Robust Facial Landmarking for Registration

Albert ALI SALAH[1], Hatice ÇINAR[2], Lale AKARUN[1], Bülent SANKUR[2]

Abstract

Finding landmark positions on facial images is an important step in face registration and normalization, for both 2D and 3D face recognition. In this paper, we inspect shortcomings of existing approaches in the literature and compare several methods for performing automatic landmarking on near-frontal faces in different scales. Two novel methods have been employed to analyze facial features in coarse and fine scales successively. The first method uses a mixture of factor analyzers to learn Gabor filter outputs on a coarse scale. The second method is a template matching of block-based Discrete Cosine Transform (DCT) features. In addition, a structural analysis subsystem is proposed that can determine false matches, and correct their positions.

Key words:

LOCALISATION ROBUSTE DE POINTS CARACTÉRISTIQUES SUR DES IMAGES FACIALES

Résumé

Trouver des points de repères est une étape importante pour l'enregistrement et la normalisation des visages et pour la reconnaissance 2D et 3D. Dans cet article, nous allons étudier les faiblesses des travaux existants dans la litterature et comparer plusieurs méthodes pour trouver automatiquement des points de repères sur des figures frontales dans des échelles différentes. Deux nouvelles méthodes furent employées pour analyser les traits faciaux dans des échelles granulaires et précises successivement. La première utilisant un mélange d'analyseurs factoriels pour procéder à des sorties de filtres de Gabor dans l'échelle granulaire. La deuxième utilise des modèles fondés sur des traits de la transformation en cosinus discrète bidimensionnelle (DCT). En plus, un sous-système d'analyse structurale a été proposé pour déterminer les points de repères non concordant et pour corriger leurs positions.

Mots clés :

1. Boaziçi University, Perceptual Intelligence Laboratory, Computer Engineering Department.
2. Bogaziçi University, Signal Processing Laboratory, Electrical and Electronic Engineering Department.

Contents

I. INTRODUCTION

Facial feature localization is a critical first step in many subsequent tasks, such as face recognition, pose normalization, expression understanding and face tracking. Although there is not yet a general consensus, the fiduciary or first tier facial features, most often cited in the literature, are the eyes or the eye corners, the tip of the nose and the mouth corners. Similarly, second tier features are the eyebrows, the bridge of the nose, the tip of the chin, etc. The importance of facial features stems from the fact that most face recognition algorithms in 2D and/or 3D rely on accurate feature localization. This step is critical not only directly for recognition techniques based on features themselves, but indirectly for the global appearance-based techniques that necessitate prior image normalization. For example, in 2D face recognition, popular recognition techniques, such as eigenfaces or Fisherfaces, are very sensitive to registration and scaling errors. For 3D face recognition, the widely used iterative closest point (ICP) registration technique requires scale-normalized faces and a fairly accurate initialization. In any case, both modalities require accurate and robust automatic landmarking.

Despite the increasing volume of related literature, facial feature localization is still an open problem. The existing localization techniques need improvements in accuracy on the one side and in their robustness in adverse conditions on the other side. Major factors handicapping facial feature localization can be listed as follows: The scale differences change the local geometry of features (e.g. curvatures), and the sample density may not always be adequate. Lighting variations and expressions cause non-linear effects on the value of face image pixels. Both in-plane and out-of-plane rotation of the head causes major changes in visual appearance. Self-occlusion and accessories such as hair, moustache and eyeglasses give rise to another gamut of problems. In 3D acquisition, non-reflecting surfaces cause outlier holes or protrusions. Finally, appearance of facial features has great variability across subjects.

Implicit in many landmarking techniques is the fact that the face has already been detected and that one may even have conducted a coarse localization that enables us to focus on the zones where the landmarks can potentially be found [13, 19, 20, 43]. Most registration or even coarse facial landmarking algorithms are guided by heuristics [2, 5, 9, 20, 22, 37, 43]. For example, in 3D, the biggest protrusion is usually taken as the tip of the nose; although, depending on the pose, chin or a streak of hair can be labeled erroneously as such [29]. In 2D, one uses vertical projection histograms to initialize the eye and mouth regions, where one assumes that the first and second histogram valleys correspond to the eye sockets

and lips, respectively [11, 12, 20, 26, 31, 37, 39, 47]. In the same vein, morphological processing of the face can be used to yield nose tip and eye sockets in 3D.

Another typical characteristic of facial landmarking is the serial search approach and the concomitant chicken-and-egg problem [2, 5, 20, 21] due to the iteratively executed normalization, registration and landmarking. For example, one often starts with one prominent landmark, say tip of the nose in 3D or eye depressions in 2D, and based on this ground-truth, proceeds to identify the other features [13]. A case in point is the mouth localization in 2D, where having first identified the eyes, one uses the perpendicular bisecting segment between the two eyes and the information of interocular distance to locate the mouth position [37]. Lastly, based on this middle-of-the-mouth reference point, the mouth corners are found.

As Brunelli and Poggio have stated [7], "features are only as good as they can be computed". Therefore, the desiderata for good facial feature localization can be stated as follows: i) Minimum possible dependence upon heuristics; ii) Robustness against pose, illumination, expression and scale variations as well as against occlusions and disturbance from facial accessories; iii) Being amenable to dynamic tracking; iv) High accuracy; v) Computational feasibility; vi) Applicability to data collected under different conditions.

In this work, we propose and compare two novel approaches for high-accuracy facial feature localization. We use low-level features to identify a host of possible coarse feature locations. The candidate locations are then sieved through a scheme that looks at the structural information among them, and then the locations are refined in coarse-to-fine iterations. These initial locations are refined by using independent Gabor features in one method and by using DCT coefficients in a second method. Our methods are based on 2D gray-level information.

The paper is structured as follows: In Section II, we review the literature on automatic facial landmarking. Section III introduces two novel facial feature localization algorithms proposed. Section IV. 1 summarizes the databases and the testing methodology. We present our simulation results in Section IV.2. Section V concludes and indicates future directions.

II. LITERATURE SURVEY ON FACIAL FEATURE DETECTION

II.1. Schemes Based on 2D Information

The various approaches in the literature for facial feature localization can be classified as appearance-based, geometric-based and structure-based. The majority of approaches use a preprocessing stage for initial coarse localization, using horizontal and vertical gray-level [4], or edge field projections, followed by some histogram valley detection filter [43]. A second commonality between methods is that most use a coarse-to-fine localization to reduce the computational load [2, 9, 13, 18, 22, 23, 31, 38, 40]. Some algorithms employ a skin colour-based scene segmentation to detect the face first [5, 11, 12, 13, 26, 47], and further colour segmentation for lip detection [27, 32].

Appearance-based approaches aim to find basis vectors to represent the face and its facial features. Examples of transformations used are principal components analysis (PCA) [1, 31], Gabor wavelets [18, 37, 38, 40], independent components analysis (ICA) [1], discrete cosine transform (DCT) [46] and Gaussian derivative filters [2, 19]. These transform features capture and model facial features under statistical variability when selected and processed with machine learning techniques like boosted cascade detectors [11, 12, 13], support vector machines (SVM) [1], and multi-layer perceptrons (MLP) [9, 31].

Geometric-based methods use prior knowledge about the face position, and constrain the landmark search by heuristic rules that involve angles, distances, and areas [37, 39, 46]. Structural information is an extension of geometric information that is used in validating localized features. For example, Wiskott et al. analyze faces with a graph that models the relative positions of fiducial points, and a set of templates for each landmark for feature response comparison [41, 42]. Other approaches that use the structural information in addition to local similarity enable more flexible graph transforms to represent displacements of fiducial point positions, and search these positions to maximize feature responses under landmark configuration constraints [13, 44]. For these models, the optimization process is plagued by local minima, which makes a good -and often manual- initialization necessary [6]. Table 1 summarizes the facial feature extraction methods in 2D face images.

II.2. Schemes Based on 3D Information

3D information is not commonly used in finding facial fiducial points, since 3D face imaging and handling of the resulting data volume are still not mainstream techniques. Furthermore outlier noise makes reliable processing difficult. However, when the scale of the faces are known, ICP can be used to register the face to a 3D template, thereby greatly constraining the possible locations for each facial landmark [24]. In [40], the method with 40-dimensional 2D Gabor jets [41] is extended to a 324-dimensional 3D jet method (36-dimensional point signatures from a 3×3 neighbourhood of each landmark). Colbry et al. employ surface curvature-based shape indices under geometrical constraints to locate features on frontal 3D faces [13]. Their method has been generalized to the multi-pose case with the aid of 2D information, such as the output of Harris corner detector on the gray-level information and related geometrical constraints. Conde et al. use SVM classifiers trained on spin images for a purely 3D approach [14]. As their proposed method requires great computational resources, they constrain the search for the landmarks by using apriori knowledge about the face. In [5], the 3D information plays a secondary or support role, in filtering out the background, and to compute intra-feature distances in geometry-based heuristics. In [16], 3D information is used to assist 2D in filtering out the background, and a comparison between 2D and 3D methods under relatively controlled illumination conditions indicates superiority of the 2D approaches. A summary of methods based on 3D information, totally or partially, is given in Table II.

TABLE I. – Summary of 2D facial landmarking methods.

Sommaire des methodes pour trouver des points de repères sur des images 2D.

Reference	Coarse Localization	Fine Localization
Chen, Zhang, Zhang [11]	Gaussian mixture based feature model + 3D shape model	
Cristinacce, Cootes, Scott [15]	Assumed given	Boosted Haar wavelet-like features and classifiers
Smeraldi, Bigun [38]	30 dimensional Gabor response of each point + SVM	Gabor responses of the complete retinal field + SVM
Feris, Gemmell, Toyama, Krüger [18]	Template matching using Hierarchical Gabor Wavelet Network (GWN) representation of faces	Template matching using Hierarchical Gabor Wavelet Network (GWN) representation of features
Lai, Yuen, Chen, Lao, Kawade [26]	Color segmentation (skin, lip) + edge map	Vertical projection of thresholded image obtained from the coarse level
Shakunaga, Ogawa, Oki [36]	PCA on canonical positions of features + structural matching	PCA
Ryu, Oh [31]	Vertical and horizontal projections of face edge map	PCA on coordinates of the feature edge map + MLP for template matching
Shih, Chuang [37]	Edge projections + geometric model of facial features	Not present
Arca, Campadelli, Lanzarotti [2]	Color segmentation (Skin and lip) + SVM	Geometrical heuristics
Zobel, Gebhard, Paulus, Denzler, Niemann [46]	Geometrical heuristics on DCT coded images + Probabilistic model, based on coupled structured representation of feature locations	Not present
Gourier, Hall, Crowley [19]	Gaussian derivatives (G_x, G_y, G_{xx}, G_{xy}, G_{yy}) + clustering to 10 centroids	Not present
Antonini, Popovici, Thiran [1]	Corner detection	Feature extraction using PCA and ICA projections of windows surrounding the corner points + SVM for template matching

TABLE II. – Facial landmarking methods that use 3D information.

Sommaire des méthodes qui utilisent l'information 3D.

Reference	Coarse Localization	Fine Localization
Boehnen, Russ [5]	Cascaded smoothing, minimum- and z-filtering + 2D and 3D geometry	
Colbry, Stockman, Jain [13]	Interpoint statistics + heuristics	Shape index + Harris edge detector
Conde et al.[14]	Curvature analysis + heuristics	Spin images + SVM
Irfanoglu, Gökberk, Akarun [24]	ICP based registration	Curvature – and surface normal-based heuristics
Çinar Akakin, Salah, Akarun, Sankur [16]	2D IMoFA-L + structural correction	IMoFA-L projection vs. DCT coefficients on 2D and 3D + SVM

III. TWO NOVEL FACIAL LANDMARKING ALGORITHMS

Similar to many other facial feature localization algorithms, our methods employ a two-stage coarse-to-fine landmarking approach: In order to build a computationally efficient system, we start by searching potential landmark zones on downsampled 2D images. The coarse landmarking is performed with the Gabor factor analysis algorithm described in Section III.1. Method 1 complements the coarse search by a structural correction scheme. The coarse level process is used as the first step in fine localization of landmarks. In Section III.2, we present Method 2, which achieves accurate landmarking by complementing the coarse localization by a method based on DCT features. These methods learn different feature models from the actual landmark neighborhoods during the training phase.

III.1. Algorithm-1: Gabor Factor Analysis and Structural Information

The first algorithm takes a generative modeling approach for landmark localization at a coarse stage. By modeling the distributions of local features with a state-of-the-art unsupervised model we achieve robust localization. Searching for each landmark independently protects the system against misleading local similarity values caused by missing or occluded landmarks. Figure 1 summarizes the landmarking scheme where a coarse localization is performed using the Gabor factor analysis algorithm, followed by a structural correction.

Gabor wavelets are often used in the literature to extract local discriminating features [28, 38]. The image is initially convolved with a Gabor kernel as below:

FIG 1. – Coarse landmark localization starts with computation of Gabor wavelets on downsampled images. Conspicuity maps for each landmark are computed by summing the likelihoods under mixtures of factor analyzers models. The most conspicuous locations are sent to the structural analysis subsystem for correction.

La localisation granulaire commence avec le calcul des ondelettes de Gabor sur la texture sous-échantillonnée. Pour chaque point de repère, une carte de saillance est obtenue par la somme des vraisemblances indiquée par les mélanges des analyseurs factoriels. Les points avec les meilleures vraisemblances sont passées au sous-système d'analyse structurale pour correction.

$$\Psi_j(\vec{x}) = \frac{\vec{k}_j \vec{k}_j^T}{\sigma^2} e^{\left(-\frac{\vec{k}_j \vec{k}_j^T \vec{x} \vec{x}^T}{2\sigma^2}\right)} \left[e^{(i\vec{k}_j \vec{x})} - e^{(\frac{-\sigma^2}{2})} \right]$$

$$\vec{k}_j = (k_{jx}, k_{jy}) = (k_v \cos \varphi_w, k_v \sin \varphi_w), k_v = 2^{-\frac{v+2}{2}} \pi, \varphi_w = w \frac{\pi}{8} \quad (1)$$

where $\vec{x} = (x, y)$ is the given pixel, $j = w + 8v$, and (w, v) defines the orientation and scale parameters of the Gabor kernels, respectively, and 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 function, known as the carrier. The term, $e^{-\sigma^2/2}$ in the square brackets compensates for the DC value.

III.1.1. Preprocessing

To reduce the computational burden, we downsample the high resolution face images. For the UND database, we have used a downsampling factor of eight (from the size 480×640 down to 60×80), which is a good compromise between precision and computational parsimony. An additional benefit of downsampling is the dramatic reduction in the number of candidate points to be tested, from approximately 300,000 to 1,500-2,000 with the addition of 3D masking. We use the 7×7 neighbourhood as a feature generation window, producing 49-dimensional feature vectors for each point at each orientation in the coarse scale (See Figure 5 for the size of the feature generation window).

The Gabor filter outputs are obtained in 8 orientations at a single scale. In the notation of Eq.1, this corresponds to scale $v \in \{3\}$ and orientations $w \in \{0, 1, 2, 3, 4, 5, 6, 7\}$. In the training phase, the landmarks u are manually labeled, and the resulting filter outputs are min-max normalized. Our simulations have indicated that using more scale parameters or larger windows do not contribute to the overall accuracy.

III.1.2. Gabor Factor Analysis

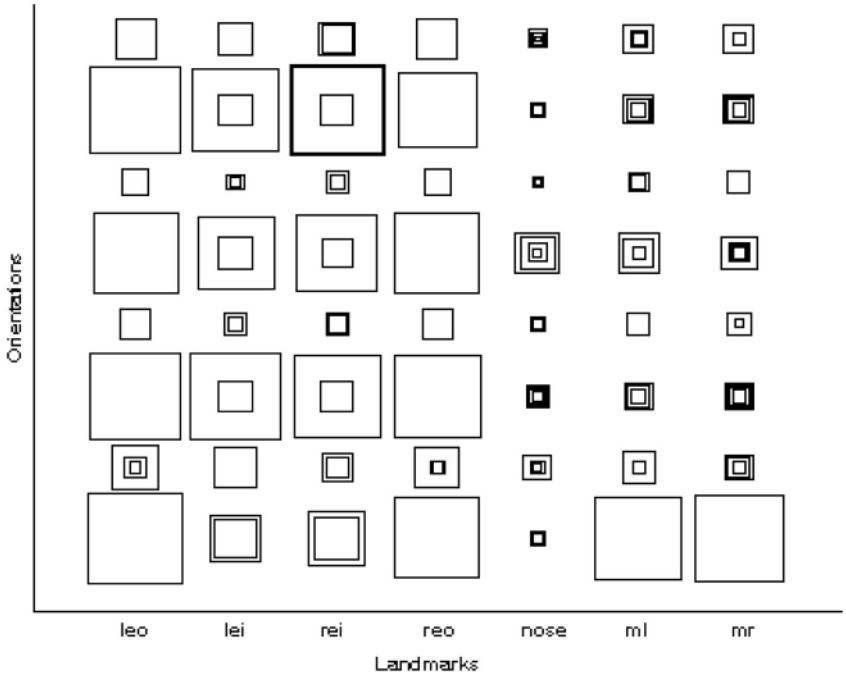

FIG 2. – This diagram shows the trained IMoFA-L model. There are 7×8 mixtures, one per landmark perorientation. The number of squares indicate the number of components, and the size of each square is proportional to the number of factors in that component.

Ce diagrame montre les modèles IMoFA-L. Il y a 7×8 melanges, un par point de repère et par orientation. Le nombre de carrés indiquant le nombre de composants, et la taille de chaque carré est proportionelles au nombre de facteurs dans le composant.

To model the distribution of these 49-dimensional features, we have employed an Incremental Mixtures of Factor Analyzers (IMoFA-L), which places a number of Gaussian distributions with arbitrary covariance on the data [34]. We opted for this method as Yang et al. have shown that a mixture of factor analyzers outperforms PCA in a face detection application [45]. The IMoFA-L model is convenient as it reduces the time and space complexity of the mixture model and it automatically finds a trade-off between accuracy and complexity (see Appendix A). Thus, complex patterns in the data are modeled with more components, and with more factors per component (e.g. mouth corners *ml* and *mr*), whereas simple patterns (e.g. nose) are modeled with smaller number of parameters. In Figure 2, the trained model is depicted graphically. If we look at the second orientation channel (second row of squares), the mixture for the outer corner of the left eye (leo) is a depicted as a single, large square, which means there is a single component with a large number of factors. The number of factors stands for the intrinsic dimensionality of the feature distribution, and a large number means there is variation in many directions. Conversely, if the number of components is

large, as in the left mouth corner (ml) for the same orientation, there exist a number of different structures, modeled as clusters in the feature space. IMoFA-L allocates these components and factors automatically.

Let u denote one of the directional Gabor feature vectors (49-dimensional in our case), that is, the $\theta' th$ Gabor filter output computed on the 49-points of a 7×7 neighborhood for any feature $1 = 1.7$. Without loss of generality, we leave out the feature subscript to simplify the equations. The mixture of Gaussians (MoG) model is written as:

$$p_\theta(u) = \sum_j p_\theta(u \mid G_j)\, p_\theta\,(G_j), \tag{2}$$

where θ subscript denotes one of the eight orientations, G_j is a Gaussian component characterized by a mean vector and a covariance matrix, and denoted by $N(\mu_j, \Sigma_j) \cdot p_\theta\,(G_j)$ is the prior probability of the component i along orientation θ, and $p_\theta(u \mid G_j)$ is the probability that a feature u is generated by that component i. Sahbi and Boujemaa have previously modeled salient features of faces with single full-covariance Gaussian distributions [33]. However, our simulations have shown that a mixture of Gaussians works better than a single full-covariance Gaussian.

Notice that Gabor orientations are treated separately to keep the overall complexity low, so that we forsake any covariance information between different orientations. The local analysis proceeds by computing the likelihood of each feature generation window for being one of the candidate features by summing the scores of the 8 orientations (each a separate mixture model). The highest likelihood score among all the visited positions indicates the landmark location, where the likelihood for a generic feature is given by:

$$L_{feature} = (u) = \prod_{\theta=1}^{8} \sum_j p_\theta(u \mid G_j)\, p_\theta\,(G_j) \tag{3}$$

This search over the face pixels is repeated for each one of the l features sought until one candidate is found for each feature. In our first tier scheme l is seven, composed of four eye corners, two mouth corners and the nose tip.

III.1.3. Correction via Structural Analysis

The refining stage of the method is based on the structural analysis that aims at detecting erroneous or missing landmarks and correcting them. In an earlier study, Burl and Perona have assumed that false alarms are distributed independently from each other and are independent from the feature location [8]. If this assumption is correct, a structural model that searches features at their expected locations will be able to single out false alarms. However, this is a simplifying assumption, and a more truthful model can take the correlations between false alarms (e.g. moustaches cause failures at both mouth corners) at the expense of being more complex. Burl and Perona have modeled the joint distribution of the landmark coordinates with a single multivariate Gaussian, and they base the affine correction of the face solely on the eye landmarks. However, their scheme fails if the eyes are incorrectly detected in the first place. Our system is more robust in that we do not assume the correctness of any particular landmark. Instead, subsets of landmarks are used to validate mutually the

location of other landmarks: that is, we have multiple sequential tests. The second difference is that we model the position of a landmark given the other landmarks, with a bivariate Gaussian distribution, making the complete model a mixture of bivariate Gaussians. The parameters of these Gaussians are learned during the training phase.

Traditional face graphs incorporate structural information in the framework of an energy minimization problem, where any deviation from the nominal distance between landmarks (conceptualized as edges of the face graph) is penalized. These approaches bring in a non-uniform energy gradient around the landmark, as perturbations of landmark locations affect the graph edge lengths in ways dependent upon the direction of the perturbation [13, 25, 42, 44]. Emphasizing a directionality for perturbations is meaningful if a large number of landmarks that follow common contour segments are modeled, as displacing a landmark in one direction makes the next landmark more likely to be displaced in that direction. However, for a small set of landmarks, this sort of constraint imposed by the face graph is not justified. Our simulations indicate that the Gaussian mixture model is very successful in modeling the landmark distributions.

In our proposed scheme, subsets of located landmarks take turns as *support sets.* For each such support set, we perform normalization involving translation, rotation and scaling with respect to the subset coordinates. The subsets are taken three at a time and then the rest of the landmarks are compared to their relative expected locations. A support set is validated if the ensemble of landmarks normalized vis-à-vis the chosen subset gives a high structural fitting score. Any incorrectly localized landmark in the support set will badly distort the positioning of the other landmarks during normalization, and result in a poor fitting score.

We learn the spatial distribution of the normalized landmarks for each possible support set, as each support set corresponds to a different normalization. If we have a support set size of i, and l landmarks in total, the number of possible support sets is $C(l, i)$, where $C(.)$ is the combination operator. For example, a support set of size three from within seven landmarks results in 35 support combinations. For each such combination, we model the distribution of the remaining landmark positions (after normalization) with a mixture of Gaussians. In the testing phase of a support set, the likelihood of the non-support feature locations is calculated.

In the normalization process, the centroid of the support set landmarks are first translated to the origin, then they are scaled so that the average distance of the support landmarks to the origin is fixed, and finally they are rotated to align one of the landmarks in the support set (arbitrarily selected to be the one with the smallest index) with the y-axis. The covariance matrix of this landmark will be singular after rotation, as it varies on a single dimension only. In order to remedy this situation, we use an additional minimization step (see Appendix B).

During testing of a support set, if at least one landmark outside the support set is *acceptable,* then the corresponding support set is accepted as correct. A non-support landmark j is assumed to be acceptable, if its likelihood under the model is higher than a fixed threshold:

$$L(l_j, \mu_j, \Sigma_j) > \tau(k) \tag{4}$$

We conceptualize this threshold as isodensity lines around its expected location, as illustrated in Figure 3. We heuristically determine the threshold in the following manner. The covariance can be visualised as an ellipsoid around the data distribution. We scale the covariance matrix by a scalar k, and obtain a larger ellipsoid (shown in Figure 3 and in Figure 7). To obtain the likelihood value on any point of this ellipsoid, we select an eigenvector of the

covariance matrix, and displace the mean in that direction by an amount proportional to the square-root of the corresponding eigenvalue. This gives us the threshold:

$$\tau(k) = L(\mu_j + v_j' \sqrt{u_j', \mu_i, \Sigma_j'}) \quad (5)$$

where Σ_j' is the scaled covariance matrix $\Sigma_j' = k^2 \Sigma_j$, u_j' is an eigenvalue of Σ_j', v_j' is the corresponding eigenvector, and L is the likelihood function, as defined previously. For our simulations we have chosen $k = 4$, although larger (but not smaller) values can be considered. Setting k to a smaller value means that the structural subsystem will label more points as outliers, and will be forced to re-estimate them. Conversely, if k is too large, some of the close outliers will be missed.

The non-support landmarks are labeled as acceptable or unacceptable according to their agreement with learned models under that support set. An unacceptable landmark, say the n'th landmark that yields a likelihood score below threshold, is replaced by the *backprojection* (i.e. reverse rotation, scaling and translation with respect to the local coordinate system) of the expected landmark location, μ_n. The support sets are searched until the non-support landmarks validate one of them (see Section IV.2.2 for a discussion of this stopping criterion).

The structural analysis subsystem can locate and correct one to three incorrect landmarks. Additionally, the landmarks that do not conform to any of the stored permutation patterns are labeled as failures. These are the cases where the local feature detector fails for some reason (e.g. because of a moustache), and it is important to be able to detect such failures as well. This part of our work is usable as a post-processing block with any landmarking method for failure detection and correction.

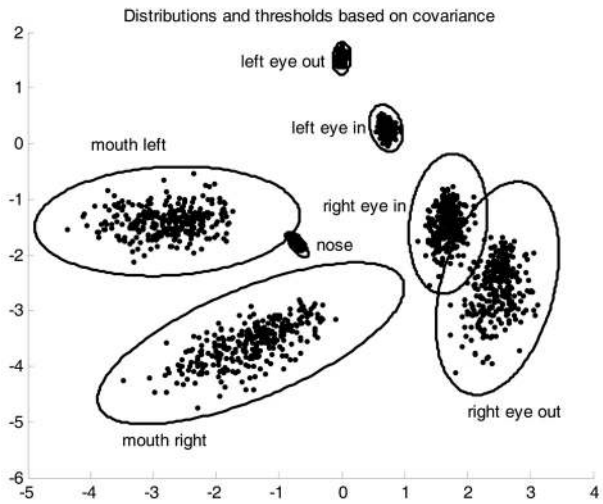

FIG 3. – Thresholds for outlier detection are shown as ellipsoids around landmark clusters. The support set consists of the corners of the left eye and the nose. As the normalization is based on the landmarks of the support set, they have smaller variations.

Les seuils de détection des valeurs abérrantes sont montrés comme des elli psoides autour de clusters de points de repères. L'ensemble de support consiste par des coins de l'œil gauche et le point de nez. Ces points ont moins des variations parce que la normalisation est basée sur eux.

III.1.4. Fine Localization

For the fine level (i.e. 480×640 images), a window around the coarse landmark location is searched for the best candidate. We have used another batch of IMoFA-L mixtures to compute the best candidate, trained on patches of fine-resolution images. For the UND database, we have selected a 9×9 search window. Note that since the upsampling factor is eight, a 9×9 window essentially corresponds to searching an area 10% larger than the corresponding coarse-level pixel found by the coarse landmarking scheme. We have evaluated window sizes up to 41×41; and observed that window sizes larger than 9×9 deteriorate performance. For the more challenging BANCA dataset, the accuracy peaked for a larger search window.

III.2. Algorithm-2: DCT-based Facial Feature Extraction

In our second approach, instead of using the Gabor wavelets, we have used comparatively low-frequency DCT features [17, 30]. In fact, DCT coefficients can also capture the statistical shape variations and can be a faster alternative for facial landmark detection, compared to local Gabor feature analysis.

III.2.1. Preprocessing

The coarse level process described in Section III.1 is used as the first step. An area of size 9×9 around the coarse landmark is probed by cutting 15×15 feature patches for each point in the area. DCT coefficients are extracted from these patches and analyzed on a block-by-block basis. Given an image block from the search window $f(x, y)$, where y, $x = \{0, 1, \ldots, K\text{-}1\}$, we decompose it in terms of orthogonal 2-D DCT basis functions. The result is a matrix $C(v, u)$ containing DCT coefficients:

$$C(v, u) = a(v)\ \alpha(u) \sum_{y=0}^{K-1} \sum_{x=0}^{K-1} f(y, x)/\beta(y, x, v, u) \text{ for } v, u = 0, 1, \ldots, K\text{-}1, \tag{6}$$

$$\text{where } \alpha(v) = \begin{cases} \sqrt{\dfrac{1}{K}} \text{ for } v = 0 \\ \sqrt{\dfrac{2}{K}} \text{ for } v = 1, 2, \ldots, K-1 \end{cases} \quad \text{and } \beta(y,x,v,u) = \cos\left[\frac{(2y+1)v\pi}{2K}\right] \cos\left[\frac{(2x+1)u\pi}{2K}\right]$$

The DCT coefficient values can be regarded as the relative presence of 2D spatial patterns contained in the visited locality. Once the coefficients of the DCT output matrix are computed, they are re-organized by a zigzag scanning pattern. This selection favours the highenergy, low-frequency coefficients, in agreement with the amount of information stored in them. The first coefficient (DC value) is removed, since it only represents the average intensity value of the block. The remaining (AC) coefficients denote the intensity changes or gray-level shape

variations over the image block. Those in the upper diagonal (119 coefficients) are chosen and z-normalized to form the feature vector.

III.2.2. DCT Feature Matching

In the training phase, k-means clustering is performed on the z-normalized DCT coefficients to obtain eight codebook vectors separately for each facial landmark. During testing, when the coarse locations of fiducial points of a test image are given, z-normalized 15×15 DCT coefficients are calculated for $9 \times 9 = 81$ neighbourhood points of the coarse location. For each of the 81 candidate points, the Euclidean distances to eight cluster centroids are computed. The minimum Euclidean distance is assigned as the matching score. Within the search neighbourhood, the candidate with the best matching score (i.e. the minimum distance to any centroid) is selected as the fine-tuned facial feature location.

IV. EXPERIMENTAL RESULTS

IV.1. Databases and Testing Methodology

We have employed the University of Notre Dame (UND) database in our experiments (Sections IV.2-IV.4) [10] and a subset of the BANCA dataset for a more challenging set of experiments (Section IV.5) [3]. The UND data consists of 942 2D images at 480×640 resolution. The ground truth is created by manually landmarking the seven points: That is, four eye corners, nose tip, mouth corners. The data are randomly split into training (707 samples), and test sets (235 samples). A validation set of 235 samples is separated from the training set to tune the model parameters whenever necessary. The face is assumed to be roughly localized. For UND, we use the associated depth mask for this purpose, which eliminates most of the background, but not the shoulders and the neck.

To measure the performance of the feature localizers, we have used two criteria:

a) The feature localization error as a function of the search basin size, the latter centered on the true landmark. As the extent of the search basin increases, the computational effort increases quadratically while the chances of finding the landmark increases. This measure is used for the coarse stage.

b) The successful landmark detection performance. A landmark is considered correctly detected if its deviation from the true landmark position is less than a given threshold, called the *acceptance threshold.* We use relative distances, where all distances are normalized to the inter-eye distance. This measure is used for the fine stage.

A good feature detector should have high accuracy and should converge to the true feature starting from any point within a large basin of attraction. If the correct localization probability flattens out after an initial monotonic growth with the increasing size of the search window, this indicates the robustness of the scheme. We contrast our method with two methods for facial feature localization: Lades et al. [25] and Wiskott et al. [41]. Lades et al. have employed Gabor wavelet jets, with five scales and eight different orientations, resulting in 40-dimensional feature vectors overall. Wiskott et al. also use Gabor jets, but extend this idea by including the phase information of Gabor jets. Both methods store the features extracted from the training set as bunch models. During testing, features extracted from the test image are compared with all templates in the bunch, and the minimum distance to any template is used for candidate selection.

IV.2. Performance of the Coarse Localization Method

V.2.1. The Gabor Factor Analysis Approach

For the Gabor factor approach as explained in Section III.1, we downsampled the images of the UND dataset down to size 60×80, and then, obtained mixture distributions of Gabor features. The test images were similarly processed. The likelihood of pixels to correspond to a facial landmark is computed using Eq.3 of Section III. 1.2. This likelihood is called the feature conspicuity map, and is obtained by summing the eight likelihood maps, one for each orientation. The per-orientation conspicuity maps for the left eye outer corner are illustrated in Figure 4, while the seven inset images in Figure 5(b) show the total (summed) conspicuity maps for each of the seven landmarks. For any one of the landmarks, the peak of the summed conspicuity map is taken as the identified facial landmark. All the identified landmarks are pictured in Figure 5(a).

FIG 4. – Likelihood based conspicuity maps obtained from IMoFA-L outputs for the outer corner of the left eye. Lighter locations indicate higher values.

Des cartes de saillance obtenues par l'application de IMoFA-L pour trouver les vraisemblances au coin extemale de l'oeil gauche. Les valeurs élevées sont plus claires.

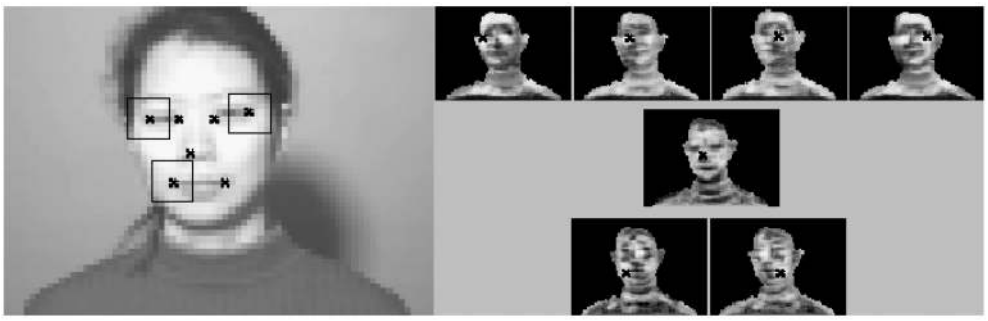

FIG 5. – a) Left image: The correctly located landmarks on the downsampled intensity image; 7×7 neighbourhoods are shown. b) Right inset images: The total conspicuity maps for each landmark. Points lacking depth are trimmed. The landmarks correspond to to global maximum of the respective maps.

a) Image gauche: Les points de repères correctement localisés sur la texture sous – échantillonée; Les voisinages sont montrés avec des boîtes 7×7. b) Images droites: Les cartes de saillance pour chaque point de repère. Les points manquant en profondeur ne sont pas traitées. Les points trouvés sont des maxima globaux.

Figure 6 shows the localization mean square error (in pixels) versus search window size in the reduced resolution face images. Note that one pixel error shown on this figure (coarse localizations) corresponds to eight pixels in the high-resolution images. The IMoFA-L coarse localization scheme, computed over 7×7 windows, outperforms the Wiscott et al. and Lades et al. local feature analysis schemes explained in Section IV. 1.

The most plausible explanation for the relative success of our method is that both bunch-based methods are template matching methods, while the IMoFA-L is a generative method that models the feature distribution probabilistically. Also, it should be noted that increasing the number of training samples is beneficial for the IMoFA-L method, but not recommended for the bunch methods, as the increased number of comparisons linearly increase computation time for only marginal improvement [41]. This observation was validated by our simulations.

IV.2.2. Structural Analysis

The structural analysis subsystem can be viewed as a hypothesis testing labeling step. Given the seven landmark candidates, the role of the structural analysis is to validate these candidates based upon configurational information. The permutations of landmark triplets are tested in descending order given by the product of reliabilities of landmarks in the corresponding support sets. Our simulations have shown that three is the optimum number of landmarks for support set size.

Once a support set is found to validate at least one other landmark, we accept it as correct, and use it to re-estimate the invalid landmarks: that is, those that did not survive the likelihood thresholding. With this scheme, the average number of support sets we need to test is 1.2 in UND, and 3.0 in the BANCA dataset. Conversely, we can evaluate all the support sets and select the best. This resulted in two per cent accuracy increase for the coarse localization, with acceptable extra computational load.

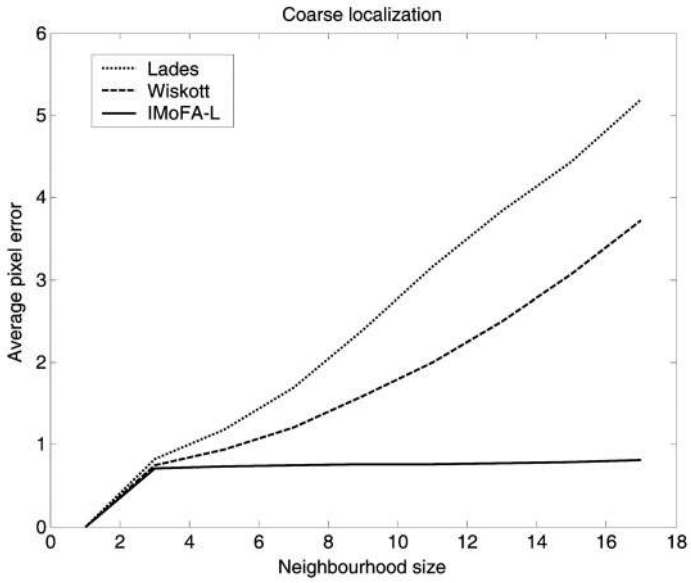

FIG 6. – Coarse localization errors (in pixels) for increasing sized search basins around true landmarks. When the search basin is large, IMoFA-L based similarity does not deviate from the landmark as the Gabor jet similarity metrics proposed in [25] and [41].

L'erreur moyenne (en pixelles) contre la taille de voisinage autour des points de repères réels sur les images granulaires. Pour un grand voisinage, IMoFA-L est plus consistante que les méthodes basées sur les jets de Gabor, proposées par [25] et [41].

For 235 test samples, the structural analysis re-estimates 120 landmark locations, and correctly tags five samples as poorly labeled (i.e. more than two landmarks out of seven are misplaced). The average localization error was shown to be around one pixel (corresponds to eight pixels in the original resolution) for neighbourhood sizes up to 18 pixels in Figure 6. The coarse landmarking is performed over the whole image, and the average localization error for all landmarks is 1.54 pixels. The usefulness of the structural analysis is proven when it reduces this error to 0.52 pixels via backprojections.

Figure 7 shows the coarse localization result on a sample image, where the structural subsystem successfully determines that the mouth corners are incorrectly labeled. The backprojection remedies this situation by asserting correctly the expected locations.

We tested the subsystem further by supplying it with systematically disrupted landmark information. We have given the system first one and then two incorrect landmarks that deviate from their true locations by at least 10 pixels. The structural subsystem was able to detect and correct 99 per cent of the cases.

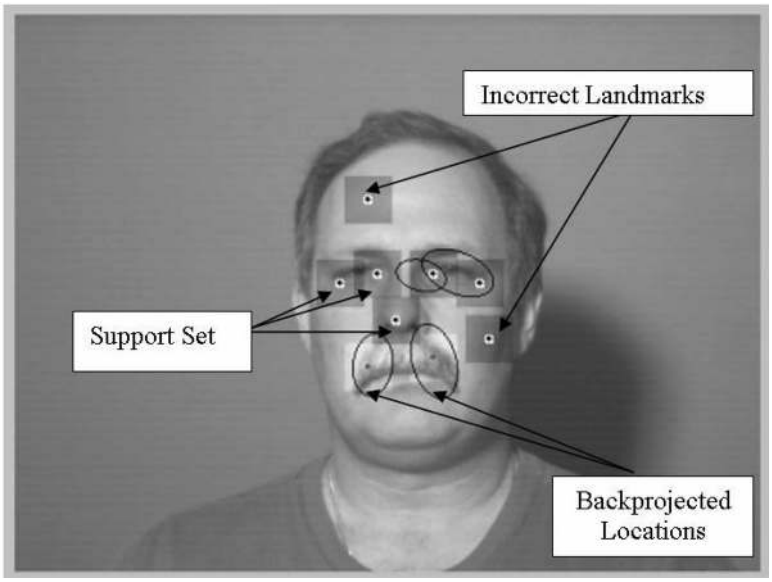

FIG 7. – The coarse localization results projected to the original image. The gray boxes are 41×41 neighbourhoods. The ellipsoids indicate the expected locations for each landmark outside the support set. Landmarks in the support set, detected incorrect landmarks and their backprojected coarse locations are shown.

Les points trouvés par le système granulaire sur l'image originale. Leurs voisinages sont montrés avec des boîtes grises 41×41. Les ellipsoïdes indiquent les locations attendues pour des points en dehors du set de support. Les points dans le set de support, les point incorrects, et leur positions corrigées sont indiqués.

IV.3. Fine Landmark Localization

The second tier of the algorithm consists in refining the landmark localization based upon the coarse homing result of the first tier. We revert thus to the original resolution of images, that is, 480×640. For each landmark, we search a 9×9 window around the coarse localization result with two different second tier methods.

The refining with the DCT method consists in searching at and around the zones indicated by the first tier landmarks. New landmark positions are searched using the DCT templates. The refinement with the IMoFA-L method is similar to the first tier procedure.

In Figure 8, it is observed that the IMoFA results are not as successful as they were in the coarse model. The reason is that the patches we have used in high-resolution images do not contain sufficiently rich structural information to learn reliable local models. Increasing patch size, on the other hand, increases the dimensionality of the problem, thus necessitating either substantially many more training samples for learning, or various constraints to guide learning.

We evaluate the proposed methods and the competitor algorithms on the high resolution images. As localization accuracy depends on the landmark type, we compare the correct localization percentages for each landmark type separately. A landmark is assumed to be correctly localized, if it has a pixel distance smaller than a threshold, whose values are given in the horizontal axis. The IMoFA method performs better than the baseline, but the DCT method seems to perform the best (See Figure 8). The method with Lades et al. similarity measure is not shown as it is always worse than Wiskott et al.'s measure. The mouth corners are harder to localize in general.

Figure 9 shows a more detailed analysis of the DCT method. Assuming that five per cent of the inter-eye distance (calculated from interpolated eye-centres) is an acceptable threshold for the deviation of the located landmark from the ground truth, and deeming three correct landmarks as sufficient (as in [13]), we have 97.5 per cent correct localization. For a 10 per cent threshold, we have 99.6 per cent correct localization. This is comparable to 99.8 per cent detection reported in [13] and 99.6 in [5] for the same data set. We believe our method is more robust, as we do not employ feature-specific heuristics. For instance in [13], the tip of the head is used to constrain the search area for the nose. However, locating the tip of the head can be even more difficult than locating the nose, because the hair is difficult to segment.

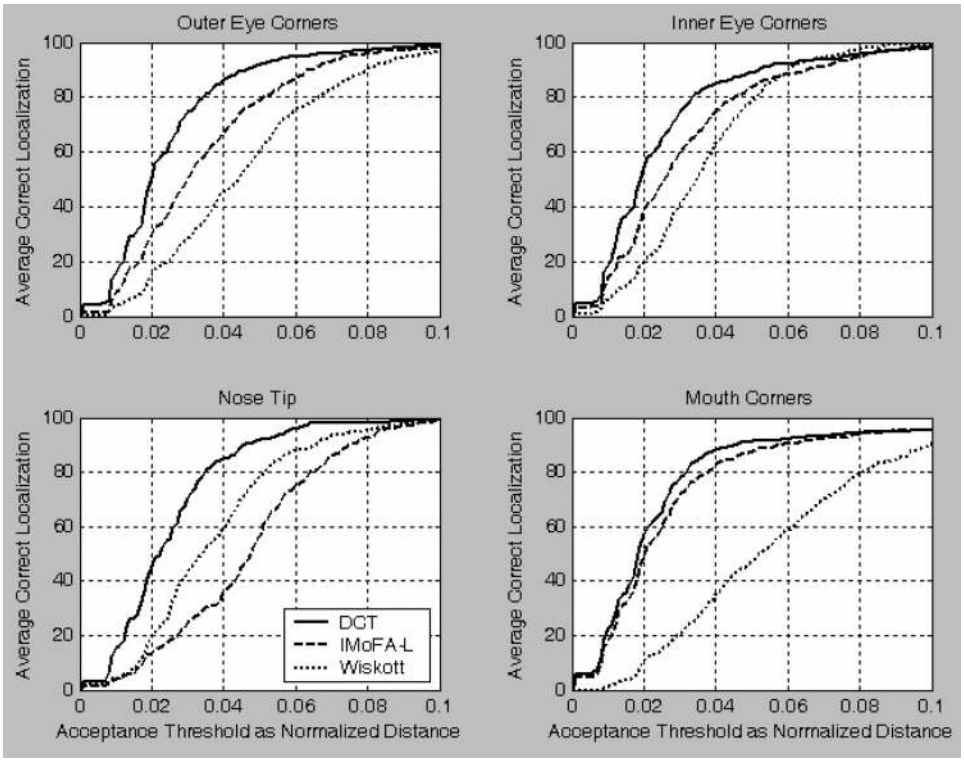

FIG 8. – Comparison of proposed methods for fine localization, around a 9×9 neighbourhood of the coarse landmarks, shown separately for each landmark type

Les comparaisons des méthodes proposées pour la localisation exacte dans un voisinage 9×9 autour des points granulaires, montrées séparément pour chaque type de points.

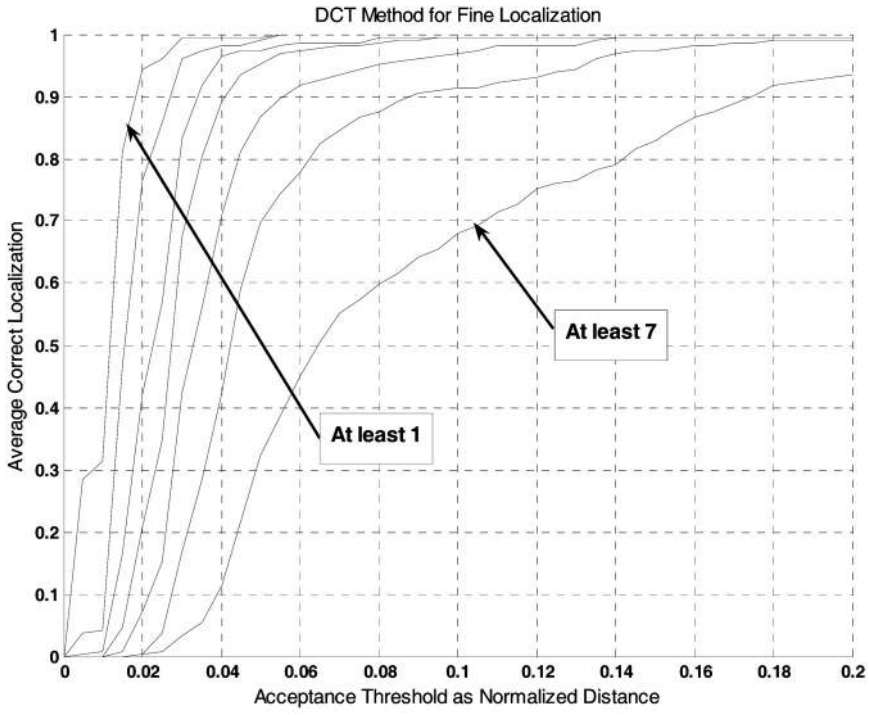

FIG 9. – The n[th] plot from the left shows the correct localization percentage, if we assume that locating *n* landmarks (out of seven) are sufficient.

La 1eme ligne montre le pourcentage de la localisation exacte, avec la supposition de trouver un nombre « n » de points (de sept point des repères en total) est suffisante.

IV.4. Running Time

To give a rough idea of the temporal complexity of different parts of the algorithm, we report running times on a computer with a 587 MHz Pentium M processor and 512 Mb RAM using non-optimized MATLAB 6.5 code. In the coarse localization part, the Gabor wavelet coefficients extraction on the 60×80 image takes 0.06 seconds for one landmark in eight orientations. Then, the computation of a single conspicuity map under the IMoFA-L model takes approximately 0.6 seconds. This is an average value, as the complexity of the model changes depending on the number of factors and components. The computation of all the conspicuity master maps takes $0.6 \times 8 \times 7 = 33.6$ seconds. The structural correction of a single test sample takes 0.3 seconds on the average. The computation of DCT coefficients for a single 9×9 window ($81 \times 25 \times 25 = 50.625$ coefficients) takes about 5 seconds, which is indicative of the computation speed of MATLAB. The template matching part for eight codebook vectors takes only 0.05 seconds.

IV.5. Experiments in Challenging Environments

To test the applicability of our method in a more challenging environment, we have performed additional tests on a subset of the English part of the BANCA dataset [3]. This dataset contains significant illumination, pose, and background and face clutter variations, such as eyeglasses and hair, and presents major challenges for feature localization. From the BANCA dataset, three representative sessions with different environmental conditions were selected. 50 subjects from each session contributed one image for training, one image for validation and one image for the test set.

For the coarse level, all parameter settings were retained. Images are downsampled by eight (from to 576×720 to 72×90) and eight Gabor channels are used. Search area is the coarsely located face. Figure 10 shows coarse localization examples on the downsampled faces for three sessions that increase in difficulty. It is observed that when eyeglasses occlude the eye corners, facial landmarks cannot be recovered.

FIG 10. – Samples from sessions 1, 5, and 10 of the English part of the BANCA dataset. The crosses are initial detections, the dots are final landmarks after structural correction. Facial hair as in sample (c) and eyeglasses as in samples (f) and (i) make detection difficult.

Les examples de la 1re, 5e et 10e session de la partie anglaise de jeu de données BANCA. Les croix sont des points initiaux, les points sont les points des repères finaux apres correction structurelle. La barbe (comme dans (c)) et les lunettes (comme dans (f) et (i)) rendent la détection difficile.

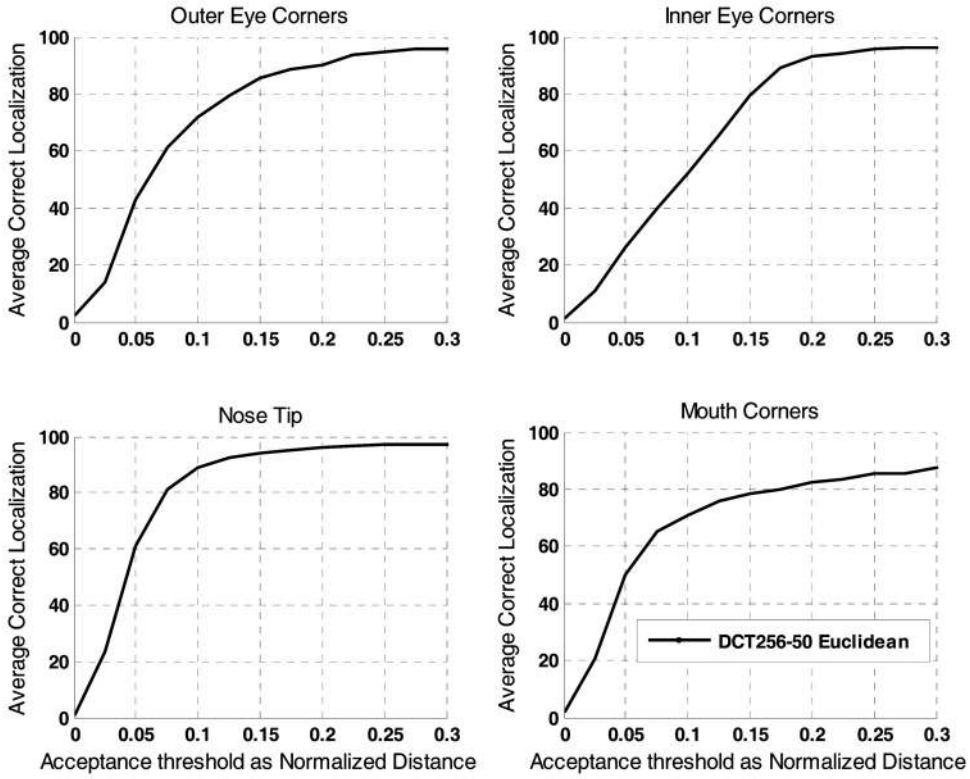

FIG 10. – Performance of DCT method for fine localization, around a 16×16 neighbourhood of the coarse landmarks, shown separately for each landmark type. In the legend DCT 256-50 means that 50 DCT coefficients out of 256 are chosen.

Figure 11. Les résultats de la méthode DCT pour la localisation exacte, dans un voisinage 16×16 autour des points granulaires, montrées séparément pour chaque type de points. DCT 256-50 veut dire que 50 coefficients de 256 sont choisis.

In order to obtain the fine localization results of BANCA database, we used the k-means classifier as mentioned in section III.2.2. Figure 11 shows the fine localization results for individual landmarks. It is observed that both eye corner and mouth corner localization results suffer due to occlusions and large pose changes.

For a five percent threshold and three correct landmarks, we have 75.5 per cent correct localization. For a 10 per cent threshold, we have 94 per cent correct localization. The drop in the accuracy is the partly attributed to the presence of eyeglasses, but also to illumination and pose changes. The lack of 3D information for eliminating the background results in an increase in the number of false positives.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced and tested a two-tier method of facial landmarking. The coarse level features are extracted by Gabor wavelets, and then each facial feature is learned by a separate model. The configurational information of the landmarks is also integrated into the localization method. The robustness and satisfactory performance of the coarse landmarking are due in part to the elimination of outlier positions with configurational information, and partly due to observing the image with 56×56 windows (7×7 after downsampling), which are sizes commensurate with the facial features.

The structural subsystem is a used as a post-processing block to detect and correct localization failures. We should also remark that the proposed method is not necessarily specific to faces, but is applicable to any other objects and defined feature sets. One byproduct of the structural analysis is that it can roughly locate facial features that are not rich in texture, like cheeks, chin, or features under total occlusion of a facial accessory.

An important aspect of our approach is its independence from any heuristics or initialization of the landmark points. Examples of heuristics are the detection of hair ceiling line and then extrapolating the midline going through the nose, or to find depressions of the vertical projection for mouth and eye socket candidates. While heuristics seem to simplify the problem, they actually push the problem back one level without solving it.

The output positions of the coarse stage are fed into the fine localization stage. This stage, using a DCT-based scheme, finds more accurate feature positions. In the final analysis, the proposed method outperforms its competitor methods in the literature ([25] and [42]).

One possible extension of our method is to 3D facial data (See [35]). The face information can be given in terms of depth information on a lattice or as 3D point cloud. The usefulness of 3D face information emerges when the 2D information is compromised (e.g. under severe illumination distortion [5]), and in assisting the 2D scheme which has to operate with inadequate information. A case in point is the nose tip, which is not particularly rich in texture information [40]. Our preliminary studies indicate that 3D provides for a marginal localization improvement over 2D methods. Thus the role of 3D depth information is limited to eliminating background clutter and to localizing some landmarks better (e.g. the nose).

The BANCA database, with pose difficulties and background clutter indicated some of the limitations of the facial feature localizer. Thus, the scheme needs to be further improved against pose self-occlusions.

VI. ACKNOWLEDGEMENTS

This work is supported by TUBITAK project grant no: 104E080 and FP6 NoE project Biosecure.

Manuscrit reçu le
Accepté le

REFERENCES

[1] ANTONINI (G.), POPOVICI (V.), THIRAN (J.P.), Independent Component Analysis and Support Vector Machine for Face Feature Extraction, *4th Int. Conf on Audio – and Video-Based Biometric Person Authentication,* pp. 111-118, Guildford, UK, 2003.

[2] ARCA (S.), CAMPADELLI (P.), LANZAROTTI (R.), A face recognition system based on automatically determined facial fiducial points, *Pattern Recognition,* 39(3), pp. 432443, 2006.

[3] The BANCA dataset – English part; http://banca.ee.surrey.ac.uk/.

[4] BASKAN (S.), BULUT (M.M.), ATALAY (V.), Projection Based Method for Segmentation of Human Face and its Evaluation, *Pattern Recognition Letters, 23,* pp. 1623-1629, 2002.

[5] BOEHNEN (C.), RUSS (T.), A Fast Multi-Modal Approach to Facial Feature Detection, *7th IEEE Workshop on Applications of Computer Vision*, pp. 135-142, Breckenridge, USA, 2005.

[6] BOLME (D.S.), Elastic Bunch Graph Matching, unpublished MS thesis, Colorado State University, 2003.

[7] BRUNELLI (R.), POGGIO (T.), Face Recognition: Features versus Templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15, pp.* 1042-1052, 1993.

[8] BURL (C.), PERONA (P.), Recognition of Planar Object Classes, in *Computer Vision and Pattern Recognition Conference*, pp. 223-230, San Francisco, USA, 1996.

[9] CAMPADELLI (P.), LANZAROTTI (R.), Localization of Facial Features and Fiducial Points, in *Proc. IASTED Int. Conf Visualization, Imaging & Image Processing,* Malaga, Spain, 2002.

[10] CHANG (K.I.), BOWYER (K.W.), FLYNN (P. J.), Multi-modal 2D and 3D Biometrics for Face Recognition, in *IEEE Workshop on Analysis and Modeling of Faces and Gestures*, pp. 187-194, Nice, France, 2003.

[11] CHEN (L.), ZHANG (L.), ZHANG (H.), ABDEL-MOTTALEB (M.), 3D Shape Constraint for Facial Feature Localization Using Probabilistic-like Output, in *6th IEEE Int. Conf on Automatic Face and Gesture Recognition,* pp.302-307, Seoul, Korea, 2004.

[12] CHEN (L.), ZHANG (L.), ZHU (L.), LI (M.), ZHANG (H.), A Novel Facial Feature Localization Method Using Probabilistic-Like Output, *Asian Conference on Computer Vision,* Jeju Island, Korea, 2004.

[13] COLBRY (D.), STOCKMAN (G.), JAIN (AK.), Detection of Anchor Points for 3D Face Verification, in *Proc. IEEE Workshop on Advanced 3D Imaging for Safety and Security, A3DISS,* San Diego, USA, 2005.

[14] CONDE (C.), SERRANO (A.), RODRIGUEZ-ARAGÔN (L.J.), CABELLO (E.), 3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification, *IEEE Conf Computer Vision and Pattern Recognition, pp.* 115, 2005.

[15] CRISTINACCE (D.), COOTES (T.), SCOTT (I.), A multi-stage approach to facial feature detection, in *Proc. British Machine Vision Conference*, pp. 23 1-240, 2004.

[16] ÇINAR AKAKIN (H.), SALAH (A. A.), AKARUN (L.), SANKUR (B.), 2D/3D Facial Feature Extraction, in *Proc. SPIE Conference on Electronic Imaging*, pp. 441-452, San Jose, USA, 2006.

[17] EKENEL (H. K.), STIEFELHAGEN (R.), Local Appearance Based Face Recognition Using Discrete Cosine Transform, *13th European Signal Processing Conf.,* Antalya, Turkey, 2005.

[18] FERIS (R. S.), GEMMELL (J.), TOYAMA (K.), KRÜGER (V.), Hierarchical Wavelet Networks for Facial Feature Localization, in *5th IEEE Int. Conf on Automatic Face and Gesture Recognition*, pp. 125-130, Washington DC, USA, 2002.

[19] GOURIER (N.), HALL (D.), CROWLEY (J.L.), Facial Features Detection Robust to Pose, Illumination and Identity, in *IEEE Int. Conf on Systems, Man and Cybernetics*, 1(10-13), pp. 617-622, 2004.

[20] GU (H.), SU (G.), DU (C.), Feature Points Extraction from Faces, *Conf on Image and Vision Computing*, pp. 154-158, New Zealand, 2003.

[21] GUNDUZ (A.), KRIM (H.), Facial Feature Extraction Using Topological Methods, in *IEEE Int. Conf on Image Processing,* 1, pp. 673-676, Barcelona, Spain, 2003.

[22] HERPERS (R.), MICHAELIS (M.), LICHTENAUER (K.-H.), SOMMER (G.), Edge and Keypoint Detection in Facial Regions, in *2»' Int. Conf. on Automatic Face and Gesture Recognition*, pp. 212-217, Killington, USA, 1996.

[23] HERPERS (R.), SOMMER (G.), An Attentive Processing Strategy for the Analysis of Facial Features, in WECHSLER (H.) et al. (eds.), *Face Recognition: From Theory to Applications*, pp. 457-468, Springer, ASI Series, 1998.

[24] IRFANOĞLU (M.O.), GÖKBERK (B.), AKARUN (L.), 3D Shape-Based Face Recognition Using Automatically Registered Facial Surfaces, in *Int. Conf of Pattern Recognition*, 1, pp. 183-186, Cambridge, UK, 2004.

[25] LADES (M.), VORBRUGGEN (J.), BUHMANN (J.) LANGE (J.), VON DER MALSBURG (C.), WURTZ (R.), KONEN (W.), Distortion Invariant Object Recognition in the Dynamic Link Architecture, *IEEE Transactions on Computers,* 42, pp. 300311,1993.

[26] LAI (J. H.), YUEN (P.C.), CHEN (W.S.), LAO (S.), KAWADE (M.), Robust Facial Feature Point Detection under Nonlinear Illuminations, in *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 168-174, Vancouver, Canada, 2001.

[27] LIEW (A.W.-C.), LEUNG (S. H.), LAU (W. H.), Segmentation of Color Lip Images by Spatial Fuzzy Clustering, *IEEE Transactions on Fuzzy Systems*, 11, pp. 542-549, 2003.

[28] LIU (C.), WECHSLER (H.), Independent Component Analysis of Gabor Features for Face Recognition, *IEEE Transactions on Neural Networks,* 14, pp. 919-928, 2003.

[29] LU (X.), JAIN (A.K.), Multimodal Facial Feature Extraction for Automatic 3D Face Recognition, Technical Report MSU-cse-05-22, Michigan State University, 2005.

[30] PAN (G.), WANG (Y.), WU (Z.), Pose-invariant Detection of Facial Features from Range Data, in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, Washington DC, USA, pp. 4171-4175, 2003.

[31] RYU (Y. S.), OH (S.Y.), Automatic Extraction of Eye and Mouth Fields from a Face Image Using Eigenfeatures and Ensemble Networks, *Applied Intelligence*, 17, pp. 171185, 2002.

[32] SADEGHI (M.), KITTLER (J.), MESSER (K.), Modelling and Segmentation of Lip Area in Face Images, *IEEE Vision, Image, and Signal Processing,* 149, n° 3, pp. 179-184, 2002.

[33] SAHBI (H.), BOUJEMAA (N.), Robust Face Recognition Using Dynamic Space Warping, in Tistarelli (M.), Bigun (J.), Jam (A.K.) (eds.), *Biometric Authentication*, *LNCS 2359*, pp. 121-132, Springer-Verlag, Berlin, Heidelberg, 2002.

[34] SALAH (A.A.), ALPAYDIN (E.), Incremental Mixtures of Factor Analyzers, in *Int. Conf on Pattern Recognition,* 1, Cambridge, UK, pp. 276-279, 2004.

[35] SALAH (A.A.), AKARUN (L.), 3D Facial Feature Localization for Registration, in Günsel et al. (eds.), *Int. Workshop on Multimedia Content Representation, Classification and Security, LHCS* vol. 4105/2006, İstanbul Turkey, pp. 338-345, 2006.

[36] SHAKUNAGA (T.), OGAWA (K.), OKI (S.), Integration of Eigentemplate and Structure Matching for Automatic Facial Feature Detection, in *3' In t. Conf on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 94-98, 1998.

[37] SHIH (F.Y.), CHUANG (C.), Automatic Extraction of Head and Face Boundaries and Facial Features, *Information Sciences*, 158, pp. 117-130, 2004.

[38] SMERALDI (F.), BIGUN (J.), Retinal Vision Applied to Facial Features Detection and Face Authentication, *Pattern Recognition Letters*, 23, pp. 463-475, 2002.

[39] SOBOTTKA (K.), PITAS (I.), A Fully Automatic Approach to Facial Feature Detection and Tracking, in BIGUN (J.), CHOLLET (G.), BORGEFORS (G.) (eds.), *Audio- and Video-based Biometric Person Authentication*, LNCS, 1206, Springer Verlag, pp. 77-84, 1997.

[40] WANG (Y.), CHUA (C.), HO (Y.), Facial Feature Detection and Face Recognition from 2D and 3D Images , Pattern *Recognition Letters, 23,* n°10, pp. 1191-1202, 2002.

[41] WISKOTT (L.), FELLOUS (J.-M.), KRUGER (N.), VON DER MALSBURG (C.), Face Recognition by Elastic Bunch Graph Matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, n° 7, pp. 775-779, 1997.

[42] WISKOTT (L.), FELLOUS (J.-M.), KRUGER (N.), VON DER MALSBURG (C.), Face Recognition by Elastic Bunch Graph Matching, in Jam (L.C.) et al. (eds.), *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, pp. 355-396, CRC Press, 1999.

[43] WONG (K.), LAM (K.), SIU (W.), An Efficient Algorithm for Human Face Detection and Facial Feature Extraction under Different Conditions, *Pattern Recognition*, 34, pp. 1993-2004, 2001.

[44] XUE (Z.), LIB (S.Z.), TEOH (E.K.), Bayesian Shape Model for Facial Feature Extraction and Recognition, *Pattern Recognition*, 36, pp. 2819-2833, 2003.

[45] YANG, (M.), AHUJA (N.), KRIEGMAN (D.), Face Detection Using Mixtures of Linear y Subspaces, in *4th Int. Conf on Automatic Face and Gesture Recognition*, pp. 70-76, Grenoble, France, 2000.

[46] ZOBEL (M.), GEBHARD (A.), PAULUS (D.), DENZLER (J.), NIEMANN (H.), Robust Facial Feature Localization by Coupled Features, in *4th IEEE Int. Conf on Automatic Face and Gesture Recognition,* Grenoble, France, 2000.

[47] ZHU (X.), FAN (J.), ELMAGARMID (A.K.), Towards Facial Feature Extraction and Verification for Omni-Face Detection in Video-Images, *Image Processing*, 2, pp. 113116,2002.

Appendix A – Mixtures of Factor Analysers

The Gabor detectors yield 49-dimensional outputs in our scheme (7×7 window) and 1,225 parameters are involved for each mixture component (49 mean and $49 \times 48/2 = 1{,}176$ covariance terms). Such high dimensional spaces run the risk of overfitting as the consequent number of model parameters is large. One solution would be restricting them to be diagonal to reduce the number sacrificing some valuable covariance shape information. A good alternative method to reduce the number of parameters is the factor analysis model, where the data (denoted with x) are assumed to be generated in a lower dimensional latent space:

$$x - \mu_j = \Lambda_j z + \varepsilon \tag{7}$$

Here, u_j denotes the d-dimensional component mean, Λ_j is the (dxp) factor loading matrix that maps the data from the p-dimensional generative space to the d-dimensional space, z are the latent variables and ε is the Gaussian sensor noise. In other words, the covariances of the Gaussian components (Σ_j) are not restricted to be spherical or diagonal, but given as

$$\Sigma_j = \Lambda_j \Lambda_j' + \Psi, \tag{8}$$

(with $p<<d$) and Ψ is the diagonal variance due to sensor noise. Thus the number of parameters is reduced from $O(d^2)$ to $O(pd)$, yet the covariances are still modeled. To automatically determine the number of components and the values of the loading factors, we employ the IMoFA-L algorithm proposed in [34]. The IMoFA-L algorithm starts with a single Gaussian component derived from a single factor placed on the data, and incrementally adds components and factors to the mixture. At each step, a multivariate kurtosis based metric is employed to split the component that looks least unimodal. For factor addition the discrepancy between the real and modeled covariances are monitored. The algorithm proceeds by interrupting EM (Expectation-Maximization procedure) to add a component to the mixture or a factor to a component. A validation set is used to control the complexity of the final model.

Appendix B – Removing the Singularity in the Structural Analysis

Let $x_j = [x_j, y_j]^T$ denote the normalized 2D coordinates of the i'th landmark with respect to the origin in a particular support set. The likelihood of a single landmark vector $\boldsymbol{x}$ under Gaussian assumption $N(\mu, \Sigma,)$ is written as

$$L(\mathbf{x}, \mu, \Sigma) = \frac{1}{} \; e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)} \tag{9}$$

If we denote the rotation function with $r(\boldsymbol{x}, \theta)$, the joint likelihood of all landmark points to be maximized, that incorporates the structural a priori information, is expressed as

$$L_{structure} = \prod_{j=1}^{l} L(r(\mathbf{x}_j, \theta), \mu_j, \Sigma_j) \tag{10}$$

Maximizing this expression is equivalent to minimizing the following expression:

$$\min_{\theta} \sum_{j=1}^{l} (r(\mathbf{x}_j, \theta) - \mu_j)^T \Sigma_j^{-1} (r(\mathbf{x}_j, \theta) - \mu_j) \tag{11}$$

with

$$r(\mathbf{x}_j, \theta) = \begin{bmatrix} x_j \cos\theta - y_j \sin\theta \\ x_j \sin\theta + y_j \cos\theta \end{bmatrix} \tag{12}$$

We use a Levenberg-Marquard (LM) procedure to minimize this expression. The first landmark in the support set is excluded from the minimization, as it varies in one dimension only. Since the initial rotation brings the landmarks very close to the global minimum and the function is smooth, two or three iterations are sufficient to find the solution. The resulting rotation is applied, and the distribution parameters of the landmarks are re-estimated. The LM procedure is only used to perturb the landmark positions to remove the singularity in the distribution of the first landmark, which is an artifact, generated by the normalization procedure.

In the 3D case, we only need to change the rotation expression accordingly:

$$r\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}, \theta, \varphi\right) = \begin{bmatrix} \cos(\varphi) - \cos(-\theta) & -\cos(\varphi)\sin(-\theta) & \sin(\varphi) \\ \sin(-\theta) & \cos(-\theta) & 0 \\ -\sin(\varphi)\cos(-\theta) & \sin(\varphi)\sin(-\theta) & \cos(\varphi) \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{13}$$

where 0 is the azimuth, and φ is the elevation angle.