CROSS-MODALITY POSE-INVARIANT FACIAL EXPRESSION

Jordan Hashemi, Qiang Qiu, Guillermo Sapiro

Department of Electrical and Computer Engineering, Duke University, USA

ABSTRACT

In this work, we present a dictionary learning based framework for robust, cross-modality, and pose-invariant facial expression recognition. The proposed framework first learns a dictionary that i) contains both 3D shape and morphological information as well as 2D texture and geometric information, ii) enforces coherence across both 2D and 3D modalities and different poses, and iii) is robust in the sense that a learned dictionary can be applied across multiple facial expression datasets. We demonstrate that enforcing domain specific block structures on the dictionary, given a test expression sample, we can transform such sample across different domains for tasks such as pose alignment. We validate our approach on the task of pose-invariant facial expression recognition on the standard BU3D-FE and MultiPie datasets, achieving state of the art performance.

***Index Terms*—** Facial expression, domain adaptive, pose-invariant, cross-modality

1. INTRODUCTION

The analysis of facial expression is studied in computer vision, psychology, psychiatry, and marketing, all of which require a facial expression recognition (FER) system to be robust to changes in pose. In particular for the psychology and psychiatry fields, risk signs of aniexty and autism can be depicted from facial expressions as the participant is looking at various stimuli [1, 2]. Robustness to pose is especially important since the experts need to analyze participants in their natural states, in other words being observed in an unconstrained manner (see [3] and [4] for examples). Many state of the art facial expression approaches focus on frontal or nearly frontal images of the face [5, 6]. Changes in head pose or facial expression cause nonlinear transformations of the face in a 2D image, making it a non-trivial task to classify expressions under varying poses [7]. Even with recent FER advancements, manually coding of facial expression is still performed in the psychiatry and psychology fields due in part to this challenge.

Approaches to handle facial expression across multiple poses fall within two main categories. The first category corresponds to approaches based on learning expression models on a discrete set of poses [8, 9]. For example, [8] employ a 2 stage approach where they first train a classifier to distinguish pose, and then train pose-dependent classifiers across expressions. The second category involves approaches that learn the mappings of the expressions as a function of pose [10, 11, 12]. Notably, [10] presents an accurate geometric based approach to first learn the transformation of facial points at any given pose to a frontal pose, then FER is performed on facial points from the projected frontal pose, thus requiring only one posed classifier. The work [12] adopts a Partial Least Squares approach, that has been explored in facial recognition, to model the relations between pairs of images of the same person at different poses and expressions.

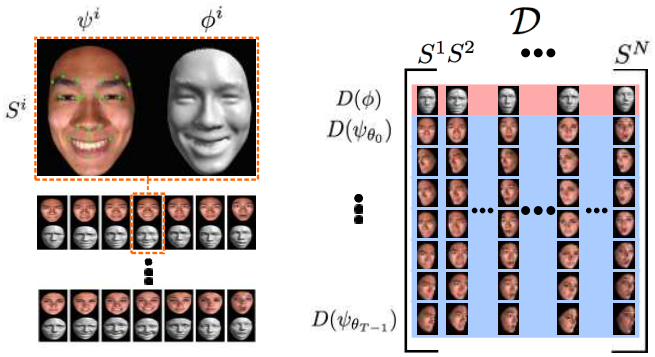


Fig. 1: Overview of multi-modality and pose-invariant dictionary construction. **(a)** Samples of 3D textured face scans from the BU3D-FE dataset, where each sample S^i can be decomposed into a 3D and 2D component, ψ^i and ϕ^i respectively. The used 19 facial landmarks are highlighted with green markers. **(b)** the dictionary is composed of blocks containing different modalities and poses. The red section represent dictionary block $D(\phi)$ containing 3D features, while the blue sections represent dictionary blocks $D(\psi_{\theta_t})$ containing 2D features from the synthesized head poses θ_t.

In addition to FER in 2D images, much attention has been focused on using 3D face scans [13, 14]. Specifically, textured 3D face scans not only contain 3D features (e.g., morphological and shape), but also 2D features (e.g., geometric and texture). Zhao *et al.* [13] have shown that when dealing with 2D and 3D features independently on a frontal face, the ordering of discriminative power for FER is morphological, shape, and texture; and combining all three feature modalities together achieves the strongest discriminative power. Although textured 3D face scans provide the most discriminative features, technology has not yet allowed for practical acquisition in unconstrained environments, such as capturing child facial behaviors in a doctor's office.

Dictionary based approaches have been extensively used for classification and regression in the areas of facial recognition and expression [15, 16]. Furthermore, one can apply

Work partially supported by NSF, NGA, AFOSR, ARO, and ONR. We thank our team on the *Duke Information and Child Mental Health Initiative* for important feedback and comments. The work with that team is the motivation behind the framework here reported.

sparse based methods by incorporating regularized penalty functions to determine sparse coefficients in a more greedy fashion [16, 17]. By encoding structure along atoms in the dictionary, such as annotating or grouping atoms in the dictionary with class labels, the sparse coefficients can provide knowledge to the class that the unseen face belongs to. Recent work has also focused on encoding structure within the atoms themselves, namely domain adaptive dictionary learning [18]. A powerful example focuses on encoding atoms so they contain blocks of features across different domains [18].

In this work, we develop a framework based on learning and applying a robust, cross-modality, and pose-invariant dictionary to the recognition of facial expressions. The presented framework first learns a dictionary that i) contains both 3D shape and morphological information as well as 2D texture and geometric information, ii) enforces coherence across both 2D and 3D modalities and different poses, and iii) is robust in the sense that a learned dictionary can be applied to multiple facial expression datasets. With our novel dictionary based approach, we achieve powerful results in the task of pose-invariant FER. The rest of the paper is organized as follows: in Section 2 we describe our approach for constructing and applying the proposed dictionary to pose-invariant FER. In Section 3 we validate our approach using two publicly available datasets: the BU-3D Facial Expression (BU3D-FE) [19] and the CMU Pose, Illumination, and Expression (MultiPie) [20]. Section 4 concludes the paper.

2. PROPOSED APPROACH

We now describe our approach, which can be separated into two main components: constructing the cross-modality and pose-invariant dictionary, and applying it to the task of pose-invariant FER. Figures 1 and 2a show the outline of our approach for dictionary construction and cross domain representation.

2.1. Pose-invariant dictionary

Given a dataset containing N textured 3D face scans under varying expressions, we define each sample $S^i = \{\phi^i, \psi^i\}$, where $i = 1, 2, ..., N$, and ϕ^i and ψ^i represent the 3D specific and 2D specific information from sample i, respectively. From a single textured 3D face scan, 2D images with varying head poses, θ, can be synthesized. In this sense, we can decompose a sample as $S^i = \{\phi^i, \psi^i_{\theta^{T-1}_{t=0}}\}$, with T different head poses, θ_0 represents a frontal face, and $\psi^i_{\theta_t}$ represents 2D specific information at pose θ_t for sample i. Note that 3D specific information does not change with varying head poses. For all samples, we define the dictionary block $D(\phi) \in \mathbb{R}^{d_m \times N}$ of extracted 3D features as

$$D(\phi) = \left[f(\phi^1),\, f(\phi^2),\, \cdots,\, f(\phi^N)\right],$$

where $f(\phi^i) \in \mathbb{R}^{d_m}$ represents the column array of computed frontal 3D features from the i^{th} sample. Similarly, for all samples with simulated head pose θ_t, we define the block $D(\psi_{\theta_t}) \in \mathbb{R}^{d_n \times N}$ of extracted 2D features as

$$D(\psi_{\theta_t}) = \left[f(\psi^1_{\theta_t}),\, f(\psi^2_{\theta_t}),\, \cdots,\, f(\psi^N_{\theta_t})\right],$$

where $f(\psi^i_{\theta_t}) \in \mathbb{R}^{d_n}$, represents the column array of computed 2D features from the i^{th} sample at pose θ_t.

The cross-modal and pose-invariant dictionary, $\mathcal{D}$, is organized by stacking the dictionary blocks (see Figure 1)

$$\mathcal{D} = [D(\phi);\, D(\psi_{\theta_0});\, D(\psi_{\theta_1});\, \cdots;\, D(\psi_{\theta_{T-1}})],$$

with the stacking operator $[D(\phi);\, D(\psi_{\theta_0})] = \begin{bmatrix} D(\phi) \\ D(\psi_{\theta_0}) \end{bmatrix}$. $\mathcal{D} \in \mathbb{R}^{(d_m + T \times d_n) \times N}$ is composed of a total of $T+1$ blocks, specifically one block containing the 3D features, $D(\phi)$, and T blocks containing the 2D features from each of the T simulated head poses, $D(\psi^i_{\theta^{T-1}_{t=0}})$. This block structure within the dictionary $\mathcal{D}$ imposes coherence across the different domains.

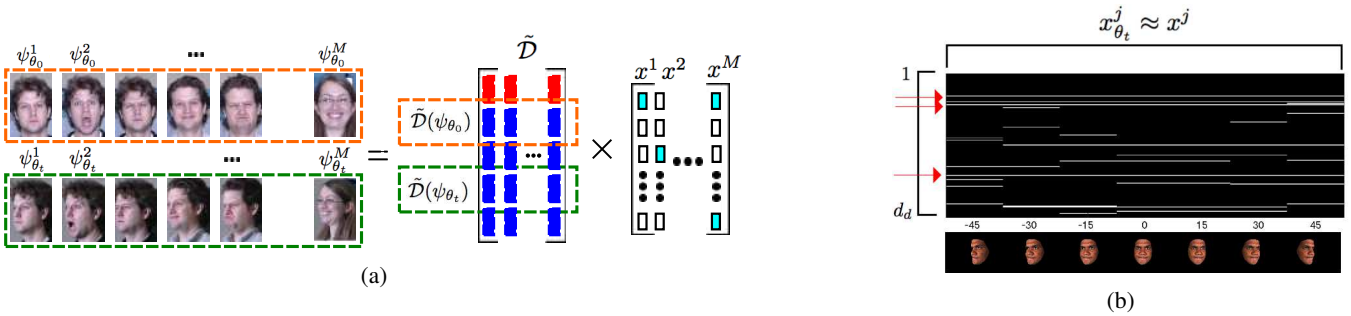
In addition, we learn a more compact dictionary by applying a dictionary learning method, such as K-SVD [21], creating a new dictionary $\tilde{\mathcal{D}} \in \mathbb{R}^{(d_m + T \times d_n) \times d_d}$ where $d_d \leq N$. Note that since the block structure still remains, the coherence between the domains is preserved: $\mathcal{D}$ is transferred to $\tilde{\mathcal{D}} = [\tilde{D}(\phi_{\theta_0});\, \tilde{D}(\psi_{\theta_0});\, \tilde{D}(\psi_{\theta_1});\, \cdots;\, \tilde{D}(\psi_{\theta_{T-1}})]$, where now $\tilde{D}(\phi) \in \mathbb{R}^{d_m \times d_d}$ and $\tilde{D}(\psi^i_{\theta_t}) \in \mathbb{R}^{d_n \times d_d}$ (see Figure 1).

2.2. Cross modality and domain representation

The learned dictionary, $\tilde{\mathcal{D}}$, contains a dense amount of expression information jointly learned across multiple domains (3D and different poses in 2D). Let unseen samples containing expression class labels and only 2D images at any pose θ_t be defined as $Q^j = \{\psi^j_{\theta_t},\, L^j\}$, where $j = 1, 2, ..., M$ represent the samples and $L^j = 1, 2, ..., C$ represents the class label of the j^{th} sample taking the values of C possible classes. The goal is to represent Q^j as a sparse linear combination of the frontal 3D and frontal 2D features in $\tilde{\mathcal{D}}$, namely $[\tilde{\mathcal{D}}(\phi_{\theta_0});\, \tilde{D}(\psi_{\theta_0})]$, since they are know to have large discrimination power for FER. Thus we wish to solve:

$$\begin{aligned} \{\tilde{Q}^j, x^j\} \;=\; & \underset{x^j, \tilde{Q}^j}{\operatorname{argmin}} \|\tilde{Q}^j - [\tilde{D}(\phi);\, \tilde{D}(\psi_{\theta_0})]\, x^j\|_2^2 \\ & s.t. \|x^j\|_0 \leq \lambda, \end{aligned} \tag{1}$$

where $x^j \in \mathbb{R}^{d_d}$ is the sparse coefficient vector, $\tilde{Q}^j \in \mathbb{R}^{(d_n + d_m)}$ is the transformed version of sample Q^j onto the domains represented by $[\tilde{D}(\phi);\, \tilde{D}(\psi_{\theta_0})]$, $\|x^j\|_0$ counts the number of non-zeros values in x^j, and λ is the imposed sparsity constant. (1) is not directly solvable since the 3D information and frontal 2D information, $\tilde{Q}^j$, and the sparse coefficient vector, x^j, are unknown. Instead we propose to represent the unknown 3D and frontal 2D information via our domain adaptive dictionary. We propose that the computed sparse coefficient vector in the known domain provided by Q^j can be directly applied to dictionary blocks in unseen domains to estimate $\tilde{Q}^j$.



(a) (b)

Fig. 2: **(a)** Overview of the proposed cross-modality and pose-invariant representation and examples from the MultiPie dataset. Given the dictionary $\tilde{D}$, we propose that the sparse coefficient vectors between the same subjects performing the same expressions at different poses or modalities will be nearly identical. The dotted colored boxes around the faces (orange and green) represent observation of the same subjects from a given expression at different poses. These observations can be represented by a linear combination of the same sparse coefficients being applied to a given sub-dictionary of $\tilde{D}$, that is represented with the respective dotted color boxes. **(b)** Example of the sparse coefficient vectors extracted from a subject performing a Disgust expression at 7 poses. Each of the 7 columns in the image correspond to a sparse coefficient vector x^j extracted at a given pose, and the rows represent weights corresponding to atoms in $\tilde{D}$. Images of the input subject are shown below each pose. Notice the horizontal line structure (depicted by the red arrows) throughout the sparse coefficient vectors at different poses, reinforcing the notation that the sparse coefficients extracted for different poses are approximately consistent thus pose invariant.

Since Q^j provides information in the domain $\psi^j_{\theta_t}$, the sparse coefficient vector can be determined from:

$$x^j \quad = \quad \underset{x^j}{\operatorname{argmin}} \|\psi^j_{\theta_t} - \tilde{D}(\psi_{\theta_t})\, x^j\|_2^2,\ s.t. \|x^j\|_0 \leq \lambda.$$

If θ_t is unknown, it can be estimated from a variety of head pose approaches [22] or by determining which domain block in $\tilde{D}$ gives the lowest reconstruction error. Due to the coherence across domains within the stacks of the dictionary $\tilde{\mathcal{D}}$, we assume the sparse coefficient vector, x^j, should not differ greatly between extracted data of the same subject but in different domains (see Figure 2a). In other words,

$$x^j \quad = \quad \underset{x^j}{\operatorname{argmin}} \|\psi^j_{\theta_t} - \tilde{D}(\psi_{\theta_t})\, x^j\|_2^2,\ s.t. \|x^j\|_0 \leq \lambda \qquad (2)$$

$$\approx \quad \underset{x^j}{\operatorname{argmin}} \|\psi^j_{\theta_{t' \neq t}} - \tilde{D}(\psi_{\theta_{t' \neq t}})\, x^j\|_2^2,\ s.t. \|x^j\|_0 \leq \lambda$$

$$\approx \quad \underset{x^j}{\operatorname{argmin}} \|\phi^j - \tilde{D}(\phi)\, x^j\|_2^2,\ s.t. \|x^j\|_0 \leq \lambda. \qquad (3)$$

This assumption is explored and further validated in Section 3.2 and Figure 2b. Equations (2) and (3) state that x^j can be determined from any domain that lies in both Q^j and $\tilde{\mathcal{D}}$. Once x^j is determined, $\tilde{Q}^j$ can be computed from (1).

3. EXPERIMENTAL VALIDATION

3.1. Datasets and setup

To evaluate our proposed method, we used two publicly available face datasets: the BU3D-FE and the MultiPie datasets. The BU3D-FE dataset consists of textured 3D face scans of 100 subjects performing 6 different expressions: Anger (AN), Disgust (DI), Fear (FE), Happy (HA), Sad (SA), Surprised (SU) at 4 different levels, and a Neutral (NE) expression (see Figure 1 for examples). For this demonstration, we only considered the data from the maximum level which corresponds to the apex of the expression. From the MultiPie dataset, we selected 2D images from 160 subjects performing 4 different expressions: DI, HA, SU, and NE at 7 different yaw angles $(0, -45, -30, -15, 15, 30, 45)$ (see Figure 2a for examples). The MultiPie dataset also contains each expression and pose at different illuminations; however, we only considered the data from the frontal illumination.

To compute features from the datasets, 49 facial landmarks were automatically extracted with the IntraFace software [23]. Faces were aligned and normalized to a mean face across the BU3D-FE dataset using the inner-eye landmarks and the spine of nose. For selection of 2D and 3D features, we followed the state of the art approach in [13], where four modalities of features consisting of morphological, shape, texture, and geometric features are computed around 19 of the facial landmark points (see Figure 1). 3D morphological features consist of 157 Euclidean distance pairs between the 19 landmarks on the range data of the faces. 3D shape features consist of multi-scale local binary pattern (LBP) patches around each of the 19 landmarks on the image of the range data. Specifically, we compute LBP with radii ranging from 1 to 5, where the total features extracted across all the patches at a given LBP scale is 4275. 2D texture features are computed in the same manner as the 3D shape features except extracted on the 2D textured images. 2D geometric features consist of the same distance pairs as the 3D morphological features, but the range value of each landmark is not considered. Principal component analysis (PCA) is performed on each modality independently, preserving at least 95% of the variation, thus reducing the dimensions of the morphological, shape, texture, and geometric features to $100, 1000, 1000, 100$ respectively. Thus in the following experiments $d_n = d_m = 1100$. For all experiments shown, 2D images containing 7 poses with yaw angles $(0, -45, -30, -15, 15, 30, 45)$ were considered.

The sparse coefficient vectors for the projected discriminant frontal $\tilde{Q}^j$ representations were determined through Orthogonal Matching Pursuit (OMP) with sparsity constant $\lambda = \frac{1}{7}\, d_d$, since the dictionary is composed of an even representation of samples across 7 expressions. For each experiment, a

Appr.	Expr.							Total
	AN	DI	FE	HA	SA	SU	NE	
3D [13]	83	87	68	93	83	95	—	85
Proposed	85	85	75	91	77	94	—	85
2D [10]	68	75	63	81	63	82	71	77
Proposed w/NE	82	85	72	91	66	94	81	82

Table 1: Comparisons of recognition rates (%) for varying expression (expr.) across different methods on BU3D-FE dataset, including a 3D specific framework [13], a pose-invariant framework [10], and our proposed approach when Neutral is and is not included. Note that [13] only considers a frontal pose and use 3D data for testing, while we adopt a more general and challenging testing setup.

single facial expression classifier was trained by applying to the extracted $\tilde{Q}^j$ representations a multi-class Support Vector Machine (SVM) [24] with a radial basis function kernel. Experiments in Section 3.2 perform a five-fold cross validation procedure to construct and test on the pose-invariant dictionary. Out of the samples chosen for constructing the dictionary, a ten-fold cross validation procedure was performed to determine the optimal SVM parameters.

3.2. Validation on BU3D-FE

We now present experiments performed on the BU3D-FE dataset. Since the 3D modalities and the 7 poses are considered for the dictionary, it contains 8 dictionary blocks (see Figure 1). Furthermore, K-SVD was applied to create a compact dictionary $\tilde{\mathcal{D}} \in \mathbb{R}^{8800\times 400}$. For testing, 2D images of expressions performed at the 7 pose angles are used. Figure 2b provides an example of the sparse coefficients vectors extracted from a given subject performing a specific expression at multiple poses. In this figure one can observe many sparse coefficients that are present throughout all of the poses (red arrows), thus illustrating that the learned dictionary is invariant to observed poses. Furthermore since we assume the sparse coefficient vector is approximately the same given any modality from a sample, then we can project a given sample to modalities that may not have been observed (e.g., projecting a posed image to one containing 3D features).

Our approach achieved high results for pose-invariant FER, achieving 82% and 85% recognition rates when Neutral is and is not considered. In Table 1 we compare our results to those of two recently published, state of the art methods, namely a pose-invariant method involving only 2D modalities [10] and a 3D specific method that only considers frontal face scans [13]. It should be noted the testing setup differed between cited methods. Rudovic *et al.* [10] provide results using manually annotated facial landmarks, and test on a wide variety of poses unseen to the training data including pitch poses. Zhao *et al.* [13] only consider 3D face scans, frontal pose, and do not classify the Neutral expression. With this said, our proposed approach therefore achieves results for FER that are on par with current state of the art approaches on the BU3D-FE dataset in a more general and challenging setting. When not including the (challenging) Neutral expression, we achieve the same recognition rate as [13] even though they only use the frontal pose and 3D data for testing.

Pose (deg)	-45	-30	-15	0	15	30	45
Baseline	67	76	90	91	91	75	64
Proposed	86	87	90	92	90	88	85

Table 2: Comparisons of recognition rates for all expressions across the 7 poses on the MultiPie dataset. Our proposed method performs consistently well across drastic pose changes and significantly outperforms the baseline at sever pose angles.

3.3. Pose-invariant FER from frontal training data

We now present an experiment that utilizes both the BU3D-FE and MultiPie datasets, in order to validate the robustness of our approach. We propose to first learn a cross-modal and pose-invariant dictionary with the textured 3D data provided by the BU3D-FE dataset. Then using the 2D images from the MultiPie dataset, we wish to train and test a FER classifier. Furthermore, we wish to demonstrate the power of our proposed dictionary formulation by learning pose-invariant FER classification models using only frontal faces from the MultiPie dataset as training samples and testing on posed 2D images from the MultiPie dataset. This is the first instance where both of these datasets are utilized simultaneously and in this fashion. Although the expressions presented in the MultiPie dataset are only a subset of those presented in the BU3D-FE dataset, we trained a pose-invariant dictionary based on the entire BU3D-FE at the 7 dynamic pose angles to demonstrate our approach's general usability. Similar to the experiments in Section 3.2, we applied K-SVD to get a final pose-invariant dictionary $\tilde{\mathcal{D}} \in \mathbb{R}^{8800\times 400}$. Five-fold cross validation is carried out on the MultiPie dataset, where at each fold 80% of the MultiPie subjects are used to train a classifier and the other 20% of the subjects at 7 different poses are used for testing. The dictionary learned from the BU3D-FE dataset did not change throughout any of the folds.

Table 2 shows the total recognition rates for all expressions across each of the 7 poses for our proposed method and a baseline method. The baseline method consisted of training a multi-class SVM for each of the expressions performed on a frontal pose using the same set of 2D features as in Section 3.1. Both methods perform very well for nearly frontal faces when the pose is between -15 and 15 degrees; however outside this range, as severe pose changes occur, our method greatly outperforms the baseline method and achieves high recognition rates similar to those of nearly frontal faces.

4. CONCLUSION

We have presented a framework for constructing and learning a cross-modality and pose-invariant dictionary for the task of facial expression recognition. Using the BU3D-FE dataset, we have shown we get results on par with current (frontal) state of the art approaches for 3D and pose-invariant expression recognition. Furthermore, we have validated the robustness of our approach by achieving high performance when two different datasets are combined. The generic nature of our approach allows for many extensions including the use of different features and modalities.

5. REFERENCES

[1] C. Nichols, L. Ibanez, J. Foss-Feig, and W. Stone, "Social smiling and its components in high-risk infant siblings without later asd symptomatology," *JADD*, vol. 44, no. 4, pp. 984–902, 2014.

[2] S. Ozonoff, A. Iosif, F. Baguio, I. Cook, M. Hill, T. Hutman, Rozga A. Roger, S., S. Sangha, M. Sigman, M. Steinfeld, and G. Young, "A prospective study of the emergence of early behavioral signs of autism a prospective study of the emergence of early behavioral signs of autism," *J Am Acad Child Adolesc Psychiatry*, vol. 49, no. 3, pp. 256–266, 2010.

[3] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, L. Yin, K. Chanho, H. Rao, J. Kim, L. Presti, Z. Jianming, D. Lantsman, J. Bidwell, and Y Zhefan, "Decoding children's social behavior," in *CVPR*, 2013, pp. 3414–3421.

[4] J. Hashemi, M. Tepper, T. Spina, A. Esler, V. Morellas, N. Papanikolopoulos, H. Egger, G. Dawson, and G. Sapiro, "Computer vision tools for low-cost and non-invasive measurement of autism-related behaviors in infants," *Autism Research and Treatment*, 2014.

[5] C. Shan, S. Gong, and P. McOwan, "Facial expression recognitino based on local binary patterns: a comprehensive study," in *Image and Vision Computing*, 2009, vol. 27, pp. 803–816.

[6] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," in *PAMI*, 2009, vol. 31, pp. 39–58.

[7] Z. Zhu and Q. Ji, "Robust real-time face pose and facial expression recovery," in *CVPR*, 2006, pp. 681–688.

[8] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," in *Computer Vision and Image Understanding*, 2011, vol. 115, pp. 541–558.

[9] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden markov model supervectors," in *ICME*, 2010, pp. 1202–1207.

[10] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *PAMI*, vol. 35, no. 6, pp. 1357 – 1369, 2013.

[11] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-invariant facial expression recognition using variable-intensity templates," *IJVC*, vol. 83, no. 2, pp. 178–194, 2009.

[12] F. Guney, N. Arar, M. Fischer, and H. Ekenel, "Cross-pose facial expression recognition," *FG*, pp. 1–6, 2013.

[13] X. Zhao, E. Dellandréa, and J. Zou, "A unified probabilistic framework for automatic 3D facial expression analysis based on a bayesian belief inference and statistical feature models," *Image and Vision Computing*, vol. 31, no. 3, pp. 231–245, 2013.

[14] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: a comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[15] S. Taheri, Q. Qiu, and R. Chellappa, "Structure-preserving sparse decomposition for facial expression analysis," *IEEE Trans Image Process*, vol. 23, no. 8, pp. 3590–3603, 2014.

[16] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Yi. Ma, "Robust face recognition via sparse representation," *PAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[18] Q. Qiu, V. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *ECCV*, 2012, pp. 631–645.

[19] L. Yin, X. Wei, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *FG*, 2006, pp. 211–216.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *FG*, 2010, pp. 807–813.

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans Signal Proccess*, vol. 54, no. 11, pp. 4311–4322, 2006.

[22] E. Chutorian and M. Trivedi, "Head pose estimation in computer vision: a survey," *PAMI*, vol. 31, no. 4, pp. 607–626, 2009.

[23] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[24] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Trans Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.