A Statistical Method for 2-D Facial Landmarking

Hamdi Dibeklioğlu, *Student Member, IEEE*, Albert Ali Salah, *Member, IEEE*, and Theo Gevers, *Member, IEEE*

***Abstract*—Many facial-analysis approaches rely on robust and accurate automatic facial landmarking to correctly function. In this paper, we describe a statistical method for automatic facial-landmark localization. Our landmarking relies on a parsimonious mixture model of Gabor wavelet features, computed in coarse-to-fine fashion and complemented with a shape prior. We assess the accuracy and the robustness of the proposed approach in extensive cross-database conditions conducted on four face data sets (Face Recognition Grand Challenge, Cohn–Kanade, Bosphorus, and BioID). Our method has 99.33% accuracy on the Bosphorus database and 97.62% accuracy on the BioID database on the average, which improves the state of the art. We show that the method is not significantly affected by low-resolution images, small rotations, facial expressions, and natural occlusions such as beard and mustache. We further test the goodness of the landmarks in a facial expression recognition application and report landmarking-induced improvement over baseline on two separate databases for video-based expression recognition (Cohn–Kanade and BU-4DFE).**

***Index Terms*—Facial feature localization, facial landmarking, factor analysis, Gabor wavelet features, mixture models, shape prior, structural analysis.**

I. Introduction

AUTOMATIC facial landmarking is an important component for face registration, analysis, and recognition methods. The pipeline of a facial-analysis method starts with face detection and often proceeds by locating several fiducial points on detected faces, also called anchor points, or landmarks. The landmarks are used for aligning the faces, also called registration, which has a significant effect on the subsequent analysis. These include (most of the time) eye and eyebrow corners, centers of irises, the nose tip, mouth corners, and the tip of the chin. While a few landmarks are sufficient for registration prior to face recognition, a greater number of landmarks are usually required (typically between 20–60) for expression analysis.

Manuscript received September 17, 2010; revised April 08, 2011 and July 05, 2011; accepted July 11, 2011. Date of publication July 29, 2011; date of current version January 18, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Patrick Flynn.

H. Dibeklioğlu is with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, 1098 XH Amsterdam, The Netherlands (e-mail: h.dibeklioglu@uva.nl).

A. A. Salah is with the Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey (e-mail: salah@boun.edu.tr).

T. Gevers is with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, 1098 XH Amsterdam, The Netherlands, and also with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (e-mail: th.gevers@uva.nl).

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/TIP.2011.2163162

Comparative studies have demonstrated that improving facial registration can have a significant impact on a face analysis system [1]. This is particularly the case for expression analysis, where the configurations of facial landmarks are indicative of deformations caused by expressions. Subsequently, deformation analysis can reveal expression categories, provided that facial landmarks are accurately detected and contain sufficient information for the recognition of a particular facial expression. However, facial landmarking is a challenging problem under changing acquisition conditions, and statistical models can fail if the variance captured during training is not rich enough to generalize to new test settings.

One may argue that models that jointly learn shape and appearance, such as the active appearance model (AAM) and the active shape model (ASM), sidestep the need for landmarking. There are no theoretical constraints to necessitate a high number of landmark annotations for these models, as the number depends on the requirements of the task at hand, yet these approaches are traditionally trained with 50–60 annotated points for face analysis [2]. Milborrow and Nicolls have proposed an extended active shape model and have shown that the point-to-point error doubles on the average when the number of landmark annotations is around ten, instead of 50–60 [3]. While these approaches provide good results in many cases, training such a system is more difficult and costly than training a landmarker for this reason. On the other hand, methods that are based on landmark-specific heuristics often give good results on a particular data set, but their reliance on various assumptions (e.g., open eyes for contrast-based eye detection) makes them vulnerable.

The contribution of this paper is a generic statistical 2-D landmarking method that empirically performs better than methods reported so far in the literature. We assess the strengths and the weaknesses of the method in the most extensive experimental setting reported in a landmarking paper to date, and by making our protocols available online, we hope to make possible a more rigorous evaluation of landmarking methods in the future. Our setup includes a large variety of imaging conditions and separate assessment for low-resolution images, pose variations, different facial expressions, and natural occlusions such as mustache and beard.

Our method follows a coarse-to-fine strategy, which reduces the computational burden and also improves the accuracy by making the method resilient to resolution changes. Statistical methods perform better on coarse scales, where adjacent pixels have less correlation. The appearance of each landmark is modeled with statistical models of Gabor wavelet features with different scale and rotation parameters, which allows straightforward probabilistic interpretation. To integrate facial morphology and to constrain the search, we use a novel structural prior, based on the assumption that the input face is frontal. Rotations will cause both shape and appearance

problems. Subsequently, we assess our method with $+10^\circ$ and $+20^\circ$ rotations and show that including even a small amount of rotated faces in the training set will improve the methods' resilience to these. We also contrast our statistical method with several alternatives and demonstrate its superiority for the present setting.

This paper is structured as follows: Section II describes related work in landmarking, followed by Section III that describes our statistical landmark localization algorithm. In Section IV, we present the experimental results with discussion of different aspects of the approach, such as the effect of resolution, natural occlusions, and rotation. We also contrast and discuss several performance criteria. Section V describes an application of facial-expression classification that demonstrates the extent of improvement by automatic landmarking and also illustrates the shortcomings of the approach. We conclude in Section VI.

II. Related Work in Landmarking

Finding facial landmarks automatically is a difficult problem, which faces all hurdles of face recognition in a smaller scale, such as illumination and occlusion problems [4]. The constellation of facial landmarks is different for each face image. A part of the difference is due to the subjective morphology of the face, as different persons have different face shapes. Even for the same person, different images will have different configurations. Another part of this difference is due to camera angle and pose differences. There are also expression-based changes (of which some part may be attributable to emotion) and measurement noise to take into account.

The appearance of each landmark and the structural relationships between landmark points (i.e., configuration) can be both taken into account in locating landmarks automatically. However, both the appearance and the structure are changed under expression variations and in different ways. If the application involves video input, it is simpler to solve the problem of landmarking on the neutral face and then track each landmark while the face is deformed under the influence of an expression. Tracking is computationally cheaper than a search for exact landmark locations, as the latter usually requires a lot of features for robust detection across different conditions.

The detection of facial landmarks is frequently performed with landmark-specific heuristics that are experimentally validated on a particular data set. For instance, vertical projection histograms of intensity values can be used to localize the eye and mouth regions [9]. Moreover, the contrast differences in the eye region were employed to train classifiers for eye detection [7]. Heuristic approaches do not need extensive training sets, but they require individual treatment of each landmark, which can mean excessive engineering for a rich set of landmarks.

The second approach is the joint optimization of structural relationships between landmark locations and local feature constraints, which are frequently conceived as distances to feature templates [10], [11]. The landmark locations are modeled with graph nodes (e.g., the elastic bunch graph), where the edges characterize pairwise distances. Two popular graph-based approaches are the AAM and the ASM [12]. The ASM models textures of small neighborhoods around landmarks and iteratively minimizes the differences between landmark points and their corresponding models. The AAM typically looks at the texture within the convex hull of landmarks, synthesizes a face image from a joint appearance and shape model, and seeks to maximize similarity to the target face iteratively.

A large number of facial landmarks (typically 50–60) are used for graph-based methods. Tong *et al.* [13] proposed a semisupervised deformation procedure to locate large numbers of facial points with the help of a few annotated images. Fewer and sparsely distributed landmarks produce a smaller number of structural constraints. Our approach does not rely on a structural model and can be employed to locate a few landmarks.

Recently, Cristinacce and Cootes proposed the constrained local model (CLM) approach, which is similar to an AAM but uses a set of local feature templates instead of modeling the appearance of the whole face [2]. Coarse alignment is via face detection, and modeled templates are matched to the image through a shape-constrained search that uses both appearance and shape information. The CLM is shown to perform better than the AAM for the facial-landmarking task. In [3], Milborrow and Nicolls propose a number of simple extensions to the ASM approach of Cootes *et al.* [14] and report further accuracy improvement over the CLM approach [2].

A number of approaches use larger sets of landmarks to track faces. For instance, Gu and Kanade [15] propose a generative shape regularization model, which is applied on automatically initialized key points to localize 83 points. In [16], Zhao *et al.* use Gabor features to align 13 control points on the face, and further 83 points are generated by constrained profile and flexible shape models. Liu proposes an adaptive algorithm that uses a generic AAM and a subject-specific appearance model together for detecting 72 points [17]. During fitting, the subject-specific model is updated by using the generic model to track the next frames. For these approaches, the precise locations of landmarks are less important than the coverage of the face. The neighboring points often have similar nondiscriminating visual features. Thus, the additional landmarks are not very beneficial for precise computations such as expression analysis, and their detection may have a large variance. It is also very difficult to verify landmarking under these conditions, as the ground-truth annotation for such large numbers of landmarks is costly to obtain and is not reliable at all. Nonetheless, these approaches can be very useful for tracking the facial boundary. In this paper, however, our aim is to find a few well-defined landmarks with as much precision as possible.

The independent detection of landmark points provides robustness against missing or occluded landmarks. When the image contains poor or wrong feature information (e.g., sunglasses masking the eyes or beard masking the mouth corners), the joint estimation of landmarks (e.g., in AAM approaches) will be problematic, unless the occlusion is detected before optimization, and the occluded landmarks are prevented from contributing to the appearance term. Yu *et al.* report about a 20% accuracy decrease for the 10% image occlusion and about 40% accuracy decrease for the 40% image occlusion under the basic active appearance model [18]. Their solution is to presegment the image to detect occlusion and to adapt the error terms corresponding to appearance parameters accordingly. A

second point is that a large number of annotations (i.e., 40–50 landmarks per face) are usually used during the training stage of the ASM and its derivatives. This kind of ground truth is not available in any of the major databases. Because of all these reasons, [4] and [8] focus on the independent detection of landmark points, and we do the same in this paper.

In [8], Vukadinovic and Pantic propose a method that uses Gabor-feature-based boosted classifiers. In this approach, the detected face is divided into regions of interest (ROI). Then, individual GentleBoost templates are used to detect landmarks within the relevant ROI independently. Modeled templates are based on grayscale texture information and Gabor wavelet features. The ROI sizes are heuristically determined, and they need to be kept large enough to deal with expressive faces. Another independent feature-point detection algorithm has been proposed in [4], which follows a coarse-to-fine strategy. The statistical modeling of Gabor wavelet features on the coarse scale is complemented with a structural analysis step and a fine-tuning step. In [4], the search for each landmark is conducted over the whole image, whereas face detection can greatly constrain the search through a shape prior. Unlike the discrete segmentation of ROI-based methods, a shape prior would provide continuous probability values for the search area. Recently, Valstar *et al.* proposed a system based on the support vector regression of Haar-like features, where the search space is constrained by Markov random fields [5]. In this paper, we introduce a probabilistic shape prior that is simple and fast to compute, and show that, through such a prior, contrasted landmarking methods become faster and more accurate.

The statistical model we employ in our approach is a mixture of factor analyzers, which is similar to a mixture of Gaussians. An earlier approach that uses the Gaussian mixture modeling of Gabor features is proposed by Hamouz *et al.*, where modeled facial features are combined for face detection but not for precise landmarking [19]. A related approach is taken in [6], which describes a weighted vector concentration scheme combined with models of histogram-of-oriented-gradient features.

The approach presented in this paper builds on [4], which was the first paper to explore incremental statistical modeling of Gabor features. This paper improves the approach presented in [4] by adapting an efficient pyramid representation, a structural prior, and uniform features in each scale. The experimental setup is much more extensive, as we report cross-database results, measure the influence of adverse conditions, report results with 22 landmarks (instead of seven), report an application to expression recognition, and compare the approach with many results from the literature.

Table I shows the test configurations of some recent landmarking approaches in the literature. Unfortunately, there is no commonly adopted protocol for evaluating and comparing landmarking methods. Therefore, we make the complete experimental protocol (the training, validation, and test partitions) for each result reported in this paper available.

TABLE I
RECENT FACIAL LANDMARKING METHODS IN THE LITERATURE

Reference	Number of Landmarks	Test Database
Valstar *et al.* [5] (2010)	22	FERET+MMI, BioID
Kozakaya *et al.* [6] (2009)	14	FERET
Milborrow & Nicolls [3] (2008)	17	BioID
Arca *et al.* [7] (2006)	16	XM2VTS, UniMiDb
Cristinacce & Cootes [2] (2006)	17	BioID, XM2VTS
Vukadinovic & Pantic [8] (2005)	20	Cohn-Kanade

III. FACIAL LANDMARKING ALGORITHM

The proposed method is a coarse-to-fine strategy for the localization of facial landmarks. The search for the points uses Gabor wavelet features with different scale and rotation parameters. A structural prior is used to integrate facial morphology. Including shape information speeds up the system and increases the accuracy.

A. Facial Model and Features

Our first assumption is that the face area is detected. We use the Viola–Jones face detection algorithm [20], which requires the face image to be roughly frontal (i.e., rotations up to $+20^\circ$ are acceptable). We assess the effect of rotations explicitly in Section IV-A. A histogram equalization is applied to the face image for damping the illumination effects. Face detection and illumination compensation are standard steps in facial landmarking. To reduce the computational complexity of the search, we prepare a three-level image pyramid from the cropped high-resolution face images. The pyramid consists of 160×224, 80×112, and 40×56 pixel images. The expected locations for each landmark is learned with respect to the face boundary from a training set of manually annotated images.

Since coarse-level images have lower pixel-to-pixel correlation, they are much more suited for statistical modeling [4]. For instance, the nose tip in the full-resolution image is a large area with almost identical pixel values. If we include enough pixels to reach a sufficiently discriminative feature vector, the dimensionality will be too large. This further complicates the training of statistical models for obvious reasons. This is the main reason for following a coarse-to-fine strategy. The first-level search is performed on the 40×56 pixel image, and the search area is constrained by the landmark occurrences in the training set. Afterward, the detected facial feature point and its immediate neighbors are passed on to the next stage. Only 6×6 pixels are processed in each of the second and third tiers, as shown in Fig. 1. In the coarse-to-fine architecture, the training time of the system and the runtime memory requirement are increased because each level requires separately learned features, but this approach dramatically reduces the time complexity of the landmark search. The three-level search is approximately 16 times faster than using the one-shot detection on the resolution of the third level. Moreover, experimental results (not shown here) demonstrate that the three-level strategy localizes the landmarks more accurately (5%–10% on average) than the one-shot detection.

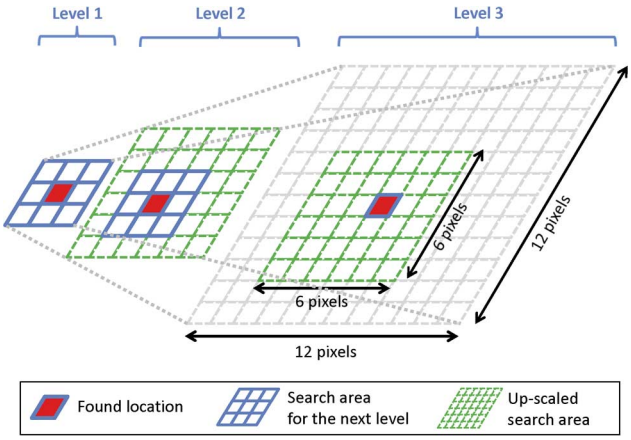


Fig. 1. Coarse-to-fine search for landmark detection in three levels.

Within search windows, independent Gabor wavelet features are extracted by convolving the candidate image patches with Gabor kernels of different orientation and frequency, i.e.,

$$\begin{aligned}\Psi_j(\vec{x}) &= \frac{\vec{k}_j\vec{k}_j^T}{\sigma^2}e^{\left(-\frac{\vec{k}_j\vec{k}_j^T\vec{x}\vec{x}^T}{2\sigma^2}\right)}\left[e^{(i\vec{k}_j\vec{x}^T)}-e^{\left(-\frac{\sigma^2}{2}\right)}\right]\\ \vec{k}_j &= (\vec{k}_{jx},\vec{k}_{jy})\\ &= (k_v\cos\phi_w, k_v\sin\phi_w)\\ k_v &= 2^{-\frac{v+2}{2}}\pi\\ \phi_w &= w\frac{\pi}{8}.\end{aligned} \tag{1}$$

In (1), $\vec{x} = (x, y)$ is a candidate landmark location, (w, v) defines the orientation and scale parameters of the Gabor kernel, and j is an index for different kernels. The standard deviation of the Gaussian function (σ) is 2π. The first factor in the Gabor kernel represents the Gaussian envelope, and the second factor represents the complex sinusoidal (carrier) function. Term $e^{-\sigma^2/2}$ in the square brackets compensates for the DC value.

Feature patches are extracted from around a 7×7 neighborhood of each landmark candidate. Consequently, 49-D feature vectors for each orientation and scale are obtained. These are extracted in eight orientations, i.e., $w \in \{0, 1, 2, 3, 4, 5, 6, 7\}$, and at three different scales, i.e., $v \in \{0, 2, 4\}$. Since we use a generative method, the training uses only the positive samples of each landmark class, obtained from ground truth. The features are minimum–maximum normalized to the [0, 1] range before statistical modeling. The search window constraint we introduce greatly reduces the area over which we compute wavelet features. Furthermore, this part of the system lends itself easily to parallel computation.

B. Statistical Feature Modeling

We model the statistical distribution of the extracted features with an incremental mixtures of factor analyzers (IMoFA) algorithm that places a number of Gaussian distributions with arbitrary covariance on the data [21]. This algorithm relies on a factor analysis formulation of the high-dimensional covariance matrix of each component in the mixture distribution, thereby adapting model complexity to the data locality. Complex models have more free parameters and consequently need a large number of training data to appropriately generalize. The IMoFA algorithm is favorable as it automatically finds a tradeoff between accuracy and complexity. It also responds to increases in the training data by an increase in complexity, instead of a tighter fit on the training set, which means there is less diminishing returns for increasing data volume and less overfitting due to model complexity.

A mixture of factor analyzers is, in essence, a mixture of Gaussians where the data are assumed to be generated in a lower dimensional manifold. For each component of the mixture, the $(d \times d)$ covariance matrix Σ is generated by a $(d \times p)$ dimensional factor matrix Λ and a $(d \times d)$ diagonal matrix Ψ, i.e.,

$$\Sigma_j = \Lambda_j\Lambda_j^T + \Psi. \tag{2}$$

Ψ is called the uniqueness, and it stands for the independent variance due to each dimension. With a small number of factors (represented with p), the model will have a much smaller number of parameters than a full Gaussian, although the covariances are modeled. In the mixture model, each distribution is potentially multimodal as we fit an arbitrary number of factor analysis components to each feature set.

The IMoFA algorithm adds components and factors to the mixture one by one, while monitoring a separate validation set for likelihood changes. Given a training set, the maximum likelihood estimates of the model parameters are calculated using the expectation-maximization (EM) algorithm, which simultaneously places the components in the input space and also finds the factors in each component, performing dimensionality reduction in each component [22]. At each step of the algorithm, the EM is interrupted to add a component to the mixture or a factor to an existing component in a greedy fashion.

Component addition is performed by selecting the component that looks least unimodal and splitting it along its principal axis. We look at a multivariate mixture kurtosis measure to decide which component to split. We compute the sample kurtosis $b_{2,d}^j$ for component j as

$$b_{2,d}^j = \frac{1}{\sum_{l=1}^N h_j^l}\sum_{t=1}^N h_j^t\left[(\boldsymbol{x}^t-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}^t-\boldsymbol{\mu}_j)\right]^2 \tag{3}$$

where $h_j^t \equiv E[\mathcal{G}_j|\boldsymbol{x}^t]$ and

$$\gamma_j = \left\{b_{2,d}^j - d(d+2)\right\}\left[\frac{8d(d+2)}{\sum_{t=1}^N h_j^t}\right]^{-\frac{1}{2}}. \tag{4}$$

The component with greatest γ_j is the one that looks least like a unimodal multivariate normal and is selected for splitting.

Factor addition is performed by finding a good initial point for a new factor and by concatenating it to the factor loading matrix Λ. This new factor is obtained by first projecting the data points $\mathcal{X}_j$ under component j (i.e., points for which the posterior probability of the component is largest) to the low-dimensional space described by Λ_j and by then taking the inverse projection to recover feature points $\mathcal{X}_j'$ in the original d-dimensional space. The new factor is then selected as the one that minimizes the compression error $\sum\|\mathcal{X}_j - \mathcal{X}_j'\|$ in the least square sense. At each

iteration, one action is performed (splitting or factor addition), and EM is run until convergence. A separate validation set is used to control the model complexity. This set is designated as *validation* in Section IV.

With this approach, the number of parameters is automatically adapted to the complexity of the feature space. The incremental approach is particularly feasible when the dimensionality is high, as opposed to starting with a large number of components and eliminating components one by one [23]. In the popular unsupervised learning approach proposed by Figueiredo and Jain, the resulting mixture model is composed of components with diagonal or full covariance matrices [23]. The IMoFA allows the exploration of many models that are in between these extremes in complexity, which leads to improved generalization capabilities. Our experiments in Section IV show that using IMoFA results in better models compared with Gaussian mixture models (GMMs), as proposed in [23].

C. Shape Prior

In this paper, we also introduce a shape prior to give weight to expected locations of each landmark. We categorize the facial feature points into two groups, i.e., as *stable* (L^s) and *unstable* (L^u) landmarks, as shown in Fig. 2. The stable landmarks are those that are relatively stable under expression- and speech-induced movement. The eye corners, nose tip, and nostril landmarks are stable. The unstable landmarks are eyebrow corners, eye pupils, upper nose saddles, the tip of the chin, and the points on the mouth. The instability of upper nose saddles is partly due expressions and partly due lack of discriminative appearance features, which makes their labeling noisy. The landmark estimation starts by locating the landmarks in the stable set and proceeds by using those to further constrain the landmarks in the unstable set. A prior distribution $p(x, y|M_l)$ is learned for each landmark l, where M_l^i denotes the particular landmark model. x and y can be defined as

$$[x \quad y] = \begin{cases} [x \quad y], & \text{if } l \in L^s \\ [x \quad y] \cdot R + [t_x \quad t_y], & \text{if } l \in L^u \end{cases}$$
$$R = \begin{bmatrix} s \cdot \cos(\Theta) & -s \cdot \sin(\Theta) \\ s \cdot \sin(\Theta) & s \cdot \cos(\Theta) \end{bmatrix} \tag{5}$$

where x and y denotes the relative coordinates of the landmark within the cropped face area. For unstable landmarks, transformation parameters (scale s, rotation Θ, and translations t_x and t_y) are estimated by aligning the stable landmarks of the related sample to the mean coordinates of the corresponding landmark learned from the training set with the Procrustes analysis [24]. Then, the shape prior for a particular landmark l is approximated with a multivariate GMM as follows:

$$S_\alpha(c) = \sum_{i=1}^{k} \pi_i \mathcal{N}\left((\mu_i - c), \Sigma_i \times 2^\alpha\right), \quad \sum_{i=1}^{k} \pi_i = 1 \tag{6}$$

where α is an amplification parameter to control the impact of the prior and c denotes the normalized location of the landmark. The parameters of the mixture (number of components k, prior π_i, mean μ_i, and covariance Σ_i), as well as an appropriate value (a positive integer smaller than 5) for α, are learned on the training set, using the validation set to control generalization. We have omitted subscripts l from (6); one such prior is learned for each landmark. The number of components for the GMM is determined using the method proposed in [23]. This algorithm uses a minimum-description-length-based criterion to assess converged models with different number of full-covariance components (all possibilities from one to seven components) and selects the best.

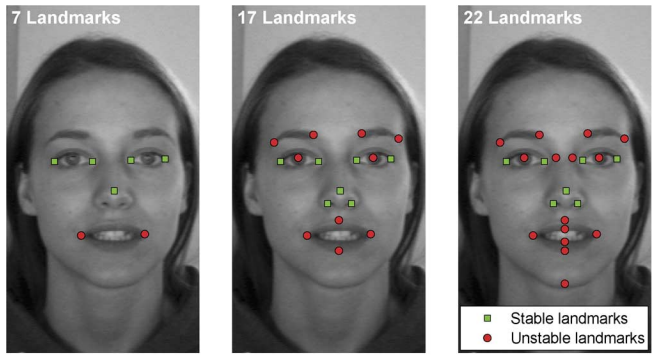


Fig. 2. Selected landmarks: (green marker) stable and (red marker) unstable landmarks.

Let u_j denote the Gabor feature vector for a given landmark l, with $j \in \{w, v\}$ denoting the orientation and the scale of the Gabor filter, respectively. Then, the selected location for a given landmark is selected to maximize the following:

$$S_{l,\alpha}(c) \left(\sum_{j \in \{w,v\}} \sum_{k=1}^{K_j} p\left(u_j | G_j^k\right) p\left(G_j^k\right) \right) \tag{7}$$

where G_j^k is a Gaussian component, defined by $N(\mu_j^k, \Sigma_j^k)$, and K_j denotes the number of components in the mixture for a given j. $p(G_j^k)$ is the prior probability of component k, and $p(u_j|G_j^k)$ is the probability that u_j is generated by component k. Because of the transformation step in the shape prior estimation, stable points are detected before the set of unstable points.

D. Structural Analysis

The shape prior introduced in (7) ensures the integrity of the landmark constellation. The effect is graphically shown in Fig. 3. In the absence of such a prior, the independent detection of landmarks can occasionally result in large errors. In [4], the GOLLUM algorithm was introduced to test the structural integrity of the landmarks. This is a very fast algorithm that solely operates on the low-dimensional shape space (i.e., 2-D coordinates of landmark points) and can be added as a postprocessing step to any landmarking algorithm. Since the temporal complexity is negligible, the results obtained with the proposed method use the GOLLUM algorithm as a postprocessing step after the first-level search. We summarize it here and refer the reader to [4] for details.

In GOLLUM, the landmarks are separated into two groups. The *support set* is a set of three correctly localized landmarks. The remaining landmarks are tested for integrity, based on the support set. Using the support set, the algorithm computes a transformation, which is affine invariant. The expected locations

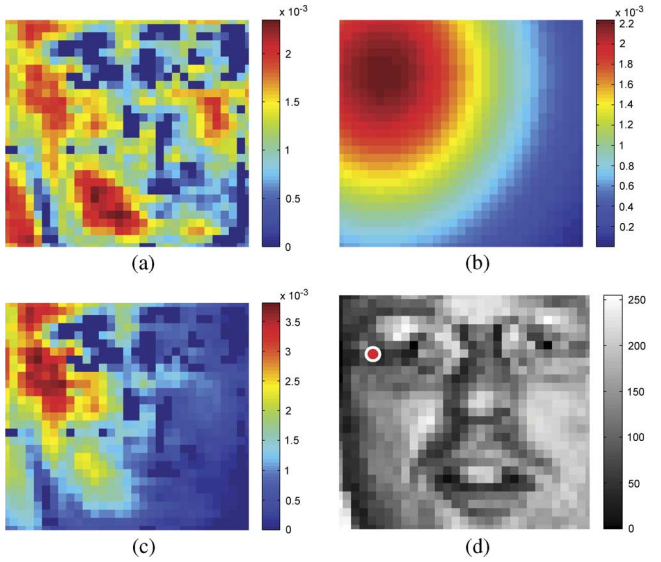


Fig. 3. First-level search for outer eye corner detection on a face with glasses. The probability map with respect to Gabor feature (a) likelihoods, (b) prior probabilities, (c) combined probability map, and (d) the detected landmark on the first-level image.

of landmarks outside the support set are learned during training from the training set. During the operation, if any of the landmarks are not close to their expected location, they can be detected and corrected with the expected location itself. Since we do not know which landmarks are correctly located, the algorithm tests all possible support sets and selects the result with maximum integrity.

The complete landmarking algorithm consists of the extraction of Gabor features in different scales and orientations; the computation of a likelihood for stable and unstable landmarks, combined with the shape prior; and the structural analysis postprocessing. We now describe the experiments conducted to verify the properties of the proposed scheme and discuss our findings.

IV. Experimental Results With Discussion

A. Experimental Setup and Data

To fully explore the system performance, we train and test our approach with challenging cross-database protocols. The particulars of the databases relevant to our discussion are given below.

AR [25]: The AR database consists of more than 4000 frontal face images from 126 subjects. Images of each subject were recorded in two sessions. These color images have a resolution of 768 × 576 pixels and include different facial expressions, illumination conditions, and occlusions. In our experiments, the AR database is used as a validation set. We use 508 images (without occlusions), which have manually annotated landmarks in the face and gesture recognition research network (FGnet) project [26].

BioID [27]: The BioID database consists of 1521 roughly frontal facial images from 23 subjects. Variations in the BioID database include different illumination, background, and face size conditions that resemble "real world" conditions. Images are grayscale and have a resolution of 384 × 286 pixels. In our tests, we use 1482 images in which faces can be detected with the Viola–Jones face detection algorithm.

Bosphorus [28]: The Bosphorus database consists of 3-D faces and corresponding texture images, specifically collected for expression analysis purposes. The subject variation in this database comprises not only various expressions and poses but also realistic occlusions such as glasses, hair tassel, and hands over the face area. The pose variations are composed of systematic rotations. The facial expressions include six universal expressions (happiness, surprise, fear, sadness, anger, and disgust), as well as expressions based on facial action units (AU) of the Facial Action Coding System [29]. There are 61 male and 44 female subjects (29 of which are professional actors and actresses) with a total number of 5102 face images. Frontal and $+10°$- and $+20°$-rotated faces are used in our landmark localization experiments. The texture images are of high quality and are acquired under controlled studio light.

Cohn–Kanade [30]: The Cohn–Kanade AU-Coded Facial Expression Database consists of approximately 500 image sequences from 100 subjects. These image sequences incorporate both single AUs and AU combinations, as well as six universal expressions. Each of these sequences starts with a neutral/nearly neutral face. The annotation of emotion-specified expressions are provided in the database. Only frontal images are open to public use, and we only use those. Image sequences were digitized with a resolution of 640 × 480 or 640 × 490 pixels. We use 249 image sequences from the Cohn–Kanade database. Two different data sets are prepared by taking the first (neutral) frame and the most extreme expression frame of these sequences, respectively.

Face Recognition Grand Challenge (FRGC) [31]: The FRGC data set of 2-D/3-D face images was collected by the University of Notre Dame, and it is one of the most prominent data sets used for face recognition. For a fair comparison with the results obtained on the Bosphorus data set, we only use the Spring 2004 subset, which is the most challenging setting. It contains 2114 face images (neutral and uncontrolled expressions) from 465 subjects. Each subject has between one to eight images. The resolution of images is 640 × 480 pixels. Facial scans are acquired with different distances to the camera, under challenging natural illumination conditions.

Extended M2VTS [32]: The extended M2VTS database contains facial images, sound files, and 3-D face models of 295 subjects. It has 2360 color images with a resolution of 432 × 346 pixels. Recordings of each subject were taken in four sessions. We have excluded the images with closed eyes, hair occlusions (on landmarks), and motion blur. The remaining 1701 images are used as the training set. To replicate the training protocol in [3], we have doubled the size of this set by mirroring images.

All these manually annotated data sets are used in landmark localization experiments. Fig. 4 illustrates the scale, pose, expression, resolution, and illumination conditions across

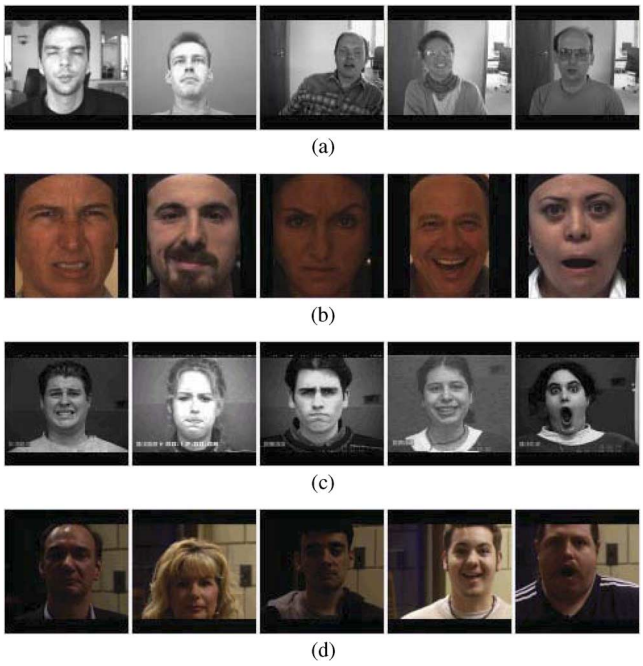


Fig. 4. Samples from (a) BioID, (b) Bosphorus, (c) Cohn–Kanade, and (d) FRGC databases. (Fourth row) Difficult illumination conditions on the FRGC (uncontrolled) set.

BioD, Bosphorus, Cohn–Kanade, and FRGC databases. We use the AR, Bosphorus, Cohn–Kanade, FRGC, and XM2VTS databases for training different versions of our automatic landmarking system. The BioID database is used for the comparison of the proposed system with state-of-the-art methods. For Cohn–Kanade and FRGC databases, the ground truth for seven landmarks are manually annotated. Additionally, we have 22 landmarks for frontal images in the Bosphorus database, 20 landmarks for the BioID database, 22 landmarks for a subset of the AR database, and 68 landmarks for the XM2VTS database. The landmarks of the BioID, AR, and XM2VTS databases are provided by the FGnet project [26]. The AR, BioID, Bosphorus, FRGC, and XM2VTS sets are composed of static images, whereas the Cohn–Kanade data set has video sequences. All models are trained with a *training* partition of the corresponding database, and the model complexity is controlled on a *validation* partition. Then, results are reported on a separate *test* partition. The sizes of training, validation, and test sets for each database are given in Table II, and their exact composition is available at the corresponding author's Web page. There are no subject overlaps that we are aware of between databases.

B. Performance Measures

In our experiments, interocular distance d_{io} is used for computing an error measure. It is the distance between left- and right-eye centers, and is regularly used in state-of-the-art studies on 2-D facial landmarking. A landmark location is accepted as correct if the distance to the ground-truth location is less than a percentage of interocular distance. This threshold is typically set to 10% or 20%, and we have set it to 10% in our tests,

TABLE II
ABBREVIATIONS; NUMBER OF MANUALLY ANNOTATED GROUND TRUTH FOR LANDMARKS (GT); AND THE SIZE OF TRAINING, VALIDATION, AND TEST SETS FOR EACH DATABASE

Database	Abbr.	Train	Validation	Test	GT
AR	AR	–	508	–	22
BioID	BioID	–	–	1482	20
Bosphorus (Frontal)	BOS	1057	529	1334	22
Bosphorus ($+10°$)	BOS-R10	37	22	46	22
Bosphorus ($+20°$)	BOS-R20	37	22	46	22
Cohn-Kanade (Neutral)	CK-N	–	–	249	7
Cohn-Kanade (Expressive)	CK-E	–	–	249	7
CK (Expressive + Neutral)	CK-All	332	166	–	7
Extended M2VTS	XM2VTS	1701	–	–	68
FRGC (Spring 2004)	FRGC	1057	529	528	7

which is the more difficult of the two. In the absence of the absolute metric ground truth, the interocular distance is a reliable measure because it provides constancy in terms of the scale. For faces scanned by calibrated 3-D sensors, absolute distances are available. It is then possible to adapt a threshold set to a certain absolute distance. Except for comparisons with other state-of-the-art methods, we calculate the average performances as the mean accuracy for the detection of the each landmark.

To compare our system with other reported results, we present the performance of our method in terms of the m_e error measure on the BioID database, as described in [2]. m_e is defined as the normalized mean distance of internal facial feature points to their ground-truth locations. Feature points that are close to the edge of the face (such as the tip of the chin) are ignored because they are easily affected by the rotation of the face, and the ground-truth annotation is noisy. Instead of providing individual landmark errors, m_e gives a mean error for the entire system, i.e.,

$$m_e = \frac{1}{n d_{\text{io}}} \sum_{l=1}^{n} d_l \tag{8}$$

where n denotes the number of landmarks and d_l values are the Euclidean point-to-point distances for each individual landmark location.

We report the average accuracy on the Bosphorus data set with both measures to emphasize the difference between these error measures.

C. Accuracy of Landmarking

We first report the accuracy of our landmarking algorithm with the Bosphorus database, for which we have the most extensive ground truth. We use 22 landmarks [see Fig. 2(c)] rolled into 12 groups, i.e., outer eyebrows, inner eyebrows, outer eye corners, inner eye corners, eye pupils, nose tip, nose saddles, nostrils, mouth corners, inner lip middles, outer lip middles, and tip of the chin. The detailed results are reported for each landmark group in Table III. The proposed system has 92.21% average accuracy when accepting points within 10% of interocular distance to the ground truth. This is the most stringent criterion used in the 2-D landmarking literature. If 3-D information is available, errors can be reported using millimetric ground

TABLE III
ACCURACY OF THE LANDMARKING ALGORITHM ON THE BOSPHORUS DATABASE

Landmarks	Success (%)	Mean Error (% of d_{io})
Outer Eye Corners	97.98	3.16 (±3.96)
Inner Eye Corners	98.46	2.54 (±3.54)
Nose Tip	94.68	3.65 (±4.07)
Mouth Corners	90.74	4.94 (±5.48)
Outer Eyebrows	91.42	4.80 (±4.95)
Inner Eyebrows	94.79	4.35 (±4.20)
Pupils	98.84	2.37 (±3.59)
Nose Saddles	86.51	4.93 (±5.74)
Nostrils	99.03	2.84 (±3.50)
Lip Outer Middles	88.76	5.85 (±9.16)
Lip Inner Middles	89.69	4.83 (±7.68)
Tip of chin	61.47	10.02 (±6.62)
Mean	**92.21**	**4.31 (±5.19)**
$\mathbf{m_e \leq 0.1}$	**99.33**	

truth. In Bosphorus and FRGC, 10% of interocular distance corresponds to 6.3 mm on the average. In [33], 12- and 16-mm precisions are used for successfully located inner eye corners and nose tip, respectively. In [34], 20-mm precision is used for successfully located landmarks. In [35], it is indicated that 99.09% of samples are located within 10-mm precision.

Table III also shows (the bottom row) the accuracy of the proposed method under the error measure $m_e \leq 0.1$. With this measure, our method reaches 99.33% accuracy on the Bosphorus set. This demonstrates the significant change in reporting accuracy for different error measures. Excluding the tip of the chin from the results also has a significant impact on reported accuracy.

D. Assessment of Generalization

The local features on which we base our analysis may depend on acquisition and preprocessing conditions of a specific database, and statistical methods may be ineffective on cross-database tests. Here, we apply cross-database tests on the statistical models learned from the FRGC, Bosphorus, and Cohn–Kanade databases to evaluate the generalization capabilities of our model.

We run the proposed algorithm for seven landmarks since we have manually annotated ground truth for only seven landmarks on the FRGC and Cohn–Kanade data sets [see Fig. 2(a)]. Landmarks are rolled into four groups, i.e., outer eye corners, inner eye corners, nose tip, and mouth corners. The correct localization accuracies for different training and test sets are given in Table IV. The reported success rates are obtained by accepting points within 10% of interocular distance to the ground truth, which is much more stringent than m_e, as we demonstrated in the previous section. The intersubject variation of the ground truth itself is about 5%–7% of interocular distance. Under the same acquisition conditions (samples from the same database), the average accuracy is 95.6% for Bosphorus and 94.5% for FRGC databases (the first two rows). These are comparable with reported state-of-the-art figures. The cross-database results are given in the remaining rows of the table.

TABLE IV
CROSS-DATABASE ACCURACY OF THE LANDMARKING ALGORITHM

Database		Success (%)			
Training	Test	O. Eye	I. Eye	Nose	Mouth Cor.
BOS	BOS	97.98	98.46	94.68	90.74
FRGC	FRGC	94.89	94.32	93.75	94.70
BOS	CK-N	92.77	98.19	98.39	91.97
BOS	CK-E	90.76	92.57	97.19	90.36
BOS	FRGC	91.86	89.11	87.50	88.83
FRGC	BOS	90.33	88.61	89.58	90.55
FRGC	CK-N	91.57	91.97	92.77	97.39
FRGC	CK-E	90.36	91.16	89.96	91.57
CK-All	BOS	89.96	89.81	91.67	86.88
CK-All	FRGC	90.91	92.42	85.80	90.15

Fig. 5. Some failure cases of the proposed method.

We do not have sufficiently many samples in the Cohn–Kanade set for training a system in conditions comparable with the FRGC and Bosphorus sets. Subsequently, we join the neutral and extreme expressions of the Cohn–Kanade (denoted with CK-All) and train a system to compare our results with the system proposed in [8], which also reports results on this database. We use the authors' own published code, trained on the Cohn–Kanade database. The cross-database results we obtain with this system are 79.8% and 77.8% for the FRGC and Bosphorus databases, respectively. Our system has 90.4% and 89.3% accuracy under similar training conditions, respectively.

The systems trained on FRGC and Bosphorus both show about 3% accuracy decrease when we compare detection on neutral Cohn–Kanade faces to detection on extreme expressions. Most of this loss is due to mouth corners, which are, for some extreme expressions outside, the search area constrained by the shape prior. When this area is enlarged, this loss is quickly alleviated, but the computation requirement is increased. Fig. 5 shows some failure cases of the proposed method.

On average, the mouth corners are detected with less accuracy than other landmarks, as they are affected more under expression changes. The variation on the training set naturally reflects on the testing conditions. Since the FRGC data set has more pose variations than the frontal subset of Bosphorus, the shape prior implicates a larger area, resulting in higher accuracy. Although not shown here, GOLLUM increases the accuracy by 1% on the average. In the absence of the introduced shape prior, GOLLUM would contribute 5%–10% depending on the landmark. The shape prior reduces its impact by eliminating large deviations from expected locations. For this reason, this is not a

TABLE V
EFFECT OF STATISTICAL FEATURE MODELING AND STRUCTURAL PRIOR

Method	Test Database			
	BOS		FRGC	
IMoFA	80.53 (±2.28)	89.28 (±1.77)	82.27 (±1.68)	90.39 (±2.24)
SVM	80.09 (±2.40)	88.24 (±1.86)	79.40 (±2.56)	88.98 (±2.35)
GMM	80.23 (±2.81)	86.13 (±2.19)	76.19 (±3.30)	84.24 (±3.60)
PCA (Mahalanobis)	76.05 (±2.64)	79.88 (±3.08)	77.52 (±2.34)	80.30 (±2.15)
	with ROI constraint	with structural prior	with ROI constraint	with structural prior

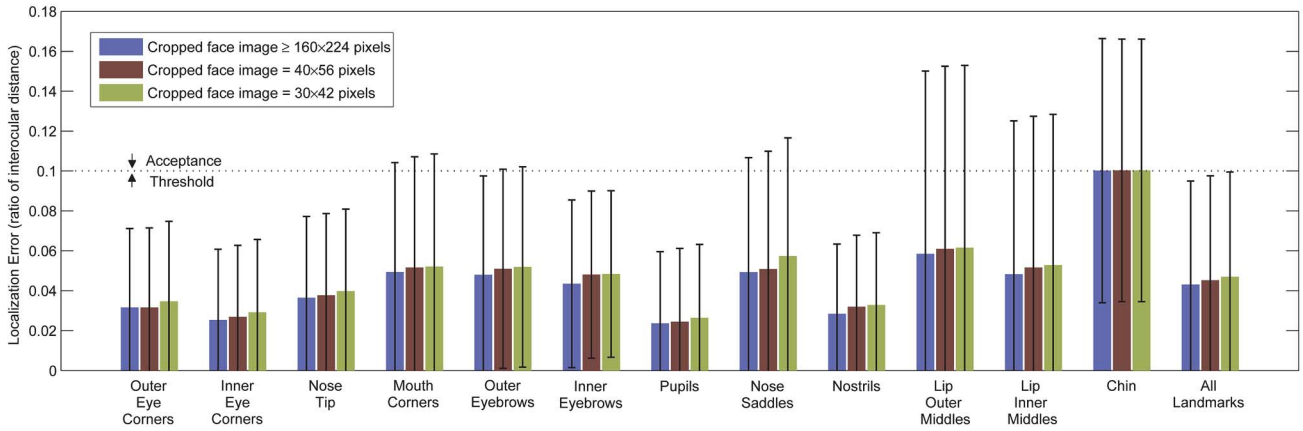


Fig. 6. Effect of image resolution on landmark localization accuracy.

very surprising result. As we noted before, we retain GOLLUM because it has negligible time complexity, both in theory (fixed) and in practice.

E. Effect of the Model Choice

To compare the robustness of the IMoFA algorithm with other popular methodologies, extracted Gabor feature vectors are modeled for seven landmarks [inner/outer eye corners, nose tip, and mouth corners, see Fig. 2(a)] with support vector machines (SVMs), GMMs, and principal component analysis (PCA). Both expressive and neutral Cohn–Kanade data sets are used to train these models. For SVM, GMM, and PCA methods, we either use a rectangular ROI constraint to focus the search (as frequently used in the literature) or use the proposed structural prior. For the SVM, the classifier is used to generate distance maps, which are then weighted with shape priors, and the maximum location is selected. To optimize the SVM configuration, different kernels with different parameters are tested on the validation set, and the configuration with the minimum validation error is selected. Training sets and patch sizes are the same for all compared methods. In the absence of the ROI or the structural prior (hence searching the landmark on the entire face), the results will be very poor. The same Gabor wavelet features are modeled with both approaches. The PCA dimensionality is selected through the screen graph by taking sufficient eigenvectors to explain 95% of the variance. The number of components for the GMM is automatically determined (between 1 and 15), as proposed in [23]. Table V shows that the IMoFA improves on the SVM, the GMM, and the PCA, and that the proposed structural prior gives better results than a ROI-based constraint.

The major advantage of using IMoFA is in its automatic parameter estimation and flexibility. Complex data relations are learned with more parameters, whereas simple structures are devoted less parameters and have better generalization. Compared with standard mixtures of Gaussians, the IMoFA explores models of intermediate complexity between a full-covariance Gaussian and a diagonal-covariance Gaussian. As the dimensionality of the feature vector is increased, the gains of such a flexibility become more marked.

F. Effect of Resolution

To assess the effect of image resolution on landmarking accuracy, we use images with different resolutions. Datasets for multiresolution analysis are prepared by resizing the Bosphorus test set. We use face images with a resolution of 30×42, 40×56, and 160×224 pixels. Fig. 6 shows that, even when the resolution of the face area is drastically reduced, the feature localization scheme successfully locates the features. This is an expected result as our multiresolution analysis starts by a resolution reduction; thus, the crucial first-level of analysis is not affected at all. We exploit the fact that the coarse level is the most adequate for statistical modeling, as pixelwise correlation is at its lowest.

G. Effect of Expressions

To evaluate the robustness of our method with respect to facial expressions, we inspect the effect of expressions on the Bosphorus database, which has rich expression variations. We group the samples in the Bosphorus database into five, i.e., as "neutral," "emotional expression," "lower-face AU,"

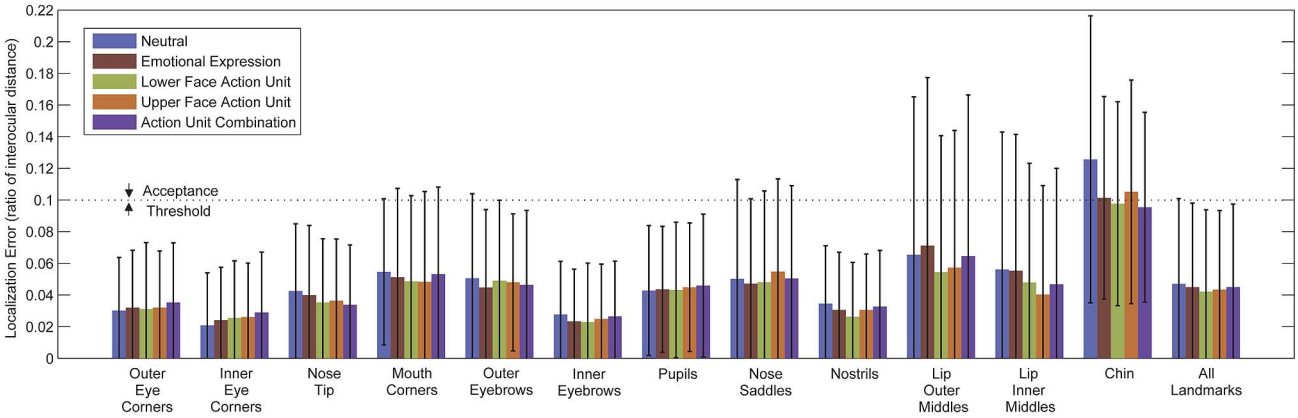


Fig. 7. Effect of facial expressions on landmark localization accuracy.

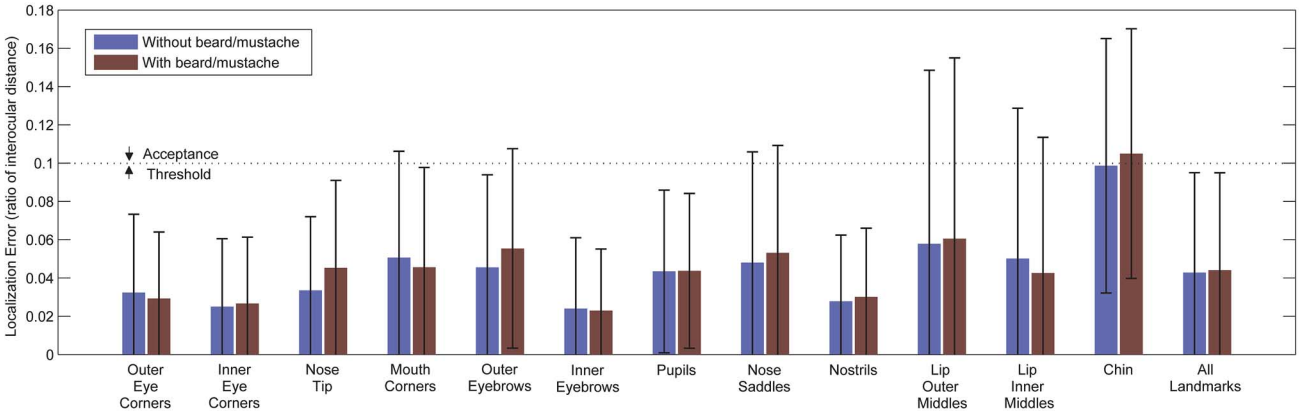


Fig. 8. Effect of beard/mustache on landmark localization accuracy.

"upper-face AU," and "AU combination." In the AU combination case, both upper and lower AUs are activated. As shown in Fig. 7, our method provides the most accurate results with lower and upper face AUs, although the differences are not great. This is partly due to the larger number of samples with lower- and upper-face AUs in the Bosphorus set. There are fewer neutral samples in the Bosphorus data set, as compared with expressive samples, which results in a more restrictive shape prior for the neutral case. Among the landmarks, the tip of chin is the most difficult one and receives the least leverage from the shape prior for the neutral images.

H. Effect of Beard and Mustache

We also analyze the effect of the beard and the mustache on the landmark localization accuracy. Roughly one third of images in the Bosphorus database have either a beard or a mustache; thus, the training data set contains enough samples to generalize. As shown in Fig. 8, the accuracy differences are smaller than differences due to landmark types, even for mouth corners and lips. We conclude that the beard and the mustache do not deteriorate the accuracy of the proposed method, provided that the training set includes sufficiently many samples for generalization.

I. Effect of Rotations

Both 2-D shape and landmark appearance change under rotations. We experimented with the rotated samples of the Bosphorus data set to see first how much our shape prior was affected by the assumption that the face is frontal when it is not. Since the number of rotated samples is much less, we take windows shifted by one pixel from the manually annotated landmark for each sample and increase the number of rotated samples ninefold. While, under $+10^\circ$ rotations, all landmarks were located by the prior, there was minor (2.4% on the average) loss for $+20^\circ$ rotations, except for the mouth corner in the rotated side (60% loss). We did not look beyond $+20^\circ$, as the standard Viola–Jones cascade fails under more severe rotations. These results strongly relate to the cascade we have used; another cascade, derived from another training set, might give different results by cropping the face area with different margins. When we add the rotated training samples to the learning set of the shape prior, all samples are located within their expected locations.

We then look at the performance of the whole system under these conditions. We have two conditions to learn the shape prior, as well as to train the appearance features via the IMoFA. The first condition is called *frontal* and is composed of only frontal training samples. The second condition is *both*, meaning that the rotated training samples are added to the frontal. We measure the effect of changing just the appearance or both shape

TABLE VI
ACCURACY OF LANDMARKING UNDER DIFFERENT ROTATION CONDITIONS

Rotation (Test)	Appearance (IMoFA)	Shape (Prior)	Success (%)						
			LOE	LIE	RIE	ROE	Nose	LMC	RMC
frontal	*frontal*	*frontal*	97.75	98.13	98.80	98.20	94.68	92.80	88.68
frontal	*both*	*frontal*	95.50	94.08	96.70	95.05	93.93	91.23	85.01
frontal	*both*	*both*	95.43	94.15	96.70	94.90	93.78	90.55	85.68
$+10^\circ$	*frontal*	*frontal*	89.13	93.48	95.65	93.48	76.09	93.48	91.30
$+10^\circ$	*both*	*frontal*	86.96	95.65	93.48	91.30	86.96	91.30	91.30
$+10^\circ$	*both*	*both*	91.30	93.48	95.65	91.30	86.96	91.30	93.48
$+20^\circ$	*frontal*	*frontal*	17.39	82.61	89.13	73.91	19.57	34.78	76.09
$+20^\circ$	*both*	*frontal*	67.39	89.13	78.26	71.74	65.22	56.52	86.96
$+20^\circ$	*both*	*both*	69.57	91.30	82.61	71.74	65.22	58.70	86.96

and appearance on frontal and rotated sets. Seven landmarks (left/right outer eye corners: LOE/ROE; left/right inner eye corners: LIE/RIE; nose tip; and left/right mouth corners: LMC/RMC) are tested for different rotation conditions in our experiments. The results are summarized in Table VI. The reported success rates are obtained by accepting points within 10% of interocular distance to the ground truth. The table demonstrates that our method has some capability in dealing with minor pose variations not present in the training set, but for improved detections, it is necessary to enrich the training set. Adding the very limited set of rotated training samples to the training set reduces the accuracy by 2.5% on the average for frontal samples and increases the accuracy by 18.9% for the $+20^\circ$-rotated samples on the average.

J. Comparison with Other Methods

Recent landmarking results in the literature vary according to the database employed for reporting. In [7], for instance, 85.8% average localization accuracy is reported for finding 24 landmarks on XM2VTS and UniMiDb databases, but the acceptance condition is not given, which we have shown to have a significant effect on the interpretation of the results. In [2], the reported localization accuracy for 22 landmarks is 95% (within 20% of interocular distance) for the BioID database and 92% for the XM2VTS database. In the experimental results presented in this paper, we use a more stringent criterion and use 10% interocular distance as the acceptance threshold. Kozakaya *et al.* report 95.1% average localization accuracy for their weighted vector concentration approach [6] on the facial-recognition-technology database.

BioID is the data set on which the most promising and recent results are reported. Therefore, we conducted a study on BioID and compared our accuracy with five recent methods denoted as *SliWiGa* [8], *CLM* [2], *AAM* [2], *Stacked Model* [3], and *BorMaN* [5]. In [8], Vukadinovic and Pantic use boosted classifiers to model Gabor features and grayscale texture values and constrain the search areas with the related ROIs. In [2], Cristinacce and Cootes use boosted Haar cascades for coarse-scale detection, fine tune the locations through a constrained local models approach, and also report results with active appearance models. In [3], Milborrow and Nicolls enable

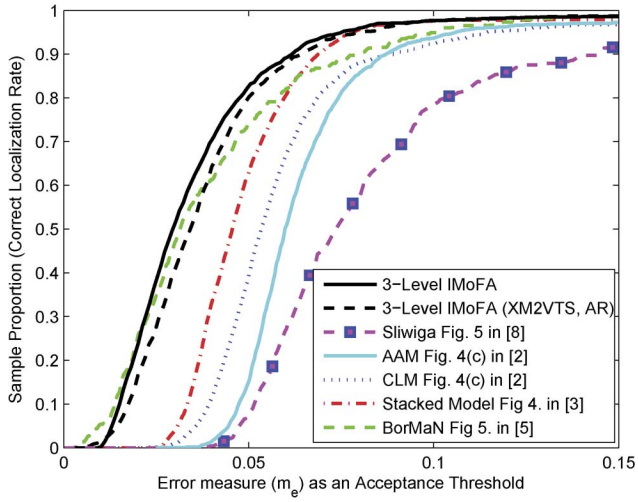


Fig. 9. Cumulative error distribution of the point-to-point error measure m_e for different methods on the BioID data set.

a number of simple extensions to the ASM approach in [14] such as stacking two ASMs in series. Haar-like features are modeled with a support vector regressor, and the search space is constrained by Markov random fields in [5]. Fig. 9 shows the cumulative error distribution of the point-to-point error measure m_e and demonstrates that the accuracy of the proposed method (shown as *3-Level IMoFA*) is good. For a fair comparison, we exclude five points (nose saddles, lip inner middles, and the tip of the chin) and give results for only 17 points, as reported in these studies [see Fig. 2(b)]. Since the competing methods use different training/validation protocols, we have included the cumulative error distributions of our method with two different training/validation protocols in Fig. 9. The system, shown as *3-Level IMoFA (BOS)*, has been trained and validated on the Bosphorus database. To replicate the training/validation schema of *Stacked Model* [3], we have trained another system with XM2VTS and have validated on the AR database [shown as *3-Level IMoFA (XM2VTS, AR)*]. Our method gives similar results with different training conditions. The training on Bosphorus provides slightly higher accuracy for low m_e values because the Bosphorus database has higher resolution and more expression variations than XM2VTS and AR databases.

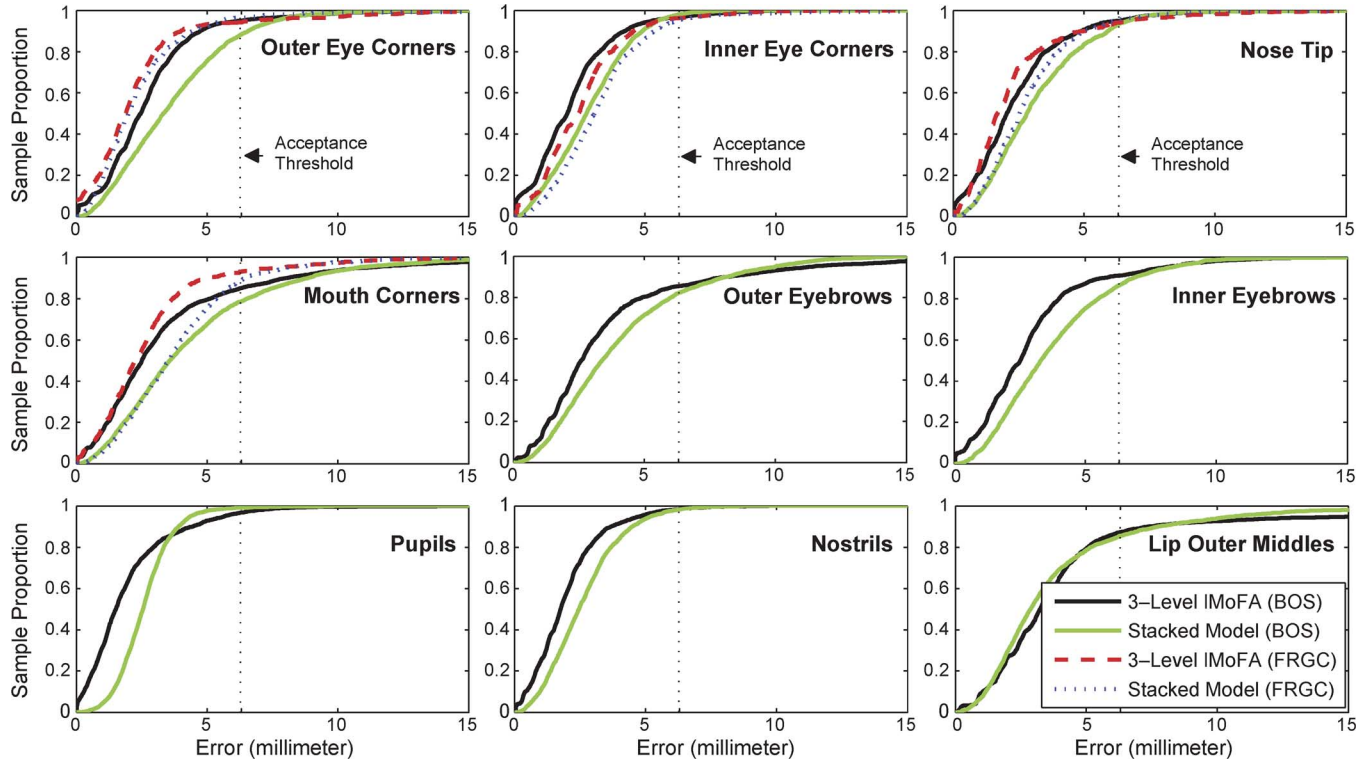
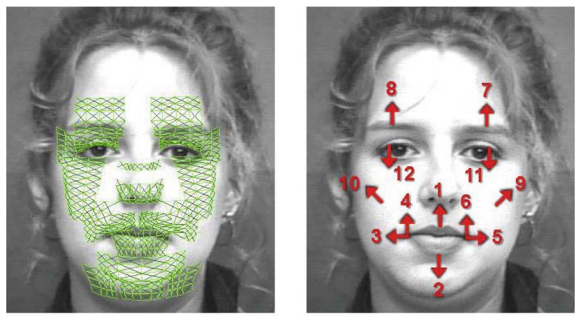


Fig. 10. Cumulative error distributions of the 17 points on the Bosphorus and FRGC data sets for different methods.

Additionally, we have focused on the best competing system, for which the training/testing software is available. Consequently, we have compared our method with *Stacked Model* under exactly the same experimental conditions. In this experiment, we use Bosphorus training and AR validation for both *3-Level IMoFA* and *Stacked Model*. Fig. 10 shows the cumulative error distributions in terms of millimeters for the 17 landmarks [see Fig. 2(b)] on the Bosphorus and FRGC data sets. Since our method does not use any (facial) model fitting, it optimizes the probable location of each landmark individually. As a result, the proposed approach more accurately performs for low errors ($\leq$8 mm). For high error values, both methods are very successful, and the small error they make can be due to labeling errors or occasional poor imaging conditions (e.g., FRGC contains samples with motion blur, where the exact landmark location is not perfectly indicated in the manual annotation).

K. Speed

While we report improved accuracy over competing methods, our current implementation is not real time. Processing speeds of *Sliwiga* [8] and *BorMaN* [5] methods were not reported in the related papers. Average time requirements of *CLM/AAM* [2] and *Stacked Model* [3] (with 3-GHz Pentium, on a still image) are given as 0.24 and 0.22 s, respectively. Our system localizes 22 landmark points on an image within 3.1 s (with a combination of nonoptimized C++ and compiled MATLAB code, including face detection) on an Intel Core2 Duo 3-GHz processor with 3 GBs of random access memory. For seven landmark points, it requires 1.22 s per image. Real-time performance can be achieved by multithreading and multicore programming, since parallel implementation is straightforward in our method. The separate likelihood computations for each Gabor feature channel can be performed in parallel. This is the most time-consuming step in the proposed algorithm.



(a) (b)

Fig. 11. (a) Bézier volume model. (b) Motion units.

V. Application: Expression Classification

We use the automatically located landmarks in an expression classification application. Accurate facial landmarks are crucial for expression analysis, and in the absence of a reliable landmarking algorithm, expression analysis requires costly manual initialization [36]. Automatic expression analysis received a lot of attention in the last few years, and there is great progress in granular methods that aim at detecting facial AUs, linking expressions to these AUs [37]. On the downside, these methods require correspondingly granular AU annotations prepared by experts for training.

TABLE VII
AVERAGE TRACKING ERROR ON THE COHN–KANADE DATABASE AS A PERCENTAGE OF INTEROCULAR DISTANCE

Support	O. Eye	I. Eye	Nose	Mouth C.	Average
Viola-Jones	7.34 (±1.23)	7.79 (±1.41)	11.99 (±1.93)	14.79 (±3.04)	10.26 (±1.90)
Auto (BOS)	5.28 (±1.24)	7.08 (±1.30)	4.12 (±1.58)	9.36 (±2.94)	6.79 (±1.79)
Auto (FRGC)	6.12 (±1.46)	7.44 (±1.42)	5.41 (±1.80)	8.28 (±3.13)	7.01 (±1.97)
Ground Truth	4.98 (±1.27)	6.73 (±1.17)	2.79 (±1.67)	7.57 (±2.77)	5.91 (±1.73)

TABLE VIII
EMOTION RECOGNITION ACCURACIES ON THE COHN–KANADE AND BU-4DFE DATABASES

Landmark Support	Test Set	Happy	Surprised	Angry	Disgusted	Frightened	Sad	Average
Viola-Jones	CK-Video	80.43	95.92	66.67	59.46	43.90	69.77	70.68
Auto (BOS)	CK-Video	89.13	95.92	69.70	72.97	63.41	72.09	78.31
Auto (FRGC)	CK-Video	92.10	100	72.73	75.68	60.98	65.12	78.86
Ground Truth	CK-Video	93.48	100	69.70	81.08	65.85	67.44	80.72
Viola-Jones	BU-4DFE	82.89	77.27	79.27	39.74	24.14	82.14	64.04
Auto (CK-All)	BU-4DFE	88.16	87.50	80.49	51.28	63.22	79.76	75.15
Auto (BOS)	BU-4DFE	93.42	94.32	82.93	66.67	62.07	82.14	80.20
Auto (FRGC)	BU-4DFE	94.74	96.59	86.59	70.51	68.97	83.33	83.44

Our prototype application aims at identifying the presence of six basic emotional expression categories. The training of such a system can be performed with a set of videos that are annotated for the main expression categories in broadly defined segments. We briefly describe the main components of this application here. Since the literature on facial-expression analysis is extensive, we refer the reader to [38] and to a more recent work [39] for related approaches.

A. Methodology

Our baseline method, described in [40], maintains a face model described by 16 surface patches embedded in Bézier volumes, as shown in Fig. 11(a). Once this model is fit to the appearance of the face, a piecewise Bézier volume deformation tracker is used to trace the motion of the facial features [41]. In the system described by [40], this tracker is initialized by an average shape model, learned during the training phase. We test the effect of our automatic landmark detection algorithm on this system.

We use the well-known thin-plate spline (TPS) algorithm for warping the generic face model to the detected landmark locations [42]. Since Cohn–Kanade and FRGC data sets have manually annotated ground truth for only seven landmarks, we use detected seven landmarks for warping [see Fig. 2(a)]. The deformation transforms the landmarks on the model to exactly match the detected landmarks. The rest of the points are deformed in accordance to their proximity to the landmarks. During the development of our algorithm, we have also tested Procrustes alignment [24] as an alternative, but our experiments showed that the TPS was more accurate in most cases. In general, we expect this to be the case when the landmarks are sufficiently accurate.

We use a simple but efficient naive Bayes classifier for categorizing expressions. One advantage of the method is that the posterior probabilities allow a soft output of the system, usable as a continuous input to any facial affect-based system. The classifier receives as input-quantized difference vectors extracted from a number of locations on the model. These are called motion units, and they are not unlike AUs but simpler and tailored toward basic expression categories [see Fig. 11(b)].

B. Database and Experiments

We use the Cohn–Kanade [30] and BU-4DFE [43] data sets for testing the facial-expression analysis system. These data sets have video sequences and are thus adequate for the dynamic measurement of expressions. The BU-4DFE data set has a more challenging nature, and the expressions are not as pronounced and well-segmented as the Cohn–Kanade set.

We use only the texture portion of the BU-4DFE data set, which contains facial expressions captured at a video rate of 25 fps. The database contains 606 facial-expression sequences (of about 100 frames) captured from 101 subjects. Four hundred ninety-five of these commence with a neutral face, and those are used for our experiments. For each subject, there are six model sequences showing all basic expressions. The texture video has a resolution of 1040×1329 pixels per frame. There are 58 female and 43 male subjects, with mixed ethnic ancestries.

We evaluate the effect of locating seven landmarks (outer eye corners, inner eye corners, nose tip, and mouth corners) on this application. Two hundred forty-nine image sequences from the Cohn–Kanade database are tested with threefold cross validation. For expression classification tests on the BU-4DFE, we use the system that is trained with the Cohn–Kanade data set.

In Table VII, we report the average tracking error on the Cohn–Kanade database. All seven landmarks are manually annotated for all frames of the database. The reported error is the average deviation of the tracked point from its ground-truth annotation as a percentage of the interocular distance. The first row is the baseline error over all frames and shows an average tracking error as 10.26% of interocular distance. As expected,

mouth corners show the highest error. Our landmark-based initialization reduces this error by 3.47% of interocular distance on the average. If the ground truth for landmark locations is made available at the onset of the algorithm, there will be additional error reduction of 0.88% of interocular distance. This indicates that there is a room for improvement on automatic landmarking.

Finally, we report the net effect on expression classification on Table VIII. For each test set, we report results without landmarking (baseline) and with landmarking trained on the FRGC, Bosphorus, or Cohn–Kanade database. Results with the landmark ground truth are reported only for the Cohn–Kanade data set since we do not have manually annotated ground truth for the BU-4DFE data set. We obtain different effects for different expressions; happiness, fear, and disgust greatly benefit from the improved alignment. Sadness does not because it is a subtle expression and it mostly relies on eyebrow movements, which is not among the detected landmarks. The results show a 8%–16% (absolute) accuracy increase from the baseline. For the Cohn–Kanade data set, a baseline accuracy of 70.68% is increased to 78.86% via FRGC training. For the BU-4DFE data set, 80.20% classification accuracy is achieved with a 16.16% (absolute) accuracy increase from the baseline via Bosphorus training.

VI. Conclusion

We have described in this paper a statistical landmark localization method with good cross-database accuracy. The three-level search for landmarks, approximately constrained by a multivariate shape prior, allows for a robust landmarking scheme. The flexible statistical models that we have used increase the accuracy of our models. Closely placed landmark points are difficult to separate with statistical landmark classification methods, as the models need to deal with idiosyncratic variations and noise and to thus be sufficiently "general" in admitting a landmark. Our paper has extensively demonstrated the possibilities and the limits of such approaches. We have assessed our method under different performance criteria and discussed the implications. The complete evaluation protocol has been made available to the reader on the author's website.

We have included additional experiments on an expression recognition application to demonstrate the significant contribution (7%–16%) of the automatic landmarking procedure over a coarse alignment following face detection. The expression recognition application is not the novelty of this paper; there are better results obtained in the literature for the much-perused Cohn–Kanade set, although cross-database evaluations of such methods are rarely given. We have also reported good results on the more challenging BU-4DFE data set. Most importantly, our expression recognition results have given a notion of the room of improvement for further explorations.

References

[1] A. A. Salah, N. Alyüz, and L. Akarun, "Registration of 3D face scans with average face models," *J. Electron. Imag.*, vol. 17, no. 011006, Mar. 2008.

[2] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proc. BMVC*, 2006, pp. 929–938.

[3] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. ECCV*, 2008, pp. 504–513.

[4] 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.

[5] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. CVPR*, 2010, pp. 2729–2736.

[6] T. Kozakaya, T. Shibata, M. Yuasa, and O. Yamaguchi, "Facial feature localization using weighted vector concentration approach," *Image Vis. Comput.*, vol. 28, no. 5, pp. 772–780, May 2010.

[7] S. Arca, P. Campadelli, and R. Lanzarotti, "A face recognition system based on automatically determined facial fiducial points," *Pattern Recognit.*, vol. 39, no. 3, pp. 432–443, Mar. 2006.

[8] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *Proc. IEEE CSMC*, 2005, pp. 1692–1698.

[9] L. Chen, L. Zhang, H. Zhang, and M. Abdel-Mottaleb, "3D shape constraint for facial feature localization using probabilistic-like output," in *Proc. AFGR*, 2004, pp. 302–307.

[10] R. Senaratne and S. Halgamuge, "Optimised landmark model matching for face recognition," in *Proc. AFGR*, 2006, pp. 120–125.

[11] L. Wiskott, J. M. Fellous, N. Krüger, 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.

[12] T. F. Cootes and C. J. Taylor, Statistical models of appearance for computer vision Manchester, U.K., 2004, Tech. Rep..

[13] Y. Tong, X. Liu, F. W. Wheeler, and P. Tu, "Automatic facial landmark labeling with minimal supervision," in *Proc. CVPR*, 2009, pp. 2097–2104.

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

[15] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *Proc. ECCV*, 2008, pp. 413–426.

[16] S. Zhao, Y. Gao, and B. Zhang, "Gabor feature constrained statistical model for efficient landmark localization and face recognition," *Pattern Recognit. Lett.*, vol. 30, no. 10, pp. 922–930, Jul. 2009.

[17] X. Liu, "Video-based face model fitting using adaptive active appearance model," *Image Vis. Comput.*, vol. 28, no. 7, pp. 1162–1172, Jul. 2010.

[18] X. Yu, J. Tian, and J. Liu, "Active appearance models fitting with occlusion," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Berlin, Germany: Springer-Verlag, 2007, pp. 137–144.

[19] M. Hamouz, J. Kittler, J. K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas, "Feature-based affine-invariant localization of faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1490–1495, Sep. 2005.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, vol. 1, pp. 511–518.

[21] A. A. Salah and E. Alpaydin, "Incremental mixtures of factor analysers," in *Proc. ICPR*, 2004, vol. 1, pp. 276–279.

[22] Z. Ghahramani and G. E. Hinton, The EM algorithm for mixtures of factor analyzers Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-96-1, 1997, (revised).

[23] M. A. T. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[24] C. Goodall, "Procrustes methods in the statistical analysis of shape," *J. Roy. Stat. Soc. B*, vol. 53, no. 2, pp. 285–339, 1991.

[25] A. Martinez and R. Benavente, The AR face data base Purdue Univ., West Lafayette, IN, CVC Tech. Rep. #24, 1998.

[26] FGnet Project [Online]. Available: http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html

[27] BioID Database [Online]. Available: http://www.bioid.com/

[28] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *BIOID*. Berlin, Germany: Springer-Verlag, 2008, pp. 47–56.

[29] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System*. Palo Alto, CA: Consult. Psychol. Press, 1978.

[30] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. AFGR*, 2000, pp. 46–53.

[31] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. J. Worek, "Overview of the face recognition grand challenge," in *Proc. CVPR*, 2005, vol. 1, pp. 947–954.

[32] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTS: The extended M2VTS database," in *Proc. AVBPA*, Washington, DC, 1999.

[33] M. Romero-Huertas and N. Pears, "3d facial landmark localisation by matching simple descriptors," in *Proc. 2nd IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, 2008, pp. 1–6.
[34] D. Colbry, G. Stockman, and A. Jain, "Detection of anchor points for 3d face verification," in *Proc. CVPR Workshops*, 2005, p. 118.
[35] X. Zhao, E. Dellandrea, and L. Chen, "A 3D statistical facial feature model and its application on locating facial landmarks," in *Proc. Adv. Concepts Intell. Vis. Syst.*, 2009, pp. 686–697.
[36] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 36, no. 1, pp. 35–43, 1999.
[37] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. CVPR*, 2005, vol. 2, p. 568.
[38] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[40] R. Valenti, N. Sebe, and T. Gevers, "Facial expression recognition: A fully integrated approach," in *Proc. ICIAPW*, 2007, pp. 125–130.
[41] H. Tao and T. S. Huang, "Connected vibrations: A modal analysis approach for non-rigid motion tracking," in *Proc. CVPR*, 1998, pp. 735–740.
[42] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.
[43] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. AFGR*, 2008, pp. 1–6.

Hamdi Dibeklioğlu (S'08) received the B.Sc. degree from Yeditepe University, Istanbul, Turkey, in 2006 and the M.Sc. degree from Boğaziçi University, Istanbul, Turkey, in 2008. He is currently working toward the Ph.D. degree in the Intelligent Systems Lab Amsterdam, University of Amsterdam, Amsterdam, The Netherlands.

His research interests include computer vision, image understanding, biometrics, pattern recognition, and intelligent human–computer interfaces.

In 2009, he was the recipient of the Alper Atalay second best student paper award at the IEEE Signal Processing and Communications Applications Conference (SIU). He served in the local organization committee of the eNTERFACE Workshop on Multimodal Interfaces in 2007 and 2010.

Albert Ali Salah (M'08) received the Ph.D. degree from Boğaziçi University, Istanbul, Turkey.

From 2007 to 2009, he was with the CWI Institute, Amsterdam, The Netherlands, and with the Informatics Institute, University of Amsterdam, Amsterdam, from 2009 to 2011. He is currently an Assistant Professor with Boğaziçi University. His research interests are biologically inspired models of learning and vision, with applications to pattern recognition, biometrics, and human behavior understanding, where he has more than 50 publications in related areas, including an edited book on computer analysis of human behavior.

Dr. Salah was the recipient of the inaugural European Biometrics Research Award of the European Biometrics Forum in 2006 for his work on facial feature localization. He cochaired the 2010 eNTERFACE Workshop on Multimodal Interfaces, as well as the International Workshop on Human Behavior Understanding in 2010 and 2011. He served as a guest editor for special issues in the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, the Journal of Ambient Intelligence and Smart Environments, and the Journal on Multimodal Interfaces.

Theo Gevers (M'01) is an Associate Professor of computer science with the University of Amsterdam (UvA), Amsterdam, The Netherlands, and a Full Professor with the Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain. At the University of Amsterdam, he is a Teaching Director of the M.Sc. degree in Artificial Intelligence. He currently holds a VICI Award (for research excellence) from the Dutch Organization for Scientific Research. He is a Cofounder and the Chief Scientific Officer of ThirdSight, a spinoff of the UvA. His main research interests are in the fundamentals of image understanding, object recognition, and color in computer vision. Furthermore, he is interested in different aspects of human behavior, specifically in emotion recognition.

Prof. Gevers is the chair for various conferences and is an associate editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. Furthermore, he is a program committee member for a number of conferences and an invited speaker at major conferences. He is a lecturer delivering postdoctoral courses given at various major conferences (the IEEE Conference on Computer Vision and Pattern Recognition; the International Conference on Pattern Recognition; SPIE; and the Computer Graphics, Imaging, and Vision).