# Facial Expression Recognition under a Wide Range of Head Poses

Radu-Laurenţiu Vieriu, Sergey Tulyakov, Stanislau Semeniuta, Enver Sangineto, Nicu Sebe

DISI, University of Trento, via Sommarive 9, 38123 Trento, Italy

*Abstract*— Most of the facial expression recognition methods assume frontal or near-frontal head poses and usually their accuracy strongly decreases when tested with non-frontal poses. Training a 2D pose-specific classifier for a large number of discrete poses can be time consuming due to the need of many samples per pose. On the other hand, 2D and 3D view-point independent approaches are usually not robust to very large head rotations. In this paper we transform the problem of facial expression recognition under large head rotations into a missing data classification problem. 3D data of the face are projected onto a head pose invariant 2D representation and in this projection the only difference between poses is due to self-occlusions with respect to the depth sensor's position. Once projected, the visible part of the face is split in overlapping patches which are input to independent local classifiers and a voting scheme gives the final output. Experimental results on common benchmarks show that our method can accurately recognize facial expressions in a much larger pan and tilt range than state-of-the-art approaches, obtaining comparable performance to the best existing systems working only in narrower ranges.

## I. Introduction

Facial expression recognition is one of the most important research topics in Human-Computer Interaction [1] and many research methods have been proposed in the last decades [1], [2], [3], [4]. However, despite the large applicative interest and the many years of research on this topic, facial expression recognition is still an open problem. One of the main open issues is the drastic drop of accuracy of the existing facial expression approaches when applied to non-frontal faces [5]. In this paper we propose to transform the problem of emotion recognition with drastic head poses into a missing data classification problem. Using a single depth image, we create the 3D point cloud of the input face. The 3D points are then projected onto a head pose invariant cylindrical surface, which is constructed using the position of the subject's eyes. When the head position is not frontal with respect to the depth sensor's plane, only part of the face is visible and can be projected onto the cylindrical surface (e.g., see Fig. 1). However, the visible part of the analyzed face is represented *in a frontal position*, which makes it possible to ignore deformations due to the viewpoint perspective, common in 2D images, and to train and test our classifier using only frontal representations. Once obtained a pose invariant representation of the face, we deal with missing information using a local classification approach: the cylindrical surface is split in overlapping rectangular patches and each patch is fed into a Random Forest classifier. More in detail, inspired by the recent success of multi-channels approaches in object detection [6], [7], we firstly compute multiple maps (channels) of the projected data, using LBP features [8] and oriented gradients. Then, each channel is divided in overlapping patches and for every patch we train a dedicated Random Forest [9]. Splitting tests associated with the nodes of the trees are based on the selection of the channel *and* the feature which *jointly* contribute the most in decreasing the entropy at that node. Moreover, each Random Forest takes a local decision depending only on the specific patch it was trained on. At testing time, we deal with large missing information by collecting the votes of the forests corresponding to only the visible patches. We show in Sec. V that our approach can deal with a range of rotations much broader than other state-of-the-art methods and obtains comparable results in the more commonly adopted ranges.

**Contributions:** (1) We propose an easy-to-compute and rotation-invariant cylindrical projection of the 3D face point cloud for face expression analysis in arbitrary head poses. (2) We deal with self-occlusions of a rotated face by means of a local classification approach transforming the head rotation problem into a missing information problem. (3) As far as we know, we are the first to adopt a joint channel-feature selection approach for facial expression recognition.

## II. Related work

Facial expression analysis methods based only on 2D intensity images can be grouped in *view-dependent* and *view independent* approaches [5]. In [10] Zheng et al. use regional covariance matrices for view independent facial expression recognition. In [11] Rudovic et al. use a discrete training set of 34 poses in conjunction with the frontal pose in order to learn a regression function which maps the facial landmark positions of a given non-frontal face to a frontal layout. At testing time, the head pose is estimated and the closest training poses are used to project the landmarks onto a frontal pose. Finally, a multi-class SVM is applied to the frontally-normalized landmark positions to categorize the facial expressions. One drawback of this work and, generally speaking, of all the view-dependent 2D approaches, is the necessity to accurately estimate the head pose and/or a large number of facial landmarks. Both of these tasks are quite difficult when using only 2D data [12].

Another problem of the 2D approaches is the need of labeled training samples in many different head poses, which are expensive to collect. In [13] Jiang and Jia propose a semi-supervised approach based on Transfer AdaBoost [14] to deal with the scarcity of non frontal training samples. However, the authors show experimental results with only a very limited discrete set of pan rotations and the method needs a (limited) number of non-frontal face training samples. Recently Huang et al. [5] proposed to use Multiset

Canonical Correlation Analysis [15] in order to exploit the correlation between facial expressions and pose labels. They show promising results with a limited set of discretized pan angles.

To avoid problems related to varying illumination conditions and to be able to synthesize multi-view projections of a face, most of the static facial expression recognition systems dealing with non frontal poses directly or indirectly exploit 3D data [16]. According to [16], these methods can be split in four main categories: *distance based*, *patch based*, *morphable models* and *2D representations*. Distance based methods extract the (3D) landmark positions of the input face and use a prefixed set of inter-landmark distances to classify the facial expressions [17], [18]. The drawback of these methods is the need of an accurate localization of many landmarks. Patch based approaches extract local features from either every point of a 3D mesh or around specific landmarks [16]. For instance, in [19] facial landmarks on the 3D surface of a face specify the positions in which patches are described by means of level curves. Probe and gallery expression samples are compared computing the geodesic distance between such curves. Note that the patches we use in our approach are computed on the 2D projection of the 3D face points (see below the 2D representation-based methods). In [20] a morphable model is fitted to the face point cloud by matching a set of landmarks, which need to be localized both on the prototypical model and on the analyzed face.

The 2D representation approaches [16] are the category most similar to our method and are based on mapping the 3D data onto a 2D representation. Once the mapping has been computed, different features can be extracted from the 2D representation. For instance, in [21] depth maps and Azimuthal Projection Distance Images are filtered with different methods, such as Gabor filters, LBP features, etc., with the goal of action unit detection. In [22] a depth map of the 3D facial meshes is computed and SIFT features are extracted in this map around specific landmarks. In our approach we do not need to accurately localize landmarks on our 2D representation and a rough estimation of the head pose together with the position of the eyes in the depth map is sufficient to compute the cylindrical projection surface (Sec. III). Moreover, our Random Forest based joint selection of features and channels makes it possible to adaptively choose among a huge number of possible features, and our experimental results (Sec. V) show that the Multi-Channel approach [23], [24], [6], [7] can be robustly adopted for facial expression analysis.

### III. HEAD POSE INVARIANT FACE REPRESENTATION

In this section we describe how the head pose invariant face representation is constructed. We assume as input a depth image of the analysed face, the positions of the centers of the eyes and the head pose. In this work, we do not address eye detection and head pose estimation problems. However, several studies have recently shown that it is possible to obtain the head pose along with the estimated positions of both eyes even when only one eye is visible.

For instance this is done on a frame-by-frame basis in [25] and using head tracking information and person-specific templates in [26]. At training time we use the ground truth values provided in the BU-3DFE and the Bosphorus datasets, while at testing time this information can be obtained via detection or tracking.

Typically, a 3D *point cloud* of a face is represented by a set of 3D coordinates and the associated color information obtained from a generic 3D scanning device. Since we do not use color in our approach, from now on we define a point cloud as a set of $m$ points $\mathcal{P} = \{p_i : i = 1, ..., m\}$, where $p_i = (x, y, z)$. $\mathcal{P}$ is projected onto a Cylindrical Head Model (CHM) using a sampling approach with the following properties. (1) It requires only two landmarks and the head pose. The latter can be represented as a transformation matrix $T_0$. (2) Rectifies the head orientation so that the representation is the same as if the face was seen from the frontal view, except for missing information caused by pose-dependent self-occlusions. (3) Handles various acquisition devices: our representation is independent of the parameters of the capturing device, which makes it possible to train the method on one dataset and test on the other (see Sec. V).

It is worth noticing that CHMs have already served as a proxy for *head pose estimation* problems in 2D [27], [28], [29]. However, while in these methods CHMs are used to track the head orientation or to estimate the head pose parameters, in our approach we use a CHM for facial expression recognition based on the above-mentioned properties. We start our sampling pipeline by applying the transformation $T_0^{-1}$, since the head pose $T_0$ is known, to make the input face frontal.

**CHM parameters.** In order to build a CHM we need to determine the longitudinal axis of the cylinder and choose a proper head radius $r_h$. To set these parameters we use anthropometric values of the human face [30]. The schematic view from top is given in Fig. 1(b). The head radius is determined as $r_h = \alpha \cdot l$, where $l$ is the interocular distance. The longitudinal axis of the cylinder from top is viewed as a point $P_0$ and is determined as the intersection of two cylinders centered in the eyes, each with radius $r_e = \beta \cdot l$. $\alpha$ and $\beta$ are determined to be consistent with average anthropometric values: $\alpha = 1.38, \beta = 1.13$.

**Sampling points.** Sampling points are generated in cylindrical coordinates according to the sampling parameters $\{R, M, N, \phi_0, \phi_1\}$, where $R = r_h$ is the CHM radius, $M$ and $N$ are the number of sampling points along the vertical and horizontal directions and $\phi_0$, $\phi_1$ are the starting and the ending azimuths of sampling in cylindrical coordinates respectively. We use: $M = 150$, $N = 120$, $\phi_0 = \frac{\pi}{20}$ and $\phi_1 = \frac{19 \cdot \pi}{20}$.

**Interpolation.** For every point $s_{ij}$ of the sampling cylinder, $i = 1, ..., M$, $j = 1, ..., N$ we select the $n$ closest points of $s_{ij}$ in $\mathcal{P}$ projected using the normal direction to $s_{ij}$. In order to determine these $n$ points, we project the points in $\mathcal{P}$ onto a surface $S_{ij}$ tangent to the cylinder in $s_{ij}$. $S_{ij}$ is defined by the orthonormal basis $[\vec{n_{s_{ij}}}, \vec{v_2}, \vec{v_3}]$, where $\vec{n_{s_{ij}}}$ is a normal vector to the cylinder at $s_{ij}$, $\vec{v_2}$ is a vector tangential
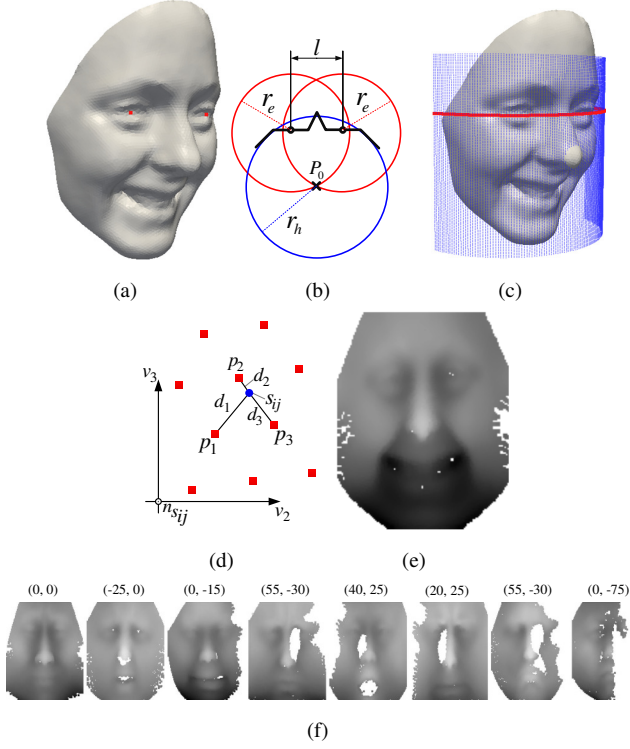
Fig. 1. Cylindrical sampling pipeline. (a) An example face scan from the BU-3DFE database with two landmarks. (b) The cylindrical head model parameters. (c) Our CHM with $150 \times 120$ sampling points imposed on the face scan. (d) Sampling point values computed based on the 3 nearest neighbors. The blue circle represents a sampling point $s_{ij}$ while the red squares are points of the source cloud. (e) An example of pose-invariant face representation. (f) Examples of sampled faces under varying head poses and facial expressions. The head rotation $(tilt, yaw)$ is given in the brackets.

to the cylinder surface and parallel to plane defined by axes $X$ and $Z$, and $\vec{v_3} = n_{\vec{s}_{ij}} \times \vec{v_2}$. Let $\widetilde{p}_k$ be the projection of $p_k$ onto $S_{ij}$ ($p_k \in \mathcal{P}$) and $V(p_k) = ||p_k - \widetilde{p}_k||$.

Red squares in Fig. 1(d) represent projected points of the face onto $S_{ij}$. The blue circle is the sampling point $s_{ij}$ of the cylinder. The normal to the sampling point $s_{ij}$ is directed toward the observer. The sampled value $V(s_{ij})$ at $s_{ij}$ is computed by interpolating the $n$ closest points of $s_{ij}$ in the set $P_c(s_{ij}) = \{\widetilde{p} : ||\widetilde{p} - s_{ij}|| \leq \epsilon\}$:

$$V(s_{ij}) = \sum_{k=1}^{n} V(p_k) \cdot w_k, \quad (1)$$

where: $w_k = \frac{D_n - d_k}{(n-1) \cdot D_n}$, $D_n = \sum_{k=1}^{n} d_k$, $p_k \in P_c(s_{ij})$ and $d_k = ||\widetilde{p}_k - s_{ij}||$. Both $||\widetilde{p}_k - s_{ij}||$ and $||\widetilde{p} - s_{ij}||$ are computed on $S_{ij}$.

Using Eq. (1), points having larger distance from $s_{ij}$ have a smaller impact on $V(s_{ij})$. By varying $n$ we can control how blurry the final image is. We set $n = 3$. The parameter $\epsilon$ controls how far a point can be from $s_{ij}$ and still have an influence in computing $V(s_{ij})$. We finally compute a Boolean mask $M$ which is false for sampling points belonging to the background or possible holes in the face, and it is simply defined as: $M(s_{ij}) := P_c(s_{ij}) \neq \varnothing$.

Using Eq. (1) it is possible to sample any value associated with a point in $\mathcal{P}$, i.e. depth or color values. In our work, we use only depth. Some examples of faces of the BU-3DFE

dataset projected onto our head pose invariant representation are shown in Fig. 1(f).

## IV. RANDOM FORESTS FOR FACIAL EXPRESSION RECOGNITION

In this section we describe the feature extraction stage, explain how the trees inside each Random Forest are built and present late-fusion strategies for the decision making process.

**Feature extraction.** Given the face representation $V$ obtained in Sec. III, we first compute 9 channels ($C_i, i = 1, .., 9$), inspired by the work of Dollar et al. [23]. We use the following channels: the initial representation itself ($V$), the magnitude of the gradients computed over $V$, 6 quantized orientations of the gradient of $V$ (split by the following values: $0, \pi/6, \pi/3, \pi/2, 2\pi/3$ and $5\pi/6$) and finally the LBP image [8] computed using $V$ (Fig. 2 (a)-(b)).

Similarly to [23], on each channel we compute generalized Haar features (*i.e.*, differences of rectangular regions of pixels), with the important difference that the rectangles of each Haar feature in our case are defined with respect to a specific patch of the channel (Fig. 2 (d)). More specifically, we first split each channel $i$ into $N$ overlapping patches, $P_n^i, n = 1, ..., N$ (Fig. 2 (c)), then extract generalized Haar features at patch level. $P_n^i$ and $P_n^j$ correspond to the same rectangular patch but are defined in two different channels.

Given channel $i$ and a patch index $n$, a feature $f(P_n)$ defined over patch $P_n$ is:

$$f_\theta(P_n) = \frac{1}{|M(R_1)|} \sum_{z \in R_1, M(z)} P_n^i(z) - \\ - \frac{1}{|M(R_2)|} \sum_{z \in R_2, M(z)} P_n^i(z), \quad (2)$$

where $R_1$ and $R_2$ are two rectangles placed over the patch surface, $M$ is the Boolean mask defined in Sec. III,
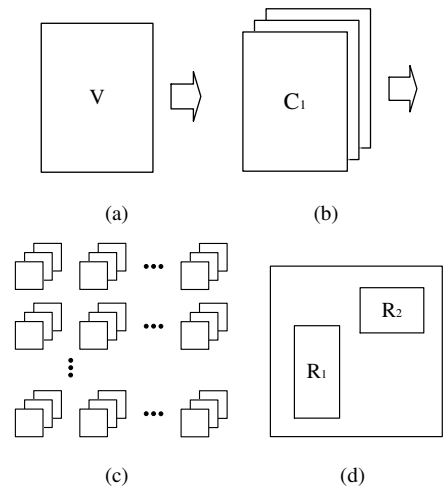


Fig. 2. Feature extraction process: the initial face representation (a) is transformed into a multiple channel representation (b), followed by a dense sampling with overlapping patches (c). (d) A patch example with two rectangular regions $R_1$ and $R_2$.

$z = (x, y)$ is a 2D coordinate inside the patch and $\theta = \{S_1^1, S_2^1, S_1^2, S_2^2, i\}$ is a parameter vector, with $S_1^k$ and $S_2^k$ being the top-left and bottom-right 2D coordinates that define $R_k$, respectively. Haar features have been proved to work remarkably well under noisy conditions and missing information [31]. Computational efficiency is guaranteed by the use of integral images.

**Random Forests.** Varying the parameters in $\theta$ we can compute an over-complete basis of features for every patch. Automatic feature selection is performed using Random Forest classifiers. Specifically, we train a dedicated Random Forest for each patch $P_n, n = 1, ..., N$, this way enabling the possibility of taking local decisions which, in turn, enhances our system with the ability of handling missing information. Random Forests have been widely used for various classification/regression tasks due to their computational efficiency and their powerful feature selection mechanism [32], [31], [33], [34]. Furthermore, Random Forests have been proven to be less prone to overfitting, in comparison with individual trees [9], and can naturally address multiple classes.

We grow each tree by sequentially splitting non-leaf nodes following a classical information gain maximization principle. At each splitting node, we define a set of $K$ binary tests $\Theta = \{t_1, t_2..., t_K\}$, with $t_k(P_n) := f_\theta(P_n) > \tau$. The parameters defining the shape of the rectangles $S_1^1, S_2^1, S_1^2, S_2^2$, as well as the values of the threshold $\tau$ are randomly generated, while $i$ spans exhaustively all the 9 channels. The patch index $n$ is fixed for a given Random Forest classifier. The winning binary test in $\Theta$ is selected as the one that maximizes the information gain associated to the splitting output. The process continues until the stopping criterion is met (*i.e.* a predefined minimum amount of samples reach the splitting node), at which point a leaf node is formed. Leaves store the class posterior probabilities computed from the samples that reach them.

**Decision fusion.** In order to obtain the final classification output, we use a weighting decision scheme that fuses probability measures produced by each patch-specific Random Forest and takes into account patch importance. Apart from the case of equal weights (**NW**), we tested one more scheme, called Performance-based (**PB**) weighting. In computing **PB** weights we use the training samples to perform a 5-fold inner loop cross validation and for each patch position we train a Random Forest classifier. Every patch-based Random Forest is *independently* tested on the validation set and its average accuracy result is finally normalized over all patches, thus obtaining the patch-specific weight.

## V. EXPERIMENTS

We perform experiments on two widely used public 3D datasets: BU3DFE [35] and Bosphorus [36]. BU3DFE contains 100 subjects (44 males and 56 females) showing the six standard expressions (*i.e.* Anger, Disgust, Fear, Happy, Sadness and Surprise), each with 4 degrees of intensity per subject and one additional Neutral example per subject, summing a total of 2500 examples. As in [11], we select only the last two most intense frames per emotion, for each

subject, resulting in 200 examples per expression plus 100 examples for the Neutral class. In order to balance the whole set, we mirror the Neutral examples, reaching a total of 1400 samples, equally distributed among the 7 classes. The Bosphorus dataset includes 105 subjects but the distribution over all the expressions is not as uniform as in the previous case. In each experiment using Bosphorus, we balance the training set such as to have the same number of examples in each class. The selected point clouds from BU3DFE are used to render all the head orientations, as described in Sec. III, whereas those from Bosphorus are rendered only in the frontal pose. In fact, to our best knowledge, there are no published results on facial expression recognition under varying head poses using Bosphorus. Thus we use this dataset only for testing our system in the frontal scenario and for training in cross-domain experiments.

### A. Parameter setting

We performed a grid search analysis in order to fix the system's key parameters. We show here only the results of the most important variables, namely the *patch size* and the *patch density* or *stride*. The patch size was varied between 20 and 60 pixels with a step of 10 pixels. For patch density, we varied the exploring stride between 5 and 15 pixels with a step of 5 pixels. A 5-fold subject independent cross validation was performed *using only frontally projected* samples of the BU3DFE dataset. As can be seen from Fig. 3(a), the best recognition rate (computed for the 7-class classification task) is obtained by using a patch size of $40 \times 40$ pixels and a stride of 5 pixels. These values are fixed throughout the entire experimental section.

Next, we tested the performance of the two weighting schemes (PB and NW, see Sec. IV) on samples corresponding to frontal head poses, both on the BU3DFE and the
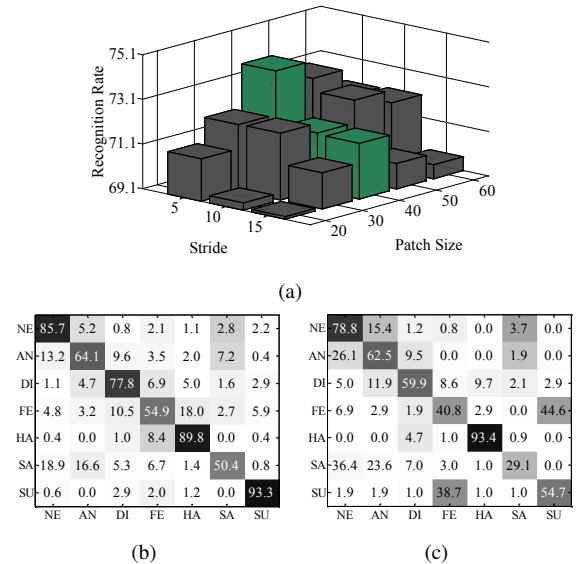


(a)



(b)      (c)

Fig. 3. (a) Parametric analysis over the patch size/density space. Recognition rate is computed over all 7 classes on BU3DFE. Confusion matrices corresponding to (b) BU3DFE $SET_1$ and (c) Bosphorus frontal for the same task.

Bosphorus dataset. As Table I shows, PB is consistently better than NW. In the rest of the paper we always use PB as the weighting scheme.

### B. Facial Expressions under varying head poses

In this section we present results corresponding to a wide range of head orientations. All the results have been obtained using the 3D point clouds taken from either BU3DFE or Bosphorus and exploiting the associated ground truth for the eyes' position and the head pose. However, producing a point cloud representation from a generic depth sensor is a trivial operation, and the eyes' localization and the head pose estimation can be reliably obtained with different methods (see Sec. III). Moreover, using synthetic data obtained from BU3DFE or Bosphorus is a widely adopted protocol (e.g., see [11]) and makes it possible to compare our method with state-of-the-art approaches. In our case, we project each point cloud example onto our head pose invariant representation using different rotation angles and then we perform our Random Forest classification. Fig. 1 (f) shows some projected examples.

We divide our experiments into three parts. In the first part (Results on $SET_1$) we adopted the protocol suggested in [11] using the BU3DFE dataset and showing experiments performed on a relatively reduced set of 247 head poses, ranging from $-45°$ to $45°$ for yaw and from $-30°$ to $30°$ for tilt, both with a step of $5°$. The second part (Results on $SET_2$) includes samples of the BU3DFE dataset projected using a much wider head pose range, $(-90°, 90°)$ for the yaw and $(-60°, 60°)$ for tilt, with the same step size. Notice that $SET_2$, summing 925 poses, includes samples for which entire facial regions disappear completely under self-occlusion, limiting the operating point of those systems that, for instance, rely on features extracted from facial landmarks. In the third part (Training on Bosphorus) we show results by training on frontal samples taken from the Bosphorus dataset. As in [11], we adopt on all the experiments a subject-independent 5-fold cross validation scheme. In our case, we used only the frontal head pose for training (training pose, $tp$) and we tested on samples belonging to one of the previously defined sets (non training poses, $ntp$). Note that in [11] 36 $tp$ poses are used.

**Results on $SET_1$:** On the reduced head-pose set we perform two different experiments. In the first we measure the influence of the facial landmark localization's precision on the recognition rate. Similarly to [11], for each test sample we corrupt the landmark positions (the eyes' center in our case) with uniformly distributed noise inside the 3D range $[-\sigma, \sigma]^3$, where $\sigma = \alpha \cdot l$, $l$ is the inter-ocular distance and $\alpha \in \{0, 0.04, 0.08, 0.12\}$. Note that in [11] the authors need to precisely localize many more landmarks (39) and that the noise levels do not include $0.12 \cdot l$. Results shown in Table II reveal superior performance of our method (RF-PB) under low noise, while achieving slightly lower but comparable recognition rates at higher noise levels. All other results in Table II, except ours, are taken from [11], to which we refer the reader for more details.

In the second experiment on $SET_1$ we compare the accuracy of our approach with state-of-the-art methods. In order to be inline with previously reported results, we separately use samples corresponding to expression intensity levels 3 and 4, respectively. As Table III shows, our proposed method produces state-of-the-art results on $SET_1$ (lines 13 and 14), while requiring the localization of only two landmarks. In particular, our system generates recognition rates superior to [11] for intensity level 3 (line 13), while for level 4, we score slightly worse. However, it is worth noticing that we use as training poses ($tp$) just the frontal representation, whereas in [11] $tp$ gathers 35 different head poses. Fig. 3(b) presents the confusion matrix obtained using $SET_1$, but combining both intensity levels. Fear and Sadness remain the mostly confused classes for BU3DFE, similarly to other work in the literature. As in Table II, we filled all results (except ours and lines 1 and 2) in Table III from [11].

**Results on $SET_2$:** Using the previously trained models, we test the performance of our system on the extended head-pose range. Results obtained on separate expression intensity levels are presented in Table III (lines 15 and 16, respectively). From the table, we can observe that the average recognition rate computed on the full set ($ntp$) is still comparable with the results obtained in much narrower head orientations. Fig. 4 shows the recognition rate distribution over the yaw/tilt space. The angle ranges are divided into blocks of equal size $15° \times 15°$ and performance is computed on samples belonging to each block. The gray area corresponds to $SET_1$, as in [11], [43], [10]. Two important observations arise from Fig. 4. **i)** Our system is able to maintain competitive recognition rates moving to severe head rotation angles. **ii)** While on the yaw space the performance distribution is symmetric, not the same can be concluded regarding the tilt space. This aspect is motivated by the fact that the mouth area is one very informative region for facial expression recognition (claim supported by the performance based weights). Under negative tilt, the mouth often suffers from self occlusion, which in turn explains the lower performance on this particular tilt range. Finally, Fig. 5 highlights the change in performance as a function of the yaw rotation angle, at constant tilt. The red lines
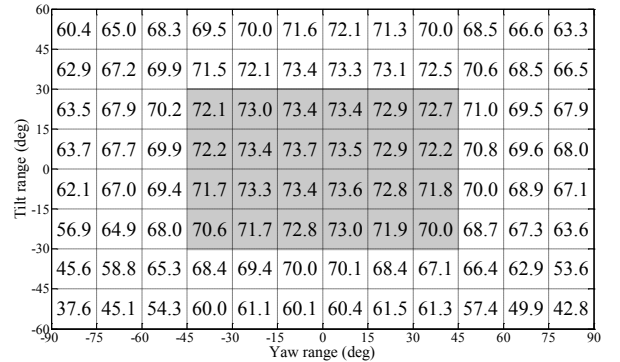


Fig. 4. Recognition rate distribution over the yaw/tilt space. The gray area shows the reduced head-pose range reported in [11] ($SET_1$).

|  | BU3DFE | Bosphorus |
|---|---|---|
| RF-PB (%) | **75.1** | **66.1** |
| RF-NW (%) | 73.9 | 64.1 |

| Method | RR (%) ($\sigma = 0$) | | RR (%) ($\sigma = 0.04 \cdot l$) | | RR (%) ($\sigma = 0.08 \cdot l$) | | RR (%) ($\sigma = 0.12 \cdot l$) | |
|---|---|---|---|---|---|---|---|---|
|  | tp | ntp | tp | ntp | tp | ntp | tp | ntp |
| 3D-PDM [37] | 62.1 | 62.3 | 61.8 | 61.0 | 59.9 | 59.2 | - | - |
| CSGPR [11] | 72.6 | 71.5 | 70.5 | 69.4 | **69.9** | **69.7** | - | - |
| RF-PB | **75.1** | **73.7** | **73.4** | **72.4** | 68.9 | 69.0 | 65.3 | **64.1** |
| RF-NW | 73.9 | 72.3 | 71.2 | 70.8 | 67.7 | 67.5 | 64.1 | 62.6 |

| No. | Method | Classifier | Features | Poses tilt | Poses pan | Poses total | Expressions no. | Expressions levels | Rec. Rates tp | Rec. Rates ntp |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gong et al. [38] | SVM | depth differences | - | $\sim 0°$ | 1 | 6 | 3,4 | 76.2% | - |
| 2 | Lemaire et al. [39] | SVM | 3D patch distances | - | $\sim 0°$ | 1 | 6 | 3,4 | 75.8% | - |
| 3 | Hu et al. [40] | PW-SVM | 41 landmarks | - | $0°, +90°$ | 5 | 6 | 1-4 | 66.7% | - |
| 4 | Moore and Bowden [8] | PW-SVM | lgbp/lbp | - | $0°, +90°$ | 5 | 6 | 1-4 | 71.1% | - |
| 5 | Hu et al. [41] | KNN | sift+lbp | - | $0°, +90°$ | 5 | 6 | 2-4 | 73.8% | - |
| 6 | Zheng et al. [42] | PW-KNN | 83 landmarks+sift | - | $0°, +90°$ | 5 | 6 | 1-4 | 78.5% | - |
| 7 | Zheng et al. [10] | linear | sift+bda/gmm | $-30°, +30°$ | $-45°, +45°$ | 35 | 6 | 4 | 68.3% | - |
| 8 | Tang et al. [43] | PW-SVM | sift+hmm | $-30°, +30°$ | $-45°, +45°$ | 35 | 6 | 4 | 75.3% | - |
| 9 | SGPR [11] | F-SVM | 39 landmarks | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 3 | 68.2% | 65.4% |
| 10 | CSGPR [11] | F-SVM | 39 landmarks | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 3 | 68.7% | 67.1% |
| 11 | SGPR [11] | F-SVM | 39 landmarks | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 4 | 75.4% | 74.2% |
| 12 | CSGPR [11] | F-SVM | 39 landmarks | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 4 | 76.5% | 76.1% |
| 13 | RF-PB | RF | multi-channel Haar | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 3 | 71.1% | 70.2% |
| 14 | RF-PB | RF | multi-channel Haar | $-30°, +30°$ | $-45°, +45°$ | 247 | 7 | 4 | 74.7% | 73.4% |
| 15 | RF-PB | RF | multi-channel Haar | $-60°, +60°$ | $-90°, +90°$ | 925 | 7 | 3 | 71.1% | 62.1% |
| 16 | RF-PB | RF | multi-channel Haar | $-60°, +60°$ | $-90°, +90°$ | 925 | 7 | 4 | 74.7% | 66.2% |

mark the boundaries of the narrower head pose range adopted in the literature, while the blue line plots the best average performance rate reported in [11]. Even at $75°$ yaw rotation of the head, our system is still able to correctly classify facial expression in more than $65\%$ of the cases.

**Training on Bosphorus:** We conducted two types of experiments using Bosphorus training data. In the first we train our system on Bosphorus and we test on the same dataset, as most of the other work does. For instance, in [44] a mean recognition rate of $60.5\%$ is reported (with only 6 classes). Even if we test with one more class, *Neutral* (hence, performing a more challenging task), we achieve a superior mean recognition rate: $66.1\%$. In the second test, we performed a cross-domain experiment, training our model on samples corresponding to frontal poses of Bosphorus and testing on samples of BU3DFE $SET_1$. We obtained a mean recognition rate of $40.7\%$, which we believe is encouraging taking into account the diversity of the two datasets.

## VI. CONCLUSIONS

We proposed a facial expression recognition approach able to deal with a very wide range of head rotations, much larger than state-of-the-art methods. Our proposal is based on a pose invariant representation in which missing information due to self occlusions is dealt with using a local classification approach. Experimental results on different benchmarks and with different rotation ranges show that our method is comparable with state-of-the-art works in commonly adopted angle ranges, being able to accurately recognize facial expressions in much broader ranges.

We are currently working on extending our approach for a specific application domain which is based on Kinect data. Despite Kinect data are very noisy, our preliminary results show that we are able to train our system on the "clean" BU3DFE dataset and test on Kinect data with only a reasonable loss in accuracy.

## VII. ACKNOWLEDGMENTS

## REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *PAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[2] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of Face Recognition*, pp. 487–519, 2011.

[3] A. Martinez and S. Du, "A model of the perception of facial expressions of emotion by humans: Research overview and perspectives," *Journal of Machine Learning Research*, vol. 13, pp. 1589–1608, 2012.

[4] F. de la Torre and J. F. Cohn, "Facial expression analysis," in *Visual Analysis of Humans*, pp. 377–409, 2011.

[5] X. Huang, G. Zhao, and M. Pietikainen, "Emotion recognition from facial images with arbitrary views," in *BMVC*, 2013.

[6] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[7] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *CVPR*, 2012.

[8] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *CVIU*, vol. 115, no. 4, pp. 541–558, 2011.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion recognition from arbitrary view facial images," in *ECCV*, 2010.

[11] 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.
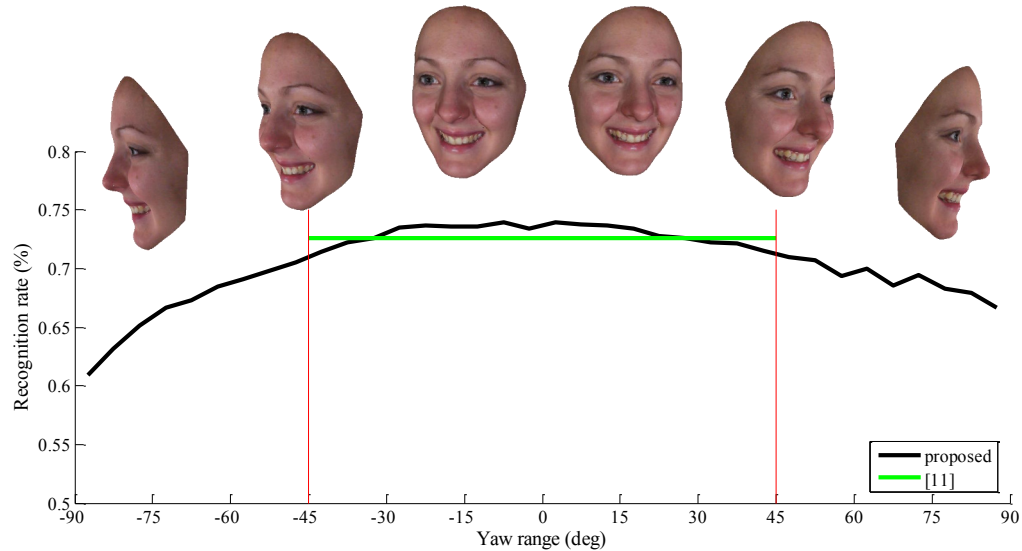
Fig. 5. Performance variation under the full range of yaw angles at 0 tilt. The red lines mark the limits of the narrower range adopted by state of the art, while the green line draws the best average performance obtained in [11], within the reported limits.

[12] E. Sangineto, "Pose and expression independent facial landmark localization using dense-SURF and the Hausdorff distance," *PAMI*, vol. 35, no. 3, pp. 624–638, 2013.

[13] B. Jiang and K. Jia, "Semi-supervised facial expression recognition algorithm on the condition of multi-pose," *J. of Information Hiding and Multimedia Signal Processing*, vol. 4, no. 3, pp. 138–146, 2013.

[14] W. Dai, Q. Yang, G. R. Xue, and Y. Yu, "Boosting for transfer learning," in *ICML*, 2007.

[15] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[16] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *IVC*, vol. 30, no. 10, pp. 683–697, 2012.

[17] H. Soyel and H. Demirel, "Facial expression recognition using 3d facial feature distances," in *ICIAR*, 2007.

[18] X. Li, Q. Ruan, and Y. Ming, "3D facial expression recognition based on basic geometric features," in *ICSP*, 2010.

[19] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3d facial expression recognition," *Pattern Recognition*, vol. 44, no. 8, pp. 1581–1589, 2011.

[20] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-d face and facial expression recognition," *Information Forensics and Security*, vol. 3, no. 3, pp. 498–511, 2008.

[21] G. Sandbach, S. Zafeiriou, and M. Pantic, "Binary pattern analysis for 3d facial action unit detection," in *BMVC*, 2012.

[22] S. Berretti, B. B. Amor, M. Daoudi, and A. D. Bimbo, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.

[23] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009.

[24] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *BMVC*, 2010.

[25] G. Fanelli, M. Dantone, and L. Van Gool, "Real time 3D face alignment with Random Forests-based Active Appearance Models," in *FG*, 2013.

[26] S. Tulyakov, R. L. Vieriu, S. Semeniuta, and N. Sebe, "Robust Real-Time Extreme Head Pose Estimation," in *ICPR*, 2014.

[27] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation.," *IEEE Trans. on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.

[28] J. Sung, T. Kanade, and D. Kim, "Pose Robust Face Tracking by Combining Active Appearance Models and Cylinder Head Models," *IJCV*, vol. 80, no. 2, pp. 260–274, 2008.

[29] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn, "Robust full-motion recovery of head by dynamic templates and re-registration techniques," *Imaging Systems and Technology*, vol. 13, pp. 85–94, 2003.

[30] C. Gordon, T. Churchill, C. Clauser, J. Mcconville, I. Tebbetts, and R. Walker, "Anthropometric survey of us army personnel: Methods and summary statistics," *U.S. Army Natick Res., Tech. Rep. NATICK/TR-89/044*, 1988.

[31] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, "Real time head pose estimation from consumer depth cameras," in *DAGM*, pp. 101–110, 2011.

[32] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pp. 106–117, 2011.

[33] C. Huang, X. Ding, and C. Fang, "Head pose estimation based on random forests for multiclass classification," in *ICPR*, 2010.

[34] V. Lepetit and P. Fua, "Keypoint recognition using randomized trees," *PAMI*, vol. 28, no. 9, pp. 1465–1479, 2006.

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

[36] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *Biometrics and Identity Management*, pp. 47–56, 2008.

[37] Z. Zhu and Q. Ji, "Robust real-time face pose and facial expression recovery," in *CVPR*, 2006.

[38] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3d face by exploring shape deformation," in *ACMMM*, 2009.

[39] P. Lemaire, B. Ben Amor, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3D facial expression recognition using a region-based approach," in *ACM J-HGBU*, 2011.

[40] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, and T. S. Huang, "A study of non-frontal-view facial expressions recognition," in *ICPR*, 2008.

[41] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *FG*, 2008.

[42] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "A novel approach to expression recognition from non-frontal face images," in *ICCV*, 2009.

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

[44] N. Vretos, N. Nikolaidis, and I. Pitas, "3d facial expression recognition using zernike moments on depth images," in *ICIP*, 2011.