Accurate Landmarking of Three-Dimensional Facial Data in the Presence of Facial Expressions and Occlusions Using a Three-Dimensional Statistical Facial Feature Model

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

***Abstract*—Three-dimensional 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 3-D facial data under challenging conditions (i.e., facial expressions and occlusions). Our approach relies on a statistical model, called 3-D statistical facial feature model, 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 3-D face databases (FRGC, BU-3-DFE, and Bosphorus) demonstrate the effectiveness of our approach, in terms of landmarking accuracy and robustness, in the presence of expressions and occlusions.**

***Index Terms*—Facial expression, fitting, landmarks, occlusion, statistical face model, 3-D face feature.**

I. Introduction

THE RECENT emergence of 3-D facial data has provided an alternative to overcome the challenges in 2-D face recognition, caused by pose changes and lighting variations [6]. Although 2.5D/3-D 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., 3-D 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/3-D face landmarks, a reliable localization of these facial feature points is essential.

A. 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, 3-D 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 3-D face landmarking, which is robust to changes in facial expressions and partial occlusions.

Face landmarking on 2-D 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 2-D statistical models, such as the popular active appearance model [12] or the more recent constrained local model (CLM) [14], which perform statistical analysis both on the facial appearance and the 2-D shape. However, since they are applied to 2-D texture images, these approaches inherit the sensitivity to lighting and pose changes.

Manuscript received December 23, 2009; revised November 18, 2010, February 14, 2011, and March 29, 2011; accepted April 8, 2011. This work was supported in part by the French National Research Agency (ANR) through the ANR Omnia project under Grant ANR-07-MDCO-009-02 and through the ANR FAR 3-D project under Grant ANR-07-SESU-004-03. This paper was recommended by Associate Editor J. Su.

X. Zhao and I. A. Kakadiaris are with the Computational Biomedicine Laboratory, Department of Computer Science, University of Houston, Houston, TX 77204-3010 USA (e-mail: zhaoxi@ieee.org; ioannisk@uh.edu).

E. Dellandréa and L. Chen are with the Université de Lyon, Centre National de la Recherche Scientifique, Ecole Centrale Lyon, Laboratoire d'InfoRmatique en Image et Systèmes d'information, Unité Mixte de Recherche 5205, 69134 Ecully Cedex, France (e-mail: emmanuel.dellandrea@ec-lyon.fr; liming.chen@ec-lyon.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2011.2148711

Research on 3-D face landmarking is rather recent. Most of the existing methods embed *a priori* knowledge on landmarks on 3-D face by computing the response to local 3-D 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 3-D 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 and Russ [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 3-D face mesh as well as Gabor jets of landmarks from the 2-D 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 and Jain [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 3-D geometry-based confidence [27], an extended elastic bunch graph [23], or a simple mean model constructed as the average 3-D position of landmarks from a learning data set [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 and Cavallaro [21] used their 3-D point distribution model (PDM) 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 3-D 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 3-D facial data under multiple orientations, taking into account missing data due to self occlusion.

B. *Proposed Approach*

In this paper, we propose a general learning-based framework for 3-D 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 3-D Statistical Facial feAture Model (SFAM), which learns both the global variations in 3-D face morphology and the local variations around each 3-D face landmark in terms of texture and geometry. To train the model, we manually labeled the target landmarks for each aligned frontal 3-D 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 3-D 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 3-D partial face instances that consist of local face regions with texture and shape (structured by their global 3-D 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 3-D face model [15] and the CLM [14] as we propose to learn, at the same time, the global variations of 3-D face morphology and the local ones in terms of texture and shape around each landmark. Fitting the SFAM on

TABLE I
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

a probe facial scan is accomplished by maximum *a posteriori* (MAP) probability. The fitted morphology instance delivers the locations of targeted landmarks. Using 3-D training faces with expressions, the SFAM has the ability to learn expression variations and generate instances with the learned variations so as to increase the *a posteriori* probability in fitting faces with expression. Furthermore, we propose to use a k-nearest neighbor (k-NN) 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 the 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 3-D facial scans.

The main contributions of this paper are the following.

1) We build an SFAM 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 partially occluded faces.

The rest of this paper is organized as follows. In Section II, our statistical model SFAM is introduced. In Section III, the objective function that combines the local shape and texture properties and the fitting algorithm are described. Section IV addresses 3-D face partial occlusion. Experimental results are discussed in Section V, while Section VI concludes this paper. Table I presents a summary of the different symbols used in this paper.

II. SFAM

Three-dimensional facial data acquired by the current 3-D imaging systems are usually noisy and may contain holes and spikes. Hence, we first preprocess all the 3-D facial scans to remove noise. Head pose and scale variations are normalized by alignment and remeshing (see Section II-A). Then, we model the variations in 3-D configurations of landmarks and their local variations in terms of texture and shape around each landmark (see Section II-B). New partial 3-D face instances can be synthesized from the learned model (see Section II-C).

A. Preprocessing the Training Facial Data

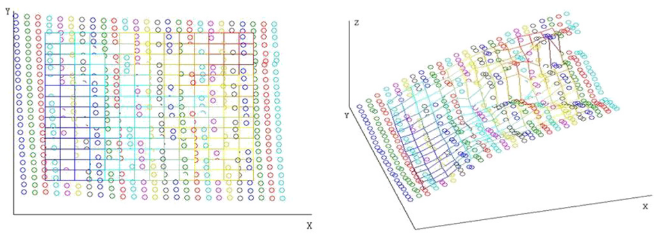
To remove the noise (e.g., spikes and holes) and enhance the quality of 3-D facial scans, we perform the following operations.

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 3-D 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 3-D 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 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 have been compensated after alignment, the SFAM can be learned 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 3-D face. This is the major difference between the proposed approach and most of the existing 3-D 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 data set, 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 data sets: FRGC, BU-3-DFE, and Bosphorus data sets. Most landmarks out of 15 (as illustrated in Fig. 5) on the FRGC data set were selected from the rigid part of the face as they were subsequently used for 3-D face recognition. On the other hand, landmarks on the BU-3-DFE and the Bosphorus data sets (as illustrated in Figs. 6 and 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



(a) (b)

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

to face scale. Second, a uniform grid is associated with each landmark. As illustrated in Fig. 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.

B. Modeling the Configurational Relationships and Local Shape and Texture Features of the Landmarks

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

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

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

We further generate the 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

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

PCA is then applied to the three vector sets $\{\boldsymbol{s_k}\}$, $\{\boldsymbol{g_k}\}$, and $\{\boldsymbol{z_k}\}$, extracted from the training 3-D facial data (k denotes the kth 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 $\bar{\boldsymbol{s}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{z}}$ are the mean landmark configuration, the mean intensity, and the mean range value, respectively, while $\boldsymbol{P_s}$, $\boldsymbol{P_g}$, and $\boldsymbol{P_z}$ are the three sets of modes of configuration, intensity, and depth variation, respectively. The terms $\boldsymbol{b_s}$, $\boldsymbol{b_g}$, and $\boldsymbol{b_z}$ 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 $\boldsymbol{\sigma_q}$.

C. Synthesizing Instances From a New Face

Given the parameters $\boldsymbol{b_s}$, a configuration instance can be generated using (3). Then, given a new facial scan, the set of scan points closest to the configuration instance is 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 (2). Then, $\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ are estimated as follows:

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

$\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ 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 (4). The landmarks, along with their local texture and local shape instances, compose a partial face instance.

III. Localizing Landmarks

The SFAM-based landmark localization procedure consists of MAP probability of landmark configuration, given a 3-D facial scan to be landmarked, and leads to optimizing an objective function. In Section III-A, we present the objective AQ14 function to be optimized, and in Section III-B, we introduce the fitting algorithm for localizing landmarks. We then discuss our assumptions in Section III-C.

A. Objective Function and MAP

We first define the objective function $f(\boldsymbol{b_s}) = p(\boldsymbol{s}|T, R, \psi)$ as the *a posteriori* probability of landmark configuration $\boldsymbol{s}$ to be maximized for a 3-D facial scan represented by its texture map T and range map R and the learned 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 (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

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

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 (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 paper, 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 (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):

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

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

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

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 $\sum_{j=1}^{K}(b_j^2/\lambda_j)$, $\sum_{i=1}^{N} F_{zi}$, and $\sum_{j=1}^{K}(b_j^2/\lambda_j)$, respectively, during the offline training.

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

B. Fitting Algorithm

Landmarking a 3-D facial scan consists of fitting the SFAM ψ while maximizing the objective function (8). First, the 3-D facial scan is preprocessed as described in Section II-A, including spike removal, hole filling, and head pose normalization. The occlusion algorithm, introduced in Section IV, 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.

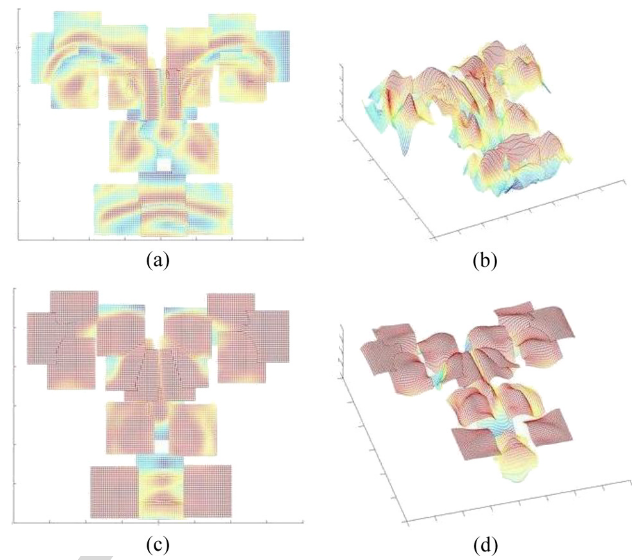


Fig. 2. Depiction of the correlation meshes from the frontal and side views. These meshes capture the similarity between instances and local facial regions in both texture and shape representations. The red color corresponds to large correlation values while blue corresponds to small correlation values. 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).

Algorithm 1 SFAM Fitting

Input: A 3-D 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 $\mathcal{S}$.

2. Synthesize texture and shape instances $\hat{G}$, $\hat{Z}$ as described in Section II-C using $\mathcal{S}$.

3. Normalize local regions around points $\mathcal{S}$ within a neighborhood large enough to cover the potential landmark locations as in Section II-A, creating a set of local mesh $\mathcal{G}$, $\mathcal{Z}$.

4. Compute correlation meshes on both texture and geometry representations (see Fig. 2) by correlating $\hat{G}$, $\hat{Z}$ with G, Z, respectively, which are different parts of $\mathcal{G}$, $\mathcal{Z}$ sampled by a sliding window (size of 15×15) on local regions (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 landmark configuration defined by the control parameters $\boldsymbol{b_s}$.

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

The optimization process in steps one and five of the algorithm is processed by the Nelder–Mead 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 nonface meshes lead the optimization to converge to an unpredictable point far from the desired minimum.

C. Discussion

To deduce (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 3-D face landmarking is actually a Markov random field (MRF) which consists of assigning a label from a set of labels $\mathcal{L}$ to each vertex of a 3-D facial scan. The set $\mathcal{L}$ encompasses all targeted landmarks (e.g., nose tip and 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–Mead 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].

IV. Occlusion Detection and Classification

Facial data analysis in the presence of partial occlusions (caused by a variety of factors such as hair, glasses, mustaches, and scarf) is a difficult problem. In 3-D facial landmarking, only occlusions which may occur in local regions around landmarks are of interest. Thus, in this paper, 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.

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 $\bar{\boldsymbol{s}}(3)$ are first computed. Then, grids (50×50) are used to remesh local regions around these points for range values (see Section II-A). 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 the 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:

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

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

where $\boldsymbol{P}_{\boldsymbol{z},\boldsymbol{i}}$ 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 the 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 II-C. 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 (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 3-D 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 data set which encompasses partially occluded 3-D 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 (8)].

V. Experimental Results

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

A. Data Sets and Experimental Setup

The FRGC data set 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 3-D facial scans having uncorrected correspondence between the range and texture maps. These scans were not used in our experiment. FRGC v2 contains 4007 facial scans from 466 persons. These 3-D facial scans were captured under different illumination conditions and contain various facial expressions (such as happiness or surprise).

The BU-3-DFE database contains data from 100 subjects [36]. Each subject performed a neutral expression and six universal expressions in front of a 3-D 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 1300 facial scans in total.

The Bosphorus data set contains 3396 facial scans from 104 subjects [37]. This data set contains not only the six universal facial expressions but also 3-D scans under realistic occlusions (e.g., glasses, hands around the mouth, and eye rubbing).

TABLE II
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 %

Moreover, the data set includes many male subjects that have moustache and beard.

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

B. Occlusion Classification Results

The proposed algorithm for occlusion detection was applied to 3-D scans from the Bosphorus data set. 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 a hand near the ocular region in addition to unoccluded 3-D scans. We experimentally set k to five in the k-NN classifier and performed a two-fold cross-validation. The confusion matrix is provided in Table II. An average classification accuracy up to 93.8% is achieved, which appears to be sufficient for the subsequent landmarking task.

C. Results on SFAM

We used 452 scans from the FRGC v1 data set 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-3-DFE data set and the first 32 subjects in the Bosphorus data set to build the SFAM-2 and SFAM-3, respectively. For every subject, 13 scans were used for training in the case of the BU-3-DFE data set (a neutral scan and the two scans for each of the six universal expressions at the intensity levels three and four), and seven scans in the case of the Bosphorus data set (a neutral scan and a scan for each of the six universal expressions). Fig. 3 illustrates the SFAM-3 learned from the Bosphorus data set containing the first mode of configuration, local texture, and local shape for variances $3 \pm \boldsymbol{\sigma}$.

D. Results on Landmarking

Using the learned statistical models, the fitting algorithm for 3-D 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.

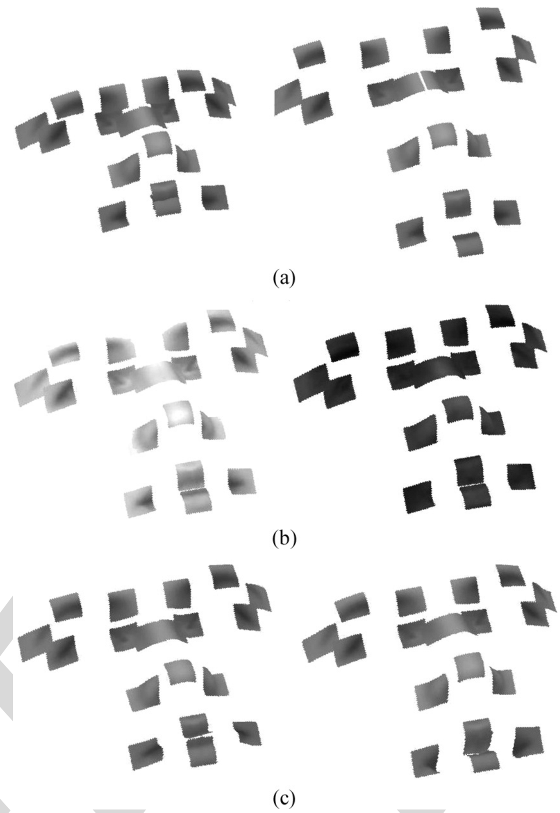
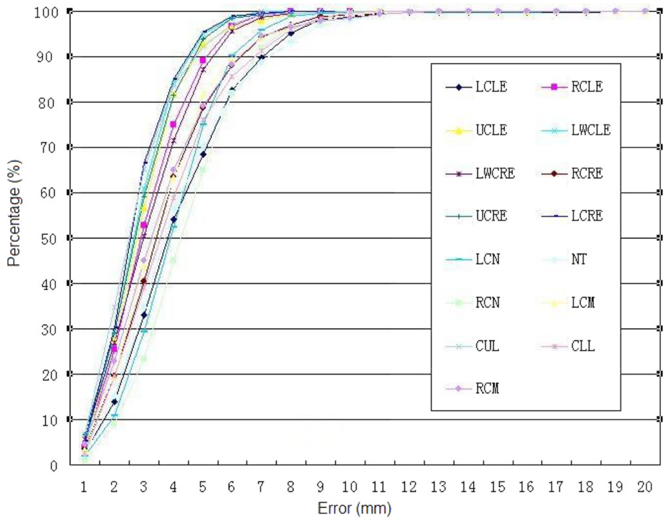
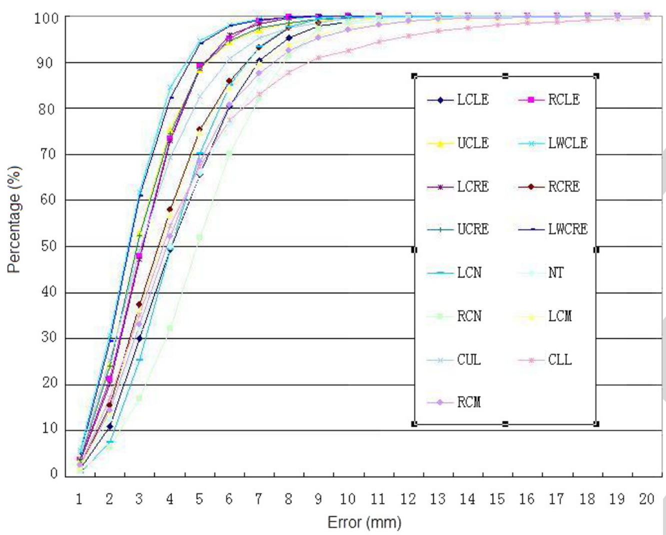


Fig. 3. SFAM learned from the Bosphorus data set. (a) First landmark configuration mode explains variations in terms of the face size and expression. (b) First texture mode explains skin color variations. (c) First range mode explains surface geometry variations, mainly in the nose and mouth regions.

Using the SFAM-1, the fitting algorithm was first applied on the remaining FRGC v1 data sets (i.e., 462 scans from subjects different from those in training). We then tested the algorithm on 1500 facial scans (randomly selected from the FRGC v2 data set) which contain illumination variations and facial expressions. Fig. 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 III summarizes the mean, the 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 deviation 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. Fig. 5 illustrates selected landmark localization results from the first two experiments.



(a)



(b)

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, and RCM—right corner of mouth.

AQ15

The third experiment was carried out on the BU-3-DFE data set. 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, 1157 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. Fig. 6 illustrates several localization examples having facial expressions. Fig. 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, and nose corner). Mouth corners and the middle of the 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 IV summarizes

TABLE III
COMPARISON OF MEAN ERROR AND STANDARD DEVIATION ASSOCIATED WITH EACH OF THE 15 LANDMARKS ON THE FRGC DATA SET

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)

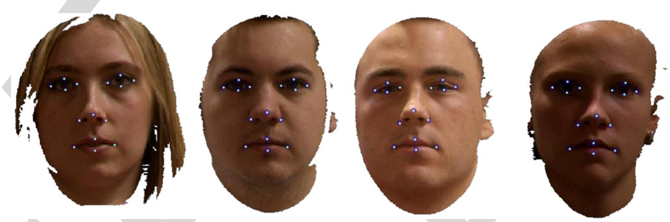


Fig. 5. Landmark localization examples from the FRGC data set.

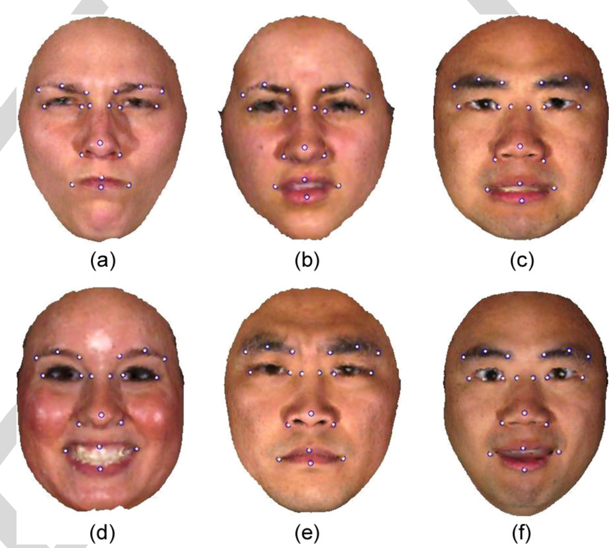


Fig. 6. Landmarking examples from the BU-3-DFE data set with expressions. (a) Anger. (b) Disgust. (c) Fear. (d) Happiness. (e) Sadness. (f) Surprise.

the mean error and the standard deviation of the proposed landmarking algorithm compared to the mean error of a PDM [21], which is trained with 150 face scans and tested on the remainder of the BU-3-DFE data set. 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.

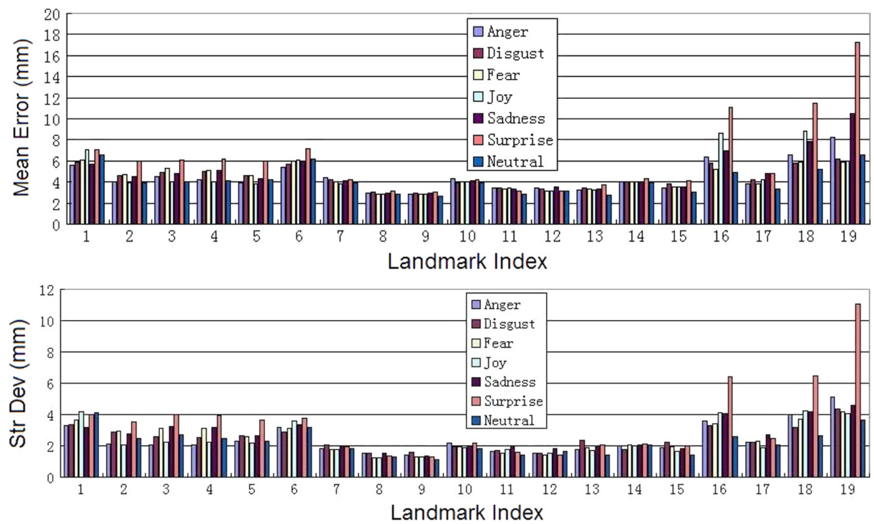


Fig. 7. Landmarking accuracy on different expressions with the BU-3-DFE data set. 1: Left corner of left eyebrow. 2: Middle of left eyebrow. 3: Right corner of left eyebrow. 4: Left corner of right eyebrow. 5: Middle of left eyebrow. 6: Right corner of right eyebrow. 7: Left corner of left eye. 8: Right 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.

TABLE IV

MEAN ERROR AND THE CORRESPONDING STANDARD DEVIATION (IN MILLIMETERS) OF THE 19 AUTOMATICALLY LOCALIZED LANDMARKS ON THE FACIAL SCANS FROM THE BU-3-DFE DATA SET (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				

TABLE V

MEAN ERROR AND THE CORRESPONDING STANDARD DEVIATION ASSOCIATED WITH EACH OF THE 19 AUTOMATICALLY LOCALIZED LANDMARKS ON THE FACIAL SCANS FROM THE BOSPHORUS DATA SET UNDER OCCLUSION

ID	Mean (Std) mm I	Mean (Std) mm II	ID	Mean (Std) mm I	Mean (Std) mm 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)			

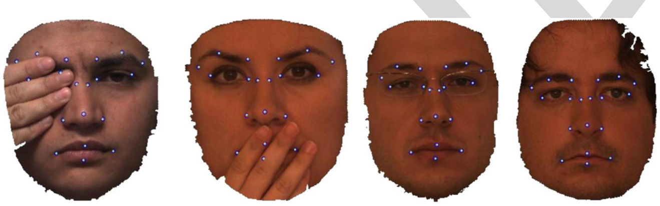


Fig. 8. Landmarking examples from the Bosphorus data set with occlusion. From left to right, faces are occluded in the eye region, in the mouth region, by glasses, and by hair.

The last experiment tested the fitting algorithm using the SFAM-3 to locate 19 landmarks on 3-D scans under occlusion from the Bosphorus data set. Fig. 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 data set. To evaluate the efficiency of our proposed occlusion classifier, the fitting algorithm was first tested with occlusion knowledge directly provided by the data set and, then, with occlusion knowledge from our occlusion detection and classification algorithm (see Table V). In both configurations, the mean errors ranged from 6 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 data set to the one provided by the proposed occlusion detection algorithm described in Section IV.

E. Discussion

We studied the influence of landmark configuration on the landmarking results (see Table VI). Three sets of landmarks, consisting of 5, 9, and 15 landmarks, respectively, were tested on 100 facial scans randomly selected from the FRGC v1 data set. The subjects depicted in these scans were different from the subjects used for training the SFAM, which is the SFAM-1 described in Section V-C. From Table VI, 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 paper based on their global configurational

TABLE VI
INFLUENCE OF LANDMARK CONFIGURATION ON MEAN ERRORS (IN MILLIMETERS)

	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)

(a) (b) (c) (d)

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.

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 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. Over 99% of the
AQ16 cases converge within 2000 iterations or 422 s on average.

Fig. 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 the 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 an 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 Fig. 5 on the same 10 facial scans randomly selected from FRGC v1. We then computed the mean error and the corresponding 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 data set.

Compared to previous 3-D face landmarking algorithms [7], [8], [10], [17], [19], [21], [31], [32], our SFAM-based algorithm is a general data-driven 3-D 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.

VI. CONCLUSION

In this paper, we have presented a general learning-based framework for 3-D 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 *a 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 the occlusion probability around each landmark can be estimated. Based on this evidence, we have also introduced a 3-D facial occlusion detection and classification algorithm which exhibited a 93.8% classification accuracy on the Bosphorus data set. This detection is based on local shape similarity between local ranges of an input 3-D facial scan and the instances synthesized from the SFAM. The effectiveness of our technique was supported by the experiments on the FRGC data set (v1 and v2), BU-3-DFE containing expressions, and the Bosphorus data set 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.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments which greatly contributed to the improvement of this paper.

REFERENCES

[1] A. A. Salah, H. Cinar, L. Akarun, and B. Sankur, "Robust facial landmarking for registration," *Ann. Telecommun.*, 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. 5th IEEE Int. Conf. Autom. Face Gesture Recog.*, 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," *Inf. Sci.*, vol. 158, pp. 117–130, Jan. 2004.

[4] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[5] C. Tu and J. J. Lien, "Automatic location of facial feature points and synthesis of facial sketches using direct combined model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1158–1169, Aug. 2010.

[6] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D +2D face recognition," *Comput. Vis. Image Understand.*, 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. Int. Conf. Biometrics: Theory, Appl., Syst.*, Crystal City, VA, Sep. 27–29, 2007, pp. 1–6.

[8] T. Faltemier, K. Bowyer, and P. Flynn, "Rotated profile signatures for robust 3D feature detection," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recog.*, Amsterdam, The Netherlands, Sep. 17–19, 2008, pp. 1–7.

[9] L. Farkas, *Anthropometry of the Head and Face*, L. G. Farkas, Ed., 2nd ed. New York: Raven, 1994.

[10] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1487–1494, Oct. 2006.

[11] D. Colbry, G. Stockman, and A. Jain, "Detection of anchor points for 3D face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, San Diego, CA, Jun. 20–25, 2005, pp. 118–124.

[12] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[14] D. Cristinacce and T. F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, 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. 26th Annu. Conf. Comput. Graph. Interactive Techn.*, Los Angeles, CA, Aug. 8–13, 1999, pp. 187–194.

[16] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, Jan. 1965.

[17] S. Jahanbin, A. C. Bovik, and H. Choi, "Automated facial feature detection from portrait and range images," in *Proc. IEEE Southwest Symp. Image Anal. 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," *Int. J. Comput. Vis.*, vol. 13, no. 2, pp. 119–152, Oct. 1994.

[19] H. Dibeklioglu, A. A. Salah, and L. Akarun, "3D facial landmarking under expression, pose, and occlusion variations," in *Proc. IEEE Int. Conf. Biometrics: Theory, Appl. Syst.*, Arlington, VA, Sep. 29–Oct. 1, 2008, pp. 1–6.

[20] X. Lu, A. Jain, and D. Colbry, "Matching 2.5D face scans to 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, Jan. 2006.

[21] P. Nair and A. Cavallaro, "3D face detection, landmark localization, and registration using a point distribution model," *IEEE Trans. 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 Recognit.*, vol. 39, no. 3, pp. 444–455, Mar. 2006.

[23] S. Jahanbin, H. Choi, R. Jahanbin, and A. C. Bovik, "Automated facial feature detection and face recognition using Gabor features on range and portrait images," in *Proc. Int. Conf. Image Process.*, 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. Int. Conf. Adv. Intell. Comput. Theories Appl.*, 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 Recognit. Lett.*, vol. 23, no. 10, pp. 1191–1202, Aug. 2002.

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

[27] C. Boehnen and T. Russ, "A fast multi-modal approach to facial feature detection," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 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 Trans. Pattern Anal. Mach. Intell.*, 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 Comput. Vis. Pattern Recog.*, San Diego, CA, Jun. 20–25, 2005, p. 168.

[30] X. Lu and A. K. Jain, "Multimodal facial feature extraction for automatic 3D face recognition," Michigan State Univ., East Lansing, MI, Tech. Rep. MSU-CSE-05-22, 2005.

[31] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *Proc. 3rd Int. Conf. Biometrics: Theory, Appl. Syst.*, Washington, DC, 2009, pp. 1–6.

[32] X. Lu and A. Jain, "Automatic feature extraction for multiview 3D face recognition," in *Proc. 7th Int. Conf. Autom. Face Gesture Recog.*, Southampton, U.K., 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. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 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. Int. Conf. Autom. Face Gesture Recog.*, Amsterdam, The Netherlands, Sep. 17–19, 2008, pp. 1–7.

[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 Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, San Diego, CA, Jan. 20–25, 2005, vol. 1, 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. 7th Int. Conf. Autom. Face Gesture Recog.*, Southampton, U.K., 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. 1st COST 2101 Workshop Biometrics Identity Manage.*, 2008, pp. 47–56.

[38] P. Soille, *Morphological Image Analysis: Principles and Applications*. New York: Springer-Verlag, 1999.

[39] S. Z. Li, *Markov Random Field Modelling in Image Analysis*, 3rd ed. New York: Springer-Verlag, 2009.

[40] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Hoboken, NJ: Wiley, 2001.

[41] S. Singer and S. Singer, "Efficient implementation of the Nelder–Mead search algorithm," *Appl. Numer. Anal. Comput. Math.*, vol. 1, no. 2, pp. 524–534, Dec. 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 3D Object Retrieval*, Munich, Germany, Mar. 30–Apr. 3, 2009, pp. 37–44.

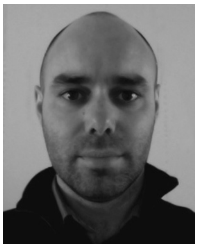
[43] P. Perakis, G. Passalis, T. Theoharis, G. Toderici, and I. A. Kakadiaris, "Partial matching of interpose 3D facial data for face recognition," in *Proc. 3rd IEEE Int. Conf. Biometrics: Theory, Appl. Syst.*, Arlington, VA, Sep. 28–30, 2009, pp. 1–8.

[44] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proc. IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[45] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

Xi Zhao (S'XX) received the B.Sc. and M.Sc. degrees (with honors) from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2003 and 2007, respectively, and the Ph.D. degree (with honors) in computer science from Ecole Centrale Lyon, Ecully, France, in 2010. AQ17

He is currently conducting research as a Postdoctoral Fellow at the Computational Biomedicine Laboratory, University of Houston, Houston, TX. His research interests include 3-D face analysis, statistical pattern analysis, and computer vision.

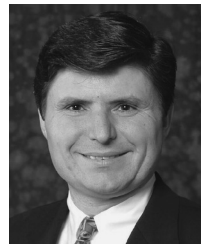


Emmanuel Dellandréa received the Master and Engineer degrees in computer science and the Ph.D. degree in computer science from Université de Tours, AQ18 Tours, France, in 2000 and 2003, respectively.

In 2004, he joined Ecole Centrale Lyon, Ecully, France, as an 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.

AQ19 **Liming Chen** (M'XX) was awarded the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984. AQ20 He obtained the M.S. and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor at the Université de Technologie de Compiègne, AQ21 Compiègne, France, and then joined Ecole Centrale de AQ22 Lyon, Ecully, France, 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 the Head of the Department of Mathematics and Computer Science since 2007. He has taken out three 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 AQ23 a (co)-principal investigator on a number of research grants from the European AQ24 Union FP programme, French research funding bodies, and local government departments. He has directed more than 15 Ph.D. theses. His current research spans from 2-D/3-D face analysis and recognition and image and video analysis and categorization to affect analysis both in image audio and video.



Ioannis A. Kakadiaris (SM'XX) AQ25 received the B.Sc. degree in physics from the University of Athens, Athens, Greece, the M.Sc. degree in computer science from Northeastern University, Boston, MA, and the Ph.D. degree from the University of Pennsylvania, Philadelphia.

He is a Hugh Roy and Lillie Cranz Cullen Professor of Computer Science, Electrical and Computer Engineering, and Biomedical Engineering at the University of Houston (UH), Houston, TX. He joined UH in August 1997 after a postdoctoral fellowship at the University of Pennsylvania. He is the founder of the Computational Biomedicine Laboratory (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, nonverbal 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 National Science Foundation Early Career Development Award, Schlumberger Technical AQ26 Foundation Award, UH Computer Science Research Excellence Award, UH Enron Teaching Excellence Award, and the James Muller Vulnerable Plaque Young Investigator Prize.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

Note that your paper will incur overlength page charges of $175 per page. The page limit for regular papers is 12 pages, and the page limit for correspondence papers is 6 pages.

AQ1 = The sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ2 = "ANR" is defined as "French National Research Agency." Please check if appropriate, and correct if necessary.
AQ3 = "ANR-07-MDCO-009-02" and "ANR-07-SESU-004-03" were captured as grants. Please check if appropriate, and correct if necessary.
AQ4 = "CNRS" was expanded as "Centre National de la Recherche Scientifique." Please check if appropriate, and correct if necessary.
AQ5 = "LIRIS" was expanded as "Laboratoire d'InfoRmatique en Image et Systèmes d'information." Please check if appropriate, and correct if necessary.
AQ6 = "UMR" was expanded as "Unité Mixte de Recherche." Please check if appropriate, and correct if necessary.
AQ7 = "Boehnen *et al.*" was changed to "Boehnen and Russ." Please check if appropriate, and correct if necessary.
AQ8 = "Lu *et al.*" was changed to "Lu and Jain." Please check if appropriate, and correct if necessary.
AQ9 = "Nair *et al.*" was changed to "Nair and Cavallaro." Please check if appropriate, and correct if necessary.
AQ10 = "MAP" is defined as "maximum a posteriori." Please check if appropriate, and correct if necessary.
AQ11 = This sentence was rephrased. Please check if the original thought was retained, and correct if necessary.
AQ12 = All occurrences of "posteriori probability" were changed to "a posteriori probability." Please check if appropriate, and correct if necessary.
AQ13 = All occurrences of "kNN" and "K-NN" were captured as "k-NN" to refer to "k-nearest neighbor" and for consistency. Please check if appropriate, and correct if necessary.
AQ14 = This sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ15 = The caption was restructured for clarity of presentation. Please check if Please check if appropriate, and correct if necessary.
AQ16 = The sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ17 = Please provide the IEEE membership history of author Xi Zhao.
AQ18 = The sentence was rephrased. Please check if the original thought was retained, and correct if necessary.
AQ19 = Please provide the IEEE membership history of author Liming Chen.
AQ20 = "Master" was changed to "M.S." Please check if appropriate, and correct if necessary.
AQ21 = Please validate the address provided for the Université de Technologie de Compiègne.
AQ22 = Please validate the address provided for Ecole Centrale de Lyon.
AQ23 = "EU" was expanded as "European Union." Please check if appropriate, and correct if necessary.
AQ24 = Please provide the expanded form of "FP."
AQ25 = Please provide the IEEE membership history of author Ioannis A. Kakadiaris.
AQ26 = "NSF" was expanded as "National Science Foundation." Please check if appropriate, and correct if necessary.

END OF ALL QUERIES

Accurate Landmarking of Three-Dimensional Facial Data in the Presence of Facial Expressions and Occlusions Using a Three-Dimensional Statistical Facial Feature Model

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

***Abstract*—Three-dimensional 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 3-D facial data under challenging conditions (i.e., facial expressions and occlusions). Our approach relies on a statistical model, called 3-D statistical facial feature model, 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 3-D face databases (FRGC, BU-3-DFE, and Bosphorus) demonstrate the effectiveness of our approach, in terms of landmarking accuracy and robustness, in the presence of expressions and occlusions.**

***Index Terms*—Facial expression, fitting, landmarks, occlusion, statistical face model, 3-D face feature.**

I. Introduction

THE RECENT emergence of 3-D facial data has provided an alternative to overcome the challenges in 2-D face recognition, caused by pose changes and lighting variations [6]. Although 2.5D/3-D 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., 3-D 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/3-D face landmarks, a reliable localization of these facial feature points is essential.

A. 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, 3-D 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 3-D face landmarking, which is robust to changes in facial expressions and partial occlusions.

Face landmarking on 2-D 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 2-D statistical models, such as the popular active appearance model [12] or the more recent constrained local model (CLM) [14], which perform statistical analysis both on the facial appearance and the 2-D shape. However, since they are applied to 2-D texture images, these approaches inherit the sensitivity to lighting and pose changes.

Manuscript received December 23, 2009; revised November 18, 2010,
February 14, 2011, and March 29, 2011; accepted April 8, 2011. This work
was supported in part by the French National Research Agency (ANR) through
AQ1 the ANR Omnia project under Grant ANR-07-MDCO-009-02 and through the
AQ2 ANR FAR 3-D project under Grant ANR-07-SESU-004-03. This paper was
AQ3 recommended by Associate Editor J. Su.

X. Zhao and I. A. Kakadiaris are with the Computational Biomedicine Laboratory, Department of Computer Science, University of Houston, Houston, TX 77204-3010 USA (e-mail: zhaoxi@ieee.org; ioannisk@uh.edu).

E. Dellandréa and L. Chen are with the Université de Lyon, Centre National
de la Recherche Scientifique, Ecole Centrale Lyon, Laboratoire d'InfoRmatique
AQ4 en Image et Systèmes d'information, Unité Mixte de Recherche 5205, 69134
AQ5 Ecully Cedex, France (e-mail: emmanuel.dellandrea@ec-lyon.fr; liming.chen@
AQ6 ec-lyon.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2011.2148711

Research on 3-D face landmarking is rather recent. Most of the existing methods embed *a priori* knowledge on landmarks on 3-D face by computing the response to local 3-D 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 3-D 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 and Russ [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 3-D face mesh as well as Gabor jets of landmarks from the 2-D 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 and Jain [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 3-D geometry-based confidence [27], an extended elastic bunch graph [23], or a simple mean model constructed as the average 3-D position of landmarks from a learning data set [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 and Cavallaro [21] used their 3-D point distribution model (PDM) 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 3-D 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 3-D facial data under multiple orientations, taking into account missing data due to self occlusion.

B. Proposed Approach

In this paper, we propose a general learning-based framework for 3-D 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 3-D Statistical Facial feAture Model (SFAM), which learns both the global variations in 3-D face morphology and the local variations around each 3-D face landmark in terms of texture and geometry. To train the model, we manually labeled the target landmarks for each aligned frontal 3-D 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 3-D 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 3-D partial face instances that consist of local face regions with texture and shape (structured by their global 3-D 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 3-D face model [15] and the CLM [14] as we propose to learn, at the same time, the global variations of 3-D face morphology and the local ones in terms of texture and shape around each landmark. Fitting the SFAM on

TABLE I
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

a probe facial scan is accomplished by maximum *a posteriori* (MAP) probability. The fitted morphology instance delivers the locations of targeted landmarks. Using 3-D training faces with expressions, the SFAM has the ability to learn expression variations and generate instances with the learned variations so as to increase the *a posteriori* probability in fitting faces with expression. Furthermore, we propose to use a k-nearest neighbor (k-NN) 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 the 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 3-D facial scans.

The main contributions of this paper are the following.

1) We build an SFAM 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 partially occluded faces.

The rest of this paper is organized as follows. In Section II, our statistical model SFAM is introduced. In Section III, the objective function that combines the local shape and texture properties and the fitting algorithm are described. Section IV addresses 3-D face partial occlusion. Experimental results are discussed in Section V, while Section VI concludes this paper. Table I presents a summary of the different symbols used in this paper.

II. SFAM

Three-dimensional facial data acquired by the current 3-D imaging systems are usually noisy and may contain holes and spikes. Hence, we first preprocess all the 3-D facial scans to remove noise. Head pose and scale variations are normalized by alignment and remeshing (see Section II-A). Then, we model the variations in 3-D configurations of landmarks and their local variations in terms of texture and shape around each landmark (see Section II-B). New partial 3-D face instances can be synthesized from the learned model (see Section II-C).

A. Preprocessing the Training Facial Data

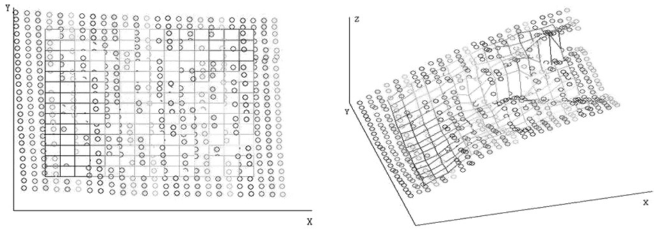
To remove the noise (e.g., spikes and holes) and enhance the quality of 3-D facial scans, we perform the following operations.

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 3-D 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 3-D 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 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 have been compensated after alignment, the SFAM can be learned 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 3-D face. This is the major difference between the proposed approach and most of the existing 3-D 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 data set, 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 data sets: FRGC, BU-3-DFE, and Bosphorus data sets. Most landmarks out of 15 (as illustrated in Fig. 5) on the FRGC data set were selected from the rigid part of the face as they were subsequently used for 3-D face recognition. On the other hand, landmarks on the BU-3-DFE and the Bosphorus data sets (as illustrated in Figs. 6 and 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



(a) (b)

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

to face scale. Second, a uniform grid is associated with each landmark. As illustrated in Fig. 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.

B. Modeling the Configurational Relationships and Local Shape and Texture Features of the Landmarks

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

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

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

We further generate the 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

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

PCA is then applied to the three vector sets $\{\boldsymbol{s_k}\}$, $\{\boldsymbol{g_k}\}$, and $\{\boldsymbol{z_k}\}$, extracted from the training 3-D facial data (k denotes the kth 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:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{P_s b_s} \tag{3}$$

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

where $\bar{\boldsymbol{s}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{z}}$ are the mean landmark configuration, the mean intensity, and the mean range value, respectively, while $\boldsymbol{P_s}$, $\boldsymbol{P_g}$, and $\boldsymbol{P_z}$ are the three sets of modes of configuration, intensity, and depth variation, respectively. The terms $\boldsymbol{b_s}$, $\boldsymbol{b_g}$, and $\boldsymbol{b_z}$ 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 $\boldsymbol{\sigma_q}$.

C. Synthesizing Instances From a New Face

Given the parameters $\boldsymbol{b_s}$, a configuration instance can be generated using (3). Then, given a new facial scan, the set of scan points closest to the configuration instance is 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 (2). Then, $\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ are estimated as follows:

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

$\boldsymbol{b_g}$ and $\boldsymbol{b_z}$ 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 (4). The landmarks, along with their local texture and local shape instances, compose a partial face instance.

III. Localizing Landmarks

The SFAM-based landmark localization procedure consists of MAP probability of landmark configuration, given a 3-D facial scan to be landmarked, and leads to optimizing an objective function. In Section III-A, we present the objective function to be optimized, and in Section III-B, we introduce the fitting algorithm for localizing landmarks. We then discuss our assumptions in Section III-C. AQ14

A. Objective Function and MAP

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

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

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 (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

$$p(\boldsymbol{s}|T, R, \psi) \propto \prod_{i=1}^{N} e^{-(\alpha\eta_i)} \prod_{i=1}^{N} e^{-(\beta\gamma_i)} \prod_{j=1}^{K} e^{\frac{-b_j^2}{\lambda_j}}$$

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

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 (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 paper, 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 (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):

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

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

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

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 $\sum_{j=1}^{K}(b_j^2/\lambda_j)$, $\sum_{i=1}^{N} F_{zi}$, and $\sum_{j=1}^{K}(b_j^2/\lambda_j)$, respectively, during the offline training.

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

B. Fitting Algorithm

Landmarking a 3-D facial scan consists of fitting the SFAM ψ while maximizing the objective function (8). First, the 3-D facial scan is preprocessed as described in Section II-A, including spike removal, hole filling, and head pose normalization. The occlusion algorithm, introduced in Section IV, 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.

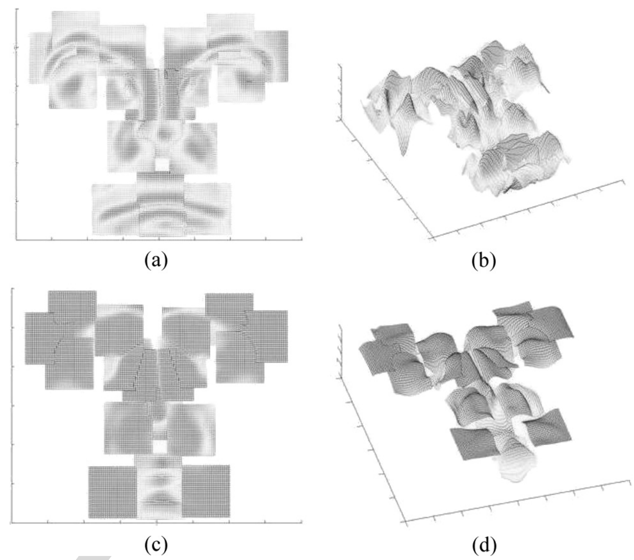


Fig. 2. Depiction of the correlation meshes from the frontal and side views. These meshes capture the similarity between instances and local facial regions in both texture and shape representations. The red color corresponds to large correlation values while blue corresponds to small correlation values. 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).

Algorithm 1 SFAM Fitting

Input: A 3-D 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 $\mathcal{S}$.

2. Synthesize texture and shape instances $\hat{G}$, $\hat{Z}$ as described in Section II-C using $\mathcal{S}$.

3. Normalize local regions around points $\mathcal{S}$ within a neighborhood large enough to cover the potential landmark locations as in Section II-A, creating a set of local mesh $\mathcal{G}$, $\mathcal{Z}$.

4. Compute correlation meshes on both texture and geometry representations (see Fig. 2) by correlating $\hat{G}$, $\hat{Z}$ with G, Z, respectively, which are different parts of $\mathcal{G}$, $\mathcal{Z}$ sampled by a sliding window (size of 15×15) on local regions (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 landmark configuration defined by the control parameters $\boldsymbol{b_s}$.

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

The optimization process in steps one and five of the algorithm is processed by the Nelder–Mead 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 nonface meshes lead the optimization to converge to an unpredictable point far from the desired minimum.

C. Discussion

To deduce (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 3-D face landmarking is actually a Markov random field (MRF) which consists of assigning a label from a set of labels $\mathcal{L}$ to each vertex of a 3-D facial scan. The set $\mathcal{L}$ encompasses all targeted landmarks (e.g., nose tip and 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–Mead 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].

IV. Occlusion Detection and Classification

Facial data analysis in the presence of partial occlusions (caused by a variety of factors such as hair, glasses, mustaches, and scarf) is a difficult problem. In 3-D facial landmarking, only occlusions which may occur in local regions around landmarks are of interest. Thus, in this paper, 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.

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 $\bar{\boldsymbol{s}}(3)$ are first computed. Then, grids (50×50) are used to remesh local regions around these points for range values (see Section II-A). 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 the 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:

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

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

where $\boldsymbol{P}_{\boldsymbol{z},\boldsymbol{i}}$ 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 the 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 II-C. 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 (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 3-D 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 data set which encompasses partially occluded 3-D 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 (8)].

V. Experimental Results

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

A. Data Sets and Experimental Setup

The FRGC data set 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 3-D facial scans having uncorrected correspondence between the range and texture maps. These scans were not used in our experiment. FRGC v2 contains 4007 facial scans from 466 persons. These 3-D facial scans were captured under different illumination conditions and contain various facial expressions (such as happiness or surprise).

The BU-3-DFE database contains data from 100 subjects [36]. Each subject performed a neutral expression and six universal expressions in front of a 3-D 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 1300 facial scans in total.

The Bosphorus data set contains 3396 facial scans from 104 subjects [37]. This data set contains not only the six universal facial expressions but also 3-D scans under realistic occlusions (e.g., glasses, hands around the mouth, and eye rubbing).

TABLE II
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 %

Moreover, the data set includes many male subjects that have moustache and beard.

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

B. Occlusion Classification Results

The proposed algorithm for occlusion detection was applied to 3-D scans from the Bosphorus data set. 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 a hand near the ocular region in addition to unoccluded 3-D scans. We experimentally set k to five in the k-NN classifier and performed a two-fold cross-validation. The confusion matrix is provided in Table II. An average classification accuracy up to 93.8% is achieved, which appears to be sufficient for the subsequent landmarking task.

C. Results on SFAM

We used 452 scans from the FRGC v1 data set 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-3-DFE data set and the first 32 subjects in the Bosphorus data set to build the SFAM-2 and SFAM-3, respectively. For every subject, 13 scans were used for training in the case of the BU-3-DFE data set (a neutral scan and the two scans for each of the six universal expressions at the intensity levels three and four), and seven scans in the case of the Bosphorus data set (a neutral scan and a scan for each of the six universal expressions). Fig. 3 illustrates the SFAM-3 learned from the Bosphorus data set containing the first mode of configuration, local texture, and local shape for variances $3 \pm \boldsymbol{\sigma}$.

D. Results on Landmarking

Using the learned statistical models, the fitting algorithm for 3-D 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.

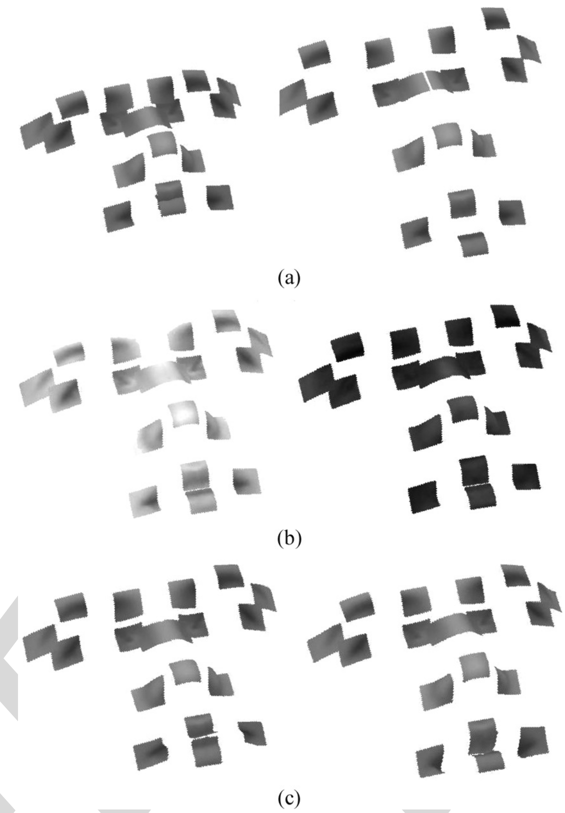


Fig. 3. SFAM learned from the Bosphorus data set. (a) First landmark configuration mode explains variations in terms of the face size and expression. (b) First texture mode explains skin color variations. (c) First range mode explains surface geometry variations, mainly in the nose and mouth regions.

Using the SFAM-1, the fitting algorithm was first applied on the remaining FRGC v1 data sets (i.e., 462 scans from subjects different from those in training). We then tested the algorithm on 1500 facial scans (randomly selected from the FRGC v2 data set) which contain illumination variations and facial expressions. Fig. 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 III summarizes the mean, the 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 deviation 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. Fig. 5 illustrates selected landmark localization results from the first two experiments.

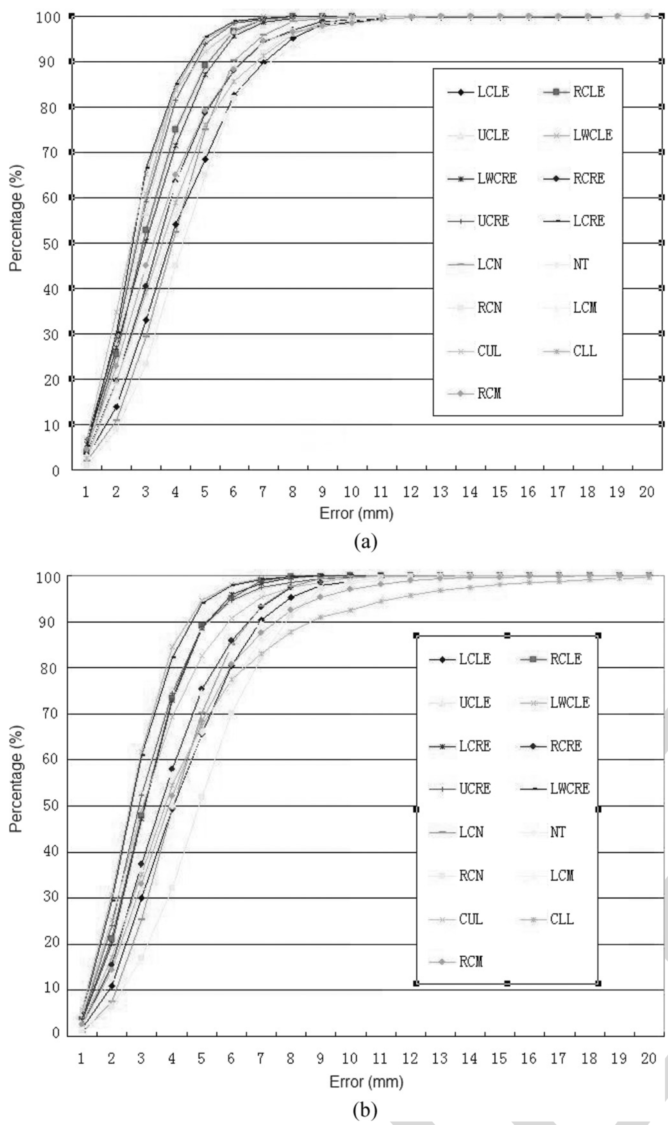


Fig. 4. Cumulative error distribution of the error for the 15 landmarks using (a) FRGC v1 and (b) FRGC v2. The symbols used are the following: LCLE—left corner of left eye, RCLE—right corner of left eye, UCLE—upper corner of left eye, LWCLE—lower corner of left eye, LCRE—left corner of right eye, RCRE—right 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
AQ15 upper lip, CLL—center of lower lip, and RCM—right corner of mouth.

The third experiment was carried out on the BU-3-DFE data set. 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, 1157 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. Fig. 6 illustrates several localization examples having facial expressions. Fig. 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, and nose corner). Mouth corners and the middle of the 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 IV summarizes the mean error and the standard deviation of the proposed landmarking algorithm compared to the mean error of a PDM [21], which is trained with 150 face scans and tested on the remainder of the BU-3-DFE data set. 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.

TABLE III
COMPARISON OF MEAN ERROR AND STANDARD DEVIATION ASSOCIATED WITH EACH OF THE 15 LANDMARKS ON THE FRGC DATA SET

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)

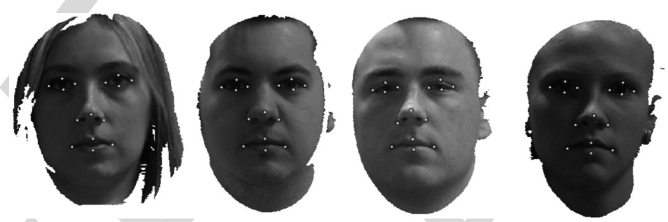


Fig. 5. Landmark localization examples from the FRGC data set.

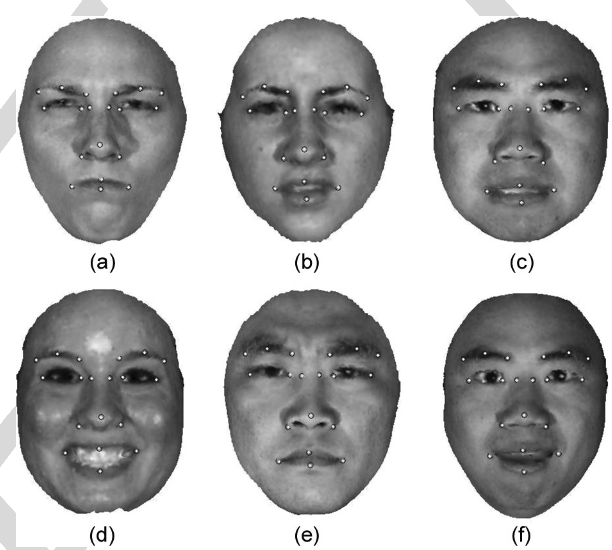


Fig. 6. Landmarking examples from the BU-3-DFE data set with expressions. (a) Anger. (b) Disgust. (c) Fear. (d) Happiness. (e) Sadness. (f) Surprise.

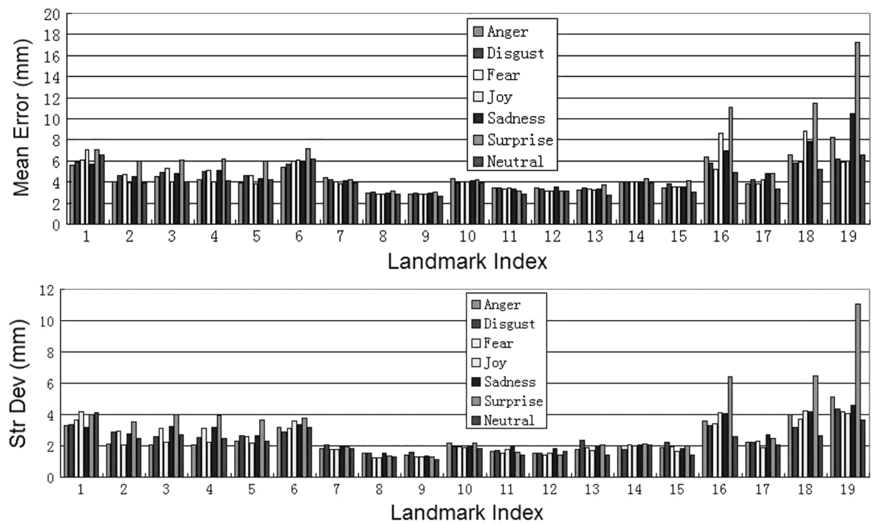


Fig. 7. Landmarking accuracy on different expressions with the BU-3-DFE data set. 1: Left corner of left eyebrow. 2: Middle of left eyebrow. 3: Right corner of left eyebrow. 4: Left corner of right eyebrow. 5: Middle of left eyebrow. 6: Right corner of right eyebrow. 7: Left corner of left eye. 8: Right 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.

TABLE IV
MEAN ERROR AND THE CORRESPONDING STANDARD DEVIATION (IN MILLIMETERS) OF THE 19 AUTOMATICALLY LOCALIZED LANDMARKS ON THE FACIAL SCANS FROM THE BU-3-DFE DATA SET (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				

TABLE V
MEAN ERROR AND THE CORRESPONDING STANDARD DEVIATION ASSOCIATED WITH EACH OF THE 19 AUTOMATICALLY LOCALIZED LANDMARKS ON THE FACIAL SCANS FROM THE BOSPHORUS DATA SET UNDER OCCLUSION

	Mean (Std) mm			Mean (Std) mm	
ID	I	II	ID	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)			

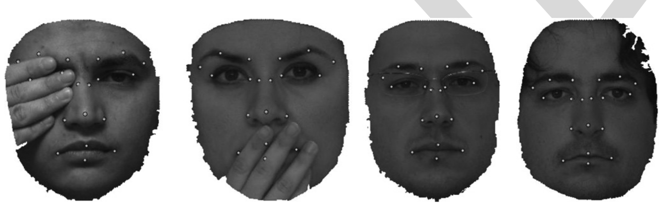


Fig. 8. Landmarking examples from the Bosphorus data set with occlusion. From left to right, faces are occluded in the eye region, in the mouth region, by glasses, and by hair.

The last experiment tested the fitting algorithm using the SFAM-3 to locate 19 landmarks on 3-D scans under occlusion from the Bosphorus data set. Fig. 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 data set. To evaluate the efficiency of our proposed occlusion classifier, the fitting algorithm was first tested with occlusion knowledge directly provided by the data set and, then, with occlusion knowledge from our occlusion detection and classification algorithm (see Table V). In both configurations, the mean errors ranged from 6 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 data set to the one provided by the proposed occlusion detection algorithm described in Section IV.

E. Discussion

We studied the influence of landmark configuration on the landmarking results (see Table VI). Three sets of landmarks, consisting of 5, 9, and 15 landmarks, respectively, were tested on 100 facial scans randomly selected from the FRGC v1 data set. The subjects depicted in these scans were different from the subjects used for training the SFAM, which is the SFAM-1 described in Section V-C. From Table VI, 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 paper based on their global configurational

TABLE VI
INFLUENCE OF LANDMARK CONFIGURATION ON MEAN ERRORS (IN MILLIMETERS)

	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)

(a) (b) (c) (d)

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.

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 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. Over 99% of the cases converge within 2000 iterations or 422 s on average.

Fig. 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 the 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 an 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 Fig. 5 on the same 10 facial scans randomly selected from FRGC v1. We then computed the mean error and the corresponding 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 data set.

Compared to previous 3-D face landmarking algorithms [7], [8], [10], [17], [19], [21], [31], [32], our SFAM-based algorithm is a general data-driven 3-D 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.

VI. CONCLUSION

In this paper, we have presented a general learning-based framework for 3-D 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 *a 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 the occlusion probability around each landmark can be estimated. Based on this evidence, we have also introduced a 3-D facial occlusion detection and classification algorithm which exhibited a 93.8% classification accuracy on the Bosphorus data set. This detection is based on local shape similarity between local ranges of an input 3-D facial scan and the instances synthesized from the SFAM. The effectiveness of our technique was supported by the experiments on the FRGC data set (v1 and v2), BU-3-DFE containing expressions, and the Bosphorus data set 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.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments which greatly contributed to the improvement of this paper.

References

[1] A. A. Salah, H. Cinar, L. Akarun, and B. Sankur, "Robust facial landmarking for registration," *Ann. Telecommun.*, 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. 5th IEEE Int. Conf. Autom. Face Gesture Recog.*, 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," *Inf. Sci.*, vol. 158, pp. 117–130, Jan. 2004.

[4] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[5] C. Tu and J. J. Lien, "Automatic location of facial feature points and synthesis of facial sketches using direct combined model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1158–1169, Aug. 2010.

[6] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D +2D face recognition," *Comput. Vis. Image Understand.*, 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. Int. Conf. Biometrics: Theory, Appl., Syst.*, Crystal City, VA, Sep. 27–29, 2007, pp. 1–6.

[8] T. Faltemier, K. Bowyer, and P. Flynn, "Rotated profile signatures for robust 3D feature detection," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recog.*, Amsterdam, The Netherlands, Sep. 17–19, 2008, pp. 1–7.

[9] L. Farkas, *Anthropometry of the Head and Face*, L. G. Farkas, Ed., 2nd ed. New York: Raven, 1994.

[10] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1487–1494, Oct. 2006.

[11] D. Colbry, G. Stockman, and A. Jain, "Detection of anchor points for 3D face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, San Diego, CA, Jun. 20–25, 2005, pp. 118–124.

[12] T. F. Cootes, G. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[14] D. Cristinacce and T. F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognit.*, 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. 26th Annu. Conf. Comput. Graph. Interactive Techn.*, Los Angeles, CA, Aug. 8–13, 1999, pp. 187–194.

[16] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, Jan. 1965.

[17] S. Jahanbin, A. C. Bovik, and H. Choi, "Automated facial feature detection from portrait and range images," in *Proc. IEEE Southwest Symp. Image Anal. 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," *Int. J. Comput. Vis.*, vol. 13, no. 2, pp. 119–152, Oct. 1994.

[19] H. Dibeklioglu, A. A. Salah, and L. Akarun, "3D facial landmarking under expression, pose, and occlusion variations," in *Proc. IEEE Int. Conf. Biometrics: Theory, Appl. Syst.*, Arlington, VA, Sep. 29–Oct. 1, 2008, pp. 1–6.

[20] X. Lu, A. Jain, and D. Colbry, "Matching 2.5D face scans to 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 31–43, Jan. 2006.

[21] P. Nair and A. Cavallaro, "3D face detection, landmark localization, and registration using a point distribution model," *IEEE Trans. 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 Recognit.*, vol. 39, no. 3, pp. 444–455, Mar. 2006.

[23] S. Jahanbin, H. Choi, R. Jahanbin, and A. C. Bovik, "Automated facial feature detection and face recognition using Gabor features on range and portrait images," in *Proc. Int. Conf. Image Process.*, 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. Int. Conf. Adv. Intell. Comput. Theories Appl.*, 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 Recognit. Lett.*, vol. 23, no. 10, pp. 1191–1202, Aug. 2002.

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

[27] C. Boehnen and T. Russ, "A fast multi-modal approach to facial feature detection," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 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 Trans. Pattern Anal. Mach. Intell.*, 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 Comput. Vis. Pattern Recog.*, San Diego, CA, Jun. 20–25, 2005, p. 168.

[30] X. Lu and A. K. Jain, "Multimodal facial feature extraction for automatic 3D face recognition," Michigan State Univ., East Lansing, MI, Tech. Rep. MSU-CSE-05-22, 2005.

[31] P. Szeptycki, M. Ardabilian, and L. Chen, "A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking," in *Proc. 3rd Int. Conf. Biometrics: Theory, Appl. Syst.*, Washington, DC, 2009, pp. 1–6.

[32] X. Lu and A. Jain, "Automatic feature extraction for multiview 3D face recognition," in *Proc. 7th Int. Conf. Autom. Face Gesture Recog.*, Southampton, U.K., 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. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 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. Int. Conf. Autom. Face Gesture Recog.*, Amsterdam, The Netherlands, Sep. 17–19, 2008, pp. 1–7.

[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 Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, San Diego, CA, Jan. 20–25, 2005, vol. 1, 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. 7th Int. Conf. Autom. Face Gesture Recog.*, Southampton, U.K., 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. 1st COST 2101 Workshop Biometrics Identity Manage.*, 2008, pp. 47–56.

[38] P. Soille, *Morphological Image Analysis: Principles and Applications*. New York: Springer-Verlag, 1999.

[39] S. Z. Li, *Markov Random Field Modelling in Image Analysis*, 3rd ed. New York: Springer-Verlag, 2009.

[40] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Hoboken, NJ: Wiley, 2001.

[41] S. Singer and S. Singer, "Efficient implementation of the Nelder–Mead search algorithm," *Appl. Numer. Anal. Comput. Math.*, vol. 1, no. 2, pp. 524–534, Dec. 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 3D Object Retrieval*, Munich, Germany, Mar. 30–Apr. 3, 2009, pp. 37–44.

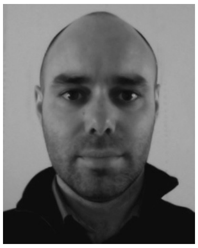
[43] P. Perakis, G. Passalis, T. Theoharis, G. Toderici, and I. A. Kakadiaris, "Partial matching of interpose 3D facial data for face recognition," in *Proc. 3rd IEEE Int. Conf. Biometrics: Theory, Appl. Syst.*, Arlington, VA, Sep. 28–30, 2009, pp. 1–8.

[44] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proc. IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[45] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

Xi Zhao (S'XX) received the B.Sc. and M.Sc. degrees (with honors) from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2003 and 2007, respectively, and the Ph.D. degree (with honors) in computer science from Ecole Centrale Lyon, Ecully, France, in 2010.

He is currently conducting research as a Postdoctoral Fellow at the Computational Biomedicine Laboratory, University of Houston, Houston, TX. His research interests include 3-D face analysis, statistical pattern analysis, and computer vision.



Emmanuel Dellandréa received the Master and Engineer degrees in computer science and the Ph.D. degree in computer science from Université de Tours, Tours, France, in 2000 and 2003, respectively.

In 2004, he joined Ecole Centrale Lyon, Ecully, France, as an 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 (M'XX) was awarded the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984. He obtained the M.S. and Ph.D. degrees in computer science from the University of Paris 6, Paris, France, in 1986 and 1989, respectively.

He first served as an Associate Professor at the Université de Technologie de Compiègne, Compiègne, France, and then joined Ecole Centrale de Lyon, Ecully, France, 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 the Head of the Department of Mathematics and Computer Science since 2007. He has taken out three 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 the European Union FP programme, French research funding bodies, and local government departments. He has directed more than 15 Ph.D. theses. His current research spans from 2-D/3-D face analysis and recognition and image and video analysis and categorization to affect analysis both in image audio and video.

Ioannis A. Kakadiaris (SM'XX) received the B.Sc. degree in physics from the University of Athens, Athens, Greece, the M.Sc. degree in computer science from Northeastern University, Boston, MA, and the Ph.D. degree from the University of Pennsylvania, Philadelphia.

He is a Hugh Roy and Lillie Cranz Cullen Professor of Computer Science, Electrical and Computer Engineering, and Biomedical Engineering at the University of Houston (UH), Houston, TX. He joined UH in August 1997 after a postdoctoral fellowship at the University of Pennsylvania. He is the founder of the Computational Biomedicine Laboratory (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, nonverbal 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 National Science Foundation 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.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

Note that your paper will incur overlength page charges of $175 per page. The page limit for regular papers is 12 pages, and the page limit for correspondence papers is 6 pages.

AQ1 = The sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ2 = "ANR" is defined as "French National Research Agency." Please check if appropriate, and correct if necessary.
AQ3 = "ANR-07-MDCO-009-02" and "ANR-07-SESU-004-03" were captured as grants. Please check if appropriate, and correct if necessary.
AQ4 = "CNRS" was expanded as "Centre National de la Recherche Scientifique." Please check if appropriate, and correct if necessary.
AQ5 = "LIRIS" was expanded as "Laboratoire d'InfoRmatique en Image et Systèmes d'information." Please check if appropriate, and correct if necessary.
AQ6 = "UMR" was expanded as "Unité Mixte de Recherche." Please check if appropriate, and correct if necessary.
AQ7 = "Boehnen *et al.*" was changed to "Boehnen and Russ." Please check if appropriate, and correct if necessary.
AQ8 = "Lu *et al.*" was changed to "Lu and Jain." Please check if appropriate, and correct if necessary.
AQ9 = "Nair *et al.*" was changed to "Nair and Cavallaro." Please check if appropriate, and correct if necessary.
AQ10 = "MAP" is defined as "maximum a posteriori." Please check if appropriate, and correct if necessary.
AQ11 = This sentence was rephrased. Please check if the original thought was retained, and correct if necessary.
AQ12 = All occurrences of "posteriori probability" were changed to "a posteriori probability." Please check if appropriate, and correct if necessary.
AQ13 = All occurrences of "kNN" and "K-NN" were captured as "k-NN" to refer to "k-nearest neighbor" and for consistency. Please check if appropriate, and correct if necessary.
AQ14 = This sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ15 = The caption was restructured for clarity of presentation. Please check if Please check if appropriate, and correct if necessary.
AQ16 = The sentence was rephrased for clarity. Please check if the original thought was retained, and correct if necessary.
AQ17 = Please provide the IEEE membership history of author Xi Zhao.
AQ18 = The sentence was rephrased. Please check if the original thought was retained, and correct if necessary.
AQ19 = Please provide the IEEE membership history of author Liming Chen.
AQ20 = "Master" was changed to "M.S." Please check if appropriate, and correct if necessary.
AQ21 = Please validate the address provided for the Université de Technologie de Compiègne.
AQ22 = Please validate the address provided for Ecole Centrale de Lyon.
AQ23 = "EU" was expanded as "European Union." Please check if appropriate, and correct if necessary.
AQ24 = Please provide the expanded form of "FP."
AQ25 = Please provide the IEEE membership history of author Ioannis A. Kakadiaris.
AQ26 = "NSF" was expanded as "National Science Foundation." Please check if appropriate, and correct if necessary.

END OF ALL QUERIES