categories according to their learning mechanisms, such as weighted view combination [McFee and Lanckriet, 2011], multi-view dimension reduction (MVDR) [Han *et al.*, 2012], and so on. In MVDR, irrelevant or redundant information in the multi-view data can be removed by leveraging the dependency, coherence, and complementarity of the different views.

Canonical Correlation Analysis (CCA) [Luo *et al.*, 2015] is one of the most representative unsupervised MVDR (UMDR) methods, but it is limited in that only data from two views can be handled. Distributed spectral embedding (DSE) [Long *et al.*, 2008] is a general UMDR approach for handling data from an arbitrary number of views. Although simple and efficient, the complementary nature of different views, which is critical in multi-view learning, is not well explored in DSE [Han *et al.*, 2012]. To address this issue, structured sparsity based UMDR (SSMVD) is proposed in [Han *et al.*, 2012], which is the most similar work to our method. SSMVD imposes a structured sparsity-inducing norm on the projection matrix which maps different patterns to the common low-dimensional space, and thus allows flexible information sharing in certain subsets of patterns across multiple views. In spite of this advantage, SSMVD still has several drawbacks: 1) the dimensions of the patterns of different views in SSMVD must be the same to construct a 2-D grid, thus a limited (certain and specified) number of relationships (e.g., rectangular groups) between the feature spaces of multiple views can be exploited; 2) the orthogonality constraint on the final common representation is replaced by a bound constraint in SSMVD for the convenience of optimization. The explainable efficiency of the factors in the final common subspace is not strong enough, and redundancy among the factors is inevitable; 3) only the information shared across different views is considered, and the sparsity structure in each view is ignored. Additionally, most of the existing UMDR methods (including CCA, DSE, and SSMVD) only consider the relationships at the inter-view level and ignore those at the intra-view level. All these problems are specifically tackled in the proposed MEUDR method.

3 Experiments

In this section, we validate the effectiveness of the proposed MEUDR method on two challenging FE video datasets. Prior to the evaluations, we present the datasets and features used, as well as our experimental settings.

3.1 Datasets

The first dataset is the facial expression (FE) dataset proposed in [Dollár *et al.*, 2005], which we call "FE05" in this paper. The FE05 dataset involves two individuals, each of whom expresses six different emotions under two lighting setups. In our experiments, we choose two subsets that belong to two different identities and have different lighting setups, making the resultant dataset sufficiently challenging. To observe the performance of the compared algorithms with respect to different numbers of labeled samples, we randomly select 0.25, 0.5, 0.75, 1.0 percent of samples (2, 4, 6, 8 out of 8 samples) for each expression as the labeled samples from the training set, and the remaining samples are regarded as unlabeled.

The second FE dataset is the Oulu-CASIA VIS (CASIA for short) database [Li *et al.*, 2013]. The CASIA dataset contains 80 subjects. We randomly separate these subjects into two groups: 70 subjects with six expressions (420 samples) for training, 10 subjects (60 samples) for testing. A further 20, 30, 50, 70 subjects are randomly chosen as labeled samples from the training set.

The five-fold cross-validation strategy is adopted for both datasets for tuning the parameters. Both support vector machine (SVM) [Tao *et al.*, 2006; Liu and Tao, 2015] and kNN classifiers are tested for recognition, and ultimately we choose kNN for FE05 and SVM for CASIA according to classification accuracy.

3.2 Feature extraction

In our experiment, four different types of visual features have been extracted: Local Binary Pattern on Three Orthogonal Planes (LBP-TOP), and three descriptors HOG, HOF, and MBH combined with dense trajectories. The dimensionality of the features is 2832, 2000, 2000 and 2000, respectively.

LBP-TOP [Zhao and Pietikainen, 2007] is one of the most widely used features in texture analysis of image sequences. It is insensitive not only to translation and rotation, but also to monotonic gray-scale changes. Dense trajectories [Wang *et al.*, 2013a] is a state-of-art approach that computes local descriptors for action recognition, and we apply it to FER. We adopt three types of descriptor to describe the video: HOG, HOF, and MBH. HOG (histograms of oriented gradients) captures static appearance information by using the orientation and magnitude of gradient; HOF (histograms of optical flow) focuses on the motion information by using the orientation and magnitude of the flow field; MBH (motion binary histograms) encodes the gradient of horizontal and vertical components of the flow to capture the relative pixel motion. Therefore, the different features are complementary to one another.

3.3 Comparison methods

We compare our method with BSV, CAT, PCA, DSE, and SSMVD. The first three are different baseline methods, and the experimental setup of the compared approaches is given as follows:

- BSV: the best single view method, i.e., performing FER by regarding each single view as the feature representation, the results of the view that achieves the best performance are reported;
- CAT: concatenating the normalized features of all different views, and then performing FER on the concatenated features;
- dPCA: the distributed PCA method, i.e., reducing the dimensionality of each view to a pre-defined value (such as 100 for both datasets in this paper) by PCA, and then concatenating all the different views as a long vector for FER.
- DSE: the distributed spectral embedding [Long *et al.*, 2008] method for UMDR. PCA is adopted for finding the patterns $A^{(v)}, v = 1, ..., V$, and the reduced dimensionality is 100.
- SSMVD: the structured sparsity-based UMDR approach proposed in [Han *et al.*, 2012]. Similar to DSE, PCA is adopted to find the patterns, the dimensionality of each of which is set to be the same (e.g., 100) so that a 2-D grid can be constructed. The trade-off parameter γ is chosen from the set $\{10^i | i = -5, -4, ..., 3, 4\}$.
- MEUDR: the proposed multi-view UDR method based on exclusive group sparsity. To reduce the time cost and also avoid over-fitting, since the number of training samples in both datasets is limited , we also apply PCA to the original features as in DSE and SSMVD. The candidate set for both trade-off parameters γ and η is $\{10^i | i = -5, -4, ..., 3, 4\}$.

3.4 Experimental results

The performance of the compared methods in relation to the dimension of the final (consensus) representation on the FE05 and CASIA dataset is shown in Figure 1 and Figure 2 respectively. Accuracy is averaged over five runs for each dimension r in $\{1, 2, 5, 8, 10, 15, 20, 30, 50, 80, 100, 150, 200, 300\}$. We summarize the performance of the various methods at their best dimensions in Table 1 and Table 2 respectively on the two datasets.

On the FE05 dataset, we observe that: 1) concatenating all features (CAT) is usually superior to using only a single view features (BSV) because more information is utilized by involving more features, although this is not true in all cases. For example, when the percentage of labeled samples is 0.5, CAT is a bit worse than BSV. By first applying PCA on each view and then concatenating, we obtain better results than BSV. This indicates that the simple concatenation strategy may fail due to over-fitting; 2) all the UMDR methods (DSE, SSMVD, and MEUDR) can be significantly better than the baselines (BSV, CAT, and dPCA), if the dimensionalities are properly set according to cross-validation. For DSE, the performance curve stops early because the solution is obtained by the SVD of an $n \times n$ matrix, and the final dimension should be less than n; 3) the accuracy of the

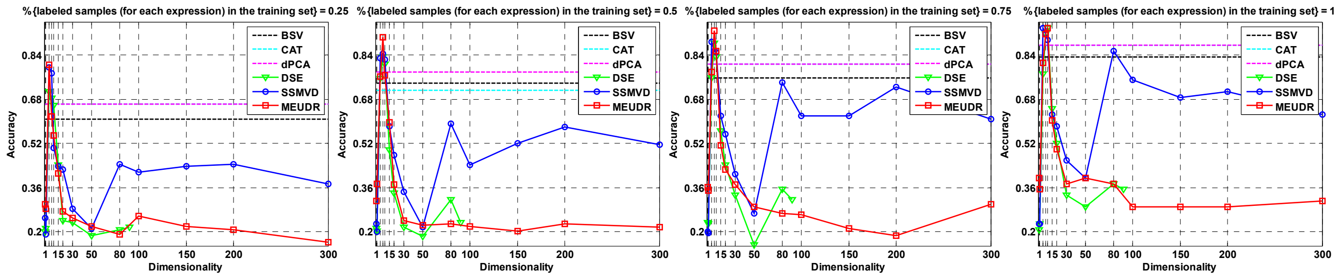

Figure 1: Classification accuracy vs. the dimensionality of the result data on the FE05 dataset.

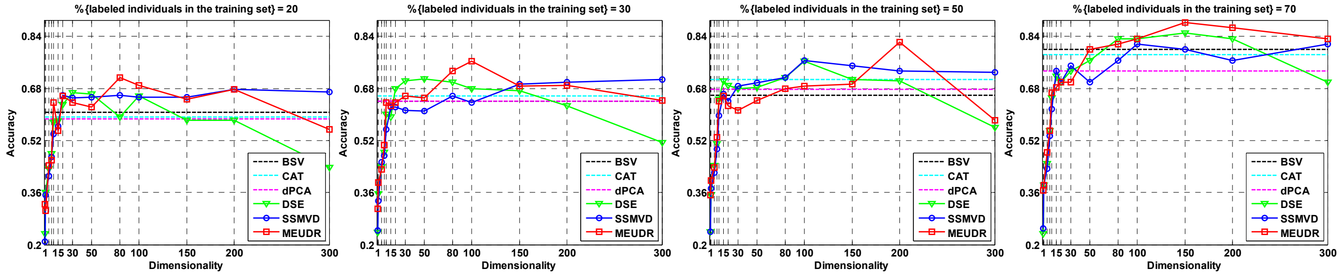

Figure 2: Classification accuracy vs. the dimensionality of the result data on the CASIA dataset.

UMDR approaches increases sharply with an increasing dimension r and then decreases significantly. The performance often peaks around the dimension $r = 8$. This demonstrates the effectiveness of the first few factors found by UMDR; 4) the performance of SSMVD is better than that of DSE because complementary nature is better explored, and the proposed MEUDR outperforms all the other approaches at their best dimensions since we allow arbitrary information to be shared on and between the feature spaces of different views. Although the performance of SSMVD is higher than that of MEUDR when r is large, the curve line is quite unstable and has large oscillations.

It can be seen from the results on the CASIA dataset that: 1) the performance of the simple concatenation strategies (CAT and dPCA) are quite unstable and are worse than BSV in many cases; 2) the improvements of the proposed MEUDR compared to DSE and SSMVD are significant at many dimensions, not only the peak of the curve.

Overall, the recognition accuracy of 93.75% on FE dataset and 88.33% on the CASIA dataset are currently, to the best of our knowledge, state-of-the-art.

4 Conclusion

Multi-view learning has become an active research topic in recent years, but few works have addressed the issue of FER. In this paper, we proposed a new method termed MEUDR for the dimensionality reduction of multi-view data. The proposed MEUDR exploits arbitrary relationships on and between the feature spaces of different views. We allow the uneven contribution of different views to the final consensus representation by learning integration weights for them. This allows the complementary nature of different views to be better exploited, hopefully achieving improved performance in FER. Two challenging video-based FER datasets were adopted to demonstrate the advantages of MEUDR. It is worth noting that the application of MEUDR is suitable for,

Table 1: The average recognition rates and standard deviations (in %) of different approaches at their best dimensions on the FE05 dataset.

	Accuracy			
Methods	0.25	0.50	0.75	1.00
BSV	60.83±2.7	73.75±1.8	75.83±1.8	83.3
CAT	66.25±3.4	71.25±2.7	80.83±0.9	87.5
dPCA	66.25±3.4	77.92±4.1	80.83±0.9	87.5
DSE	70.83±0.0	83.33±0.0	88.33±1.1	93.7
SSMVD	79.58±1.7	84.58±1.1	88.75±1.8	93.7
MEUDR	**80.42±1.8**	**90.42±1.1**	**92.92±1.1**	**93.8**

Table 2: The average recognition rates and standard deviations (in %) of different approaches at their best dimensions on the CASIA dataset.

	Accuracy			
Methods	20	30	50	70
BSV	60.67±3.0	64.00±2.5	66.00±0.9	80.0
CAT	59.33±1.5	65.67±1.9	70.67±2.2	78.3
dPCA	58.67±2.7	64.00±1.9	67.67±2.5	73.3
DSE	69.33±3.5	71.33±0.7	75.00±0.0	83.3
SSMVD	67.67±0.9	70.67±0.9	76.67±0.0	81.6
MEUDR	**71.33±0.7**	**76.33±1.8**	**82.33±2.2**	**88.3**

but not limited to, FER. In any scenario that requires the dimensionality reduction of multi-view data, MEUDR is a suitable candidate.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61374006, Major Program of National Natural Science Foundation of China under Grant 11190015, the Scientific Research Foundation of Graduate School of Southeast University YBJJ1518, the Fundamental Research Funds for the Central Universities, the Program Sponsored for Scientific Innovation Research of College Graduates in Jiangsu Province under Grant KYLX15_0113, and Australian Research Council Projects DP-140102164, FT-130101457, LE140100061.

References

[Ambadar *et al.*, 2005] Zara Ambadar, Jonathan W Schooler, and Jeffrey F Cohn. Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[Cai *et al.*, 2014] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*. AAAI Press, 2014.

[Dollár *et al.*, 2005] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS Workshop*, pages 65–72, 2005.

[Gu *et al.*, 2012] Wenfei Gu, Cheng Xiang, YV Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *PR*, 45(1):80–91, 2012.

[Han *et al.*, 2012] Yahong Han, Fei Wu, Dacheng Tao, Jian Shao, Yueting Zhuang, and Jianmin Jiang. Sparse unsupervised dimensionality reduction for multiple view data. *IEEE TCSVT*, 22(10):1485–1496, 2012.

[Huang *et al.*, 2015] Jin Huang, Feiping Nie, and Heng Huang. A new simplex sparse learning model to measure data similarity for clustering. In *IJCAI*, pages 3569–3575. AAAI Press, 2015.

[Klaser *et al.*, 2008] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 275–1, 2008.

[Kong *et al.*, 2014] Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via $\mathrm{l}_{1,2}$-norm. In *NIPS*, pages 1655–1663, 2014.

[Li *et al.*, 2013] Stan Z Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *CVPR Workshops*, pages 348–353, 2013.

[Liu and Tao, 2015] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE TPAMI*, 2015.

[Long *et al.*, 2008] Bo Long, S Yu Philip, and Zhongfei (Mark) Zhang. A general model for multiple view unsupervised learning. In *SIAM ICDM*, pages 822–833, 2008.

[Luo *et al.*, 2015] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE TKDE*, 27(11):3111–3124, 2015.

[Luo *et al.*, 2016] Yong Luo, Yonggang Wen, Dacheng Tao, Jie Gui, and Chao Xu. Large margin multi-modal multitask feature extraction for image classification. *IEEE TIP*, 25(1):414–427, 2016.

[McFee and Lanckriet, 2011] Brian McFee and Gert Lanckriet. Learning multi-modal similarity. *JMLR*, 12:491–523, 2011.

[Moore and Bowden, 2011] S Moore and R Bowden. Local binary patterns for multi-view facial expression recognition. *CVIU*, 115(4):541–558, 2011.

[Nie *et al.*, 2014] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *ACM SIGKDD*, pages 977–986. ACM, 2014.

[Tao *et al.*, 2006] Dacheng Tao, Xiaoou Tang, Xuelong Li, and Xindong Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE TPAMI*, 28(7):1088–1099, 2006.

[Tao *et al.*, 2007] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE TPAMI*, 29(10):1700–1715, 2007.

[Tao *et al.*, 2009] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. Geometric mean for subspace selection. *IEEE TPAMI*, 31(2):260–274, 2009.

[Wang *et al.*, 2013a] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[Wang *et al.*, 2013b] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, pages 352–360, 2013.

[Xie *et al.*, 2014] Liping Xie, Haikun Wei, Wankou Yang, and Kanjian Zhang. Video-based facial expression recognition using histogram sequence of local gabor binary patterns from three orthogonal planes. In *CCC*, pages 4772–4776, 2014.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE TPAMI*, 37(12):2531–2544, 2015.

[Zhao and Pietikainen, 2007] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6):915–928, 2007.