Image and Vision Computing 58 (2017) 38–46

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Pose-specific non-linear mappings in feature space towards multiview facial expression recognition☆

Mahdi Jampour*, Vincent Lepetit, Thomas Mauthner, Horst Bischof

Graz University of Technology, Graz, Austria

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 31 March 2016
Accepted 5 May 2016
Available online 13 May 2016

Keywords:
Non-frontal facial expression recognition
Sparse coding
Non-linear transformation
Robust arbitrary view facial expression recognition

ABSTRACT

We introduce a novel approach to recognizing facial expressions over a large range of head poses. Like previous approaches, we map the features extracted from the input image to the corresponding features of the face with the same facial expression but seen in a frontal view. This allows us to collect all training data into a common referential and therefore benefit from more data to learn to recognize the expressions. However, by contrast with such previous work, our mapping depends on the pose of the input image: We first estimate the pose of the head in the input image, and then apply the mapping specifically learned for this pose. The features after mapping are therefore much more reliable for recognition purposes. In addition, we introduce a non-linear form for the mapping of the features, and we show that it is robust to occasional mistakes made by the pose estimation stage. We evaluate our approach with extensive experiments on two protocols of the BU3DFE and Multi-PIE datasets, and show that it outperforms the state-of-the-art on both datasets.

1. Introduction

With the ever growing importance of Human–Computer Interfaces, facial expression recognition (FER) is one of the important challenges of computer vision. Even if recognition in frontal views, either based on appearance or geometry already performs very well [1,2,3,4,5], having a frontal view is an unrealistic assumption for real-world applications, and multiview facial expression recognition (MFER) is still very challenging as facial features important for recognition are likely to be hidden.

To date, the most successful methods [6,7,8] map facial features extracted from non-frontal views to the corresponding features in the frontal view: by mapping all the available training data to a common referential one can generalize better. However, Ref. [6] used the same mapping regardless of the pose of the head; Ref. [7] proposed a complex method that relies on a time-consuming optimization process.

We therefore propose to learn several mappings, each specific to a pose from a discrete set of possible poses of the face. To know which mapping to use for a new input image, we simply rely on another classifier to predict the pose of the face. Since the mappings are adapted to the pose of the input face, this approach yields significantly better results than using a single mapping. Fig. 1 illustrates our claim that local mappings can provide more reasonable results than a global one. We explore two different forms to perform the mappings in feature space: We first consider a simple linear mapping, and we introduce non-linear mappings based on Taylor expansion.

We evaluate this approach with extensive experiments on two protocols of the BU3DFE and Multi-PIE datasets. Our evaluation shows that simple linear transformations for the mappings are enough for our approach to outperform the state-of-the-art on both datasets. When non-linear mappings are used, we improve the results even further.

This paper is an extension of our previous work [9], where we introduced pose specific linear mappings. Here we introduce non-linear mappings, and provide an extensive evaluation, and an in-depth discussion.

In the remainder of the paper, we first discuss related work. We then explain how we predict the pose of the faces in the input images. After that, we describe the different mappings and the facial expression recognition, and finally we provide extensive evaluation of our approach.

2. Related work

Facial expression recognition has many exciting and various applications, including Human–Computer Interaction (HCI), psychology, games, children education, etc., and the literature is

☆ This paper has been recommended for acceptance by Vitomir Štruc.

* Corresponding author.
E-mail addresses: jampour@icg.tugraz.at (M. Jampour), lepetit@icg.tugraz.at (V. Lepetit), mauthner@icg.tugraz.at (T. Mauthner), bischof@icg.tugraz.at (H. Bischof).

http://dx.doi.org/10.1016/j.imavis.2016.05.002
0262-8856/

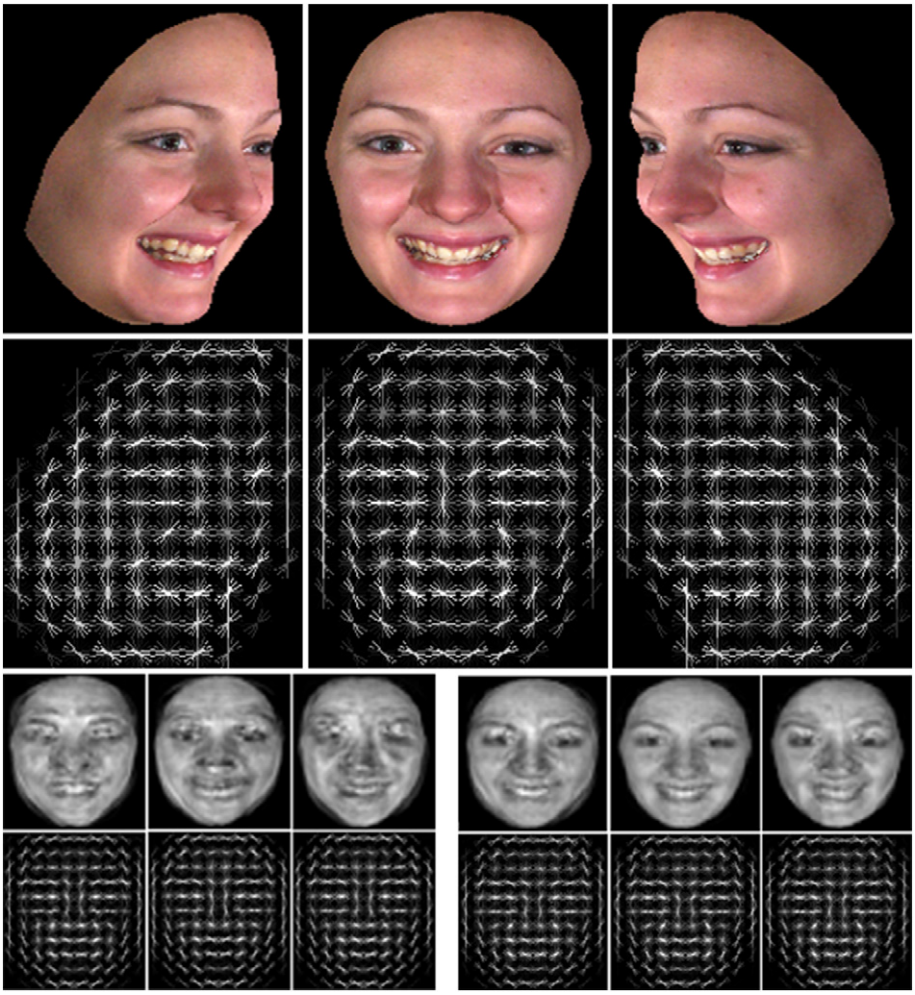

Fig. 1. Comparison between global and pairwise transformations: (a) three samples from different viewpoints and their HOG features; their reconstructions using (b) a global mapping and (c) a pairwise mapping. Note that the bitmap images are provided for visualization only. We use a concatenation of HOG and LBP features instead of raw features for recognition.

very broad. Some approaches explicitly consider the main facial components, mouths, eyes, etc. and extract them from the input images. Most of the geometric approaches use Facial Action Units (AUs) which are part of Facial Action Coding Systems (FACS) introduced by Ekman et al. [10] to recognize expressions. AUs are observable components of facial movement that acted by a group of facial muscles [11,12,13,14]. Other approaches [15,16,17,18,19] rely only on texture information and use local descriptors such as SIFT, Gabor, HOG, Pyramid HOG (PHOG), LBP, Pyramid LBP (PLBP). Some hybrid methods exploit both geometric and texture information [6,20,21,22].

Recently, researchers turned to the multiview facial expression recognition problem, where the face is not necessary frontal. This problem is of course much more challenging than recognizing facial expression from frontal, as the perspective can deform the expressions, or even hide some features.

One of the first attempts for non-frontal facial expression recognition was Ref. [23], which proposed to consider the 2D facial feature displacements of 38 facial landmarks as features to perform the recognition and they investigated various classifiers on the BU3DFE dataset. Rudovic et al. [24] proposed a mapping model between the facial points from non-frontal views to a frontal referential. They used Coupled Scaled Gaussian Process Regression (CSGPR) for the mapping, and multiclass LDA for estimating the head poses.

Other approaches rely on appearance only. For instance, Zheng [7] proposed a Group Sparse Reduced-Rank Regression based method (GSRRR). Sparse SIFT descriptors and LBP features are extracted. Feature vectors are synthesized using kernel reduced-rank regression to obtain feature vectors corresponding to different facial views. The facial expression category is finally obtained by solving GSRRR optimization problem. This approach obtains very good results, however it is very computationally expensive. The same author proposed a discriminant analysis theory (BDA/GMM) [15] which optimizes the upper bound of the Bayes error derived by Gaussian mixture model but it only outperformed the baseline.

Hesse et al. [16] evaluated various descriptors such as SIFT, LBP and DCT extracted around facial landmarks and classify then using ensemble SVM. They showed that DCT based features yield better performance than SIFT or LBP. Another approach proposed by Moore and Bowden [25] first estimates the pose orientation directly from the image and then applies a pose-dependent classifier to recognize the facial expressions. This method is simple and fast as it is based on the LBP features. However, it is sensitive to occlusion, and more importantly to the low number of training data for each specific head pose.

Huang et al. [18] proposed a discriminative framework based on multi-set canonical correlation analysis (MCCA) and proposed a multiview model theorem for facial expression recognition with

arbitrary views. Their method respects the intrinsic and discriminant structure of samples. They obtained discriminative information from facial expression images based on the Discriminative Neighbor Preserving Embedding (DNPE). In practice we found that, as CCA maps both source (e.g. X) and target (e.g. Y) sets into a new intermediate space like U and V which are more related to each other, their discrimination regarding to the expressions will decrease. Therefore, the overall recognition rate may not improve or even lessen.

On the other hand, sparse coding has been used intensively for face recognition [26], and facial expression recognition [9,19,27], and has been shown to be a successful encoding technique. Zhang et al. [26] explained why sparsity could improve discrimination and how regression could be used to solve a classification problem. Timofte et al. [28] proposed an efficient sparse-based model and showed a significant speed up. Ref. [19] improved an existing facial expression recognition model using generic sparse coding features. 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 the sparse codes to make the possibility of multiview processing.

Our method is related to these techniques as it benefits from sparse coding applied to image features. However we propose to learn several mappings, each specific to a pose from a discrete set of possible poses of the face. We show that this allows us to circumvent the limited amount of training data while performing a discriminative recognition of the facial expression regardless of the head pose.

3. Facial expression recognition in frontal views

In this section, we briefly describe our method for facial expression recognition in frontal views. We used Chehra face detector [29] which is a fully-automatic real-time face and landmark detector to obtain face region. We then extract the appearance and semi-geometric information of the faces using a concatenation of HOG [30] and LBP [31] descriptors, which we fed to a classification method after dimensionality reduction. We will extend this method to multiple viewpoints in Section 4.

3.1. Image feature extraction and concatenation

We concatenate HOG [30] and LBP [31] feature vectors for each face images as they are most popular and successful feature descriptors in face analysis. HOG is an image descriptor based on local histograms of oriented gradients and LBP is a descriptor based on pixel intensities. Let X be a set of aligned vectorized features with size $(q \times L)$ where q is the dimension of the feature vectors and L the number of samples. X_θ is a subset of facial features in the 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 $I_k^{\theta_i} \in IR^{(q\times 1)}$. Note that I_k^0 and $I_k^{\theta_i}$ are vectorized features of the kth facial expression image of the training data from the same person in different poses. Subsequently X_0 is a set of vectorized features related to the frontal faces. Based on this, for a given image, we therefore obtain a feature vector $x = [h^T; l^T]^T$. x is a very long vector: In our implementation its size is 5480, where the first 2232 dimensions come from HOG, and the remaining 3248 from LBP. We explain below how we decrease the dimensionality of our feature vectors by means of sparse coding.

3.2. Compact feature vectors based on sparse coding

Sparse representation approximates an input vector by a sparse linear combination of codebooks, or *atoms*, based on a compact dictionary D. It has more flexibility than Principal Component Analysis (PCA), for example because it does not impose that the codebooks are orthogonal. While sparse representation is a successful method for recognition purposes, the solutions relying on basic features can be expensive in terms of memory usage because it involves large feature vectors. To improve the memory usage, we are interested in finding a reconstructive dictionary given training feature vectors $\{x_i\}_i$ by minimizing

$$\min_{D,\,S} \|X - DS\|_2^2 \quad \text{such that} \quad \forall i \quad \|s_i\|_0 \le \Gamma, \tag{1}$$

where $D \in IR^{q\times s}$ is the dictionary, each column of it corresponding to a codebook vector, $X = [..x_i..]$ is a matrix whose columns are the feature vectors, and $S \in IR^{s\times N}$ is a matrix of encoding coefficients. Γ is the sparsity constraint factor, defining the maximum number of non-zero coefficients in each column s_i of S. We apply K-SVD [32] as the dictionary learning algorithm and orthogonal matching pursuit (OMP) [33] as an efficient way to solve the sparse coding S for X given a fixed dictionary D. In practice, we use a dictionary made of 200 codebooks, and $\Gamma = 50$.

Given a feature vector x computed for an image, we could obtain a corresponding vector s by applying OMP again. However, OMP requires heavy computations to provide a sparse representation, while previous researches [26,34] show that the sparsity constrain is not be needed during the reconstruction for classification purposes.

Therefore, given the codebook D created by Eq. (1), and a feature vector x, we reformulate the solution of finding the encoding s as Ref. [28] to prepare a fast version of sparse coding:

$$\arg\min_s \|x - Ds\|_2^2 + \lambda \|s\|_2^2 . \tag{2}$$

The second term adds regularization to the retrieved vector s. s can be computed efficiently in closed-form as:

$$s = (D^T D + \lambda \mathbf{I})^{-1} D^T x. \tag{3}$$

3.3. Frontal facial expression recognition

Finally, given an input image of a frontal face, we apply a multiclass linear SVM [35] on the corresponding x or s feature vector. The meta-parameters of the SVM were obtained by k-fold ($k = 5$) cross-validation of the training part of the dataset. In other words, we had to use different datasets for evaluations. All non-frontal data are mapped to the frontal view using our proposed approach, which is explained in detail by the next section. The "frontalized" data is employed for facial expression recognition with a basic frontal facial expression recognition system. Fig. 2 compares our proposed mapping approach to the simple pose-specific approach without any mapping. The performances are reported in the Experimental results section. In the next section, we explain how we extend the approach described in this section to multiple viewpoints.

4. Multiview facial expression recognition

In this section, we propose to learn a mapping specific to the pose of the input image. To select which mapping we need to apply at runtime, we also train a classifier to recognize the head pose in input images. In the following, we first explain how we estimate the head pose and then explain how we compute the mapping functions. We also introduce kernel-based non-linear transformations to capture complex mappings.

4.1. Splitting data and pose estimation

Splitting data into the several smaller subsets is a convenient idea to improve the recognition accuracy. We learn our system to classify input test sample based on the head poses as we have the

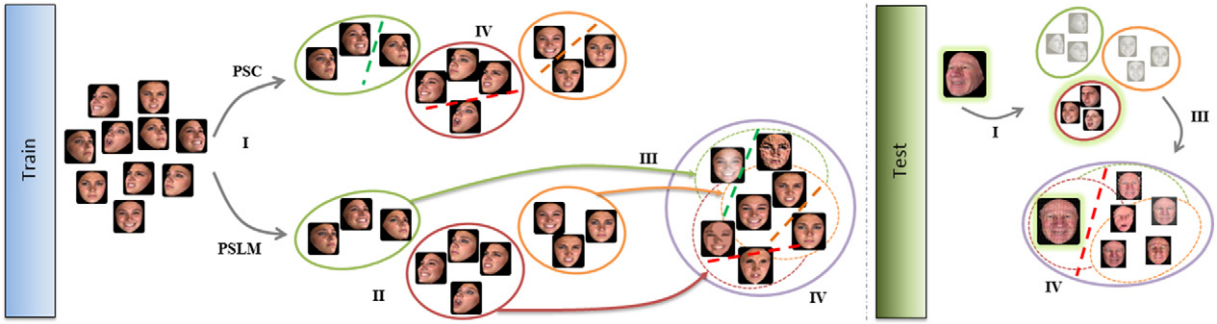

Fig. 2. Comparison between Pose Specific Classification (PSC) and Pose specific linear mapping (PSLM). I) Splitting data into the subsets based on the viewpoints by means of supervised classification. II) Learn mapping functions by transforming non-frontal subsets to the frontal. III) Map to the frontal view, and IV) expression classification. For testing, we first estimate the head pose of the input sample then map it to the frontal and finally classify for expressions.

information of head poses during the training. Although, we can use both supervised or unsupervised methods for data assortment, we choose supervised linear SVM as a well-known classification method. In other words, we use all training faces with re-labeling them by means of viewpoints information to learn our system for head poses detection. Therefore, a new unseen test sample must be first classified into the correct subset and then we employ correspondence mapping model for transferring into the frontal. Thus, we define $C_i \forall i \in [1 : k]$ class labels related to the k viewpoints for different subsets of training data to learn our head poses classifier.

4.2. Pose specific linear mapping (PSLM)

In this subsection, we introduce a linear mapping that maps the features extracted from an image to a non-frontal view to an approximation of the features that would have been extracted if the view was frontal. This mapping can be computed by solving:

$$\arg\min_{M} \|X_0 - MX_{0^c}\|, \tag{4}$$

where X_0 is the matrix of the feature vectors for the training images of frontal views, and X_{0^c} the matrix of the feature vectors for the training images of non-frontal views. M can actually be computed in closed-form using the well-known formula:

$$M = X_0(X_{0^c}^T X_{0^c})^{-1} X_{0^c}^T. \tag{5}$$

This approach was used for example in Refs. [36,37]. However it seems clear that it is better to have one transformation specific for the pose of the face in the input image.

We therefore compute a linear mapping M_θ specific for each pose θ of the input view. M_θ can be estimated during an offline stage using the same formula as before:

$$M_\theta = X_0(X_\theta^T X_\theta)^{-1} X_\theta^T, \tag{6}$$

where X_θ is the matrix of the feature vectors for the training images with the head that is under pose θ.

Given a feature vector x_θ computed for an image of a face seen under pose θ, we can predict the corresponding feature vector $\tilde{x}_{\theta\to 0}$ as if the face was under a frontal view in the image by computing:

$$\tilde{x}_{\theta\to 0} = M_\theta x_\theta.$$

Similarly, we can compute linear mappings T_θ^S that can be applied to sparse feature vectors s_θ.

4.3. Kernel-based pose specific non-linear mapping (KPSNM)

A linear mapping, even a pose-specific one as described in the previous subsection, can capture variations from the non-frontal view to the frontal one only in a very limited way. We therefore introduce a more complex mapping, based on polynomial kernels. Let us first introduce h_n, a function that applies to a feature vector such that:

$$h_n(x) = \begin{bmatrix} x^{\cdot 0} \\ x^{\cdot 1} \\ \vdots \\ x^{\cdot n} \end{bmatrix}, \tag{7}$$

where $x^{\cdot i}$ is the element-wise exponent i applied to each element of x. Let $h_n(X_\theta)$ be the matrix whose columns are the results of $h_n(.)$ applied to the feature vectors for the training images the head is seen under pose θ.

We can now compute a new mapping M_θ^h that applied to polynomial kernels of the feature vectors:

$$M_\theta^h = X_0(h_n(X_\theta)^T h_n(X_\theta) + \lambda I)^{-1} h_n(X_\theta)^T. \tag{8}$$

The term λI is required for regularization, otherwise the system would be under-constrained. In practice we use $n = 10$, which we empirically found to provide a good trade-off between mapping accuracy and non-overfitting.

Given a feature vector x_θ computed for an image of a face seen under pose θ, we can use this new mapping to predict the corresponding feature vector $\tilde{x}_{\theta\to 0}$ as if the face was under a frontal view in the image by computing:

$$\tilde{x}_{\theta\to 0} = M_\theta^h h_n(x_\theta).$$

5. Experimental results on BU3DFE and Multi-PIE

We performed quantitative and qualitative extensive experiments on two popular datasets. Before comparing our results with those from several state-of-the-art approaches, we describe here the protocols we used for each dataset.

5.1. BU3DFE dataset

BU3DFE is a publicly available dataset containing 3D scanned faces of 100 subjects with six basic expressions. More details can be found in Ref. [38]. As usually done, we rendered multiple views

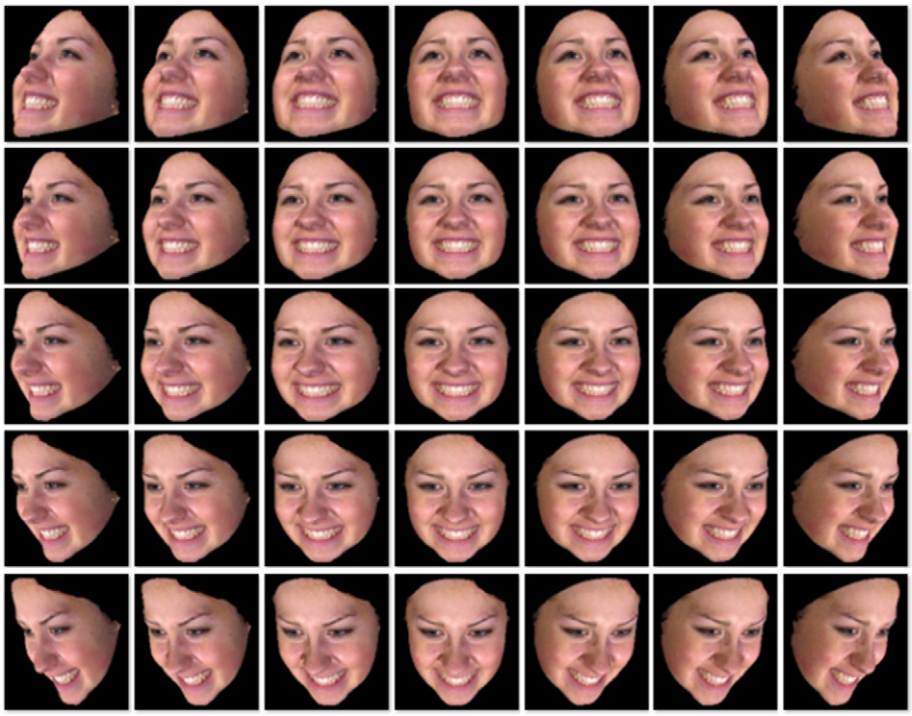

Fig. 3. BU3DFE first protocol (P1), containing 35 viewpoints.

of the 3D faces from seven pan angles (0°, ±15°, ±30°, ±45°) and five tilt angles (0°, ±15°, ±30°), which gives 35 poses we used to compare our results with the state-of-the-art [15,19,39,40,41]. Fig. 3 illustrates rendered faces for a subject sample regarding to the head pose.

In addition we rendered images from 5 angles: 0°, 30°, 45°, 60° and 90° as a second protocol to compare our model with papers that applied this protocol [18,42]. A specific subject in this protocol with variant head pose is demonstrated in Fig. 4. Therefore, as there are 6 expressions for 100 subjects over the highest level of expression intensity in 35 viewpoints, we have 21,000 samples for the first protocol. Similarly, with 6 expressions for 100 subjects over the highest level of expression intensity in 5 viewpoints therefore, we have 3000 samples for the second protocol of BU3DFE.

5.2. Multi-PIE dataset

CMU Multi-PIE is a multi-purpose dataset for facial analysis containing 337 subjects taken across 15 different viewpoints in four recording sessions [43]. Pose variations are between −90° and +90° with steps of 15°. There are therefore 13 different viewpoints per subject 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 the facial regions using a semi-automatic algorithm into 175 × 200 images. We considered the two following protocols for the Multi-PIE dataset. The first protocol is the one used in Refs. [9,18] and contains 13 viewpoints, and the second protocol is the one used in Refs. [7,25] and contains 7 viewpoints. Fig. 5, shows a subject sample of Multi-PIE in variant poses.

5.3. Soft learning using nearest neighbors

Working with smaller subsets as we do has the advantage of dealing with smaller ranges of variation. However it also reduces the amount of training data. We therefore propose to exploit training data from neighborhoods. Therefore our training data is augmented by means of the samples from similar poses. We consider four neighbors “lower”, “upper”, “right”, and “left” poses. To obtain the neighborhoods, after detecting head pose, we select the neighbors based on our prior knowledge about the head poses and neighborhoods. For instance, we know from BU3DFE dataset that four neighbors of 'Frontal' view are the viewpoints at left with head pose (−15°, 0), right with head pose (15°, 0), above with head pose (0, −15°) and at the bottom with head pose (0, 15°). Therefore, with the similar knowledge we determine the neighborhoods for each detected

Fig. 4. BU3DFE second protocol (P2), containing 5 viewpoints.

Fig. 5. First protocol of Multi-PIE containing 13 viewpoints, the second protocol of Multi-PIE is the same but first 7 viewpoints.

head pose. Moreover, we do not consider related neighbor if there is no neighborhood. The Experimental results section shows that this strategy slightly improves our overall accuracy.

5.4. Experimental results

We obtained our best results with a dictionary size of 200 with sparsity (Γ =)50. We evaluate our proposed approaches and provide comparisons between them, they are Pose Specific Classification (PSC) which is a basic pose-dependent classification without any transformation as baseline. Pose specific linear mapping (PSLM), explained in Section 4.2 which maps faces from each viewpoint to frontal and evaluate them for facial expression. A version of PSLM with Sparse Features is presented as PSLM-SF in Section 3.2 and the improved version of PSLM-SF in terms of time complexity proposed as Fast PSLM-SF or FPSLM-SF in the same subsection, finally the kernel-based non-linear version of PSLM described in Section 4.3 as KPSNM.

The performances of PSLM-SF and FPSLM-SF are very close to each other, but FPSLM-SF is faster at run-time: As shown in Table 1, FPSLM-SF is much faster than all the other approaches on all protocols for both BU3DFE and Multi-PIE datasets because it does not use OMP, decreases the feature dimensionality from 5480 to 200, and relies on the fast ridge regression step, while having better results than several related works. Note that the running times given in Table 1 for the experiments already performed for Ref. [9] are different from the ones reported in Ref. [9] as we used a better computer. We ran all the evaluations again on the same machine for fair comparison with our new KPSNM method.

Therefore, Table 1 provides comparison between PSC, PSLM, PSLM-SF, FPSLM-SF and KPSNM methods in terms of time complexity. Where the best method is FPSLM-SF while the best accuracy is provided by KPSNM. The KPSNM needs about 86 ± 5 ms but FPSLM-SF performs the result in 43 ± 7 ms for an input face sample. This means that FPSLM-SF is obviously closer than others to real-time computations. Moreover, PSLM results on both datasets show that it is better than sparse-based methods concerning the accuracy while KPSNM which is a kernel-based non-linear version of PSLM outperforms all methods with 79.26% on BU3DFE-P1, 78.79% on BU3DFE-P2, 82.43% on Multi-PIE-P1 and 83.09% on Multi-PIE-P2. A detailed comparison between proposed methods is shown in Table 2. In addition, Fig. 6 gives four confusion matrices for the KPSNM method. It can be seen that most of the confusion is between sadness and anger on both protocols of BU3DFE and similarly the most confusion on Multi-PIE protocols is between disgust and squint. The best recognized expressions are surprise and then happiness due to the clear variations.

5.5. 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 3 shows

Table 1
Proposed approaches running time (seconds) evaluation on the BU3DFE and Multi-PIE datasets. Note that the number of samples is shown under the protocols. The best performance are shown in bold.

Dataset	BU3DFE		Multi-PIE	
Test samples	P1 (35 vp)	P2 (5 vp)	P1 (13 vp)	P2 (7 vp)
Methods	4200	600	1885	1015
PSLM-SF	407	56	171	90
FPSLM-SF	**149**	**23**	**95**	**40**
PSC	211	29	107	54
PSLM	324	43	150	81
KPSNM	390	53	156	83

Table 2
Proposed approaches accuracy on BU3DFE and Multi-PIE datasets. The best performance are shown in bold.

Dataset	BU3DFE		Multi-PIE	
Methods	P1 (35 vp)	P2 (5 vp)	P1 (13 vp)	P2 (7 vp)
PSLM-SF	76.04	75.16	74.61	77.04
FPSLM-SF	77.61	75.63	75.20	76.89
PSC	77.66	76.36	80.94	82.07
PSLM	78.04	77.87	81.96	82.55
KPSNM	**79.26**	**78.79**	**82.43**	**83.09**

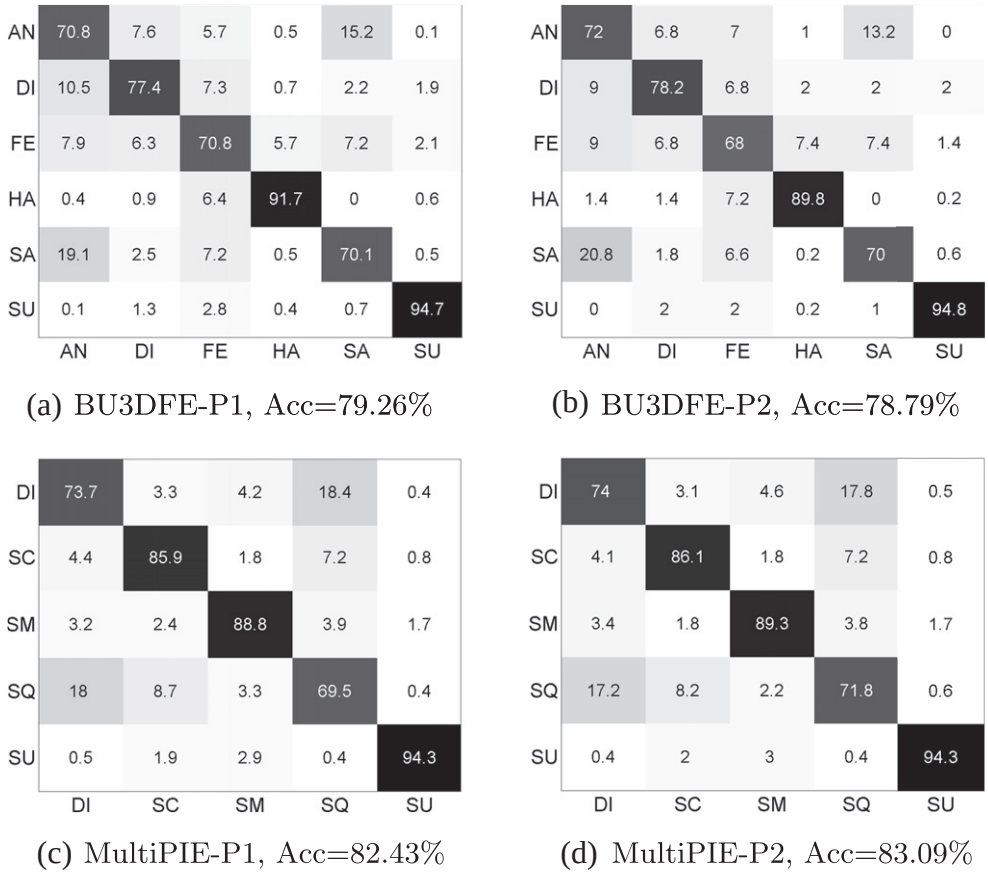

Fig. 6. Confusion matrices for the KNPSR method and (a),(b) the two protocols of BU3DFE and (c),(d) the two protocols of the Multi-PIE dataset.

that KPSNM outperforms the state-of-the-art for all protocols of BU3DFE and Multi-PIE. For instance, Ref. [44] proposed an approach similar to our PSC method introduced in Ref. [9] but based on a new descriptor (LGBP). They reported 80.17% accuracy on Multi-PIE dataset with seven viewpoints similar to Multi-PIE-P2 however they used six expressions from 100 subjects. Ref. [7] reported 81.7% on the same dataset with GSRRR method whereas our KPSNM reaches 83.09% for the same seven viewpoints. While Table 3 shows that our non-linear mapping-based approach is the best technique for MFER, we point out that the main important points of the method we proposed are its applicability, simplicity and high accuracy which are desirable for real applications. Ref. [24] also achieves good performance on the Multi-PIE dataset, however they do not consider Scream and Squint, which appear to be the most confusing ones in our experiments. We focus here on the settings already used by popular works to perform reasonable comparisons. In the next section, we propose three challenging evaluations in order to introduce and compare the robustness of MFER methods.

Table 3
Comparing our PSR and KNPSR methods with the state-of-the-art. The best performance are shown in bold.

Methods	Dataset/protocol	Accuracy
BDA/GMM by [15]	BU3DFE-P1	68.20
EHMM by [41]	BU3DFE-P1	75.30
GSCF by [19]	BU3DFE-P1	76.10
SSVQ by [39]	BU3DFE-P1	76.34
SSE by [40]	BU3DFE-P1	76.60
PSR by [9]	BU3DFE-P1	78.04
KPSNM [ours]	BU3DFE-P1	**79.26**
LBP^{ms} by [25]	BU3DFE-P2	72.43
DNPE by [18]	BU3DFE-P2	72.47
LPP by [42]	BU3DFE-P2[a]	73.06
LGBP by [25]	BU3DFE-P2	77.67
PSR by [9]	BU3DFE-P2	77.87
KPSNM [ours]	BU3DFE-P2	**78.79**
DNPE by [18]	Multi-PIE-P1[b]	76.83
PSR by [9]	Multi-PIE-P1	81.96
KPSNM [ours]	Multi-PIE-P1	**82.43**
PSLM [ours]	Multi-PIE-P2	82.55
KPSNM [ours]	Multi-PIE-P2	**83.09**

[a] With 4 level of intensities.
[b] 100 subjects instead of our protocol with 145 subjects.

6. Robustness

In this section, we consider three important challenges for practical multiview facial expression recognition: Presence of occlusion, amount of training data, and head pose estimation errors. There is no standard protocols, and we first detail the choices we made before comparing with previous methods.

6.1. Evaluation of occlusion presence

For this experiment we inserted white square blocks of various sizes (40 × 40, 50 × 50 and 60 × 60, in 200 × 220 face images in BU3DFE-P1) at random places. Fig. 7 shows examples of occluded faces. Table 4 summarizes the results of the proposed approaches with the first protocol of BU3DFE (35 viewpoints) on these perturbed

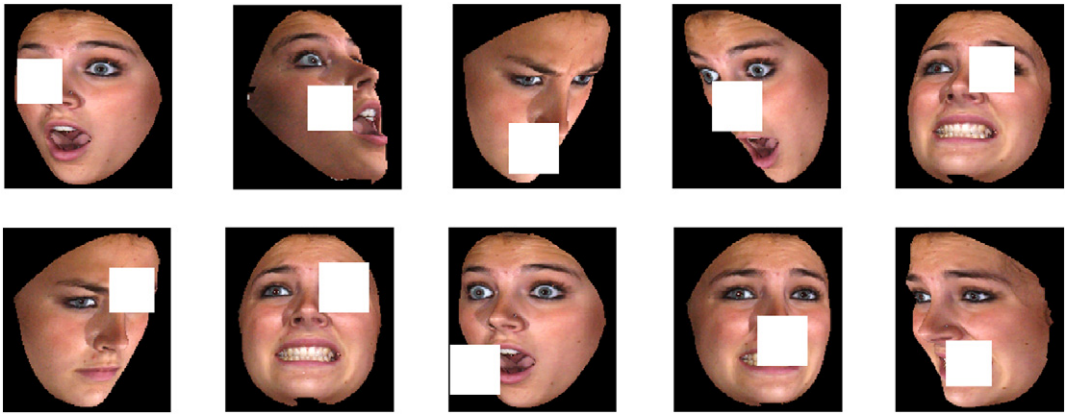

Fig. 7. Face samples after an artificial occluder are added at a random location.

images. All methods decrease slightly, those using sparse representations decrease more, but KPSNM still performs the best. Our PSLM and KPSNM methods are not much influenced by these artificial occlusions. By contrast, sparse coding-based methods cannot handle this amount of noise.

6.2. Evaluation of reducing training data

Acquiring labeled data is always cumbersome, and similar accuracy using less training data is desirable. In this experiment we evaluate our approaches by removing some viewpoints from the training data for the first protocol of BU3DFE, while testing on all 35 viewpoints.

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 that 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), as shown in Fig. 8. The results are illustrated in Table 5, which shows our approaches' robustness on reducing training data. KPSNM and PSLM are more stable than the other methods while we reduce 28.5% of training data the expression recognition of KPSNM and PSLM is 77.10% and 76.76% respectively, which are still better than the state-of-the-art.

Table 4
Evaluation of the influence of occlusions on BU3DFE-P1. We considered three different block sizes 40 × 40, 50 × 50 and 60 × 60. The best performance are shown in bold.

Occlusion size / Methods	Ground truth	40 × 40	50 × 50	60 × 60
PSLM-SF	76.04	61.65	56.64	50.66
FPSLM-SF	77.61	61.83	56.19	50.46
PSC	77.66	69.15	65.30	61.50
PSLM	78.04	71.10	67.44	63.32
KPSNM	**79.26**	**72.08**	**67.89**	**63.76**

6.3. 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 2 were obtained by an automatic viewpoint classification and therefore already included a small amount of head pose errors, 10.84% for BU3DFE-P1, 2.43% for BU3DFE-P2 and less than 1% for both protocols of the Multi-PIE dataset.

In this experiment, we artificially add two levels of pose estimation noise: During testing we randomly replace each viewpoint estimation by one of its neighboring ones, 15 or 30° farther, therefore taking wrong classifiers in PSC, wrong transformations and classifiers in PSLM, PSLM-SF, FPSLM-SF and KPSNM. Table 6 shows averaged results over 8 runs of selecting wrong neighboring poses. It illustrates that all our regression-based approaches are almost perfectly stable with respect to pose estimation errors. The PSC approach decreased as it is trained purely on view specific data.

7. Conclusions and future work

In this paper, we have proposed several pose specific mapping approaches for multiview facial expression recognition. Sparsity-based methods are faster than other methods, while PSLM and KPSNM are more accurate approaches and outperform state-of-the-art methods. Moreover, we showed that our PSLM and KPSNM methods are stable under head pose estimation errors, partial occlusion, and small training datasets. Compensating for occluded facial parts and evaluating our mapping approaches in the wild are possible directions for future work.

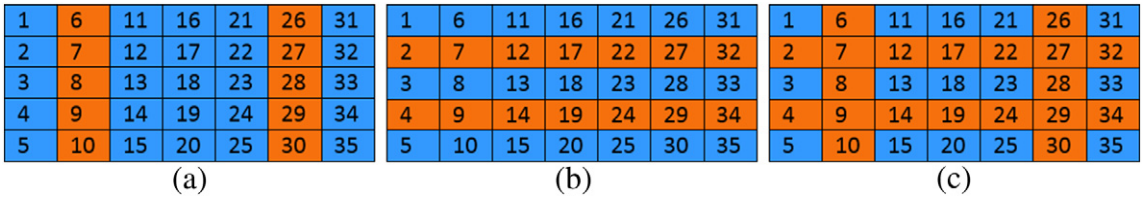

Fig. 8. Ignoring some training data to test the need for large training sets. (a) Ignoring two columns of training data: viewpoints 6–10 and 26–30); (b) ignoring two rows of 2 and 4 from training data; (c) ignoring two rows and two columns of training data.

Table 5
Evaluation of the influence of reducing the amount of training data on the approaches we propose. The best performance are shown in bold.

Reduced training / Methods	Ground truth	28.50%	40%	57.14%
PSLM-SF	76.04	70.33	70.05	68.22
FPSLM-SF	77.61	73.07	72.20	70.85
PSC	77.66	74.88	74.68	72.81
PSLM	78.04	76.76	76.50	75.10
KPSNM	**79.26**	**77.10**	**76.80**	**75.40**

Table 6
Influence of the error when estimating the head pose on the recognition of the facial expressions for the methods we propose. The best performance are shown in bold.

Used viewpoint / Methods	Ground truth	First level of neighborhood	Second level of neighborhood
PSLM-SF	76.04	72.94	72.66
FPSLM-SF	77.61	74.20	73.03
PSC	77.66	66.33	50.48
PSLM	78.04	77.10	74.18
KPSNM	**79.26**	**78.19**	**76.64**

Acknowledgments

This work was supported by following Austrian Research Promotion Agency (FFG) projects FACTS (832045), DIANGO (840824) and Vision+ (836630).

References

[1] L. Zhang, D. Tjondronegoro, V. Chandran, Facial expression recognition experiments with data from television broadcasts and the World Wide Web, Image Vis. Comput. 32 (2) (2014) 107–119.
[2] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 39–58.
[3] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D.N. Metaxas, C. Neidle, Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions, Image Vis. Comput. 32 (10) (2014) 671–681.
[4] O. Rudovic, V. Pavlovic, M. Pantic, Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On, 2012. pp. 2634–2641.
[5] W. Liu, C. Song, Y. Wang, Facial expression recognition based on discriminative dictionary learning, Pattern Recognition (ICPR), 2012 21St International Conference On, 2012. pp. 1839–1842.
[6] O. Rudovic, I. Patras, M. Pantic, Regression-based multi-view facial expression recognition, Pattern Recognition (ICPR), 2010 20Th International Conference On, 2010. pp. 4121–4124.
[7] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Trans. Affect. Comput. 5 (2014) 71–85.
[8] A. Dhall, K. Sikka, G. Littlewort, R. Goecke, M. Bartlett, A discriminative parts based model approach for fiducial points free and shape constrained head pose NorMalisation in the wild, Applications of Computer Vision (WACV), 2014 IEEE Winter Conference On, 2014. pp. 1028–1035.
[9] M. Jampour, T. Mauthner, H. Bischof, Pairwise linear regression: an efficient and fast multi-view facial expression recognition, Automatic Face and Gesture Recognition (FG), 2015 11Th IEEE International Conference and Workshops On, 2015. pp. 1–8.
[10] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press.
[11] W. Chu, F.D. la Torre, J. Cohn, Selective transfer machine for personalized facial action unit detection, Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference On, 2013. pp. 3515–3522.
[12] X. Ding, W.-S. Chu, F.D. la Torre, J. Cohn, Q. Wang, Facial action unit event detection by cascade of tasks, Computer Vision (ICCV), 2013 IEEE International Conference On, 2013. pp. 2400–2407.
[13] X. Huang, G. Zhao, M. Pietikäinen, W. Zheng, Dynamic facial expression recognition using boosted component-based spatiotemporal features and multi-classifier fusion, Advanced Concepts for Intelligent Vision Systems: 12Th International Conference, 2010. pp. 312–322.
[14] I. Kotsia, S. Zafeiriou, I. Pitas, Texture and shape information fusion for facial expression and facial action unit recognition, Pattern Recogn. 41 (2008) 833–851.
[15] W. Zheng, H. Tang, Z. Lin, T. Huang, Emotion recognition from arbitrary view facial images, ECCV 2010: 11Th European Conference on Computer Vision, 2010. pp. 490–503.
[16] N. Hesse, T. Gehrig, G. Hua, H. Ekenel, Multi-view facial expression recognition using local appearance features, Pattern Recognition (ICPR), 2012 21st International Conference on, 2012. pp. 3533–3536.
[17] R.A. Khan, A. Meyer, H. Konik, S. Bouakaz, Framework for reliable, real-time facial expression recognition for low resolution images, Pattern Recogn. Lett. 34 (2013) 1159–1168.
[18] X. Huang, G. Zhao, M. Pietikäinen, Emotion recognition from facial images with arbitrary views, British Machine Vision Conference (BMVC), 2013. pp. 76.1–76.11.
[19] U. Tariq, J. Yang, T. Huang, Multi-view facial expression recognition analysis with generic sparse coding feature, ECCV 2012: 12Th European Conference on Computer Vision, 2012. pp. 578–588.
[20] S. Taheri, P.K. Turaga, R. Chellappa, Towards view-invariant expression analysis using analytic shape manifolds, Automatic Face Gesture Recognition and Workshops (FG 2011) 2011 IEEE International Conference On, 2011. pp. 306–313.
[21] A. Moeini, H. Moeini, K. Faez, Unrestricted pose-invariant face recognition by sparse dictionary matrix, Image Vis. Comput. 36 (2015) 9–22.
[22] B.B. Amor, H. Drira, S. Berretti, M. Daoudi, A. Srivastava, 4D facial expression recognition by learning geometric deformations, IEEE Transactions on Cybernetics 44 (2014) 1–16.
[23] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, T. Huang, A study of non-frontal-view facial expressions recognition, Pattern Recognition (ICPR), 2008 19Th International Conference On, 2008. pp. 1–4.
[24] O. Rudovic, M. Pantic, I. Patras, Coupled Gaussian processes for pose-invariant facial expression recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1357–1369.
[25] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vis. Image Underst. 115 (2011) 541–558.
[26] D. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition? Computer Vision (ICCV), 2011 IEEE International Conference On, 2011. pp. 471–478.
[27] Facial Expression Recognition Using Sparse Coding, 2013.
[28] R. Timofte, V. De, L.V. Gool, Anchored neighborhood regression for fast example-based super-resolution, Computer Vision (ICCV), 2013 IEEE International Conference On, 2013. pp. 1920–1927.
[29] S.C.A. Asthana, S. Zafeiriou, M. Pantic, Incremental face alignment in the wild, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference On, 2014. pp. 1–8.
[30] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference On, 2005. pp. 886–893.
[31] T. Ojala, M. Pietikäinen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, Pattern Recognition (ICPR), 1994 12Th International Conference On, 1994. pp. 582–585.
[32] M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (2006) 4311–4322.
[33] J. Tropp, A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (2007) 4655–4666.
[34] R. Rigamonti, M. Brown, V. Lepetit, Are sparse representations really relevant for image classification? Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On, 2011. pp. 1545–1552.
[35] C. Chih-Chung, L. Chih-Jen, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27.
[36] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 2106–2112.
[37] H. Bristow, S. Lucey, Regression-based image alignment for general object categories, Arxiv Preprint Arxiv:1407.1957.
[38] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3D facial expression database for facial behavior research, Automatic Face Gesture Recognition, 2006. FG '06. 7Th IEEE International Conference On, 2006. pp. 211–216.
[39] U. Tariq, J. Yang, T. Huang, Automatic Face and Gesture Recognition (FG), 2013 10Th IEEE International Conference and Workshops On, 2013. pp. 1–6.
[40] U. Tariq, J. Yang, T. Huang, Supervised super-vector encoding for facial expression recognition, Pattern Recogn. Lett. 46 (2014) 89–95.
[41] H. Tang, M. Hasegawa-Johnson, T. Huang, Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors, Multimedia and Expo (ICME), 2010 IEEE International Conference On, 2010. pp. 1202–1207.
[42] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, T. Huang, Multi-view facial expression recognition, Automatic Face Gesture Recognition, 2008. FG '08. 8Th IEEE International Conference On, 2008. pp. 1–6.
[43] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (2010) 807–813.
[44] S. Moore, R. Bowden, Multi-view pose and facial expression recognition, British Machine Vision Conference (BMVC), 2010. pp. 11.1–11.11.