Nonlinear Topological Component Analysis: Application to Age-Invariant Face Recognition

1
Author(s)

Bouchaffra, D. ; Division of Design of Intelligent Machines, Center for Development of Advanced Technologies, Algiers 16303, Algeria.

We introduce a novel formalism that performs dimensionality reduction and captures topological features (such as the shape of the observed data) to conduct pattern classification. This mission is achieved by: 1) reducing the dimension of the observed variables through a kernelized radial basis function technique and expressing the latent variables probability distribution in terms of the observed variables; 2) disclosing the data manifold as a 3-D polyhedron via the (alpha)-shape constructor and extracting topological features; and 3) classifying a data set using a mixture of multinomial distributions. We have applied our methodology to the problem of age-invariant face recognition. Experimental results obtained demonstrate the efficiency of the proposed methodology named nonlinear topological component analysis when compared with some state-of-the-art approaches.

Published in:
Neural Networks and Learning Systems, IEEE Transactions on (Volume:PP , Issue: 99)

Page(s):
1

Date of Publication :
11 August 2014

ISSN :
2162-237X

Publisher:
IEEE

DOI:
10.1109/TNNLS.2014.2341634

Nonlinear Topological Component Analysis: Application to Age-Invariant Face Recognition

Djamel Bouchaffra[1], *Senior Member, IEEE*

***Abstract*—We introduce a novel formalism that performs dimensionality reduction and captures topological features (such as the shape of the observed data) to conduct pattern classification. This mission is achieved by: (i) reducing the dimension of the observed variables through a kernelized radial basis functions (KRBF) technique and expressing the latent variables probability distribution in terms of the observed variables; (ii) disclosing the data manifold as a 3D polyhedron via the α-shape constructor and extracting topological features; and (iii) classifying a data set using a mixture of multinomial distributions. We have applied our methodology to the problem of age-invariant face recognition. Experimental results obtained demonstrate the efficiency of the proposed methodology named Nonlinear Topological Component Analysis (NTCA) when compared to some state-of-the-art approaches.**

***Index Terms*—nonlinear mapping, topological manifold, kernel trick, α-shape signatures, multinomial mixtures, face recognition across ages, facial aging.**

I. INTRODUCTION

Principal component analysis (PCA) remains an invaluable tool when one acquires measures on a number of observed variables and wish to generate a smaller number of artificial variables (called principal components) that will account for most of the variance. This variable reduction is motivated by the belief that some observed variables measure the same construct and therefore exhibit significant correlations. The principal components may then be utilized as a predictor set in subsequent analyses. PCA has been exploited in a nonlinear way by means of the kernel trick to derive Kernel-based PCA (KPCA) [1]. Other prominent nonlinear techniques include manifold learning such as Isomap [2], Locally Linear Embedding (LLE) [3], Hessian LLE [4], Laplacian Eigenmaps [5], and Local Tangent Space Alignment (LTSA) [6] have been proposed in the machine learning literature. These techniques build a low-dimensional data representation using a cost function that maintains local properties of the data. These approaches can be perceived as defining a graph-based kernel for KPCA. Likewise, graph and network embedding techniques have also been introduced as a general framework for dimensionality reduction for the purpose of visualization [7], [8], [9], [10], [11], [12]. Other formalisms such as "Generative Topographic Mapping" (GTM) [13] have been put forward as an attempt to reveal the inverse nonlinear relationship between the set of latent variables and the set of observed variables. Unlike PCA, these latter mappings assume an *underlying causal structure* within the observed variables and thus represent a building block of an exploratory factor analysis (FA). However, most of these techniques are oriented towards the search of a dimension reduction criterion; they have difficulty providing a means that permits the disclosure of the low-dimensional shape formed by the data set. Unveiling shapes would allow the exploitation of mathematical concepts such as homeomorphism, homotopy equivalence and topological invariance [14] that are precious in many pattern recognition problems such as classification of 3D protein folds or brain repair after injury.

The main motivation of this research consists of extending dimensionality reduction techniques to derive topological features (e.g., as the shape of the observed data) and to merge them with statistical information. The approach that we propose is viewed as: (1) an integration of a kernelized radial basis functions (KRBF) dimensionality reduction method, (2) a construction of α-shapes and extraction of topological features (signatures); and (3) an object classification based on a mixture of multinomial distributions. One would be able to gain an insight into the morphology of shapes formed by the data set in a latent space for a sharper classification task. We show how statistical information can be merged seamlessly with topological knowledge using a multinomial mixture probability distribution. This allows comparing several shapes in the latent variable space. The entire methodology that we have named "Nonlinear Topological Components Analysis" (NTCA) is undertaken by: (i) mapping the original data set to a Hilbert space using a quadratic function: This allows to obtain a more accurate separation of the observations; (ii) mapping the obtained set to a set spanned by kernelized radial basis functions, (iii) projecting linearly this latter set on a latent space; (iv) building a hierarchy of α-shapes associated to the points of the latent space; and (iv) extracting topological signatures from the α-shapes for the purpose of classification. In other words, (i), (ii), and (iii) represent a dimensionality reduction technique (that we have named: "Kernelized Radial Basis Functions (KRBF)) whereby each point of the original data set is mapped nonlinearly to a point in a low dimensional latent vector space. The cloud of points obtained in the latent space is assigned a hierarchy of α-shapes that represents the original data set. Finally, a group of topological traits are extracted

[1] Dr. D. Bouchaffra is currently a Director of Research and the Head of the Division "Design of Intelligent Machines" located at the Center for Development of Advanced Technologies, Algiers (e-mail: dbouchaffra@ieee.org). He was a former CS Professor in Michigan, USA.

by averaging over the α-shapes to form a template for object identification. Applications of the NTCA formalism are diverse: Biometrics and Forensics can benefit from NTCA since the original data set depicting a topological manifold can be built from different images of the same biometric sample (e.g., fingerprint, face) of an individual. The proposed approach maps this multidimensional manifold to a 3D α-shape which is a 3D-polyhedron (n-polytope). Another application would consist of reconstructing 3D-MRI images of damaged organs (such as the brain) in order to identify the causes of the injury, and find its location using the geometrical concept of α-shapes.

The organization of this manuscript is described as follows: Section II addresses some basics about the kernel functions and the notion of α-shapes. The concept of nonlinear topological component analysis that covers the data distribution in the latent space is introduced in Section III. The parameter estimation of the latent space data distribution is carried out in Section IV. The object classification phase based on multinomial mixtures distribution is laid out in Section V. The application in face identification across ages and the experiments conducted are the object of Section VI. Finally, the conclusion and directions for future work are presented in Section VII.

II. SOME BASICS

In order to gain insight into the proposed methodology, we propose to introduce some basics about kernels and the α-shapes geometrical constructor.

A. *Kernel basics*

We consider the data set $\mathfrak{D}$ as a set composed of n points $t_1,t_2,\ldots,t_n$, in which each $t_i \in \mathbb{R}^D$.

Definition 1. A kernel k(.,.) on a domain $\mathfrak{D}$ is a real positive function k: $\mathfrak{D} \times \mathfrak{D} \rightarrow \mathbb{R}^+$ that is *continuous*, *symmetric* and *positive semi-definite* in a certain sense.

Let's denote the data matrix as: $\mathbf{D} = (D^{(1)}|D^{(2)}|,\ldots,|D^{(D)})$ in which each column of **D** corresponds to a variable $D^{(i)}= (t_{1i},t_{2i},\ldots,t_{ni}) \in \mathbb{R}^n$, and each row is a data point $t_i = (t_{i1},t_{i2},\ldots,t_{iD}) \in \mathbb{R}^D$. Given a set of data points $t_1,t_2,\ldots,t_n \in \mathfrak{D}$, and the (n x n) matrix $K_{i,j} = k(t_i,t_j)$ (called the Gram matrix with respect to $t_1,t_2,\ldots,t_n$), we have the following definition:

Definition 2. The kernel function k is positive definite, if $\forall n \in \mathbb{N}$, its Gram matrix with respect to all $t_1,t_2,\ldots,t_n$ is a positive definite matrix, i.e.,

$$\forall(\lambda_1,\lambda_2,\ldots,\lambda_n) \in \mathbb{R}^n, \sum_{i,j}\lambda_i\lambda_j K_{i,j} \geq 0, \text{ and if}$$
$$\sum_{i,j}\lambda_i\lambda_j K_{i,j} = 0 \Rightarrow \lambda_1 = \lambda_2 \ldots = \lambda_n = 0.$$

Theorem. [Kernel Trick Theorem]. Given a positive definite kernel k(.,.) on a domain $\mathfrak{D}$, one can build a mapping φ from $\mathfrak{D}$ into a Hilbert space $\mathcal{H}_k$, in which the function k represents a dot product <,> such that:

$$<\varphi(t_1), \varphi(t_2)> = \varphi(t_1)^T.\varphi(t_2) = k(t_1,t_2), \forall(t_1,t_2) \in \mathfrak{D}^2.$$

The proof of this theorem can be found in [15].

The idea behind the kernel trick is to replace a nonlinear model defined in the original space by a linear model defined in the target space. The original space is mapped to the target space using a nonlinear transformation based on suitably chosen basis functions. Therefore, the inner product of basis functions $\varphi(t_1)^T.\varphi(t_2)$ is replaced by a kernel function $k(t_1,t_2)$, between instances in the original input space. So, instead of mapping two instances t_1 and t_2 to the Hilbert space and performing a dot product there, one directly applies the kernel function in the original space. Amongst the most used kernels, one can cite the Gaussian kernel $k(x,y) = k(\|x-y\|) = \exp\{-(x-y)^T\Sigma^{-1}(x-y)\}$ which is an example of radial basis function kernel and the polynomial kernel $k(x,y) = (ax^T.y + c)^d$, where a is the slope parameter, c is a constant term and d is the polynomial degree. There are several other kernels proposed in the literature [16], however, it is important to underscore that the choice of a "best kernel" depends on the type and structure of the data at hand. For example there exist kernels for similarities [17], strings [18], graphs [19] and other data structures.

B. *Alpha-shapes formalism*

The concept of α-shapes represents an efficient means for creating shapes at different levels of resolution α out of point sets [20], [21]. The α-shapes' formalism can be viewed as a "soft" generalization of the standard notion of convex hull assigned to a set of points. In the sequel, we provide some basic definitions that explain briefly α-shapes and how they are constructed.

Definition 3. Let E be a finite set of points in $\mathbb{R}^3$, the Voronoi cell of $u \in E$ (or Voronoi Complex) is defined as:

$$V_u = \{x \in \mathbb{R}^3 : \forall v \in E - \{u\}, \|x-u\| \leq \|x-v\|\}.$$

The set of V_u represents convex polyhedral or an empty set since the set of points that are equidistant from two points in E forms a hyperplane.

Definition 4. Let E be a finite set of points in $\mathbb{R}^3$, the union of balls centered at the points in E with radius $\sqrt{\alpha}$ for $\alpha \geq 0$ is denoted by $B^\alpha(E)$, i.e.

$$B^\alpha(E) = \{ x \in \mathbb{R}^3 : \exists u \in E \text{ such that } \|u-x\|^2 \leq \alpha\}.$$

Definition 5. Let E be a set of points in $\mathbb{R}^3$, and $\alpha \geq 0$, the α-complex of E is the dual complex of the Voronoi diagram of E restricted to the union of balls $B^\alpha(E)$. The restricted Voronoi cell of $u \in E$ is: $V_u^\alpha = V_u \cap B^\alpha(E)$.

The α-shape is the underlying space corresponding to the α-complex. The α-shape concept represents a formalization of the intuitive notion of "shape" for spatial point set data. *An α-shape is a concrete geometric object that is uniquely defined for a particular set of points*. The parameter α controls the *desired level of details of the shape*. There are several algorithms that construct a family of α-shapes for a

given set of size n in a worst-case time complexity equal to $O(n^2)$ [21]. The α-shapes define a hierarchy of shapes from a set of points that allows features multiscale modeling that are very useful in macromolecule structure exploration as well as in facial aging (identifying changes of human facial compartments: a human face is made of compartments of fat). The α-shapes insert a ball of radius $\sqrt{\alpha}$ around each point and build a simplicial complex that respects the intersections among these balls. The simplicial space formed is defined as the α-shapes.

α-Shapes Construction: The construction of α-shapes is based on the notion of "ball" known also as "generalized disk". Given a set of points and a specific α value, the α-shape is constructed using the following scheme:

1. Each point X_u in the embedded set is assigned a vertex u.
2. An edge is created between two vertices u and v whenever there exists a generalized disk of radius $\sqrt{\alpha}$ containing the entire set of points and which has the property that X_u and X_v lie on its boundary.

However, the notion of generalized disk is defined as follows:

- If α > 0 but very large then the generalized disk is a half-plane (very large radius!).
- If α > 0 but not large then the generalized disk is a closed ball of radius $\sqrt{\alpha}$.
- If α < 0 the generalized disk is the complement of a closed ball of radius $\sqrt{-\alpha}$.

An example that depicts sub-images of the letter "a" at different levels of resolution (details) corresponding to different α values is shown in Fig. 1.

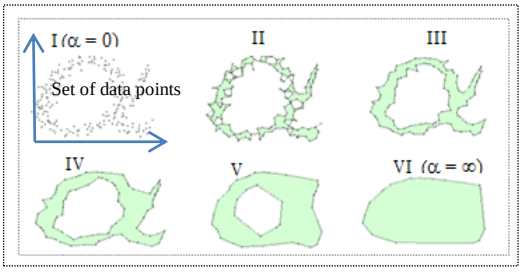

Fig. 1. A discrete set of data points that depicts the α-shape associated to the letter "a" from finer (α = 0) to cruder (α = ∞) levels of details.

As α increases more and more vertices are connected in the graph of these sub-images which results in the obtention of finer shapes (at high resolution). This process proceeds until the value of α becomes very large (very large radius in which the generalized disk is a half-plane) and cruder shapes start to appear (low resolution). There are different types of signatures assigned to α-shapes one can capture. There are those which are *metric-based* (e.g., volume, area and length), *combinatorial-based* (e.g., number of tetrahedral) and *topological-based*: (e.g., number of voids, number of gaps, each known as a *hole or a genus*). Fig. 2 pinpoints these topological features.

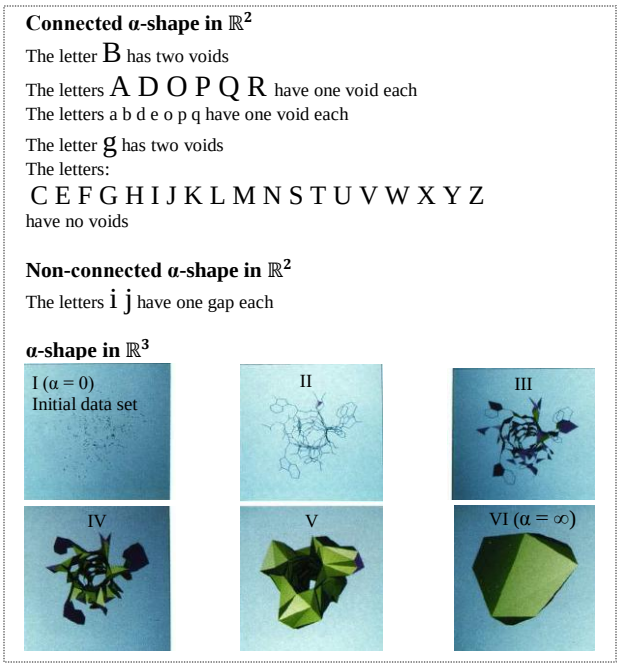

Fig. 2. Voids, and gaps formed by the alphabetical letters. An α-shape in $\mathbb{R}^3$ can have voids and gaps, but it can also have tunnels.

III. KERNELIZED RADIAL BASIS FUNCTIONS DIMENSIONALITY REDUCTION

A. *Mapping data space to latent space*

The mission in KRBF is to find a representation for the distribution p(x) of low dimensional latent variables in terms of a number D of data variables $t = (t_1, t_2, \ldots, t_D)$. The computation of this distribution allows the creation of the set of points $\mathcal{X}$ that represents the data set in the latent space. The assumption of the existence of a low-dimensional data manifold governed by some latent variables is explained by the fact that: (i) data exhibit substantial correlations between the variables and (ii) some data are noisy due to defective sensors. A nonlinear mapping between the original data set and the latent variable space $\mathcal{X}$ is performed by first mapping the data set $\mathfrak{D}$ of $\mathbb{R}^D$ to a subset $\mathfrak{D}_e$ (e stands for extended) of a high dimensional Hilbert space $\mathcal{H}_k$ (endowed with a positive definite kernel function k on $\mathfrak{D}$) using a feature space mapping φ. The subset $\mathfrak{D}_e$ is in turn mapped to a subset B spanned by radial basis functions and finally B is projected on the latent space $\mathbb{R}^L$ via a nonlinear transformation χ to derive the set $\mathcal{X}$. The image x of an element $t \in \mathfrak{D}$ is computed as: $x = (\chi \circ \pi \circ \varphi)(t)$, whereby $z = \varphi(t) \in \mathfrak{D}_e \quad (t \in \mathfrak{D})$, and $\chi(\pi(z)) = \boldsymbol{W}.\pi(z)$, where $\pi = (\pi_i)$ (i=1,…,Q) designates a set of fixed nonlinear basis functions and **W** is an (LxQ) matrix of parameters. Finally, the α-shape geometric descriptor is invoked to compute a family of L-dimensional α-shapes (defined in $\mathcal{A}^{(\alpha)}$) assigned to the subset of data points $\mathcal{X}$ (refer to Fig. 3). The dimensionality L of the latent space is less than the dimensionality D of the data space and therefore less than the dimensionality H of the Hilbert space. Topological signatures extracted from the α-shapes (polyhedra)

characterize the data manifold. These signatures can be used for a classification or a regression task (refer to Fig. 4).

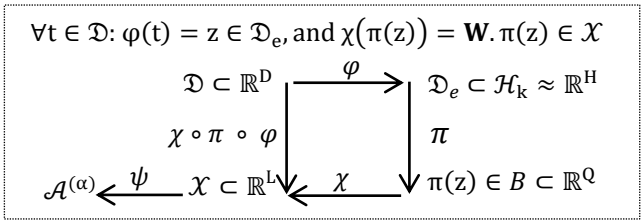

Fig. 3. Nonlinear dimensionality reduction mapping (KRBF) whose latent variable data are used to build the α-shapes of $\mathcal{A}^{(\alpha)}$.

It should be borne in mind that the motivation for this work is twofold: (i) it provides a 3D-visualization in the form of α-shapes (in the latent space) of the data topological manifold. (ii) It creates a methodology that merges seamlessly statistical data and topological features within a single probabilistic framework.

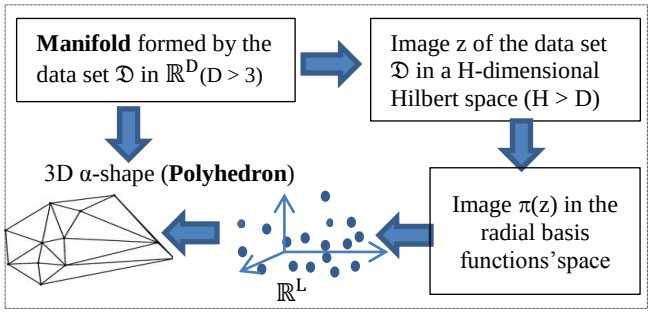

Fig. 4. Nonlinear mapping between the data topological manifold and an α-shape (3D polyhedron) via a latent space.

B. *Latent space probability distribution*

We start by defining a probability distribution p(t) in which $t \in \mathfrak{D}$, this will infer a corresponding probability distribution p(x) that generates the set $\mathcal{X}$ in the latent space $\mathbb{R}^L$:

$$p(x) = \int_{\mathfrak{D}} p(x|t, \mathbf{W}, \beta) p(t) dt \,. \quad (1)$$

If one chooses a radial symmetric Gaussian centered on $\chi(\varphi(t))$ for the conditional density $p(x|t, \mathbf{W}, \beta)$ (noise model), then:

$$p(x|t, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{L|2} \exp\left\{\frac{-\beta}{2} \|\chi(\varphi(t)) - x\|^2\right\}.$$

Since $z = \varphi(t)$ and $\chi(z) = \mathbf{W}.\pi(z)$, then one can write:

$$p(x|t, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{L|2} \exp\left\{\frac{-\beta}{2} \|\mathbf{W}.\pi(z) - x\|^2\right\}. \quad (2)$$

If one selects the radial basis functions centered on z_{c_i} for $\pi(z)$, then $\pi(z) = (\pi_i(z)) = (\exp\{-\epsilon_i \|z - z_{c_i}\|^2\})$, (3) (i=1,…,Q), where: $z_i = \varphi(t_i)$ and $\epsilon_i = 1/N^2 Max_j^2(z_{ij})$. The vectors z_{c_i} represent a set of Q centroids obtained during a clustering process of the z_i vectors using OPTICS and k-means [22]. Using the kernel trick theorem, the nonlinear basis functions can be written as:

$$\pi_i(z) = \exp\{-\epsilon_i[k(t,t) - 2k(t,t_i) + k(t_i,t_i)\}, \quad (4)$$

with: $k(t, t_i) = (t^T.t_i)^2 = \langle z, z_i \rangle = \langle \varphi(t), \varphi(t_i) \rangle$ and

$$\varphi(t) = [t_1^2, t_2^2, \dots, t_D^2, \sqrt{2}\, t_1 t_2, \dots, \sqrt{2}\, t_1 t_D, \sqrt{2}\, t_2 t_D, \dots, \sqrt{2}\, t_{D-1} t_D]^T. \quad (5)$$

We can assume that the prior p(t) is a sum of M delta functions:

$$p(t) = \frac{1}{M} \sum_{l=1}^{l=M} \delta(t - t_l), \quad (6)$$

where $\{t_l\}$ is a set of points on a regular grid in the data space. Therefore, equation (1) can be written:

$$p(x) = \int_{\mathfrak{D}} p(x|t, \mathbf{W}, \beta) \frac{1}{M} \sum_{l=1}^{l=M} \delta(t - t_l) dt. \quad (7)$$

This latter expression can be written as:

$$p(x) = p(x|\mathbf{W}, \beta) = \frac{1}{M} \sum_{l=1}^{l=M} p(x|t_l, \mathbf{W}, \beta). \quad (8)$$

Finally, the density model is expanded into:

$$p(x) = p(x|\mathbf{W}, \beta)$$

$$= \frac{1}{M} \sum_{l=1}^{l=M} \left(\frac{\beta}{2\pi}\right)^{L|2} \exp\left\{\frac{-\beta}{2} \|\mathbf{W}.\pi(z_l) - x\|^2\right\}. \quad (9)$$

IV. KRBF PARAMETERS

A. *Expectation-Maximization Parameters Estimation*

For a given latent variable set $\mathcal{X}=\{x_1,\dots,x_N\}$, one can determine the parameter matrix **W**, and the inverse variance using maximum-likelihood estimation. It is often convenient to maximize the log-likelihood:

$$\mathcal{L}(\mathbf{W}, \beta) = \ln\left[\prod_{i=1}^{i=N} p(x_i|\mathbf{W}, \beta)\right], \quad (10)$$

which is equivalent to:

$$\mathcal{L}(\mathbf{W}, \beta) = \sum_{i=1}^{i=N} \ln[p(x_i|\mathbf{W}, \beta)]. \quad (11)$$

Finally, the likelihood function that needs to be maximized with respect to **W** and β is:

$$\mathcal{L}(\mathbf{W}, \beta) =$$

$$\sum_{i=1}^{i=N} \ln\left[\frac{1}{M} \sum_{l=1}^{l=M} \left(\frac{\beta}{2\pi}\right)^{L|2} \exp\left\{\frac{-\beta}{2} \left\|\sum_j w_{ij} \exp\left[-\epsilon(\|t_l\|^2 - 2(t_l^T.t_j) + \|t_j\|^2\right] - x\right\|^2\right\}\right]. \quad (12)$$

Since the density model consists of a sum of M mixtures, the EM algorithm is more suitable for parameter estimation. The maximization of the likelihood $\mathcal{L}(\mathbf{W}, \beta)$ can be viewed as a missing data problem in which the mixture j which generated a latent variable is unknown. In the E-step of the EM algorithm, we use $\mathbf{W}^c$ and β_c (the superscript c stands for current) to evaluate the responsibilities Resp. (posterior

probabilities) of each Gaussian mixtures j for every latent point x_n associated to a data point t_n using Bayes' rule:

$$\underset{j\to x_n}{\mathrm{Resp}}(\mathbf{W}^c,\beta^c) = p(t_j|x_n,\mathbf{W}^c,\beta^c)$$

$$= \frac{p(x_n,\mathbf{W}^c,\beta^c|t_j).\,p(t_j)}{p(x_n,\mathbf{W}^c,\beta^c)} \quad (13)$$

$$= \frac{p(x_n|\mathbf{W}^c,\beta^c,t_j).\,p(\mathbf{W}^c,\beta^c|t_j).\,P(t_j)}{\sum_{p=1}^{p=M} p(x_n|t_p,\mathbf{W}^c,\beta^c).\,p(t_p,\mathbf{W}^c,\beta^c)} \quad (14)$$

$$= \frac{p(x_n|\mathbf{W}^c,\beta^c,t_j).\,p(\mathbf{W}^c,\beta^c,t_j)}{\sum_{p=1}^{p=M} p(x_n|t_p,\mathbf{W}^c,\beta^c).\,p(t_p,\mathbf{W}^c,\beta^c)}. \quad (15)$$

We now consider the expectation of the complete data log-likelihood in the form:

$$\mathcal{L}_{\mathrm{complete}}(\mathbf{W},\beta) = \sum_{n=1}^{n=N}\sum_{i=1}^{i=M} \underset{i\to x_n}{\mathrm{Resp}}(\mathbf{W}^c,\beta^c).\ln\{p(x_n|t_i,\mathbf{W},\beta)\}. \quad (16)$$

Maximizing (1) with respect to the variable W, one obtains:

$$\mathcal{L}_{\mathrm{complete}}(\mathbf{W},\beta) = \sum_{n=1}^{n=N}\sum_{j=1}^{j=M} \frac{p(x_n|\mathbf{W}^c,\beta^c,t_j).\,p(\mathbf{W}^c,\beta^c,t_j)}{\sum_{p=1}^{p=M} p(x_n|t_p,\mathbf{W}^c,\beta^c).\,p(t_p,\mathbf{W}^c,\beta^c)}.\ln\{p(x_n|t_j,\mathbf{W},\beta)\}$$

with the following expression for: $p(x_n|t_j,\mathbf{W},\beta)$:

$$p(x_n|t_j,\mathbf{W},\beta) = \left(\frac{\beta}{2\pi}\right)^{L|2}.\exp\left\{\frac{-\beta}{2}\left\|\mathbf{W}.\pi(z_j)-x_n\right\|^2\right\}. \quad (18)$$

This latter expression can be written as:

$$\sum_{n=1}^{n=N}\sum_{i=1}^{i=M} \underset{i\to x_n}{\mathrm{Resp}}(\mathbf{W}^c,\beta^c)\{\mathbf{W}^{\mathrm{next}}\pi(z_i)-x_n\}\pi^{\mathrm{T}}(z_i) = 0 \quad (19)$$

with: $z_i = \varphi(t_i)$. Finally, this equation is translated into the following matrix form:

$$\mathbf{\Pi}^{\mathrm{T}}.\mathbf{\Lambda}.\mathbf{\Pi}.(\mathbf{W}^{\mathrm{next}})^{\mathrm{T}} = \mathbf{\Pi}^{\mathrm{T}}.\mathbf{R}^c.\mathbf{X}^c, \quad (20)$$

- the matrix $\mathbf{\Pi}$ is a M x Q matrix with elements $\pi_{ij} = \pi_j(z_i)$,
- the matrix $\mathbf{\Lambda}$ is an M x M diagonal matrix: $\mathbf{\Lambda}_{jj} = \sum_{n=1}^{n=N} \mathrm{Resp}_{j\to x_n}(\mathbf{W},\beta)$,
- the matrix $\mathbf{W}$ is a L x Q matrix,
- the matrix $\mathbf{R}$ is a M x N matrix with elements $\underset{i\to x_n}{\mathrm{Resp}}$,
- the matrix $\mathbf{X}$ is a N x L with elements x_{ij}.

Solving equation (20) for $(\mathbf{W}^{\mathrm{next}})^{\mathrm{T}}$, yields:

$$(\mathbf{W}^{\mathrm{next}})^{\mathrm{T}} = (\mathbf{\Pi}^{\mathrm{T}}.\mathbf{\Lambda}.\mathbf{\Pi})^{-1}.\mathbf{\Pi}^{\mathrm{T}}.\mathbf{R}^c.\mathbf{X}^c. \quad (21)$$

Similarly, the maximization of the complete log-likelihood with respect to β yields the following equation:

$$\beta^{\mathrm{next}} = \left(\frac{1}{N.L}\sum_{n=1}^{n=N}\sum_{k=1}^{k=M} \mathrm{Resp}_{k\to x_n}(\mathbf{W}^c,\beta^c)\|\mathbf{W}^{\mathrm{next}}.\pi(z_k)-x_n^c\|^2\right)^{-1}. \quad (22)$$

The initial values for the parameters of the matrix $\mathbf{W}$ are computed so that initially the NTCA model's performance gets as close as possible to the PCA's performance.

B. Number of Parameter in KRBF

We define in this section all parameters involved in the KRBF dimensionality reduction method. We also show how they are computed.

- D is the dimension of the input data vectors t_i's; this dimension value depends on the application at hand.
- H = D(D+1)/2 represents the dimension of the vector z_i: it is the sum of an arithmetic series.
- Q is the dimension of the vectors $\pi(z_i)$ which consists of the numbers of groups obtained via a clustering procedure of the vectors z_i's. It corresponds to the cardinality of the radial basis set.
- The initial elements of the (LxQ) matrix $\mathbf{W}$ are the L eigenvectors of the covariance matrix of the vectors $\pi_i(z)$ computed using principal components analysis (minimize information loss during the mapping from the $\pi_i(z)$ in the Q-dimensional space to the new L (L<Q) -dimensional space $\mathbb{R}^L$).
- L is the dimension of the space generated by the x_i vectors; its value is set to 3 for visualization purposes.
- $\epsilon_i = 1/N^2 \mathrm{Max}_j^2(z_{ij})$, for scaling the exponential.
- M= N is the number of vectors $t_i = |\mathfrak{D}|$.
- β is initially set to 1.

Fig. 5 depicts the KRBF algorithm with all its steps.

KRBF Algorithm

Begin

1. Compute the vectors t_i (i=1,...,N) representing the data using a feature extractor technique.
2. Cluster the vectors z_i's into Q groups using OPTICS and k-means.
3. Define the initial matrix $\mathbf{W}$ (using a linear mapping such as PCA as explained above) and the initial β parameter.
4. Compute the initial set $\mathcal{X}$ whose matrix is: $\mathbf{X} = \mathbf{W}.\mathbf{\Pi}$

Repeat

5. Compute $(\mathbf{W}^{\mathrm{next}})^{\mathrm{T}}$, $\mathbf{X}^{\mathrm{next}}$ and β^{next} using equation (21) and (22), respectively

Until convergence // optimal parameters $\mathbf{W}^*$ and β^* are obtained.

6. Store the optimal matrix $\mathbf{X}$

End.

Fig. 5. The different steps involved in the KRBF algorithm.

V. TOPOLOGICAL FEATURE EXTRACTION

As outlined in section II, the mission in this step consists of exploiting dimensionality reduction techniques to reveal the α–shape low-dimensional layout (3D polyhedron) of the data

topological manifold living in $\mathbb{R}^D$. The thrust consists of unwrapping this data manifold's shape in the latent space $\mathbb{R}^L$ (L is often equal 3 for visualization) and represent these data in terms of topological features extracted from this α-shape. *Any modification in the data manifold is viewed as a deformation of this L-dimensional α-shape polyhedron*. One can extract "signatures" of α-shapes such as *metric properties*: (volume, area and length), *combinatorial properties*: (number of tetrahedral, number of triangles, number of edges, number of vertices) and *topological properties*: (number of components, number of independent tunnels, number of voids, number of gaps). These signatures (or features) characterize an α-shape. A "void" might be encountered in a connected α-shape in $\mathbb{R}^2$, whereas a "gap" is encountered only if we allow for non-connected shapes. However, the notion of "tunnel" (or open doors) exists only in the 3D α-shape context. All these signatures are computed as average over a set of predefined α values.

VI. Classification/Testing

Beside the fact that NTCA is endowed with the power of representing the input data set in a low-dimensional latent space; it also allows classification of this data set when this set defines a physical object. The α-shape formalism provided a way to represent any set by its shape and thereby renders classification feasible. The NTCA classification consists of assigning one and only one class ω* (from a set of c predefined classes in Ω) to this α-shape signature pattern. Given a set of data points $\mathfrak{D}$, the NTCA methodology starts by invoking the KRBF dimensionality reduction procedure. This latter requires the estimation of the probability density function p(x) defined in the latent space $\mathcal{X}$. Once the set $\mathcal{X} \subset \mathbb{R}^L$ assigned to the set $\mathcal{D}$ is generated in the latent space, the α-shapes geometric constructor is invoked to draw the 3D-polyhedron associated with the nonlinear topological manifold representing the data set $\mathfrak{D}$. Finally, topological signatures inherent to the manifold's α-shape characterizing the input object to classify are extracted and stored into a set S. The classification problem can be recast into: Determine the class ω* among c target classes (from a gallery) assigned to an input (or probe) data point set $\mathcal{D} = \{t_1, \dots, t_N\} \subset \mathbb{R}^D$ such that:

$$\omega^* = argmax_{\omega_i} P[\omega_i | t_1, t_2, \dots, t_N]. \quad (23)$$

Because the data set $\mathcal{D}$ is mapped nonlinearly to the latent set $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathbb{R}^L$, this classification problem is equivalent to the determination of the class ω* such that:

$$\omega^* = \text{argmax}_{\omega_i} P[\omega_i | \{x_1, x_2, \dots, x_N\}], \quad (24)$$

where the set $\{x_1, x_2, \dots, x_N\}$ is generated using the KRBF algorithm. However, because our goal is to disclose topological information conveyed by the α-shapes, a set of signatures are extracted for the purpose of classification. It is worth highlighting the fact that the entire subset $\mathcal{X} \subset \mathbb{R}^L$ is mapped to only one α-shape whose topological signature set S_j^α contains average values of signatures of the α-shape at some different resolutions α. Finally, one can restate the classification problem with respect to the signatures set as follows: Determine the class ω* among c target classes assigned to a set of points $\mathcal{D} = \{t_1, \dots, t_N\} \subset \mathbb{R}^D$ such that:

$$\omega^* = argmax_{\omega_i} P[\omega_i | S_j^\alpha], \quad (i, j = 1, \dots, c), \alpha \geq 0. \quad (25)$$

Equation (25) requires the estimation of the posterior probability model $P[\omega_i | S_j^\alpha]$. Through Bayes' rule; one can transform the problem into an estimation of the class conditional probability $P[S_j^\alpha | \omega_i]$:

$$P[\omega_i | S_j^\alpha] = \frac{P(S_j^\alpha | \omega_i).P(\omega_i)}{\sum_{i=1}^{i=c} P(S_j^\alpha | \omega_i).P(\omega_i)}. \quad (26)$$

To obtain an expression of the class-conditional probability, we make an analogy between the model we are investigating and the *word/document model*. In this analogy, an α-shape described by a set of signatures corresponds to a document represented by a set of words. As outlined in Section II, signatures are features such as "independent tunnels", "voids", "triangles", "edges", and "vertices"; they are the words of an α-shape. These words are assumed to be generated independently ("bag of words document model"). This latter assumption favors the use of a *multinomial distribution*. Each signature is a word; therefore, signature histograms are "term frequency" distributions. Likewise, because statistical similarities among the signature histograms of α-shapes designating similar objects are observed, we can assume that signature histograms of similar α-shapes are generated by similar multinomial distributions. *Consequently, a multinomial mixture model should be adequate to model the topological signature distribution in each class* $\omega_i \subset \Omega$. In other words, we make the assumption that the distribution of an α-shape signature set in each class ω_i is constituted of p multinomial distributions.

- Let $A^{(\alpha)} = \{a_1, \dots, a_{|A(\alpha)|}\}$ be the set of α-shapes, each of which is described by topological signatures.
- Let $S_j^\alpha = \{s_1^\alpha, \dots, s_T^\alpha\}$ be the set of topological signatures assigned to the α-shape with a specific α value.

Fig. 6 depicts the topological feature frequency distribution in which n_{ij}^α represents the frequency of the topological signature s_j^α in the α-shape a_i^α at a given α value.

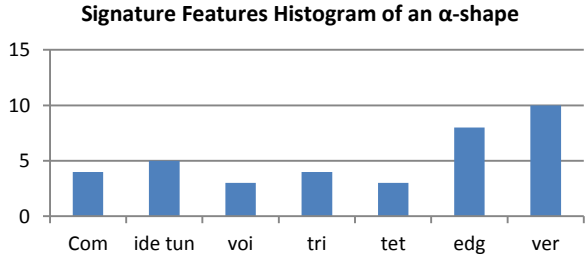

Fig. 6. Unimodal topological signatures histogram (number of: components (com), independent tunnels (ide tun), voids (voi), triangles (tri), edges (edg), vertices (ver)) of an α-shape.

- Let $\mathcal{C} = \{c_1, \dots, c_P\}$ be the set of P mixtures (clusters) within the class ω_i.

The probability $P(a_i^\alpha | \omega_i)$ that the α-shape a_i^α (at resolution α) of a class ω_i is generated by a multinomial mixture is:

$$P(a_i^{\alpha}|\omega_i) = P(S_j^{\alpha}|\omega_i) = \sum_{k=1}^{k=P} P(c_{k,i}) \prod_{j=1}^{j=T} P(s_j^{\alpha}|c_{k,i})^{n_{ij}^{\alpha}}, \quad (27)$$

where:

- $c_{k,i}$ is the mixture coefficient for the k-th mixture in class ω_i.
- T is the number of topological signatures.
- $P(c_{k,i})$ is the probability that the α-shape a_i^{α} assigned to the class ω_i belong to the cluster $c_{k,i}$.
- $P(s_j^{\alpha}|c_{k,i})$ is the probability that the topological signature s_j^{α} appears in the cluster $c_{k,i}$.
- P probabilities $P(c_{k,i})$ and PxT probabilities $P(s_j^{\alpha}|c_{k,i})$ are the parameters of the multinomial mixtures. For each class ω_i; they are subject to:

$$\sum_{k=1}^{k=P} P(c_{k,i}) = 1, \quad \sum_{j=1}^{j=T} P(s_j^{\alpha}|c_{k,i}) = 1 \quad (28)$$

The maximum likelihood estimation consists of maximizing:

$$P(A^{(\alpha)}) = \prod_{i=1}^{i=|A^{(\alpha)}|} P(a_i^{\alpha}|\omega_i), \quad (29)$$

subject to the constraints in the system of equations (28).
The EM algorithm is used for parameters estimation. It is worth noting that this classification problem is enriched by the topological knowledge S_j^{α} revealed by the 3D α-shape polyhedron defined in the latent space. This synergy between probability and topology is exhibited by the NTCA formalism to perform a classification or a regression task.

VII. FACIAL IDENTIFICATION ACROSS AGES

To validate the proposed methodology, we have selected one of the most difficult problems in robust biometrics known as "face identification across ages".

A. Problem Statement and Background

Given a face sample of an individual at age interval A_0; one investigates if this input face is associated with any of a large number of enrolled faces of individuals. However, some face images of the same individual at different age intervals A_i are *among the enrollees.* In other words, given two face images, can one infer that they represent the same individual but at different age intervals? (Refer to Fig. 7). Several features that are responsible for a facial appearance: (i) the 3D structure of human faces, (ii) the reflective peculiarity of facial skin and (iii) the bilateral symmetry in the structure of facial features, in combination with scene-centric attributes such as view-point and illumination. One of the main challenges in face identification stems from the alteration of facial appearances. This modification is explained by the combination of several factors such as illumination, facial expression, occlusions and age. However, the power of face identification across ages is mostly dependent on the ability of modeling some singularities caused by face wrinkles as well as bones structural deformation (eg., jaw bones) of the enrollee face. The process of aging remains the major cause of significant deformation in the appearance (face texture) and anatomy of human faces. Aging is therefore one of the most compelling challenges for automatic human-identification systems and forensics investigation.

Fig. 7. Face images of the same individual; at different age intervals [Afghan Girl: National Geographic June 1985]

Compared to standard face recognition [23], [24], [25], [26], [27], [28] in which aging is not relevant, age-invariant face recognition remains an intricate task due to an imperfect modeling of facial aging. According to Gerontologists, aging is only partially understood. Within this context, Juefei-Xu et al., [29] introduced a technique that utilizes periocular region for age invariant face recognition. They have applied Walsh-Hadamard transform encoded local binary patterns on periocular regions. In [30], Ramanathan et al. presented a survey containing a thorough analysis of the problem of facial aging. Several different paradigms and views of this problem have been described in this report. Yang and Ai [31] laid out a learning scheme that permits the face images classification to be based on their age group. Their approach employs a local binary pattern (LBP) as an image operator and derives the LBP histogram that was utilized for texture characterization. They subsequently selected a sequence of local features and performed age classification. Geng et al., [32] modeled the aging pattern, defined as the sequence of individual's face images at different ages, by building a representative subspace of these images. The unseen face image age is predicted via the projection in this subspace that can rebuild the face image optimally. The position of the face image in that aging pattern will then disclose its age. Park et al. [33] proposed a 3D modeling technique and a simulated method for age-invariant face recognition. They have expanded shape modeling from 2D to 3D domain in order to inject capability for compensating for pose and lighting variations. Li et al. [34] proposed a discriminative model by representing each face via a densely sampled local feature description scheme. They have used a scale invariant feature transform and multi-scale local binary patterns as local descriptors. Finally, Bouchaffra [12], [35], [36] devised a methodology that converts a dynamic Bayesian network into a family of α-shapes and applied a Gaussian mixture model for age-invariant face identification.

A manifold assigned to an individual is described by images (set of points) of the same individual at different ages of his/her life. α-shapes computed at different resolutions capture this manifold through an averaged signature set. Each individual has his/her own α-shape. Hence, the mission consists of: (i) Extracting facial features by applying a traditional facial features extraction method. These facial feature vectors represent the data points of the set $\mathfrak{D} \subset \mathbb{R}^D$; (ii) building for each enrollee the set $\mathcal{X}$ using the probability distribution p(x) of the data manifold; (iii) computing the α-shape (3D polyhedron) assigned to each individual's nonlinear manifold for visualization, and (iv) extracting α-shape topological signatures for a face identification task. In

the next section, we show how these low-level facial features are built to fill out the data set $\mathfrak{D} \subset \mathbb{R}^D$.

B. Facial Feature and Data Set Generation

To illustrate the entire NTCA approach, we extract a set of facial features using the traditional local binary pattern (LBP) method [37]. A face image is first divided into non-overlapping k small rectangular regions. For each pixel in a region, one compares the pixel to each of its 8 neighbors. If the center pixel's value is greater than the neighbor's value, this pixel value is replaced by "1", otherwise, it is replaced by "0". The thresholded neighbors provide an 8-digit binary sequence which is converted into a decimal number (called LBP label), refer to Fig. 8.

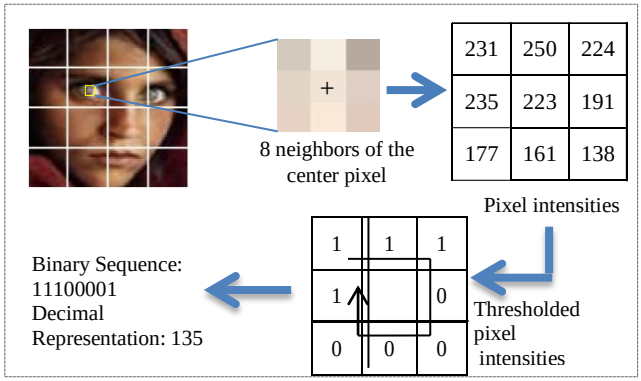

Fig. 8. Basic LBP operator: The neighbors are thresholded, and their binary sequence is converted into a decimal LBP label.

256 possible patterns ranging from 0 to 255 can be extracted in the case of 8 neighbors. A LBP bit sequence corresponding to thresholded pixel intensities matrix is said to be *uniform* if the bits change less than two times (e.g., 11100001 has two changes, between the third and the forth bit and between the seventh and the eighth bit). This operator has been extended whereby different neighborhood sizes are considered. In general, $LBP_{P,R}$ refers to the neighbors size of P equally spaced pixels on a circle of radius R that form a circularly symmetric neighbor set. As these points do not necessarily fall in the center of a pixel of the image, their values are obtained using bilinear interpolation (on-grid vs. off-grid), refer to Fig. 9.

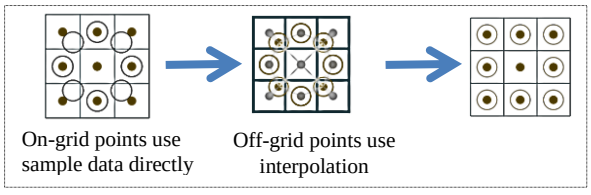

Fig.9. The extended $LBP_{8,1}$ corresponding to P=8 and R=1.

Every region has an associated histogram depicting the frequency of each LBP label. Each histogram has 59 bins: There are 56 bins that correspond to 56 uniform bit sequences that have exactly two bit changes, 2 bins that correspond to the two bit sequences that undergo no change. These two sequences are assigned to bin 0 and bin 58. The non-uniform sequences are assigned to bin 59. The feature vector is a concatenation (putting side by side) of all histograms corresponding to the k regions of an image has therefore 59 components.

C. Experiment

We have used the LBP method available at http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab. This method extracts 59 facial features (D=59), H=59x60/2=1770. The value of L is set to 3, M =N, and Q=3. The number of points N in $\mathfrak{D}$ is set according to the face database considered. The face identification phase was conducted by averaging the topological signatures S_j^α assigned to an individual over 4 α-shapes (4 α-values). Most of the topological signatures have been computed from the online repository [38]. Furthermore, the value for p (number of multinomial mixtures) was set to 4 during classification. Finally, testing was undertaken by computing the class ω_i (individual) whose posterior probability $P[\omega_i|S_j^\alpha]$ is maximum. Experimentation was carried out using MATLAB 7.10 (64-bit) on a 3.20 GHZ Intel(R) Core(TM) i3 CPU and 4GB of RAM. This face identification phase was conducted on standard public domain databases (with and without aging effect) that are presented in the next sections.

C.1. Data Collection

- *Non-Aging Database*

In order to analyze the efficiency of the NTCA model without aging effect, we have conducted a face identification task on a non-aging database in which a set of same age face images is assigned to each individual. We have therefore selected the GeorgiaTech database that comprises 750 images (15 color jpeg at resolution 640x480 pixels) representing 50 people. These individuals are from different races and genders and with different age intervals. Each individual from this database has 15 images of him/herself at the same age with different lighting conditions, facial expressions and head poses. Fig. 10 shows a GeorgiaTech database sample.

Fig. 10. Sample of face images from the GeorgiaTech database.

- *Aging Databases*

In order to investigate whether the NTCA model could identify individuals at different age intervals, we have conducted our experiment on two aging databases: MORPH [39] and FGNET [40]. The MORPH database contains images of adults at different ages. This database is divided into two albums MORPH-album 1 and MORPH-album 2. MORPH-album 1 contains 1690 digitized images of 515 individuals (subjects) that are between 15 and 68 years. MORPH-album 2 (original release) contains 15,204 images of more than 4000 individuals that are in the range 18-50 years. This data set was recently updated by adding new images of the existing individuals and acquiring new individuals with at least 3 images that were not part of the initial album 2. The extended album 2 comprises over 94,000 images of over 24,000 individuals between 16 and 77

years. Both albums contain an average of 3 to 4 images per individual. The FGNET (Face and Gesture Recognition Research Network) aging database contains 1002 images of 82 individuals whose age is in the range 2-69 years with 6 to 18 images per individual. For a better performance analysis, it is desired that all these images should not possess large variations in pose, expression, and illumination. Fig. 11 shows image samples from MORPH and FGNET databases.

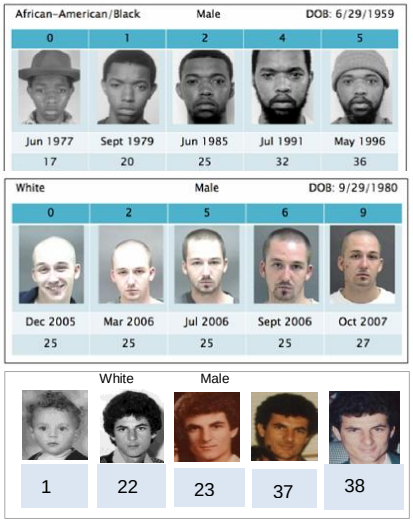

Fig. 11. Sample of face images from: (a) MORPH-album 1, (b) MORPH-album 2, and (c) FGNET. The light blue row in the bottom of each sequence represents the ages. (a) and (b) are extracted from [41].

C.2. Evaluation Criteria

It is worth underscoring' that the process of evaluating the NTCA algorithm is based on three subsets of the entire aging or non-aging databases: (i) *the training set* T, (ii) *the gallery set* G, and (iii) *the probe set* P. The usage of T and P is identical to that of training and test data pertaining to the classical machine learning area. The performance of NTCA is rated with respect to how well the facial images in P are matched to the facial images in G. We can state for example that the performance is flawless on a given probe set P, if every image $I_{i,j} \in P$ is matched to a facial image $I_{i,k} \in G$. In other words, the probe facial image $I_{i,j}$ is matched to a different image $I_{i,k}$ of the same individual: the i^{th} individual. Consequently, the sets G and P should be disjoint. It is often the case that the set G contains no more than one image for each individual enrolled in the database [42]. This is why it is desirable to mention the number of individuals, their age separation, and the number of images in probe and gallery for reproducibility promotion and performance comparison.

C.3. Training and Testing

The face image feature vectors were generated as described in section VI (A.2.) only to build the data set $\mathfrak{D}$ for each individual in the targeted database.

- *Case of GeorgiaTech Database*

The GeorgiaTech database was split into a training set and an independent test set. To learn the NTCA model, we have selected (25x15=375) images from the 25 individuals allocated to training with 15 images per individual. However, the test set was chosen from the remaining 25 individuals. The test set for each individual was divided into a gallery set of 8 images and a probe set of 7 images. Each unique individual's signature set that makes the gallery is built from the 8 images and represent an individual's prototype. However, to obtain a sufficient quantity of points for building the α-shape assigned to a prototype, we have invoked $LBP_{P,R}$ for different values of the radius R with the same neighborhood size P. This variation of the radius allows keeping the same vector dimension. Therefore, for each individual's prototype, we have collected 20 additional feature vectors that produce 7x20=140 points. Likewise, each unique probe signature set of an individual among the remaining 7 images is constructed from 1 image, which also constitutes a tally of 1x20=20 points.

- *Case of MORPH Database*

MORPH-album 1 was partitioned into a training set and an independent test set. The training set was used to learn the NTCA model parameters. We have selected 514 images (257x2) from the 257 individuals allotted to training with 2 images per individual. The age difference between these two images is less than or equal to 7 years. However, the test set was divided into a gallery set and a probe set compiled from the remaining 257 individuals. Each unique individual's signature set of the gallery is extracted from the α-shape that is built based on 1 image. This α-shape contains (1x20=20) points. Furthermore, each unique probe signature set assigned to an individual is built from the remaining 1 image, which also represents an α-shape of (1x20) points. The age difference for the same individual in both sets does not exceed 7 years. Similarly, MORPH-album 2 was divided into a training set and an independent test set. We have chosen 20000 images of 10000 individuals (2 images per individual) for training. Likewise, the remaining 10000 individuals were split into a gallery set and a probe set. The same procedure as in MORPH-album 1 has been adopted to build the unique individual's signature set of the gallery. In other words, each of the probe and gallery signature set has been built based on 1 image (which means an α-shape of 20 points) for each individual. Because we have considered 2 images per individual, both the gallery set and the probe set contain 10000 images each.

- *Case of FGNET Database*

The FGNET database was broken up into a training set and an independent test set. We have used 41 individuals that represent 246 images (41x6) for training the NTCA model. The age difference between two images amongst the 6 images considered for each individual does not exceed 12 years. The remaining 41 individuals were employed for testing. The test set for each individual was divided into a gallery set of 3 images and a probe set of 3 images. Each unique individual's α-shape in the gallery was (i) computed based on 3 images representing 60 points (3x20=60), and (ii) stored as a prototype (refer to Fig. 12).

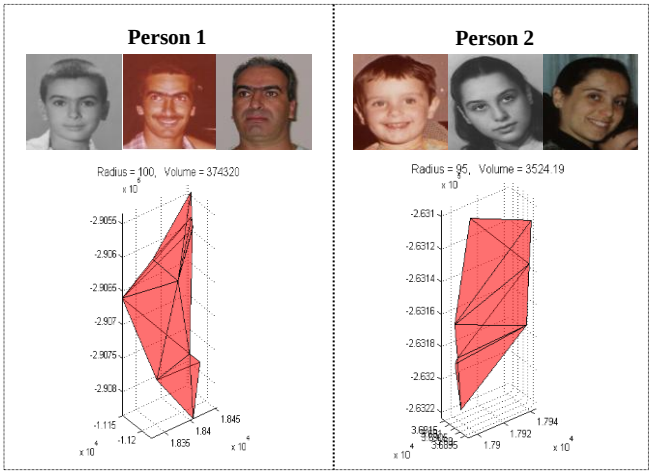

Fig. 12. Alpha-shapes of two different individuals computed and stored as prototypes in the FGNET gallery.

The same procedure was undertaken for the probe set. In conclusion, the total number of images in probe and gallery is 246. The distribution of images in the training, probe and gallery sets per individual is illustrated in Table 1.

TABLE 1

NUMBER OF POINTS USED DURING TRAINING, PROBE AND GALLERY

	MORPH-album 1	MORPH-album 2	FGNET	GeorgiaTech
Training	2x20=40	2x20=40	6x20=120	15x20=300
Probe	1x20=20	1x20=20	3x20=60	7x20=140
Gallery	1x20=20	1x20=20	3x20=60	8x20=160

Finally, testing is conducted the same way within the three databases-the set of signatures of a probe image is assigned the identity ω* (class) contained in the gallery set that maximizes equation (27). In order to measure the power of generalization of the NTCA classifier, we used a 5-fold cross-validation estimation technique in all databases. We have concatenated the training set and the testing set to form one unique set that we partitioned into three subsets (training, gallery, and probe). Each time, we have respected the cardinality of each subset. For example, in the case of MORPH-album 1, for each individual, we ensured that only 1 image is used as training, 1 image as probe and 1 image as a gallery. This three subsets partitioning is repeated five times (5-folds). The actual performance of an algorithm is always rated relative to how well the images in probe are matched to the images in gallery. The relative frequency of correct matches for each individual is divided by the total number of individuals in the probe set to produce global face identification accuracy of the NTCA on a particular test set.

C.4. Identification Results

We have first run the NTCA model on the GeorgiaTech database. This experiment provides separate clues; it tells whether NTCA is capable to operate efficiently in presence of variations in facial expression, head pose, and illumination condition, but without aging effect. We have compared the results with the TDBN [12] (one of the state-of-the-art models) on the five different folds. We have used the TDBN model that we have recently developed as a benchmark to ensure reproducibility in terms of folds. Table 2 depicts the performance results obtained. NTCA has outperformed TDBN in all fold scenarios with an average performance increase of 2.5%. This gain in accuracy is explained by: (i) the amount of images available in this database, (ii) the effect of the extended LBP (through a variation of the radius and the neighborhood size) used to build the data manifold, and (iii) the power of the multinomial mixture during classification.

TABLE 2

ACCURACY (%) OF TDBN AND NTCA USING DIFFERENT TRAINING/GALLERY/PROBE FOLDS (RANK 1)

	Georgia Tech	
Rank 1	**TDBN**	**NTCA**
Fold 1	79.30	83.30
Fold 2	77.20	82.70
Fold 3	82.60	85.10
Fold 4	80.50	82.60
Fold 5	81.10	84.30
Mean	**80.14**	**83.60**
Standard Deviation (STD)	2.02	1.07

To acquire a keen understanding of the performance behavior, we have compared NTCA with the TDBN model on both the MORPH and the FGNET databases to test whether NTCA is more sensitive to the aging effect, facial expressions and illumination conditions. This experiment was conducted using the 5 folds described above. The results obtained are illustrated in Table 3. This table indicates that the highest average accuracy achieved by NTCA at rank 1 (all ages) is when tested on the MORPH-album 2 database. This peak is obtained using fold 3. It also shows that the NTCA has outperformed the TDBN's model in every fold scenario. Since our contribution throughout this paper is threefold: (i) Dimensionality reduction method that we have named "Kernelized Radial Basis Functions" (KRBF); (ii) the construction of alpha-shapes formed by a set of points in the latent space $\mathcal{X}$; the extraction of topological features; and (iii) the use of a mixture of multinomial distributions *(bag of words document model)* in the context of age-invariant face classification, therefore to compare our KRBF with other competitive techniques such as Park and al. (2010) [33], Li and al. (2011) [34], and Bouchaffra (2012) [12], we have brought these techniques inside our platform. In other words:

- In the case of Park and al., we have replaced our KRBF with the Principal Component Analysis (PCA) method used by Park and left our multinomial mixture classifier unchanged.
- In the case of Li and al., we have replaced our KRBF with the Multi-Feature Discriminant Analysis (MFDA) method and left our multinomial mixture classifier unchanged.
- In the case of Bouchaffra, we have replaced our KRBF with the Maximum Weighted Cut (MWC) method and left our multinomial mixture classifier unchanged.

We have reported in Table 4 the performance of this test experiments.

TABLE 3

IDENTIFICATION ACCURACY (%) OF TDBN AND NTCA USING 5-FOLDS ON TWO DIFFERENT AGING DATABASES (RANK 1)

	FGNET		MORPH DATABASE			
			MORPH-1		MORPH-2	
	TDBN	**NTCA**	**TDBN**	**NTCA**	**TDBN**	**NTCA**
Fold 1	36.50	49.60	62.20	72.50	54.40	81.80
Fold 2	33.60	48.20	58.90	70.40	52.70	84.50
Fold 3	37.30	50.50	68.60	73.20	63.40	86.80
Fold 4	40.50	48.90	61.40	69.50	55.70	83.60
Fold 5	38.90	47.60	60.10	70.90	56.70	82.30
Mean	**37.36**	**48.96**	**62.24**	**71.30**	**56.58**	**83.80**
STD	2.60	1.14	3.77	1.52	4.09	1.98

TABLE 4

COMPARISON OF KRBF DIMENSIONALITY REDUCTION TECHNIQUE WITH SOME STATE-OF-THE-ARTS APPROACHES.

	Approach	**Features extraction**	**Dimensionality reduction**	**Face matcher**	**Database (#individuals;# images) in probe and gallery**	**Rank 1 recognition accuracy reported (%)**
Park and al. (2010)[33]	Aging patterns are learned via PCA coefficients in separated 3D shape and texture given 2D database	AAM[1]	PCA[2]	Face VACS from Cognitec	FGNET (82;82)	37.40
					MORPH-album 1 (10000;20000)	66.40
					MORPH-album 2 (10000;20000)	79.80
Li and al. (2011)[34]	Discriminative analysis method in which a face is represented by densely sample local descriptors	MLBP[3]+SIFT[4]	MFDA[5]	Combination of Multiple LDA-based classifiers	FGNET (82;82)	47.50
					MORPH-album 2 (10000;20000)	83.90
Bouchaffra (2012)[12]	Aging patterns based on network embedding and α-shape constructors	DWT[6]	MWC[7]	Maximum a-posteriori probability using Gaussian mixture	FGNET (82;82)	38.20
Proposed method (NTCA)	Aging patterns based on nonlinear topological component analysis (dimensionality reduction and extraction of α-shape signatures)	LBP	KRBF	Maximum a-posteriori probability using a mixture of multinomial distribution	FGNET (82;82)	48.96
					MORPH-album 1 (612;612)	71.30
					MORPH-album 2 (10000;20000)	83.80
Proposed method where KRBF is replaced by PCA		LBP	PCA	Maximum a-posteriori probability using a mixture of multinomial distribution	FGNET (82;82)	35.15
					MORPH-album 1 (612;612)	63.33
					MORPH-album 2 (10000;20000)	74.55
Proposed method where KRBF is replaced by MFDA		LBP	MFDA	Maximum a-posteriori probability using a mixture of multinomial distribution	FGNET (82;82)	43.55
					MORPH-album 1 (612;612)	66.44
					MORPH-album 2 (10000;20000)	78.22
Proposed method where KRBF is replaced by MWC		LBP	MWC	Maximum a-posteriori probability using a mixture of multinomial distribution	FGNET (82;82)	36.52
					MORPH-album 1 (612;612)	62.10
					MORPH-album 2 (10000;20000)	75.15

1: Active Appearance Model; **2:** Principal Component Analysis; **3:** Multi-scale Local Binary Patterns; **4:** Scale Invariant Feature Transform
5: Multi-Feature Discriminant Analysis; **6:** Discrete Wavelet Transform; **7:** Maximum Weighted Cut.

The results indicate that the accuracy of the KRBF dimensionality reduction method is almost always higher compared to the accuracy of these state-of-the-art dimensionality reduction approaches. It is also worth noting that the mixture of multinomial distributions classifier has impacted the results but not as much as the feature extraction techniques. This observation confirms the impact of the fusion of topology and statistics in the entire NTCA face identification methodology. Likewise, we have depicted the CMC curves for both the NTCA model (whose rank 1 recognition accuracy is 83.80%) and the FaceVACS of Cognitec [42] (whose rank 1 recognition accuracy is 79.80%), when tested on MORPH-album 2 (refer to Fig. 13). We have also plotted a 3D graph of the Recognition Accuracy of NTCA and FaceVACS of Cognitex [43] as a function of the False Accept Rate (FAR) and the Number of top n-matches (refer to Fig. 14). The value n is called the *rank* of the match; it indicates how many gallery images have to be inspected in order to acquire a desired performance level. NTCA has significantly outperformed FaceVACS for $n \leq 17$ and $FAR \geq 20$ and meagerly outside this range when tested on MORPH-album 2.

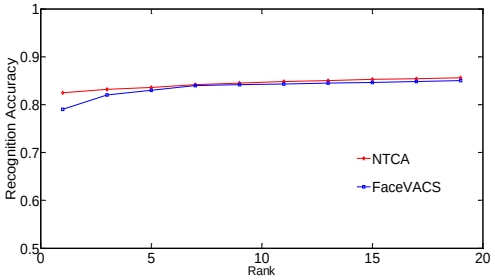

Fig. 13. CMC curves of NTCA and FACEVACS.

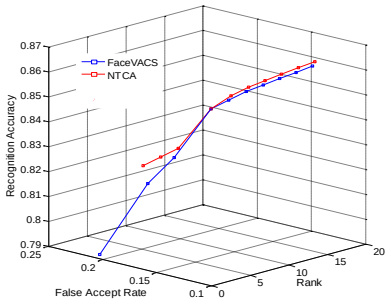

Fig. 14. 3D plot of face recognition rate as a function of the false accept rate and the rank of NTCA and FaceVacs.

Furthermore, in order to evaluate the contribution of the mapping φ (kernel step) from $\mathfrak{D}$ to $\mathfrak{D}_e$, we have run KRBF by dropping φ. In other words, we projected the vector $t = [t_1,...,t_D]$ directly to the Euclidean space $\mathbb{R}^3$. The results depicted in Table 5 show a significant recognition improvement in favor of using the kernel step. These results are consistent with the theoretical expectations.

TABLE 5

COMPARISON BETWEEN KRBF PERFORMANCE WITH AND WITHOUT THE KERNEL STEP ON DIFFERENT DATABASES

Databases	Accuracy (%) (KRBF with Kernel Step)	Accuracy (%) (KRBF without Kernel Step)
MORPH-album1 (612;612)	71.30	65.10
MORPH-album 2 (10000;20000)	83.80	79.20
FGNET (82;82)	48.96	41.30

Finally, to test the efficiency of KRBF, we have run NTCA by replacing KRBF with the traditional Isomap (with a neighborhood k=7) dimensionality reduction technique. The comparison results when tested on the FGNET database have shown a worsening in recognition performance using Isomap: (48.96% for NTCA with KRBF vs. 45.20% for NTCA with Isomap).

VIII. CONCLUSION AND FUTURE WORK

We have developed a nonlinear age-invariant face recognizer that we named *Nonlinear Topological Component Analysis*. This mission has been achieved by: (i) nonlinearly mapping the data topological manifold to a low dimensional latent variable space; (ii) extracting topological features through the computation of the α-shape polyhedron formed by these latent variables; and (iii) performing shape classification using a mixture of multinomial distributions. The results obtained have demonstrated the potential of the NTCA approach as a whole. *The benefit of expressing p(x) is invaluable since it allows the estimation of some parameters and the generation of other points in case where face images of the same individual are lacking. It also renders the more general weighted α-shapes concept exploitable.* However, our methodology can still be improved through the use of a more stable clustering algorithm of the z_i vectors. Furthermore, a classification with a reject option should be devised to meet the demands of a large scale application.

References

[1] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863-874, Mar. 2007.

[2] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, Dec. 2000.

[3] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, Dec. 2000.

[4] D. Donoho and C. Grimes, "Hessian Eigenmaps: new tools for nonlinear dimensionality reduction," in *Proceedings of National Academy of Science*, 2003, pp. 5591-5596.

[5] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," in *Advances in Neural Information Processing Systems*, vol. 14, D. G. Dietterich, S. Becker, Z. Ghahramani, Eds. The MIT Press, pp. 585-591, 2002.

[6] Z. Zhang and Z. Hongyuan, "Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2005.

[7] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, Jan. 2007.

[8] G. Di Battista, P., Eades. R. Tamassia and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[9] D. Harel, and Y. Koren, "Graph Drawing by High-Dimensional Embedding," in *Proceedings of 10th International Symposium of Graph Drawing*, London, UK, Feb. 2002, pp. 207-219.

[10] C. C. Lin and H. C. Yen, "A New Force-Directed Graph Drawing Method based on Edge-Edge Repulsion Method," in *Proceedings of the Ninth IEEE ICCV*, Beijing, China, Oct. 2005, pp. 329-334.

[11] M. Chen, Q. Yang and X. Tang, "Directed Graph Embedding," in *Proceedings of the 20th IJCAI*, Hyderabad, India, Jan. 2007, pp. 2707-2712.

[12] D. Bouchaffra, "Mapping Dynamic Bayesian Networks to α-Shapes: Application to Human Faces Identification Across Ages," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1229-1241, Aug. 2012.

[13] C. M. Bishop, M. Svensén, and C. K. I. Williams, "The Generative Topographic Mapping," *Neural Computation*, vol. 10, no 1, pp. 215-234, January 1998.

[14] M. Gemignani, *Elementary Topology*. Second edition, New York: Dover Publications, Inc., 1990.

[15] M. Aizerman, E. Braverman and L. Rozonoer, "Theoretical foundations of the potential function method in Pattern Recognition Learning," *Automation and Remote Control*, vol. 25, no. 1, pp. 821-837, 1964.

[16] Y. Y. Lin, T. L. Liu and C. S. Fuh, "Multiple Kernel Learning for Dimensionality Reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147-1160, June 2011.

[17] M. Gonen and E. Alpaydin, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12 (July), pp. 2211-2268, 2011.

[18] P. P. Kuksa, "2D similarity kernels and representations for sequence data," in *Proceedings of The Learning Workshop*, Snowbird, Utah, April 3-6, 2012.

[19] Rupp, "Graph Kernels," in *Statistical and Machine Learning Approaches for Network Analysis*, M. Dehmer and S. C. Basak Eds. John Wiley & Sons, Inc., Hoboken, New-Jersey, USA, 2012.

[20] H. Edelsbrunner and E. P. Mücke, "Three-Dimensional Alpha Shapes," *ACM Transactions on Graphics*, vol. 13, no. 1, pp. 43-72, 1994.

[21] T. K. Dey, J. Giesen and M. John, "Alpha-Shapes and Flow Shapes Are Homotopy Equivalent," in *Proceedings of the 35th ACM Symp. on the Theory of Computing*, San Diego, CA, June 2003, pp. 493-502.

[22] M. Ankherst, M. Breunig, H. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure", in *Proceedings of the ACM SIGMOD'99 Int. Conf. on Manag. of Data*, Philly, PA, 1999, pp. 49-60.

[23] A. V. Nefian, "Embedded Bayesian networks for face recognition," in *Proceedings of the IEEE Int. Conf. Multimedia Expo*, Lausanne, Switzerland Nov. 2002, pp. 133–136.

[24] D. Bouchaffra and A. Amira, "Structural Hidden Markov Models for Biometrics: Fusion of Face and Fingerprint," *Special Issue of Pattern Recognition Journal (Elsevier)*, vol. 41, no. 3, pp. 852-867, 2008.

[25] F. Samaria, *Face Recognition using Hidden Markov Models.* Ph.D. Thesis, Eng. Dept., Cambridge Univ., Cambridge, U.K., 1994.

[26] D. Bouchaffra, "Conformation-based Hidden Markov Models: Application to Human Face Identification," *IEEE Transactions on Neural Networks*, vol. 21, no. 4, pp. 597-608, April 2010.

[27] W. Zhao, R. Chellappa, *Face Processing: Advanced Modeling and Methods.* Academic Press Inc., USA, 2006.

[28] S.Z. Li, A.K. Jain, *Handbook of Face Recognition.* Springer, New York, March 2005.

[29] F. Juefei-Xu, L. Khoa, M. Savvides, T.D. Bui, and C.Y. Suen, "Investigating Age-Invariant Face Recognition based on Periocular Biometrics," in *Proceedings of the IEEE 2011 International Joint Conference on Biometrics*, Washington DC, 11-13 Oct. 2011, pp. 1-7.

[30] N. Ramanathan, R. Chellapa and S. Biswas, "Computational Methods for Modeling Facial Aging: A Survey," *Journal of Vis. Lang. and Comput.*, vol. 20, no. 3, pp. 131-144, 2009.

[31] Z. Yang and H. Ai, "Demographic Classification with Local Binary Patterns," in *Proceedings of the 3rd IEEE Int. Conf. on Biometrics*, San Francisco, CA, 2007, pp. 464-473.

[32] A. X. Geng, Z. H. Zhou and K. Smith-Miles, "Automatic Age Estimation Based on Facial Aging Patterns," *IEEE TPAMI*, vol. 29, no. 12, pp. 2234-2240, Dec. 2007.

[33] U. Park, Y. Tong, and A. K. Jain, "Age-Invariant Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 947-954, May 2010.

[34] Z. Li, U. Park, and A. K. Jain, "A Discriminative Model for Age Invariant Face Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1028-1037, Sept. 2011.

[35] D. Bouchaffra, "Topological Dynamic Bayesian Networks," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, Aug. 2010, pp. 898-901.

[36] D. Bouchaffra, "Topological Dynamic Bayesian Networks: Application to Facial Aging," in *Proceedings of the IEEE Computer Society on CVPR Workshops*, San Francisco, June 13-18, 2010, pp. 1-8.

[37] A. V. Malkapurkar, R. Patil, S. Murarka, "A New Technique for LBP Method to Improve Face Recognition," in *Int. Journal of Emerging Technology and Advanced Engineering*, vol. 1, no. 1, Nov. 2011.

[38] Alpha-shapes software [Online]. Available at: http://biogeometry.duke.edu/software/alphashapes/index.htm

[39] K. Ricanek Jr and T. Tesafaye, "MORPH: A Longitudinal Image Database of Normal Adult Age-Progression," in *Proceedings of IEEE 7th International Conference on Automatic Face and Gesture Recognition*, UK, Apr. 2006, pp. 341-345.

[40] "FG-Net aging database" [Online]. Available at: http://www.sting.cycollege.ac.cy/~alanitis/fgnetaging/

[41] A. W. Rawls and K. Ricanek Jr., "MORPH: Development and Optimization of a Longitudinal Age Progression Database", in *Proceedings of BioID_MultiComm*, J. Fierrez, J.Ortega-Garcia, A. Esposito Eds. Madrid, Spain, September 16-18, 2009.

[42] J. R. Beveridge, K. She, B. A. Draper and G. H. Givens, "Parametric and Nonparametric Methods for the Statistical Evaluation of Human ID Algorithms," in *Proceedings of the Third Workshop on Empirical Evaluation Methods in Computer Vision*, Kauai, HI, Dec. 2001, pp. 1-14.

[43] FaceVACS Software Developer Kit, Cognitec Systems Gbmh, http://www.cognitec-systems.de