Robust representation and recognition of facial emotions using extreme sparse learning

Li, Jun; Teoh, Eam Khwang; Nandakumar, Karthik; Shojaeilangari, Seyedehsamaneh; Yau, Wei-Yun

2015

Shojaeilangari, S., Yau, W.-Y., Nandakumar, K., Li, J., & Teoh, E. K. (2015). Robust representation and recognition of facial emotions using extreme sparse learning. IEEE transactions on image processing, 24(7), 2140-2152.

https://hdl.handle.net/10356/103128

https://doi.org/10.1109/TIP.2015.2416634

Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning

Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, *Senior Member, IEEE,* Karthik Nandakumar, *Member, IEEE,* Li Jun, and Eam Khwang Teoh, *Member, IEEE*

***Abstract*—Recognition of natural emotions from human faces is an interesting topic with a wide range of potential applications like human-computer interaction, automated tutoring systems, image and video retrieval, smart environments, and driver warning systems. Traditionally, facial emotion recognition systems have been evaluated on laboratory controlled data, which is not representative of the environment faced in real-world applications. To robustly recognize facial emotions in real-world natural situations, this paper proposes an approach called Extreme Sparse Learning (ESL), which has the ability to jointly learn a dictionary (set of basis) and a non-linear classification model. The proposed approach combines the discriminative power of Extreme Learning Machine (ELM) with the reconstruction property of sparse representation to enable accurate classification when presented with noisy signals and imperfect data recorded in natural settings. Additionally, this work presents a new local spatio-temporal descriptor that is distinctive and pose-invariant. The proposed framework is able to achieve state-of-the-art recognition accuracy on both acted and spontaneous facial emotion databases.**

***Index Terms*—Emotion recognition, Facial emotion, Pose-invariance, Dictionary learning, Sparse representation, Extreme learning machine, Extreme sparse learning.**

I. Introduction

Facial emotion recognition in uncontrolled environments is a very challenging task due to large intra-class variations caused by factors such as illumination and pose changes, occlusion, and head movement. The accuracy of a facial emotion recognition system generally depends on two critical factors: (i) extraction of facial features that are robust under intra-class variations (e.g. pose changes), but are distinctive for various emotions, and (ii) design of a classifier that is capable of distinguishing different facial emotions based on noisy and imperfect data (e.g., illumination changes and occlusion).

The objective of the present work is to develop a facial emotion recognition system that is capable of handling variations in facial pose, illumination, and partial occlusion. The proposed system robustly represents the facial emotions using a novel spatio-temporal descriptor based on Optical Flow (OF), which is distinctive and pose-invariant. Robustness to pose variations is achieved by extracting features that depend only on relative movements of different facial regions. However, the feature encoding may fail in the case of extreme poses, where some parts of the face are not visible in the recorded images. To recognize the emotions in the presence of self-occlusion and illumination variations, we combine the idea of sparse representation with Extreme Learning Machine (ELM) to learn a powerful classifier that can handle noisy and imperfect data.

Sparse representation [1], [2] is a powerful tool for reconstruction, representation, and compression of high-dimensional noisy data (such as images/videos and features derived from them) due to its ability to uncover important information about signals from the base elements or dictionary atoms. While the sparse representation approach has the ability to enhance noisy data using a dictionary learned from clean data, it is not sufficient because our end goal is to correctly recognize the facial emotion. In a sparse-representation-based classification task, the desired dictionary should have both representational ability and discriminative power. Since separating the classifier training from dictionary learning may cause the learned dictionary to be sub-optimal for the classification task, we propose to jointly learn a dictionary (which may not be necessarily over-complete) and a classification model. To the best of our knowledge, this is the first attempt in the literature to simultaneously learn the sparse representation of the signal and train a ***non-linear*** classifier based on sparse codes.

The key contributions of this paper are as follows:

- A pose-invariant OF-based spatio-temporal descriptor, which is able to robustly represent facial emotions even when there are head movements while expressing an emotion. The proposed descriptor is capable of characterizing both the intensity and dynamics of facial emotions.
- A new classifier called Extreme Sparse Learning (ESL) is obtained by adding the ELM error term to the objective function of the conventional sparse representation to learn a dictionary that is both discriminative and reconstructive. This combined objective function (containing both linear and non-linear terms) is solved using a novel approach called Class Specific Matching Pursuit (CSMP). A kernel extension of the above framework called Kernel ESL (KESL) has also been developed.

II. Related Work

Facial emotion is an important cue for assessment of human affective behavior. While various techniques have been proposed for vision-based facial emotion recognition, majority of them focus on emotion recognition based on static images and ignore the temporal component of such a dynamic event

S. Shojaeilangari and E. K. Teoh are with the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore, e-mail: seyedehs1@e.ntu.edu.sg, EEKTEOH@ntu.edu.sg

W.-Y. Yau and J. Li are with the Institute for Infocomm Research, A*STAR, Singapore, e-mail: {wyyau, jli}@i2r.a-star.edu.sg

K. Nandakumar is with the IBM Research Collaboratory, Singapore, e-mail: nkarthik@sg.ibm.com

[3], [4]. However, research on the human visual system has demonstrated that better judgment of the facial emotion is achieved when the temporal information is taken into account [5]. Techniques that exploit the dynamics of facial emotion include hidden Markov models [6], dynamic Bayesian networks [7], geometrical displacement [8], and dynamic texture descriptors [9]. A comprehensive literature survey on facial emotion recognition can be found in [10]. However, most of the existing techniques are applicable only for laboratory-controlled data and are not able to deal with natural settings.

The following sub-sections present a review of (a) pose-invariant methods for feature extraction and (b) relevant works on sparse representation based classification.

A. *Pose-Invariant Feature Extraction*

Although facial emotion recognition has been extensively studied in the past, most of the existing feature extraction approaches require frontal facial images and even small changes in facial pose may reduce their effectiveness. Only a few researchers have attempted to solve the facial pose challenge.

In [11], a probabilistic method based on 2D geometrical features was proposed for pose-invariant facial emotion recognition. The locations of 39 landmarks were extracted from an expressed facial image with arbitrary head pose. The coupled scaled Gaussian process regression model was then applied to normalize the facial pose. Although the model was trained based on only a few discrete head poses, the method has ability to deal with continuous head pose variations. But the method requires accurate localization of facial landmarks, which is a very challenging task for automatic emotion recognition.

A face representation scheme using the regional covariance matrix was proposed in [12]. A dimensionality reduction step is then applied to the resulting features based on discriminant analysis. An effective approach was further proposed to find the optimal discriminant vectors. The key advantage of this method is that it does not need any facial alignment and feature point localization, which are both challenging tasks. However, this method is only applicable for facial emotion recognition based on static images.

A technique called variable-intensity template was proposed in [13] to obtain a person specific model for describing various facial emotions. The variable intensity templates define how the intensity of multiple facial points varies for an observed emotion. This method is able to detect the facial emotion and estimate the pose simultaneously within the framework of particle filtering. While this method is simple and has low computational cost, it is quite sensitive to errors in interest point localization and misalignment.

Since the dynamics of facial emotion is critical for a reliable facial emotion analysis, a variety of approaches focus on motion and OF based feature extraction [14], [15]. However, our proposed dynamic descriptor is different from existing OF based representations in three aspects: (i) we propose a new set of spatio-temporal features to capture the dynamic information hidden in a flow field, (ii) the extracted features are encoded effectively to achieve pose-invariance, and (iii) only the statistics of the extracted features is retained as discriminative information for further processing.

B. *Sparse Representation based Classification*

While learning a dictionary directly from the training data usually leads to satisfactory reconstruction from sparse codes, adding a specific discriminative criterion to dictionary training can improve the discriminative ability of the method and lead to better classification results. Recently, several methods have been developed to train a classification oriented dictionary. These methods can be divided into three broad categories.

- The first category of methods directly forces the dictionary atoms to be discriminative and uses the reconstruction error for the final classification [16], [17].
- The second approach makes the sparse coefficients discriminative by incorporating the classification error term into the dictionary learning and indirectly propagates the discrimination power to the overall dictionary [18]–[20]. Most of the techniques proposed in the literature for sparse representation based classification (including the one proposed in this paper) fall under this second category, where the classifier is simultaneously trained along with Dictionary Learning (DL).
- The third category includes methods that apply the discriminative criterion for coefficients, but the classifier is not necessarily trained along with DL. They either use the reconstruction error based classification or employ other classifiers on the resulting sparse representation [1], [21].

An example of the first approach is the scheme proposed in [16] for learning the class specific sub-dictionaries by incorporating a penalty term to make the sub-dictionaries incoherent. Another example of this approach is the classification-oriented DL model proposed in [17], which learns a class-specific dictionary (named particularity) to capture the most discriminative information of each category, and also a common pattern dictionary (named commonality) that only contributes the essential representation for all the data.

A supervised DL method that incorporates a logistic loss function to the classical reconstructive term to simultaneously learn a classifier was introduced in [18]. This work also proposed a general formulation of supervised DL and an efficient algorithm for solving the corresponding optimization criterion. The Discriminative K-Singular Value Decomposition (D-KSVD) method was proposed in [19] by introducing a discriminative term into the conventional objective function of K-SVD. The dictionary learned by this method was claimed to be both reconstructive and discriminative. The Label Consistent K-SVD (LCKSVD) algorithm proposed in [20] trains a discriminative dictionary utilizing the class label information of each dictionary atom. This algorithm incorporates sparse coding error and classification error criterion into a unified objective function, which is optimized using the K-SVD algorithm. The LCKSVD algorithm efficiently learns an over-complete, compact, and discriminative dictionary and a multi-class linear classifier simultaneously. However, the method cannot be directly extended to learn a non-linear classifier, which is required when the data is not linearly separable.

A good example of third category is the Fisher Discriminative Dictionary Learning (FDDL) proposed in [21]. In this method, dictionary learning based on Fisher discrimination criterion is used to improve the classification performance. The method aims to learn a structured dictionary, where the criterion imposed on sparse coding causes the sparse coefficients to have small within-class scatter but large between-class variance. Another example of the third approach is the scheme proposed in [1] for signal classification, which combines a reconstructive approach with a discriminative term using linear discriminant analysis and a predefined dictionary.

To the best of our knowledge, none of the existing methods can learn a non-linear classifier in the context of simultaneous sparse coding and classifier training. Learning such a non-linear classifier is not only an interesting research topic, but also very important in many real-world applications where the observations are not probably linearly separable. This paper is the first research work that explores how to simultaneously learn the sparse representation of the signal and train a non-linear classifier to be discriminative for sparse codes.

III. Proposed Methodology: Feature Extraction

We propose a set of pose-invariant features derived based on the optical flow (OF) extracted from the videos. To begin with, we define a new face coordinate system on the image plane as shown in Fig. 1(a). The algorithm[1] proposed in [22] is used for detecting the eyes and the nose tip. While the nose tip is considered as the origin of the face coordinate system, the reference vector connecting the nose tip to the midpoint between the centres of the two eyes is considered as the positive y-axis.

In order to compute the dynamic features, we start by computing the OF of a given video based on both brightness and gradient constancy assumption, combined with a discontinuity-preserving spatio-temporal smoothness constraint[2] [23]. Let $\mathbf{U}(\mathbf{P}, t_i)$ represent the flow vector (u, v) at pixel location $\mathbf{P} = (p_x, p_y)$ at time t_i. Note that all the pixel locations in the feature extraction stage are defined with respect to the new face coordinate system.

A. Optical Flow Correction

Since we are only interested in the local motion of facial components resulting from the act of expressing an emotion, global motion of the head is subtracted from the flow vector.

$$\mathbf{U}_{emo} = \mathbf{U}_{tot} - \mathbf{U}_{head}, \tag{1}$$

where $\mathbf{U}_{emo}$ is the emotion-related OF that we intend to measure, $\mathbf{U}_{tot}$ is the overall OF, and $\mathbf{U}_{head}$ is the OF representing the global head movement. Since head movement does not necessarily imply change in pose, the above optical flow correction only has a limited impact on pose invariance.

To measure $\mathbf{U}_{head}$, we divide the face into a few regions and compute the average flow vector in each region. If the angle difference between the flow vector at individual pixels and the corresponding average flow vector of the region is less than a threshold for a majority of the pixels, the average flow vector is considered as $\mathbf{U}_{head}$ for each pixel in that region. Otherwise, $\mathbf{U}_{head}$ is set to zero for all the pixels in that region. Note that in all the subsequent processing steps, $\mathbf{U}(\mathbf{P}, t_i)$ indicates only the emotion-related OF (i.e, $\mathbf{U}_{emo}$) and not the overall OF.

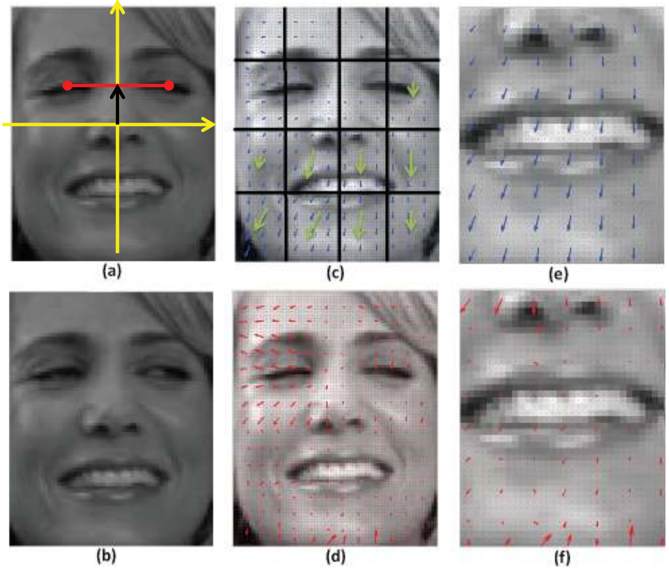


Fig. 1. Optical flow correction for head movement. (a) and (b) are two consecutive frames in a video; (a) also shows the proposed face coordinate system, where the nose tip is considered as the origin and the reference vector connects the nose tip to the midpoint between the centres of the two eyes; (c) total optical flow ($\mathbf{U}_{tot}$) illustrated in blue and the head movement optical flow ($\mathbf{U}_{head}$) indicated in green; (d) emotion related optical flow ($\mathbf{U}_{emo}$) illustrated in red; (e) $\mathbf{U}_{tot}$ of mouth region; and (f) $\mathbf{U}_{emo}$ of mouth region.

Fig. 1 shows an example of head movement correction using the above method. As shown in Figures 1(a) and 1(b), the emotion does not change between the two successive frames, but there is a slight head movement. Fig. 1(c) shows the region-wise estimate for $\mathbf{U}_{head}$. For some regions, $\mathbf{U}_{head}$ is not shown because majority of the movements in these regions are not coherent ($\mathbf{U}_{head} = \mathbf{0}$). For better illustration, we zoomed out the OF of the mouth region before and after correction in Figures 1(e-f). Fig. 1(f) shows that the emotion-related OF ($\mathbf{U}_{emo}$) is almost zero in the mouth region.

B. Spatio-Temporal Features

Four pose-invariant features are proposed for encoding the motion information of facial components. The first feature is the divergence of the flow field, which measures the amount of local expansion or contraction of the facial muscles.

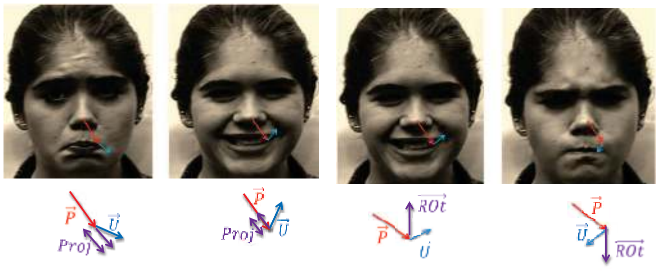
$$Div(\mathbf{U}(\mathbf{P}, t_i)) = \frac{\partial u(\mathbf{P}, t_i)}{\partial x} + \frac{\partial v(\mathbf{P}, t_i)}{\partial y}, \tag{2}$$

where $\frac{\partial u(\mathbf{P}, t_i)}{\partial x}$ and $\frac{\partial v(\mathbf{P}, t_i)}{\partial y}$ are the partial derivatives of u and v components of the OF along the x and y directions, respectively. We used a simple Prewitt operator to compute the gradient of the OF.

The second feature captures the local spin around the axis perpendicular to the OF plane and is referred to as $Curl$. It

[1] The source code for nose point detection is available at: http://humansensing.cs.cmu.edu/intraface/download.html

[2] The source code for OF extraction is available at: http://www.mathworks.com/matlabcentral/fileexchange/17500-high-accuracy-optical-flow



(a) (b)

Fig. 2. (a)Illustration of projection feature ($Proj$) for sad (left) and happy (right) emotions; and (b) illustration of rotation feature (Rot) for happy (left) and anger (right) emotions.

is useful to measure the dynamics of the local circular motion of the facial components.

$$Curl(\mathbf{U}(\mathbf{P},t_i)) = \frac{\partial v(\mathbf{P},t_i)}{\partial x} - \frac{\partial u(\mathbf{P},t_i)}{\partial y}, \tag{3}$$

where $\frac{\partial v(\mathbf{P},t_i)}{\partial x}$ and $\frac{\partial u(\mathbf{P},t_i)}{\partial y}$ are the partial derivatives of v and u components of the OF along the x and y directions, respectively.

The third feature is the scalar projection of the OF vector $\mathbf{U}$ onto the unit position vector $\hat{\mathbf{P}}$, where $\hat{\mathbf{P}} = \mathbf{P}/\|\mathbf{P}\| = (\hat{p}_x, \hat{p}_y)$.

$$Proj(\mathbf{U}(\mathbf{P},t_i)) = \mathbf{U}\cdot\hat{\mathbf{P}} = u\hat{p}_x + v\hat{p}_y. \tag{4}$$

This $Proj$ feature captures the amount of expansion or contraction of each point with respect to the nose point. For example, the “happy” and “sad” emotions can be distinguished by this feature clearly. Fig. 2(a) shows how the sign and magnitude of the $Proj$ feature changes for a sample lip point (the magnitude is exaggerated for better illustration) depending on the facial emotion.

The rotation (Rot) feature is the defined as the cross product of the unit position vector $\hat{\mathbf{P}}$ and OF vector $\mathbf{U}$ as follows:

$$Rot(\mathbf{U}(\mathbf{P},t_i)) = \hat{\mathbf{P}}\times\mathbf{U} = v\hat{p}_x - u\hat{p}_y. \tag{5}$$

Since both $\hat{\mathbf{P}}$ and $\mathbf{U}$ lie on the image plane, their cross product is a vector perpendicular to the image (x-y) plane. For simplicity, we consider only the coefficient of this cross-product vector and treat it as a scalar quantity. The Rot feature measures the amount of clockwise or anti-clockwise rotation of each facial point movement with respect to the position vector. Fig. 2(b) illustrates the usefulness of this feature in distinguishing between “happy” and “anger” emotions. As shown in this figure, the sign and magnitude of Rot feature are different for a sample lip point (the magnitude is exaggerated for better illustration) depending on the facial emotion.

C. Spatio Temporal Descriptor Construction

A spatio-temporal descriptor is obtained by concatenating the spatio-temporal features extracted at each local region in the video. Fig. 3 illustrates the construction of the spatio-temporal descriptor. The local regions are determined by dividing the volumetric data into M 3D blocks (could be overlapping or non-overlapping) as shown in Fig. 3a. To preserve the geometric information of descriptors, each block is further divided into N 3D cells as illustrated in Fig. 3b.

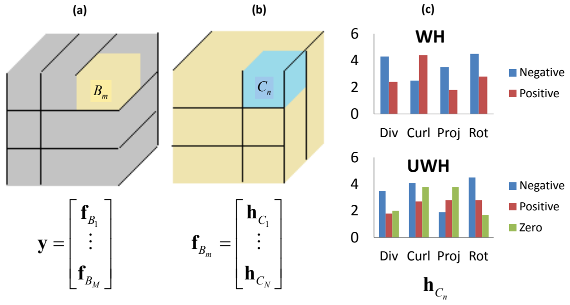

(a)
(b)
(c)
B_m
C_n
$\mathbf{y} = \begin{bmatrix}\mathbf{f}_{B_1}\\ \vdots \\ \mathbf{f}_{B_M}\end{bmatrix}$
$\mathbf{f}_{B_m} = \begin{bmatrix}\mathbf{h}_{C_1}\\ \vdots \\ \mathbf{h}_{C_N}\end{bmatrix}$
WH
6
4
2
0
Div
Curl
Proj
Rot
Negative
Positive
UWH
6
4
2
0
Div
Curl
Proj
Rot
Negative
Positive
Zero
$\mathbf{h}_{C_n}$

Fig. 3. Spatio-temporal descriptor construction. (a) Volume data is divided into a number of 3D blocks ($\boldsymbol{B}_m$) and the final descriptor ($\mathbf{y}$) is a concatenation of features from all the blocks. (b) Each block is further divided into number of 3D cells ($\boldsymbol{C}_n$) and the feature vector of each block ($f_{\boldsymbol{B}_m}$) is a concatenation of all the cell histograms within that block. (c) Weighted and un-weighted histograms are calculated for each cell based on the four spatio-temporal features and concatenated to obtain the cell histogram.

Two types of histograms, namely, weighted histogram (WH) and un-weighted histogram (UWH), are used to aggregate the features in each cell. The weighted histogram characterizes the magnitude of emotion, i.e., it differentiates a subtle emotion from an exaggerated emotion. Each WH consists of two bins - positive and negative bins, and the magnitude of the associated features is used to vote for each bin. The un-weighted histogram ignores the magnitude of the emotion and attempts to characterize its dynamics. It involves three bins related to positive, negative, and zero features. Equal vote is assigned for each bin, which means that the total number of positive, negative, and zero features are counted. The UWH minimizes the effect of changes in the emotion speed[3] by considering only the sign (positive, negative, or zero) of the features and ignoring their magnitude. Thus, the two histograms (WH and UWH) encode complementary information that can potentially improve the classification performance.

WH and UWH are computed for each cell based on the four spatio-temporal features (Div, $Curl$, $Proj$, and Rot) described earlier. The concatenation of all these eight histograms is considered as the final descriptor of the corresponding cell, $\mathbf{h}_{\mathcal{C}_n}$, $n = 1, 2, \cdots, N$. The dimensionality of each cell descriptor is 20 as shown in Fig. 3c. The concatenation of all the cell descriptors gives the block descriptor, $\mathbf{f}_{\mathcal{B}_m}$, $m = 1, 2, \cdots, M$. The concatenation of all the block descriptors results in the final spatio-temporal descriptor ($\mathbf{y}$) representing the given video sequence. The dimensionality of the spatio-temporal descriptor $\mathbf{y}$ is $20MN$.

IV. Proposed Methodology: Recognition Framework

In this section, we propose a dictionary-based classification method called Extreme Sparse Learning (ESL) to recognize

[3] The emotion speed is inversely proportional to time lapse (or number of frames) between the start of the expression and the peak expression.

facial emotions in real-world natural situations. The proposed approach combines the discriminative power of Extreme Learning Machine (ELM) with the reconstruction capability of sparse representation. The key motivation behind the use of sparse representation is its inherent ability to reconstruct the original signals from noisy and imperfect samples (in this context, imperfect data may refer to cases with large pose variations, occlusion, and illumination changes) based on a learned dictionary [2]. By simultaneously learning a dictionary for sparse representation and a classification model, the proposed ESL algorithm is able to implicitly handle illumination and occlusion changes. Before introducing the ESL, we briefly present the concepts underlying sparse representation and ELM in the following sub-sections.

A. Sparse Representation and Dictionary Learning

The basic assumption underlying sparse representation is that natural signals or images can be efficiently approximated by linear combination of a few elements (so called atoms) of a dictionary. One of the critical issues in sparse representation is the choice of the dictionary. The dictionary can be obtained by either applying predefined transforms to the data (e.g., Fourier transforms) or directly learning from training data. Since Dictionary Learning (DL) directly from the training data usually leads to a satisfactory reconstruction, we applied this technique in our proposed ESL algorithm.

Let $\boldsymbol{Y}$ be a set of $\mathcal{S}$ input signals of dimension $\mathcal{N}$, i.e. $\boldsymbol{Y} = [\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_\mathcal{S}] \in \mathbb{R}^{\mathcal{N}\times\mathcal{S}}$. Learning a reconstructive dictionary for sparse representation of $\boldsymbol{Y}$ can be accomplished by solving the following problem:

$$\min_{\boldsymbol{X},\boldsymbol{D}} \left(\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_2^2\right) \quad s.t. \quad \|\mathbf{x}_i\|_0 \leq \mathcal{N}_0, \tag{6}$$

where $\boldsymbol{D} = [\mathbf{d}_1\mathbf{d}_2\cdots\mathbf{d}_\mathcal{M}] \in \mathbb{R}^{\mathcal{N}\times\mathcal{M}}$ is the learned overcomplete dictionary, $\boldsymbol{X} = [\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_\mathcal{S}] \in \mathbb{R}^{\mathcal{M}\times\mathcal{S}}$ is the sparse code matrix of the input signals, $\mathcal{N}_0$ is the sparsity constraint, and $\|\cdot\|_0$ is the ℓ^0 pseudo norm that counts the number of non-zero elements. The K-SVD algorithm [24] is an efficient technique for solving the optimization problem in (6).

B. Extreme Learning Machine (ELM)

ELM is considered to be a state-of-the-art classification technique, especially for multi-class classification problems. ELM requires fewer optimization constraints in comparison to Support Vector Machines (SVM), which results in simple implementation, fast learning, and better generalization performance [25]. Therefore, ELM is a good choice for the problem of facial emotion recognition. We believe that combining ELM with sparse representation and dictionary learning can lead to further improvement in recognition performance.

The objective function of ELM can be summarized as:

$$\min_{\beta} \left(\|H(\boldsymbol{X})\beta - \mathbf{Z}\|_2^2 + \|\beta\|_2^2\right), \tag{7}$$

where $\boldsymbol{X}$ denotes the set of training samples, H is the hidden layer output matrix ($H(\boldsymbol{X}) \in \mathbb{R}^{\mathcal{S}\times\mathcal{L}}$, where $\mathcal{L}$ is the number of nodes in the hidden layer), β is the output weight vector of length $\mathcal{L}$, and $\mathbf{Z}$ is the vector of class labels of length $\mathcal{S}$.

The minimal norm least squares method can be used to solve the above optimization problem, whose solution is represented as follows:

$$\begin{aligned} \text{when } \mathcal{S} < \mathcal{L} : \beta = H^\dagger\mathbf{Z} = &\quad H^T\left(\tfrac{I}{c} + HH^T\right)^{-1}\mathbf{Z}, \\ \text{when } \mathcal{S} > \mathcal{L} : \beta = H^\dagger\mathbf{Z} = &\quad \left(\tfrac{I}{c} + H^TH\right)^{-1}H^T\mathbf{Z}, \end{aligned} \tag{8}$$

where $H^\dagger$ is the Moore-Penrose generalized inverse of matrix H, H^T is the transpose of the matrix H, I is the identity matrix, and c is a the user-specified parameter added to the formulation for better generalization performance [25].

It has been shown in the literature that a wide variety of feature mappings including random hidden nodes and kernels can be utilized for ELM [25]. For unknown feature mappings, kernels can be applied interchangeably, resulting in the Kernel ELM (KELM) approach. We conducted preliminary experiments to evaluate various types of kernels and activation functions. Based on the results of these experiments, we chose a sigmoid activation function for the hidden nodes of the ELM and a polynomial kernel for KELM.

C. Extreme Sparse Learning (ESL)

Separating the classifier training from dictionary learning may lead to a scenario where the learned dictionary is not optimal for the classification task. Therefore, we propose to jointly learn the dictionary and the classification model for better performance. Learning a discriminative dictionary for sparse representation of $\boldsymbol{Y}$ can be accomplished by solving the following optimization problem:

$$\min_{\boldsymbol{X},\boldsymbol{D},\beta} \mathcal{E}_t(\boldsymbol{X},\boldsymbol{D},\beta,\boldsymbol{Y},\mathbf{Z}), \text{ where}$$

$$\begin{aligned} \mathcal{E}_t(\boldsymbol{X},\boldsymbol{D},\beta,\boldsymbol{Y},\mathbf{Z}) &= (\mathcal{E}_r + \gamma_1\mathcal{E}_c + \gamma_2\mathcal{E}_s), \\ \mathcal{E}_r(\boldsymbol{X},\boldsymbol{D},\boldsymbol{Y}) &= \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_2^2, \\ \mathcal{E}_c(\boldsymbol{X},\beta,\boldsymbol{Y},\mathbf{Z}) &= \|H(\boldsymbol{X})\beta - \mathbf{Z}\|_2^2 + \|\beta\|_2^2, \\ \mathcal{E}_s(\boldsymbol{X}) &= \|\boldsymbol{X}\|_1. \end{aligned} \tag{9}$$

In the above equation, $\mathcal{E}_t$ represents the overall objective function, $\mathcal{E}_r$ measures the reconstruction error, $\mathcal{E}_c$ represents the ELM optimization constraints, and $\mathcal{E}_s$ represents the sparsity constraint. The notation $\|\cdot\|_1$ indicates the ℓ^1 norm that simply sums up the absolute value of the elements. Note that formulating the sparsity constraint using ℓ^1 norm simplifies the optimization problem without affecting the sparse representation significantly [26]. The parameters γ_1 and γ_2 are regularization terms, which control the relative contributions of reconstruction error, classification error, and sparsity constraints to the final objective function.

The framework in (9) is referred to as Extreme Sparse Learning (ESL). When kernels are incorporated in the above framework, we refer to it as Kernel ESL (KESL). Fig. 4 shows a high-level outline of the proposed ESL recognition framework. The associated training and classification algorithms are presented in Algorithm 1 and Algorithm 2, respectively.

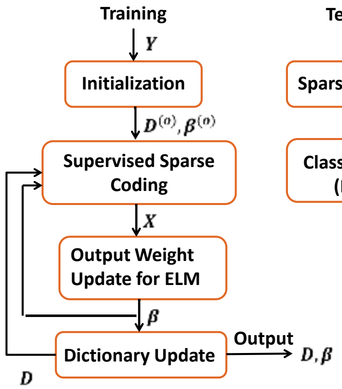


Fig. 4. High-level outline of the proposed ESL framework.

As shown in Fig. 4, there are three main steps involved in ESL training: supervised sparse coding, ELM output weight update, and dictionary optimization. The supervised sparse coding step learns the sparse code matrix $\boldsymbol{X}$ corresponding to the input signals $\boldsymbol{Y}$ based on the given dictionary $\boldsymbol{D}$ and the ELM output weight vector β. Based on this estimated sparse code matrix, the ELM output weight vector β is updated in the second step. These two steps are repeated until the first stopping criterion is met. The first stopping criterion is triggered when the value (e_1) of the objective function $\mathcal{E}_t$ falls below a threshold (ϵ_1) or when the maximum number of iterations (η_1) is reached. The output of this inner loop is the estimated sparse code matrix $\boldsymbol{X}$ and the updated ELM weight vector β. Finally, the dictionary atoms are updated based on the sparse code matrix $\boldsymbol{X}$ to obtain the updated dictionary $\boldsymbol{D}$. All three steps are iteratively repeated, until the second stopping criterion is met. The second stopping criterion is triggered when the value (e_2) of the objective function $\mathcal{E}_t$ falls below a threshold (ϵ_2) or when the maximum number of iterations (η_2) is reached. The output of ESL training is the learned dictionary $\boldsymbol{D}$ and the ELM output weight β.

For classification, the sparse coefficient vector $\mathbf{x}$ of the given test sample $\mathbf{y}$ is first estimated using the learned dictionary $\boldsymbol{D}$. The sparse vector $\mathbf{x}$ is then classified by ELM with output weight vector β as shown in Algorithm 2.

1) Supervised Sparse Coding: Class Specific Matching Pursuit (CSMP): The most critical step in the proposed ESL training algorithm is supervised sparse coding, which estimates a sparse code matrix $\boldsymbol{X}$ that simultaneously minimizes the reconstruction error based on the given dictionary $\boldsymbol{D}$ and the classification error based on the given ELM output weight vector β. In this section, we propose a novel algorithm called Class Specific Matching Pursuit (CSMP) to perform supervised sparse coding.

The objective of supervised sparse coding can be summarized by the following equation:

$$\min_{\boldsymbol{X}} \mathcal{E}_t(\cdot, \boldsymbol{D}, \beta, \boldsymbol{Y}, \mathbf{Z}) \tag{10}$$

Although different methods have been suggested in the literature for supervised sparse coding [19], [20], [27], [28], none of them can be applied for (10) directly due to the presence of the non-linear term $H(\boldsymbol{X})$. In this paper, we have developed an algorithm for supervised sparse coding inspired by the simultaneous sparse approximation algorithm [1] and Simultaneous Orthogonal Matching Pursuit (S-OMP) [27]. While S-OMP seeks to find a set of dictionary atoms that best represents all the signals irrespective of their class labels, the proposed method attempts to find a fixed set of atoms that can optimize the objective function in (10) for all signals that belong to the same class. Therefore, the proposed method is referred to as Class Specific Matching Pursuit (CSMP).

Algorithm 1 Training Algorithm for Extreme Sparse Learning

```
Input: Training set $\boldsymbol{Y}$, vector of class labels $\mathbf{Z}$ corresponding to $\boldsymbol{Y}$, regularization terms $\gamma_1$ and $\gamma_2$, stopping criterion for inner loop $(\epsilon_1, \eta_1)$, and stopping criterion for outer loop $(\epsilon_2, \eta_2)$
Output: Dictionary $\boldsymbol{D}$ and ELM output weight vector $\beta$
Initialize $\boldsymbol{D} \leftarrow \boldsymbol{D}^{(0)}$, $\beta \leftarrow \beta^{(0)}$, $i_2 \leftarrow 0$
repeat
  $i_1 \leftarrow 0$
  repeat
    $i_1 \leftarrow i_1 + 1$
    $\boldsymbol{X} \leftarrow \min_{\boldsymbol{X}} \mathcal{E}_t(\cdot, \boldsymbol{D}, \beta, \boldsymbol{Y}, \mathbf{Z})$
    $\beta \leftarrow \min_{\beta} \mathcal{E}_c(\boldsymbol{X}, \cdot, \boldsymbol{Y}, \mathbf{Z})$
    $e_1 \leftarrow \mathcal{E}_t(\boldsymbol{X}, \boldsymbol{D}, \beta, \boldsymbol{Y}, \mathbf{Z})$
  until $(e_1 < \epsilon_1)$ OR $(i_1 > \eta_1)$
  $i_2 \leftarrow i_2 + 1$
  $\boldsymbol{D} \leftarrow \min_{\boldsymbol{D}} \mathcal{E}_r(\boldsymbol{X}, \cdot, \boldsymbol{Y})$
  $e_2 \leftarrow \mathcal{E}_t(\boldsymbol{X}, \boldsymbol{D}, \beta, \boldsymbol{Y}, \mathbf{Z})$
until $(e_2 < \epsilon_2)$ OR $(i_2 > \eta_2)$
return $\boldsymbol{D}$ and $\beta$
```

Algorithm 2 Classification Algorithm for Extreme Sparse Learning

```
Input: Test sample $\mathbf{y}$, learned dictionary $\boldsymbol{D}$, ELM output weight vector $\beta$, total number of classes $\omega$, and regularization term $\gamma$
Output: Class label $z$ of test sample
$\mathbf{x} \leftarrow \min_{\mathbf{x}} \left(\|\mathbf{y} - \boldsymbol{D}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1\right)$
$z \leftarrow \min_{z} |H(\mathbf{x})\beta - z|$, $z \in \{1, 2, \cdots, \omega\}$
return $z$
```

Algorithm 3 shows the steps involved in the CSMP method. The basic idea underlying the matching pursuit process is to sequentially find atoms in the dictionary and the corresponding sparse coefficients that minimize the objective function $\mathcal{E}_t$ defined in (9) and (10). This process is repeated for each class. More specifically, during each iteration j of the while loop in Algorithm 3, one atom (with index λ_j) is chosen from the list of unselected atoms (denoted by Ω_j) and added to the selected index list (denoted by Λ_j). This selection is done based on the value of the total error E_t, which is a weighted combination of the reconstruction error E_r, the classification error E_c, and the sparsity constraint E_s. The innermost for loop in Algorithm 3 iterates through all the unselected atoms and computes the value of E_t based on each one of them. The sparse code matrix $(\widehat{\boldsymbol{X}})$ is then updated based on the subset of atoms given by

Λ_j and the new value (e_3) of the objective function $\mathcal{E}_t$ is computed. The above iterative process is repeated until all the atoms in the dictionary are exhausted or e_3 becomes less than the stopping threshold ϵ_3.

Algorithm 3 Class Specific Matching Pursuit

Input: Training set $\boldsymbol{Y}$, vector of class labels $\mathbf{Z}$ corresponding to $\boldsymbol{Y}$, total number of classes ω, number of dictionary atoms $\mathcal{M}$, regularization terms γ_1 and γ_2, stopping threshold ϵ_3, dictionary $\boldsymbol{D}$, ELM output weight vector β, and sparse code matrix from previous iteration $\boldsymbol{X}^{(old)}$

Output: Updated sparse code matrix $\boldsymbol{X}^{(new)}$

Notation: Let $\boldsymbol{A}_\Theta$ denote a sub-matrix containing only the columns of matrix $\boldsymbol{A}$ whose indices are included in set Θ, $\boldsymbol{A}_{*,q}$ denote the q-th column of $\boldsymbol{A}$, $\boldsymbol{A}_{q,*}$ denote the q-th row of $\boldsymbol{A}$, $\boldsymbol{A}^\dagger$ denote the pseudo-inverse of matrix $\boldsymbol{A}$, and ϕ is the empty set

```
Initialize $\boldsymbol{X}^{(new)} \leftarrow \boldsymbol{X}^{(old)}$
for $i = 1$ to $\omega$ do
  $\Theta_i \leftarrow$ Indices of all training samples that belong to class $i$
  $\mathcal{S}_i \leftarrow$ number of elements in the set $\Theta_i$
  $\mathbf{Z}_i \leftarrow \mathcal{S}_i$-dimensional vector with all elements taking value $i$
  Initialize $\widehat{\boldsymbol{X}} \leftarrow \boldsymbol{0}$, $\widehat{\boldsymbol{X}} \in \mathbb{R}^{\mathcal{M} \times \mathcal{S}_i}$
  Initialize $\Lambda_0 \leftarrow \phi$, $\Omega_1 \leftarrow \{1, 2, \cdots, \mathcal{M}\}$, $j \leftarrow 1$
  while $j \leq \mathcal{M}$ do
    $m \leftarrow$ number of elements in the set $\Omega_j$
    $\Delta \leftarrow \boldsymbol{D}_{\Omega_j}, \Delta \in \mathbb{R}^{\mathcal{N} \times m}$
    $\Gamma \leftarrow \Delta^\dagger \boldsymbol{Y}_{\Theta_i}, \Gamma \in \mathbb{R}^{m \times \mathcal{S}_i}$
    for $k = 1$ to $m$ do
      $E_r \leftarrow \|\boldsymbol{Y}_{\Theta_i} - \Delta_{*,k}\Gamma_{k,*}\|_2^2$
      $E_s \leftarrow \|\Gamma_{k,*}\|_1$
      $\kappa \leftarrow k$-th element of set $\Omega_j$
      $\widetilde{\boldsymbol{X}} \leftarrow \widehat{\boldsymbol{X}}$
      $\widetilde{\boldsymbol{X}}_{\kappa,*} \leftarrow \Gamma_{k,*}$
      $E_c \leftarrow$ ELM classification error for class $i$ based on $\widetilde{\boldsymbol{X}}$ and $\beta$
      $E_t(k) \leftarrow E_r + \gamma_1 E_c + \gamma_2 E_s$
    end for
    $\lambda_j \leftarrow \Omega_j(argmin_k\ E_t(.))$
    $\Lambda_j \leftarrow \Lambda_{j-1} \cup \lambda_j$
    $\Omega_{j+1} \leftarrow \Omega_j \setminus \lambda_j$
    $\widehat{\boldsymbol{X}}_{\Lambda_j} \leftarrow (\boldsymbol{D}_{\Lambda_j})^\dagger \boldsymbol{Y}_{\Theta_i}$
    $e_3 \leftarrow \mathcal{E}_t(\widehat{\boldsymbol{X}}, \boldsymbol{D}, \beta, \boldsymbol{Y}_{\Theta_i}, \mathbf{Z}_i)$
    if $(e_3 < \epsilon_3)$ then
      break while loop
    end if
    $j \leftarrow j + 1$
  end while
  $\boldsymbol{X}_{\Theta_i}^{(new)} \leftarrow \widehat{\boldsymbol{X}}$
end for
return $\boldsymbol{X}^{(new)}$
```

2) Dictionary Update Stage: The objective of this step is to find the dictionary $\boldsymbol{D}$ that minimizes the reconstruction error of signal $\boldsymbol{Y}$ estimated by the sparse code matrix $\boldsymbol{X}$ as follows:

$$\min_{\boldsymbol{D}} \mathcal{E}_r(\boldsymbol{X}, \cdot, \boldsymbol{Y}). \tag{11}$$

We used the classical "Projected Gradient Descent" method [18] to solve this optimization problem. Given a dictionary $\boldsymbol{D}^{(old)}$ from the previous iteration, it is updated as follows:

$$\boldsymbol{D}^{(new)} = \mathcal{H}(\boldsymbol{D}^{(old)} - \tau \nabla \mathcal{E}_r), \tag{12}$$

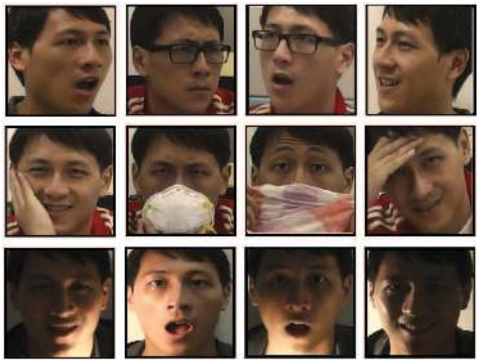


Fig. 5. Sample frames of a single subject from our own collected data. The top row shows some examples of pose variation, the middle row depicts occlusion examples, and the bottom row includes illumination variations.

where $\mathcal{H}$ is a simple normalizing function that forces each dictionary atom to be of unit norm, τ is the step size, $\nabla \mathcal{E}_r = 2(\boldsymbol{D}^{(old)}\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{X}^T$, and $\boldsymbol{X}^T$ is the transpose of matrix $\boldsymbol{X}$. We follow the Barzilai-Borwein technique to calculate the optimal step size iteratively [29].

Since our final objective is not perfect reconstruction, we have not used an over-complete dictionary. Indeed, over-completeness is not always necessary for the classification task as long as discriminative features are captured in the sparse coding procedure [18].

V. Database Description & Pre-processing

We have evaluated the performance of the proposed spatio-temporal descriptor and ESL algorithm on four databases for the facial emotion recognition task.

A. Cohn-Kanade (CK$^+$) Database

The CK$^+$ dataset [30] contains acted emotional data captured under controlled environmental conditions. It consists of 309 video sequences, where each sequence is labeled with one of the six basic emotions (joy, surprise, anger, fear, disgust, and sadness). Since the location of nose point is provided in the database, preprocessing involves only face alignment based on constant distance between the two eyes.

B. Extended Cohn-Kanade (ECK$^+$) Database

To evaluate the robustness of our proposed approach under difficult environmental conditions, the CK$^+$ database is appended with data collected in our lab, which includes 42 samples of head pose variation, 15 samples of illumination changes, and 18 samples of facial occlusion. Three subjects were asked to show one emotion (anger, happiness, sadness, and surprise) from neutral to apex. Some sample frames from our database are depicted in Fig. 5. The localization of nose points and face cropping have been performed manually.

C. AVEC2011 Database

The Audio Visual Emotion Challenge (AVEC 2011) database [31] contains spontaneous emotional states in naturalistic situations. It consists of 95 videos recorded at 49.979

frames per second. Binary labels along the four affective dimensions (activation, expectation, power and valence) are provided for each video frame. The data is divided into 3 subsets: training, development, and test, containing 31, 32, and 11 sequences, respectively. Due to processor and memory constraints, we sample the videos in the training and development sets as follows. We partition each video into segments containing 60 frames with 20% overlap between the segments. Only 10 frames per segment are selected for processing (by downsampling at a rate of 6), resulting in 1550 frames for each video. Information about the position of the face and eyes are provided in the database. Thus, the preprocessing stage includes only normalization of the face to achieve a constant distance between the two eyes.

D. EmotiW Database

The Emotion Recognition in the Wild (EmotiW) dataset [32] contains realistic challenges like pose variations, various illumination conditions, occlusion, and spontaneous emotions. EmotiW database is a collection of short video clips collected from some popular movies, where the actor is expressing one of seven emotions (anger, disgust, fear, happy, neutral, sad, and surprise) under near real-world conditions. EmotiW consists of three sets for training, validation, and testing including 380, 396, and 312 video clips, respectively. We use the faces detected by a simple eye based alignment method as provided by the organizers of EmotiW2013. This database is challenging due to the following reasons:

- Due to large pose variations, the face detection failure rate is quite high.
- Many video clips consist of more than one human subject, making it difficult to isolate the subject of interest.
- There is a wide difference in the way that the same emotion is expressed by the various subjects. Some of emotions are very confusing and hard even for a human expert to identify correctly.

VI. Results and Discussion

We systematically evaluate each component of the proposed facial emotion recognition framework. Note that the proposed algorithms were implemented using Matlab 7.11.0 running on a Core i5 CPU (2.8 GHz with 16 GB RAM).

A. Parameter setting and initialization

For OF extraction, we use the default setting of parameters in [23] as it was claimed that the algorithm is insensitive to parameter variations. We also observed that the number of regions (equivalently, the block size used for computing the average flow vector) used in optical flow correction does not have a significant impact on the final classification performance.

We carried out a preliminary experiment to determine the numbers of blocks and cells to be used in feature extraction and evaluated its effect on feature dimension and classification performance. Based on these results, the volume data is partitioned into 100 blocks ($10 \times 10 \times 1$) and 4 cells ($2 \times 2 \times 1$) per block, because this setting achieves the highest classification accuracy.

TABLE I
Parameter Setting for ESL and KESL.

Method	**Databases**					
	ECK$^+$		**AVEC 2011**		**EmotiW**	
	ESL	**KESL**	**ESL**	**KESL**	**ESL**	**KESL**
No. of atoms ($\mathcal{M}$)	100	100	150	150	200	200
γ_1	2	1.5	1	2	2	1
γ_2	0.1	0.05	0.05	0.5	0.05	0.01
c	16	1	100	8	16	8
d	-	0.5	-	0.25	-	0.125
n	-	8	-	8	-	6

For the classification part, we performed a greedy search to select the ELM parameter c (refer to (8)), kernel parameters, and regularization parameters γ_1 and γ_2 (refer to (9)). We used the polynomial kernel ($\mathcal{K}(\mathbf{a}, \mathbf{b}) = (\mathbf{a}.\mathbf{b} + d)^n$) for all experiments using KESL, SVM, and KELM. The optimal values of all the classification parameters for the different databases are summarized in Table I.

The sensitivity of the proposed algorithm with respect to the regularization parameters γ_1 and γ_2 was analyzed on the ECK$^+$ database. Our experiments indicate that the best performance is obtained by setting $\gamma_1 > 1 > \gamma_2$. In other words, the best performance is obtained when the classification error term (with weight γ_1) is assigned the highest priority, the reconstruction error term is assigned the second priority, and the sparsity constraint (with weight γ_2) is assigned the least priority. This result is intuitive because we are primarily interested in classification accuracy for this application. However, for other applications such as noise reduction, the priorities may change. Therefore, the regularization terms should be set according to the application.

We need to initialize the dictionary $\boldsymbol{D}$ and the ELM output weight β in Algorithm 1. There are some suggestions for dictionary initialization in the literature [20], [21]. In our experiments, a sub-matrix of $\boldsymbol{Y}$ containing randomly selected input samples from all the classes is treated as the initial $\boldsymbol{D}^{(0)}$. To obtain $\beta^{(0)}$, we first compute the initial sparse matrix $\boldsymbol{X} = (\boldsymbol{D}^{(0)})^{\dagger}\boldsymbol{Y}$, where † represents the pseudo-inverse, and then apply (8) to get the initial ELM output weight.

One important parameter is the number of dictionary atoms ($\mathcal{M}$). If this parameter is too small, the dictionary will not be very representative. On the other hand, the execution time would be prohibitively large for a large number of dictionary atoms. In fact, the number of dictionary atoms should be set depending on the characteristics of the database. If the same emotion is expressed differently by different subjects (as in the EmotiW database), the number of dictionary atoms should be on the higher side. However, if the subjects show the same emotion in a similar fashion (as in the CK$^+$ database), the number of dictionary atoms need not be large. Fig. 6 plots the recognition rates of ESL as a function of the number of dictionary atoms for different databases. We can observe that the recognition rate increases with the number of dictionary atoms only up to a point (100 atoms for ECK$^+$, 150 atoms for AVEC2011, and 200 atoms for EmotiW in Fig. 6).

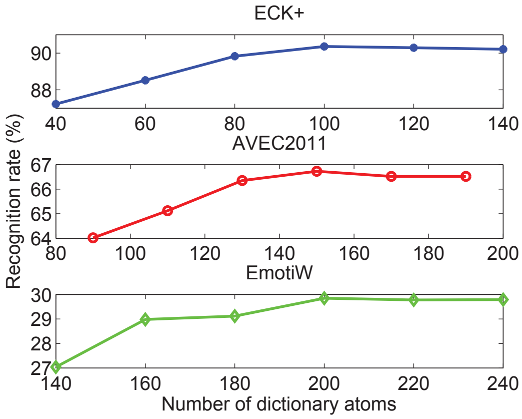


Fig. 6. Recognition rate of ESL as a function of the number of dictionary atoms.

TABLE II
EFFECT OF DIFFERENT COMBINATIONS OF FEATURES ON THE CLASSIFICATION ACCURACY. THESE RESULTS ARE OBTAINED BASED ON THE CK$^+$ DATABASE USING SVM AS THE CLASSIFIER.

Feature Combination	**Recognition rate (%)**
Div (WH+UWH)	92.42
$Curl$ (WH+UWH)	91.05
$Proj$ (WH+UWH)	84.75
Rot (WH+UWH)	88.57
Div+$Curl$+$Proj$ (WH+UWH)	94.83
Div+$Curl$+Rot: (WH+UWH)	93.38
Div+$Proj$+Rot: (WH+UWH)	92.37
$Curl$+$Proj$+Rot: (WH+UWH)	92.95
All 4 features (WH only)	90.66
All 4 features (UWH only)	91.91
All 4 features (WH+UWH)	**95.33**

B. Performance of the Spatio-Temporal Descriptor

To evaluate the accuracy of the proposed spatio-temporal descriptor for the facial emotion recognition task, we conducted a series of experiments on the CK$^+$ database. In all these experiments, support vector machine (SVM) with polynomial kernel is used as the classifier.

First, we evaluated the performance based on the two types of histograms (WH and UWH) individually. The results in the last three rows of Table II clearly show that both WH and UWH encode complementary information, which can potentially improve the classification performance. Secondly, we evaluated the performance based on each feature (Div, $Curl$, $Proj$, and Rot)) individually. As shown in Table II, the features contain complementary information and an ensemble of all these features gives better classification performance compared to any subset of these four features.

Finally, we compared the performance of the proposed spatio-temporal descriptor to two other successful dynamic descriptors in this field: Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [9] and Local Phase Quantization on Three Orthogonal Planes (LPQ-TOP) [33]. The source codes of these methods are publicly available[4]. We have attempted to set the parameters of each method to get the best result. Experiments were conducted using 5-fold cross-validation and the average results are reported. The execution time was also measured for the extraction of descriptors from a volume data of size ($100 \times 100 \times 10$). As shown in Table III, the recognition rate of the proposed descriptor is significantly better than the other two descriptors, but at the cost of increased execution time. The dimensionality of the proposed descriptor is also less compared to the other two methods.

TABLE III
COMPARISON OF THE PROPOSED SPATIO-TEMPORAL DESCRIPTOR TO OTHER DYNAMIC DESCRIPTORS. THESE RESULTS ARE OBTAINED BASED ON THE CK$^+$ DATABASE USING SVM AS THE CLASSIFIER.

Method	**No. of Features**	**Recognition rate (%)**	**Time complexity (Sec)**
LBP-TOP [9]	17700	89.31	4.26
LPQ-TOP [33]	76800	89.17	**2.19**
Proposed Descriptor	8000	**94.48**	13.12

To illustrate the robustness of the descriptor to facial pose variations, we show the features extracted from the lip segment of both frontal and non-frontal faces for happy and surprise emotions in Fig. 7. As we can observe in this figure, the feature histograms are similar across pose changes, but dissimilar for different emotions. For instance, if we compare the Curl feature in the weighted histograms corresponding to the happy emotion (shown in Fig. 7(a)), the negative bin has a higher weight than the positive one for both frontal face and non-frontal faces. This is in contrast to the Curl feature extracted from surprise emotion (shown in Fig. 7(b)), where the positive bin has a higher weight compared to negative bin across both the poses. Similar phenomenon was also observed in the case of other features from different emotions, which indicates that the proposed features indeed have some view invariance properties.

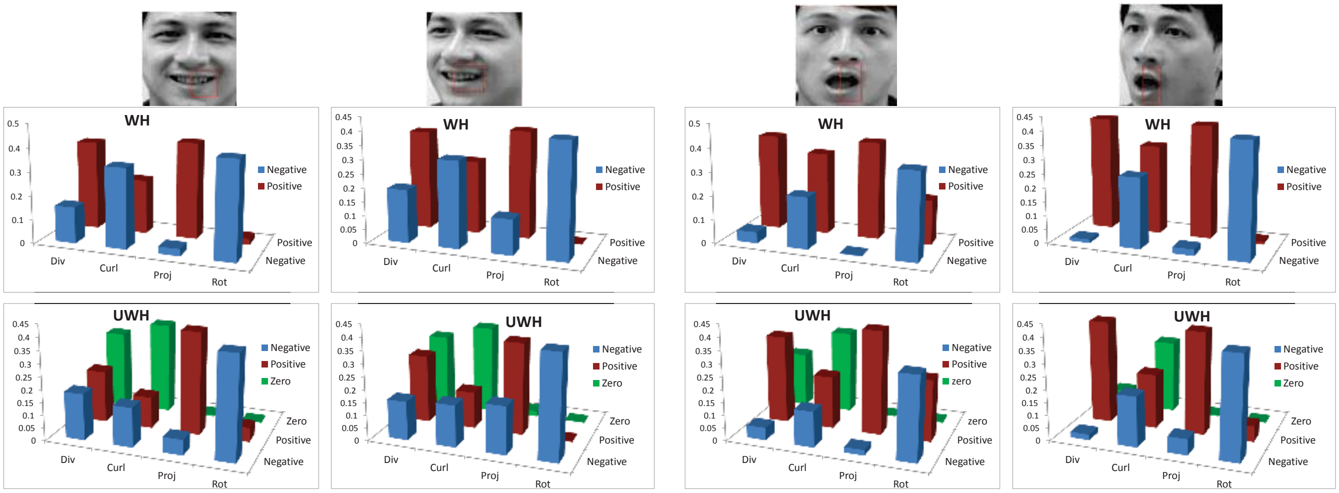
C. Performance of ESL

Table IV compares the performance of the proposed ESL and KESL classifiers to other classifiers in term of recognition accuracy and time complexity of training and testing phases individually. These results are based on 5-fold cross-validation using the proposed spatio-temporal descriptor as the feature vector for all the classifiers. Apart from standard classifiers such as SVM, ELM and KELM, we have also implemented the Sparse ELM (SELM) method, which performs sparse coding and classification individually in the training phase. Note that the learned dictionary and output weights of SELM will be completely different from those obtained through ESL, which simultaneously optimizes the sparse representation and classifier weights as defined in (9).

The results in Table IV show that ESL and KESL classifiers typically have higher recognition rates compared to ELM and SELM classifiers. The KESL algorithm results in the highest classification accuracy among all the methods, but this improvement comes at the cost of increased time complexity.

While a number of researchers have reported the performance of their facial emotion recognition algorithms on the

[4] http://www.ee.oulu.fi/~gyzhao/LBP_Book.htm



(a)

(b)

Fig. 7. Pose-invariant descriptor for happy and surprise emotions. (a) Features extracted from the lip segment from frontal (left) and non-frontal (right) faces with happy emotion. (b) Features extracted from the lip segment from frontal (left) and non-frontal (right) faces with surprise emotion.

TABLE IV
COMPARISON OF THE PROPOSED ESL APPROACH WITH OTHER STANDARD CLASSIFIERS ON ALL DATABASES.

Method	**Recognition rate (%)**					**Time complexity (Sec)**	
	ECK$^+$	**AVEC 2011**		**EmotiW**			
		Val	Test	Val	Test	Train	Test
SVM	90.60	62.61	56.52	27.42	22.75	0.96	0.14
ELM	81.25	59.04	52.48	23.18	19.23	0.26	0.04
SELM	86.19	60.14	54.18	24.64	26.93	110.55	7.42
ESL	90.36	**66.73**	59.21	29.85	27.88	1403.51	5.06
KELM	90.12	60.79	55.19	27.88	23.72	**0.06**	**0.04**
KESL	**92.74**	65.92	**61.82**	**31.34**	**29.81**	1629.20	5.02

TABLE V
COMPARISON OF THE PROPOSED ESL APPROACH WITH OTHER STATE-OF-THE-ART CLASSIFIERS BASED ON SPARSE CODING. THESE RESULTS ARE OBTAINED ON THE ECK$^+$ DATABASE USING THE PROPOSED SPATIO-TEMPORAL DESCRIPTOR AND A COMMON EXPERIMENTAL SETUP.

Method	SRC [2]	DKSVD [19]	LCKSVD [20]	FDDL [21]	ESL	KESL
Recognition rate (%)	89.00	90.11	90.38	91.93	90.36	**92.74**

CK$^+$ benchmark database, these results are not directly comparable to those reported in Table IV. This is due to the large differences in the experimental setup (e.g., pre-processing steps, feature extraction method, number of sequences used for training and evaluation, etc.). Therefore, to obtain a meaningful comparison of the proposed ESL classifier with other state-of-the-art classifiers that involve sparse coding, we have evaluated some of the successful techniques reported in the literature using a common experimental setup. Though the source codes for these methods are publicly available[5], they have not been written on the same platform. Therefore, we do not compare the computational cost of these methods and limit ourselves to comparing them only in terms of classification accuracy. Note that we have optimized the parameters of each method via greedy search. From Table V, we observe that the recognition rates of the proposed method are quite comparable to the state-of-the-art methods.

Tables VI and VII summarize the accuracy of the proposed emotion recognition system (using the novel spatio-temporal descriptor and KESL classifier) to other reported results on the AVEC 2011 and EmotiW databases, respectively. Note that we have included the results of only the vision part and ignore the results that require the audio modality. The results in Tables VI and VII show that the performance of the proposed emotion recognition system is quite comparable to the best results achieved in both these competitions.

TABLE VI
PERFORMANCE COMPARISON ON THE TEST SET OF AVEC 2011 DATABASE.

Method	Baseline [31]	[34]	[35]	[36]	Proposed system
Recognition rate (%)	46.2	61.0	55.9	51.8	**61.8**

Unlike the CK$^+$ database, the AVEC 2011 and EmotiW

TABLE VII
PERFORMANCE COMPARISON ON THE TEST SET OF EMOTIW DATABASE.

Method	Baseline [32]	[37]	[38]	[39]	[40]	[41]	Proposed system
Recognition rate (%)	22.75	24.04	24.68	24.68	**35.58**	29.81	29.81

[5] SRC:http://www.mathworks.com/matlabcentral/fileexchange/35813-sparse-representations-classifier
DKSVD, LCKSVD: http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html
FDDL: http://www4.comp.polyu.edu.hk/~cslzhang/code.htm

databases are representative of real-world applications because they contain samples with natural or spontaneous emotions that do not exhibit sharp facial changes from the start to apex of an expression. This is one of the main reasons for the huge difference between the recognition rates reported in the fourth (AVEC 2011 - Test) and sixth (EmotiW - Test) columns of Table IV compared to those reported in the second column (ECK$^+$) of Table IV. However, the proposed system is able to achieve recognition rates that are comparable to the state-of-the-art performance reported on these two databases. This shows that the proposed emotion recognition system is indeed capable of recognizing natural emotions with subtle changes in facial expression, although there is a scope for significant improvement in the recognition accuracy in such scenarios.

A possible limitation of the proposed emotion recognition system is the need for database-specific tuning of parameters as described in section VI-A. However, it must be emphasized that this phenomenon is not unique to the proposed system, but is common to most pattern recognition systems. In fact, we have performed database-specific parameter tuning via greedy search for all the competing systems reported in Tables III, IV, and V. In the case of AVEC 2011 and EmotiW databases, the primary purpose of having a validation or development subset is to allow tuning of parameters before the model is evaluated on the test set. Thus, the comparison of results in Tables VI and VII is fair, with all the competing approaches being allowed the luxury of parameter tuning.

Another approach to measure the sensitivity to various parameters is to evaluate the generalization performance on unseen test data. While the results in Tables IV-VII demonstrate that the proposed ESL and KESL algorithms have good generalization performance, we observe that the accuracy improvement is not very significant in most cases. For example on the ECK$^+$ database, the accuracy of KESL is only $\approx 2\%$ better than the competing approaches. However, we believe that the advantage of the proposed algorithms can be readily observed if we train the classifier on clean data and test it using samples with large intra-class variations (i.e., the training set is no longer representative of the test set). To investigate this claim, we train the ESL and KESL classifiers on the original CK$^+$ data and then test it using our own collected samples (occlusion, pose variation, and illumination changes). Fig. 8 illustrates the results of this experiment, which clearly demonstrates the advantage of the ESL and KESL methods over other classifiers in terms of generalization performance. The reason for this better generalization performance is that ESL algorithms do not directly use the noisy input samples, but only the sparse coefficients based on a learned dictionary. Thus, ESL algorithms are indeed more robust to noisy and imperfect test data, but this comes at the cost of longer execution times during training and testing.

D. Analysis of Failure Cases

We have analyzed the errors on the EmotiW database and identified three main sources of failure of the proposed emotion recognition system.

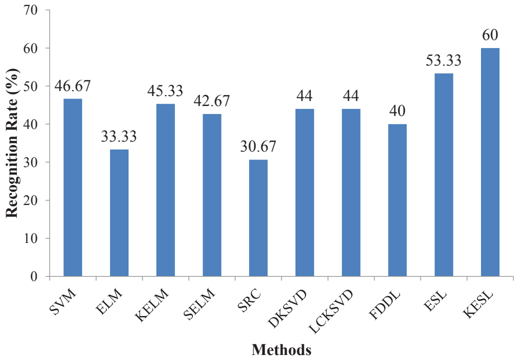
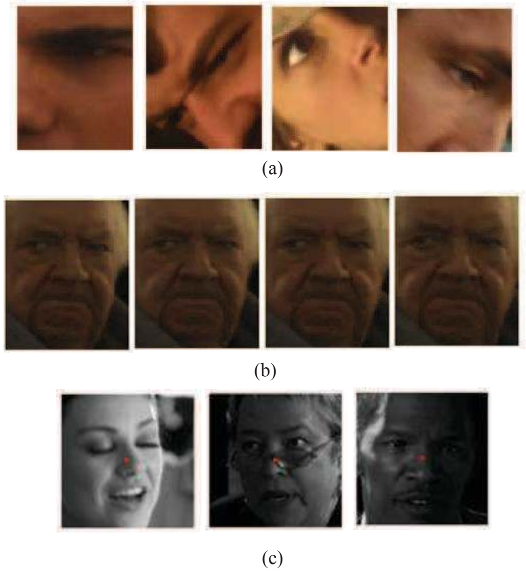


Fig. 8. Comparison of accuracy when the different classifiers are trained on the original CK$^+$ database and applied to data collected by us, which includes occlusion, head posed variations, and illumination changes.



(a)

(b)

(c)

Fig. 9. Sources of failure of the proposed emotion recognition system. (a) Failure of face detector to detect the whole face; (b) failure to detect all faces in an emotion sequence, leading to a scenario where the detected faces do not capture the dynamics of the emotion; (c) wrong reference point localization. The detected nose points are represented as red dots.

1) Failure of face detector and reference point detection:

The first and primary source of error is caused by the failure of the face detector to correctly detect the faces. For example, on the validation set of the EmotiW database, no face was detected for 60 out of the 389 sequences, which accounts for a 15.42% absolute reduction in the final recognition rate.

False detection of non-faces is another source of error. To overcome this problem, we could filter out the non-face samples using methods similar to the one used in [37]. We also identified cases where the face detector detected only a part of face instead of whole face as shown in Fig. 9(a).

Another source of error is the failure of the face detector to detect all faces in a given emotion sequence. This may lead to scenario where there is a lack of faces that depict an emotion from start to apex. For example, Fig. 9(b) shows the faces

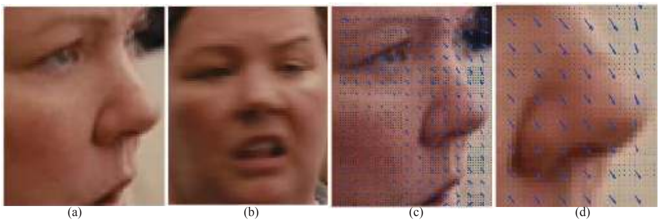


Fig. 10. Failure to estimate the correct optical flow; (a)-(b) two consequent frames; (c) estimated optical flow; (d) magnification of optical flow for nose region indicating errors in OF extraction.

detected in a whole sequence. One can readily observe that all the faces in Fig. 9(b) represent the apex of the emotion and there is no facial movement due to emotion. There are many similar cases in this database. Since the proposed descriptor only encodes the dynamic information (motion), it fails to accurately extract the target features if the emotion is not captured from start to apex. Consequently, many sequences with other emotions will be classified as neutral emotion. Incorporating the static features in addition to the motion features will mitigate this problem.

Errors in nose point localization also affect the feature extraction process. Fig. 9(c) shows a few sample faces with wrong reference point localization.

2) Failure of optical flow: When the range of head pose variation is very large, the OF algorithm will fail to correctly compute the flow field. Fig. 10 illustrates an example of OF failure due to large head movement. As shown in this figure, the head was turned left, but the estimated OF dose not capture this information. We zoomed out the OF around the nose region for better illustration in Fig. 10d.

VII. Conclusions and Future Work

In this paper, we proposed a novel classification scheme called ESL, which is motivated by the recent advancements in the field of sparse representation and supervised dictionary learning. ESL incorporates reconstruction properties of sparse representation and discriminative power of a nonlinear ELM for robust classification. In addition, we proposed a novel OF-based spatio-temporal descriptor for pose invariant facial emotion detection. We have performed extensive experiments on both acted and spontaneous emotion databases to evaluate the effectiveness of the proposed feature extraction and classification schemes under different scenarios.

Our results clearly demonstrate the robustness of the proposed emotion recognition system, especially in challenging scenarios that involve illumination changes, occlusion, and pose variations. The limitations include the higher computational cost for both feature extraction and classification as well as the need to optimize many parameters. Furthermore, there is still a large room for improvement in the recognition accuracy when dealing with natural or spontaneous emotions. Possible ways to improve the proposed emotion recognition framework include: (i) combining the proposed spatio-temporal descriptor with static (appearance) based features to deal with failure in motion feature (e.g., optical flow) extraction, (ii) use of motion exaggeration techniques to improve the recognition accuracy for subtle facial emotions, and (iii) enhancing the OF correction model to remove the effect of facial muscle movement caused due to the person speaking.

Acknowledgment

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore. The authors would like to thank the reviewers for their valuable comments.

References

[1] K. Huang and S. Aviyente, “Sparse Representation for Signal Classification,” in *Adv. NIPS*, 2006.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi, “Robust Face Recognition via Sparse Representation,” *IEEE Trans. PAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[3] D. H. Kim, S. U. Jung, and M. J. Chung, “Extension of Cascaded Simple Feature based Face Detection to Facial Expression Recognition,” *Pattern Recog. Letters*, vol. 29, no. 11, pp. 1621–1631, 2008.

[4] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, “Facial Expression Recognition using Radial Encoding of Local Gabor Features and Cassifier Synthesis,” *Pattern Recog.*, vol. 45, no. 1, pp. 80–91, 2012.

[5] T. Wehrle, S. Kaiser, S. Schmidt, and K. R. Scherer, “Studying the Dynamics of Emotional Expression Using Synthesized Facial Muscle Movements,” *J. Pers. Soc. Psychol.*, vol. 78, no. 1, pp. 105–119, 2000.

[6] P. S. Aleksic and A. K. Katsaggelos, “Automatic Facial Expression Recognition using Facial Animation Parameters and Multistream HMMs,” *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 1, pp. 3–11, 2006.

[7] Y. Zhang and Q. Ji, “Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences,” *IEEE Trans. PAMI*, vol. 27, no. 5, pp. 699–714, 2005.

[8] I. Kotsia and I. Pitas, “Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines,” *IEEE Trans. IP*, vol. 16, no. 1, pp. 172–187, 2007.

[9] G. Zhao and M. Pietikainen, “Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions,” *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[10] T. Wu, S. Fu, and G. Yang, “Survey of the Facial Expression Recognition Research,” *Advances in Brain Inspired Cognitive Systems*, vol. 7366, pp. 392–402, 2012.

[11] O. Rudovic, M. Pantic, and I. Patras, “Coupled Gaussian Processes for Pose-invariant Facial Expression Recognition,” *IEEE Trans. PAMI*, vol. 35, no. 6, pp. 1357–1369, 2013.

[12] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, “Emotion Recognition from Arbitrary View Facial Images,” in *ECCV*, vol. 6316, 2010, pp. 490–503.

[13] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, “Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates,” *Int. J. Computer Vision*, vol. 83, no. 2, pp. 178–194, 2009.

[14] A. Snchez, J. V. Ruiz, A. B. Moreno, A. S. Montemayor, J. Hernndez, and J. J. Pantrigo, “Differential Optical Flow Applied to Automatic Facial Expression Recognition,” *Neurocomputing*, vol. 74, no. 8, pp. 1272–1282, 2011.

[15] R. Niese, A. Al-Hamadi, A. Farag, H. Neumann, and B. Michaelis, “Facial Expression Recognition Based on Geometric and Optical Flow Features in Colour Image Sequences,” *Computer Vision, IET*, vol. 6, pp. 79–89, 2012.

[16] I. Ramirez, P. Sprechmann, and G. Sapiro, “Classification and Clustering via Dictionary Learning with Structured Incoherence and Shared Features,” in *CVPR*, 2010, pp. 3501–3508.

[17] D. Wang and S. Kong, “A classification-oriented dictionary learning model: Explicitly learning the particularity and commonality across categories,” *Pattern Recog.*, vol. 47, no. 2, pp. 885–898, 2014.

[18] J. Mairal, F. Bach, and J. Ponce, “Task-Driven Dictionary Learning,” *IEEE Trans. PAMI*, vol. 34, no. 4, pp. 791–804, 2012.

[19] Q. Zhang and B. Li, “Discriminative K-SVD for Dictionary Learning in Face Recognition,” in *CVPR*, 2010, pp. 2691–2698.

[20] Z. Jiang, Z. Lin, and L. S. Davis, “Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition,” *IEEE Trans. PAMI*, vol. 35, no. 11, pp. 2651–2664, 2013.

[21] M. Yang, L. Zhang, X. Feng, and D. Zhang, “Fisher Discrimination Dictionary Learning for Sparse Representation,” in *ICCV*, 2011, pp. 543–550.

[22] X. Xiong and F. De La Torre, “Supervised Descent Method and Its Applications to Face Alignment,” in *CVPR*, 2013, pp. 532–539.

[23] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High Accuracy Optical Flow Estimation Based on a Theory for Warping,” in *ECCV*, vol. 3024, 2004, pp. 25–36.

[24] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,” *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[25] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme Learning Machine for Regression and Multiclass Classification,” *IEEE Trans. Syst. Man Cybern., Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[26] D. Donoho and X. Huo, “Uncertainty Principles and Ideal Atomic Decomposition,” *IEEE Trans. on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[27] J. Tropp, A. C. Gilbert, and M. Strauss, “Algorithms for Simultaneous Sparse Approximation. Part I: Greedy Pursuit,” *IEEE Trans. Signal Processing*, vol. 86, pp. 572–588, 2006.

[28] T. Blumensath and M. E. Davies, “Iterative Hard Thresholding for Compressed Sensing,” *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[29] Z. Xie and S. Chen, “SCIHTBB: Sparsity Constrained Iterative Hard Thresholding with BarzilaiBorwein Step Size,” *Neurocomputing*, vol. 74, no. 17, pp. 3663–3676, 2011.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-specified Expression,” in *CVPR*, 2010, pp. 94–101.

[31] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “AVEC 2011 The First International Audio/Visual Emotion Challenge,” in *Affective Computing and Intelligent Interaction*, vol. 6975, 2011, pp. 415–424.

[32] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, “Emotion Recognition In The Wild Challenge 2013,” in *ICMI*, 2013.

[33] J. Bihan, M. F. Valstar, and M. Pantic, “Action Unit Detection using Sparse Appearance Descriptors in Space-Time Video Volumes,” in *IEEE Conf. Face and Gesture Recog.*, 2011, pp. 314–321.

[34] G. Ramirez, T. Baltruaitis, and L.-P. Morency, “Modeling Latent Discriminative Dynamic of Multi-dimensional Affective Signals,” in *Affective Computing and Intelligent Interaction*, vol. 6975, 2011, pp. 396–406.

[35] A. Cruz, B. Bhanu, and S. Yang, “A Psychologically-Inspired Match-Score Fusion Model for Video-Based Facial Expression Recognition,” in *Affective Computing and Intelligent Interaction*, vol. 6975, 2011, pp. 341–350.

[36] M. Glodek et al., “Multiple Classifier Systems for the Classification of Audio-Visual Emotional States,” in *Affective Computing and Intelligent Interaction*, vol. 6975, 2011, pp. 359–368.

[37] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, “Partial Least Squares Regression on Grassmannian Manifold for Emotion Recognition,” in *Int. Conf. Multimodal Interaction*, 2013, pp. 525–530.

[38] M. Day, “Emotion Recognition with Boosted Tree Classifiers,” in *Int. Conf. Multimodal Interaction*, 2013, pp. 531–534.

[39] T. R. Almaev, A. Yuce, A. Ghitulescu, and M. F. Valstar, “Distribution-based Iterative Pairwise Classification of Emotions in the Wild using LGBP-TOP,” in *Int. Conf. Multimodal Interaction*, 2013, pp. 535–542.

[40] S. E. Kahou et al., “Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video,” in *Int. Conf. Multimodal Interaction*, 2013, pp. 543–550.

[41] T. Gehrig and H. K. Ekenel, “Why is Facial Expression Analysis in the Wild Challenging?” in *Int. Conf. Multimodal Interaction*, 2013, pp. 9–16.

Seyedehsamaneh Shojaeilangari received her B.S. and M.S. in the biomedical engineering from Amirkabir University of Technology, Iran in 2006 and 2009 respectively. She is currently a Ph.D candidate at the School of Electrical & Electronic Engineering (EEE) at Nanyang Technological University (NTU), Singapore. Her research interests include computer vision, pattern recognition, image processing, and video processing.

Wei-Yun Yau received his BEng (Electrical) from the National University of Singapore (1992), MEng degree (1995) and PhD degree (1999) from NTU, Singapore. From 1997 to 2002, he was with the Centre for Signal Processing, Singapore leading the R&D effort in biometrics. Currently, he is a Programme Manager with the Institute for Infocomm Research, leading the research in the Robotics programme. His research interest includes biometrics, active vision system, personalized media and interactive TV.

Karthik Nandakumar received his B.E. degree (2002) from Anna University, India, M.S. degrees in Computer Science (2005) and Statistics (2007), and Ph.D. degree in Computer Science (2008) from Michigan State University. From 2008 to 2014, he was a Scientist at the Institute for Infocomm Research. Currently, he is a Research Staff Member at the IBM Research, Singapore. His research interests include pattern recognition, biometric authentication, image processing, and computer vision.

Jun Li received his B.S. in mechanical and electrical engineering and M. Eng. in the biometrics processing from the University of Science and Technology of China, Hefei, China, in 1997 and 2002, respectively. In 2007, he obtained his Ph.D. degree in the School of EEE, NTU, Singapore. He is currently with Institute for Infocomm Research. His research interests include biometrics, object detection and facial analytics.

Eam Khwang Teoh received his BE and ME degrees in Electrical Engineering from the University of Auckland, New Zealand in 1980 and 1982, respectively and the PhD degree in Electrical & Computer Engineering from the University of Newcastle, Australia in 1986. Since 1985, he has been with the School of EEE, Nanyang Technological University, Singapore as a Lecturer, Senior Lecturer, and currently an Associate Professor. His research interests include computer vision, pattern recognition, biometric recognition, and medical image processing.