U.Porto Journal of Engineering, 2:2 (2016) 39-54
ISSN 2183-6493
DOI 10.24840/2183-6493_002.002_0005

Received: 17 October, 2015
Accepted: 29 March, 2016
Published: 29 July, 2016

Multimodal Hierarchical Face Recognition using Information from 2.5*D* Images

João C. Monteiro[1], Tiago Freitas[2], Jaime S. Cardoso[3]
INESC-TEC, Porto, Faculty of Engineering, University of Porto, Porto, Portugal ([1]dee12007@fe.up.pt, [2]bio11051@fe.up.pt, [3]jaime.cardoso@inesctec.pt)

Abstract
Facial recognition under uncontrolled acquisition environments faces major challenges that limit the deployment of real-life systems. The use of 2.5*D* information can be used to improve discriminative power of such systems in conditions where RGB information alone would fail. In this paper we propose a multimodal extension of a previous work, based on SIFT descriptors of RGB images, integrated with LBP information obtained from depth scans, modeled by an hierarchical framework motivated by principles of human cognition. The framework was tested on EURECOM dataset and proved that the inclusion of depth information improved significantly the results in all the tested conditions, compared to independent unimodal approaches.

Subject Headings. Biometrics, Information Processing
Author Keywords. Biometrics, 3*D* Face Recognition, Information Fusion, Partial Data

1. Introduction

Over the past few years, the issue of face recognition has been in the spotlight of many research works in pattern recognition, due to its wide array of real-world applications. The face is a natural, easily acquirable trait with a high degree of uniqueness, representing one of the main sources of information during human interaction. These marked advantages, however, fall short when images of limited quality, acquired under unconstrained environments, are pre-sented to the system.

Whereas technological improvements in image capturing and transmitting equipment managed to attenuate most noise factors, partial face occlusions still pose a genuine challenge to automated face recognition (Li et al., 2014).

Facial occlusions may occur due to a multiplicity of deliberate or unintentional reasons. Whereas accessories, such as sunglasses and scarves, and facial hair represent quite common sources of occlusion in daily life, they can also be explored by bank robbers and shop thieves in an attempt to avoid recognition. Furthermore, the use of some accessories might be enforced in restricted environments (such as medical masks in hospitals and protection helmets in construction areas) or by religious or cultural constraints (Min et al., 2014a). The fact that humans perform and rely on face recognition routinely and effortlessly throughout their daily lives leads to an increased interest in replicating this process in an automated way, even when such limitations are known to frequently occur.

Even though there is no consensus in the cognitive science field as to how the human brain recognizes faces, either based on their individual local features or, more holistically, on the basis of their overall shape, several works have shown that both levels of information play a non-negligible role in human face perception (Schwaninger et al., 2007, Gold et al., 2012). In previous works (Monteiro and Cardoso, 2015a,b), the authors explored the global precedent

hypothesis for human perception as the basis for a new decision strategy to guide the face recognition process, in an hierarchical manner, in RGB color images. Such hypothesis claims that face recognition is performed by the human brain in a global-to-local flow, with holistic information gaining precedence over a more detailed local analysis.

In the present work, we built upon the aforementioned previous research, incorporating information from the three-dimensional structure of the face, through the use of 2.5*D* depth images acquired using the Microsoft Kinect low-cost sensor. By exploring information from an extra spatial dimension we aim to grant the original algorithm with higher robustness in scenarios, such as critically low illumination, where the acquisition of color images is severely limited. With this goal in mind, we performed a detailed analysis of the state-of-the-art works on 3*D* face recognition, in order to identify trends of research to help guide the design of the extension of the referred previous works, as well as defining future prospects of research.

We start by presenting in Section 2 a thorough review of the state-of-the-art concerning the evolution of face recognition to 3*D* scenarios, with special focus to recent works on Kinect depth images. We then present, in Section 3, a detailed description of the extension of the original hierarchical algorithm to incorporate depth information. The most relevant preliminary results are presented and discussed in Section 4, while the main conclusions and prospects for future work can be found in Section 5.

2. State-of-the-art in 3*D* Face Recognition

As referred in the previous section, face recognition is a challenging pattern recognition problem especially in the presence of variations in illumination conditions, occlusions, pose and facial expression changes and disguises.

Due to the inherent 3*D* structure of the face, changes in illumination and non-frontal pose from the individuals could lead to changes in some facial features, thus conditioning the performance of the system. To overcome the decrease of performance in these situations, 3*D* face recognition can be used to improve the recognition rate, giving a more robust facial description and not being affected by illumination variation, leading possibly to a greater discriminative power. There are two main ways of representing 3D facial structure (Abate et al., 2007): the 2.5-Depth images and 3*D* images. The 3*D* images retain all the facial geometry information, whereas the 2.5*D* or range images are a bi-dimensional representation of a set of 3*D* points in which each pixel in the $X0Y$ plane stores the depth *z* value. The disadvantage of this representation is that it only takes information from one point of view, allowing only a single facial model. Also, the 3*D* image depends only on internal anatomical structure while 2.5*D* scans are affected by environmental conditions and external appearance. Both of these representations can increase the performance of recognition algorithms, but it is important to evaluate in which type of systems the acquisition of 3*D* facial data poses a feasible challenge. Table 1 lists some of most often used acquisition solutions used in the creation of most datasets found in literature.

Sensor	Type	Resolution	Working Distance	Price ($)
Minolta Sensors [Min et al., 2014b]	3*D* Laser Scanning	0.041-0.22	∼*2.5*	25000
3dMDface [Min et al., 2014b]	Vision Cameras	<0.2	——	10k - 20k
CyberWare 3030RGB/PS [Min et al., 2014b]	Low-Intensity Laser Light Source	0.08 - 0.3	0.35	∼72000
Inspeck Mega Capturer II [Hiremath and Manjunatha, 2013]	Structured-Light	0.7	1.1	Not Available
Kinect [Min et al., 2014b]	IR laser Emitter	∼1.5 - 50	0.5 - 4.5	149.99
SoftKinetic DS325 [Mracek et al., 2014]	Diffused Laser	14 at 1 m distance	0.15-1	259
Structure [Gutfeter and Pacut, 2015]	IR Structured Light	0.5 - 30	3.5	379
PrimeSense Carmine [Min et al., 2012]	IR Light Source	0.1 - 1.2	3.5	Not Available

Table 1: List of some sensors used in 3*D* Facial Recognition

The offered solutions can be either stereoscopic camera systems, structured light systems or laser range systems, obtaining both 3*D* and intensity information. While the Minolta and Inspeck sensors are generic 3*D* sensors, CyberWare and 3dMD were designed specifically to face 3*D* scanning. All of those solutions are very precise, yet they are also very expensive. The natural evolution of the 3*D* sensors is towards low cost sensors, with a decrease in the resolution, that could be used for the creation of real-time systems that are cheap but at the same time robust enough to perform face recognition in adverse conditions. Kinect is one of the most used sensors, and it contains an infra-red (IR) laser emitter and an IR camera in addition to a RGB camera. The RGB camera captures the RGB images directly, whereas the laser emitter and IR camera work together to capture the depth map. The depth map is obtained via a triangulation process based on those two sensors. First the IR laser projects a predesigned pattern of spots in the scene (using a raster) and the reflection of the pattern is captured by the IR camera. (Min et al., 2014b) Although the Kinect is the most commonly used low-cost sensor for this type of applications SoftKinetic DS325 (Mracek et al., 2014), Structure sensor (mobile sensor in tablets) (Gutfeter and Pacut, 2015) and PrimeSense (Min et al., 2012) (bought recently by Apple but currently not acquirable) also have been used in some facial recognition datasets, being an alternative to the Microsoft sensor.

Using these sensors, many datasets are available for algorithm testing. All these datasets can be divided in two groups: the high-resolution scans that use high-quality and expensive scans like Minolta and 3dMDface Systems, and the low-resolution scans that use low-cost sensors with lower precision and resolution like Kinect, SoftKinectic and Structure sensors. The databases of facial surfaces should have a large variety of subjects and conditions in order to simulate the most important challenges in facial recognition (pose, facial expression, illumination and occlusion). The information relative to these sensors was obtained from (Min et al., 2014b), where most of this sensors and databases were analysed in detail.

The first datasets created for the 3*D* facial recognition problem used high precision sensors. Some of the most important datasets are the Bosphorus, York, FRGC, GavabDB, Binghamton

University, Texas-3*D*, UMB-DB and 3*D*-RMA (Abate et al., 2007). All these datasets use expensive and high resolution sensors. Alongside the evolution of sensors towards the low-cost, lower resolution and faster acquisitions, recent databases were also constructed with this type of sensors. Some examples are the CurtinFaces (Li et al., 2013), NASK-StructureFacebase (Gutfeter and Pacut, 2015), BIWI Kinect Head Pose Dataset (Hayat et al., 2015), UWA Kinect dataset (Hayat et al., 2015), FaceWareHouse (Cao et al., 2014) and EURECOM dataset (Min et al., 2014b). Although the number of 2D+3*D* datasets are still in low number comparatively to the 2D datasets, these databases are increasing in number and in variety and are fundamental to the testing and assessment of performance of new algorithms. The specifications of these datasets are shown on Table 2.

Dataset	Texture	3*D* Sensor	Scans	Subjects	Expression	Illumination	Pose	Occlusion	Video
CurtinFaces [Li et al., 2013]	Yes	Kinect	>5000	52	Yes	Yes	Yes	Yes	No
NASK-StructureFacebase [Gutfeter and Pacut, 2015]	Yes	Structure	330	13	No	No	Yes	No	Yes
BIWI Head-Pose [Hayat et al., 2015]	Yes	Kinect	>15000	20	No	No	Yes	No	No
UWA Kinect [Hayat et al., 2015]	Yes	Kinect	15000	48	Yes	No	Yes	No	No
FaceWareHouse [Cao et al., 2014]	Yes	Kinect	3000	150	Yes	No	No	No	Yes
EURECOM [Min et al., 2014b]	Yes	Kinect	> 450	52	Yes	Yes	Yes	Yes	Yes

Table 2: List of some databases created for the assessment of 3*D* Facial Recognition algorithms

Through the analysis of the Table 2 we can observe that EURECOM database seems to be the most complete database, although the number of scans is limited. A test with different type of sensors and conditions is crucial for a construction of a good framework. The generation of 2.5*D* or 3*D* datasets leads to a necessary adaptation of the frameworks designed for 2D images to be capable to receive tridimensional information as input. Most of the datasets presented above were built due to the need of achieving and objective assessment of how newly designed algorithms worked on a variety of new 3*D* Face recognition problems. There are three main types of approaches for this pattern recognition problem: 2D Based, 3*D* based and multimodal. The first type uses synthetic 3*D* face models to increase the robustness in respect to pose variations as well in changes in illumination and facial expression. 3*D*-based methodologies don't use intensity information and only use 3*D* or 2.5*D* data for the algorithms. Finally, the multimodal approaches take advantage of information from both previous approaches in order to attempt fusion of the first two types described earlier.

The 2D based approaches were in the genesis of the 3*D* facial recognition and, although they only use a 2D input query face, a 3*D* model is used to improve the robustness of a system. Many approaches like (Blanz and Vetter, 2003), (Lu et al., 2004) and (Hu et al., 2004) in which many virtual 3*D* models are generated to simulate the variations in pose and facial expression. The problems with these approaches are the several limitations in constructing a model from a single frame, and the non proximity to reality of the generated models.

The use of methodologies based only on 3*D*, thus called unimodal, has shown to be a good alternative to RGB in conditions of varying illumination, facial expression and pose. The main

problem with such approaches concerns the need of a correct alignment of 3*D* data between two face surfaces. In 1992, Besl (Besl and McKay, 1992) introduced the Iterative Closest Point (ICP) to perform a correct alignment of facial models. One of the first works with 3*D* facial recognition was introduced by Gordon (Gordon, 1991), based on the calculation of distance measures between some regions (like shape of forehead, jaw line, eye corner cavities and cheeks). A few years later, Tanaka (Tanaka et al., 1998) proposed a curvature-based approach. By extracting the principal curvatures and their orientations in a facial model, some features are extracted and mapped on two unit spheres Extended Gaussian Images (EGI). Chua (Chua et al., 2000) found some regions (nose, eye socket and forehead) and a Point Signature two-by-two comparison among different facial expressions of the same person and similarity measure is used with a rank vote process using a training indexed table. The Local shape descriptors in these type of scans were introduced by Moreno (Moreno et al., 2003) where different regions are segmented using the median signs and the Gaussian curvatures, isolating regions with significant curvatures. Some features are extracted (areas, distances, angles, area ratios, mean of areas, mean curvatures, variances, etc.) in order to achieve a good description of these regions. In recent works, Rui Min (Min et al., 2012), using the Apple PrimeSense, proposed a canonical face based system using only frontal pose images. The facial region obtained is divided on nose, eye region, cheeks and the remaining parts (each region is associated with a respective weight). A feature vector is formed containing the L_2 distances between each facial region and their corresponding areas. Naveen (Naveen and Moni, 2015) in FRAV3D database proposed the use of 2D spectral and 2D spatial domain information to solve the problem of facial recognition, based on 2D DWT (Discrete Wavelet Transform) and 2D DCT (Discrete Cosine Transform). Using landmark detection based on three principal curvatures, Tang (Tang et al., 2015) determines the geometric properties of each vertex using an asymptotic cone in order to generate three curvature faces to which are applied Local Normal Patterns. Neto (Cardia Neto and Marana, 2015), used 3*D*-Local Binary Patterns (LBP) with Histogram Oriented Gradients (HOG) as approach on Kinect Eurecom images. Bondi (Bondi et al., 2015) also used real-time Kinect sequences by generating high resolution models every-time someone passes through the sensor. Keypoints are detected using SIFT and spatial clustering, used in pairs to evaluate the facial curves between pairs of points.

The inclusion of two modalities has shown to be the most promising for real-time systems and uncontrolled environments. The results have shown to be always improved with the fusion of 2D and 3*D* modalities. (Abate et al., 2007)

In 2003, Chang et al. (Chang et al., 2003) investigated the benefits of integrating 3*D* data (using a Minolta Vivid 900 sensor) with 2D images, using PCA separately on 2D and 3*D* data. The authors state that 2D and 3*D* individually get similar performances, but when combined (with a simple weighting system) they get a significant increase in the performance. Tsalakanidou et al. (Tsalakanidou et al., 2003) applied Eigenfaces on both 2.5*D* and 2D scans. Here the multimodal approach has shown significant improvements when compared with independent 2.5*D* and 2D recognition. Later, in his works, Mian (Mian et al., 2007) (Mian et al., 2008) proposed some new approaches for multimodal face recognition. In 2007 (Mian et al., 2007), using a combination between 3*D* Spherical Face Representation (SFR) and 2D SIFT, big part of the candidate faces are removed from the query images. Then the eyes-forehead and the nose regions are automatically segmented. One year later he proposed a new method using a new keypoint detection using the high shape variations in 3*D* data and a Local Feature Matching (Mian et al., 2008) based on tensor representation for depth data. Using Kinect Scans, Li et al. (Li et al., 2013) obtained canonical frontal views (shape and textural). Here the RGB data is

also transformed to the discriminant color space and a sparse representation classifier (SRC) is applied in both types of scans. In high resolution scans Hiremath et al. (Hiremath and Manjunatha, 2013), used Radon transform on both texture and depth images in order to obtain binary maps to crop the facial region. Gabor features are extracted from both type of scans and obtained vectors in which PCA is applied as the input in an AdaBoost classifier that selects the most discriminant features. Also using Kinect, Ajmera (Ajmera et al., 2014) proposed the use of SURF-based descriptors in Kinect scans (tested on EURECOM and CurtinFaces datasets). Here, images with variation in pose are generated, and SURF is also used for face matching independently on depth and intensity images. Mrácek (Mracek et al., 2014) used Gabor and Gauss-Laguerre filters to describe texture and depth information.

In recent works, Elaiwat (Elaiwat et al., 2015) in high resolution scans, used curvelet coefficients to represent the facial geometrical features, to identify keypoints and extract local information about its neighbourhood. Nair (Naveen et al., 2015), used a Local Polynomial Approximation Filter (LPA) to obtain directional faces. These faces are optimized using the Intersection of Confidence Interval Rule (ICI) and feature extraction is done using mLBP. Krishnan (Krishnan and Naveen, 2015) introduced a new framework using entropy maps of the texture and depth maps independently and using saliency maps on texture images. Dai et al. (Dai et al., 2015), using Kinect data, proposed a new local descriptor for feature extraction after the use of Gabor filters: Enhanced Local Mixed Derivative Pattern (ELMDP). Finally, Bennamoun (Hayat et al., 2015) proposed a new raw depth pose estimation, not assuming a strong statistical relationship between the training data and the query faces, followed by the application of a Riemannian manifold for feature selection.

As we can observe, new developments show the use of unimodal 3*D* and multimodal approaches for developing face recognition frameworks, although the use of the multimodal ones seem to be the most promising strategies for real-time systems. Table 3 summarizes the most relevant information extracted from the works described above, regarding a series of parameters (feature extraction, classifiers, datasets ...) whose rational choice must be thought of when designing and assessing a new approach to 3*D* face recognition.

3. Proposed Methodology

3.1. Original algorithm overview

The hierarchical recognition algorithm that we work upon on the present work was first proposed and explored in two previous works (Monteiro and Cardoso, 2015a,b) and is schematically represented in Figure 1. Figures 1(a) and 1(b) depicts the enrollment process in the proposed approach. During enrollment, a new individual's biometric data is added into a previously existent database of individuals. For each individual, a hierarchical ensemble of *M* partial face models is trained. The *M* individual-specific models are built by maximum a posteriori (MAP) adaptation of the corresponding set of *M* universal background models (UBM) using individual-specific data. The UBM is a representation of the distribution that a biometric trait presents in the universe of all individuals. MAP adaptation works as a specialization of the UBM based on each subject's biometric data. The idea of MAP adaptation of the UBM was first proposed by Reynolds (Reynolds et al., 2000), for speaker verification.

Author	Type	Feature Extraction	Classifier	Dataset	r_1
[Gordon, 1991]	*3D*	Distance Measures	Euclidean Distance	8 subjects (23 scans)	97.00
[Tanaka et al., 1998]	*3D*	Curvature features	Fisher Spherical Approximation	37 subjects	100
[Chua et al., 2000]	*3D*	Point Signature Comparison	Ranked vote	6 subjects (24 scans)	100
[Moreno et al., 2003]	*3D*	Geometric statistics	Euclidean Distance	GavabDB	78.00
[Min et al., 2012]	*2.5D*	*L*2 Distances	Euclidean Distance	20 subjects	100
[Naveen and Moni, 2015]	*2.5D*	DWT + DCT	Euclidean Distance	FRAV3D	96
[Tang et al., 2015]	*3D*	Principal Curvatures + LNP	Weighted Sparse Representation	FRGC	93.33
[Cardia Neto and Marana, 2015]	*2.5D*	*3D*-LBP+HAOG	SVM	Eurecom	~98
[Bondi et al., 2015]	*2.5D*	SIFT an Curvatures	RANSAC + Distance and Salience Metric	Kinect Sequences (16 subjects)	100
[Chang et al., 2003]	MM	PCA	Mahalanobis Distance	366 subjects (676 scans)	98.8
[Tsalakanidou et al., 2003]	MM	Eigenfaces	Euclidean Distance	XM2VTS	98.75
[Mian et al., 2007]	MM	*3D*-SFR and 2D-SIFT	Modified ICP	FRGC	98.31
[Mian et al., 2008]	MM	Tensor Representation + 2D SIFT	4 Different Similarity Measurements	FRGC	98.6
[Li et al., 2013]	MM	-	SRC	CurtinFaces	96.7
[Hiremath and Manjunatha, 2013]	MM	Gabor Features	Nearest Neighbor	Texas *3D*, Bosphorus and CASIA *3D*	99.5
[Ajmera et al., 2014]	MM	SURF	Nearest Neighbor	Eurocom and CurtinFaces	89.28 and 98.07
[Mracek et al., 2014]	MM	Gabor/Gauss-Laguerre features	Correlation Metric	Kinect, Kinectic Dataset, FRGC	<89
[Elaiwat et al., 2015]	MM	Curvelet Coefficients	Cosine Distance	FRGC, BU-3DFE, Bosphorus	99.2, 95.1, 91
[Naveen et al., 2015]	MM	mLBP	Euclidean Distance	FRAV3D	<91.88
[Krishnan and Naveen, 2015]	MM	Saliency + Entropy + HOG	Tree Bagger	FRAV3D, CurtinFaces	92
[Dai et al., 2015]	MM	ELMDP/Gabor features	Nearest Neighbor	CurtinFaces	~95
[Hayat et al., 2015]	MM	Riemannian manifold	SVM	BIWI Kinect, CurtinFaces, UWA Kinect	94.737

Table 3: Summary of the most relevant works concerning 3*D* face recognition

The database is probed during the recognition process to assess either the validity of an identity claim (verification) or the k most probable identities (identification) given an unknown sample of biometric data. In the aforementioned previous works, the authors proposed an innovative approach to the recognition process based on the global precedence hypothesis of face perception by the human brain. Recognition is performed hierarchically, as depicted in Figure 1(c), with global models taking precedence over more detailed ones. Partial models are hierarchically organized into levels. Each level is composed by a set of non-superimposing subregions, I_l of equal size, with subregions at the same level summing to the full-face image, I_0.

During recognition, a test image from an unknown source follows the hierarchical flow depicted in Figure 1(c), until a decision can be made with a significant degree of certainty. The significance of a decision carried out at a single level is defined through the analysis of the likelihood-ratio values obtained for each possible identity claim, through the computation of a certainty index, c_m:

$$c_m = s_{t*,m} - \frac{1}{T-1} \sum_{t=1, t \neq t^*}^{T} s_{t,m} \quad (1)$$

where $s_{t*,m}$ represents the highest observed likelihood-ratio value (true identity) and the average of all other values (average impostor) is represented by $\frac{1}{T-1}\sum_{t=1,t\neq t^*}^{T} s_{t,m}$. If the c_m value exceeds a previously optimized threshold, θ_l, the maximum likelihood-ratio decision is accepted.

When $c_m > \theta_l$, however, the algorithm will consider that an analysis at a more detailed level is necessary to achieve a decision with a higher degree of confidence. At this point, the algorithm proceeds to the next level, working on subregions I_{1-2}, the second in the hierarchical chain depicted in Figure 1. When one level is composed by multiple subregions, each one of them is treated independently, and only the maximum c_m value among them is considered.

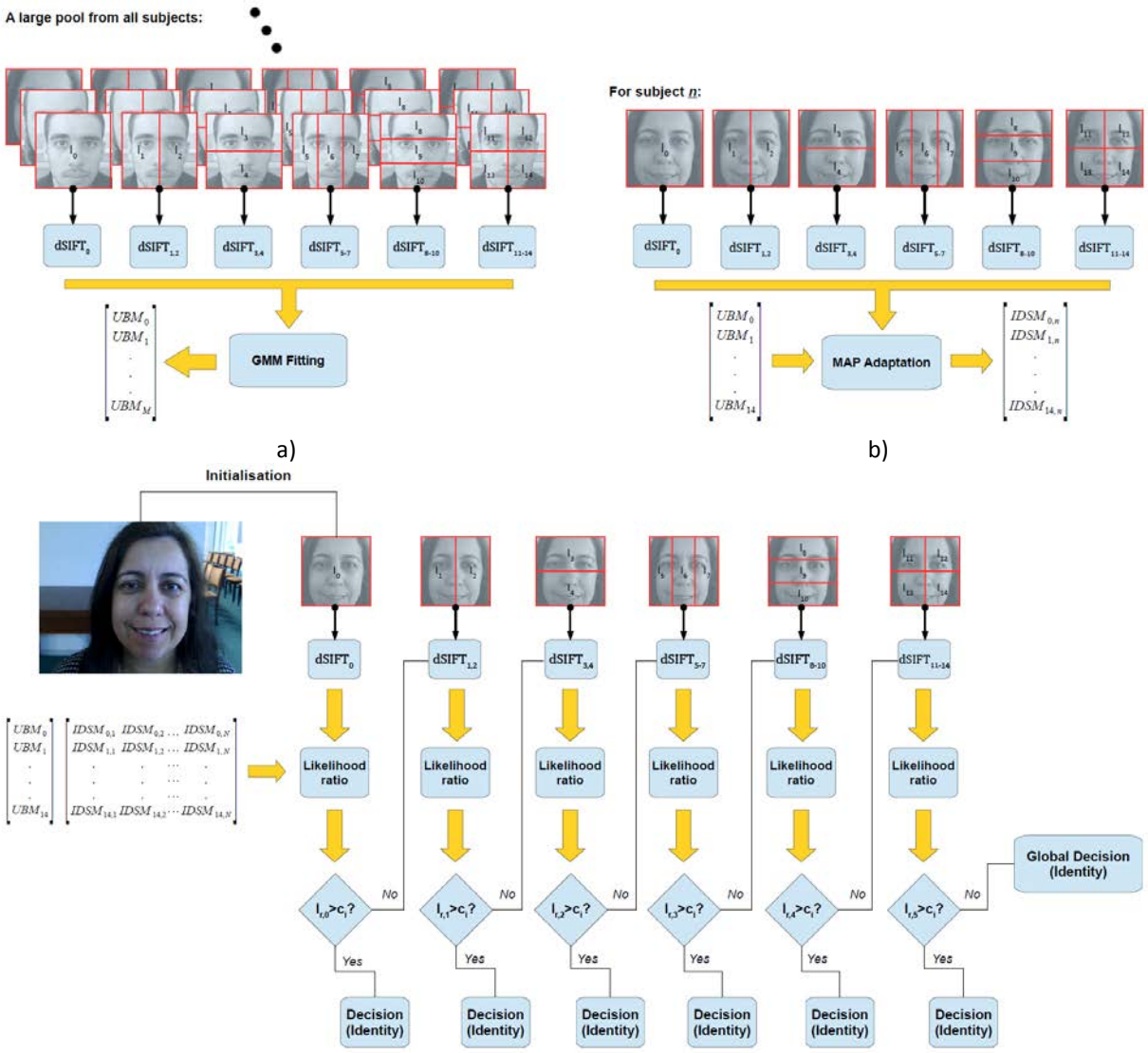

Figure 1: Schematic representation of the proposed algorithm and its main blocks:
a) training of the universal background models using data from multiple individuals; b) maximum a posteriori (MAP) adaptation of the universal background models (UBM) to generate individual specific models; and (c) testing with new data from unknown sources
From (Monteiro and Cardoso, 2015a)

All models are trained using Gaussian Mixture Models (GMM) and sets of SIFT keypoint descriptors for feature representation. In the next section we present some alterations to this choices in order to adapt the outlined framework to depth images.

3.2. Proposed extension to depth images

On the present work we carried out some preliminary experiments using the framework detailed in the last section, but using Kinect depth images as the input for the whole system. The architecture of the system remains as described above and depicted in Figure 1. By analysis of some of the recent works listed in Section 2, we decided to focus the extension of our framework on feature description. With this in mind, two feature descriptors were chosen to describe Kinect depth images, taking the place of the original SIFT keypoint descriptor from the original works:

- **Dense SIFT grid**: while the original SIFT algorithm includes a keypoint detection block, the noisy nature of depth images, associated to their low intrinsic detail, might severely hinder the correct functioning of this detection. Therefore, we used a dense grid of equally separated keypoints to compute the SIFT descriptors and guarantee that enough information is present for robust modeling.
- **Local Binary Patterns (LBP)**: as an alternative to dense SIFT, we also perform uniform LBP description locally on a set of 4 × 4 sub-images. The resulting histograms are then concatenated to achieve a full description of the image. We chose LBP not only due to its vast array of applications in computer vision in works concerning texture description, but also because of the promising performance it presented in some recent datasets built with Kinect depth images (Min et al., 2014b).

With this extension we end up with two instances of the whole hierarchical framework, based on either RGB or depth images. In the next section we discuss how information from both sources is integrated into a single decision.

3.3. Multimodal fusion

In this work we performed fusion at the score level, using the likelihood-ratio values from two hierarchical pipelines: one for RGB images, s_{RGB}, and one for depth images, s_d. The final fusion score, s_f, is obtained by a weighted averaging of the two scores, $s_f = w_{RGB} \times s_{RGB} + w_d \times s_d$. The optimal values for the weighting parameters were found through grid search, under the constraint $\Sigma_i w_i = 1$.

4. Results and Discussion

4.1. EURECOM Dataset

The experiments were conducted in the EURECOM database. Using Kinect Sensor, this database has a set of well-aligned 2D, 2.5*D*, 3*D* and video data. It includes scans from 52 subjects (38 males and 14 females) from two sessions interleaved from 5 to 14 days. Each session has nine types of scans that include: neutral face, open mouth, smiling, strong illumination, occlusion with sunglasses, occlusion by hand, occlusion by paper, right face profile and left face profile. The acquisition environment is controlled in terms of luminosity, with the individuals always in a range from 0.7 to 0.9 meters to the sensor. A blank background was chosen to make the processing of the data easier. An example of the 2D and 2.5*D* images from a single individual is presented in Figure 2. We chose to not consider the profile images as the designed framework is still limited as far as pose variations are concerned.

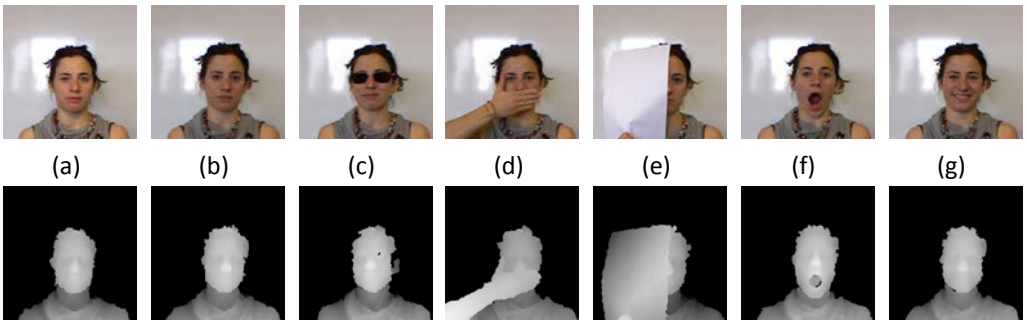

Figure 2: Example images from the EURECOM dataset, for a single subject: (a-g) RGB (h-n) depth

	LO	OE	OM	OP	OPM	S
L_0	0.990	0.962	0.856	0.240	0.865	0.962
L^*	0.990	0.962	0.952	0.625	0.913	0.990
$L_{0.2}$	1.000	0.988	0.964	0.554	0.976	0.988
LGBP [Min et al., 2014b]	0.990	0.904	0.990	0.817	0.952	1.000
SIFT [Min et al., 2014b]	0.837	0.712	.885	0.375	0.913	99.04

Table 4: Main results obtained using RGB images and SIFT feature description

	LO	OE	OM	OP	OPM	S
L_0	0.721	0.615	0.308	0.048	0.462	0.692
L^*	0.721	0.615	0.308	0.048	0.490	0.731
$L_{0.2}$	0.606	0.423	0.087	0.010	0.375	0.675
SIFT [Min et al., 2014b]	0.069	0.020	0.010	0.029	0.020	0.049

Table 5: Main results obtained using depth images and dense SIFT feature description

4.2. Experimental setup

In our framework, neutral face images were used for the training of the models and the remaining scans were used as query faces inputted in the system (profile images were eliminated). The images were manually cropped in order to only analyze the facial region.

We chose to assess the rate of correctly identified individuals, by checking if the true identity is present among the *N* highest ranked identities. The *N* parameter is generally referred to as rank. This allows us to define the Rank-1 recognition rate, r_1, as the recognition rate at N=1.

4.3. Performance analysis

The main results obtained with the framework detailed above, for both RGB and depth images, are summarized in Tables 4-7. For each tested scenario we present the individual performance observed for each condition present in the EURECOM dataset: light on (LO), occluded eyes (OE), occluded mouth (OM), occluded paper (OP), open mouth (OPM) and smile (S). For each of such conditions and scenarios we define three reference values extracted from our framework:

- **Full-face**, L_0: performance observed when considering only the first level of the hierarchical framework.
- **Optimal**, L^*: performance observed for the full hierarchical framework, optimized with regard to the θ_l parameter.
- **Reject option**, $L_{0.2}$: performance observed with the option of not-classifying a image if it reaches the last level of the hierarchical framework with no certain classification being achieved. We choose to assess performance in the specific case of 20% rejection.

From the results obtained using RGB images, we can conclude that our 2D approach has a similar or higher performance in all tested conditions, even though a fair comparison can only be performed between our results and the SIFT approach presented in (Min et al., 2014b), while the LGBP results are displayed on Table 4 because they were the best ones obtained in the same work. The SIFT algorithm tested by Min et al. (Min et al., 2014b), was outperformed by our GMM modeling approach to SIFT description. Only in the face occluded with paper, can

we observe worse results when compared to the literature. This indicates that more work needs to be done to overcome this drawback condition.

Tables 5 and 6 summarize the most relevant results concerning the application of our hierarchical framework to depth images. One common observation that can be made is that the application of the rejection mode alternative doesn't bring about any improvement, as it did on the original RGB scenario. This might relate to the higher probability of getting strong false positives from depth images. A very high score in a wrong identity exerts a strong limitation over the computation of the quality criterium defined on Equation 1. It is also interesting to note how the optimal performance from the whole hierarchical flow shows very little improvement for all the test scenarios, using dense SIFT, when compared to the holistic representation from the first level. The advantages of using our approach for this specific modeling strategy can, therefore, be questioned. However, when comparing to the traditional SIFT detector and descriptor, used by Min et al. (Min et al., 2014b), we can see that our approach of modeling a densely sampled grid of SIFT descriptors achieves a considerably higher performance and should be considered in its simplest form (using only the holistic representation from the first level) as an aid to more traditional RGB-based approaches.

	LO	**OE**	**OM**	**OP**	**OPM**	**S**
L_0	0.779	0.587	0.240	0.058	0.462	0.683
L^*	0.798	0.635	0.433	0.106	0.538	0.788
$L_{0.2}$	0.793	0.694	0.106	0.071	0.524	0.783
LBP [Min et al., 2014b]	0.837	0.789	0.519	0.125	0.827	0.837

Table 6: Main results obtained using depth images and LBP feature description

	LO	**OE**	**OM**	**OP**	**OPM**	**S**
2D-SIFT + 2.5*D*-LBP	1.000	0.981	0.952	0.625	0.933	0.990
Fusion LGBP [Min et al., 2014b]	1.000	0.894	0.981	0.856	0.981	1.000
Fusion LBP [Min et al., 2014b]	0.990	0.933	0.962	0.817	0.981	1.000

Table 7: Multimodal fusion results obtained using information from RGB and depth images

When considering the LBP results, presented in Table 6, we might observe that, opposed to the dense SIFT modeling, the hierarchical framework brings about considerable increase in performance for almost all test scenarios. All observed results are also slightly to considerably better than the performance observed for their SIFT counterpart, corroborating the observations presented by Min et al. (Min et al., 2014b) in the original work on the EURECOM dataset. For that reason the multimodal fusion shown below are obtained by considering the original SIFT formulation for 2D images and only the LBP version of the framework applied to depth 2.5*D* images.

The analysis of the multimodal fusion results, presented in Table 7, shows that significant improvement was obtained with respect to both 2D and 2.5*D* unimodal alternatives, for all test scenarios except OP, where the discrepancy between the individual performances of the unimodal approaches serves as a simple justification for this observation. When comparing with the results from the state-of-the-art we can observe that, once again, besides the OP scenario, we achieved performance either in the same range or slightly better than the ones

reported in literature. With this observations in mind we can readily conclude that our framework follows the trend observed in previous works, where multimodal fusion of multiple sources of information leads to an improvement over all individual performances. If we manage to improve the individual performances of each framework we should, thus, be able to also improve the discriminative power of multimodal fusion and, consequently, increase the real-life applicability of systems based on such approaches.

5. Conclusions and Future Work

In the present work we propose an extension of some works on hierarchical face recognition to 2.5*D* Kinect depth images. We approach this problem by proposing alternative feature description strategies, such as dense SIFT and LBP. We achieved or improved over state-of-the-art performance in most tested scenarios.

However, some potential improvements can be easily suggested in order to achieve higher performance and also to further assess the effectiveness of the proposed algorithm in a higher variety of interesting scenarios. First of all, exploring feature descriptors other than the proposed ones, or even exploring fusion of multiple features might result in a more complete description and, thus, result in better performance in a wider variety of acquisition conditions. One of such conditions, not yet tested due to its non-existence on the EURECOM dataset, is severe low illumination. This case, in theory, represents a situation where 3*D* information should severely improve over RGB alone.

References

Abate, Andrea F., Michele Nappi, Daniel Riccio and Gabriele Sabatino. 2007. "2D and 3D face recognition: A survey". *Pattern Recognition Letters* no. 28 (14):1885-1906. Accessed July 19, 2016. http://dx.doi.org/10.1016/j.patrec.2006.12.018.

Ajmera, Rahul, Aditya Nigam and Phalguni Gupta. 2014. "3D face recognition using kinect". In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, Article No. 76. ACM. Accessed July 19, 2016. http://dx.doi.org/10.1145/2683483.2683559.

Besl, P. J. and N. D. McKay. 1992. "A method for registration of 3-D shapes". In *Sensor Fusion Iv : Control Paradigms and Data Structures - Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) 1611*, 586-606. International Society for Optics and Photonics. Accessed July 20, 2016. http://dx.doi.org/10.1117/12.57955.

Blanz, V. and T. Vetter. 2003. "Face recognition based on fitting a 3D morphable model". *IEEE Transactions on Pattern Analysis and Machine Intelligence* no. 25 (9):1063-1074. Accessed July 19, 2016. http://dx.doi.org/10.1109/TPAMI.2003.1227983.

Bondi, E., P. Pala, S. Berretti and A. Del Bimbo. 2015. "Reconstructing high-resolution face models from Kinect depth sequences acquired in uncooperative contexts". In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015*, 1-6. IEEE. Accessed July 19, 2016. http://dx.doi.org/10.1109/FG.2015.7284882.

Cao, C., Y. Weng, S. Zhou, Y. Tong and K. Zhou. 2014. "FaceWarehouse: A 3D facial expression database for visual computing". *IEEE Transactions on Visualization and Computer Graphics* no. 20 (3):413-425. Accessed July 19, 2016. http://dx.doi.org/10.1109/TVCG.2013.249.

Chang, Kyong I., Kevin W. Bowyer and Patrick J. Flynn. 2003. "Face recognition using 2D and 3D facial data". Paper presented at ACM Workshop on Multimodal User Authentication.

Chin-Seng, Chua, Han Feng and Ho Yeong-Khing. 2000. "3D human face recognition using point signature". In *Proceedings: Fourth IEEE International Conference on Automatic Face and*

Gesture Recognition, 2000., 233-238. IEEE. Accessed July 19, 2016. http://dx.doi.org/10.1109/AFGR.2000.840640.

Dai, X., S. Yin, P. Ouyang, L. Liu and S. Wei. 2015. "A multi-modal 2D + 3D face recognition method with a novel local feature descriptor". In *2015 IEEE Winter Conference on Applications of Computer Vision*, 657-662. IEEE. Accessed July 19, 2016. http://dx.doi.org/10.1109/WACV.2015.93.

Elaiwat, S., M. Bennamoun, F. Boussaid and A. El-Sallam. 2015. "A curvelet-based approach for textured 3D face recognition". *Pattern Recognition* no. 48 (4):1235-1246. Accessed July 19, 2016. http://dx.doi.org/10.1016/j.patcog.2014.10.013.

Gold, Jason M., Patrick J. Mundy and Bosco S. Tjan. 2012. "The perception of a face is no more than the sum of its parts". *Psychological Science* no. 23 (4):427-434. Accessed July 19, 2016. http://dx.doi.org/10.1177/0956797611427407.

Gordon, G. G. 1991. "Face recognition based on depth maps and surface curvature". In *Geometric Methods in Computer Vision - Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) 1570*, 234-247. International Society for Optics and Photonics. Accessed July 20, 2016. http://dx.doi.org/10.1117/12.48428.

Gutfeter, W. and A. Pacut. 2015. "Face 3D biometrics goes mobile: Searching for applications of portable depth sensor in face recognition". In *2015 IEEE 2nd International Conference on Cybernetics (CYBCONF)*, 489-494. IEEE. Accessed July 19, 2016. http://dx.doi.org/10.1109/CYBConf.2015.7175983.

Hayat, Munawar, Mohammed Bennamoun and Amar A. El-Sallam. 2016. "An RGB–D based image set classification for robust face recognition from Kinect data". *Neurocomputing* no. 171:889-900. Accessed July 20, 2016. http://dx.doi.org/10.1016/j.neucom.2015.07.027.

Hiremath, P. S. and Manjunatha Hiremath. 2013. "3D face recognition based on depth and intensity gabor features using symbolic PCA and AdaBoost". *International Journal of Signal Processing, Image Processing and Pattern Recognition* no. 6 (5):1-12. Accessed July 20, 2016. http://dx.doi.org/10.14257/ijsip.2013.6.5.01.

Krishnan, Poornima and S. Naveen. 2015. "RGB-D face recognition system verification using kinect and FRAV3D databases". *Procedia Computer Science: Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India* no. 46:1653-1660. Accessed July 20, 2016. http://dx.doi.org/10.1016/j.procs.2015.02.102.

Li, B. Y. L., A. S. Mian, W. Liu and A. Krishna. 2013. "Using kinect for face recognition under varying poses, expressions, illumination and disguise". In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, 186-192. IEEE. Accessed July 20, 2016. http://dx.doi.org/10.1109/WACV.2013.6475017.

Li, YueLong, Li Meng, JuFu Feng and JiGang Wu. 2014. "Downsampling sparse representation and discriminant information aided occluded face recognition". *Science China Information Sciences* no. 57 (3):1-8. Accessed July 20, 2016. http://dx.doi.org/10.1007/s11432-013-4856-z.

Lu, Xiaoguang, Rein-Lien Hsu, Anil K. Jain, Behrooz Kamgar-Parsi and Behzad Kamgar-Parsi. 2004. "Face recognition with 3D model-based synthesis". In *Proceedings: Biometric Authentication: First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004.*, 139-146. Berlin, Heidelberg: Springer Berlin Heidelberg. Accessed July 20, 2016. http://dx.doi.org/10.1007/978-3-540-25948-0_20.

Mian, A., M. Bennamoun and R. Owens. 2007. "An efficient multimodal 2D-3D hybrid approach to automatic face recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence* no. 29 (11):1927-1943. Accessed July 20, 2016. http://dx.doi.org/10.1109/TPAMI.2007.1105.

Mian, Ajmal S., Mohammed Bennamoun and Robyn Owens. 2008. "Keypoint detection and local feature matching for textured 3D face recognition". *International Journal of Computer Vision* no. 79 (1):1-12. Accessed July 20, 2016. http://dx.doi.org/10.1007/s11263-007-0085-5.

Min, R., J. Choi, G. Medioni and J. L. Dugelay. 2012. "Real-time 3D face identification from a depth camera". In *2012 21st International Conference on Pattern Recognition (ICPR)*, 1739-1742. IEEE. Accessed July 20, 2016. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6460486.

Min, R., N. Kose and J. L. Dugelay. 2014. "KinectFaceDB: A kinect database for face recognition". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* no. 44 (11):1534-1548. Accessed July 20, 2016. http://dx.doi.org/10.1109/TSMC.2014.2331215.

Min, Rui, Abdenour Hadid and Jean-Luc Dugelay. 2014. "Efficient detection of occlusion prior to robust face recognition". *The Scientific World Journal* no. 2014:10. Accessed July 20, 2016. http://dx.doi.org/10.1155/2014/519158.

Monteiro, João and Jaime Cardoso. 2015a. "A cognitively-motivated framework for partial face recognition in unconstrained scenarios". *Sensors* no. 15 (1):1903. Accessed July 20, 2016. http://dx.doi.org/10.3390/s150101903.

———. 2015b. "A framework for face recognition in occlusion scenarios". Paper presented at 2015 Doctoral Consortium on Engineering (DCE 2015), in FEUP - Porto, Portugal, 11-12 June 2015.

Moreno, Ana Belén, Angel Sánchez, José Fco. Vélez and Fco. Javier Díaz. 2003. "Face recognition using 3d surface-extracted descriptors". Paper presented at Irish Machine Vision and Image Processing Conference, in University of Ulster - Northern Ireland - 3-5 September 2003.

Mráček, Štěpán, Martin Drahanský, Radim Dvořák, Ivo Provazník and Jan Vana. 2014. "3D face recognition on low-cost depth sensors". In *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1-4. IEEE. Accessed July 20, 2016. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7029424&tag=1.

Naveen, S. and R. S. Moni. 2015. "A robust novel method for face recognition from 2D depth images using DWT and DCT fusion". *Procedia Computer Science: Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India* no. 46:1518-1528. Accessed July 20, 2016. http://dx.doi.org/10.1016/j.procs.2015.02.072.

Naveen, S., S. S. Nair and R. S. Moni. 2015. "3D face recognition using optimised directional faces and fourier transform". In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1856-1861. IEEE. Accessed July 20, 2016. http://dx.doi.org/10.1109/ICACCI.2015.7275888.

Neto, João Baptista Cardia and Aparecido Nilceu Marana. 2015. "3DLBP and HAOG fusion for face recognition utilizing Kinect as a 3D scanner". In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 66-73. ACM. Accessed July 19, 2016. http://dx.doi.org/10.1145/2695664.2695807.

Reynolds, Douglas A., Thomas F. Quatieri and Robert B. Dunn. 2000. "Speaker verification using adapted gaussian mixture models". *Digital Signal Processing* no. 10 (1):19-41. Accessed July 20, 2016. http://dx.doi.org/10.1006/dspr.1999.0361.

Schwaninger, Adrian, Sandra Schumacher, Heinrich Bülthoff and Christian Wallraven. 2007. "Using 3D computer graphics for perception: the role of local and global information in face processing". In *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, 19-26. ACM. Accessed July 20, 2016. http://dx.doi.org/10.1145/1272582.1272586.

Tanaka, H. T., M. Ikeda and H. Chiaki. 1998. "Curvature-based face surface recognition using spherical correlation. Principal directions for curved object recognition". In *Proceedings: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998.*, 372-377. IEEE. Accessed July 20, 2016. http://dx.doi.org/10.1109/AFGR.1998.670977.

Tang, Y., X. Sun, D. Huang, J. M. Morvan, Y. Wang and L. Chen. 2015. "3D face recognition with asymptotic cones based principal curvatures". In *2015 International Conference on Biometrics (ICB)*, 466-472. IEEE. Accessed July 20, 2016. http://dx.doi.org/10.1109/ICB.2015.7139111.

Tsalakanidou, F., D. Tzovaras and M. G. Strintzis. 2003. "Use of depth and colour eigenfaces for face recognition". *Pattern Recognition Letters* no. 24 (9–10):1427-1435. Accessed July 20, 2016. http://dx.doi.org/10.1016/S0167-8655(02)00383-5.

Yuxiao, Hu, Jiang Dalong, Yan Shuicheng, Zhang Lei and zhang Hongjiang. 2004. "Automatic 3D reconstruction for face recognition". In *Proceedings: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.*, 843-848. IEEE. Accessed July 20, 2016. http://dx.doi.org/10.1109/AFGR.2004.1301639.