Pairwise Linear Regression: An Efficient and Fast Multi-view Facial Expression Recognition

Mahdi Jampour, Thomas Mauthner and Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria

***Abstract*— Multi-view facial expression recognition (MFER) is an active research topic in facial analysis. In fact, not only the accuracy but also time complexity is desirable for real applications. In this paper, we introduce a new fast and robust approach for recognizing facial expressions in arbitrary views. Our approach relies on learning linear regressions between pairs of non-frontal and frontal sets to virtually compensate occluded facial parts. We learn linear regression for projecting from non-frontal to frontal views. Such approximated frontal training features are applied for training view specific facial expression classifiers. We propose a number of different variants of our approach, including sparse encoding and ridge-regression for feature representation. While classical pose specific methods strongly depend on the quality of the pose estimation step, our approaches maintain their superior behavior even under severe pose noise. We evaluate on both BU3DFE and Multi-PIE datasets and outperform the state-of-the-art in classification accuracy, even with a simple pose specific baseline method, while being extremely robust to feature noise and erroneous viewpoint estimation with our pairwise regression approaches.**

I. INTRODUCTION

Multi-view Facial Expression Recognition (MFER) has attracted significant interest in facial analysis due to its applications in human computer interaction, education, robotics, games, medicine and psychology [1]. Most of the existing approaches work on frontal or near to frontal views [2], [3], [4] whereas in real-world applications, a frontal view is an unrealistic assumption and limits the applicability. For this reason, non-frontal analysis is now one of the active challenges related to facial expression recognition, which needs not only an effective recognition approach, but also a method for compensating missing information (i.e. non-frontal counterpart). This is a challenging problem because some of the facial features which are necessary for recognition are not or not completely available due to the face orientation. For example, eyebrows, which are very important for recognizing facial expression, may not be visible in a non-frontal face.

Lets assume that, we have pairwise sets of non-frontal and frontal views during the training that provide the ability of learning features similarities and regressions between non-frontal and frontal data. This regression from non-frontal to frontal views should fulfill two tasks. First, we would like to compensate invisible facial features in non-frontal views by related visible frontal features given during training of classifiers. Second, we assume that this transformation of all features into an approximated frontal view space diminish the impact of erroneous face alignments or pose estimation errors during test. For this purpose, we employ linear regression with/without an arbitrary intermediate encoding process (e.g. sparse coding or ridge regression) which can provide an elegant approximated transformation on learning collections (e.g. non-frontal to frontal facial expressions). We choose linear regression as a tradeoff between the capability to approximate non-available facial features in arbitrary views, and run time efficiency. In statistics, linear regression is an approach for modeling the relationship between two sets of data (e.g. X and Y) which the first is explanatory variables and the second is dependent variable. In our work, non-frontal features are explanatory variables and correspondence frontal features are dependent variable which means that we want to regress non-frontal features to the frontal features, as they are simpler to analysis for facial expression recognition. The complexity of multi-view facial expression recognition, with all possible variations in viewpoints and expressions, hampers the direct regression from arbitrary views to the frontal one. Therefore, we divide the viewpoint space into several subsets, and estimate the regressions between those subsets and the corresponding frontal data individually.

This defines the overall processing pipeline applied within all approaches proposed in this paper (see Fig. 1). Training data is clustered according to viewpoint information. A regression between this non-frontal and corresponding frontal data is learned individually for each pair (see Section II-B). An optional intermediate step could be applied in the form of encoding features by a global sparse dictionary as described within Section II-C and II-D. Finally, facial expression classifiers are trained separately on the regressed frontal data of each viewpoint cluster. During testing we estimate the closest viewpoint, apply the corresponding regression model and expression classifier. The main advantage over a standard pose-specific expression classifier, as described within Section II-A, lies in the robustness to erroneous viewpoint estimations. Our regressions to a common frontal view acts as a kind of regularization and smoothing, where discriminative expression features are preserved while differences caused by viewpoint variations are compensated. Our findings and results within our experiments in Section IV prove these assumptions.

Our contributions in this paper are as follows: First we propose a simple and straightforward pose-specific classification (PSC) approach that is able to outperform many state-of-the art approaches on two widely used FER datasets. Second we introduce pose specific linear regression (PSR) to frontal data which performs favorable compared to PSC

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

and significantly better than the state-of-the-art. Moreover, as an useful intermediate step, sparse coding with a global dictionary is proposed by applying K-SVD and OMP in training and testing respectively, for our linear regression of sparse features approach (PLRSF). Finally we replace the sparse encoding OMP step by a non-sparse ridge-regression (FPLRSF), which improves performance while being much more run-time efficient in comparison to OMP. To show the efficiency and robustness of our approach, an extensive investigation is provided on the BU3DFE and Multi-PIE datasets. We show that our approach outperforms state-of-the-art on both BU3DFE and Multi-PIE and that regression of frontal viewpoint features improves classification results and enforces robustness to feature noise and pose estimation failures.

A. Related works

Facial expression recognition (FER) could be broadly categorized into the three categories: 1) Geometric-based methods [5], [6], [7], 2) Appearance-based methods [8], [9], [10], [11], [12], and 3) hybrid methods which use both texture and shape information [13]. Recent work on geometric-based method includes regression-based of different mapping functions of geometric features which proposed by Rudovic et al. [6] and mapped 2D facial points from non-frontal to frontal view and then used new mapped points for expression recognition. Another related geometric-based facial expression recognition proposed by Hu et al. [7] which calculates the geometric 2D displacement of facial features between expressions and neural at the corresponding angles. They normalized extracted distances to zero mean and unit variance and used this information for classification. They also investigated different classifiers (linear Bayes, Quadratic Bayes, Parzen classifier and SVM) in their work. On the other hand, Some recent appearance-based methods address the problem of multi-view facial expression recognition where Zheng et al [8] proposed discriminant analysis theory (BDA/GMM) by optimizing upper bound of the Bayes error which is derived using Gaussian mixture model. They employed dense SIFT as feature descriptor and then transfer it into the regional covariance matrix (RCM) representation of facial image. Hesse et al. [9] evaluated different descriptors such as SIFT, LBP and DCT in their algorithm where the proposed multi-view facial expression recognition system extracts local appearance features around facial landmarks and then classifies them using ensemble SVM. Moore et al. [10] proposed a two-step algorithm which first uses a pose classifier to detect head orientation and then in the second stage, a pose-dependent expression classifier recognizes facial expressions. In another arbitrary view facial expression recognition model, Huang et al. [11] proposed a multi-view discriminative framework using multi-set canonical correlation analysis (MCCA) and the multi-view model theorem for facial expression recognition with arbitrary views. Their algorithm respects the intrinsic and discriminant structure of samples. They obtained discriminative information from facial expression images based on the discriminative neighbor preserving embedding (DNPE). Tariq et al. [12] proposed a multi-view facial expression recognition model using generic sparse coding feature which is state-of-the-art on BU3DFE. They applied sparse coding features of dense SIFT on the facial images in a three level spatial pyramid and then encode the local features into sparse codes. Their method could improve with linear regression and also a generic sparse coding model with K-SVD, of course suffers from time complexity. Moreover, to the best of our knowledge, hybrid methods have not applied for MFER yet. Nevertheless, recent papers are also used sparse representation for facial applications [14], [15], [16] which show that is a successful encoding for facial features where [14] proposed a sparse representation for face recognition. They explained that why sparsity could improve discrimination and how regression could be used to solve a classification problem. Timofte et al. [15] proposed an efficient sparse-based model and showed that regression transformation can improve the time complexity in both global and anchored neighborhood regression which are much faster than other related works. An important yet relatively unexplored approach is to employ pose specific linear regression which is challenging due to the partial linear regression. In previous method, regression-based approach ([6]) was obtained global transformation which is not as well as PSR (pose specific regression) or partial linear regression. Similarly, the approach that used sparse coding feature [12] did not profits the linear regression. Therefore, to address the above problems, this paper proposes to integrate them in a sequence as describe in the following.

II. Multi-view Facial Expression Recognition

In this section, the goal is to compensate not-available facial features which are important for MFER. We show that how we can perform non-available facial features using pose specific linear regression in order to perform more accurate facial expression recognition in arbitrary views. We also discuss the partial linear regression of sparse coding and how we can improve it in terms of running time.

A. Pose Specific Classification

Let X be a set of aligned vectorized features between frontal and non-frontal views which is extracted by appearance-based descriptors from the faces is described in section IV with size $(q \times 1)$. X_{θ_i} is a subset of facial features in X from viewing angle θ_i, where $X_{\theta_i} = [I_1^{\theta_i}, I_2^{\theta_i}, \ldots, I_N^{\theta_i}]$ is a matrix of size $(q \times N)$, and refers to the N vectorized facial features denoted by $I_k^{\theta_i} \in \mathbb{R}^{(q \times 1)}$. Note that I_k^0 and $I_k^{\theta_i}$ are vectorized features of the k^{th} facial expression image of the training data from the same person in different poses. Based on this, we define pairwise sets of training data, X_0 and X_{θ_i}, where the former is the set of frontal and latter is a set of correspondence non-frontal features. Pose Specific Classification (PSC) is a simple idea that split data X into the several subsets X_{θ_i} based on the viewpoints. It is basically divided in two steps: 1) Supervised splitting data into the smaller groups based on the viewpoints via classification, 2)

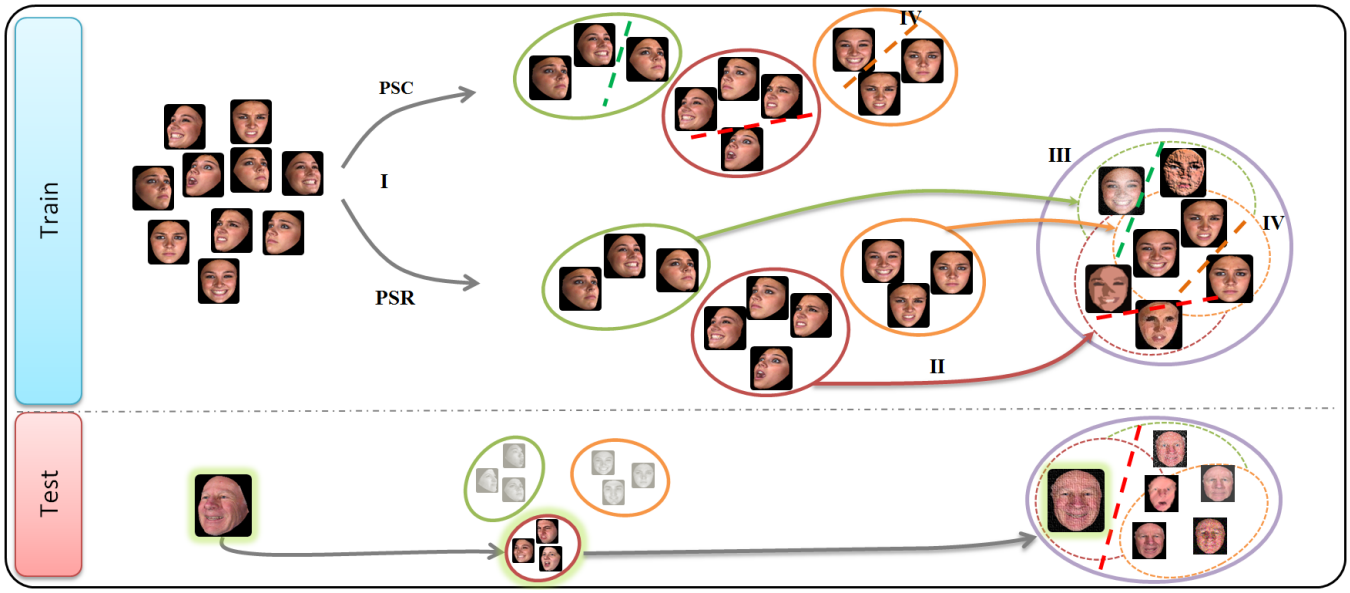

Fig. 1: Comparison between pose specific classification (PSC) and our approaches by using regression between pairwise sets of training data (PSR). I) Training features are clustered according to viewpoints. II) Projections between non-frontal and corresponding frontal samples are learned. III) Created approximated frontal samples, and IV) learn expression classifiers per viewpoint (PSC) or approximated frontal projections per viewpoint (PSR). For testing, best matching viewpoint is selected and corresponding projection and classifiers are applied.

Classifying each class for expression recognition. PSC can be improved using regression projection which is introduced in the following.

B. Pose Specific Regression

The linear regression is a successful idea for face recognition [17] we define a Pose Specific Regression (PSR) which is an efficient solution for MFER. We need to approximate regression projection between a pair of training data which can mathematically be formulated as:

$$\underset{P}{\operatorname{argmin}} \|X_0 - PX_\theta\| \tag{1}$$

Where the linear projection P can be estimated by Eq. 2, which is the closed form solution for Eq. 1.

$$P = X_0(X_\theta^T X_\theta)^{-1} X_\theta^T \tag{2}$$

$$\hat{X}_\theta = PX_\theta \tag{3}$$

Therefore, Eq. 3 is the linear regression which approximates frontal features $\hat{X}_\theta$ from non-frontal ones using projection P. A huge number of data with a lot of different properties affect the global linear regression P due to the different properties (e.g. viewpoints, expressions, gender, age, skin color, etc.) Therefore, the transformation between non-frontal and frontal is clearly not linear. Intuitively, the projection error could be decreased when we increase the number of projections reasonably; this means, splitting data into several meaningful parts and making correspondent piecewise projections leads to reduce the overall projection error compared to using one global projection. Therefore, linear regressions between specific non-frontal sets X_{θ_i} and the frontal set X_0 estimated for all subsets by:

$$P_i = X_0(X_{\theta_i}^T X_{\theta_i})^{-1} X_{\theta_i}^T \quad i = 1, 2, \ldots, M \tag{4}$$

$$\hat{X}_{\theta_i} = P_i X_{\theta_i} \tag{5}$$

Where $\hat{X}_{\theta_i}$ refers to approximation of frontal features by i^{th} linear regression from correspondence non-frontal features using projection P_i, M is also number of viewpoints. X can be classified using a supervised learning like PSC into the several meaningful subsets. Therefore, PSR is summarized as:

Step 1: Classifying features into the M subsets according to viewpoints.
Step 2: Approximating M piecewise projections by linear regression from each subset to frontal individually.
Step 3: Estimating projected facial features by Eq. 5.
Step 4: M Classifiers for facial expression recognition.

In addition, we propose pose specific linear regression with sparse codes namely Partial Linear Regression of Sparse Codes (PLRSF) which is more efficient than PSR in terms of memory usage in the following.

C. Partial Linear Regression of Sparse Features

PSR is an efficient approach for MFER but as it uses basic features, it is expensive in terms of memory usage due to the large feature vectors. Therefore, sparse representation is a successful alternative that could help us to improve our solution. We are interested in finding a reconstructive dictionary given the training features X by minimizing:

$$\arg\min_{S} \|X - DS\|_2^2 \quad s.t. \|s_i\|_0 \leq \Gamma \tag{6}$$

Where $D \in \mathbb{R}^{(q\times s)}$ is the dictionary, each column representing a code book vector, and $S \in \mathbb{R}^{(s\times N)}$ the matrix of encoding coefficients. Γ is the sparsity constraint factor, defining the maximum number of non-zero coefficients per sample. We apply K-SVD [18] as dictionary learning algorithm and orthogonal matching pursuit (OMP) [19] as an efficient way for solving the coding of new test samples, given a fixed dictionary. Again, we define M partial projections which approximate linear regression for each part of data; thus let S_0 be a set of sparse features of frontal facial expressions and $S_{\theta_1}, S_{\theta_2}, \ldots, S_{\theta_M}$ are M sets of sparse features of non-frontal facial expressions where all sets have the same number of samples. Eq. 4 and 5 could be rewritten for sparse features as:

$$P_i = S_0(S_{\theta_i}^T S_{\theta_i})^{-1} S_{\theta_i}^T \quad i = 1, 2, \ldots, M \tag{7}$$

$$\hat{S}_{\theta_i} = P_i S_{\theta_i} \tag{8}$$

Again P_i is i^{th} projection which has been estimated using correspondent sparse features. $\hat{S}_{\theta_i}$ defines the projected sparse codes, and the approximated features of the projected frontal view can be reconstructed using the global dictionary D with:

$$\hat{X}_{\theta_i} = D\hat{S_{\theta_i}} \tag{9}$$

We explain how regression model can improve the time complexity of PLRSF in next section.

D. Fast Partial Linear Regression of Sparse Features

As mentioned in previous section, OMP is used to find best encoding of a new test sample regarding to the dictionary D. However, OMP is an extension of Matching Pursuit (MP) with better results than standard MP but it needs more computation. It has been shown by previous research [14] that the sparsity constrain must not be needed during the reconstruction. Therefore, given a sparse code book created by Eq. 6, we can reformulate the solution of finding best encoding S_{θ_i} by replacing l0-norm for the coefficients, which is be called Ridge Regression [15] as:

$$\arg\min_{S_{\theta_i}} \|X_{\theta_i} - DS_{\theta_i}\|_2^2 + \lambda \|S_{\theta_i}\|_2 \tag{10}$$

Where D is global dictionary and X_{θ_i} is a set of input data regarding to the i^{th} pose. However, it eliminates the rules leading to sparsity but we are using l2-norm because of two clear reasons: first, to avoid over fitting during the regression; second, to stabilize projections specially when we now there are collinearity between the frontal and correspondence non-frontal features. Moreover, as mentioned before, the sparsity constrain must not be needed during the reconstruction due to using regression in our work. The parameter λ also allows us to detract the singularity problem. Eq. 10 is a ridge regression model and the solution is given by least square solution as:

$$\widetilde{S}_{\theta_i} = (D^T D + \lambda \mathbf{I})^{-1} D^T X_{\theta_i} \quad i = 1, 2, \ldots, M \tag{11}$$

$$\hat{S}_{\theta_i} = P_i \widetilde{S}_{\theta_i} \tag{12}$$

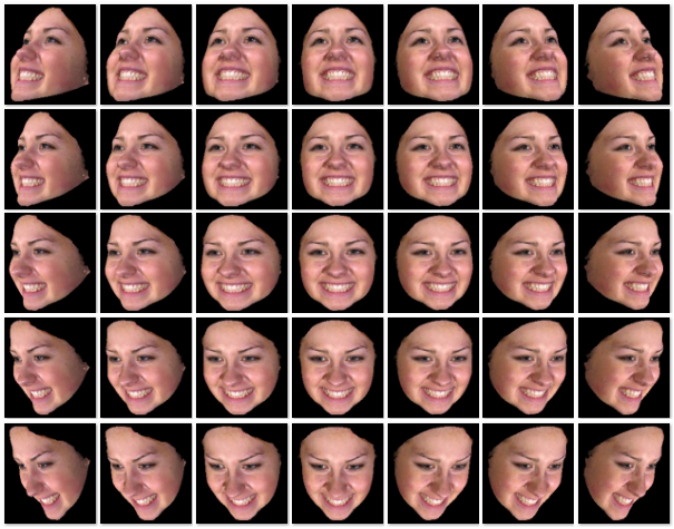

Fig. 2: Multi-view rendered faces of a sample from BU3DFE in 35 viewpoints (Protocol 1).

Fig. 3: Multi-view rendered faces of a sample from BU3DFE in 5 viewpoints (Protocol 2).

where $\widetilde{S}_{\theta_i}$ is the approximated representation of input data onto the dictionary proposed, instead of using OMP. Although, the dictionary can be compute offline but it is clear that input test samples should be transformed to sparse representation during the processing (online) therefore proposed algebraic computation in Eq. 11 is much faster than OMP which is an iterative algorithm and updates coefficients after every steps. Investigations of Fast Partial Linear Regression of Sparse Features (FPLRSF) approach are described in section IV.

III. MFER DATASETS

In order to demonstrate the performance of our model, we evaluate our approach on BU3DFE and Multi-PIE datasets, which are the most popular datasets for multi-view facial expression recognition. We have fully automatically extracted appearance features from BU3DFE and semi-automatic feature extraction for Multi-PIE dataset therefore it is possible to have some errors specially regarding to the head poses. Both BU3DFE and Multi-PIE are introduced briefly in the following.

A. BU3DFE dataset

BU3DFE is a publicly available dataset containing 3D scanned faces of 100 subjects with six basic expressions, namely anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU) in 4 levels of expression intensities plus one neutral (NE) which means, there are 25

Fig. 4: Multi-view semi-automatically cropped faces of a sample from Multi-PIE in 13 viewpoints.

samples per person and 2500 frontal samples totally. More details can be found in [20]. We rendered multiple views from the 3D faces in seven pan angles include: -45°, -30°, -15°, 0°, 15°, 30°, 45° and five tilt angles which are -30°, -15°, 0°, +15°, +30° to compare our results with the state-of-the-art [12], [8], [21], [22], In addition we generated views for 0°, 30°, 45°, 60° and 90° pan angles to compare our model with papers applied a different protocol [11], [23]. Fig. 2 and Fig. 3 illustrates multi-view rendered images of a sample 3D face model in first and second protocol where the first protocol contains 35 viewpoints and the second protocol has 5 viewpoints. Therefore, as there are 6 expressions for 100 subjects over the highest level of expression intensity in 35 viewpoints, we have 21000 samples in the first and 3000 samples in the second protocol.

B. Multi-PIE dataset

CMU Multi-PIE is a multi-purpose dataset in facial analysis containing 337 subjects taken across 15 different viewpoints in four recording sessions [24]. Pose variations are between -90° to 90° with an interval of 15° which means there are 13 different viewpoints for subjects and two other cameras are used to simulate a typical surveillance camera view. It contains five facial expressions: disgust (DI), scream (SC), smile (SM), squint (SQ) and surprise (SU) plus neutral. In order to evaluate our model, we first select all subjects where all of their expressions are available; therefore 145 subjects were selected. Then, we cropped facial regions using a semi-automatic algorithm into the size of 175 $\times$ 200 pixels. An example of cropped faces is shown in Fig. 4

IV. Experimental and Results

Our proposed MFER models can be separated into the three modules of a pipeline procedure that we follow on the standard evaluation scheme as:

a) Feature extraction: We apply a concatenation of HOG [25] and LBP [26] features. HOG is a gradient-based descriptor and it is stable on illumination variation. Moreover, it is a fast descriptor in comparison to the SIFT and LDP (Local Directional Pattern) due to the simple computation. On the other hand, LBP is a common texture-based descriptor which is used widely in face analysis. It has been shown that a concatenation of HOG and LBP can improve human detection performance by [27]. In our experiments, the extracted features are considered as feature vectors for every facial image in every viewpoint without any concern about head pose or expressions where the cell size considered for both HOG and LBP is 25 pixels, therefore, the overall dimensionality is 5480 in total that first 2232 dimensions are computed by HOG and the rest 3248 dimensions via LBP.

b) Projections: Linear regression projection in Eq. 2 estimates projection from non-frontal to frontal view. It gives us the ability of approximating non-available features from related available features. Therefore, we employ a pairwise sets of basic features in training data to make projections for PSR which is explained in section II-B and a pairwise sets of sparse features to make projections for PLRSF and FPLRSF which we described in section II-C and II-D.

c) Classification: All samples are projected to the frontal, as described in section II-B, and linear SVM [28] is applied for classifications during all experiments. Moreover, we consider 5-fold cross validation for both BU3DFE and Multi-PIE datasets where the highest level of expression intensity on BU3DFE is employed in our experiments. All evaluations are performed on a machine with the same resources supported by an Intel 2.53 GHz dual core and 4 GB RAM with a 64-bit operating system. The results are given in the following.

A. Parameters and settings

We propose three regression based approaches PSR, PLRSF and FPLRSF introduced in II-B, II-C and II-D respectively. Parameters like dictionary size and sparsity in K-SVD are evaluated, where the best result is achieved by a dictionary size of 200 with sparsity 50 (75% of dictionary elements are used for encoding). The performance of PLRSF and FPLRSF is very close to each other, although, FPLRSF is slightly better than PLRSF in accuracy but significantly better in running time. FPLRSF is the best method for multi-view facial expression recognition concerning time complexity, due to the tremendous reduction of feature dimensionality and the fast ridge regression step, while having better result than state-of-the-art on BU3DFE. Moreover, PSR results on both datasets show that is the best method concerning the accuracy. A detailed comparison between proposed methods in both accuracy and time complexity is shown in Table I. As can be seen, PSR has highest accuracy but it is not as fast as FPLRSF whereas FPLRSF has the best running time and outperforms the state-of-the-art.

B. Experimental Results

In this section, we evaluate the performance of our approaches (PLRSF, FPLRSF and PSR) on two protocols of BU3DFE and one protocol of Multi-PIE datasets. It is important to note that images in BU3DFE are 3D scans created from real faces and therefore contain challenging issues like: age, ethnicity, skin, gender, personality, etc.

Dataset / Method	BU3DFE- Protocol 1		BU3DFE- Protocol 2		Multi-PIE	
	Accuracy	Time (sec)*	Accuracy	Time (sec)*	Accuracy	Time (sec)*
PSC	77.66	532	76.36	73	80.94	265
PSR	78.04	880	77.87	85	81.96	282
PLRSF	76.04	569	75.16	77	74.61	240
FPLRSF	77.61	393	75.63	49	75.20	162

* Running time is for all test samples (4200 samples for BU3DFE and 1885 samples for MultiPIE)

TABLE I: Accuracy and time complexity of proposed methods on BU3DFE and Multi-PIE datasets. The PSR outperforms other methods and FPLRSF has slightly better performance than PLRSF while it is significantly faster.

For the first protocol, which is BU3DFE-P1 (35 viewpoints), our overall accuracy rate for PLRSF, FPLRSF and PSR are 76.04%, 77.61% and 78.04% where PSR and FPLRSF are the best on accuracy and running time respectively. FPLRSF is not only faster than PSR and PSC but also faster than PLRSF in this protocol. Performing comparison of our approaches over the variations in pan and tilt, illustrated in Fig. 5, note that the results in the Fig. 5 (a) are averaged across corresponding tilt angles and the results in Fig. 5 (b) are averaged across corresponding pan angles. Some related works evaluated their results on the second protocol of BU3DFE. The overall performance of PLRSF, FPLRSF and PSR on this protocol are 75.16%, 75.63% and 77.87% respectively, where their running time for 4200 test samples are about 77, 49 and 85 seconds respectively. Fig. 5 (c) shows comparison between proposed approaches on the second protocol of BU3DFE (5 viewpoints). Multi-PIE dataset is our third case study which is a multi-purpose popular dataset used also for MFER. We have evaluated our approaches on this dataset. The overall performances and time complexities of the proposed methods are provided in Table I where PLRSF achieved 74.61% accuracy rate in 240 seconds, FPLRSF obtained 75.20% accuracy in 162 seconds and PSR in 282 seconds with 81.96% accuracy outperforms not only the other proposed approaches but also the state-of-the-art about 5%. Reported running time on this dataset is for 1885 test samples. Again PSR and FPLRSF are clearly efficient approaches in case of accuracy and time complexity for multi-view facial expression recognition; Fig. 6 shows the performance of proposed approaches across 13 viewpoints of Multi-PIE dataset. With the above comparisons, we can see that our regression approaches are elegant and successful ideas for MFER. Moreover, the main important points of PSR are its applicability, simplicity and high accuracy which are desirable for real applications.

C. Comparison with the state-of-the-art

In this section, we compare our approach with the state-of-the-art on both protocols of BU3DFE and Multi-PIE. Table II illustrates that PSR outperforms the state-of-the-art in all protocols of BU3DFE and Multi-PIE. FPLRSF also outperforms other methods on both protocol of BU3DFE. However, there is no information about the time complexity of other related works but while the state-of-the-art used generic sparse coding [12], it could be compared with our PLRSF model which applies OMP. We showed that FPLRSF is not only slightly better than the PLRSF in terms of accuracy but also significantly faster than sparse-based methods that use OMP to represent facial features for multi-view facial expression recognition. Moreover, Moore et al. [10] proposed an approach similar to the PSC model based on a new descriptor (LGBP) and reported 80.17% accuracy rate on Multi-PIE dataset with only 7 viewpoints which is still less than our regression-based PSR approach with 81.96%, tested on 13 viewpoints.

V. Invariance analysis and Robustness

In this section we investigate our proposed approaches with respect to noisy data where three kinds of evaluations are performed: 1) Influence of occlusion, 2) Reducing training data and 3) Head pose analysis. All of these three evaluations are important challenges for which we provide the details and results in the following:

A. Evaluation of occlusion presence

In this experiment, we have randomly included a white square block in different sizes of 40×40, 50×50 and 60×60 pixel where the original face image is in size of 200×220. As our PSR model is based on the regression transformation and it can virtually perform unavailable/invisible features, the performance is not influenced much; however, sparse coding based methods cannot handle this amount of noise. Figure 7 shows some samples of occluded faces in different size of white block and its random position. Table III summarizes the results of the proposed approaches on first protocol of BU3DFE (35 viewpoints). All methods decrease slightly, while interestingly the sparse representations are more influenced by occlusion, but PSR again performs the best.

B. Evaluation of reducing training data

Without any doubt, a large number of training data can increases the complexity and classification time, therefore, similar accuracy using tiny training data is desirable. In this experiment we evaluate our approaches by deleting some viewpoints within training data in the first protocol of BU3DFE, while testing on all 35 viewpoints. This reduces the number of projections and classifiers available for PSR

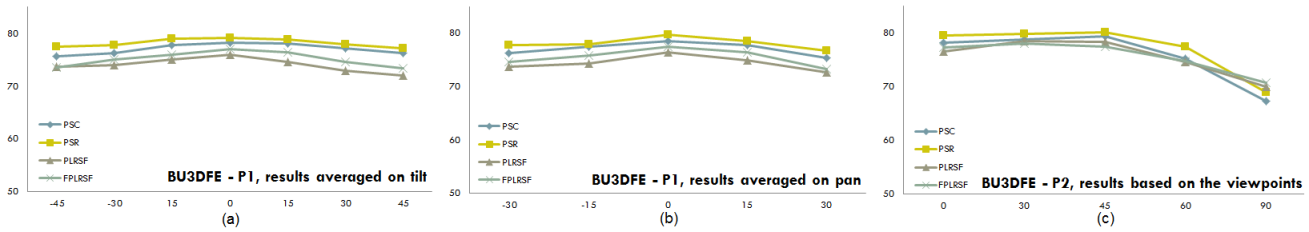

Fig. 5: Proposed methods (PSR, PSC, PLRSF and FPLRSF) performance on the first protocol of BU3DFE: (a) averaged on pan (b) averaged on tilt, (c) performance on the second protocol of BU3DFE.

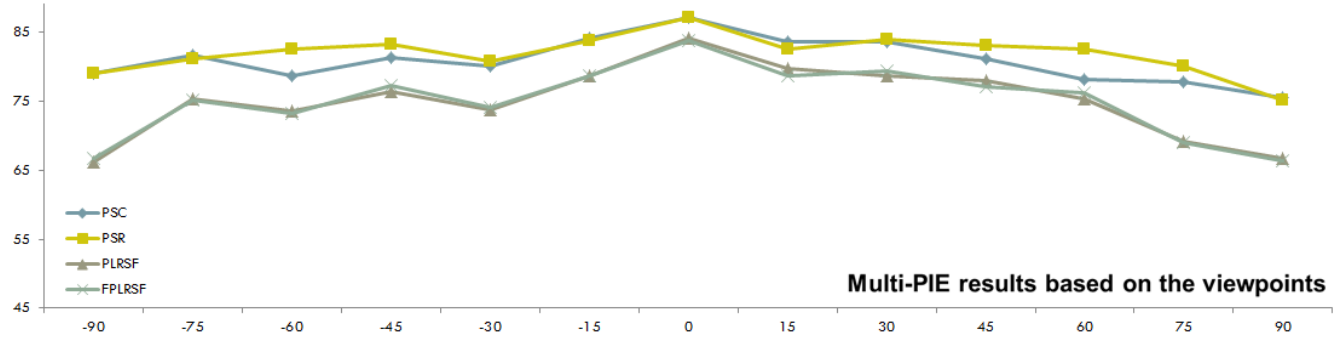

Fig. 6: Proposed methods (PSR, PSC, PLRSF and FPLRSF) performance on the Multi-PIE dataset based on the viewpoints.

Method	Dataset	Accuracy
Tariq et al. [21]	BU3DFE-P1	76.34
Tariq et al. [12]	BU3DFE-P1	76.10
Zheng et al. [8]	BU3DFE-P1	68.20
Tang et al. [22]	BU3DFE-P1	75.30
FPLRSF [ours]	BU3DFE-P1	**77.61**
PSR [ours]	BU3DFE-P1	**78.04**
Huang et al. [11]	BU3DFE-P2	72.47
Hu et al. [23]	BU3DFE-P2	74.46
FPLRSF [ours]	BU3DFE-P2	**75.63**
PSR [ours]	BU3DFE-P2	**77.87**
Huang et al. [11]	Multi-PIE	76.83
FPLRSF [ours]	Multi-PIE	**75.20**
PSR [ours]	Multi-PIE	**81.96**

TABLE II: Comparison of proposed PSR and FPLRSF with the state-of-the-art.

and FPLRSF, and proves the robustness and generalization capabilities of our proposed ideas. For this purpose we have ignored (a) two columns, (b) two rows and (c) two columns plus two rows of viewpoints in this protocol which means we have ignored 10 viewpoints (i.e. 4800 samples) in task (a), 14 viewpoints (i.e. 6720 samples) in task (b) and finally 20 viewpoints or 57.1% of training data in task (c). Table IV shows the results of our approaches with reducing training data. As can be seen, in all tasks PSR is more stable than other methods and if we reduce 40% of training

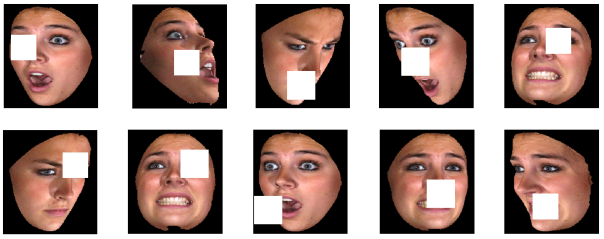

Fig. 7: Occluded face samples in different viewpoints.

data the expression recognition is 76.50%, which is still better than the state-of-the-art. The missing training data is better compensated by our projections to a common frontal representation than with PSC.

C. Evaluation of head poses estimation error

As we process features based on the viewpoints, robustness to erroneous viewpoint estimations is critical for robust results. The experimental results in Table I are generated by an automatic viewpoint classification and therefore already included a small amount of head pose errors. Nevertheless, in this experiment, we artificially create two levels of pose estimation noise, which means during testing we randomly replace each viewpoint estimation by one of its neighboring ones, 15 or 30 degrees farther (see Fig. 2), therefore taking wrong classifiers in PSC, and wrong projections and classifiers in PSR, PLRSF and FPLRSF. Table V shows averaged results over 8 runs of selecting wrong neighboring poses. It can be seen that all our regression-based approaches are almost stable with respect to pose errors as expected due to the projection to a common frontal view. The PSC approach decreases as it is trained purely on view specific data.

Block size / Method	Without occlusion	40×40	50×50	60×60
PSC	77.66	69.15	65.30	61.50
PSR	78.04	71.10	67.44	63.32
PLRSF	76.04	61.65	56.64	50.66
FPLRSF	77.61	61.83	56.19	50.46

TABLE III: Occlusion evaluation of proposed approaches on BU3DFE-P1, three different block size 40×40, 50×50 and 60×60 is applied for robustness evaluation.

Reduced (%) / Method	Without ignoring	28.5% (10 vp)	40% (14 vp)	57.14% (20 vp)
PSC	77.66	74.88	74.68	72.81
PSR	78.04	76.76	76.50	75.10
PLRSF	76.04	70.33	70.05	68.22
FPLRSF	77.61	73.07	72.20	70.85

TABLE IV: Influence of reducing training data on proposed approaches evaluated on BU3DFE-P1.

Noise / Method	Ground truth	Proposed setting	First level	Second level
PSC	79.52	77.66	66.33	50.48
PSR	80.04	78.04	77.10	74.18
PLRSF	78.33	76.04	72.94	72.66
FPLRSF	79.86	77.61	74.20	73.03

TABLE V: Head poses error evaluation of proposed approaches on BU3DFE-P1 investigated on one and two wrong viewpoints farther.

VI. Conclusion & Future work

In this paper, we introduced three linear regression based approaches: Pose Specific Regression (PSR), Partial Linear Regression of Sparse Features (PLRSF) and Fast Partial Linear Regression of Sparse Features (FPLRSF) for multi-view facial expression recognition. Our approaches are capable to estimate non-available/invisible information by using projections which are learned with regression-based models. We have shown that the proposed PSR and FPLRSF models for multi-view facial expression recognition outperform not only PLRSF but also the state-of-the-art approaches. Moreover, FPLRSF time complexity is significantly better than other methods and it can be applied in real-world applications. We have also shown that our regression-based approaches are almost stable with small occlusions; reduce training data or severe head poses error. Investigation of non-linear projections for approximation of non-frontal to frontal views would be also a possible direction for future works.

VII. Acknowledgment

Mahdi Jampour would like to thanks Iran Ministry of Science, Research and Technology for scholarship #89100016. This work was supported by the Austrian Research Promotion Agency (FFG) projects DIANGO (840824) and Vision+ (836630).

References

[1] Zeng, Z. and Pantic, M. and Roisman, G.I. and Huang, T.S., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", *PAMI*, vol. 31, 2009, pp 39-58.

[2] Rudovic, O. and Pavlovic, V. and Pantic, M., "Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation", *CVPR*, 2012, pp 2634-2641.

[3] Weifeng Liu and Caifeng Song and Yanjiang Wang, "Facial expression recognition based on discriminative dictionary learning", *ICPR*, 2012, pp 1839-1842.

[4] Xu, Liefei and Mordohai, Philippos, "Automatic Facial Expression Recognition using Bags of Motion Words", *BMVC*, 2010, pp 13.1-13.13.

[5] Taheri, Sima and Turaga, Pavan K. and Chellappa, Rama, "Towards view-invariant expression analysis using analytic shape manifolds", *FG*, 2011. pp 306-313.

[6] Rudovic, O. and Patras, I. and Pantic, M., "Regression-Based Multi-view Facial Expression Recognition", *ICPR*, 2010, pp 4121-4124.

[7] Yuxiao Hu and Zeng, Z. and Lijun Yin and Xiaozhou Wei and Tu, J. and Huang, T.S., "A study of non-frontal-view facial expressions recognition", *ICPR*, 2008, pp 1-4.

[8] Zheng, Wenming and Tang, Hao and Lin, Zhouchen and Huang, ThomasS., "Emotion Recognition from Arbitrary View Facial Images", *ECCV*, 2010, pp 490-503.

[9] Hesse, N. and Gehrig, T. and Hua Gao and Ekenel, H.K., "Multi-view facial expression recognition using local appearance features", *ICPR*, 2012, pp 3533-3536.

[10] Moore, Stephen and Bowden, Richard, "Multi-View Pose and Facial Expression Recognition", *BMVC*, 2010.

[11] X. Huang and G. Zhao and M. Pietikinen, "Emotion recognition from facial images with arbitrary views", *BMVC*, 2013, pp 76.1-76.11.

[12] Tariq, Usman and Yang, Jianchao and Huang, ThomasS., "Multi-view Facial Expression Recognition Analysis with Generic Sparse Coding Feature", *ECCV*, 2012, pp 578-588.

[13] I. Kotsia and Stefanos Zafeiriou and Ioannis Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition", *Pattern Recognition*, vol 41, 2008.

[14] Zhang, D. and Meng Yang and Xiangchu Feng, "Sparse representation or collaborative representation: Which helps face recognition?", *ICCV*, 2011, pp 471-478.

[15] Timofte, R. and De, V. and Van Gool, L., "Anchored Neighborhood Regression for Fast Example-Based Super-Resolution", *ICCV*, 2013, pp 1920-1927.

[16] Abdolali, M. and Rahmati, M., "Facial expression recognition using sparse coding", *MVIP*, 2013, pp 150-153.

[17] Naseem, I. and Togneri, R. and Bennamoun, M., "Linear Regression for Face Recognition", *PAMI*, vol 32, 2010, pp 2106-2112.

[18] Aharon, M. and Elad, M. and Bruckstein, A., "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", *Signal Processing, IEEE Transaction on*, vol 54, 2006, pp 4311-4322.

[19] Tropp, J.A. and Gilbert, A.C., "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit", *Information Theory, IEEE Transaction on*, vol 53, 2007, pp 4655-4666.

[20] Lijun Yin and Xiaozhou Wei and Yi Sun and Jun Wang and Rosato, M.J., "A 3D facial expression database for facial behavior research", *FG*, 2006, pp 211-216.

[21] Tariq, U. and Jianchao Yang and Huang, T.S., "Maximum margin GMM learning for facial expression recognition", *FG*, 2013, pp 1-6.

[22] Hao Tang and Hasegawa-Johnson, M. and Huang, T., "Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors", *ICME*, 2010, pp 1202-1207.

[23] Yuxiao Hu and Zeng, Z. and Lijun Yin and Xiaozhou Wei and Xi Zhou and Huang, T.S., "Multi-view facial expression recognition", *FG*, 2008, pp 1-6.

[24] Gross, R., Matthews, I., Cohn, J. F., Kanade, T., Baker, S., "Multi-PIE", *Image and Vision Computing*, vol 28, 2010, pp 807-813.

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

[26] T. Ojala, M. Pietikinen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions", *ICPR* 1994, pp 582 - 585.

[27] Xiaoyu Wang, Han T.X., Shuicheng Yan, An HOG-LBP human detector with partial occlusion handling, *ICCV*, 2009, pp 32-39.

[28] Chang, Chih-Chung and Lin, Chih-Jen, "LIBSVM: A library for support vector machines", *ACM T. Intell. Syst. Technol.*, vol 2, 2011.