Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis

Ping Liu[1], Joey Tianyi Zhou[2], Ivor Wai-Hung Tsang[3], Zibo Meng[1], Shizhong Han[1], and Yan Tong[1]

[1]Department of Computer Science, University of South Carolina, USA
[2] Center for Computational Intelligence, Nanyang Technology University, Singapore
[3]Center for Quantum Computation and Intelligent Systems, University of Technology, Australia

Abstract. Studies in psychology show that not all facial regions are of importance in recognizing facial expressions and different facial regions make different contributions in various facial expressions. Motivated by this, a novel framework, named Feature Disentangling Machine (FDM), is proposed to effectively select active features characterizing facial expressions. More importantly, the FDM aims to disentangle these selected features into non-overlapped groups, in particular, *common features* that are shared across different expressions and *expression-specific features* that are discriminative only for a target expression. Specifically, the FDM integrates sparse support vector machine and multi-task learning in a unified framework, where a novel loss function and a set of constraints are formulated to precisely control the sparsity and naturally disentangle active features. Extensive experiments on two well-known facial expression databases have demonstrated that the FDM outperforms the state-of-the-art methods for facial expression analysis. More importantly, the FDM achieves an impressive performance in a cross-database validation, which demonstrates the generalization capability of the selected features.

1 Introduction

Facial activity is one of the most important cues to perceive emotion and intention of a human. Accurate and reliable analysis of facial expressions is imperative to fulfill the demands of emerging applications, such as online/remote education, interactive games, intelligent transportation systems, and many other HCI related applications.

Previous work has shown that not all facial regions but only a few make contributions for expression analysis [3]. More specifically, the most important facial features are extracted from the regions like mouth and eyes [3], since the muscular movements in these regions invoke the expression. These discoveries indicate that features employed in facial expression analysis are sparse and thus, it is important to select the features that are the most effective to characterize facial expressions. To capture the sparsity pattern in features, sparse-coding

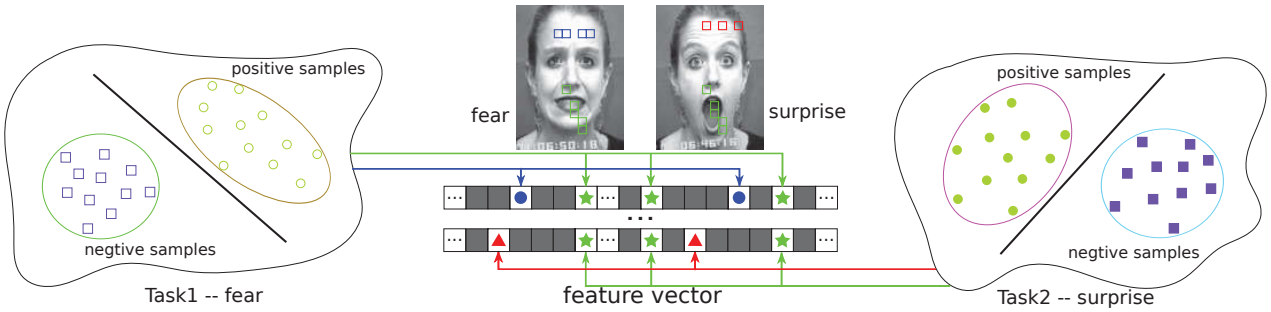

Fig. 1. An FDM performs feature selection and disentangling, simultaneously, using two expressions as an example. Green boxes in the two face images represent the common features shared across the two expressions (fear and surprise) with the corresponding bits (marked by green stars) activated in the feature vectors; blue or red boxes denote expression-specific features for the target expression (fear or surprise) with the corresponding bits (marked by blue circles or red triangles) activated. The bits marked by dark gray in the feature vector correspond to inactivated features for the target expression. Best viewed in color.

based feature learning approaches [33,8,31,14,16,2,39,32] have been employed to extract underlying "edge-like" features from facial images.

Furthermore, the evocation of different expressions may involve the muscular movements from the same facial regions, which could be treated as common features across different expressions. The existence of common features implies the relationships between different tasks, i.e., recognizing different expressions. These relationships cannot be captured in a single-task learning (STL) framework, where each facial expression is recognized individually. In contrast, multi-task learning (MTL) [29] is more suitable to exploit the potential information shared between related tasks to enhance the recognition performance for all target expressions.

Most recently, Zhong et al [40] proposed to divide the sparse features into two groups, i.e., *common features* and *expression-specific features*, through a two-stage multi-task sparse learning (MTSL) framework and achieved promising results in facial expression analysis. Specifically, *common features* that are active for all expressions are extracted by an MTSL model considering all expressions; while *expression-specific features*, learned by a separate MTSL model, are active in recognizing a specific facial expression and are important for face verification as well. Since *common features* and *expression-specific* features are learned sequentially and independently, these two groups can be overlapped.

Intuitively, it is desired to disentangle expression-specific features from the common features. Inspired by the recent work of Sparse Support Vector Machine (SSVM) [24], which employed a novel feature selection vector to precisely control the sparsity of the selected features, we propose a unified MTL framework, named *Feature Disentangling Machine* (FDM), to simultaneously select and disentangle common features and expression-specific features. As illustrated in Fig 1, a set of common features represented by green boxes in the two face images are effective to recognize both *fear* and *surprise* expressions;

while expression-specific features (represented by blue boxes for *fear* and red boxes for *surprise*) are only employed when recognizing the target expression.

Compared with the previous work [40], a novel loss function is proposed in the FDM together with a set of novel constraints to precisely control sparsity and naturally disentangle active features into common and expression-specific groups. Hence, features in any two groups are mutually exclusive. To the best of our knowledge, this is the first work to achieve this. By utilizing an MTL setting, FDM is capable of making fully use of underlying commonality between different facial expressions, which intends to enhance the generalization capability of the selected features. Furthermore, FDM is a general framework and can be applied to various multi-task problems.

Extensive experiments on two well-known facial expression databases have shown that the FDM outperforms the state-of-the-art methods in facial expression analysis. More importantly, in a cross-database experimental validation, the features selected for the Extended Cohn-Kanade (CK+) database [10,17] are also effective in recognizing expressions for the JAFFE database [18].

2 Related Work

As detailed in the surveys [19,35], extensive efforts have been devoted to facial expression analysis. As detailed in the surveys [19,35], extensive efforts have been devoted to facial expression analysis. Generally, facial expression recognition can be performed in three major steps: feature extraction/learning, feature selection, and classifier construction.

First, features are extracted from static images or videos to capture facial changes in appearance or geometry, which are related to a target expression. These features can be human-crafted including Gabor wavelet coefficients [37,36,25,1], Haar features [27,30], Histograms of Oriented Gradients (HOG) [9,4], histograms of Local Binary Patterns (LBP) [38,26,22], or learned in a data-driven manner including sparse-coding based approaches [33,8,31,14,16,2,39,32,15] and deep learning framework [21].

As shown in the psychological studies [3], information extracted around nose, eyes, and mouth is more critical for expression analysis. Moreover, the activation of facial regions varies among different expressions. Consequently, features selected from different facial regions should make different contributions in expression analysis. To achieve this goal, boosting-based feature selection approaches, which aim to automatically adjust the weights of features, have been employed [1,27]. Zafeiriou and Pitas [34] proposed discriminant expression-specific graphs to select expression-specific facial landmarks. However, these approaches have been performed in an STL setting, where each target expression has been treated independently despite the fact that the same set of facial muscles can be contracted when activating different facial expressions [10]. More recently, Zhong et al [40] proposed a two-stage MTSL framework to sequentially locate common and specific facial patches, which are discriminative to all expressions and a specific expression, respectively.

Compared with the previous work in feature selection, the proposed FDM considers the interactions between expression-specific and common features among

different expressions in an MTL framework. More specifically, our work differs from the two-stage MTSL-based method [40] in two aspects. First, feature selection and disentangling are performed jointly in a unified framework, via minimizing a novel loss function. Second, the proposed constraints ensure that there is no overlapping for any two feature groups.

Given the selected features and a training dataset, a pre-specified classifier is employed to construct a facial expression recognizer for a target expression.

3 Methodology

In this section, we first give a brief review on SSVM [24], based on which the proposed FDM is derived. Then, the proposed FDM framework will be presented together with an efficient algorithm for solving the FDM.

3.1 A Brief Review on Sparse Support Vector Machine

Given a set of N labeled images $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in R^m$ is a feature vector[1] extracted from the i^{th} sample and $y_i \in \{\pm 1\}$ represents the expression label, a linear decision hyperplane with the corresponding weight vector $\mathbf{w} = [w_1, ..., w_m]^T \in R^m$ can be estimated in a linear SVM through minimizing an objective function as follows:

$$\min_{\mathbf{w}} \Omega\left(\|\mathbf{w}\|_p\right) + \gamma \sum_{i=1}^{N} loss(-y_i \mathbf{w}^T \mathbf{x}_i) \tag{1}$$

where γ is a positive parameter to balance the complexity of the model and the fitness of the decision hyperplane. $loss(\cdot)$ is a loss function, where various choices of loss, such as quadratic loss and 0-1 loss, can be employed. Among them, hinge loss has been proven to be effective in classification problems, and is adopted in this work. $\Omega(\mathbf{w})$ is a penalty term to control the characteristics of $\mathbf{w}$. For example, l_1 norm [15] and mixed l_{21} norm [40] have been employed to model the sparsity patterns in features.

Recently, Tan et al [24] proposed SSVM, which employed a *feature selection* vector $\mathbf{d} = [d_1, \cdots, d_m]^T \in \mathcal{D}$, where $\mathcal{D} = \left\{\mathbf{d} | \sum_{j=1}^m d_j \leq \tau, d_j \in \{0,1\}\right\}$, to select a subset of features for classification. Through specifying the value of parameter τ, the sparsity pattern in the data can be well controlled so that

$$\mathbf{w}^T \mathbf{x} = (\tilde{\mathbf{w}} \circ \mathbf{d})^T \mathbf{x} = \tilde{\mathbf{w}}^T (\mathbf{d} \circ \mathbf{x}) \tag{2}$$

where $d_j = 1$ when the j^{th} feature is selected and otherwise $d_j = 0$; "$\circ$" denotes the element-wise multiplication.

[1] In this work, histograms of LBP features have been employed to represent images, while other features such as HOG and Gabor wavelet features can be employed as well. Implementation details of LBP features can be found in Section 4.

Hence, the objective function (Eq. 1) of SSVM [24] is formulated as:

$$\begin{aligned} &\min_{\mathbf{d}\in\mathcal{D}} \min_{\tilde{\mathbf{w}},\boldsymbol{\epsilon},\rho} \frac{1}{2}\|\tilde{\mathbf{w}}\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^{N}\epsilon_i^2 - \rho \\ s.t. \quad & y_i\tilde{\mathbf{w}}^T(\mathbf{x}_i \circ \mathbf{d}) \geq \rho - \epsilon_i, i = 1, \cdots, N. \end{aligned} \tag{3}$$

where $\boldsymbol{\epsilon}$ and ρ are parameters used to generate a soft margin for non-separable classification problems. Then, the optimal $\tilde{\mathbf{w}}$ can be used to construct the classifier with a subset of selected features specified by $\mathbf{d}$.

The SSVM proposed in [24] has been proven to be efficient and effective in various classification problems. However, it is in an STL setting and has not considered the interconnections between related classification problems, while facial expression recognition has been shown benefiting from MTL by exploiting information shared between different expressions [40]. By taking advantage of the underlying shared commonality, we propose an FDM to disentangle expression-specific and common features by extending the SSVM to an MTL framework. By introducing a novel joint objective function and novel constraints, the FDM aims to capture and utilize the interconnections among multiple expressions via the common features.

3.2 Formulation for the FDM

To simplify the discussion, we only discuss an MTL expression recognition problem considering two target expressions, denoted as E_1 and E_2, at the same time. Then, each image sample can be represented by a triplet $\left\{\mathbf{x}_i, y_i^{E_1}, y_i^{E_2}\right\}, i = 1, \cdots, N$, with two expression labels ($y_i^{E_1}$ and $y_i^{E_2}$). Specifically, if only one of the target expressions, e.g., E_1, is activated in the image, $y_i^{E_1} = 1$ and $y_i^{E_2} = -1$, and vice versa; while if neither E_1 nor E_2 is activated, both the expression labels are set to -1.

In order to select *expression-specific features*, two *expression-specific feature selection* vectors denoted as $\mathbf{d}^{E_1}$ and $\mathbf{d}^{E_2}$ are introduced for tasks E_1 and E_2, respectively. In addition, a *common feature selection* vector denoted as $\mathbf{d}^{E_c}$ is used to select common features that are effective and shared in recognizing all expressions.

Therefore, the objective function in Eq. 3 can be extended to recognizing both expressions simultaneously in an MTL framework as follows:

$$\begin{aligned} &\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\in\mathcal{D}\}} \min_{\{\mathbf{w}^{E_1},\mathbf{w}^{E_2},\boldsymbol{\epsilon}^{E_1},\boldsymbol{\epsilon}^{E_2},\rho_1,\rho_2\}} \frac{1}{2}\left(\|\mathbf{w}^{E_1}\|_2^2 + \|\mathbf{w}^{E_2}\|_2^2\right) + \frac{\gamma}{2}\sum_{i=1}^{N}\left[(\epsilon_i^{E_1})^2 + (\epsilon_i^{E_2})^2\right] - (\rho_1+\rho_2) \\ s.t. \quad & y_i^{E_1}(\mathbf{w}^{E_1})^T\left[\mathbf{x}_i \circ \left(\mathbf{d}^{E_1} + \mathbf{d}^{E_c}\right)\right] \geq \rho_1 - \epsilon_i^{E_1}, i = 1, \cdots, N, \\ & y_i^{E_2}(\mathbf{w}^{E_2})^T\left[\mathbf{x}_i \circ \left(\mathbf{d}^{E_2} + \mathbf{d}^{E_c}\right)\right] \geq \rho_2 - \epsilon_i^{E_2}, i = 1, \cdots, N. \end{aligned} \tag{4}$$

where the sparsity of features is controlled in the three feature selection vectors (i.e., $\mathbf{d}^{E_1}$, $\mathbf{d}^{E_2}$, and $\mathbf{d}^{E_c}$) by three parameters τ_1, τ_2, and τ_c, respectively as follows

$$\sum_{j=1}^{m} d_j^{E_1} \leq \tau_1 \qquad \sum_{j=1}^{m} d_j^{E_2} \leq \tau_2 \qquad \sum_{j=1}^{m} d_j^{E_c} \leq \tau_c \qquad d_j^{E_1}, d_j^{E_2}, d_j^{E_c} \in \{0,1\} \quad (5)$$

Furthermore, to ensure that there is no intersection between any two sets of features, i.e., a feature can be selected by at most one subset, we propose a novel constraint formulated as:

$$d_j^{E_1} + d_j^{E_2} + d_j^{E_c} \leq 1 \qquad j = 1, \cdots, m. \quad (6)$$

Therefore, by minimizing the proposed joint objective function (Eq. 4) with novel constraints (Eq. 5 and 6), the FDM framework is capable of simultaneously finding the optimal hyperplanes (represented by $\mathbf{w}^{E_1}$ and $\mathbf{w}^{E_2}$), expression-specific features, and common features, for classifying the two expressions. In the discussion below, we will present an efficient algorithm to solve the FDM.

3.3 Algorithm for Solving Feature Disentangling Machine

To solve the FDM, the Lagrange multiplier method with KKT condition is employed to transform the original inner problem (Eq. 4) into its dual formulation, and then the solution to the original problem can be found by solving the corresponding dual problem as:

$$\begin{aligned}
&\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \\
&\min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} -\frac{1}{2}\left\|\sum_i \alpha_i y_i^{E_1}\left[\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\alpha}^T\boldsymbol{\alpha} \\
&\qquad\qquad - \frac{1}{2}\left\|\sum_i \beta_i y_i^{E_2}\left[\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\beta}^T\boldsymbol{\beta} \\
&s.t. \quad \sum_{i=1}^{N}\alpha_i = 1, \quad \sum_{i=1}^{N}\beta_i = 1, \quad \alpha_i > 0, \quad \beta_i > 0, \qquad for\ i = 1,\cdots,N, \\
&\qquad \{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\} \in \mathcal{D}, \\
&where \qquad \mathcal{D} = \{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\} | \sum_{j=1}^{m} d_j^{E_1} \leq \tau_1, \sum_{j=1}^{m} d_j^{E_2} \leq \tau_2, \sum_{j=1}^{m} d_j^{E_c} \leq \tau_c, \\
&\qquad d_j^{E_1} + d_j^{E_2} + d_j^{E_c} \leq 1, \qquad d_j^{E_1}, d_j^{E_2}, d_j^{E_c} \in \{0,1\}, \qquad for\ j = 1,\cdots,m\}
\end{aligned} \quad (7)$$

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are dual variable vectors for the inequality constraints in the inner minimization problem (Eq. 4).

The saddle point problem (7) can be lower bounded by:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} L_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \\
&\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}} -\frac{1}{2}\left\|\sum_i \alpha_i y_i^{E_1}\left[\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\alpha}^T\boldsymbol{\alpha} \\
&\qquad\qquad -\frac{1}{2}\left\|\sum_i \beta_i y_i^{E_2}\left[\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\circ\mathbf{x}_i\right]\right\|^2 - \frac{1}{2\gamma}\boldsymbol{\beta}^T\boldsymbol{\beta} \qquad (8)\\
&s.t.\ \sum_{i=1}^{n}\alpha_i = 1,\ \sum_{i=1}^{n}\beta_i = 1,\ \alpha_i > 0,\ \beta_i > 0,\ for\ i = 1,...,N,\ \{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D}
\end{aligned}$$

By bringing an additional variable θ, the above optimization problem becomes:

$$\max_{\theta,\boldsymbol{\alpha},\boldsymbol{\beta}} -\theta : \theta \geq -L_{\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}}(\boldsymbol{\alpha},\boldsymbol{\beta}), \qquad \forall\{\mathbf{d}_t^{E_1},\mathbf{d}_t^{E_2},\mathbf{d}_t^{E_c}\}\in\mathcal{D} \tag{9}$$

which is a convex Quadratically Constrained Quadratic Programming (QCQP) problem.

Define $\mu_t \geq 0$ as the dual variable for each constraint in Eq. 9 [24], the Lagrangian of Eq. 9 can be rewritten as:

$$\begin{aligned}
\min_{\mu\in\mathcal{M}}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} &\frac{-1}{2}(\boldsymbol{\alpha}\circ\mathbf{y}^{E_1})^T\Big(\sum_t \mu_t X_t^{E_1}{X_t^{E_1}}^T + \frac{1}{\gamma}\mathcal{I}\Big)(\boldsymbol{\alpha}\circ\mathbf{y}^{E_1}) \\
&\frac{-1}{2}(\boldsymbol{\beta}\circ\mathbf{y}^{E_2})^T\Big(\sum_t \mu_t X_t^{E_2}{X_t^{E_2}}^T + \frac{1}{\gamma}\mathcal{I}\Big)(\boldsymbol{\beta}\circ\mathbf{y}^{E_2}) \\
where X_t^{E_1} &= \left[\mathbf{x}_1\circ(\mathbf{d}_t^{E_1}+\mathbf{d}_t^{E_c}),\cdots,\mathbf{x}_N\circ(\mathbf{d}_t^{E_1}+\mathbf{d}_t^{E_c})\right]^T \qquad (10)\\
X_t^{E_2} &= \left[\mathbf{x}_1\circ(\mathbf{d}_t^{E_2}+\mathbf{d}_t^{E_c}),...,\mathbf{x}_N\circ(\mathbf{d}_t^{E_2}+\mathbf{d}_t^{E_c})\right]^T \\
\mathcal{M} &= \{\mu|\sum\mu_t = 1,\mu_t\geq 0\}
\end{aligned}$$

where $\mathcal{I}$ represents an identity matrix.

Eq. 10 is a Multiple Kernel Learning (MKL) problem, in which the kernel matrix $\sum_t \mu_t X_t^{E_1}{X_t^{E_1}}^T$ and $\sum_t \mu_t X_t^{E_2}{X_t^{E_2}}^T$ are both convex combinations of $|\mathcal{D}|$ base kernel matrices $X_t^{E_1}{X_t^{E_1}}^T$ and $X_t^{E_2}{X_t^{E_2}}^T$. However, not all constraints in Eq. 9 are active at optimality. Therefore, the problem can be solved efficiently and effectively by cutting plane algorithm [11]. The overall algorithm of solving FDM is described here and summarized in Algorithm 1.

Denote the subset of constraints by $\mathcal{C}\in\mathcal{D}$. First, the dual variables α_i and β_i are set to $\frac{1}{N}$ for $i = 1,\cdots,N$ for initialization. A combination of feature selection vectors, denoted as $\hat{\mathcal{D}} = \{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D}$, which violates the constraints (Eq. 5

and 6) the most, are obtained. Then, two steps, i.e., estimating the new $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with MKL and finding the most violated feature selection vectors ($\hat{\mathcal{D}}$), run alternatively until converge. By introducing constraints for the *expression-specific* feature selection vectors ($\mathbf{d}^{E_1}$ and $\mathbf{d}^{E_2}$) and the *common* feature selection vector $\mathbf{d}^{E_c}$, it is ensured that the features in any two subsets are mutually exclusive. This is known as "Feature Disentangling".

MKL with a subset of kernel matrices Inspired by previous work on SSVM [24], we apply SimpleMKL [20] to solve the MKL problem defined on the subset of kernel matrices selected in $\mathcal{C}$.

In this step, since the feature selection vectors ($\mathbf{d}^{E_1}$, $\mathbf{d}^{E_2}$ and $\mathbf{d}^{E_c}$) are fixed, we can solve the MKL problem corresponding to the following primal optimization problem:

$$\begin{aligned}
\min_{\mu\in\mathcal{M},\mathbf{w}^{E_1},\mathbf{w}^{E_2},\rho_1,\rho_2,\boldsymbol{\epsilon}^{E_1},\boldsymbol{\epsilon}^{E_2}} \; & \frac{1}{2}\sum_{t=1}^{K}\frac{1}{\mu_t}\|\mathbf{w}^{E_1}\|^2+\frac{\gamma}{2}\sum_{i=1}^{N}\left(\epsilon_i^{E_1}\right)^2-\rho_1 \\
& +\frac{1}{2}\sum_{t=1}^{K}\frac{1}{\mu_t}\|\mathbf{w}^{E_2}\|^2+\frac{\gamma}{2}\sum_{i=1}^{N}\left(\epsilon_i^{E_2}\right)^2-\rho_2 \\
s.t. \sum_{t=1}^{K}\left(\mathbf{w}^{E_1}\right)^T & \left[y_i^{E_1}\mathbf{x}_i\circ\left(\mathbf{d}^{E_1}+\mathbf{d}^{E_c}\right)\right]\geq\rho_1-\epsilon_i^{E_1} \qquad \forall i=1,\cdots,N \\
\sum_{t=1}^{K}\left(\mathbf{w}^{E_2}\right)^T & \left[y_i^{E_2}\mathbf{x}_i\circ\left(\mathbf{d}^{E_2}+\mathbf{d}^{E_c}\right)\right]\geq\rho_2-\epsilon_i^{E_2} \qquad \forall i=1,\cdots,N
\end{aligned} \tag{11}$$

Since $\mathbf{d}^{E_1}$, $\mathbf{d}^{E_2}$and $\mathbf{d}^{E_c}$are fixed here, Eq. 11 actually becomes a combination of two SSVMs ($SSVM^{E_1}$ and $SSVM^{E_2}$). The problem can be solved by optimizing the parameters of $SSVM^{E_1}$ and $SSVM^{E_2}$ in an iterative way. For solving each sub problem $SSVM^{E_1}$ or $SSVM^{E_2}$, we employ SimpleMKL [20], following [24].

Finding the most violated feature selection vectors by a Knapsack problem solver To find $\hat{\mathcal{D}}=\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D}$ in Eq. 9, we propose to solve the equivalent optimization problem:

$$\begin{aligned}
\max_{\{\mathbf{d}^{E_1},\mathbf{d}^{E_2},\mathbf{d}^{E_c}\}\in\mathcal{D}} \; & \frac{1}{2}\sum_{j=1}^{m}(c_j^{E_1})^2(d_j^{E_1}+d_j^{E_c})+\frac{1}{2}\sum_{j=1}^{m}(c_j^{E_2})^2(d_j^{E_2}+d_j^{E_c}) \\
where \quad & c_j^{E_1}=\sum_{i=1}^{N}\alpha_i y_i x_{ij} \qquad c_j^{E_2}=\sum_{i=1}^{N}\beta_i y_i x_{ij} \\
s.t. \; & \sum_{j=1}^{m}d_j^{E_1}\leq\tau_1;\ \sum_{j=1}^{m}d_j^{E_2}\leq\tau_2;\ \sum_{j=1}^{m}d_j^{E_c}\leq\tau_c; \\
& d_j^{E_1}+d_j^{E_2}+d_j^{E_c}\leq 1;\ d_j^{E_1},d_j^{E_2},d_j^{E_c}\in\{0,1\};\ for\ j=1,\cdots,m
\end{aligned} \tag{12}$$

Based on Eq. 12, the problem becomes a binary and linear programming problem, more specifically, the Knapsack Problem [28]. Various methods such as dynamic programming and greedy algorithm have been proposed to solve this problem efficiently and effectively. In this work, we adopt the optimization toolbox [7] provided by MATLAB to solve the problem.

Initialize α_i, β_i as $\frac{1}{N}$ for $i = 1, \cdots, N$; Find the most violated feature selection vectors $\hat{\mathcal{D}} = \{\mathbf{d}^{E_1}, \mathbf{d}^{E_2}, \mathbf{d}^{E_c}\}$ and let $\mathcal{C} = \{\hat{\mathcal{D}}\}$;
repeat
  Initialize $\mu = [1]^T$;
  repeat
    Find the optimal μ, α and β by simpleMKL
  until *convergence*;
  Find the most violated feature selection vectors $\hat{\mathcal{D}} = \{\mathbf{d}^{E_1}, \mathbf{d}^{E_2}, \mathbf{d}^{E_c}\}$
  make $\mathcal{C} = \mathcal{C} \cup \{\hat{\mathcal{D}}\}$
until *convergence*;

Algorithm 1: Algorithm of Feature Disentangling Machine

Once the optimal solution of feature selection vectors are estimated, we employ expression-specific features specified by $\mathbf{d}^{E_1}$ or $\mathbf{d}^{E_2}$, together with common features specified by $\mathbf{d}^{E_c}$ to train a classifier to recognize the target expression. In this work, the LIBLinear software [6] is adopted for constructing classifiers.

3.4 Computational Complexity

As shown in Algorithm 1, in each iteration, two major steps run alternatively: solving two sub problems of MKL and searching for the most violated feature selection vectors $\hat{\mathcal{D}}$. For the first step, linear base kernels are employed; and the LIBLinear [6], which scales linearly in the number of samples N and the feature dimensions m, is adopted in solving the two sub problems. Hence, the time complexity of the first step is $O(mN)$. For the second step, we employed the MATLAB function *bintprog* in our experiments, which uses a linear programming-based branch-and-bound algorithm with polynomial complexity [7]. Other methods like dynamic programming could be used to achieve a linear complexity of $O(m\tau)$, where τ is the maximum capacity of the Knapsack problem. For a multiclass problem with K classes, the overall computational complexity can be $O(K^2 m(N + \tau))$ using an one-versus-all strategy.

4 Experimental Results

In order to evaluate the proposed FDM framework, extensive experiments have been performed on two well-known facial expression databases: Extended Cohn-Kanade (CK+) database [10,17] and JAFFE database [18].

Preprocessing Images and Feature Extraction For preprocessing purpose, the face regions across different facial images were aligned to remove the scale and positional variance [2] and then cropped to 96×64. Each cropped face image

[2] In this work, the face region was roughly aligned based on eye positions detected by an eye detector.

was further divided into 7×7 non-overlapped patches. In this work, a uniform $\text{LBP}_{8,1}$ pattern was employed to compute LBP features at each pixel location. Then, a histogram with 59 bins were calculated for each image patch. Hence, each image was represented by a feature vector with $2891(7 \times 7 \times 59)$ features. This preprocessing strategy was adopted for both databases we employed.

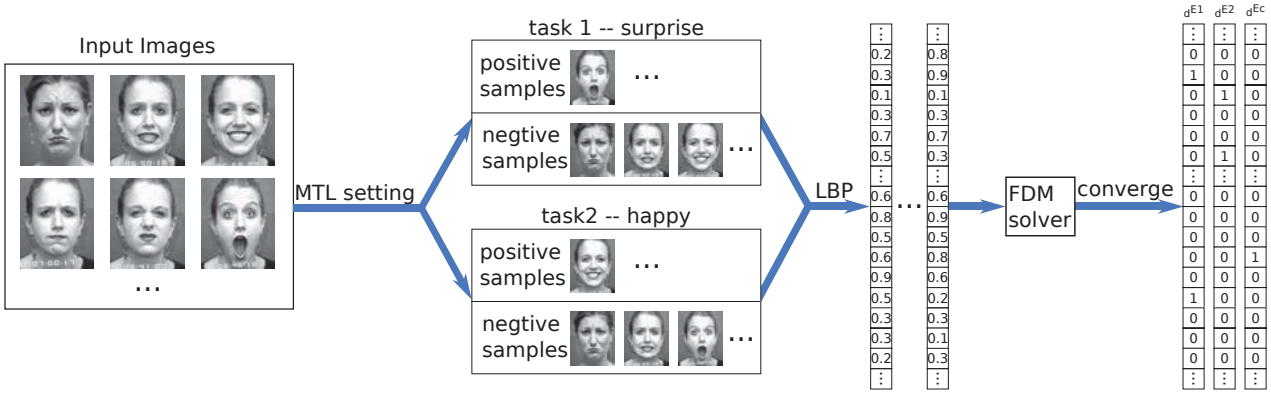

Fig. 2. An illustration of the FDM training procedure.

Multi-task Learning Configuration and Experimental Setup As shown in Fig. 2, an FDM is trained to select features for a pair of expressions simultaneously. For the FDM, the parameters τ_1 and τ_2 are set to 500 and the parameter τ_c is set to 250 empirically in all experiments. For recognizing P expressions, there are a total of $\binom{P}{2}$ FDMs needed. [3] As a result, $P-1$ sets of features can be selected and extracted for each expression, from each of which a binary classifier can be trained. In this work, the L_2-loss L_2-regularized SVM implemented in the LIBLinear [6] was adopted for constructing the binary classifiers. Given a testing image, the expression label was estimated using an average rule such that the final recognition score is an average of the $P-1$ classification scores.

4.1 Experiments on the CK+ Database

The CK+ database contains 327 expression-labeled image sequences, each of which has one of 7 expressions, i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise activated. For each image sequence, only the last frame (the peak frame) is provided with an expression label. To collect more image samples from the database, we selected the last three frames from each image sequence. In addition, we also collected the first frame from each of the 327 labeled sequences for "neutral" expression. Through this way, an experimental data set named *CK-DB* with a total of 1308 images is built. Then, we employed an 8-fold cross-validation strategy. The CK-DB was divided into 8 subsets, where the subjects in any two of subsets were mutually exclusive. For each run, 7 subsets were employed for training and the remaining one for testing. We performed such 8 runs by enumerating the subset used for testing; and the recognition performance was computed as the average of the 8 runs.

[3] $E_1 - E_2$ and $E_2 - E_1$ are treated as the same combination.

Performance Evaluation on the CK-DB We first compared the proposed *FDM* framework with five baseline methods. The first method, denoted as *LOG*, employed an L_2 regularized logistic classifier. The second method, denoted as L_2L_2, employed an L_2-loss SVM with L_2 regulation. Both *LOG* and L_2L_2 have no feature selection ability. The third method, denoted as L_2L_1, employed an L_2-loss SVM with L_1 regulation. The fourth method employed the *SSVM* [24]. The fifth method, denoted as FDM_{wocf}, only employed the expression-specific features selected by FDM for recognition. All the baseline methods, except the *LOG*, employed the hinge loss. The regulation term is $\|\mathbf{w}\|_{l_1}$ for L_1 regularization and $\|\mathbf{w}\|_{l_2}$ for L_2 regularization.

Quantitative experimental results were reported in terms of *average classification rate*, *hit rate*, *false positive rate*, *F1 score*, and *Area Under Curve (AUC) score*. As shown in Fig. 3, the proposed *FDM* outperformed all baseline methods drastically in terms of the average classification rate (**0.977**), the average hit rate (**0.978**), the average false positive rate (**0.023**), the average F1 score (**0.908**) and the average AUC score (**0.989**) of the 6 basic expressions, i.e., anger, disgust, fear, happiness, sadness, and surprise [4].

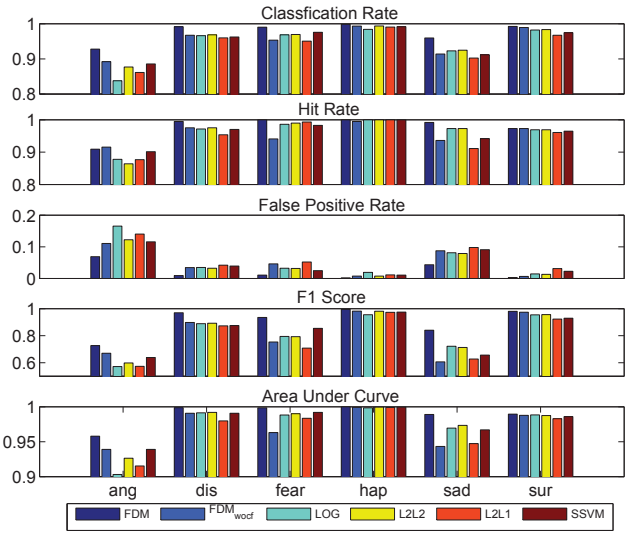

Fig. 3. From top to bottom, performance comparison on the CK-DB in terms of average classification rate, hit rate, false positive rate, F1 score and AUC score for 6 basic expressions. Best viewed in color.

From Fig. 3, we can find that the FDM yielded a significant improvement in F1 score compared to the methods without feature selection, i.e., *LOG* (0.818) and L_2L_2 (0.822), which demonstrated the effectiveness of feature selection and disentangling. Not surprisingly, the FDM outperformed the methods with feature selection in an STL setting, i.e., L_2L_1 (0.779) and *SSVM* (0.821) thanks to the multi-task learning; it also outperformed the one without common features, i.e., FDM_{wocf} (0.814), which demonstrates the importance of the common features in expression recognition.

Furthermore, we compared the proposed FDM method with the state-of-the-art methods evaluated on CK+ or the original Cohn-Kanade database [10] [5] including methods without feature selection denoted as PGKNMF [32], selecting features by AdaBoost in an STL setting (AdaGabor [1] and LBPSVM [23]), and selecting and disentangling features sequentially based on MSTL denoted as CSPL [40]. The experimental results reported in their papers were used directly

[4] We did not recognize the "contempt" and "neutral" for a fair comparison with the state-of-the-art methods evaluated on the original Cohn-Kanade database [10].

[5] Cohn-Kanade database [10] is an early version of CK+ and contains a subset of CK+ data (i.e., 320 image sequences with expression labels [23]).

Table 1. Performance comparison on the CK+ database in terms of average classification rate for 6 expressions.

CSPL [40]	AdaGabor [1]	LBPSVM [23]	PGKNMF [32]	**FDM**
0.899	0.933	0.951	0.835	**0.977**

for comparison. As shown in Table 1, our proposed FDM framework outperformed all the state-of-the-art methods in comparison in terms of the average classification rate (**0.977**). It is worth to mention that the FDM performed better than the MSTL-based method (CSPL) because of the jointly selecting and disentangling the common and expression-specific features.

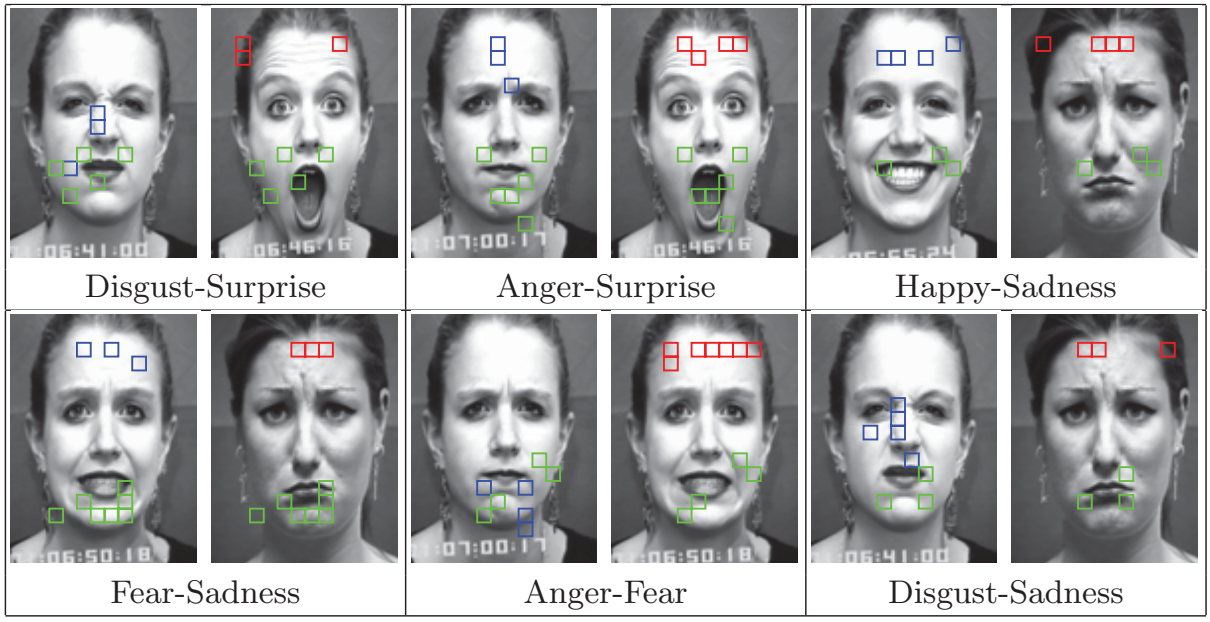

Fig. 4. An illustration of the selected image patches for recognizing the six basic expressions in CK+ database. Green boxes represent common features selected for the pair of expressions; blue or red boxes represent expression-specific features for the target expressions in the pair, respectively. For example, for the pair of *Anger-Surprise*, the features in the green boxes are selected to recognize both anger and surprise; the features in the blue boxes are only sensitive to anger, while the features in the red boxes are only sensitive to surprise. Best viewed in color.

Analysis on Patch Selection Results in the CK+ Database To analyze what information each selected patch provides for expression recognition, a data analysis on the patch selection results was performed. As shown in Fig. 4, patches selected through FDM were marked by boxes. Only a few patches are selected for each target expression, which demonstrates the sparsity in active features. Specifically, patches enclosed in green boxes were selected as *common features* for both expressions in the FDM, while the patches enclosed in red or blue boxes were selected as expression-specific features for the corresponding expression, respectively. These selected patches contain the most discriminative information to characterize the corresponding expression.

From Fig. 4, we can find that most of the selected common patches are located around lip, which coincides with the psychological studies [3]. Furthermore, the expression-specific patches for the target expression are closely related to the facial Action Units (AUs) [5] that describe the corresponding expression. For example, the expression-specific patches selected for *anger* are either located around the lip (in the middle of the second row in Fig. 4), which are related to AU23 (Lip tighten), or around the eyebrows (in the middle of the first row in Fig. 4), which are related to AU4 (Brow Lowerer). AU23 and AU4 are the primary AUs to describe the anger expression [17]. Similar results can be found in other expressions.

4.2 Experiments on the JAFFE Database

The proposed FDM framework has been also evaluated on the JAFFE database, which consists of 213 images from 10 Japanese female subjects. For each subject, there are 3 or 4 examples of each of the six basic expressions and "neutral" expression. The experimental results on the JAFFE database are used to demonstrate the cross-database generalization ability of the FDM.

Cross-database Validation To evaluate the generalization ability, we performed a cross-database validation, where the features were selected and/or the classifiers were trained on the CK+ database; while the performance was tested on the JAFFE database. Particularly, we employed two different experimental settings, namely FDM^{CJ} and FDM^{CC}. In FDM^{CJ}, the features were selected by an FDM trained from CK-DB, while the classifiers were trained on JAFFE database using a leave-one-subject-out training/testing strategy. In FDM^{CC}, we employed the selected features and the trained classifiers from CK-DB to perform test directly on the JAFFE database [6].

The generalization across database is usually low in previous work. Shan et al [23] trained selected LBP features using SVMs on Cohn-Kanade database and tested the trained system on the JAFFE database. An average classification rate about 41% for 7 expressions (6 basis expressions and neutral) was obtained in their work [23]. From Table 2, we can find that both FDM^{CJ} and FDM^{CC} performed much better than [23]. Furthermore, even the training was performed on the CK-DB exclusively in FDM^{CC}, the proposed method could achieve satisfactory recognition performance on the JAFFE database. This further demonstrated that the features selected and disentangled by the FDM captured the most discriminative information for expression analysis, which can be generalized across different data sets.

Performance Evaluation on the JAFFE Database In addition, we also evaluated the FDM trained and tested on the JAFFE database with a leave-one-subject-out training/testing strategy. To make a fair comparison, we only compared with the-state-of-the-art methods employing the leave-one-subject-out

[6] In order to recognize the neutral expression, feature selection and classifier training were performed on the CK-DB for FDM^{CC} and FDM^{CJ}.

Table 2. Cross-database validation, trained on the CK+ database and tested on the JAFFE database, in terms of average classification rate for 7 expressions (6 basis expressions and neutral). In [23], LBP features were employed and fed into SVM with three different kernels, i.e., linear, polynomial, and RBF, respectively. In FDM^{CJ}, the features were selected from CK+, but the classifiers were trained from the JAFFE; while FDM^{CC} employed the selected features and the trained classifiers using CK+.

Ada+SVM(Linear) [23]	Ada+SVM(Poly) [23]	Ada+SVM(RBF) [23]	FDM^{CJ}	FDM^{CC}
0.404	0.404	0.413	**0.901**	**0.882**

Table 3. Performance comparison on the JAFFE database in terms of average classification rate for 7 expressions (6 basis expressions and neutral).

SLLE [13]	SFRCS [12]	Ada+SVM(RBF) [23]	FDM^{JJ}
0.868	0.860	0.810	**0.897**

strategy and recognizing 7 expressions (six basis expressions plus "neutral"). As shown in Table 3, the FDM^{JJ} outperformed the other methods in comparison.

Note that the performance of the FDM^{CC} shown in Table 2 is similar to that of the FDM^{JJ} trained on the JAFFE Database; and the FDM^{CJ} achieved the best performance among all the methods reported in both Table 2 and 3. This implies that the FDM can be adopted in a transfer learning framework. It would be especially useful when the image dataset that was used to train the original classifiers cannot be accessed. Better recognition performance can be achieved by employing the features selected by the original classifiers together with a few labeled images in the new application.

5 Conclusion and Future Work

In this work, we propose a novel FDM framework to perform feature selection and disentangling simultaneously and jointly through a multi-task learning framework. Specifically, two types of feature selection vectors are proposed for common features and expression-specific features, respectively. Furthermore, a novel loss function and a set of constraints are formulated to precisely control the sparsity and ensure non-intersection between different feature groups. In this way, the most discriminative features for recognizing the target expression can be selected and categorized into non-overlapped groups. As demonstrated in our experiments, the proposed FDM outperformed all methods in comparison including the state-of-the-art techniques evaluated on two public facial expression databases. More importantly, the FDM yields impressive performance in the cross-database validation. In the future, we will evaluate the FDM on spontaneous facial displays. Furthermore, the FDM will be adopted to recognize facial action units (AUs), where common features learned by the FDM are expected to capture correlations among AUs.

6 Acknowledgments

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

References

1. Bartlett, M.S., Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I., Movellan, J.R.: Recognizing facial expression: Machine learning and application to spontaneous behavior. In: CVPR. vol. 2, pp. 568–573 (2005)
2. Bociu, I., Pitas, I.: A new sparse image representation algorithm applied to facial expression recognition. In: MLSP. pp. 539–548 (2004)
3. Cohn, J.F., Zlochower, A.: A computerized analysis of facial expression: Feasibility of automated discrimination. American Psychological Society (1995)
4. Dahmane, M., Meunier, J.: Emotion recognition using dynamic grid-based HoG features. In: FG (March 2011)
5. Ekman, P., Friesen, W.V., Hager, J.C.: Facial Action Coding System: the Manual. Research Nexus, Div., Network Information Research Corp., Salt Lake City, UT (2002)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Machine Learning Research 9, 1871–1874 (2008)
7. Grace, A., Works, M.: Optimization Toolbox: For Use with MATLAB: User's Guide. Math works (2013)
8. H, M.M., Mu, Z., L, V.K., Mohammad, M.S., F, C.J.: Facial action unit recognition with sparse representation. In: FG. pp. 336–342. IEEE (2011)
9. Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X., Huang, T.S.: Multi-view facial expression recognition. In: FG. pp. 1–6 (2008)
10. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: FG. pp. 46–53 (2000)
11. Kelley, Jr, J.E.: The cutting-plane method for solving convex programs. Journal of the Society for Industrial & Applied Mathematics 8(4), 703–712 (1960)
12. Kyperountas, M., Tefas, A., Pitas, I.: Salient feature and reliable classifier selection for facial expression classification. Pattern Recognition 43(3), 972–986 (2010)
13. Liang, D., Yang, J., Zheng, Z., Chang, Y.: A facial expression recognition system based on supervised locally linear embedding. Pattern Recognition Letters 26(15), 2374–2389 (2005)
14. Lin, Y., Song, M., Quynh, D., He, Y., Chen, C.: Sparse coding for flexible, robust 3d facial-expression synthesis. Computer Graphics and Applications 32(2), 76–88 (2012)
15. Liu, P., Han, S., Tong, Y.: Improving facial expression analysis using histograms of log-transformed nonnegative sparse representation with a spatial pyramid structure. In: FG. pp. 1–7. IEEE (2013)
16. Liu, W., Song, C., Wang, Y.: Facial expression recognition based on discriminative dictionary learning. In: ICPR (2012)
17. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression. In: CVPR Workshops. pp. 94–101 (2010)
18. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE T-PAMI 21(12), 1357–1362 (1999)
19. Pantic, M., Pentland, A., Nijholt, A., Huang, T.S.: Human computing and machine understanding of human behavior: A survey. In: Huang, T.S., Nijholt, A., Pantic, M., Pentland, A. (eds.) Artificial Intelligence for Human Computing. LNAI (2007)
20. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: Simplemkl. J. Machine Learning Research 9(11) (2008)

21. Ranzato, M., Susskind, J., Mnih, V., Hinton, G.: On deep generative models with applications to recognition. In: CVPR. pp. 2857–2864. IEEE (2011)
22. Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., Prevost, L.: Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In: FG Workshops. pp. 860 – 865 (2011)
23. Shan, C., Gong, S., McOwan, P.: Facial expression recognition based on Local Binary Patterns: A comprehensive study. J. IVC 27(6), 803–816 (2009)
24. Tan, M., Wang, L., Tsang, I.W.: Learning sparse svm for feature selection on very high dimensional datasets. pp. 1047–1054 (2010)
25. Tian, Y., Kanade, T., Cohn, J.F.: Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In: FG. pp. 229–234 (May 2002)
26. Valstar, M.F., Mehu, M., Jiang, B., Pantic, M., Scherer, K.: Meta-analyis of the first facial expression recognition challenge. IEEE T-SMC-B 42(4), 966–979 (2012)
27. Whitehill, J., Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Towards practical smile detection. IEEE T-PAMI 31(11), 2106–2111 (Nov 2009)
28. Wolsey, L.A.: Integer programming. IIE Transactions 32(273-285), 2–58 (2000)
29. Xue, Y., Liao, X., Carin, L., Krishnapuram, B.: Multi-task learning for classification with dirichlet process priors. J. Machine Learning Research 8, 35–63 (2007)
30. Yang, P., Liu, Q., Metaxas, D.N.: Boosting coded dynamic features for facial action units and facial expression recognition. In: CVPR. pp. 1–6 (June 2007)
31. Ying, Z.L., Wang, Z.W., Huang, M.W.: Facial expression recognition based on fusion of sparse representation. In: Huang, D.S., Zhang, X., Reyes García, C., Zhang, L. (eds.) Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, pp. 457–464. LNCS (2010)
32. Zafeiriou, S., Petrou, M.: Nonlinear non-negative component analysis algorithms. IEEE T-IP 19(4), 1050–1066 (2010)
33. Zafeiriou, S., Petrou, M.: Sparse representations for facial expressions recognition via L1 optimization. In: CVPR Workshops. pp. 32–39 (2010)
34. Zafeiriou, S., Pitas, I.: Discriminant graph structures for facial expression recognition. IEEE T-Multimedia 10(8), 1528–1540 (2008)
35. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE T-PAMI 31(1), 39–58 (Jan 2009)
36. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE T-PAMI 27(5), 699–714 (May 2005)
37. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: FG. pp. 454–459 (1998)
38. Zhao, G., Pietiäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE T-PAMI 29(6), 915–928 (June 2007)
39. Zhi, R., Flierl, M., Ruan, Q., Kleijn, W.: Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE T-SMC-B (99), 1–15 (2010)
40. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.: Learning active facial patches for expression analysis. In: CVPR (2012)