Neural Computing and Applications (2020) 32:17909–17926
https://doi.org/10.1007/s00521-019-04529-7

DEVELOPING NATURE-INSPIRED INTELLIGENCE BY NEURAL SYSTEMS

Simultaneous multi-descent regression and feature learning for facial landmarking in depth images

Janez Križaj[1] · Peter Peer[2] · Vitomir Štruc[1] · Simon Dobrišek[1]

Received: 12 August 2019 / Accepted: 5 October 2019 / Published online: 23 October 2019

Abstract
Face alignment (or facial landmarking) is an important task in many face-related applications, ranging from registration, tracking, and animation to higher-level classification problems such as face, expression, or attribute recognition. While several solutions have been presented in the literature for this task so far, reliably locating salient facial features across a wide range of posses still remains challenging. To address this issue, we propose in this paper a novel method for automatic facial landmark localization in 3D face data designed specifically to address appearance variability caused by significant pose variations. Our method builds on recent cascaded regression-based methods to facial landmarking and uses a gating mechanism to incorporate multiple linear cascaded regression models each trained for a limited range of poses into a single powerful landmarking model capable of processing arbitrary-posed input data. We develop two distinct approaches around the proposed gating mechanism: (1) the first uses a gated multiple ridge descent mechanism in conjunction with established (hand-crafted) histogram of gradients features for face alignment and achieves state-of-the-art landmarking performance across a wide range of facial poses and (2) the second simultaneously learns multiple-descent directions as well as binary features that are optimal for the alignment tasks and in addition to competitive landmarking results also ensures extremely rapid processing. We evaluate both approaches in rigorous experiments on several popular datasets of 3D face images, i.e., the FRGCv2 and Bosphorus 3D face datasets and image collections F and G from the University of Notre Dame. The results of our evaluation show that both approaches compare favorably to the state-of-the-art, while exhibiting considerable robustness to pose variations.

Keywords Facial landmarking · Feature learning · Hand-crafted features · Pose variations

1 Introduction

Face alignment or facial landmarking refers to the task of locating salient facial features in facial images, which is of paramount importance in various applications including face registration and recognition [21, 41], expression recognition [49], face tracking [37], normalization of facial pose, size, and expressions [16], face synthesis from morphable models and facial animation [23], to name a few. In real-world scenarios where face images are often acquired in uncontrolled conditions, one has to deal with various unfavorable factors that adversely affect landmarking performance including pose, expression, and illumination variations as well as partial occlusions of the facial areas. These factors influence the appearance of the facial features in traditional 2D images, e.g., [12] but also in 3D (or better said 2.5D) face data used in this work.[1] Although some of the existing landmark localization procedures promise to be (at least partially) robust to some of the factors mentioned above (e.g., [9, 25, 36]), reliable localization of facial landmarks in the presence of highly variable nuisance factors still remains a considerable challenge.

✉ Janez Križaj
janez.krizaj@fe.uni-lj.si

[1] Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia

[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia

[1] To be strict, we consider 2.5D images in this work, but will use the terms 3D images and depth images interchangeably to refer to this type of data throughout the paper.

With the advancement of 3D acquisition technology, landmark localization on 3D facial data has recently been researched extensively [13, 36]. Many of the 3D landmarking techniques proposed in the literature in the last few years rely on the so-called cascaded regression framework, where facial landmarks are estimated by regressing from facial features to landmark locations in a cascaded (iterative) manner [2]. Techniques following this framework made considerable advancements toward robust facial landmarking, although they generally still use handcrafted features, such as the scale-invariant feature transforms (SIFT) or histograms of oriented gradients (HOG) [2, 3, 39]. Additionally, these methods typically rely on a single regression model in each stage of the cascade to estimate the facial landmarks regardless of the facial characteristics. However, as facial appearance is a complex function of various factors, such as facial shape, pose, incident illumination, expression, and occlusion, a single model is often not sufficient to capture the broad range of variability commonly encountered with facial-image data and to robustly estimate the location of the most salient facial features.

To address this problem, we propose in this paper a novel gating mechanism that incorporates multiple cascaded regression-based models each trained for a narrow range of facial posses into a single (coherent) landmarking model that is able to reliably estimate the location of salient facial features from arbitrary-posed input face data. The combination of simpler view-specific landmarking approaches provides the combined gating-based model the necessary expressive power to describe the considerable appearance variability typically seen with 3D face data captured under different facial poses and consequently allows it to reliably estimate the landmark locations regardless of the facial pose of the input image. The model is partially motivated by the success of earlier methods designed for 2D images that combine multiple landmarking models trained for face alignment of different views, e.g., [4, 15, 29, 47, 50], but unlike these early methods does not rely on parametric appearance or shape models.

We develop two distinct facial landmarking approaches around the proposed gating mechanism. The first relies on a combination of the Gated multiple RIdge Descent (GRID) mechanism and established HOG features and as illustrated in Fig. 1 achieves remarkable landmarking performance across a broad range of pose variation. Even for poses with yaw rotations of up to $\pm$ 90°, the model is still able to reliably estimate the location of salient facial features. The second approach again relies on the introduced gating mechanism, but in addition to the cascaded regression models needed for face alignment of each group of poses, it also learns a feature representation that is used with the regression models for landmark estimation. Specifically,

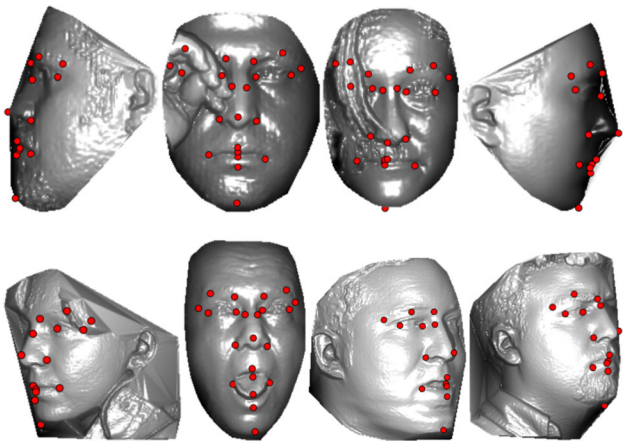


Fig. 1 Sample results of the landmarking GRID approach proposed in this paper. Our model is able to reliably estimate the location of salient facial features in 3D face data even in the presence of large pose variations, i.e., with yaw angles up to $\pm$ 90°

the model Simultaneously learns MUltiple-descent directions as well as binary Features (SMUF) that are optimal for the alignment task and due to their binary nature also ensure extremely fast execution times. This second approach follows recent trends in computer vision and aims to learn the feature representation that is optimal for a given tasks, but different from deep learning models that are typically used for feature learning [40, 42], it uses a computationally much simpler scheme, where binary features are learned based on a learning objective that can be solved using standard optimization procedures.

To evaluate the proposed landmarking approaches, we conduct experiments on multiple datasets of 3D face images, i.e., the Face Recognition Grand Challenge v2 (FRGCv2) [26], the Bosphorus 3D face dataset [31], and the University of Notre Dame dataset (collections F and G, hereinafter referred to as UND dataset) [46]. We present extensive experiments and comparisons with state-of-the-art methods from the literature. The results of our evaluation show that the GRID ensures state-of-the-art performance for facial landmark localization in 3D face data across pose, while SMUF yields not only competitive landmarking accuracy but is also extremely fast.

Our main contributions in this paper are as follows:

- We propose a gating mechanism for face alignment in 3D face data that allows us to combine multiple alignment models and foster the combined power of the combined models for face alignment across pose. The use of multiple models adds to the overall localization performance, since each model needs to account only for a limited set of plausible facial variations. With this approach, reliable landmarking is possible even under large head rotations such as profile facial images, with

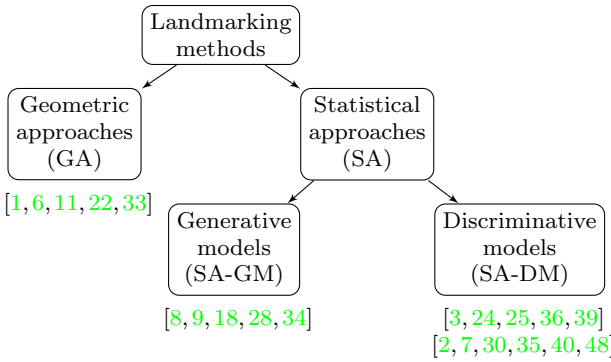


Fig. 2 Taxonomy of 3D landmarking approaches as discussed in "Related work" section. Recent work is largely focusing on statistical approaches, where landmarking is learned from annotated training data using either generative or discriminative models

yaw rotations up to ± 90° (see Fig. 1), where many competing methods fail.

- We develop two distinct landmarking approaches based on the introduced gating mechanism, where one is optimized for performance and the second one is optimized for both performance and speed. We evaluate both approaches in rigorous experiments on multiple face datasets and report competitive performance in comparison to competing methods from the literature.
- We study different configurations of the proposed approaches and investigate their behavior when localizing specific facial landmarks.

The rest of the paper is organized as follows: In Sect. 2, we describe prior work in the field of facial landmarking with the goal of providing the necessary context for our contributions. In Sect. 3, we present our gating mechanism and the GRID and SMUF alignment techniques. We describe experiments conducted to evaluate the performance of the proposed methods in discuss results in Sect. 4. We conclude the paper with some final remarks and future research challenges in Sect. 5.

2 Related work

Numerous methods have been proposed in the literature for the task of automatic facial landmark localization over recent years. In this section, we present a brief overview of these methods with a focus on alignment techniques that work on 3D images. These techniques can be categorized in various ways, but here we chose to perform a categorization as shown in Fig. 2. We classify existing techniques into two groups: (1) techniques that are entirely dependent on geometric information and derive prior knowledge about the facial structure and location of facial landmarks by defining a number of heuristic rules and (2) techniques that rely on trained statistical models. The latter group of techniques is further divided according to the type of the model utilized into generative and discriminative methods. A high-level comparison of the related works discussed in this section is given in Table 1.

2.1 Geometric approaches

Geometric approaches to facial landmarking are generally training free and depend solely on the geometric information such as surface curvature or shape index values. A number of rules and heuristics encodes the prior knowledge about the relationships between adjacent landmarks (e.g., the nose tip lies on the face symmetry axis, eyes are located above the nose tip, etc.). In most cases, the rules used to define the location of facial landmarks require the face to be in upright and near-frontal positions. Moreover, a common downside of these methods is that the landmarks are detected in sequence (commonly starting by detecting a nose tip) and the success rate of finding the next landmark in the sequence is dependent on successfully locating the preceding landmark in the sequence. With these methods, an incorrect detection of one landmark affects the detection of all subsequent landmarks.

Exemplar geometric methods [1, 6, 11, 22, 33] start by detecting the nose tip and use its location to constrain the search space of the remaining landmarks. Landmark detection can be grounded on the analysis of Gaussian curvatures [11], profile curvatures [6], x and y coordinate projections of the depth data [33], shape index values and facial symmetry lines [1], or horizontal slices of range images [22], to name a few.

2.2 Statistical approaches

Statistical landmarking approaches also exploit local shape information around candidate landmark locations. Additionally, these methods derive some prior knowledge from the training data about the location constraints and encode the acquired knowledge into a statistical model. Thus, these methods require a training set of facial images with annotated landmarks. Unlike training-free geometric approaches, statistics are utilized uniformly for all landmarks, as there is no need for specific rules for each individual landmark. Since all landmarks are handled simultaneously approaches from this group are typically more robust to local distortions, missing data, and occlusions of individual landmarks. However, the fact that statistical methods generally address a complete set of landmarks defined by the model could prove to be a problem when a large number of landmarks is (self-)occluded or data are missing from the input images due to

Table 1 Summary of the existing 3D facial landmark detection methods

Author	Type[a]	Procedure	Learning algorithm	Features
Mian et al. [22]	GA	Nose tip detection through the analysis of horizontal slices	Training free	Depth data
Faltemier et al. [6]	GA	Rotated profile signatures	Training free	Rightmost 3D profile lines
Gupta et al. [11]	GA	ICP coarse alignment, heuristic rules	Training free	Surface curvatures
Segundo et al. [33]	GA	Clustering-based face detection, heuristic rules	Training free	Depth relief curves, surface curvatures
Alyüz et al. [1]	GA	Facial symmetry axis, heuristic rules	Training free	Shape index, Gaussian curvature
Passalis et al. [24]	SA-DM	Candidate landmarks fitting to facial landmark model	PCA	Shape index, spin images
Zhao et al. [48]	SA-DM	3D statistical facial feature model	PCA-based learning	Range map, intensity map
Fanelli et al. [7]	SA-DM	Random forest-based voting approach	Random forests	Binary tests
Fanelli et al. [8]	SA-GM	Random forests-based regression, AAM	Random forests	Binary tests with trees in forest
Perakis et al. [25]	SA-DM	Candidate landmark fitting to facial landmark model	PCA	Shape index, spin images
Smolyanskiy et al. [34]	SA-GM	2D AMM and 3D morphable face model	PCA	RGBD values
Liu et al. [18]	SA-GM	Normalized cross correlation and depth-based AAM	PCA	Shape and depth values
Song et al. [35]	SA-GM	Local coordinate coding	Coupled dict. learning	Spin images, synthesized features
Cao et al. [3]	SA-DM	Cascaded regression	Linear regression	Shape indexed and depth features
Sukno et al. [36]	SA-DM	Combinatorial search constrained by a flexible shape model	PCA	APSC descriptors
Camgöz et al. [2]	SA-DM	Cascaded ridge regression	Ridge regression	Multi-scale HOG
Feng et al. [9]	SA-DM	Cascaded collaborative regression	Weighted ridge regression	Dynamic multi-scale HOG
Rai et al. [28]	SA-GM	3D constrained local models	ICA, point dist. model	LBP descriptors
Wang et al. [39]	SA-DM	Joint head pose and facial landmark regression	Classif. guided casc. regression	Random forest feature selection
Liu et al. [17]	SA-DM	Hidden Markov models (HMMs)	HMM learning	Spin image
Wang et al. [40]	SA-DM	CNN-based feature extraction and landmark regression	Pre-trained CNN fine-tuning	CNN-based global and local features

[a]For an explanation of the respective method types, see Fig. 2

acquisition errors. Recently, efforts have been made to handle such problems, e.g., [36] use a flexible shape model that works even with an incomplete set of landmarks.

In terms of the type statistical model used, approaches from this group can be divided into techniques that rely either on generative or on discriminative models. We discuss both types of techniques in the following two subsections.

2.2.1 Generative approaches

Landmark locations can be modeled by generative models, such as active appearance models, active shape models [8, 18, 34], or morphable models [9]. Techniques from this group often learn face appearances (and/or shape) by conducting principal component analysis (PCA) [38] on manually annotated and aligned training data. Given a test

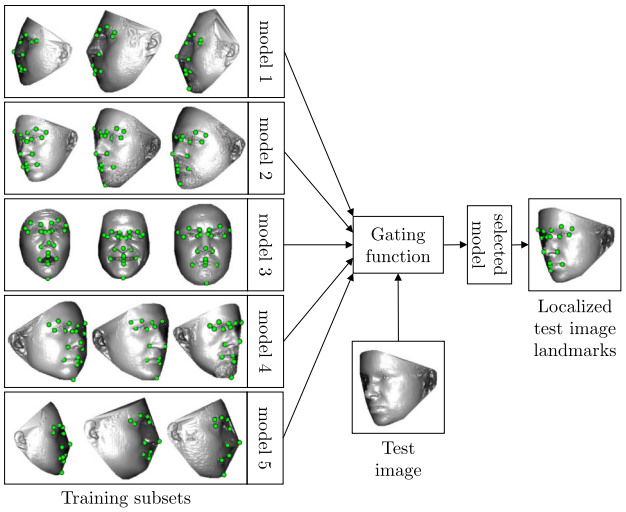


Fig. 3 Schematic representation of the gating mechanism used in this work. Multiple landmarking models (each containing a cascade of regression models) are trained during the learning stage. At run time, a gating function is used to select the landmarking model that best fits the characteristics of the test data

image, alignment/fitting is achieved by minimizing the difference between the current estimate of the appearance (and/or shape) and the test face. Generally, generative approaches are often computationally expensive and may perform poorly in the presence of occlusions, pose expression, and illumination variations due to the involved fitting procedure.

2.2.2 Discriminative approaches

More recent methods to face alignment focus mostly on discriminative approaches that learn a mapping function that predicts the shape, i.e., landmark locations, directly from corresponding image features. Methods from this category typically offer better landmark localization performance when compared to generative models, especially for faces with greater variability in appearance [2, 3, 9, 36]. These methods commonly incorporate shape constraints into the models and use local descriptors that are more robust to appearance variations than conventional depth/intensity pixel values used with generative approaches. Discriminative approaches include random forests [7], graph matching [30], cascaded regression [2, 3, 9, 39], specifically tailored shape models [24, 25, 36, 48], hidden Markov models [17], and convolutional neural networks [40].

The landmarking techniques proposed in this work fall into the group of discriminative approaches. They build on recent face alignment techniques that rely on cascaded regression models that have proven highly successful for landmark localization from 2D face images, e.g., [44, 45]. However, compared to these models, our solutions exhibit unique features, such as the novel gating mechanism for exploiting multiple pose-specific landmarking models and the ability to incorporate task-specific binary features into the landmarking procedure.

3 Methodology

In this section, we describe GRID and SMUF, two novel facial landmarking approaches built around the gating mechanism illustrated in Fig. 3. As can be seen, the gating mechanism partitions the search space for the landmark localization procedure into a number of sub-domains, where each sub-domain encompasses a range of similar facial poses. A separate landmarking model is then trained for each of the sub-domains, and the gating mechanism is used to select the most suitable landmarking model for the given test image. Based on this overall framework, we develop two distinct landmarking techniques, which are described next.

3.1 GRID description

We design GRID (Gated Multiple Ridge Descent) in line with the powerful cascaded regression framework to face alignment, where landmark locations (or in other words, the facial shape) are estimated by regressing from facial features to landmark locations in a cascaded manner. In the first step of this framework, features are extracted from some initial landmark configuration (estimated from the training data) and a regression model is applied on the extracted features to predict landmark updates to better align the landmarks with the actual test image. The update results in a new landmark configuration that forms the basis for the next step in the cascade. The entire procedure is then repeated multiple times and, thus, sequentially refines the predicted locations of the facial landmarks in the test image.

With GRID, we train multiple cascaded regression models and integrate them into a gated approach that is robust to pose variations. While different regression models and feature representation have been proposed in the literature for facial landmarking, we built GRID around the supervised descent method (SDM) from [44] that has not proven successful only for facial landmark localization in 2D images [43], but also for alignment of 3D face images, as we have shown in [2, 14].

3.1.1 Background

To train the regression models needed for landmarking, SDM requires a number of facial images $\{\boldsymbol{I}_n\}_{n=1}^{N}$, where

each image $\boldsymbol{I}$ has L landmarks annotated in the form of a shape vector $\boldsymbol{x}_* \in \mathbb{R}^{2L\times 1}$. The landmark localization task is then posed as a minimization problem over $\Delta\boldsymbol{x}$:

$$\arg\min_{\Delta\boldsymbol{x}} \|h(\boldsymbol{I}, \boldsymbol{x}_1 + \Delta\boldsymbol{x}) - \boldsymbol{\phi}_*\|^2, \tag{1}$$

where h is a feature extraction function, $\boldsymbol{\phi}_* = h(\boldsymbol{I}, \boldsymbol{x}_*)$ are features extracted around the ground truth landmarks $\boldsymbol{x}_*$, $\boldsymbol{x}_1$ is an initial landmark configuration, and $\Delta\boldsymbol{x}$ is a landmark update (known for the training data).

Equation (1) represents a nonlinear least squares problem and, in general, has no closed-form solution. However, it was shown in [43] that the problem can be solved through a cascade of least squares regression problems. Thus, for each step k in the cascade, the solution of the least squares problem is given in the form of a regression matrix $\boldsymbol{R}_k$ (also referred to as a descent map, DM) that can be used to predict the update of the landmark locations from the current image features. The learning algorithm is formulated as a minimization of the loss between the true shape updates $\Delta\hat{\boldsymbol{x}}_k^n = \boldsymbol{x}_k^n - \boldsymbol{x}_*^n$ and the expected updates over all training images, i.e.,

$$\arg\min_{\boldsymbol{R}_k} \sum_n \left\|\Delta\hat{\boldsymbol{x}}_k^n - \boldsymbol{R}_k\left(\boldsymbol{\phi}_k^n - \overline{\boldsymbol{\phi}_*}\right)\right\|^2, \tag{2}$$

where n is a training-image index and $\overline{\boldsymbol{\phi}_*}$ is an average feature vector computed from the ground truth locations $\boldsymbol{x}_*^n$ over all training images. Equation (2) now represents a sequence of ordinary least squares regression problems that can be solved in closed form.

During test time, the algorithm starts with some initial landmark locations $\boldsymbol{x}_1$, for which the face shape (landmark configuration) is defined by the average landmark locations of training images and the position of the face shape is determined by the face detection procedure[2], and sequentially updates the initial estimate to obtain the final landmark locations, i.e.,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{R}_k\left(\boldsymbol{\phi}_k - \overline{\boldsymbol{\phi}_*}\right), \tag{3}$$

so that the final shape $\boldsymbol{x}_k$ converges to $\boldsymbol{x}_*$ for all training images. The number of steps K in the cascade, where $k = 1, 2, \ldots K$ commonly varies depending on the implementation, but usually values of K are between 3 and 10.

3.1.2 Ridge regression

The original SDM [43] formulation uses a least squares solution of Eq. (2) to learn the DMs, i.e.,

$$\boldsymbol{R}_k = \hat{\boldsymbol{X}}_k\hat{\boldsymbol{\Phi}}_k^\top\left(\hat{\boldsymbol{\Phi}}_k\hat{\boldsymbol{\Phi}}_k^\top\right)^{-1}, \tag{4}$$

[2] The average shape is always placed consistently with respect to the detected facial region.

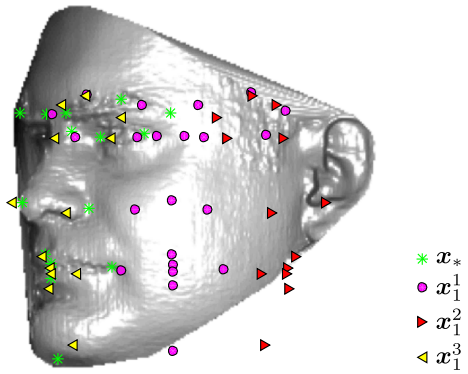


Fig. 4 Multiple landmark initializations $\{\boldsymbol{x}_1^z\}_{z=\{1:3\}}$ and the ground truth landmarks $\boldsymbol{x}_*$ superimposed on an example test image. The gating mechanism used in this work determines the pose of the test image (and consequently selects a landmarking cascade) by comparing the features extracted from different shape initializations of the test image to the average features extracted from the true landmark locations of all images in the pose-specific training sets

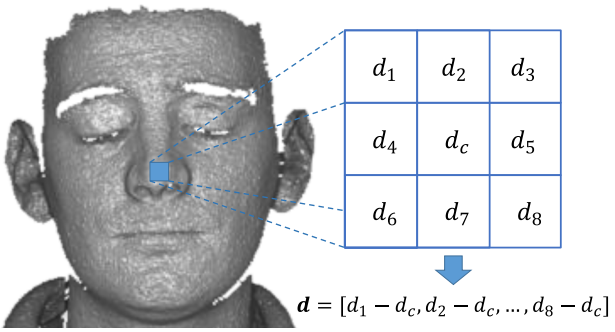


Fig. 5 We use depth-difference vectors $\boldsymbol{d}$ as the basis for binary feature calculation as proposed in [19, 20]. The example image above shows how one such vector is computed for a selected landmark. The local pixel neighborhood shown here is of size 3×3 and is selected only for illustration purposes. We use larger neighborhoods for the actual SMUF implementation

where $\hat{\boldsymbol{X}}_k$ is a shape matrix with nth column $\Delta\hat{\boldsymbol{x}}_k^n$ and $\hat{\boldsymbol{\Phi}}_k$ is a feature matrix with nth column $\boldsymbol{\phi}_k^n - \overline{\boldsymbol{\phi}_*}$. To solve Eq. (4), one needs to compute the inverse of $\hat{\boldsymbol{\Phi}}_k^\top\hat{\boldsymbol{\Phi}}_k$, which, however, may be singular when the size of the feature vectors is too large or when the features are correlated. To overcome this issue, the original SDM applies PCA [38] to the image features before inverting the matrix.

However, we have shown in [2] that better landmarking performance is achieved if ridge regression is used in the original feature space instead of least squares regression in the PCA subspace. The optimization function in (2) in this case can be written as

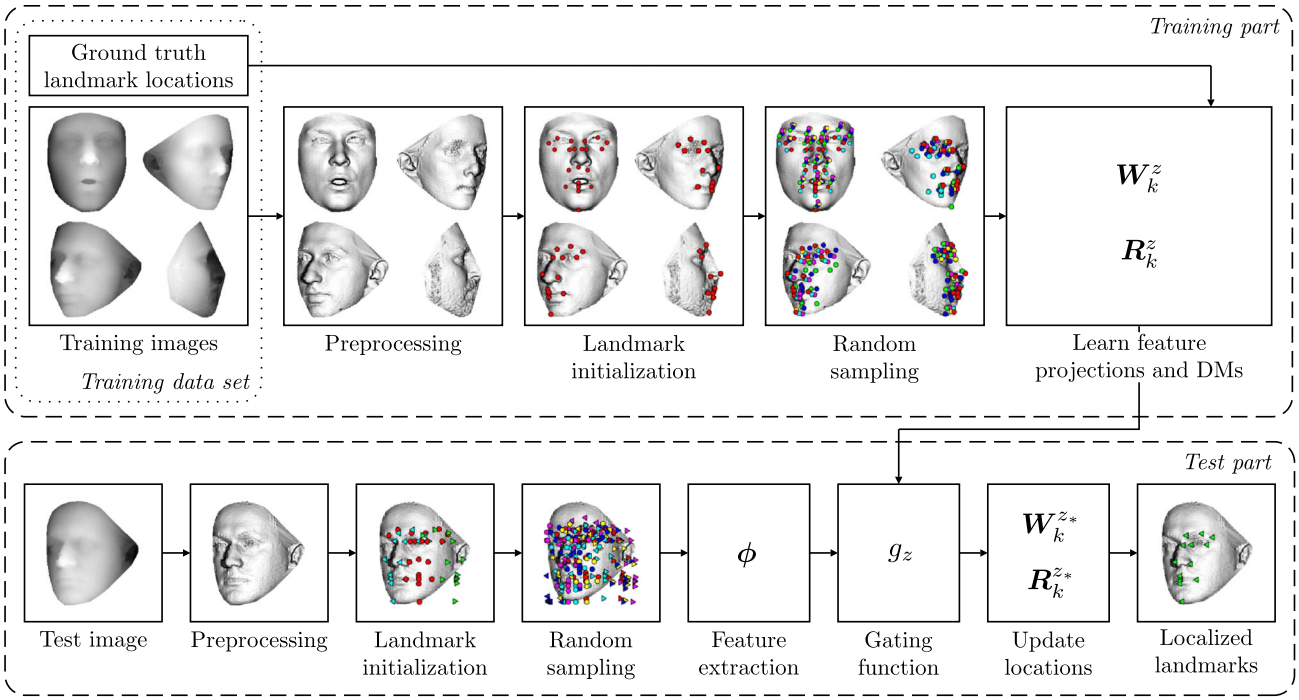


Fig. 6 Overview of the training and testing stages of the GRID and SMUF landmarking techniques. Both techniques use a similar processing pipeline, but the SMUF approach also learns features (marked by $\boldsymbol{W}_k^z$) in each stage of the training procedure in addition to the landmarking cascade (marked by $\boldsymbol{R}_k^z, k = 1, 2, \ldots, K$) learned by GRID

$$\arg\min_{\boldsymbol{R}_k} \sum_n \left\| \hat{\boldsymbol{x}}_k^n - \boldsymbol{R}_k \left(\boldsymbol{\phi}_k^n - \overline{\boldsymbol{\phi}_*} \right) \right\|^2 + \gamma_k \|\boldsymbol{R}_k\|^2, \tag{5}$$

where γ_k denotes a regularization factor and the solution of Eq. (5) is computed as

$$\boldsymbol{R}_k = \hat{\boldsymbol{X}}_k \hat{\boldsymbol{\Phi}}_k^\top \left(\hat{\boldsymbol{\Phi}}_k \hat{\boldsymbol{\Phi}}_k^\top + \gamma_k \mathbf{I} \right)^{-1}, \tag{6}$$

where $\mathbf{I}$ is an identity matrix. The regularization factor $\gamma_k \geq 0$ controls the general instability of the least squares estimate. Selecting a suitable value for γ_k avoids over-fitting and helps to produce estimates of $\boldsymbol{R}_k$ that generalize better to unseen data.

3.1.3 Gated multiple ridge descent

Experimental results in [2, 43, 45] have shown that the original SDM achieves remarkable landmarking performance on various 2D and 3D datasets. However, it still tends to perform poorly when, for example, large head rotations are present in the facial data [45]. Such rotations cause complex facial appearance variations that are difficult to model and hard to account for when using only a single DM in each step of the landmarking cascade.

To increase the robustness of the model to pose variations, we propose to exploit multiple DMs $\{\boldsymbol{R}_k^z\}_{z=\{1:Z\}}$ such that each of the Z DMs accounts for a specific range of head rotations, as illustrated in Fig. 3. Toward this end, we partition the available training images $\{\boldsymbol{I}_n\}_{n=1}^N$ into Z pose-specific subsets and train separate regression cascades for each subset in line with Eq. (6).

Once all Z cascades (series of DMs) are trained, a gating function g_z is used to select the most suitable DM (from the Z DMs available in the first cascade stage) for a given test image. The selection procedure begins by computing

Table 2 Overview of the datasets used for experimentation

Database	#images	#subjects	Variability
FRGCv2 [26]	4007	975	Expression
Bosphorus [31]	4666	105	Expression, occlusion, and orientation
UND [46]	1680	537	Orientation

The FRGCv2 dataset is among the most frequently used datasets of 3D face images, whereas the Bosphorus and UND datasets contain challenging images with a high degree of variability in face orientations and are, hence, well suited for our experiments

Table 3 Mean localization errors (and standard deviations) for the GRID and SMUF methods on the Bosphorus dataset

Variation	Number of descent maps					
	1 DM		3 DMs		5 DMs	
	GRID	SMUF	GRID	SMUF	GRID	SMUF
Frontal	3.0 ± 1.7	3.1 ± 1.8	3.0 ± 1.7	3.1 ± 1.8	3.0 ± 1.7	3.1 ± 1.8
Yaw $\leq \pm 10°$	3.1 ± 1.8	3.2 ± 1.9	3.1 ± 1.8	3.5 ± 3.0	3.1 ± 1.8	3.5 ± 3.0
Yaw $\leq \pm 20°$	3.2 ± 1.9	3.6 ± 2.5	3.3 ± 2.0	3.7 ± 2.9	3.3 ± 2.0	3.7 ± 2.9
Yaw $\leq \pm 30°$	3.5 ± 2.1	5.2 ± 5.1	3.4 ± 2.0	3.7 ± 2.7	3.4 ± 2.0	3.7 ± 2.7
Yaw $\leq \pm 45°$	5.2 ± 4.1	12.1 ± 11.4	3.4 ± 2.1	3.8 ± 2.6	3.4 ± 2.1	3.9 ± 3.1
Yaw $\leq \pm 90°$	9.6 ± 10.1	15.6 ± 13.1	5.2 ± 5.1	7.8 ± 8.5	3.5 ± 2.1	4.0 ± 3.3
Expressions	3.4 ± 2.0	3.6 ± 2.3	3.4 ± 2.0	3.6 ± 2.3	3.4 ± 2.0	3.6 ± 2.3
Occlusions	3.9 ± 2.5	3.9 ± 2.5	3.9 ± 2.5	3.9 ± 2.5	3.9 ± 2.5	3.9 ± 2.5

Results are reported for different variants of both landmarking techniques implemented with 1, 3, or 5 regression cascades. The best overall performance is achieved with 5 cascades for both techniques. The results also show that the gating function always selects the correct cascade—observe results for frontal images across the different landmarking variants

features $\{\boldsymbol{\phi}_1^z\}_{z=\{1:Z\}}$ from the initial landmark locations $\{\boldsymbol{x}_1^z\}_{z=\{1:Z\}}$ in the test image. Here, the initial landmark locations $\boldsymbol{x}_1^z$ (see Fig. 4) are computed by averaging the ground truth shapes over all training images from the zth training subset:

$$\boldsymbol{x}_1^z = \{\overline{\boldsymbol{x}_*^n}\}_{n\in z}. \tag{7}$$

The most fitting DM for the given test image is then selected based on the output of the gating function g_z:

$$g_z(\boldsymbol{\phi}_1^z) = \sqrt{\frac{1}{m}\left(\boldsymbol{\phi}_1^z - \overline{\boldsymbol{\phi}_*}\right)^\top \boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\phi}_1^z - \overline{\boldsymbol{\phi}_*}\right)}, \tag{8}$$

where m is the feature vector length (in the case of SIFT $m = 128 \times L$) and $\overline{\boldsymbol{\phi}_*}$ and $\boldsymbol{\Sigma}_*$ are the average and the covariance matrix of the ground truth features over the subset z, respectively. We select the subset z_*, for which the gating function outputs the lowest value:

$$z_* \in \{1, \ldots, Z\} : g_{z_*} = \min_z(g_z). \tag{9}$$

By doing so, we reliably choose the DM $\boldsymbol{R}_1^{z_*}$ that has been trained on images with similar face orientations to the orientation of the test face. For all the subsequent steps k, we use the DMs $\boldsymbol{R}_k^{z_*}$ that correspond to the z_*th subset selected in the first step $k = 1$ and for efficiency reasons due not change the regression cascade in subsequent steps. Location updates on a given test image are thus computed as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{R}_k^{z_*}\left(\boldsymbol{\phi}_{k-1}^{z_*} - \overline{\boldsymbol{\phi}_*}\right). \tag{10}$$

The described procedure results in significantly improved performance in the case of large head rotations as shown later in Sect. 4.

It needs to be noted that we rely on HOG features to implement the feature extraction function h in GRID. We select HOG features because of their proven performance in prior landmarking models, e.g., [2, 14].

3.2 SMUF description

The GRID landmarking approach presented in the previous section relies on HOG features to encode the appearance of the facial landmarks during face alignment. With SMUF (Simultaneous MUlti-descent regression and binary Feature learning), we take a step further and try to learn facial features that are optimal for face alignment. We choose to learn binary features, due to their simplicity and most of all computational simplicity. In the following subsection, we first review the idea of binary feature learning and then develop SMUF approach that jointly learns a landmarking model as well as corresponding binary features that are optimal for this task.

3.2.1 Binary feature learning

Hand-crafted binary features, such as local binary patterns (LBPs) [27] represent powerful image descriptors that have proven highly effective in various computer vision tasks. These features typically rely on pixel comparisons within a local neighborhood and heuristic rules to encode the pixel comparisons into binary codes. As such, they may be suboptimal and better features could potentially be constructed by learning binary features based on some dedicated learning objective.

Gong et al. [10], for example, propose a learning objective where binary features are learned from an initial image representation $\boldsymbol{d}$, such that the quantization error is

minimized. Since binary features $\boldsymbol{\phi}$ (containing only 0s and 1s) can be computed from $\boldsymbol{d}$ as

$$\boldsymbol{\phi} = 0.5(\,\mathrm{sgn}\,(\boldsymbol{W}^\top \boldsymbol{d}) + 1), \tag{11}$$

where $\boldsymbol{W}$ is a matrix of hash functions that defines the length of the binary code and $\mathrm{sgn}\,(.)$ stands for the signum function. The learning objective L_q that needs to be minimized over $\boldsymbol{W}$ on some training data can be written as

$$L_q = \left\|\boldsymbol{\phi} - 0.5 - \boldsymbol{W}^\top \boldsymbol{d}\right\|^2. \tag{12}$$

It was shown by Lu et al. [19, 20] that descriptive binary image features can be computed based on the above quantization scheme if pixel (or depth in our case) difference values are used as input $\boldsymbol{d}$ for binarization. For SMUF, we follow this approach and compute one depth-difference vector $\boldsymbol{d}$ for each considered landmark, as illustrated in Fig. 5.

3.2.2 Simultaneous DM and feature learning

The learning objective presented in the previous section is focused on representational power, as the binary features are computed in a manner that minimizes a quantization loss. To make the features useful for landmarking, we now formulate a joint optimization function that allows us to simultaneously learn a regression cascade and corresponding binary features that are optimal for the landmarking task.

Let $\boldsymbol{D}_k = [\boldsymbol{d}_k^1, \ldots, \boldsymbol{d}_k^{LN}]$ be a set of depth-difference vectors extracted from patches centered at the facial landmarks $\boldsymbol{X}_k = [\boldsymbol{x}_k^1, \ldots, \boldsymbol{x}_k^{LN}]$ and k stands for the cascade stage, $k = 1, 2, \ldots, K$. The depth-difference-vector matrix $\boldsymbol{D}_k$ is mapped to a binary feature matrix $\boldsymbol{\Phi}_k$ as follows:

$$\boldsymbol{\Phi}_k = 0.5(\,\mathrm{sgn}\,(\boldsymbol{W}_k^\top \boldsymbol{D}_k) + 1), \tag{13}$$

where $\boldsymbol{W}_k$ is a feature projection matrix and $\mathrm{sgn}\,(.)$ is again the signum function. To learn $\boldsymbol{W}_k$, we formulate the following optimization problem by rewriting (5) into matrix form and extending it with the additional constraint C_2:

$$\begin{aligned}\underset{\boldsymbol{R}_k, \boldsymbol{W}_k}{\arg\min}\, C &= C_1 + \lambda C_2 \\ &= \left\|\hat{\boldsymbol{X}}_k - \boldsymbol{R}_k \tilde{\boldsymbol{\Phi}}_k\right\|^2 + \gamma \|\boldsymbol{R}_k\|^2 \\ &\quad + \lambda \left\|\boldsymbol{R}_{\boldsymbol{k}} (\tilde{\boldsymbol{\Phi}}_k - 0.5 - \boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k)\right\|^2,\end{aligned} \tag{14}$$

where

$$\tilde{\boldsymbol{\Phi}}_k = 0.5(\,\mathrm{sgn}\,(\boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k) + 1) \tag{15}$$

and $\hat{\boldsymbol{D}}_k = \boldsymbol{D}_k - \boldsymbol{D}_*$, where $\boldsymbol{D}_*$ are the depth-difference values of the ground truth landmark locations. As already emphasized above, the objective of C_2 is to minimize the quantization loss between the original depth-difference values and the binarized features, so that most of the depth-difference energy can be preserved in the learned binary features.

We find optimal values for $\boldsymbol{R}_k$ and $\boldsymbol{W}_k$ by an iterative optimization procedure, where $\boldsymbol{W}_k$ is initialized to a random orthogonal matrix. If we assume a fixed $\boldsymbol{W}_k$ and compute a partial derivative of C in (14) with respect to $\boldsymbol{R}_k$ and set the derivative to zero, we obtain the following solution for $\boldsymbol{R}_k$:

$$\begin{aligned}\boldsymbol{R}_k = &\hat{\boldsymbol{X}}_k \tilde{\boldsymbol{\Phi}}_k^\top \Big[\tilde{\boldsymbol{\Phi}}_k \tilde{\boldsymbol{\Phi}}_k^\top + \gamma_k \mathbf{I} \\ &+ \lambda (\tilde{\boldsymbol{\Phi}}_k - 0.5 - \boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k)(\tilde{\boldsymbol{\Phi}}_k - 0.5 - \boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k)^\top\Big]^{-1}.\end{aligned} \tag{16}$$

In the next step, we aim to learn $\boldsymbol{W}_k$ with a fixed $\boldsymbol{R}_k$ and, hence, rewrite (14) as follows:

$$\begin{aligned}\underset{\boldsymbol{W}_k}{\arg\min}\, C = &\left\|\hat{\boldsymbol{X}}_k - \boldsymbol{R}_k \boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k\right\|^2 + \gamma \|\boldsymbol{R}_k\|^2 \\ &+ \lambda \left\|\boldsymbol{R}_{\boldsymbol{k}} (\tilde{\boldsymbol{\Phi}}_k - 0.5 - \boldsymbol{W}_k^\top \hat{\boldsymbol{D}}_k)\right\|^2.\end{aligned} \tag{17}$$

If we differentiate (17) with respect to $\boldsymbol{W}_{\boldsymbol{k}}$ and set the derivative to zero, we obtain the following update rule for $\boldsymbol{W}_k$:

$$\boldsymbol{W}_k = \left[(\boldsymbol{R}_k^{-1} \hat{\boldsymbol{X}}_k + \lambda (\tilde{\boldsymbol{\Phi}}_k - 0.5)) \hat{\boldsymbol{D}}_k\right]^\top / (1 + \gamma + \lambda). \tag{18}$$

The two optimization steps from (16) to (18) are then repeated until both $\boldsymbol{R}_{\boldsymbol{k}}$ and $\boldsymbol{W}_k$ converge.

Once a stable version of $\boldsymbol{R}_{\boldsymbol{k}}$ and $\boldsymbol{W}_k$ are obtained, we compute the shape update in accordance with

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \boldsymbol{R}_k \tilde{\boldsymbol{\Phi}}_k \tag{19}$$

and repeat the entire procedure for the next stage in the cascade. Note that because $\tilde{\boldsymbol{\Phi}}_k$ is binary, location updates can be computed extremely quickly by simply summing up (specific) rows from $\boldsymbol{R}_k$.

Finally, we compute separate regression cascades and projection matrices for each of the Z training subsets, that is, for each considered group of poses, and integrate the computed cascades into the overall SMUF approach using the same gating mechanism as described above for the GRID approach.

3.3 Training and testing of GRID and SMUF

The overall processing pipeline for the SMUF landmarking approach is shown in Fig. 6. The procedure for GRID is identical, except for the fact that no features are learned during training.

The training stage for both methods begins by preprocessing all N training images $\{\boldsymbol{I}_n\}_{n=1:N}$ where a depth component of the surface normal is computed in each pixel

instead of using original depth values. In each image, the face is detected using a simple clustering procedure [32] and initial landmark locations $\boldsymbol{x}_1^n$ are set based on the detected facial area. To capture the variance of the face detection procedure and to enlarge the amount of training data, we define additional initial landmark locations for each training image by randomly sampling scale and displacement parameters for the detected area from a normal distribution. Starting from the initial locations matrix $\boldsymbol{X}_1$ along with the ground truth locations $\boldsymbol{X}_*$, a number of DMs $\boldsymbol{R}_k^z$ (and for SMUF also projection matrices $\boldsymbol{W}_k^z$) are learned. The updates (16) and (18) are iteratively recomputed till convergence (we empirically estimated that 4 steps are sufficient) for each shape update step k and each of the Z training subsets.

When a test image is presented to the landmarking procedure, it goes through the same face detection, preprocessing, and feature extraction steps as the training images. DMs and feature projections are then selected as described in Sect. 3.1.3 and the final landmark locations are computed based on (10). The pseudocode of the GRID method is summarized in Algorithm 1, while the steps for the SMUF method are outlined in Algorithm 2.

Algorithm 1: GRID

procedure `Training()`
- **Input :** Training images $[\boldsymbol{I}_1, \cdots, \boldsymbol{I}_N]$ with annotated landmarks $\boldsymbol{X}_* = [\boldsymbol{x}_*^1, \cdots, \boldsymbol{x}_*^N]$ and initial locations $\boldsymbol{X}_1 = [\boldsymbol{x}_1^1, \cdots, \boldsymbol{x}_1^N]$.
- **Output:** Descent maps $\boldsymbol{R}_k^z$.
- **for** $z = 1 : Z$ **do**
  - Select z-th training subset.
  - **for** $k = 1 : K$ **do**
    - Extract HOG features $\boldsymbol{\Phi}_k^z$.
    - Update $\boldsymbol{R}_k^z$ using (6).
    - Update shape $\boldsymbol{X}_k^z$ using (3).

procedure `Evaluation()`
- **Input :** Test image $\boldsymbol{I}$, initial landmark locations computed by (7), descent maps $\boldsymbol{R}_k^z$.
- **Output:** Final landmark locations.
- **for** $k = 1 : K$ **do**
  - **if** $k == 1$ **then**
    - **for** $z = 1 : Z$ **do**
      - Extract HOG features $\boldsymbol{\Phi}_1^z$.
      - Compute gating function using (8).
    - Select corresponding sub-domain z_* according to (9).
  - Extract HOG features $\boldsymbol{\Phi}_k^{z_*}$.
  - Update shape in accordance with (10).

Algorithm 2: SMUF

procedure `Training()`
- **Input :** Training images $[\boldsymbol{I}_1, \cdots, \boldsymbol{I}_N]$ with annotated landmarks $\boldsymbol{X}_* = [\boldsymbol{x}_*^1, \cdots, \boldsymbol{x}_*^N]$ and initial locations $\boldsymbol{X}_1 = [\boldsymbol{x}_1^1, \cdots, \boldsymbol{x}_1^N]$.
- **Output:** Descent maps $\boldsymbol{R}_k^z$, feature projections $\boldsymbol{W}_k^z$.
- **for** $z = 1 : Z$ **do**
  - Select z-th training subset.
  - **for** $k = 1 : K$ **do**
    - Initialize $\boldsymbol{W}_k^z$ to a random orthogonal matrix.
    - **for** $u = 1 : U$ **do**
      - Extract features $\tilde{\boldsymbol{\Phi}}_k^z$ according to (15).
      - Update $\boldsymbol{R}_k^z$ with fixed $\boldsymbol{W}_k^z$ using (16).
      - Update $\boldsymbol{W}_k^z$ with fixed $\boldsymbol{R}_k^z$ using (18).
    - Update shape $\boldsymbol{X}_k^z$ according to (19).

procedure `Evaluation()`
- **Input :** Test image $\boldsymbol{I}$, initial landmark locations computed by (7), descent maps $\boldsymbol{R}_k^z$, feature projections $\boldsymbol{W}_k^z$.
- **Output:** Final landmark locations.
- **for** $k = 1 : K$ **do**
  - **if** $k == 1$ **then**
    - **for** $z = 1 : Z$ **do**
      - Extract features $\tilde{\boldsymbol{\Phi}}_1^z$ according to (15).
      - Compute gating function value using (8).
    - Select corresponding sub-domain z_* according to (9).
  - Extract features $\hat{\boldsymbol{\Phi}}_k^{z_*}$ using (15).
  - Update landmark locations in accordance with (10).

4 Experiments

In this section, we evaluate the proposed GRID and SMUF landmarking approaches and compare them to the state-of-the-art. We report landmarking performance in accordance with the standard methodology used in this area [32] for all experiments. Specifically, we use the localization error, i.e., the Euclidean distance in mm between the location of the detected landmark and the manually annotated ground truth landmark, for performance reporting. Additionally, we also compute the mean localization error over all landmarks of each test face for some of the experiments.

4.1 Experimental datasets

We conduct experiments with three popular datasets of 3D face images: the FRGCv2 dataset, the Bosphorus 3D face dataset, and the UND dataset. We chose these datasets

because they are among the most frequently used 3D face datasets and because they contain challenging 3D images with a high degree of variability in face orientations and are, therefore, well suited for assessing the robustness to such variations. The main characteristics of the datasets are summarized in Table 2.

The FRGCv2 dataset contains 4007 3D face images of 466 individuals. Images of the dataset were acquired with a laser-based Konica Minolta Vivid 910 scanner. Subjects exhibit minor pose variations and various facial expressions. We utilize the ground truth landmarks (8 landmarks per face) from [25], which were manually annotated on a subset of 975 images from 149 subjects.

The Bosphorus dataset consists of 4666 face samples from 105 subjects. Each sample includes a 2D color image, a 3D point cloud, and 24 manually annotated landmarks (in our experiments, we exclude ear dimple landmarks and use the remaining 22 landmarks). Next to expression and occlusion variations, images in the dataset also exhibit large variations in pose. Images from the dataset were captured using a structured-light-based Inspeck Mega Capturor II Digitizer.

The UND dataset contains 1680 semi-profile and profile 3D face images of 537 subjects. For our experiments, we use a subset of 236 images with yaw rotations of $\pm 45^\circ$ and 174 images with yaw rotations of $\pm 60^\circ$ along with the manual annotations (8 landmarks for frontal faces and 5 for non-frontal faces) also provided by [25]. Images from this dataset were captured by the same acquisition device as used with FRGCv2.

4.2 Performance evaluation on the Bosphorus dataset

In the first series of experiments, we evaluate the performance of GRID and SMUF on the Bosphorus dataset which is particularly suitable to assess the robustness to large pose variations. We perform experiments using a twofold cross validation setup using half of the images for training and the other half for testing. To increase the size of training data, we extend the training set by horizontally flipping each of the available training images. We form test sets with respect to the yaw rotation angle or the presence of expressions/occlusions. We implement both methods with $K = 7$ cascade stages and use this setup also for all the following assessments.

The results of this series of experiments are summarized in Table 3. For both GRID and SMUF, we train three landmarking variants, each with a different number of DMs. The first column in Table 3 marked 1 DM corresponds to the variant that uses only one DM that was trained on images with 22 annotated landmarks (these are generally images of near-frontal faces, since large rotations lead to self-occlusions and fewer annotations). The second column represents the GRID and SMUF variants with 3 DMs: one DM is computed in the same way as in the variants in column 1, while the second and the third DMs are computed using images with the head rotations up to 45° to the left and right, respectively. The variants in the third column correspond to the setup in Fig. 3 and contain an additional two DMs corresponding to head rotations in the ranges of $[45^\circ, 90^\circ]$ and $[-45^\circ, -90^\circ]$. The DMs of the near-frontal images are trained using 22 landmarks per face image, while the DMs of rotated images are trained using 14 landmarks per face as some of the landmarks in these images are typically self-occluded.

Fig. 7 Illustration of the face detection procedure used on the FRGCv2 and UND datasets. The procedure uses a simple k-means clustering approach (with $k = 3$) and selects the cluster with the lowest mean depth as the face region. The figure shows Input image (left), color coded clusters (middle), and cropped and smoothed face image (right)

As expected, we can observe that the robustness to face rotations is significantly increased when more DMs are utilized. With the GRID and SMUF variants with 5 DMs, we achieve reliable landmark localization even on profile face images with yaw rotations up to $\pm\ 90^\circ$. It can also be seen from the last two rows in Table 3 that the same localization errors are obtained for all three variants when evaluated on the frontal face images with expression and occlusion variations. This indicates that the expressions and occlusions do not affect the DM selection process since in all cases the frontal DM is correctly chosen by the gating function.

When comparing the performance of SMUF and GRID, we can see that, in general, GRID ensures slightly better localization results than SMUF for all implemented variants. However, while there is an evident trend toward lower average localization errors for GRID, it is clear from Table 3 that the performance differences are statistically not significant. Thus, we can conclude that for the Bosphorus dataset, both techniques perform more or less equal.

4.3 Evaluation on the FRGCv2 and UND datasets

In the second series of experiments, we evaluate GRID and SMUF on the joint FRGCv2 and UND datasets. Contrary

to the Bosphorus dataset, where images contain solely the head regions and, therefore, using a face detector is not required; images from the FRGCv2 and UND datasets may also contain parts of the upper body, and thus, face detection is needed to initialize the landmark locations. In this series of experiments, we, hence, employ a simple face detector that relies on k-means clustering similar to the one presented in [32]. Setting the number of clusters to $k = 3$ and including several heuristic conditions, this detector divides a 3D image into three regions that most likely correspond to the background, body, and head/face regions. The face region candidate is then selected as the cluster with the lowest mean depth value (yellow region in Fig. 7). By doing so, a few other minor parts of the image may be selected besides the face region that is later reliably discarded by retaining only the largest connected area (right image in Fig. 7).

The face detector introduces additional variability into the facial regions, since the detected face may still include smaller parts of the upper body, neck, and hair. For that reason, we also report face mis-detection rates and selection rates for this experiment. Face mis-detection rate is defined as the percentage of images with the discrepancy between the location of face detection box and the locations of ground truth landmarks. The selection rate is defined as the percentage of images where the correct DM has been selected by the gating function, where we define a DM as incorrect if the DM has been trained on right profile face images while the corresponding test image is facing left or vice versa. The localization errors are then computed

Table 4 Mean localization errors (and corresponding standard deviations) on the FRGCv2 and UND datasets

Dataset	Detection rate	Selection rate	Localization error		
			GRID	SMUF	Perakis [25]
DB00F	99.6	100.0	3.2 ± 1.7	3.4 ± 2.2	5.0 ± 1.9
DB00F-neut.	99.6	100.0	3.0 ± 1.6	3.2 ± 1.8	4.5 ± 1.5
DB00F-mild	99.7	100.0	3.3 ± 1.6	3.5 ± 1.9	5.0 ± 1.5
DB00F-extr.	99.4	100.0	4.0 ± 2.1	4.2 ± 3.0	6.3 ± 2.6
DB00F45RL	98.3	99.1	3.3 ± 2.0	4.0 ± 2.9	5.0 ± 1.9
DB45R	99.1	97.4	3.2 ± 2.1	3.7 ± 2.6	5.0 ± 1.9
DB45L	98.3	99.1	3.8 ± 2.1	4.1 ± 2.7	4.8 ± 1.9
DB60R	98.9	96.6	3.7 ± 1.6	4.2 ± 2.5	5.0 ± 1.8
DB60L	98.9	95.4	4.2 ± 1.9	5.1 ± 4.0	5.3 ± 2.5

Results are reported for specific subsets in accordance with the protocol from [25]. The subsets feature images with different levels of expression variations (DB00F), and yaw rotations of 45° and 60° to the right (R), the left (L), or in both directions (RL). The results show that both GRID and SMUF offer favorable performance when compared to the method of Perakis et al. [25]

Table 5 Localization errors of GRID and SMUF in comparison to the state-of-the-art on non-frontal facial datasets for 10 common facial landmarks. GRID and SMUF significantly outperform competing methods on all experimental datasets. When comparing the learned binary features from SMUF to hand-crafted LBP features, we observe better performance for the learned binary features. GRID and SMUF again perform similarly for all landmarks

Landmarks	Method and database								
	Sukno et al. [36]	Creusot et al. [5]	Passalis et al. [24]	Perakis et al. [25]	LBP	GRID		SMUF	
	Bosphorus	Bosphorus	FRGC +UND	FRGC +UND	Bosphorus	FRGC +UND	Bosphorus	FRGC +UND	Bosphorus
Inner eye c.	2.9 ± 2.0	4.1 ± 2.6	6.4 ± 3.0	4.8 ± 2.7	2.8 ± 3.4	2.7 ± 1.5	2.0 ± 1.1	3.2 ± 2.1	2.4 ± 1.9
Outer eye c.	5.1 ± 3.7	6.3 ± 4.0	6.6 ± 3.7	5.7 ± 3.9	3.5 ± 3.8	2.8 ± 1.9	2.5 ± 1.4	3.4 ± 2.3	2.9 ± 2.8
Nose tip	2.3 ± 1.8	4.3 ± 2.6	4.6 ± 3.0	4.4 ± 2.7	4.9 ± 6.3	3.5 ± 2.8	3.7 ± 2.8	4.8 ± 4.4	3.9 ± 3.5
Nose c.	3.0 ± 1.9	4.2 ± 2.4	n/a	n/a	4.0 ± 4.7	n/a	3.0 ± 2.0	n/a	3.5 ± 2.9
Mouth c.	6.1 ± 5.1	8.0 ± 5.4	5.8 ± 3.9	5.0 ± 2.9	3.4 ± 4.3	2.9 ± 2.0	2.5 ± 1.6	3.4 ± 2.5	2.7 ± 2.1
Chin tip	7.6 ± 6.7	15.4 ± 10.5	6.6 ± 3.5	4.8 ± 3.5	5.5 ± 4.8	4.8 ± 2.1	4.8 ± 3.4	6.2 ± 3.7	5.1 ± 3.5

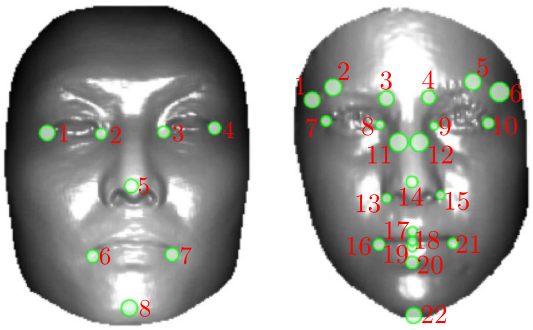
exclusively on the images with correct face detections and DM selections. This type of reporting is adapted from [25], which we use for baseline comparison in this experiment.

The results of the experiments are presented in Table 4. For details on the dataset abbreviations in the first column, please refer to [25], since the experimental setup and the landmark annotations are adopted from there. In short, however, DB00F denotes an image subset with varying facial expressions, which is further partitioned into neutral (neut.), mild, and extreme (extr.) facial expressions. The remaining image subsets contain faces with 45° or 60° yaw rotations either to the right (R), the left (L), or both (RL). As the GRID and SMUF methods require non-frontal images to train some of the DMs, our experimental setup differs from [25] only in the construction of training set where we also employ images from the Bosphorus dataset.

Detection and selection rates are consistently above 95% for all subsets as it can be observed from the first two columns in Table 4. Localization errors of our two landmarking approaches are compared to Perakis et al. [25] (last column) which, to the best of our knowledge, achieves the highest performance in the literature on these datasets. The results show the robustness of our methods to both expression variations and to rotations. The mean localization error is under 6 mm on all tested subsets for both GRID and SMUF. Since the training data is taken from the Bosphorus dataset (acquired with a different 3D camera), the results also imply good generalization to data from different sensors. All experiments from this section were performed using 5 DMs, as we observed earlier in Sect. 4.2 that this setting is the most robust to rotation variations.

4.4 Comparison to the state-of-the-art

In the next series of experiments, we compare the performance of GRID and SMUF to the performance of state-of-the-art landmarking methods from the literature. Specifically, we select the method of Sukno et al. [36], the technique of Creusot et al. [5], and the landmarking approaches of Passalis et al. [24] and Perakis et al. [25] for our comparison. To the best of our knowledge, these landmarking methods are the only ones that were evaluated on both frontal as well as rotated 3D facial images. Following the experimental protocols of other authors, we used the DB00F45RL subset when performing experiments on the FRGCv2+UND database and used the entire database for experimentation on the Bosphorus dataset. Additionally, we also implement our gated landmarking approach with hand-crafted binary features, that is, with LBPs (uniform, neighborhood size of 8 and radius of 1) to capitalize on the usefulness of learning binary features instead of using off-the-shelf binary feature extractors.



(a) GRID, FRGC/UND **(b)** SMUF, Bosphorus

Fig. 8 Mean localization errors achieved by GRID and SMUF for individual landmarks of the FRGCv2+UND and Bosphorus datasets. The size of the circles corresponds to the localization errors. The numbering of the landmarks as shown here is also used in Figs. 9 and 10. The lowest errors are achieved on distinct landmarks with corner-like properties, e.g., the mouth corners

The results of the comparison are shown in Table 5. We observe that on the Bosphorus dataset both GRID and SMUF significantly outperform the competing methods from the literature and achieve not only lower average localization errors, but also significantly smaller standard deviations on these errors. The only exception here are the nose tip and corners, where the method of Sukno performs similarly or slightly better. We also see similar results for the FRGC-UND dataset, where both GRID and SMUF achieve a considerable reduction in the localization errors for all considered landmarks compared to the state-of-the-art.

When comparing the learned binary features used in SMUF to the hand-crafted LBP features, we also see an obvious performance improvement in the learned features, re-enforcing our assumption that learning binary features is beneficial for face alignment. The comparison between GRID and SMUF shows a similar picture as in the previous series of experiments, where GRID was found to perform slightly better than SMUF, but not significantly so.

4.5 Landmark analysis

In this section, we evaluate how the overall localization performance varies across the individual landmarks for both GRID and SMUF. Figure 8 illustrates the mean localization errors achieved for the individual landmarks—the size of the circles is proportional to the errors. It can be observed that the landmarks corresponding to the nose tip and eye and mouth corners exhibit low localization errors. This is expected as these landmarks correspond to well-pronounced facial parts with distinctive "corner-like" shapes. Contrarily, landmarks relating to nose saddle points, the chin tip, and eyebrow points correspond to

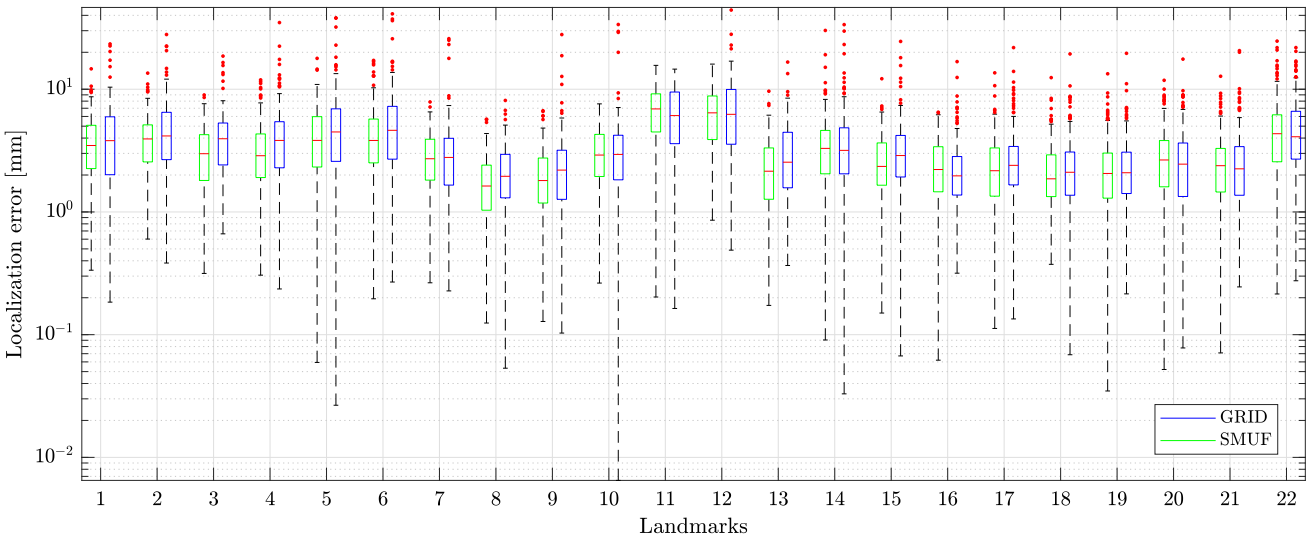


Fig. 9 Localization errors in the form of box and whiskers plots for the Bosphorus dataset achieved with the GRID and SMUF landmarking techniques. The results show that the lowest localization errors with both methods are achieved on distinct landmarks with corner-like characteristics, such as the eye or mouth corners or the nose tip. The figure is best viewed in color

indistinctive "edge-like" local shapes and, therefore, result in high localization errors. This observation is also supported by the box plots in Figs. 9 and 10 that show localization errors of individual landmarks on the Bosphorus and the FRGCv2+UND databases. The presented behavior is consistent for both evaluated methods. Note that the number of landmarks in Figs. 9 and 10 depends on the employed dataset and not on the chosen landmarking method.

To further analyze the landmarking performance of GRID and SMUF and their generalization ability, we also performed a cross-database experiment, where we built the test set with images from the FRGCv2+UND dataset, while the training set was generated using images from the Bosphorus dataset. Results relating to the cross-database experiment are illustrated by the green box plots in Fig. 10. When compared to the experiment where both training and test sets are from the FRGCv2+UND dataset, we observe a slight increase in localization errors for most of the

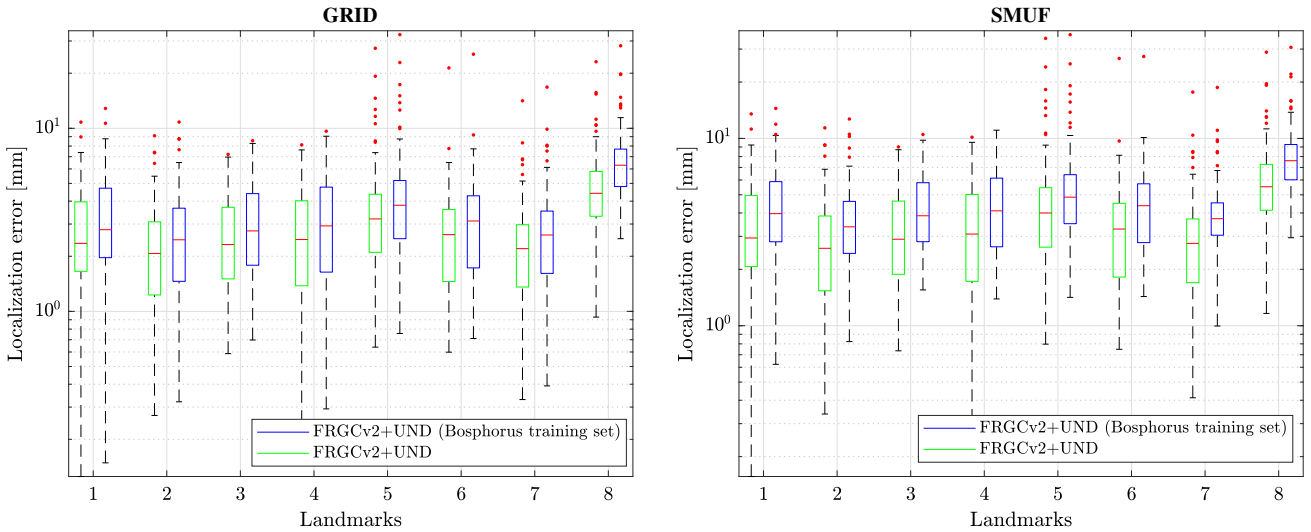


Fig. 10 Localization errors in the form of box and whiskers plots achieved on the FRGCv2+UND dataset with the GRID and SMUF landmarking techniques. Results are also presented for a cross-dataset experiment, where the landmarkers are trained on the Bosphorus datasets and are evaluated on the FRGV2+UND dataset. Lower errors are again achieved on distinct landmarks. The methods generalize well to novel datasets with the median errors for the cross-dataset experiment being only slightly larger than for the within-dataset experiments for the majority of landmarks. The figure is best viewed in color

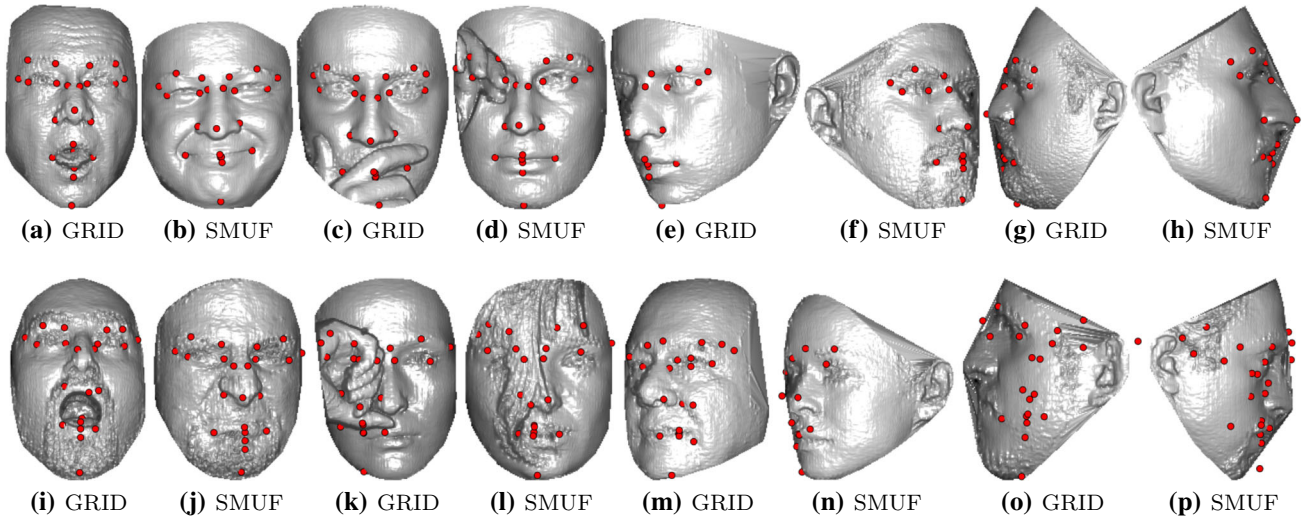


Fig. 11 Exemplar landmark detection results on the Bosphorus database: the first row depicts randomly chosen test samples (**a**–**h**), while the second row includes samples with high localization errors due to expressions (**i**, **j**), occlusions (**k**, **l**), head rotations (**m**, **n**), and incorrectly selected descent maps (**o**, **p**)

landmarks, except the chin tip where the difference between mean errors is larger. We presume that the high mean error for the chin tip comes from the increased appearance variability caused by the face detection procedure needed for the FRGCv2+UND data. (see Fig. 12g). Since such variability is not present in the training set from the Bosphorus dataset, the landmarking procedures cannot learn to accommodate for the inaccuracies of the face detector. In terms of comparison of GRID and SMUF, we see no significant difference in their performance in these experiments.

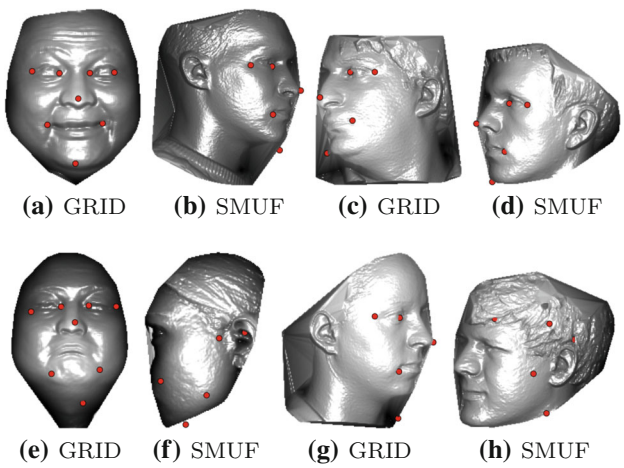


Fig. 12 Exemplar landmark detection results on the UND and FRGCv2 datasets: the first row depicts test images with typical localization performance, while images from the second row are selected among the worst samples measured by the localization error

4.6 Qualitative evaluation

In this section, we qualitatively assess the landmarking performance of the proposed landmarking methods. Figures 11 and 12 show exemplar face images from the Bosphorus and UND datasets with localized landmarks marked by red dots. The top rows of both figures contain samples with typical localization performance, where we can see that the method possesses stable performance in the presence of different types of variability, such as expressions, partial occlusions, and head rotations. However, there are some cases where landmarks are poorly localized. Such samples with large localization errors are exposed in the second rows of Figs. 11 and 12, e.g., large occlusions of face areas (Fig. 11k, l) can cause increased localization errors of visible landmarks. Some of the localization errors originate from poor face detection and cropping, where an image can contain also non-head regions (Fig. 12e, g) or parts of the face area are cropped out (Fig. 12f). Misselected descent maps can also be the cause of landmark localization errors (Fig. 11o, p, h).

4.7 Computational cost

In the last series of the experiments, we evaluate the time needed by the GRID and SMUF methods to localize landmarks on a single test image on average. We compute the average processing time over 100 randomly selected test images from the Bosphorus dataset. The size of the input images is 250×200 and we compute the locations of all 22 landmarks during the benchmark. A PC with the following specifications is used for the assessment: Intel Xeon CPU 2.67 GHz with 12 GB RAM. Both landmarking techniques are implemented using Matlab, run entirely on

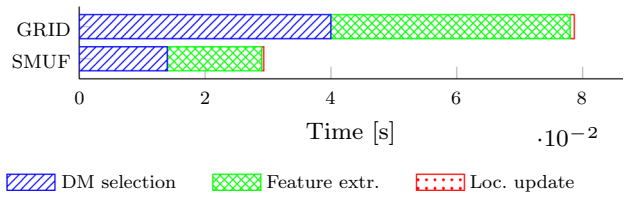


Fig. 13 Average running time of the GRID and SMUF methods to localize landmarks on one face image (computed over 100 randomly selected test images). For the benchmarking, images from the Bosphorus dataset were used and 22 landmarks were predicted. The results show that SMUF is around 3× faster than GRID, but ensures only slightly higher localization errors

CPU and could be further sped up if implemented with a compiled language such as C/C++. We start from detected and localized face regions and measure the time for feature extraction, DM selection, and location updates, which take less than 3×10^{-2} s for SMUF (see Fig. 13) and little less than 8×10^{-2} s for GRID. When compared to hand-crafted features, the learned binary features can be extracted almost 3 times faster than HOG features and 15 times faster than LBP features.

5 Conclusion and future work

We have presented two approaches to facial landmark localization from 3D face images, GRID, and SMUF, that are robust to rotations, facial expressions, and partially also to occlusions. We proposed a gating mechanism that allowed us to incorporate multiple pose-specific landmarking models (based on HOG features) into the alignment procedure and also developed a simultaneous descent map and binary feature learning algorithm around the proposed gating mechanism. To assess performance, we evaluated the developed landmarking techniques on three challenging datasets, containing 3D face images with large head rotations. Our results showed that the proposed solutions exhibit high robustness to different types of appearance variations and display competitive performance when compared to the state-of-the-art. Both of the proposed approaches need only a fraction of second to compute the landmarks on a given face image and could run in real time in combination with a suitable 3D sensor. Both methods exhibit a comparable performance, with a slight, although not statistically significant, advantage of GRID over SMUF. Therefore, in the case when fast processing times are required, it is preferable to use SMUF.

As a part of our future work, we plan to combine the proposed landmarking methods with face frontalization (or pose correction) procedures and incorporate all developed methods into pose-invariant 3D face recognition systems.

Acknowledgements This research was supported in parts by the ARRS (Slovenian Research Agency) Research Program P2-0250 (B) Metrology and Biometric Systems, the ARRS Research Program P2-0214 (A) Computer Vision, and the RS-MIZŠ and EU-ESRR funded GOSTOP.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alyüz N, Gökberk B, Akarun L (2010) Regional registration for expression resistant 3-D face recognition. IEEE Trans Inf Forensics Secur 5(3):425–440
2. Camgöz NC, Štruc V, Gökberk B, Akarun L, Kindiroğlu AA (2015) Facial landmark localization in depth images using supervised ridge descent. In: IEEE international conference on computer vision workshop (ICCVW), pp 378–383
3. Cao Y, Lu BL (2015) Neural information processing. In: Proceedings of 22nd international conference intensity-depth face alignment using cascade shape regression, ICONIP 2015, November 9–12, Part IV, chap. Springer, Cham, pp 224–231
4. Cootes TF, Walker K, Taylor CJ (2000) View-based active appearance models. In: Proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE. pp 227–232
5. Creusot C, Pears N, Austin J (2013) A machine-learning approach to keypoint detection and landmarking on 3D meshes. Int J Comput Vision 102(1):146–179
6. Faltemier TC, Bowyer KW, Flynn PJ (2008) Rotated Profile Signatures for robust 3D feature detection. In: 8th IEEE international conference on automatic face gesture recognition, 2008. FG '08, pp 1–7
7. Fanelli G, Dantone M, Gall J, Fossati A, Gool L (2012) Random forests for real time 3D face analysis. Int J Comput Vis 101(3):437–458
8. Fanelli G, Dantone M, Gool LV (2013) Real time 3D face alignment with Random Forests-based active appearance models. In: 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pp 1–8
9. Feng ZH, Hu G, Kittler J, Christmas W, Wu XJ (2015) Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. IEEE Trans Image Process 24(11):3425–3440
10. Gong Y, Lazebnik S, Gordo A, Perronnin F (2013) Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans Pattern Anal Mach Intell 35(12):2916–2929
11. Gupta S, Markey MK, Bovik AC (2010) Anthropometric 3D face recognition. Int J Comput Vision 90(3):331–349
12. Johnston B, de Chazal P (2018) A review of image-based automatic facial landmark identification techniques. EURASIP J Image Video Process 1:86. https://doi.org/10.1186/s13640-018-0324-4

13. Kendrick C, Tan K, Walker K, Yap MH (2018) Towards real-time facial landmark detection in depth data using auxiliary information. Symmetry 10(6):230. https://doi.org/10.3390/sym10060230
14. Križaj J, Emeršič Ž, Dobrišek S, Peer P, Štruc V (2018) Localization of facial landmarks in depth images using gated multiple ridge descent. In: 2018 IEEE international work conference on bioinspired intelligence (IWOBI), IEEE. pp 1–8
15. Li SZ, Zhang HJ, Cheng QS et al (2002) Multi-view face alignment using direct appearance models. In: Proceedings of fifth IEEE international conference on automatic face gesture recognition, IEEE. pp. 324–329
16. Liu F, Zhao Q, Liu X, Zeng D (2018) Joint face alignment and 3D face reconstruction with application to face recognition. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2018.2885995
17. Liu J, Zhang L, Chen X, Niu J (2017) Facial landmark automatic identification from three dimensional (3D) data by using Hidden Markov Model (HMM). Int J Ind Ergon 57:10–22
18. Liu R, Hu R, Yu H (2014) 3D face registration by depth-based template matching and active appearance model. In: Sixth international conference on wireless communications and signal processing (WCSP), pp 1–6
19. Lu J, Liong VE, Zhou J (2015) Simultaneous local binary feature learning and encoding for face recognition. In: Proceedings of the IEEE international conference on computer vision, pp 3721–3729
20. Lu J, Liong VE, Zhou J (2018) Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. IEEE Trans Pattern Anal Mach Intell 40(8):1979–1993
21. Masi I, Wu Y, Hassner T, Natarajan P (2018) Deep face recognition: a survey. In: 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), pp 471–478 . https://doi.org/10.1109/SIBGRAPI.2018.00067
22. Mian A, Bennamoun M, Owens R (2007) An efficient multimodal 2D–3D hybrid approach to automatic face recognition. IEEE Trans Pattern Anal Mach Intell 29(11):1927–1943
23. Park J, Park J (2018) A framework for virtual 3D manipulation of face in video. In: 2018 IEEE conference on virtual reality and 3D user interfaces (VR), pp 649–650 . https://doi.org/10.1109/VR.2018.8446445
24. Passalis G, Perakis P, Theoharis T, Kakadiaris IA (2011) Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Trans Pattern Anal Mach Intell 33(10):1938–1951
25. Perakis P, Passalis G, Theoharis T, Kakadiaris I (2013) 3D facial landmark detection under large yaw and expression variations. IEEE Trans Pattern Anal Mach Intell 35(7):1552–1564
26. Phillips PJ, Flynn PJ, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: IEEE computer society conference on computer vision and pattern recognition, CVPR, vol 1, pp 947–954
27. Pietikäinen M, Hadid A, Zhao G, Ahonen T (2011) Computer vision using local binary patterns, vol 40. Springer, Berlin
28. Rai MCE, Tortorici C, Al-Muhairi H, Werghi N, Linguraru M (2016) Facial landmarks detection using 3D constrained local model on mesh manifold. In: 2016 IEEE 59th International Midwest symposium on circuits and systems (MWSCAS), pp 1–4
29. Ramanan D, Zhu X (2012) Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition (CVPR), Citeseer, pp 2879–2886
30. Romero M, Pears N (2009) Landmark localisation in 3D face data. In: Sixth IEEE international conference on advanced video and signal based surveillance, 2009. AVSS '09. pp 73–78
31. Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B, Akarun L (2008) Biometrics and identity management, chap. In: Bosphorus database for 3D face analysis, Springer, Berlin, pp 47–56
32. Segundo MP, Queirolo C, Bellon ORP, Silva L (2007) Automatic 3D facial segmentation and landmark detection. In: 14th international conference on image analysis and processing, ICIAP 2007, pp 431–436
33. Segundo MP, Silva L, Bellon ORP, Queirolo CC (2010) Automatic face segmentation and facial landmark detection in range images. IEEE Trans Syst Man Cybern Part B (Cybern) 40(5):1319–1330
34. Smolyanskiy N, Huitema C, Liang L, Anderson SE (2014) Real-time 3D face tracking based on active appearance model constrained by depth data. Image Vis Comput 32(11):860–869
35. Song M, Tao D, Sun S, Chen C, Maybank SJ (2014) Robust 3D face landmark localization based on local coordinate coding. IEEE Trans Image Process 23(12):5108–5122
36. Sukno FM, Waddington JL, Whelan PF (2015) 3-D facial landmark localization with asymmetry patterns and shape regression from incomplete local features. IEEE Trans Cybern 45(9):1717–1730
37. Sánchez-Lozano E, Tzimiropoulos G, Martinez B, Torre FD, Valstar M (2018) A functional regression approach to facial landmark tracking. IEEE Trans Pattern Anal Mach Intell 40(9):2037–2050. https://doi.org/10.1109/TPAMI.2017.2745568
38. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cognit Neurosci 3(1):71–86
39. Wang J, Zhang J, Luo C, Chen F (2017) Joint head pose and facial landmark regression from depth images. Comput Vis Media 3(3):229–241
40. Wang K, Zhao X, Gao W, Zou J (2018) A coarse-to-fine approach for 3D facial landmarking by using deep feature fusion. Symmetry 10(8):308
41. Wu Y, Ji Q (2019) Facial landmark detection: a literature survey. Int J Comput Vis 127(2):115–142
42. Xia J, Cao L, Zhang G, Liao J (2019) Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2909327
43. Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 532–539
44. Xiong X, De la Torre F (2014) Supervised descent method for solving nonlinear least squares problems in computer vision. CoRR
45. Xiong X, De la Torre F (2015) Global supervised descent method. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2664–2673
46. Yan P, Bowyer K (2005) Empirical Evaluation of Advanced Ear Biometrics. In: IEEE computer society conference on computer vision and pattern recognition—workshops, 2005. CVPR workshops, p 41
47. Yu X, Huang J, Zhang S, Yan W, Metaxas DN (2013) Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: Proceedings of the IEEE international conference on computer vision, pp 1944–1951
48. Zhao X, Dellandrea E, Chen L, Kakadiaris IA (2011) Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional

statistical facial feature model. IEEE Trans Syst Man Cybern Part B (Cybern) 41(5):1417–1428
49. Zhao X, Zou J, Li H, Dellandréa E, Kakadiaris IA, Chen L (2016) Automatic 2.5-D facial landmarking and emotion annotation for social interaction assistance. IEEE Trans Cybern 46(9):2042–2055. https://doi.org/10.1109/TCYB.2015.2461131
50. Zhou Y, Zhang W, Tang X, Shum H (2005) A bayesian mixture model for multi-view face alignment. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE. vol 2, pp 741–746