Accurate landmarking of 3D facial data in the presence of facial expressions and occlusions using a 3D statistical facial feature model

Xi Zhao, *Student Member, IEEE,* Emmanuel Dellandréa, Liming Chen, *Member, IEEE* and Ioannis A. Kakadiaris, *Senior Member, IEEE*

Abstract—3D face landmarking aims at automatically localizing facial landmarks and has a wide range of applications (e.g., face recognition, face tracking and facial expression analysis). Existing methods assume neutral facial expressions and unoccluded faces. In this paper, we propose a general learning-based framework for reliable landmark localization on 3D facial data under challenging conditions (i.e., facial expressions and occlusions). Our approach relies on a statistical model, called 3D Statistical Facial feAture Model (SFAM), which learns both the global variations in configurational relationships between landmarks and the local variations of texture and geometry around each landmark. Based on this model, we further propose an occlusion classifier and a fitting algorithm. Results from experiments on three publicly available 3D face databases (FRGC, BU-3DFE, Bosphorus) demonstrate the effectiveness of our approach, in terms of landmarking accuracy and robustness, in the presence of expressions and occlusions.

Index Terms—3D face feature, landmarks, statistical face model, fitting, occlusion, facial expression.

1 Introduction

The recent emergence of three-dimensional facial data has provided an alternative to overcome the challenges in 2D face recognition, caused by pose changes and lighting variations [6]. Although 2.5D/3D face data acquisition is known to be insensitive to changes in lighting conditions, the data need to be pose normalized and correctly registered for further face analysis (e.g., 3D face matching [20], tracking [33], recognition [26] [28], and facial expression analysis [34]). As most of the existing registration techniques assume the availability of some 2.5D/3D face landmarks, a reliable localization of these facial feature points is essential.

1.1 Related work

Although there is no general consensus yet, we consider stable facial landmarks to be the fiducial points defined by anthropometry [9] that have consistent reproducibility even in adverse conditions such as facial expression or occlusion. Stable facial landmarks generally include the nose tip, the inner eye corners, the outer eye corners and the mouth corners. Such landmarks are not only characterized by their own properties, in terms of local texture and local shape, but are also characterized by their global structure resulting from the morphology of the face. Therefore, local feature information and the configurational relationships of landmarks are jointly important for accurate and robust face landmarking. This finding is coherent with human studies on face analysis suggesting that both local features and configurational relationships are important [44].

Despite the increasing amount of related literature, 3D face landmarking is still an open problem. Current face landmarking techniques lack both accuracy and robustness, particularly in the presence of lighting variations, head pose variations, scale changes, facial expressions, self-occlusions and occlusion by accessories (e.g., hair, moustache and eyeglasses) [1]. This paper proposes a data-driven general framework for precise 3D face landmarking, which is robust to changes in facial expressions and partial occlusions.

• *Xi Zhao and I.A. Kakadiaris are with the Computational Biomedicine Lab (CBL), Department of Computer Science, University of Houston, Houston, TX 77204-3010. Emmanuel Dellandréa and Liming Chen are with the Université de Lyon, CNRS, Ecole Centrale Lyon, LIRIS, UMR5205, F-69134, France.*
E-mail: zhaoxi@ieee.org, emmanuel.dellandrea@ec-lyon.fr, liming.chen@ec-lyon.fr, ioannisk@uh.edu.

Face landmarking on 2D facial texture images has been extensively studied [1] and several approaches have been proposed. These approaches can be classified into appearance-based [2], geometry-based [3] and structure-based approaches [4], [5]. Interesting approaches include 2D statistical models such as the popular Active Appearance Model [12] or the more recent Constrained Local Model [14], which perform statistical analysis both on the facial appearance and the 2D shape. However, since they are applied to 2D texture images, these approaches inherit the sensitivity to lighting and pose changes.

Research on 3D face landmarking is rather recent. Most of the existing methods embed *a priori* knowledge on landmarks on 3D face by computing the response to local 3D shape-related features (e.g., spin image [28], [42], [43], effective energy [10], Gabor filtering [7], [11], generalized Hough Transform [24], local gradients [19], HK curvature [22], shape index [20], [42], [43], curvedness index [21] and radial symmetry [29]). While these approaches enable a rather accurate detection of landmarks that are shape prominent (e.g., the nose tip or the inner corners of eyes), their localization accuracy drastically decreases for other less prominent landmarks.

As current 3D imaging systems can deliver registered range and texture images, a straightforward method to discriminate a landmark is to accumulate evidence from both face representations (i.e., face geometry and texture). Boehnen *et al.* [27] computed the eye and mouth maps based on both color and range information. Wang *et al.* [25] used a "point signature" representation to code a 3D face mesh as well as Gabor jets of landmarks from the 2D texture image. Gabor wavelet coefficients [1], [23] were used to model the local appearance in the texture map and local shape in a range map around each landmark. Lu *et al.* [32] proposed to compute and fuse the shape index response (range) and the cornerness response (texture) in local regions around seven feature points.

As the combinations of candidate landmarks resulting from shape and/or texture related descriptors are generally important, some studies also proposed to make use of the structure between landmarks. This is accomplished by using heuristics [21], a 3D geometry-based confidence [27], an extended elastic bunch graph [23], or a simple mean model constructed as the average 3D position of landmarks from a learning dataset [30]. However, there is no technique that best takes into account both the configurational relationships between landmarks and the local properties in terms of geometric shape/texture around each landmark.

Furthermore, only few of the aforementioned studies address the issue of face landmarking in the presence of facial expressions or occlusions. Nair *et al.* [21] used their 3D Point Distribution Model to locate five landmarks (the two outer eye points, the two inner eye points and the nose tip) under facial expressions with a locating accuracy ranging from 8.83 mm for the nose tip to 20.46 mm for the right outer eye point. However, all the five landmarks were located on stable face regions during facial expressions. Dibeklioglu *et al.* [19] studied 3D facial landmarking under expression, pose and occlusion variations. They built statistical models of local features around landmark locations using a mixture of factor analysis in order to determine landmark locations on a coarse level. Heuristics were then applied to locate the nose tip at a fine level. Using the configurational relationships and geometry features, Perakis *et al.* [42], [43] addressed landmarking on 3D facial data under multiple orientations, taking into account missing data due to self occlusion.

1.2 The proposed approach

In this paper, we propose a general learning-based framework for 3D face landmarking which combines both configurational relationships between the landmarks and their local properties in a principled way, through optimization of a global objective function. This data-driven based approach aims to overcome the shortcomings of the previous feature-based approaches that require the embedding of a discriminative prior knowledge for each landmark. Instead, it relies on a statistical model, called 3D Statistical Facial feAture Model (SFAM), which learns both the global variations in 3D face morphology and the local variations around each 3D face landmark in terms of texture and geometry. To train the model, we manually labeled the target landmarks for each aligned frontal 3D face. Preprocessing is first performed to enhance the quality of facial scans and then the scans are remeshed to normalize the face scale. The SFAM is then constructed by applying Principle Component Analysis (PCA) to the global 3D face landmark configurations, the local texture and the local shape around each landmark from the training facial data. PCA-based learning is popular for face recognition since human faces are similar and hence it is quite reasonable to assume that the properties

of facial features follow a Gaussian distribution, as demonstrated by previous studies (e.g., eigenfaces [45]). In our approach, only the salient variation modes (95% of the variation) for the three representations (morphology, texture and geometry) are retained. By varying the control parameters of SFAM, different 3D partial face instances that consist of local face regions with texture and shape (structured by their global 3D morphology) can be generated. In this paper, we have used a simple local range map and an intensity map to characterize the local shape and texture properties around each landmark. Alternatively, the SFAM may use all the aforementioned descriptors of local features around each landmark (e.g., mean and Gaussian curvature or shape index for local shape characterization, and Gabor jets or cornerness response for local texture description). An interesting property for the characterization of the local shape around a landmark is that the descriptor is sufficiently robust against shape deformation, which typically occurs in facial expressions. Popular geometric descriptors (e.g., shape index or HK curvatures) provide an accurate local shape description and are sensitive to geometric shape differences. However, when the normalized correlation is used as the similarity measure, local shape properties described by raw range maps are less discriminative with respect to identity and deformations. Similarly, the description of local texture should be tolerant to changes caused by lighting or expressions. A similar reasoning also applies to using the raw texture maps for texture characterization. The combination of raw texture maps and the similarity measure relieves, to some extent, the effect of lighting conditions and expressions on texture. Our experiments indicate that the use of a local raw range map and a local raw texture map around each landmark provides a good tradeoff between computational efficiency and robustness. Although a comprehensive study of the selection of robust local features is needed, it is beyond the scope of this paper.

Our learning-based framework can be considered as a natural extension of the morphable 3D face model [15] and the constrained local model (CLM) [14] as we propose to learn at the same time, the global variations of 3D face morphology and the local ones in terms of texture and shape around each landmark. Fitting the SFAM on a probe facial scan is accomplished by maximizing the *a posteriori* probability (MAP). The fitted morphology instance delivers the locations of targeted landmarks. Using 3D training faces with expressions, the SFAM has the ability to learn expression variations and generate instances with the learnt variations so as to increase the posteriori probability in fitting faces with expression. Furthermore, we propose to use a k-nearest neighbors (kNN) classifier to identify the partially occluded faces and the type of occlusion. A histogram of the similarity map between the local shapes of the target face and shape instances from SFAM is used as the input. This information about occlusions is also integrated into the objective function used in the fitting process to handle landmarking on partially occluded 3D facial scans.

The main contributions of this work are the following:

1) We build a statistical facial feature model that elegantly combines the global and local features extracted from three facial representations.
2) An occlusion detection and classification algorithm is proposed to detect occlusions and classify them into different types, thereby providing occlusion information to the fitting algorithm.
3) A fitting algorithm is proposed to locate landmarks through optimizing an objective function, implemented on local patch-based correlation meshes. In addition, the fitting algorithm incorporates occlusion knowledge and thus is able to locate landmarks on partial occluded faces.

The rest of the paper is organized as follows. In Section 2, our statistical model SFAM is introduced. In section 3, the objective function that combines the local shape and texture properties and the fitting algorithm are described. Section 4 addresses 3D face partial occlusion. Experimental results are discussed in Section 5, while Section 6 concludes the paper. Table 1 presents a summary of the different symbols used in this paper.

TABLE 1
Summary of Symbols

Symbols	Description
$\boldsymbol{s}$	3D facial landmark configuration vector
$\boldsymbol{g}$	Intensity vector
$\boldsymbol{z}$	Geometry vector
ψ	SFAM
$\boldsymbol{P}$	Learnt modes of variations
$\boldsymbol{b}$	SFAM parameters
T	Texture map of a 3D facial scan
R	Range map of a 3D facial scan
$\boldsymbol{m}$	Occlusion mask

2 Statistical Facial feAture Model

Three dimensional facial data acquired by the current 3D imaging systems are usually noisy and may

contain holes and spikes. Hence, we first preprocess all the 3D facial scans to remove noise. Head pose and scale variations are normalized by alignment and remeshing (Section 2.1). Then, we model the variations in 3D configurations of landmarks and their local variations in terms of texture and shape around each landmark (Section 2.2). New partial 3D face instances can be synthesized from the learnt model (Section 2.3).

2.1 Preprocessing the training facial data

To remove the noise (e.g., spikes and holes) and enhance the quality of 3D facial scans, we perform the following operations:

1) Median Cut: spikes are detected by checking the discontinuity of points and are removed by the application of a median filter.
2) Hole Filling: holes that are caused by the 3D scanner and the removed spikes are located on the range maps of facial scans by a morphological reconstruction [38] and filled by cubic interpolation. The open mouth is excluded from this preprocessing step by estimating the size of the hole corresponding to the open mouth region with an empirically set threshold.

Although faces are usually scanned from a frontal viewpoint, variations in head pose still exist and interfere with the learning of global variations in 3D facial morphology. Consequently, these variations may perturb the learning of local shape and texture variations. To compensate for head pose variations, the facial data are first translated close to the origin of the camera coordinate system. The Iterative Closest Point (ICP) algorithm [18] is then used to minimize the difference between the two point clouds of the new scan and the selected facial scan, which holds a frontal and straight pose. Since the head pose variations has been compensated after alignment, SFAM can be learnt with more accurate variations in local face texture and geometry.

To train the model, the targeted anthropometric landmarks have to be manually labeled for each aligned frontal 3D face. This is the major difference between the proposed approach and most of the existing 3D face landmarking algorithms. Instead of directly embedding *a priori* knowledge on landmarks into the landmarking algorithm, we propose a data-driven approach which, through statistical learning, encodes into a model discriminatory information of targeted landmarks, in terms of their global configurational relationships as well as the properties of local texture and shape around each landmark. For any given training dataset, the set of targeted landmarks can be easily changed according to the particular application. This general characteristic of the proposed approach is demonstrated in our experiments on three different public datasets: FRGC, BU-3DFE and Bosphorus datasets. Most landmarks out of 15 (as illusitrated in Figure 5) on the FRGC dataset were selected from the rigid part of the face as they were subsequently used for 3D face recognition. On the other hand, landmarks on the BU-3DFE and the Bosphorus datasets (as illustrated in Figure 6, 8) encompass anthropometric points from all facial regions as they are used for facial expression analysis.

To learn the local geometry and texture around each landmark, it is necessary to have the same number of points in a local region and have a dense correspondence among different faces. However, changes due to face scale and subject identity make this normalization difficult. Therefore, we use uniform grids to remesh local regions around landmarks. First, all the points are sampled from point clouds within a specified distance from each landmark. The number of sampled points, or the point density, in local regions varies from face to face due to face scale. Second, a uniform grid is associated with each landmark. As illustrated in Figure 1, each grid is centered at its corresponding landmark with a size of 15×15 (225 nodes on a grid) and a resolution of 1 mm (the intervals of grids on the X,Y dimensions are fixed to 1 mm). The z values of a node (and the associated intensity values) on a grid are interpolated from the range values of sampled points. Using this normalization, a fixed number of points can be obtained regardless of face scale and subject identity. Thus, the point-to-point correspondence among faces is established easily and efficiently.

2.2 Modeling the configurational relationships and local shape and texture features of the landmarks

Once a 3D facial scan is preprocessed, 3D coordinates of all the landmarks (3D morphology), are concatenated into a vector $\boldsymbol{s_i}$, which describes the configurational relationships among local regions.

$$\boldsymbol{s_k} = (x_1, y_1, z_1, x_2, y_2, z_2, ..., x_N, y_N, z_N)^T \quad (1)$$

where N is the number of landmarks (e.g., in this paper, N = 15 or 19).

We further generate two vectors $\boldsymbol{g_k}$ and $\boldsymbol{z_k}$ by concatenating intensity and range values on all the grids on a face (M is the number of interpolated points collected from all the local regions). The $\boldsymbol{z_k}$ vectors capture the variations of local geometric shapes around each landmark while the $\boldsymbol{g_k}$ vectors capture the local texture properties.

$$\boldsymbol{g_k} = (g_1^k, g_2^k, ..., g_M^k)^T, \ \boldsymbol{z_k} = (z_1^k, z_2^k, ..., z_M^k)^T \quad (2)$$

PCA is then applied to the three vector sets $\{\boldsymbol{s_k}\}$, $\{\boldsymbol{g_k}\}$ and $\{\boldsymbol{z_k}\}$, extracted from the training 3D facial data (k denotes the k^{th} training example). Thus, three linear models are built by retaining 95% of the variance in landmark configurations as well as local texture and shape around each landmark. The three models are represented as follows:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{P_s b_s} \quad (3)$$

$$\boldsymbol{g} = \bar{\boldsymbol{g}} + \boldsymbol{P_g b_g}, \ \boldsymbol{z} = \bar{\boldsymbol{z}} + \boldsymbol{P_z b_z} \quad (4)$$

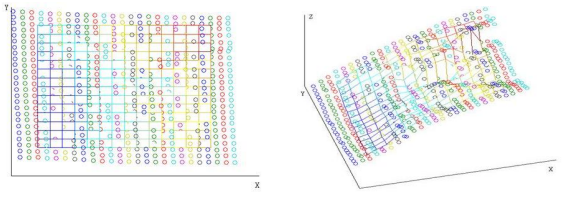
where $\bar{\boldsymbol{s}}$, $\bar{\boldsymbol{g}}$, $\bar{\boldsymbol{z}}$ are the mean landmark configuration, the mean intensity and the mean range value, respectively, while $\boldsymbol{P_s}$, $\boldsymbol{P_g}$, $\boldsymbol{P_z}$ are the three sets of modes of configuration, intensity and depth variation, respectively. The terms $\boldsymbol{b_s}$, $\boldsymbol{b_g}$, $\boldsymbol{b_z}$ are the corresponding sets of control parameters. All individual components in $\boldsymbol{b_s}$, $\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ are independent. We further assume that all the $\boldsymbol{b_q}$-parameters, where $\boldsymbol{b_q} \in (\boldsymbol{b_s}, \boldsymbol{b_g}, \boldsymbol{b_z})$ follow a Gaussian distribution with zero mean and standard deviation $\boldsymbol{\sigma_q}$.

2.3 Synthesizing instances from a new face

Given the parameters $\boldsymbol{b_s}$, a configuration instance can be generated using Eq. 3. Then, given a new facial scan, the set of scan points closest to the configuration instance are computed. Based on these points, the vectors $\boldsymbol{g^n}$ and $\boldsymbol{z^n}$ are obtained by applying the process described in the training phase (Eqs. (2). Then $\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ are estimated, as follows:

$$\boldsymbol{b_g} = \boldsymbol{P_g}^T(\boldsymbol{g^n} - \bar{\boldsymbol{g}}), \ \boldsymbol{b_z} = \boldsymbol{P_z}^T(\boldsymbol{z^n} - \bar{\boldsymbol{z}}), \quad (5)$$

$\boldsymbol{b_g}, \boldsymbol{b_z}$ are limited to the range $[-3\boldsymbol{\sigma}, 3\boldsymbol{\sigma}]$. Then using these constrained $\boldsymbol{b_g}$ and $\boldsymbol{b_z}$, we can generate texture and shape instances $\hat{\boldsymbol{g}}^{\boldsymbol{n}}$ and $\hat{\boldsymbol{z}}^{\boldsymbol{n}}$ by using Eqs. (4). The landmarks, along with their local texture and local shape instances, compose a partial face instance.



(a) (b)

Fig. 1. Scale normalization in a local region associated to the left corner of the left eye from (a) the frontal view and (b) side view. Circles denote sampled points from the 3D face model and the grid is composed of the interpolated points. Interpolation is also performed on the point intensity values

3 Localizing Landmarks

The SFAM-based landmark localization procedure consists of maximizing *a posteriori* probability (MAP) of landmark configuration given a 3D facial scan to be landmarked and leads to optimizing an objective function. In Section 3.1, we present the objective function to be optimized and in Section 3.2, we introduce the fitting algorithm for localizing landmarks. We then discuss our assumptions in Section 3.3.

3.1 Objective function and MAP

We first define the objective function $f(\boldsymbol{b_s}) = p(\boldsymbol{s}|T, R, \psi)$ as *the posteriori* probability of landmark configuration $\boldsymbol{s}$ to be maximized for a 3D facial scan represented by its texture map T and range map R, and the learnt statistical model SFAM ψ. Using the Bayes rule we obtain:

$$\begin{aligned} p(\boldsymbol{s}|T, R, \psi) &= p(T, R, \boldsymbol{s}, \psi)/p(T, R, \psi) \\ &\propto p(T, R|\boldsymbol{s}, \psi)p(\boldsymbol{s}|\psi) \\ &\propto p(T|\boldsymbol{s}, \psi)p(R|\boldsymbol{s}, \psi)p(\boldsymbol{s}|\psi) \end{aligned} \quad (6)$$

where $p(T|\boldsymbol{s}, \psi)$ and $p(R|\boldsymbol{s}, \psi)$ are the probabilities of having the facial texture T and the range R given a landmark configuration $\boldsymbol{s}$ and SFAM ψ, respectively. We assume that the random variables R and T from the different facial representations are independent within a local face region. The term $p(\boldsymbol{s}|\psi)$ denotes the probability of having a landmark configuration $\boldsymbol{s}$ given the SFAM ψ. Thus, the prior $p(\boldsymbol{s}|\psi)$ can be estimated using the assumption of Gaussian distribution on the corresponding control parameters b_j in the third term of Eq. 7.

The probabilities $p(T|\boldsymbol{s}, \psi)$ and $p(R|\boldsymbol{s}, \psi)$ can be estimated using the Gibbs-Boltzmann distribution as described in Eq. 7:

$$\begin{aligned} p(\boldsymbol{s}|T, R, \psi) &\propto \prod_{i=1}^{N} e^{-(\alpha\eta_i)} \prod_{i=1}^{N} e^{-(\beta\gamma_i)} \prod_{j=1}^{K} e^{\frac{-b_j^2}{\lambda_j}} \\ \log p(\boldsymbol{s}|T, R, \psi) &\propto \\ &\sum_{i=1}^{N} (-\alpha\eta_i) + \sum_{i=1}^{N} (-\beta\gamma_i) - \sum_{j=1}^{K} \frac{b_j^2}{\lambda_j} \end{aligned} \quad (7)$$

where N is the number of local regions, η_i and γ_i are the energy functions of the associated local region i in terms of texture and range properties, respectively, given the landmark configuration $\boldsymbol{s}$ and the SFAM ψ, and α and β are weight constants. The third term in Eq. 7 represents the Mahalanobis distance [13], where K is the number of retained landmark configuration modes and λ_j denotes the corresponding eigenvalue in the landmark configuration model. b_j denotes the control parameter that generates the landmark configuration $\boldsymbol{s}$ given the statistical model ψ. For the energy functions η_i and γ_i, high energies occur when the corresponding local texture T_i and range R_i do not match the texture and range instances which are generated by the SFAM ψ given the landmark configuration $\boldsymbol{s}$. In this work, instead of using the distances in these energy functions to express the degree of mismatch, we use a similarity measure, namely the normalized correlations defined in Eq. 9, and derive the following objective function $f(\boldsymbol{b_s})$ (thereby changing the polarity of the terms associated with η_i and γ_i):

$$f(\boldsymbol{b_s}) = \alpha \sum_{i=1}^{N} m_i F_{gi}(s_i) + \beta \sum_{i=1}^{N} m_i F_{zi}(s_i) - \sum_{j=1}^{k} \frac{b_j^2}{\lambda_j} \tag{8}$$

where F_{gi} and F_{zi} are explained in Eq. 9 and m_i is introduced to address partially occluded facial data. The term m_i is the probability of the region around the i^{th} landmark being unoccluded. The term s_i denotes the landmark location from the morphology model. Specifically,

$$F_{gi} = \left\langle \frac{\boldsymbol{g_i}}{\|\boldsymbol{g_i}\|}, \frac{\hat{\boldsymbol{g}}_{\boldsymbol{i}}}{\|\hat{\boldsymbol{g}}_{\boldsymbol{i}}\|} \right\rangle, F_{zi} = \left\langle \frac{\boldsymbol{z_i}}{\|\boldsymbol{z_i}\|}, \frac{\hat{\boldsymbol{z}}_{\boldsymbol{i}}}{\|\hat{\boldsymbol{z}}_{\boldsymbol{i}}\|} \right\rangle \tag{9}$$

where $\langle \cdot, \cdot \rangle$ is the inner product and $\|\cdot\|$ is the L_2 norm. The values of α and β are fixed and are computed as the ratios of $\sum_{i=1}^{N} F_{gi}$ and $\sum_{j=1}^{K} \frac{b_j^2}{\lambda_j}$, $\sum_{i=1}^{N} F_{zi}$ and $\sum_{j=1}^{K} \frac{b_j^2}{\lambda_j}$, respectively, during the off-line training.

In this work, we have used a simple occlusion classification algorithm which delivers a binary value for m_i: 0 if the local region is occluded and 1 if the region is not occluded.

3.2 Fitting algorithm

Landmarking a 3D facial scan consists of fitting the SFAM ψ while maximizing the objective function (Eq. 8). First, the 3D facial scan is preprocessed as described in the Section 2.1, including spike removal, hole filling and head pose normalization. The occlusion algorithm, introduced in Section 4, is then applied to identify the occluded local regions and then used to set the corresponding m_i coefficients to zero. Therefore, only the unoccluded local regions are considered in the fitting process. The algorithm works in a straightforward manner, and is described in Algorithm 1.

The optimization process in steps one and five of the algorithm is processed by the Nelder-Meade simplex algorithm [16]. Once convergence is reached, the instance $\boldsymbol{s}$ resulting from the optimized $\boldsymbol{b_s}$ indicates the location of landmarks. For partially occluded faces, occluded landmarks and their corresponding local meshes are excluded from the optimization process. In the case of incorrect occlusion classification, local non-face meshes lead the optimization to converge to an unpredictable point far from the desired minimum.

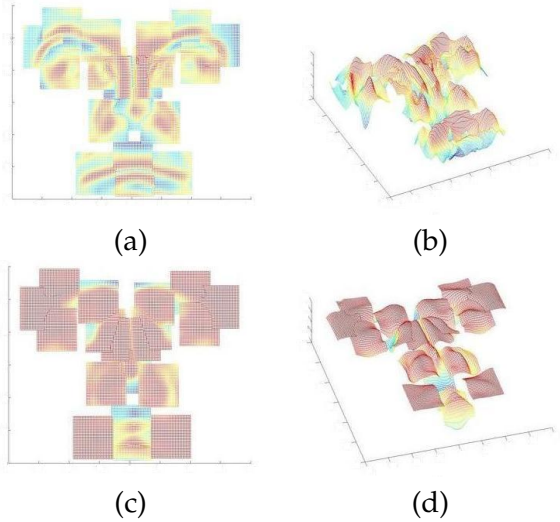


Fig. 2. Depiction of the correlation meshes from the frontal and side views. These meshes capture the similarity between instances and local facial regions in both texture and shape representations. The red color corresponds to large correlation values while blue corresponds to small correlation values. Large values on the correlation meshes correspond to large probabilities of finding landmarks on their locations. The meshes are in four dimensional space, where the first three dimensions are x, y, z and the last dimension represents correlation values. In these figures, we display the correlation values instead of z. (a,b) Two viewpoints of the same correlation mesh capturing the similarity of texture (intensity) instances from SFAM and local texture regions (intensity) on a given face. (c,d) Correlation mesh capturing the similarity of shape (range) instances from SFAM and the local face shapes (range).

3.3 Discussion

To deduce Eq. 7 we assumed that the probabilities $p(T|\boldsymbol{s},\psi)$ and $p(R|\boldsymbol{s},\psi)$ follow a Gibbs-Boltzmann distribution. This assumption is reasonable and motivated by the fact that the problem of 3D face landmarking is actually a Markov Random Field (MRF) which consists of assigning a label from a set of labels $\mathscr{L}$ to each vertex of a 3D facial scan. The set $\mathscr{L}$ encompasses all targeted landmarks (e.g., nose tip, eye corners) and a null value labeling any vertex which is not the location of a targeted landmark. Then, the theorem of the equivalence between MRFs and Gibbs distributions defined by Hammersley and Clifford [39], implies that the probabilities $p(T|\boldsymbol{s},\psi)$ and $p(R|\boldsymbol{s},\psi)$, follow a Gibbs-Boltzmann distribution [40].

We also used the Nelder-Meade simplex algorithm [16], which is one of the best known algorithms for multidimensional unconstrained optimization without derivatives. This method does not require any derivative information and is widely used to solve parameter estimation and statistical problems of similar nature [41].

4 Occlusion detection and classification

Facial data analysis in the presence of partial occlusions (caused by a variety of factors such as hair, glasses, mustaches, scarf) is a difficult problem. In 3D facial landmarking, only occlusions which may occur in local regions around landmarks are of interest. Thus, in this work we adopt an approach to classify the occlusion type and provide a set of binary values to local regions: either occluded or not occluded. Alternatively, we may compute a probability associated with a local region being occluded or a measure indicating roughly the extent to which a local region is occluded.

Algorithm 1 SFAM Fitting

Input: A 3D scan and a trained SFAM

1. Optimize the morphology parameters $\boldsymbol{b_s}$ to minimize the distance between corresponding morphology instances and their closest points on the input facial data, and obtain a set of points $\mathscr{S}$
2. Synthesize texture and shape instances $\hat{G},\hat{Z}$ as described in Section 2.3 using $\mathscr{S}$
3. Normalize local regions around points $\mathscr{S}$ within a neighborhood large enough to cover the potential landmark locations as in Section 2.1, creating a set of local mesh $\mathscr{G},\mathscr{Z}$
4. Compute correlation meshes on both texture and geometry representations (Figure 2) by correlating $\hat{G},\hat{Z}$ with G,Z, respectively, which are different parts of $\mathscr{G},\mathscr{Z}$ sampled by a sliding window (size of 15×15) on local regions (Eq. 9)
5. Optimize the morphology parameters $\boldsymbol{b_s}$ to reach the maximum of the sum of values on the two correlation meshes while minimizing the Mahalanobis distance associated with the landmarks configuration defined by the control parameters $\boldsymbol{b_s}$.

Output: Optimized morphology parameters $\boldsymbol{b_s}$

To perform occlusion detection, features from the range map are extracted as the presence of occlusion definitively changes local shape. Therefore, given a new facial scan, its closest points to the mean landmark configuration $\bar{\boldsymbol{s}}$ (Eq. 3) are first computed. Then, grids (50×50) are used to remesh local regions around these points for range values (Section 2.1). The size of local regions is chosen to be large enough to account for variations due to scale and subject changes as well as to cover the local regions near landmarks for occlusion detection.

For each local region i, processing is performed in a sliding window manner (the size of the sliding window is the same as the size of the local regions considered in SFAM). At each step, we compute a local depth map $\boldsymbol{Z_\alpha}$ and its local shape instance $\boldsymbol{Z_\beta}$ to further obtain a similarity L_S as follows:

$$\boldsymbol{b_\alpha} = \boldsymbol{P}_{\boldsymbol{z,i}}^{T}(\boldsymbol{Z_\alpha} - \bar{\boldsymbol{z}}_{\boldsymbol{i}}),\ \boldsymbol{Z_\beta} = \bar{\boldsymbol{z}}_{\boldsymbol{i}} + \boldsymbol{P_{z,i}}\boldsymbol{b_\beta} \tag{10}$$

$$L_S = \left\langle \frac{\boldsymbol{Z_\alpha}}{\|\boldsymbol{Z_\alpha}\|}, \frac{\boldsymbol{Z_\beta}}{\|\boldsymbol{Z_\beta}\|} \right\rangle \tag{11}$$

where $\boldsymbol{P_{z,i}}$ is the submatrix composed of the rows in $\boldsymbol{P_z}$ associated with local region i. The term $\bar{\boldsymbol{z}}_{\boldsymbol{i}}$ is the subvector composed of rows in $\bar{\boldsymbol{z}}$ also associated with local region i. The term $\boldsymbol{b_\beta}$ is obtained by limiting $\boldsymbol{b_\alpha}$ within the boundary as described in Section 2.3. In the case of occlusion, $\boldsymbol{b_\alpha}$ does not necessarily obey a Gaussian distribution and may be distributed far away from the mean value. Thus, by boundary limitation, the instances $\boldsymbol{Z_\beta}$ are different from the occluded local shape $\boldsymbol{Z_\alpha}$ leading to a low similarity value in Eq. 11.

The local similarity value L_S is computed for all points in a local region, leading to a local similarity map. We then build a histogram of L_S values using 50 bins to represent the values ranging from -1 to 1. Since most values in the local similarity map are close to 1, we allocate more bins near 1. Then, the histograms computed from all the local regions are concatenated into a single feature vector. Partially occluded 3D facial scans in the training set are manually labeled according to a given occlusion type (i.e., occlusion in the ocular region, occlusion in the mouth region, occlusion by glasses, or unoccluded). The distance between histograms is computed using the Euclidean metric, and the classification is performed using a simple K-NN classifier.

In our experiments, we used the Bosphorus dataset which encompasses partially occluded 3D facial scans according to several occlusion patterns. We preset a set of binary values indicating the occlusion state in each local region for each occlusion pattern. By classifying facial scans into these states, we can thus obtain a list of local regions that are occluded (m_i in Eq. 8).

5 Experimental results

The proposed statistical learning based framework for 3D facial landmarking was applied on three datasets, namely the FRGC [35], BU-3DFE [36] and Bosphorus [37] dataset. In Section 5.1, we describe the datasets and the experimental setup, and present the various experimental results in the following sections. These results are further discussed in Section 5.5.

5.1 Datasets and experimental setup

The FRGC dataset includes two versions. FRGC v1 contains 953 scans from 275 people, captured under controlled illumination conditions and generally neutral expressions [35]. However, these 953 facial scans have slight head pose and scale variation. In addition, FRGC v1 contains 33 noisy 3D facial scans having uncorrected correspondence between the range and texture maps. These scans were not used in our experiment. FRGC v2 contains 4,007 facial scans from 466 persons. These 3D facial scans were captured under different illumination conditions and contain various facial expressions (such as happiness or surprise).

The BU-3DFE database contains data from 100 subjects [36]. Each subject performed a neutral expression and six universal expressions in front of a 3D scanner. Each of these six universal expressions (happiness, disgust, fear, anger, surprise and sadness) is displayed with four levels of intensity. In our experiments, we have used the neutral facial data and facial data with expressions in the two high-level intensities from all the subjects, resulting in 1,300 facial scans in total.

The Bosphorus dataset contains 3,396 facial scans from 104 subjects [37]. This dataset contains not only the six universal facial expressions, but also 3D scans under realistic occlusions (e.g., glasses, hands around mouth and eye rubbing). Moreover, the dataset includes many male subjects that have moustache and beard.

TABLE 2
Confusion Matrix of occlusion classification

	Eye	Mouth	Glass	Unoccluded
Eye	93.3 %	2.2 %	2.4 %	2.1 %
Mouth	1.0 %	97.4 %	1.6 %	0.0 %
Glass	7.3 %	3.3 %	84.4 %	4.5 %
Unoccluded	0.0 %	0.0 %	0.0 %	100.0 %

As illustrated in Figure 5, 6, 8, we manually labeled 15 facial landmarks in the FRGC dataset and used 19 labeled landmarks in the BU-3DFE and Bosphorus datasets. They were used as ground truth for learning the SFAM model and testing our landmark fitting algorithm. These three landmark datasets contain some common landmarks, such as eye corners and mouth corners, which are sensitive to facial expressions.

5.2 Occlusion Classification Results

The proposed algorithm for occlusion detection was applied to 3D scans from the Bosphorus dataset. In our experiment, we excluded partial occlusions by hair as they do not occur in the landmark regions. We have considered partial occlusions caused by glasses, a hand near the mouth region and hand near the ocular region in addition to unoccluded 3D scans. We experimentally set K to 5 in the K-NN classifier and performed a two-fold cross-validation. The confusion matrix is provided in Table 2. An average classification accuracy up to 93.8% is achieved, which appears to be sufficient for the subsequent landmarking task.

5.3 Results on SFAM

We used 452 scans from the FRGC v1 dataset to build the SFAM-1 model by learning the local properties around 15 landmarks and their configurational relationships. The training facial scans have limited illumination variations and do not contain facial expressions.

Furthermore, we used facial scans from 11 subjects in the BU-3DFE dataset and the first 32 subjects in the Bosphorus dataset to build the SFAM-2 and SFAM-3, respectively. For every subject, 13 scans were used for training in the case of the BU-3DFE dataset (a neutral scan and the two scans for each of the six universal expressions at the intensity level three and four), and seven scans in the case of the Bosphorus dataset (a neutral scan and a scan for each of the six universal expressions). Figure 3 illustrates the SFAM-3 learnt from the Bosphorus dataset containing the first mode of configuration, local texture and local shape for variances $3 \pm \boldsymbol{\sigma}$).

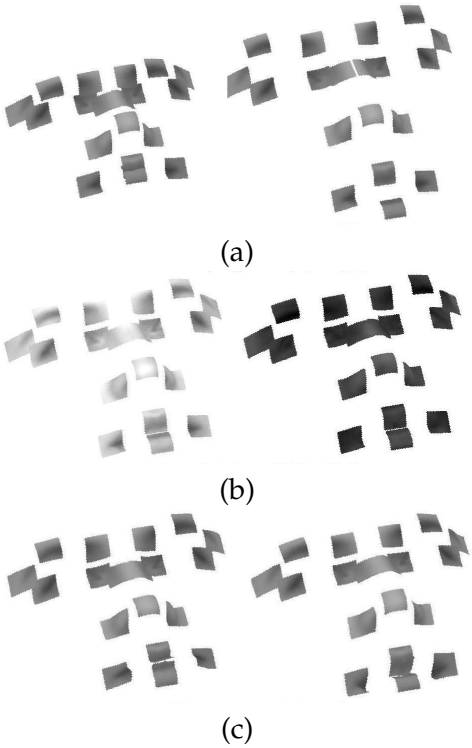


Fig. 3. SFAM learnt from the Bosphorus dataset. (a) The first landmark configuration mode explains variations in terms of the face size and expression. (b) The first texture mode explains skin color variations. (c) The first range mode explains surface geometry variations, mainly in the nose and mouth regions.

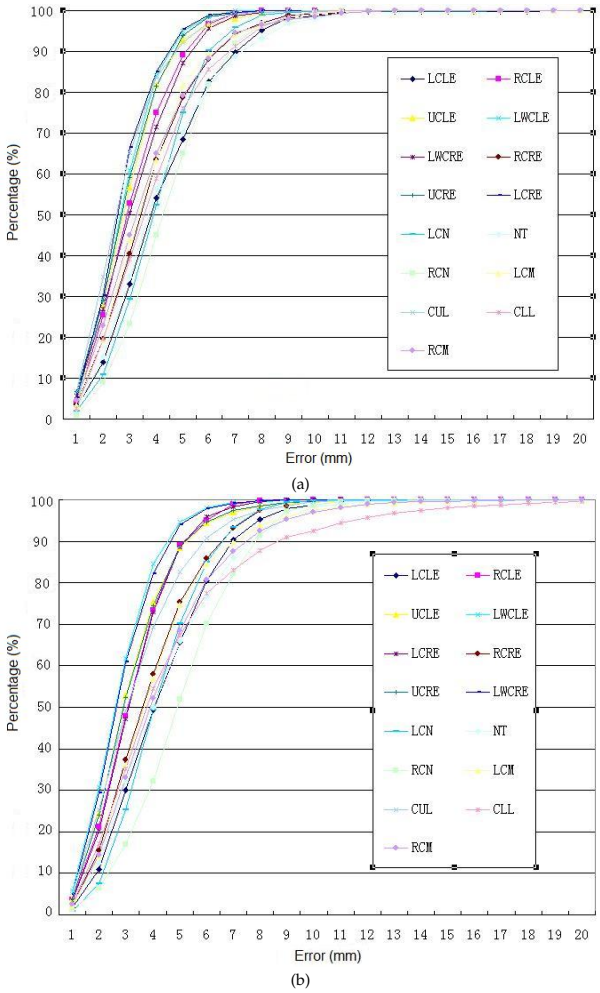


Fig. 4. Cumulative error distribution of the error for the 15 landmarks using (a) FRGC v1 and (b) FRGC v2. The symbols used are the following: LCLE: left corner of left eye; RCLE: right corner of left eye; UCLE: upper corner of left eye; LWCLE: lower corner of left eye; LCRE: left corner of right eye; RCRE: right corner of right eye; UCRE: upper corner of right eye; LWCRE: lower corner of right eye; LCN: left corner of nose; NT: nose tip; RCN: right corner of nose; LCM: left corner of mouth; CUL: center of upper lip; CLL: center of lower lip; RCM: right corner of mouth.

5.4 Results on landmarking

Using the learnt statistical models, the fitting algorithm for 3D face landmarking was evaluated on three different experimental setups. In all these experiments, the errors were computed as the Euclidean distance between the automatically localized and the corresponding manually labeled landmarks.

Using the SFAM-1, the fitting algorithm was first applied on the remaining FRGC v1 datasets (i.e., 462 scans from subjects different from those in training). We then tested the algorithm on 1,500 facial scans (randomly selected from the FRGC v2 dataset) which contain illumination variations and facial expressions. Figure 4 depicts the cumulative distribution of the fitting error for all 15 landmarks. Note that, most landmarks were automatically localized within 9 mm in both tests. Table 3 summarizes the mean, standard deviation of localization errors associated with each landmark tested on FRGC v1 and FRGC v2, and a comparison with the result achieved by a curvature analysis-based landmarking method [31]. The first two columns show the mean and the standard deviation of localization error for each landmark (d_i) from our method while the third column depicts the results achieved by the curvature analysis-based method. Note that the mean localization error of all landmarks is less than 5 mm. An increase in the mean and the standard deviation of errors generated in the experiment on FRGC v2 compared with FRGC v1 was mainly caused by uncontrolled illumination and facial expressions on tested facial scans. Compared to curvature-analysis-based method, which only uses geometry knowledge on faces, the proposed approach can locate a larger number of landmarks. The mean and standard de-

viation in localization errors from our method were smaller when compared to those obtained from the curvature analysis-based method except for the nose tip, which is the most shape salient landmark on a face. Figure 5 illustrates selected landmark localization results from the first two experiments.

The third experiment was carried out on the BU-3DFE dataset. Recall that 143 facial scans from the first five male subjects and six female subjects were used for training the SFAM-2. From the remaining 89 subjects 1,157 facial scans in total were used for testing. Each tested subject has a neutral expression and the six universal facial expressions at the intensity levels three and four. Figure 6 illustrates several localization examples having facial expressions. Figure 7 depicts the effect of expressions on landmarking accuracy. Note that landmarks with less deformation in expressions were better localized, (i.e., eye corner, nose tip, nose corner). Mouth corners and the middle of lower lip were detected with the worst accuracy and the largest standard deviation was observed in scans displaying surprise because of the large mouth displacement and ample deformation in this region. Table 4 summarizes the mean error and the standard deviation of the proposed landmarking algorithm compared to the mean error of a Point Distribution Model (PDM) [21], which is trained with 150 face scans and tested on the remainder of the BU-3DFE dataset. Because of the use of local texture and geometry knowledge in our approach, there is a significant decrease in the localization errors. The mean error for all 19 landmarks is within 10 mm while most of standard deviations are lower than 5 mm. The localization accuracy of landmarks in the rigid face region is comparable to those of the corresponding landmarks automatically localized in FRGC.

The last experiment tested the fitting algorithm using the SFAM-3 to locate 19 landmarks on 3D scans under occlusion from the Bosphorus dataset. Figure 8 illustrates several localization examples under occlusion. This experiment was carried out on 292 scans from all the subjects excluding the ones used for training in the Bosphorus dataset. To evaluate the efficiency of our proposed occlusion classifier, the fitting algorithm was first tested with occlusion knowledge directly provided by the dataset and then with occlusion knowledge from our occlusion detection and classification algorithm (Table 5). In both configurations, the mean errors ranged from 6 mm to 11 mm. Meanwhile, 71.4% of the landmarks were localized with a 10 mm precision and 97% of the landmarks were located with a 20 mm precision. Note that, there is only a slight increase on mean error and standard deviation on average when we switch the accurate knowledge on occlusion as provided by the dataset to the one provided by the proposed occlusion detection algorithm described in Section 4.

TABLE 3
A comparison of Mean error and standard deviation associated with each of the 15 landmarks on the FRGC dataset

ID	Mean (std) mm I	II	III
LCLE	4.17 (2.13)	4.31 (2.05)	7.87 (4.06)
RCLE	3.07 (1.42)	3.21 (1.44)	3.68 (1.98)
UCLE	2.92 (1.39)	3.17 (1.66)	- (-)
LWCLE	2.76 (1.21)	2.75 (1.31)	- (-)
LCRE	3.15 (1.56)	3.24 (1.43)	3.75 (1.96)
RCRE	3.67 (1.90)	3.89 (2.04)	6.59 (3.42)
UCRE	2.84 (1.45)	3.18 (1.63)	- (-)
LWCRE	2.68 (1.21)	2.83 (1.38)	- (-)
LSN	3.96 (1.65)	4.21 (1.71)	6.50 (5.36)
NT	4.11 (2.20)	4.43 (2.56)	1.93 (1.16)
RSN	4.39 (1.85)	5.07 (2.36)	6.81 (5.31)
LCM	3.61 (1.92)	4.09 (2.32)	9.10 (7.58)
CUL	2.74 (1.42)	3.37 (1.89)	- (-)
CLL	3.81 (1.97)	4.65 (3.41)	- (-)
RCM	3.58 (1.99)	4.34 (2.50)	8.83 (7.59)

The index of the landmarks is the abbreviation of the legend in Figure 4. Column I summarizes the result obtained by testing our methods on FRGCv1; column II summarizes the result obtained by testing our methods on FRGCv2; column III summarizes the results of the curvature analysis-based landmarking approach in [31]. The two tests in II and III on FRGCv2 are conducted on the same testing set. When landmarking results are not available for a point, the symbol "-" is used.

5.5 Discussion

We studied the influence of landmark configuration on the landmarking results (Table 6). Three sets of landmarks, consisting of 5, 9 and 15 landmarks, respectively, were tested on 100 facial scans randomly selected from the FRGC v1 dataset. The subjects depicted in these scans were different from the subjects used for training the SFAM, which is the SFAM-1 described in Section 5.3. From Table 6, it is evident that the mean errors remain stable (with a slight decrease in some cases), when the number of landmarks increases from 5 to 15. Meanwhile, there exists an upper bound on the number of landmarks, which depends upon the distinctiveness of landmarks so far characterized in this work based on

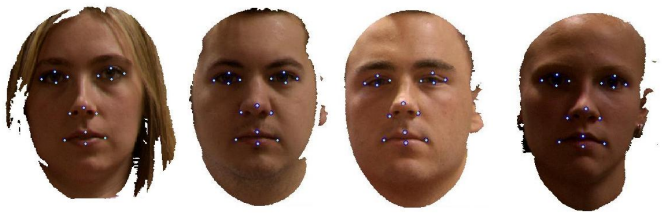


Fig. 5. Landmark localization examples from the FRGC dataset.

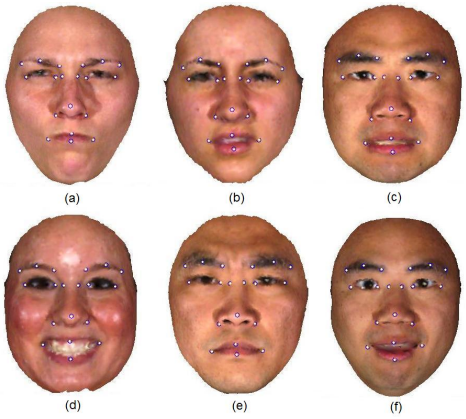


Fig. 6. Landmarking examples from the BU-3DFE dataset with expressions. (a) anger, (b) disgust, (c) fear, (d) happiness, (e) sadness, and (f) surprise.

their global configurational relationships and their local properties in terms of texture and geometric shape.

The computation time of the proposed algorithm for localizing landmarks on a scan (coded in Matlab) is around 10 *min* on a desktop PC with Intel Core i7-870 CPU and 8 *GB* RAM. The time consumed in Step 1 of the fitting algorithm is 130 s on average. It takes 70 s to 96 s to compute the correlation meshes in Step 4, depending on the density of the point clouds. The computation time for the optimization of the objective function mainly depends on the speed of convergence. In over 99% of the cases converges within 2,000 iterations or 422 s on average.

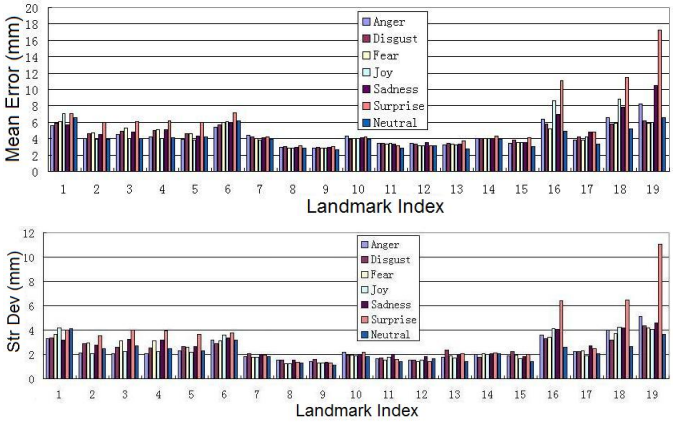


Fig. 7. Landmarking accuracy on different expressions with the BU-3DFE dataset. 1: left corner of left eyebrow, 2: middle of left eyebrow, 3: right corner of left eyebrow, 4: left corner of right eyebrow, 5: middle of left eyebrow, 6: right corner of right eyebrow, 7: left corner of left eye, 8: right corner of left eye, 9: left corner of right eye, 10: right corner of right eye, 11: left nose saddle, 12: right nose saddle, 13: left corner of nose, 14: nose tip, 15: right corner of nose, 16: left corner of mouth, 17: middle of upper lip, 18: right corner of mouth, 19: middle of lower lip

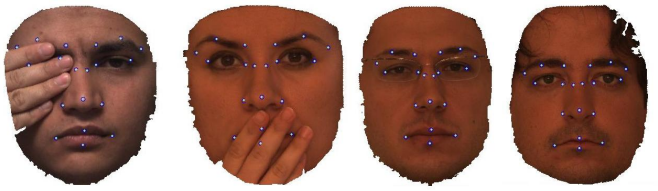


Fig. 8. Landmarking examples from the Bosphorus dataset with occlusion. From left to right, faces are occluded in the eye region, the mouth region, by glasses and by hair.

Figure 9 illustrates several failure cases of landmarking under different conditions. Cases (a) and (b) are mainly due to ample deformation on the mouth region when faces display exaggerated expressions. The morphology model in SFAM learns major variation modes from a mixture of expressions and subject identities and does not contain a specific mode for deformation caused by a specific facial expression. When fitting a SFAM on a facial scan having exaggerated facial morphology deformation (e.g., when displaying happiness and surprise), the fitting algorithm sometimes cannot generate morphology instances which approximate these extreme deformations in the mouth region. Cases (c) and (d) are mainly due to information loss in the fitting process when occlusion occurs. The occluded local regions are excluded in the fitting algorithm. Thus, the prediction of morphology parameters uses less information and is not as accurate and robust to local minima as the prediction when there is no occlusion.

We also studied the reproducibility and the corresponding accuracy of manual landmarking. For this purpose, 11 subjects were asked to manually label the 15 landmarks as defined in Figure 5 on the same 10 facial scans randomly selected from FRGC v1. We then computed the mean error and the cor-

TABLE 4
Mean error and the corresponding standard deviation (mm) of the 19 automatically localized landmarks on the facial scans from the BU-3DFE dataset (all expressions included).

ID	Mean	Std	Mean	ID	Mean	Std	Mean
1	6.26	3.72	-	11	3.30	1.70	-
2	4.58	2.82	-	12	3.27	1.56	-
3	4.87	2.99	-	13	3.32	1.94	-
4	4.88	2.97	-	14	4.04	1.99	8.83
5	4.51	2.77	-	15	3.62	1.91	-
6	6.07	3.35	-	16	7.15	4.64	-
7	4.11	1.89	20.46	17	4.19	2.34	-
8	2.93	1.40	12.11	18	7.52	4.75	-
9	2.90	1.36	11.89	19	8.82	7.12	-
10	4.07	2.00	19.38				

The index of landmarks used in this table is the same as defined in Figure 7. The mean error and Std in the first two columns are the results from our approach, the mean error in the third row are the results on a Point Distribution Model based approach [21]. The two tests were conducted on the same database with comparable experimental setups.

TABLE 5
Mean error and the corresponding standard deviation associated with the each of the 19 automatically localized landmarks on the facial scans from the Bosphorus dataset under occlusion

ID	Mean (Std) mm I	II	ID	Mean (Std) mm I	II
1	9.66 (6.08)	11.95 (8.85)	11	7.50 (3.60)	7.56 (3.88)
2	8.29 (3.92)	8.47 (4.39)	12	7.58(3.63)	6.92 (4.02)
3	7.33 (3.41)	7.15 (3.36)	13	6.35(3.11)	7.19 (2.99)
4	7.02 (3.23)	6.77 (3.38)	14	8.46(3.64)	8.39 (3.64)
5	8.21 (4.27)	8.20 (4.45)	15	8.03(3.31)	7.79 (3.36)
6	9.74 (5.23)	10.05(6.08)	16	7.96(4.18)	9.75 (6.28)
7	7.01 (3.77)	8.83 (6.37)	17	8.67(4.84)	9.01 (4.93)
8	6.25 (3.42)	6.87 (4.21)	18	8.21(4.25)	9.65 (4.97)
9	6.44 (3.08)	6.51 (3.58)	19	10.41(5.37)	10.61 (5.61)
10	7.46 (3.56)	7.86 (4.73)			

The index of landmarks used in Figure 4 is used in this table too. Column I displays the testing result using occlusion information provided by the dataset while column II displays localizing result using occlusion information provided by the proposed occlusion detection and classification algorithm. In the latter test, the knowledge of occlusion by hair (not considered by our occlusion detection and classification algorithm) was provided by the dataset

responding standard deviation of these manually labeled landmarks based on their mean landmark positions. The mean error of these manually labeled 15 landmarks was 2.49 mm with the associated standard deviation at 1.34 mm. In comparison, our localization technique achieved a mean error of 3.43 mm with the corresponding standard deviation of 1.68 mm on the same dataset.

Compared to previous 3D face landmarking algorithms [7], [8], [10], [17], [19], [21], [31], [32], our SFAM-based algorithm is a general data-driven 3D landmarking framework which encodes the configurational relationships of the landmarks and their local properties in terms of texture and shape by a statistical learning approach instead of using heuristics directly embedded within the algorithm. Thus, our algorithm is more flexible and enables localizing landmarks which are not necessarily shape prominent or texture salient.

TABLE 6
The influence of landmark configuration on mean errors (mm)

	Mean(Std) mm I	II	III
LCLE	- (-)	4.96 (2.33)	4.79 (2.15)
RCLE	3.20 (1.73)	3.15 (1.70)	3.14 (1.70)
UCLE	- (-)	- (-)	2.74 (1.30)
LWCLE	- (-)	- (-)	2.46 (1.32)
LCRE	3.60 (1.61)	3.56 (1.63)	3.56 (1.61)
RCRE	- (-)	3.73 (1.77)	3.57 (1.55)
UCRE	- (-)	- (-)	2.66 (1.08)
LWCRE	- (-)	- (-)	2.49 (1.15)
LSn	- (-)	3.92 (1.51)	3.91 (1.52)
NT	4.72 (2.58)	4.46 (2.63)	4.67 (2.51)
RSN	- (-)	4.55 (2.01)	4.41 (2.19)
LCM	3.89 (2.57)	4.07 (2.54)	3.89 (2.57)
CUL	- (-)	- (-)	2.70 (1.62)
CLL	- (-)	- (-)	4.10 (2.18)
RCM	3.77 (2.55)	3.71 (2.55)	3.75 (2.56)

The index of the landmarks is the abbreviation of the legend in Figure 4. Case I, II, III display three different landmark sets, which consist of 5, 9, 15 landmarks, respectively.

6 Conclusion

In this paper, we have presented a general learning-based framework for 3D face landmarking which proposes to characterize, through a statistical model called SFAM, the configurational relationships between the landmarks as well as their local properties in terms of texture and shape. The fitting algorithm locates the landmarks by maximizing the posteriori probability through the optimization of an objective function. The effectiveness of the framework has been demonstrated in the presence of facial expressions and partial occlusions. Consideration of both the global and local properties helps to characterize landmarks deformed under expressions. Furthermore, partial occlusion can be easily taken into account in the objective function provided that occlusion probability around each landmark can be estimated. Based on this evidence, we have also introduced a 3D facial occlusion detection and classification algorithm which exhibited a 93.8% classification accuracy on the Bosphorus dataset. This detection is based on local shape similarity between local ranges of an input 3D facial scan and the instances synthesized from SFAM. The effectiveness of our technique was supported by the experiments on the FRGC dataset (v1 and v2), BU-3DFE containing expressions and the Bosphorus dataset containing partial occlusion.

In this paper, local range and texture maps were used as simple descriptors of local shape and texture around a landmark. In future work, we plan to further improve landmark localization accuracy in considering other descriptors. We also plan to study the generalization capability of the proposed method.

7 Acknowledgment

The authors would like to express their special thanks to the anonymous reviewers for their comments which greatly contributed to the improvement of this paper. This work was supported in part

Fig. 9. Selected examples of failure cases. Facial data with (a) surprise, (b) happiness, (c) occlusion in mouth region and (d) occlusion in eye region.

by the French Research Agency within the ANR Omnia project (ANR-07-MDCO-009-02) and ANR FAR 3D project (ANR-07-SESU-004-03).

REFERENCES

[1] A. A. Salah, H. Cinar, L. Akarun, and B. Sankur, "Robust facial landmarking for registration," *Annals of Telecommunications*, vol. 62, no. 1-2, pp. 1608–1633, 2007.

[2] R. S. Feris, J. Gemmell, K. Toyama, and V. Kruger, "Hierarchical wavelet networks for facial feature localization," in *Proc. 5^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, Washington DC, May 20-21 2002, pp. 125–130.

[3] F. Y. Shih and C. Chuang, "Automatic extraction of head and face boundaries and facial features," *Information Sciences*, vol. 158, pp. 117 – 130, 2004.

[4] L. Wiskott, J. M. Fellous, N. Kruger, and C. V. D. Malburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[5] C. Tu and J. J. Lien, "Automatic location of facial feature points and synthesis of facial sketches using direct combined model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,*, vol. 40, no. 4, pp. 1158–1169, 2010.

[6] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[7] J. D'House, J. Colineau, C. Bichon, and B. Dorizzi, "Precise localization of landmarks on 3D faces using Gabor wavelets," in *Proc. International Conference on Biometrices: Theory, Applications, and Systems*, Crystal City, VA, Sep. 27-29 2007.

[8] T. Faltemier, K. Bowyer, and P. Flynn, "Rotated profile signatures for robust 3D feature detection," in *Proc. 8^{th} IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, Sep. 17-19 2008.

[9] L. Farkas, *Anthropometry of the head and face*, Second ed., L. G. Farkas, Ed. Raven Press, 1994.

[10] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 62–73, 2006.

[11] D. Colbry, G. Stockman, and A. Jain, "Detection of anchor points for 3D face verification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, Jun. 20-25 2005, pp. 118-124.

[12] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[14] D. Cristinacce and T. F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, Jan. 2008.

[15] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26^{th} Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, CA, Aug. 8–13 1999, pp. 187–194.

[16] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer journal*, vol. 7, pp. 308–313, 1965.

[17] S. Jahanbin, A. C. Bovik, and H. Choi, "Automated facial feature detection from portrait and range images," in *Proc. IEEE Southwest Symposium on Image Analysis and Interpretation*, Santa Fe, NM, Mar. 24-26 2008, pp. 25–28.

[18] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 119–152, 1994.

[19] H. Dibeklioglu, A. A. Salah, and L. Akarun, "3D facial landmarking under expression, pose, and occlusion variations," in *Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, Sep. 29-Oct. 1 2008.

[20] X. Lu, A. Jain, and D. Colbry, "Matching 2.5D face scans to 3D models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 31–43, 2006.

[21] P. Nair and A. Cavallaro, "3D face detection, landmark localization, and registration using a point distribution model," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 611–623, Jun. 2009.

[22] A. Colombo, C. Cusano, and R. Schettini, "3D face detection using curvature analysis," *Pattern Recognition*, vol. 39, no. 3, pp. 444–455, 2006.

[23] S. Jahanbin, H. Choi, R. Jahanbin, and A. C. Bovik, "Automated facial feature detecton and face recognition using gabor features on range and protrait images," in *Proc. International Conference on Image Processing*, San Diego, CA, 2008, pp. 2768–2771.

[24] V. Bevilacqua, P. Casorio, and G. Mastronardi, "Extending hough transform to a points cloud for 3D-face nose-tip detection," in *Proc. International Conference of Advanced Intelligent Computing Theories and Applications*, Shanghai, China, Sep. 15-18 2008, pp. 1200–1209.

[25] Y. Wang, C. Chua, and Y. Ho, "Facial feature detection and face recognition from 2D and 3D images," *Pattern Recognition Letters*, vol. 23, no. 10, pp. 1191–1202, 2002.

[26] B. Gokberk, H. Dutagaci, A. Ulas, L. Akarun, and B. Sankur, "Representation plurality and fusion for 3D face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 1, no. 38, pp. 155–173, 2008.

[27] C. Boehnen and T. Russ, "A fast multi-modal approach to facial feature detection," in *Proc. IEEE Workshop on Applications of Computer Vision*, Breckenridge, CO, Jan. 5-7 2005, pp. 135–142.

[28] I. A. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640–649, Apr. 2007.

[29] M. L. Koudelka, M. W. Koch, and T. D. Russ, "A prescreener for 3D face recognition using radial symmetry and the Hausdorff fraction," in *Proc. Workshop of Computer Vision and Pattern Recognition*, San Diego, CA, Jun. 20-25 2005.

[30] X. Lu and A. K. Jain, "Multimodal facial feature extraction for automatic 3D face recognition," Michigan State University, Technical Report, 2005.

[31] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *Proc. 3^{rd} International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, 2009.

[32] X. Lu and A. Jain, "Automatic feature extraction for multiview 3D face recognition," in *Proc. 7^{th} International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, Apr. 2-6 2006, pp. 585–590.

[33] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Proc. 10^{th} European Conference on Computer Vision*, Marseille, France, October 12-18, 2008, pp. 58–71.

[34] R. Niese, A. A. Hamadi, F. Aziz, and B. Michaelis, "Robust facial expression recognition based on 3D supported feature extraction and svm classification," in *Proc. International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, Sep. 17-19 2008.

[35] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, San Diego, CA, Jan 20-25 2005, pp. 947–954.

[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7^{th} International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, Apr. 10-12 2006, pp. 211–216.

[37] A. Savran, N. Alyuz, H. Dibeklioglu, O. Celiktutan, B. Gokberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. First COST 2101 Workshop on Biometrics and Identity Management*, Roskilde University, Denmark, 2008.

[38] P. Soille, *Morphological image analysis: principles and applications*. Springer-Verlag, 1999.

[39] S. Z. Li, "Markov random field modelling in image analysis," 3^{rd} edition, Springer, 2009.

[40] R. Duda, P. Hart, and D. Stork, *Pattern classification*, Wiley, Ed. Wiley, 2001.

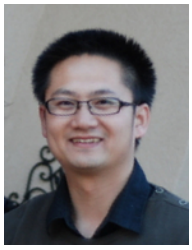
[41] S. Singer and S. Singer, "Efficient implementation of the nelder-mead search algorithm," *Applied Numerical Analysis and Computational Mathematics*, vol. 1, no. 3, pp. 524–534, 2004.

[42] P. Perakis, T. Theoharis, G. Passalis, and I. A. Kakadiaris, "Automatic 3D facial region retrieval from multi-pose facial datasets," in *Proc. Eurographics Workshop on 3D Object Retrieval*, Munich, Germany, Mar. 30 - Apr. 3 2009, pp. 37–44.

[43] P. Perakis, G. Passalis, T. Theoharis, G. Toderici, and I. A. Kakadiaris, "Partial matching of interpose 3D facial data for face recognition," in *Proc. 3^{rd} IEEE International Conference on Biometrics: Theory, Applications and Systems*, Arlington, VA, Sep. 28-30 2009.

[44] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel, "Face recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[45] M. Turk and A. Pentland, "Eigenfaces for recognition," in *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.



Xi Zhao Xi Zhao received his BSc and MSc degrees (with honors) from the School of Electronic & Information Engineering, Xi'an Jiaotong University, China, in 2003 and 2007, respectively. He obtained his PhD (with honors) in Computer Science from Ecole Centrale Lyon in 2010. He is currently conducting research as a Post-doctoral fellow at the Computational Biomedicine Lab (CBL) in University of Houston. His research interests include 3D face analysis, statistical pattern analysis and computer vision.

Emmanuel Dellandréa Emmanuel Dellandréa has received the Master and Engineer degrees in Computer Science in 2000 and the PhD in Computer Science in 2003, from Université de Tours in France. He joined in 2004 Ecole Centrale Lyon as Associate Professor. His research interests include multimedia analysis, affective computing and particularly affect recognition both in image and audio signals as well as facial expression analysis.

Liming Chen Prof. Liming Chen was awarded a joint BSc degree in Mathematics and Computer Science from the University of Nantes in 1984. He obtained a Master degree in 1986 and a PhD in computer science from the University of Paris 6 in 1989. He first served as associate professor at the Université de Technologie de Compiègne, then joined Ecole Centrale de Lyon as Professor in 1998, where he leads an advanced research team on multimedia computing and pattern recognition. From 2001 to 2003, he also served as Chief Scientific Officer in a Paris-based company, Avivias, specialized in media asset management. In 2005, he served as Scientific expert multimedia in France Telecom R&D China. He has been Head of the department of Mathematics and Computer science from 2007. Prof. Liming Chen has taken out 3 patents, authored more than 100 publications and acted as chairman, PC member and reviewer in a number of high profile journals and conferences since 1995. He has been a (co)-principal investigator on a number of research grants from EU FP programme, French research funding bodies and local government departments. He has directed more than 15 PhD theses. His current research spans from 2D/3D face analysis and recognition, image and video analysis and categorization, to affect analysis both in image, audio and video.

Ioannis A. Kakadiaris Prof. Ioannis A. Kakadiaris is a Hugh Roy and Lillie Cranz Cullen Professor of Computer Science, Electrical & Computer Engineering, and Biomedical Engineering at the University of Houston. He joined UH in August 1997 after a postdoctoral fellowship at the University of Pennsylvania. Ioannis earned his B.Sc. in physics at the University of Athens in Greece, his M.Sc. in computer science from Northeastern University and his Ph. D. at the University of Pennsylvania. He is the founder of the Computational Biomedicine Lab (www.cbl.uh.edu) and in 2008 directed the Methodist-University of Houston-Weill Cornell Medical College Institute for Biomedical Imaging Sciences (IBIS) (ibis.uh.edu). His research interests include biometrics, non-verbal human behavior understanding, computational life sciences, energy informatics, computer vision and pattern recognition. Dr. Kakadiaris is the recipient of a number of awards, including the NSF Early Career Development Award, Schlumberger Technical Foundation Award, UH Computer Science Research Excellence Award, UH Enron Teaching Excellence Award, and the James Muller Vulnerable Plaque Young Investigator Prize.