Person-Independent Facial Expression Recognition Using Low-Rank Expression and Sparse Expression Residual

Songfan Yang, *Member, IEEE,* Le An, *Member, IEEE,* Yinjie Lei, *Member, IEEE,* and Bir Bhanu, *Fellow, IEEE*

***Abstract*—Person-independent expression recognition aims to recognize facial expressions of an *unseen* subject. This remains a challenging problem due to that the facial muscle motion for a typical facial expression, which should be similar irrespective of the subject's identity, is difficult to quantify and generalize across population. Therefore, trained classifiers can easily over-fitted to person-specific facial appearances rather than to person-independent muscle motion from expressions. This problem becomes even more challenging when the rigid head motion and non-rigid facial muscle motion are observed in uncontrolled videos. To tackle these challenges, we model an facial expression video as a combination of two parts: 1) a set of low-rank components that represent person-independent expression prototypes, termed as Low-Rank Expression (LRE), and 2) a set of sparse appearance residuals that capture the muscle motion of individual frames in video, namely Sparse Expression Residual (SER). To avoid over-fitting, we explicitly model the person-independent transformation of individual frames with respect to a canonical reference space, in terms of an objective function to be minimized. We show that this objective function can be simplified and solved by an efficient sparse low-rank recovery algorithm. A comparative study between the state-of-the-art models and the proposed LRE representation shows the merits of our approach in representing expressions in a person-independent manner. We also demonstrate that the SER can also be utilized to further improve the recognition performance, when both spatial and temporal texture features are extracted out of SER. Extensive qualitative and quantitative results demonstrate the superiority of our approach as compared to most recent methods.**

***Index Terms*—Facial expression recognition, person-independent, low-rank expression, sparse expression residual**

I. Introduction

AFFECTIVE computing has been attracting an increasing amount of attentions in both psychology and computer science studies [1]. As a primary channel to convey emotion, facial expression contains rich information for communication [2], [3]. In particular, the automatic analysis of facial expressions plays an important role in human-computer interaction and human-centered interface design. Exploring the affective property of these multimedia data source enables applications such as large-scale user behavior analysis [4], patient rehabilitation [5] in recent years.

S. Yang and Y. Lei are with the College of Electronics and Information Engineering at Sichuan University, Chengdu, China 610064. E-mail: {syang, yinjie}@scu.edu.cn

L. An is the Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521 USA. E-mail: lan004@ucr.edu

B. Bhanu is with the Center for Research in Intelligent Systems, University of California, Riverside, CA 92521 USA. E-mail: bhanu@cris.ucr.edu

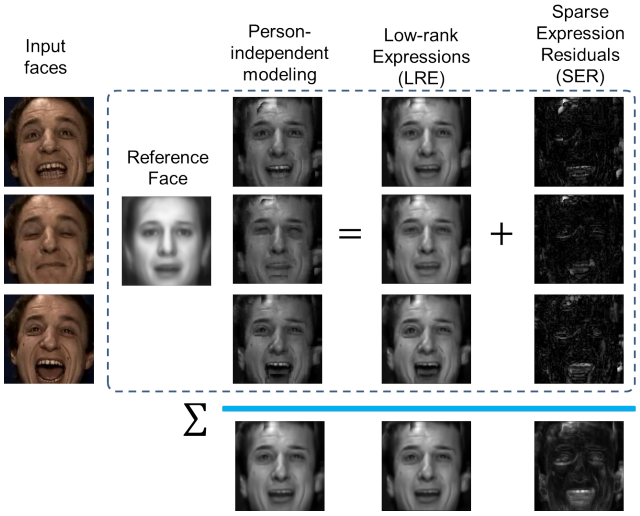


Fig. 1. The overview of our approach. We observe that: 1) each frame of a facial expression sequence shares an *global* model (*e.g.*, lip corner expansion for *Happiness* in this case); and 2) each frame possesses its own *local* properties (*e.g.*, various muscle motion intensities), which describe the facial dynamics. Due to the face anatomy constraints [7], the muscle motion that characterize the local deformation is sparsely distributed. Inspired by these observations, we model each frame as: 1) a Low-Rank Expression (LRE) that represents the global expression appearance estimated from the correlated appearances across the entire sequence; and 2) an additive Sparse Expression Residual (SER) that represents the deviations from the low-rank structure. Moreover, the person-independent factor is modeled as a transformation with respect to a canonical space (reference face), avoiding over-fitting for classification. Both LRE and SER provide viable manners to extract meaningful facial expression dynamics, yielding superior recognition performances. To appreciate the information captured by the LRE and SER representations, the bottom row shows the mean representation of the decomposed sequence. The mean LRE retains the global appearance and the mean SER captures the local motion around mouth and eye regions.

Although a rich body of research on facial expression recognition in controlled environments was conducted [3], it is still challenging to maintain the same level of cross-dataset performance in more realistic scenarios. A recent expression analysis challenge [6] indicates that person-independent expression recognition in video remains a difficult problem in uncontrolled settings, due to the following reasons:

1) **Compound motion.** In real-world scenarios, the facial expression appearance in video consists of rigid head motion and non-rigid muscle motion. The rigid head motion includes both in-plane and out-of-plane rotations. Aligning 2D faces by rectifying the out-of-plane rotation is considerably more difficult than dealing with in-plane

rotation due to the complexity in 3D human face. Most approaches in the literature only recover the in-plane rotation by computing an affine transformation based on facial feature correspondences [8], [9]. Furthermore, segmenting the non-rigid muscle motion from the rigid head motion, which reveals the true facial expression, is an inherently challenging problem from 2D images.

2) **Person-specific effect.** Psychological study shows that there are six basic prototype expressions that have similar properties across various cultures, namely, *Happiness*, *Sadness*, *Surprise*, *Fear*, *Anger*, and *Disgust* [10]. Many vision-based expression analyses rely on these basic prototypes [2]. An "ideal" expression recognition system is expected to eliminate cultural-independent and even person-independent information. However, typical facial expression recognition systems use generic appearance or geometry-based features to characterize an expression, without taking into account the person-specific property [8], [11]. Thus, person-specific information is retained and the classifier may fail to recognize person-independent expressions. This results in an over-fitted classification model that ultimately degrades the recognition accuracy on unseen testing subjects. There has been some pioneer work [12], [13] that attempts to alleviate the person-specific effect, and the improved results have verified that person-independent modeling is a key issue in obtaining robust expression recognition systems.

In this paper, the problem we intend to solve is as follows: given a video sequence of a facial expression after temporal segmentation, the goal is to classify each segment of this video as one of the six categorical expression, *e.g.*, *Happiness* or *Anger*.

Previous approaches for this task can be categorized into three strategies for video-based expression recognition. In the first category, a wide range of methods [8], [11] process each frame individually and a majority voting rule is applied to determine the expression type for the entire sequence. In the second category, temporal features are used to capture the muscle dynamics of facial expressions [14], [15]. These approaches typically require accurate facial alignment in order to generate meaningful dynamic features. In the third category, an entire video sequence is summarized into an image-based representation that captures the expression appearance [12], [13]. These approaches completely abandon the dynamic information, and cannot model the temporal characteristic of facial expressions. In essence, the aforementioned strategies summarize the video-based expression data at different resolution.

In this work, we depart from all the previous strategies and analyze the facial expression at a finer resolution. We observe that, on one hand, expressions in individual frames from one video sequence are correlated and possess similar *global* characteristics (*e.g.*, lip corner expansion for *Happiness*). Thus, the correlated appearances should form a low rank structure shared by all frames. On the other hand, each sample has its own *local* properties (*e.g.*, various muscle motion intensities for different expression stages such as *onset*, *apex*, or *offset*), which could be used to describe facial dynamics. Due to the face anatomy constraints [7], the muscle motion that characterizes the local deformation is sparsely distributed.

Inspired by these observations, we model each frame as 1) a Low-Rank Expression (LRE) that represents the global expression appearance from the correlated frames; and 2) an additive sparse expression residual (SER) that represents the deviations from the low-rank structure. As shown in Fig. 1, both static and dynamic features are extracted from LRE and SER representations for expression recognition. In addition, we explicitly model the person-independent facial expression appearance transformation in a low-rank recovery framework. The person-independent factor is modeled as a non-linear transformation function, which can be approximated by matching a target face to a canonical reference face. Concretely, we approximate this non-linearity by the correspondence of two SIFT feature descriptors, namely SIFT-flow [16]. We then show that the person-independent transformation is independent of the low-rank recovery procedure, leading to an efficient algorithm in the pursuit of the sparse low-rank facial expression appearances.

In what follows, Section II reviews the related work in the literature and highlights our contributions. Subsequently, Section III details the modeling of expression images as LRE and SER representations, and contrast them with previous approaches in theory and practice. Section IV demonstrates the effectiveness of LRE+SER modeling via extensive experimental results. Finally, Section V concludes this paper.

II. Related Work and Our Contributions

We first review the related works in the literature for both facial expression recognition and sparse representation, and highlight our contributions thereafter.

A. Facial Expression Recognition

Automatic facial expression analysis has been studied extensively. Earlier work mainly focuses on recognizing expressions from *static* images, where both facial geometry and textures are used for this task [2], [17], [18]. Recent work in expression analysis deals with videos [19], [8], [20], [9]. These systems treat each frame individually and attempt to recognize facial expression or Action Units (AUs) [7]. For feature representations in facial expression analysis, the texture features can be described using Haar-like features [21], Gabor filters [22], [8], Local Binary Patterns (LBP) [23], [24], Edge Orientation Histogram [25], Histogram of Oriented Gradients (HOG) [26], [27], [28], and Local Phase Quantization (LPQ) [29], etc. Moreover, different modalities can be used to improve the recognition performance, e.g., profile-view [30], infrared images [31], depth signal [32], audio-visual data [32], [33], [34], etc.

As pointed out in [35] and [36], facial muscle *dynamics* play an important role in spontaneous expressions. The pioneer work by Yacoob and Davis [37] describes the facial dynamics using optical flow. It is assumed that subjects are with controlled frontal head pose, and thus, a simple thresholding procedure can rule out small motion cause by

noise or tracking errors. This assumption immediately raises concern since optical flow highly relies on accurate face tracking and alignment. Moreover, this work ignores the facial texture information, which is important in characterizing facial dynamics. For example, Zhao and Pietikäinen [14] extend LBP to the temporal space, namely three orthogonal planes (LBP-TOP), by taking into account the co-occurrence of the patterns. It combines appearance and motion together and achieves outstanding performance for facial expression recognition under controlled settings, *e.g.*, the Cohn-Kanade database [38]. Similarly, various texture descriptors, *e.g.*, LPQ-TOP [39], can be extended to spatial-temporal settings, yielding a superior performance. The muscle dynamics can also be characterized by the motion of facial landmarks (*e.g.*, eyebrow corner, mouth corner, etc.). Valstar and Pantic [11] track a set of landmarks over time and infer facial expressions based on the dynamics of landmark.

Facial dynamics are easier to capture in controlled settings [38], where subjects are constraint to have limited head motion and exaggerated facial muscle motion. However, in a more realistic and uncontrolled scenario (*e.g.*, FERA dataset [9]), extracting facial dynamics highly relies on accurate facial landmark tracking and face alignment. The facial expression in real world consists of the "compound motion" from the rigid head motion and deformable facial muscle motion. Ideally, the head motion should be corrected to fully characterize the non-rigid muscle motion, which is directly related to the expressions. It is shown in [40], [12] that without a reliable alignment, using the dynamic feature is even inferior than simply using static features. Therefore, in this work, we establish a framework that aligns expression images under compound motion from region-based correspondences (in contrast to point-based correspondences), which helps to extract much more effective dynamic features.

B. Sparse Representation

Recently, sparse modeling has witnessed its success in diverse applications such as image restoration [41], face recognition [42], and human action recognition [43]. In particular, sparse representation combined with low-rank modeling can be used to extract robust features directly from an image in a matrix form, such as the transform invariant low-rank textures [44], group alignment by sparse and low-rank decomposition [45]. In the facial expression related domain, the method proposed in [46] jointly recovers a dictionary and a set of sparse coefficients to efficiently synthesize 3-D facial expressions. In [47] the mid-level features are learned by sparse coding technique for multi-view expression recognition. It uses sparse codes of local descriptors (such as SIFT) to build features for expression recognition. Different from these approaches, in this work we focus on recovering 2D representations that are of low-rank property for a better performance in facial expression recognition.

C. Our Contributions

The contributions of this work are the follows:

1) We observe that each frame of a prototype expression sequence is correlated and composed of a global and a local component. We then propose to model the generic person-independent facial expressions video in a sparse low-rank recovery framework. The sequence is decomposed into two components, termed Low-Rank Expression (LRE) and Sparse Expression Residual (SER), which jointly help to extract more effective dynamic facial features, leading to a superior recognition performance.
2) We show that a special case of person-independent transformation, via SIFT-flow, results in a low-rank expression model, which visually resembles the Emotion Avatar Image (EAI) representations [12]. As shown by the experimental results, this transformation can well handle unseen data during testing.

III. Person-independent Expression Modeling

A. The Generalized Model

We assume that a video is temporally segmented such that each segment can be described by one prototype expression or an AU. We then seek to decompose each frame of the sequence into two representations, LRE and SER, to capture the global appearance of the segment and non-rigid muscle motion for each frame, respectively. Formally, given a sequence of video segment, we denote $\mathbf{I}_1, \ldots, \mathbf{I}_n$ as n image observations of the sequence. Then each frame can be decomposed as

$$\mathbf{I}_i = \mathbf{I}_i^0 + \mathbf{R}_i, \quad i = 1 \cdots n, \tag{1}$$

where sparse residual is modeled by the additive term $\mathbf{R}_i$. $\mathbf{I}_i^0$ is the difference between $\mathbf{I}_i$ and $\mathbf{R}_i$. Furthermore, we denote $\mathbf{v} \in \mathbb{R}^m$ to be the vectorized observation of $\mathbf{I}_i$ and $\mathbf{e}_i$ to be the vectorized version of $\mathbf{R}_i$. Stacking multiple such vectors results in an observation matrix

$$\mathbf{D} = [\mathbf{v}_1, \cdots, \mathbf{v}_n] = \mathbf{A} + \mathbf{E}, \tag{2}$$

where $D \in \mathbb{R}^{m \times n}$, and $\mathbf{A} = [\mathbf{v}_1^0, \cdots, v_n^0]$ is a low-rank matrix that models the dominant structure of the facial expression sequence. $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_n]$ is a sparse matrix with large non-zero elements representing the motion of the sequence. Thus, the problem that we try to solve is

$$\min_{\mathbf{A},\mathbf{E}} rank(\mathbf{A}) + \gamma \|\mathbf{E}\|_0, \quad \text{s.t.} \quad \mathbf{D} = \mathbf{A} + \mathbf{E}, \tag{3}$$

where $\|\cdot\|_0$ is the ℓ_0 norm counting the number of non-zero entries in the error term $\mathbf{E}$. γ is a scaling parameter which controls a weight preference of the rank of $\mathbf{A}$ and sparsity of $\mathbf{E}$. Although (3) is an intuitive formulation, it is a non-convex and NP-hard problem which is difficult to solve. As suggested by many works in sparse low-rank recovery [48], [49], [50], (3) can be relaxed by its convex surrogate as

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \gamma \|\mathbf{E}\|_1, \quad \text{s.t.} \quad \mathbf{D} = \mathbf{A} + \mathbf{E}, \tag{4}$$

where $\|\cdot\|_*$ is the nuclear norm that computes the sum of all singular values of $\mathbf{A}$, and $\|\mathbf{E}\|_1 = \sum_{ij} |\mathbf{E}_{ij}|$. This new

objective is convex and can be solve by an efficient first-order accelerated proximal gradient algorithm [49], [51].

B. Reference-based Person-Independent Transformation

Since a face in each frame $\mathbf{I}_i$ contains pose and identity variations, we intend to carry out the expression recognition task in a canonical space. Specifically, we model the person-independent factor by mapping each face to a reference face model, $\mathbf{I}_r$. This strategy has been demonstrated effective in expression recognition, particularly in predicting expressions for unseen subjects [12], [13]. We model the mapping by a transformation function applied to each frame of the entire sequence. Thus, the objective is written as

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \gamma\|\mathbf{E}\|_1 \quad \text{s.t.} \quad \phi(\mathbf{D}) = \mathbf{A} + \mathbf{E}. \tag{5}$$

Here, $\phi(\mathbf{D})$ is defined as a generic function that transforms each face $\mathbf{I}_i$ in $\mathbf{D}$ to the canonical space. One instantiation of the non-linear mapping, which we employ in this work, is SIFT-flow that performs structural matching [16].

SIFT-flow was originally designed to align an image to its plausible nearest neighbor which can have large variations. It robustly matches dense structural SIFT features between two images, while maintaining spatial discontinuities. The SIFT [52] features are first extracted in a pixel-wise fashion. For every pixel in an image, the neighborhood (*e.g.*, 16×16) is divided into a 4×4 cell array. The orientation of each cell is quantized into 8 bins, generating a $4 \times 4 \times 8 = 128$-dimensional vector as the SIFT representation for a pixel. Subsequently, a dense correspondence is established to match the two SIFT images. Similar to optical flow, the objective energy function is designed as

$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min(\|\mathbf{s}_1(\mathbf{p}) - \mathbf{s}_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|_1, t) + \tag{6}$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) + \tag{7}$$

$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} \min(\alpha\,|u(\mathbf{p}) - u(\mathbf{q})|, d) + \min(\alpha\,|v(\mathbf{p}) - v(\mathbf{q})|, d), \tag{8}$$

where $\mathbf{p} = (x, y)$ is the grid coordinates of the images, and $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector at $\mathbf{p}$. $u(\mathbf{p}), v(\mathbf{p})$ are the flow vectors for x direction and y direction, respectively. $\mathbf{s}_1$ and $\mathbf{s}_2$ are two SIFT images to be matched. ε contains all the spatial neighbors (a four-neighbor system is used in this paper). The *data term* in (6) is a SIFT descriptor match constraint that enforces the match along the flow vector $\mathbf{w}(\mathbf{p})$. The *small displacement constraint* in (7) allows the flow vector to be as small as possible when no other information is available. The *smoothness constraint* in (8) takes care of the similarity of flow vectors for adjacent pixels. In this objective function, the truncated ℓ_1 norm is used in both the data term and the smoothness term with t and d as the thresholds for matching outliers and flow discontinuities, respectively. η and α are weight parameters for the small displacement and smoothness constraint, respectively. The dual-layer loopy belief propagation is used in additional to the coarse-to-fine flow matching scheme to optimize the objective function efficiently.

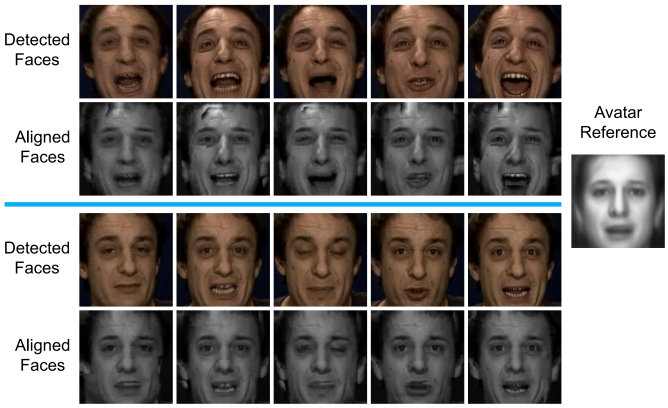


Fig. 2. Reference-based alignment via SIFT-flow warping [16]. The *Avatar Reference* is a canonical face representation [12]. After warping, the head rotation is compensated and person-specific information is attenuated in both sequences.

Fig. 2 shows SIFT-flow warping of faces with respect to the Avatar Reference (AR). Face in each frame is extracted using standard Viola-Jones face detector [53]. The reference face image is chosen to be the level-1 AR image [12] generated from the FERA-GEMEP dataset [9]. AR is essentially a face model that reflects the expression and identity of the entire population in the dataset. It is computed by an iterative algorithm that estimates the reference model and the individual expression model simultaneously. It is computed offline and has been demonstrated to perform well across different datasets [12].

Since we transform each frame to the same reference model, facial features for the entire sequence are coarsely aligned. It is also observed from Fig. 2 that person identity information is attenuated as the appearance of both sequences resembles the AR after warping. In addition, head rotation (especially out-of-plane rotation) is alleviated via this non-linear transformation.

We write $\mathbf{f}_s^i$ (shorthanded for $\mathbf{f}_s(\mathbf{I}_i, \mathbf{I}_r)$) as the SIFT-flow field given by matching a face $\mathbf{I}_i$ to the reference face $\mathbf{I}_r$. Stacking multiple flow vectors in the entire sequence results in $\mathbf{S} = [\mathbf{f}_s^1, \cdots, \mathbf{f}_s^n]$. Thus, our optimization problem can be write as

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \gamma\|\mathbf{E}\|_1, \quad \text{s.t.} \quad \mathbf{D}^* = \mathbf{A} + \mathbf{E}, \tag{9}$$

where $\mathbf{D}^* = \mathbf{D} + \mathbf{S}$. Now, with this linear constraint, our problem can be solved by an efficient Accelerated Proximal Gradient (APG) algorithm [51]. The APG relaxes (9) to

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \gamma\|\mathbf{E}\|_1 + \frac{1}{\mu}\mathbf{f}(\mathbf{A}, \mathbf{E}), \tag{10}$$

where $\mathbf{f}(\mathbf{A}, \mathbf{E}) = \frac{1}{2}\|\mathbf{D}^* - \mathbf{A} - \mathbf{E}\|_F^2$ penalizes the objective function based on the equality constraint, $\mu > 0$ is a relaxation parameter such that any solution to (10) approaches the solution set of (9) when μ approaches 0. The AGP algorithm then iteratively forms separable quadratic approximation to

$\mathbf{f}(\mathbf{A}, \mathbf{E})$ at a special sequence of points $\mathbf{Y}^k = (\mathbf{Y}^k_{\mathbf{A}}, \mathbf{Y}^k_{\mathbf{E}})$ for fast convergence [51]. Hence, $\mathbf{A}^{k+1}$ and $\mathbf{E}^{k+1}$ at the next iteration is obtained by

$$\min_{\mathbf{A},\mathbf{E}} \|\mathbf{A}\|_* + \gamma\|\mathbf{E}\|_1 + \frac{\rho}{2\mu}\|(\mathbf{A}, \mathbf{E}) - (\mathbf{G}^k_{\mathbf{A}}, \mathbf{G}^k_{\mathbf{E}})\|^2_F \quad (11)$$

where $(\mathbf{G}^k_{\mathbf{A}}, \mathbf{G}^k_{\mathbf{E}}) = \mathbf{Y}^k - \rho^{-1}\nabla\mathbf{f}|_{\mathbf{Y}^k}$; ρ is set based on the *Lipschitz* constant. $\mathbf{A}^{k+1}$ and $\mathbf{E}^{k+1}$ are computed by soft-thresholding the singular values of $\mathbf{G}^k_{\mathbf{A}}$ and the entries of $\mathbf{G}^k_{\mathbf{E}}$, respectively, and the soft-thresholding function is defined as

$$\mathbb{T}_\xi(x) = \begin{cases} sign(x)(|x| - \xi), & |x| > \xi \\ 0, & |x| > \xi \end{cases}. \quad (12)$$

The aforementioned procedures are summarized in Algorithm 1. Parameters are empirically chosen in our experiments based on the analysis in [49], *i.e.*, $\delta \leq 10^{-5}$, $\eta = 0.9$, $\mu_0 = 0.99\|\mathbf{D}\|_2$, where $\|\cdot\|_2$ is the spectral norm of a matrix. We refer interested readers to [49] for more details on the parameter selection.

Algorithm 1 Low-Rank Expression Decomposition

Input: Expression data matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$, SIFT-flow warping matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$

$\mathbf{D}^* = \mathbf{D} + \mathbf{S}$
$\mathbf{A}_0, \mathbf{A}_{-1} \leftarrow \mathbf{0}; \mathbf{E}_0, \mathbf{E}_{-1} \leftarrow \mathbf{0}; t_0, t_{-1} \leftarrow 1; \bar{\mu} \leftarrow \delta\mu_0$.
while not converged **do**
  step 1: compute proximal points:
  $\mathbf{Y}^k_{\mathbf{A}} \leftarrow \mathbf{A}^k + \frac{t_{k-1}-1}{t_k}(\mathbf{A}_k - \mathbf{A}_{k-1})$;
  $\mathbf{Y}^k_{\mathbf{E}} \leftarrow \mathbf{E}^k + \frac{t_{k-1}-1}{t_k}(\mathbf{E}_k - \mathbf{E}_{k-1})$;
  step 2: compute gradient:
  $\mathbf{G}^k_{\mathbf{A}} \leftarrow \mathbf{Y}^k_{\mathbf{A}} - \frac{1}{2}(\mathbf{Y}^k_{\mathbf{A}} + \mathbf{Y}^k_{\mathbf{E}} - \mathbf{D}^*)$;
  $\mathbf{G}^k_{\mathbf{E}} \leftarrow \mathbf{Y}^k_{\mathbf{E}} - \frac{1}{2}(\mathbf{Y}^k_{\mathbf{A}} + \mathbf{Y}^k_{\mathbf{E}} - \mathbf{D}^*)$;
  step 3: soft-thresholding:
  $(\mathbf{U}, \mathbf{S}, \mathbf{V}) \leftarrow$SVD$(\mathbf{G}^k_{\mathbf{A}})$
  $\mathbf{A}_{k+1} \leftarrow \mathbf{U}\mathbb{T}_{\mu^k/\rho}(\mathbf{S})\mathbf{V}^\top$;
  $\mathbf{E}_{k+1} \leftarrow \mathbb{T}_{\gamma\mu^k/\rho}(\mathbf{G}^k_{\mathbf{E}})$;
  step 4: update:
  $t_{k+1} \leftarrow \frac{1+\sqrt{4t_k^2+1}}{2}$;
  $\mu_{k+1} \leftarrow \max(\eta\mu_k, \bar{\mu})$;
  $k \leftarrow k + 1$;
end while
Output: $\mathbf{A} \leftarrow \mathbf{A}_k$, $\mathbf{E} \leftarrow \mathbf{E}_k$.

C. Feature Extraction and Classification

Now that we have decomposed each frame of an expression sequence into two separate modules, we extract various texture feature on both modules in both static and dynamic fashions.

Static Features: to characterize the entire sequence using static feature, there are two strategies in general. First, one can extract features from individual frames and train a frame-based classifier. A final decision can be fused at the decision level by majority voting. Second, the feature extraction and classification can be conducted on the mean representation of the entire sequence. We have tested both strategies and observed superior performance and faster processing speed for the latter strategy in most cases. Besides, this strategy is also used in [12]. Thus, we only report the results on the mean representation for the static feature extraction in this paper. The choices of feature descriptors include:

1) LBP uses the local contrast statistics to characterize the image texture. In this work, we use the uniform LBP operator [54] that generates 59 basic patterns for a region of interest.
2) Local Phase Quantization (LPQ): proposed by Ojansivu *et al.* in [29], LPQ descriptor is insensitive to image blur. The spatial blurring is modeled by the multiplication of the original image and a point spread function (PSF) in frequency domain. The key observation is that, when the PSF is centrally symmetric, the phase of the original image is invariant.

Dynamic Features: the LBP and LPQ can be extended to the temporal domain, capturing the dynamics of the texture. We consider two variants, namely TOP-LBP [14] and TOP-LPQ [39] for dynamic feature extraction. In essence, the temporal dimension is included in addition to the original 2D image based descriptor, *i.e.*, the XY 2-D plane is extended to three orthogonal directions, namely XY, XT, and YT, where "T" denotes the temporal axis. The pattern co-occurrences of each plane are computed by their corresponding histograms and all histograms are then stacked to a single feature vector. This enlarges the feature dimension by three times as compared to the static feature.

In the training process, a Support Vector Machine (SVM) [55] with linear kernel is used to train a multi-class classifier in one-vs-all fashion. During the test phase, the same steps for decomposition and feature extraction are taken. The final prediction is made based on the prediction of the trained SVM classifier.

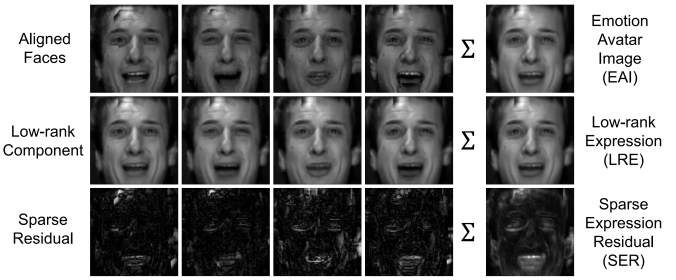


Fig. 3. The LRE and SER decomposition. A low-rank component and a sparse residual component are jointly estimated and decomposed from a sequence of aligned faces. Although individual frame differs from each other, the mean LRE visually resembles the EAI representation [12].

D. Relation to Emotion Avatar Image (EAI)

EAI was proposed in [12] for the international competition on FERA Challenge [6]. It is a robust image representation that summarizes facial expression videos and it generalizes well on person-independent expression recognition. The EAIs are

iteratively generated along with the AR representation, which is canonical representation that captures the appearance of the entire dataset. The generalization of AR is also demonstrated in [12] such that the recognition performance is not degraded when AR is computed using outside datasets. As shown from row 1 of Fig. 3, the EAI is computed from the mean of the aligned faces with respect to a certain level of AR (Level-1 in this case). It can be considered as a Maximum Likelihood (ML) estimate of the aligned sequences.

On the other hand, the low-rank component and sparse residual of individual frame is recovered by Algorithm 1. As seen from row 2 of Fig. 3, the low-rank components are jointly estimated from all frames of a sequence. Although individual frame appears to be different, its corresponding low-rank component resembles the ML estimate of the aligned faces, *i.e.*, EAI (last column of Fig. 3). More sample visualizations are shown in Fig. 4 for various sequences.

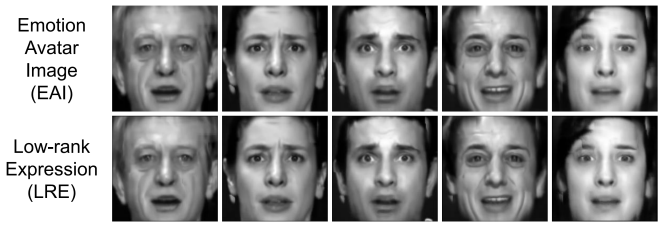


Fig. 4. The comparison of Low-Rank Expressions (LRE) and Emotion Avatar Images (EAI). Despite of being generated from different approaches, their appearances resemble each other at pixel level.

As mentioned earlier, the expression appearance captured by the LRE reveals the underlining expression. For example, the expression *Happiness* can be inferred from the LRE in Fig. 3 based on its lip corner expansion and cheek raise. However, LRE discards the expression dynamics of every individual. Fortunately, the sparse residual recovers the deviation of an aligned face from an LRE frame. Sparse residual models the subtle muscle motion of each particular frame from the underlining expression.

IV. Experimental Results

We conduct experiments on the *uncontrolled* FERA-GEMEP dataset [56]. Qualitative results are first provided to visually demonstrate the validity of our method. The quantitative results then show that our method not only outperform the state-of-the-art approaches, but also has the potential for improvement with additional data. We also provide results on the *controlled* CK+ dataset [38], for the completeness of the experiment.

A. Experimental Setup

After faces are detected from the sequences, we follow the standard image pre-processing steps in competing methods such as [40], [12] for a better and fair comparison. The detected face images are resized to 200×200 pixels, and then divided into non-overlapping blocks of size 20×20.

TABLE I
The rank of person-independent test in FERA Challenge [40]

Teams 1	Accuracy
UCR	0.75
UCSD-CERT	0.71
UIUC-UMC	0.66
KIT	0.66
ANU	0.65
NUS	0.64
QUT-CMU	0.62
UCL	0.61
UMont	0.58
MIT-Cambridge	0.45
Baseline	0.44

Texture features are extracted at each local block and then concatenated to form a final feature vector for classification. The feature extraction is conducted on both LRE and SER modalities, respectively. To prevent over-fitting, Principal Component Analysis (PCA) is carried out and the space with 99% of the variation is retained for dimension reduction.

For static features, the settings are the following:

1) LBP: the uniform LBP operator [54] generates 59 basic patterns for a local patch. Given our region segmentation schema, the total feature dimension is $59 \times 10 \times 10 = 5900$.
2) LPQ: we use the implementation described in [29]. The feature size for each patch is 256, yielding a feature dimension of 25600 for 10×10 blocks.

For dynamic features, the settings are the following:

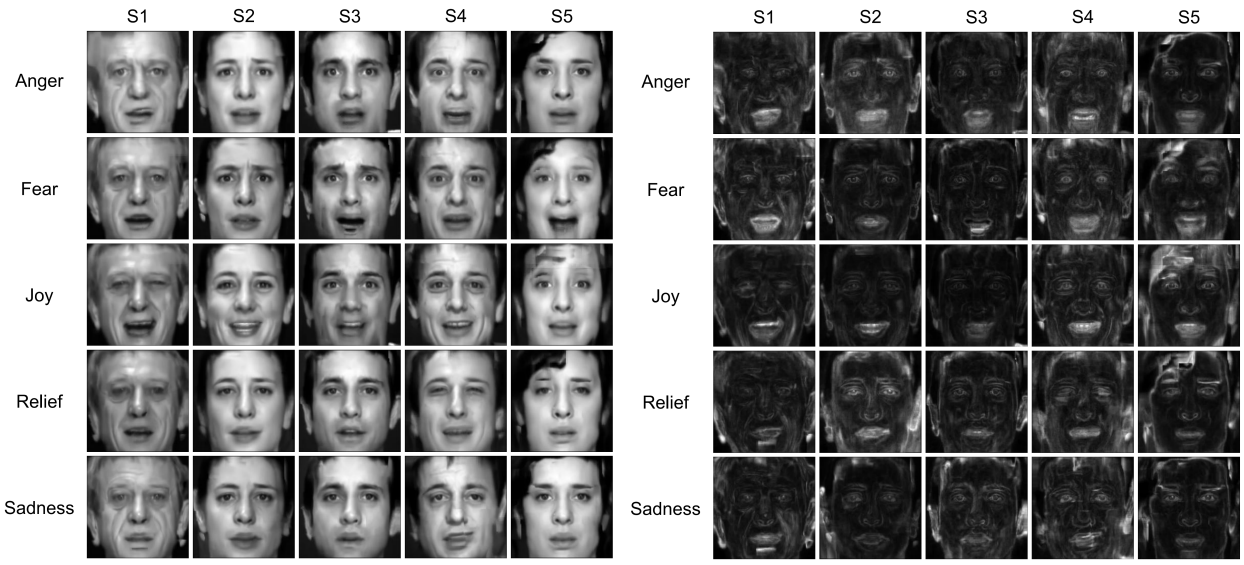
1) TOP-LBP: since the feature dimension enlarges by a multiple of 3 compared with LBP, the total feature dimension is $3 \times 5900 = 17700$.
2) TOP-LPQ: similar to TOP-LBP, for each sequence the TOP-LPQ feature dimension is $3 \times 25600 = 76800$.

In the following sections, we report results of various combinations of static and dynamic features on both LRE and SER representations.

B. Uncontrolled Data: FERA-GEMEP Dataset

The FERA-GEMEP [56] is an uncontrolled facial expression dataset, *i.e.*, in the process of data collection, subjects are asked to convey one of the following five emotions at a time: *Anger*, *Fear*, *Joy*, *Relief*, and *Sadness*, without any control of their head or body movement. Each video segment contains one subject displaying expressions that correspond to one emotion. The average video length is about 2 seconds with 30 frames per second (FPS). Most subjects are uttering some meaningless phrases while displaying expressions, and there are 3 to 5 videos for each subject with the same emotion. All experiments in this paper are designed to be *person-independent* (*i.e.*, no testing subject is included in the training set), in order to validate the generalization ability of the proposed algorithm.

The original FERA-GEMEP data contain a training and a testing set, where there are 7 subjects (3 males and 4 females) in the training set, and 6 subjects (3 males and 3 females) in the testing set. The facial expression videos



(a) The Low-Rank Expression (LRE) representations

(b) The Sparse Residual Expression (SRE) representations

Fig. 5. The sparse decomposition of a facial expression sequence. Each grid is an image representation of a video sequence. (a) The LRE representations. (b) The corresponding SRE representations.

of both sets are provided and the expression label is only available for the training set. The participants can use other publicly available data in addition to FERA training set for training. The predictions on the test set are submitted to the FERA organizer for an independent evaluation. Table I shows the challenge participants[1] and the corresponding person-independent test results.

Since the ground truth labels are only available for the training set, we evaluate our algorithm using this set. Note that although our results are not directly comparable to those in Table I, we compare our method with the best performer in Table I, *i.e.*, EAI [12], under the protocol of this paper. Specifically, we carry out leave-one-subject-out cross-validation using both static and dynamic features extracted from the LREs and SREs and these features are evaluated both individually and jointly. Fig. 5 provides a visualization of the mean LRE and SRE representations of sample expression videos of various subjects for all emotion categories in the training set. The classification results are shown in Table II and Table III, respectively. The symbol "–" means a particular module is not used for feature extraction. In addition, we report the results of the best performing approach [12] in FERA challenge as a baseline comparison with same settings. We should address that unlike most of the other methods which use additional training data, this work and [12] solely rely on the FERA training set.

Several interesting findings can be observed from the results in Table II and Table III. *First*, the performance using static features on LRE only is on par with the EAI representation in [12], 0.68 for LBP and 0.7 for LPQ. This is congruent with our qualitative analysis in Section III-D such that EAI and mean LRE have very similar appearance. *Second*, when considering single module performance, LRE results are consistently higher than SER ones (for both static and dynamic feature using both LBP and LPQ), which means LREs are more discriminative than SERs. This shows that if only single module is to be considered, global structure of a facial expression is more informative than its local muscle motion. *Third*, combining both module almost always improves the recognition performance. For example, in Table III, single module performances for LRE and SER using LPQ features are 0.7 and 0.62, respectively. Combining both modules improves the accuracy to 0.75, which demonstrates that the information captured by both modules complements each other and improved results are obtained by exploiting the local muscle motion. *Fourth*, incorporating dynamic features further improves the performance. The best performance is achieved when dynamic features are extracted from both the LRE and SER modules. This informs us that the person-independent modeling is an effective strategy, which not only diminishes the person-specific information, but also provides a reasonable means for dynamic feature extraction.

The confusion matrix for the best-performance setting, *i.e.*, LRE+SER with TOP-LPQ, is tabulated in Table. IV. *Anger* and *Fear* are misclassified to each other more often because the global structure of these two expressions are more similar compared with other expressions. This is qualitatively demonstrated in Fig. 5(a), where the appearance of the first two rows resemble each other more than other rows. It is also shown in [40] that *Fear* is indeed the most challenging case for this dataset. However, incorporating local muscle dynamics helps better characterizing the *Fear* expression.

[1]Teams are ranked based on the FERA Challenge person-independent test. **UCR**: University of California at Riverside; **UCSD-CERT**: University of California at San Diego; **UIUC-UMC**: University of Illinois at Urbana-Champaign; University of Missouri; **KIT**: Karlsruhe Institute of Technology; **ANU**: Australian National University; **NUS**: National University of Singapore; **QUT-CMU**: Queensland University of Technology; Carnegie Mellon University; **UCL**: University College London; **UMont.**: University of Montreal; **MIT-Cambridge**: Massachusetts Institute of Technology; University of Cambridge.

TABLE II
CLASSIFICATION ACCURACY ON FERA-GEMEP DATASET USING LBP

LRE	SER	Accuracy	LRE	SER	Accuracy
LBP	–	0.68	–	LBP	0.6
TOP-LBP	–	0.67	–	TOP-LBP	0.62
LBP	LBP	0.7	LBP	TOP-LBP	0.72
TOP-LBP	LBP	0.7	TOP-LBP	TOP-LBP	**0.73**
UCR [12]: LBP		0.68	UCR [12]: TOP-LBP		0.69

TABLE III
CLASSIFICATION ACCURACY ON FERA-GEMEP DATASET USING LPQ

LRE	SER	Accuracy	LRE	SER	Accuracy
LPQ	–	0.7	–	LPQ	0.62
TOP-LPQ	–	0.72	–	TOP-LPQ	0.65
LPQ	LPQ	0.75	LPQ	TOP-LPQ	0.77
TOP-LPQ	LPQ	0.76	TOP-LPQ	TOP-LPQ	**0.78**
UCR [12]: LPQ		0.7	UCR [12]: TOP-LPQ		0.73

To thoroughly evaluate our approach, we carry out more experiments with analysis. Unless mentioned otherwise, the following analysis uses this best-performing setting.

Effect of Training Size with a Fixed Subject Pool: we observe that the performance of EAI (*i.e.*, 0.7 as shown in Table III) is inferior to that in the FERA challenge [12] (*i.e.*, 0.75 as shown in Table I). We believe that this is due to a smaller training size in our person-independent experimental setting. To verify this hypothesis, we carried out a leave-one-subject-out cross-validation with respect to different training sizes using various representations. For a specific training set size, we randomly select the training data according to person-independent constraint from the subject pool to train a classifier. This experiment is repeated for 10 times for each training set size, and the average classification rate is reported in Fig. 6.

As seen in Fig. 6, the classification rate steadily raises as the training set size increases for all three curves. This informs us that given more training data from the same training subject pool, the performance improves. Additionally, it is observed that the increase rate for LRE+SER, especially when training size becomes larger. This demonstrates that the SER indeed captures more discriminative facial muscle motion for expression recognition and provides useful and complementary information to LRE. The performance of LRE+SER is not saturated as the training size reaches the maximum (130 in this experiment), suggesting a potential performance gain using more data.

TABLE IV
CONFUSION MATRIX FOR THE FERA TRAINING DATASET USING LER+SER. (AN=ANGER, FE=FEAR, JO=JOY, RE=RELIEF, SA=SADNESS)

Prediction \ True Label	An	Fe	Jo	Re	Sa
An	**62.1**	13.2	4.8	3.6	
Fe	16.4	**71.4**	3.2	1.8	
Jo	7.5	5.5	**85.5**	3.6	
Re	5.8	1.8	4.8	**78.4**	8.6
Sa	8.2	8.1	1.6	12.5	**91.4**
Average			**77.8**		

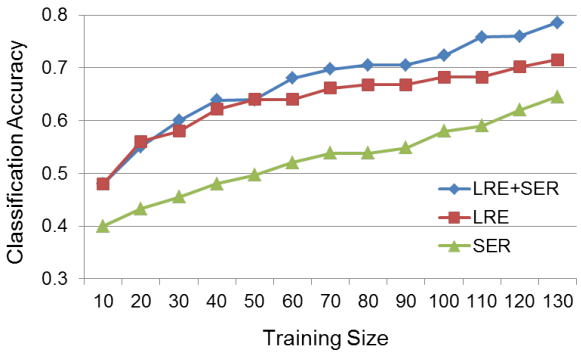


Fig. 6. The person-independent recognition accuracy with various training size from a fixed subject pool. The feature in use is TOP-LPQ. The performance of LRE is spurred by the additional information captured in SER.

Effect of Subject Pool Size: under a person-independent setting, it is also interesting to discover the impact of the number of subjects in the training set. Thus, for each leave-one-subject-out test using LRE+SER, we vary the number of subjects, k, in training from 1 to 6 (since there are 7 subjects in the dataset), and conduct the test for $\binom{6}{k}$ times. For example, when 2 subjects are in the training, there can be $\binom{6}{2} = 15$ possible combinations of subjects, we record the test results for all combinations. Fig. 7 shows an increasing trend for recognition accuracy as the number of subjects in training increases. This shows that the performance is likely to be improved as the number of training subjects increases. In addition, the decreasing variance of the classification accuracy can be observed, suggesting a more stable performance as the training subjects pool grows.

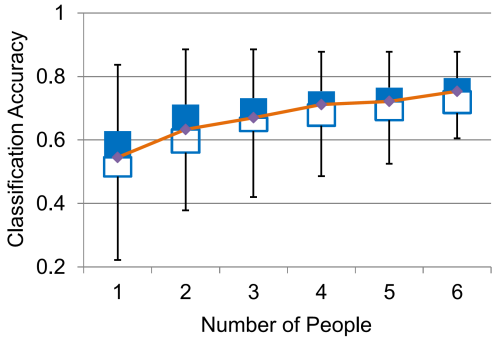


Fig. 7. The boxplot of person-independent recognition accuracy using LRE+SER and TOP-LPQ with various number of people in the training set. The mean performance at the second quartile are connected to illustrate the trend. In general, the performance increases and becomes more stable when the number of training subjects grows.

C. Controlled Data: CK+ dataset

The Extended Cohn-Kanade dataset (CK+) [38] is a facial expression dataset with 7 categories of expressions, namely, *Anger*, *Contempt*, *Disgust*, *Fear*, *Happy*, *Sadness*, and *Surprise*. This is a controlled dataset since the subjects constrained their head pose on purpose, and the expression displayed always follows the "neutral-onset-apex" pattern. These explicit control may not only hide the true muscle contraction of an

expression, but also changes the expression dynamics [57], [35]. Although the philosophy of our algorithm is not designed for this fully controlled data, for rigorous validation of our method, we also carry out person-independent tests and evaluate our LRE+SER representation using LPQ features on 316 sequences from 123 subjects in this dataset. LPQ with same setting is chosen for a better comparison with the other methods.

TABLE V
CONFUSION MATRIX FOR CK+ DATASET USING LER+SER. (AN=ANGER, CO=CONTEMPT, DI=DISGUST, FE=FEAR, HA=HAPPY, SA=SADNESS, SU=SURPRISE)

		True Label						
		An	Co	Di	Fe	Ha	Sa	Su
Prediction	An	**75**	22.2	1.8	8		25.9	
	Co	2.3	**72.2**				3.7	1.2
	Di	4.5		**94.6**		1.6	3.7	2.4
	Fe	2.3			**68**		3.7	1.2
	Ha	4.5		1.8	8	**96.8**		
	Sa	9.1			4		**48.1**	
	Su	2.3	5.6	1.8	12	1.6	14.8	**95.2**
Average					**85.1**			

As seen in confusion matrix in Table V, the LRE+SER representation achieves 85.1% accuracy on average, outperforming an accuracy of 82.6% by EAI reported in [12]. This again demonstrates that the information captured by the SER improves the expression recognition accuracy. The algorithm in [14] using the same dynamic texture feature, has a better performance on CK+ dataset. We believe it is because the experiment in [14] was designed with 10-fold cross validation. Although there is no subject that displays the same expression twice in CK+, this experiment setting does not strictly exclude testing subject in the training set. On the contrary, our experimental design is strictly person-independent.

Besides, the CK+ data has only one or a few apex expression for each "neutral-onset-apex" sequence. This unrealistic setting reduces the intensity of the expression captured by LREs, which appears to be more neutral compared with the mean LRE of FERA-GEMEP data. Fig. 8 shows a visualization of mean LRE representations for sample sequences from the CK+ dataset.

In addition, through a diagnostic experiment, we have found that the classification performance has a strong correlation with the training size for each class category, as shown in Fig. 9. We plot the classification accuracy along with the training set size for each category in Fig. 9. The categories such as *Disgust*, *Happy*, and *Surprise*, which have more training data, achieve a much higher accuracy. This also shows the potential of improvement for our algorithm given more training samples for the under-performing categories.

V. CONCLUSIONS

We have proposed a low-rank sparse recovery framework to effectively extract dynamic muscle motion for unconstrained person-independent facial expression recognition in videos. Our approach decomposes a facial expression sequence into two representations, namely the Low-Rank Expression (LRE) and the Sparse Expression Residual (SER). The LRE recovers the underlying expression appearance shared in the entire expression sequence, while the SER captures the local muscle motion. The LRE and SER complement each other for a better expression recognition performance. Extensive experimental results have demonstrated the effectiveness of the proposed method and superior performance has been observed as compared to the state-of-the-art techniques. In addition, our approach has the potential of improvement given more samples from a larger training population.

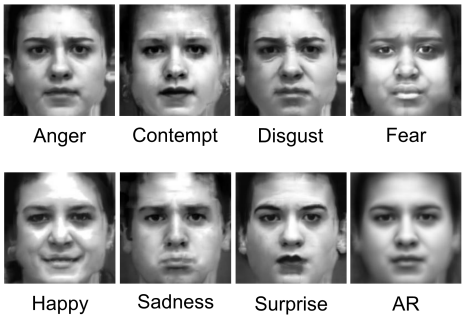


Fig. 8. Sample mean LRE representations from the CK+ dataset. The lower right grid shows the Avatar Reference generated from [12]. Although the person-specific information is attenuated as the appearance of different subjects visually resemble each other, the discriminative expression information is hidden and difficult to be distinguished. In general, they appear to be more "neutral" compared with the mean LRE representations of the sequences in FERA-GEMEP dataset as shown in Fig. 5(a).

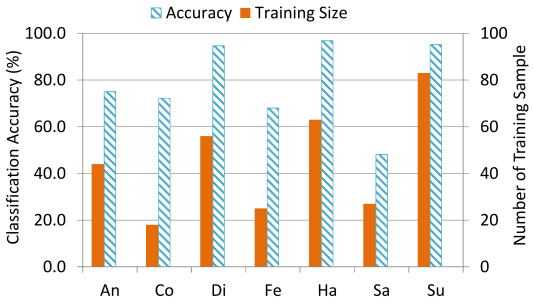


Fig. 9. Relationship between the classification accuracy and the training size for each expression category. The accuracy is generally higher for categories with more training data.

REFERENCES

[1] R. W. Picard, "Affective computing: Challenges," *International Journal of Human-Computer Studies*, 2003.
[2] M. Pantic and L. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000.
[3] F. De la Torre and J. F. Cohn, *Guide to Visual Analysis of Humans: Looking at People*. Springer, 2011, ch. Facial Expression Analysis.
[4] D. McDuff, R. Kaliouby, T. Senechal, D. Demirdjian, and R. Picard, "Automatic measurement of ad preferences from facial responses gathered over the internet," *Image and Vision Computing*, 2014.
[5] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. de C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset," *IEEE Trans*, 2015.
[6] *FERA2011: Facial Expression Recognition and Analysis Challenge*, http://sspnet.eu/fera2011/.

[7] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[8] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "Computer Expression Recognition Toolbox," in *Proc. FG*, 2011.

[9] M. Valstar, B. Jiang, M. Méhu, M. Pantic, and K. Scherer, "The First Facial Expression Recognition and Analysis Challenge," in *Proc. FG Workshop on FERA Challenge*, 2011.

[10] A. Fridlund, P. Ekman, and H. Oster, "Facial expressions of emotion: Review literature 1970-1983," in *In A. W. Siegman & S. Feldstein (Eds.), Nonverbal behavior and communication*, 1987.

[11] M. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, 2012.

[12] S. Yang and B. Bhanu, "Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[13] M. Dahmane and J. Meunier, "Prototype-based modeling for facial expression analysis," *IEEE Trans. Multimedia*, 2014.

[14] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007.

[15] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, 2011.

[16] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense Correspondence across Scenes and its Applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011.

[17] I. Essa and A. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997.

[18] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1999.

[19] S. Zafeiriou and I. Pitas, "Discriminant graph structures for facial expression recognition," *IEEE Trans. Multimedia*, 2008.

[20] R. El Kaliouby and P. Robinson, "Mind Reading Machines: Automated Inference of Cognitive Mental States from Video," in *Proc. SMC*, 2004.

[21] J. Whitehill and C. Omlin, "Haar features for facs au recognition," in *Proc. FG*, 06.

[22] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1999.

[23] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, 2009.

[24] B. Jiang, M.Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proc. FG*, 2011.

[25] K. Levi and Y.Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. CVPR*, 2004.

[26] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. CVPR*, 2005.

[27] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion Recognition Using PHOG and LPQ features," in *Proc. FG Workshop on FERA Challenge*, 2011.

[28] M. Dahmane and J. Meunier, "Emotion Recognition using Dynamic Grid-based HoG Features," in *Proc. FG Workshop on FERA Challenge*, 2011.

[29] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in *Proc. ICISP*, 2008.

[30] Y. Huang, Y. Li, and N. Fan, "Robust symbolic dual-view facial expression recognition with skin wrinkles: Local versus global approach," *IEEE Trans. Multimedia*, 2010.

[31] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, 2010.

[32] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Trans. Multimedia*, 2010.

[33] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, 2013.

[34] Q. Mao, MingDong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, 2014.

[35] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[36] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the Enigmatic Face: the Importance of Facial Dynamics to Interpreting Subtle Facial Expressions," *Psychological Science*, 2005.

[37] Y. Yacoob and L. Davis, "Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1996.

[38] P. Lucey, J. 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 Workshop*, 2010.

[39] J. Päivärinta, E. Rahtu, and J. Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[40] M. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, 2012.

[41] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. CVPR*, 2008.

[42] J. Wright, A. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009.

[43] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *Proc. ICCV*, 2011.

[44] Z. Zhang, X. Liang, A. Ganesh, and Y. Ma, "TILT: Transform Invariant Low-rank Textures," *Int. J. Comput. Vis*, 2012.

[45] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[46] Y. Lin, M. Song, D. T. P. Quynh, Y. He, and C. Chen, "Sparse coding for flexible, robust 3d facial-expression synthesis," *IEEE Computer Graphics and Applications*, 2012.

[47] U. Tariq, J. Yang, and T. Huang, "Multi-view facial expression recognition analysis with generic sparse coding feature." Springer Berlin Heidelberg, 2012.

[48] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *Journal of the ACM*, 2011.

[49] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *UIUC Technical Report UILU-ENG-09-2214*, 2009.

[50] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *Proc. CVPR*, 2010.

[51] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, 2009.

[52] D. Lowe, "Object Recognition from Local Scale-invariant Features," in *Proc. ICCV*, 1999.

[53] P. Viola and M. Jones, "Robust Real-time Face Detection," *Int. J. Comput. Vis*, 2004.

[54] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.

[55] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[56] *Facial Expression Recognition and Analysis Challenge data*, http://sspnet.eu/fera2011/fera2011data/.

[57] M. S. Bartlett, G. Littlewort, B. Braathen, T. J. Sejnowski, and J. R. Movellan, "A prototype for automatic recognition of spontaneous facial actions," in *Proc. NIPS*, 2003.