Robust 3D face recognition using adapted statistical models

Janez Križaj, Simon Dobrišek, Vitomir Štruc and Nikola Pavešić

Faculty of Electrical Engineering, University of Ljubljana
Tržaška 25, SI-1000 Ljubljana, Slovenia

E-mail: {janez.krizaj, simon.dobrisek, vitomir.struc, nikola.pavesic}@fe.uni-lj.si

Abstract

The paper presents a novel framework to 3D face recognition that exploits region covariance matrices (RCMs), Gaussian mixture models (GMMs) and support vector machine (SVM) classifiers. The proposed framework first combines several 3D face representations at the feature level using RCM descriptors and then derives low-dimensional feature vectors from the computed descriptors with the unscented transform. By doing so, it enables computations in Euclidean space, and makes Gaussian mixture modeling feasible. Finally, a support vector classifier is used for identity inference. As demonstrated by our experimental results on the FRGCv2 and UMB databases, the proposed framework is highly robust and exhibits desirable characteristics such as an inherent mechanism for data fusion (through the RCMs), the ability to examine local as well as global structures of the face with the same descriptor, the ability to integrate domain-specific prior knowledge into the modeling procedure and consequently to handle missing or unreliable data.

1 Introduction

Personal recognition based on 3D facial images is becoming increasingly popular mainly due to its potential market value and its desirable characteristics, such as inherent robustness to illumination or pose changes. Nevertheless, there are still a number of open issues, as emphasized by various surveys, e.g., [1], pertaining mainly to recognition in the presence of varying facial expressions, partial occlusions of the facial area and the overall reliability of the recognition procedure.

In this paper we build upon the framework originally introduced in [2] and present an updated framework for 3D face recognition, which again capitalizes on region covariances and Gaussian mixture models (GMMs), but adds several new steps to the processing chain. Within the proposed framework a 3D face image is first represented by a number of region covariance matrices computed from regions of different sizes. These matrices ensure that the 3D face images are described in a highly discriminative manner and form the foundation for our modeling procedure. As the matrices reside in Riemannian and not Euclidean space they cannot be subjected directly to the GMM construction step. To overcome this issue, we employ the unscented transform [3] and produce a number of feature vectors from each of the region covariance matrices and finally use these vectors as input to our modeling stage. Once the GMM model is computed from the input image a SVM classification scheme is used to make a decision regarding the identity of the 3D face image originally presented to the recognition framework. The proposed framework has a number of desirable characteristics, such as an inherent mechanism for data fusion (through the region covariance matrices), the ability to examine the facial images at different levels of locality, to handle missing data, etc.

The rest of the paper is structured as follows: in Section 2 we introduce the proposed framework and describe all of its parts and characteristics. In Section 3 we present some experiment and highlight the merits of the proposed methodology. We conclude the paper in Section 4.

2 Proposed methodology

2.1 Overview

Fig. 1 depicts a simplified block diagram of the proposed 3D face recognition framework. The first procedural step of the framework is the acquisition of the 3D face image. The data acquisition step is followed by a registration and preprocessing procedure, where the facial region is cropped from the 3D scan and any holes and spikes potentially present are removed. In the next step, the preprocessed 3D facial data is mapped into a data structure that we will refer to as a *composite representation* in the remainder of the paper. The composite representation is then processed on a block-by-block basis and region covariance matrices (RCM) are extracted from each examined block. Note that different from most other feature extraction techniques, RCM descriptors can be extracted from regions of variable sizes, thus, allowing to extract discriminative information from the data at local as well as global levels. Furthermore, the descriptors provide an elegant way of combining different representations of 3D data into a coherent feature vector. After the RCM extraction procedure each face is represented by a number of RCM descriptors, whose distribution can be modeled

This work has been supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the postdoctoral project BAMBI (ARRS ID Z2-4214), and European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT). The authors additionally appreciate the support of COST Actions IC 1106 and IC1206.

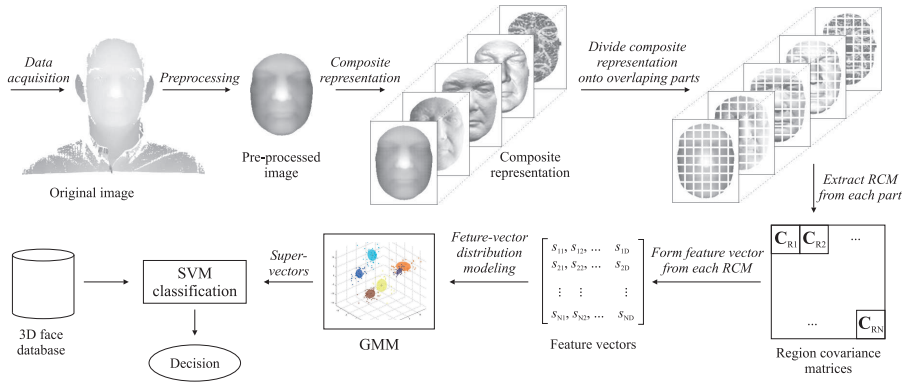

Figure 1: Conceptual diagram of the proposed framework

by a GMM. Here, GMMs are selected for modeling purposes, since they allow to incorporate prior knowledge into the modeling procedure and are easily adopted to handle unreliable or missing (i.e., occluded) data. Finally, a SVM-based classification scheme is employed to classify the super-vectors derived from the GMMs. In the remainder we elaborate on all presented steps.

2.2 Data preprocessing

Raw 3D face images, which represent the input to our framework, are initially low-pass filtered to remove any spikes potentially present. The z values (depth components) are then interpolated and uniformly re-sampled. In the last step, the depth data is smoothed with a mean filter. Automatic localization of the face is performed using a simple clustering procedure, where all $[x, y, z]$ vectors are clustered into three distinct clusters and the largest is retained as the facial area. Note that this procedure is far from perfect; however, it is highly robust and has proven "good" enough for the proposed framework.

2.3 Data representation

Let $\boldsymbol{I}$ represent a preprocessed and localized face depth image of size $w \times h$. A $w \times h \times d$ dimensional composite representation $\boldsymbol{F}$ is then constructed from the depth image $\boldsymbol{I}$ (see Fig. 1) based on the following expression:

$$\boldsymbol{F}(x, y) = \phi(\boldsymbol{I}, x, y), \tag{1}$$

where the function ϕ derives a d-dimensional vector $\boldsymbol{f} = \boldsymbol{F}(x, y)$ from a pixel at position (x, y) of $\boldsymbol{I}$. The vector $\boldsymbol{f}$ can be derived by concatenating different representations of the image $\boldsymbol{I}$ at (x, y). These representations include depth values, color information, pixel coordinates, values of image gradients, higher order derivatives, filter responses, differential-geometry descriptors, surface normals, etc. In other words, the composite representation $\boldsymbol{F}$ represents a $w \times h \times d$ tensor, with w and h representing its spatial coordinates and d denoting the number of representations combined in the tensor (see Fig. 1).

2.4 Region Covariance Matrix

Once the composite representation $\boldsymbol{F}$ is constructed from the given 3D face image, RCM descriptors can be computed from it and used to derive feature vectors for the modeling technique. Formally, any rectangular region $\boldsymbol{R} \subset \boldsymbol{F}$, comprising a set of vectors $\{\boldsymbol{f}\}_{k=1\ldots n}$, can be represented by a $d \times d$ covariance matrix [4]: $\boldsymbol{C}_R = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{f}_k - \boldsymbol{\mu}_r)(\boldsymbol{f}_k - \boldsymbol{\mu}_r)^T$, where $\boldsymbol{\mu}_r$ is the mean vector of $\{\boldsymbol{f}\}_{k=1\ldots n}$. The diagonal entries of $\boldsymbol{C}_R$ represent the variance of each feature and the non-diagonal entries represent their respective correlations. Extracting the covariance of an inhomogeneous area results in a strictly symmetric and positive semi-definite matrix with constant dimension that models the properties of the specified region. If no location-related representations, such as spatial coordinates and alike, are used for the construction of the composite representation then the RCM descriptor is both rotation as well as scale invariant.

2.5 The Unscented Transform

Covariance matrices do not lie on Euclidean space (e.g. the space is not closed under multiplication with negative scalars). Since we plan to use the computed RCM descriptors as input for our GMM modeling procedure, we exploit the Unscented Transform (UT) [3] to approximate the RCM descriptors in Euclidian space. The transform is capable of generating a specific set of vectors $\boldsymbol{w}_i$ (for $i = 0, 1, ..., 2d + 1$) from each region covariance matrix $\boldsymbol{C}_R$ with the distribution of the set of vectors approximating the distribution characterized by $\boldsymbol{C}_R$. However, unlike $\boldsymbol{C}_R$, the vectors $\boldsymbol{w}_i$ reside in Euclidean space. The UT is similar to Monte Carlo methods with the difference that the vectors are not generated randomly.

Given the region covariance matrix $\boldsymbol{C}_R$ and assuming an underlying Gaussian distribution $p(\boldsymbol{\mu}, \boldsymbol{C}_R)$, the unscented transform generates a set of $2d + 1$ vectors $\boldsymbol{w}_i$ as follows: $\boldsymbol{w}_0 = \boldsymbol{\mu}$,, $\boldsymbol{w}_i = \boldsymbol{\mu} + (\sqrt{\alpha \boldsymbol{C}_R})_i$, $\boldsymbol{w}_{i+d} = \boldsymbol{\mu} - (\sqrt{\alpha \boldsymbol{C}_R})_i$, where $i = 1 \ldots d$ and $(\sqrt{\alpha \boldsymbol{C}_R})_i$ defines the i-th column of the square root of $\boldsymbol{C}_R$. The scalar α is a weight for the elements in the covariance matrix and is set to $\alpha = 2$ in the case of the Gaussian distribution.

Each of the $(2d+1)$ vectors $\boldsymbol{w}_i$ resides in a d-dimensional Euclidean space, where L^2 distance computations can be applied. To obtain a single feature vector from each RCM, we concatenate all feature vectors extracted from a given RCM into one $1 \times d(2d + 1)$-dimensional feature vector $\boldsymbol{s}' = [\boldsymbol{w}_0^T \boldsymbol{w}_1^T ... \boldsymbol{w}_{2d+1}^T]^T$ that is first projected into a PCA (principal component analysis) subspace and finally used as input to the modeling procedure: $f_{PCA} : \boldsymbol{s}' \in \mathbb{R}^{d(2d+1)} \mapsto \boldsymbol{s} \in \mathbb{R}^{d'}$; $d' \ll d(2d + 1)$.

2.6 Modeling and Classification

In the last procedural step of our system, we model the distribution of the local feature vectors, extracted from the 3D face images by means of RCM descriptors, the unscented transform and PCA using GMMs. Formally, a GMM $\lambda = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ is defined as a linear combination of K multivariate Gaussian probability density functions (PDFs)

$$p(\boldsymbol{s}|\lambda) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{s}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\{\pi_k\}_{k=1}^{K}$ denote the weights of the mixture model, $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ denote the mean vectors and $\{\boldsymbol{\Sigma}_k\}_{k=1}^{K}$ represent diagonal covariance matrices of the GMM.

Given a set of local feature vectors $\boldsymbol{\Psi} = \{\boldsymbol{s}_n\}_{n=1}^{N}$, a GMM is constructed by determining its parameters based on maximization of the log-likelihood: $\log p(\boldsymbol{\Psi}|\lambda) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\boldsymbol{s}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Maximum likelihood (ML) solutions for the model parameters are found via the Expectation-Maximization (EM) algorithm in our case initialized with the K-means clustering algorithm.

When building subject-specific GMMs[1], there is usually not enough data available to estimate the parameters of the GMM reliably. Therefore, a universal background model (UBM) is typically constructed first and then adapted with subject-specific data. A UBM is itself a GMM representing generic, person independent feature characteristics. The parameters of the UBM are estimated via the ML paradigm on all available training data. Once the UBM is build, subject-specific GMM are computed via maximum a posteriori (MAP) adaptation [5], where only the mean vectors $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ are adapted, by iterative evaluation of

$$\hat{\boldsymbol{\mu}}_k = (1-\alpha)\boldsymbol{\mu}_k + \alpha \boldsymbol{\mu}_k^{EM}. \tag{3}$$

The mean vectors from the constructed GMM are stacked one after the other to form the so-called *super-vector* of the given 3D face image.

Once the super-vector is derived from the input 3D face image of a given subject, it can be used to train a classifier for this specific subject. During the enrollment phase, when the subject is first presented to the system, a SVM [6] classifier is trained for that subject and a decision hyperplane between the super-vector of the "enrollee" and the super-vectors from a set of development images is constructed. In the test phase, the claimant is accepted or rejected based on the distance of the claimant's super-vector to the decision hyperplane. Note that within-class covariance normalization (WCCN) as well as rank normalization of the data are used prior to SVM training and classification.

2.7 Characteristics of the Proposed Approach

The framework introduced in the previous sections exhibits some highly useful characteristics due to the use of RCM descriptors and GMMs: *i) data fusion:* RCM descriptors are capable of combining various 3D representations into a single coherent descriptor and can, thus, be treated as an efficient data fusion scheme; *ii) invariance:* RCM descriptors do not encode information relating to the ordering or number of feature vectors in the region from which they were computed and are, hence, scale and rotation invariant to some extent; *iii) computation:* since RCM descriptors are computable regardless of the number of feature vectors used for their computation, they can easily handle missing data (i.e., holes in the face scans or regions on the borders of the face scans) in the feature extraction step; *iv) local vs. global representation:* the size of the RCM-derived feature vectors does not depend on the size of the region from which they were extracted; feature vectors of equal dimensions can, therefore, be computed from variable sized image blocks; *v) robustness:* GMM-based systems treat data (i.e., feature vectors) as independent and identically distributed (i.i.d.) observations and, hence, present 3D facial images in the form of a number of orderless blocks; this characteristic is reflected in good robustness to imperfect face alignment, pose changes, occlusions and expression variations; and *vi) missing data:* GMM makes it easy to handle missing/unreliable data during the modeling step and allow for the inclusion of domain-specific prior knowledge into the modeling procedure.

[1] A subject-specific GMM in this context is a GMM constructed from one 3D face image of a specific subject.

3 Experiments and results

For the experiments presented in remainder two databases were adopted - the FRGCv2 and UMB databases [7], [8].

The images in the FRGCv2 database exhibit minor pose variations and major expression variations. FRGCv2 includes 4007 3D face images of 466 subjects. The database is used in the experimental configuration usually referred to as the *all vs. all* configuration. Here, more than 16 million verification attempts are conducted and consequently used to generate performance metrics. The second database used in the experiments, the UMB database, is composed of 1473 images (3D + color 2D) of 143 subjects. This database was acquired with a particular focus on facial occlusions that can occur in real-world scenarios. There are 590 images with facial areas occluded by different objects, such as hair, eyeglasses, hands, hats, scarves, etc. Occlusions cover, on average, 42% of the face area, with a maximum of about 84%. For the UMB database, all face scans marked as occluded were matched against all face scans marked as neutral and non-occluded. Hence, this database was used to assess the robustness of the proposed framework. Note that ZT-score normalization was used as a standard procedure for all techniques when generating the results.

The first crucial step of the proposed framework is the construction of the composite representation $\boldsymbol{F}$ to be used in all further processing steps. To select appropriate representations for this purpose, the first experiments evaluate different possibilities, such as pixel coordinates (x, y), depth values $\boldsymbol{I}$, shape index values $\boldsymbol{I}_s$, Gaussian curvature values $\boldsymbol{I}_g$, mean curvature values $\boldsymbol{I}_m$, minimum curvature values $\boldsymbol{I}_{min}$, maximum curvature values

Table 1: Verification rate (%) at a 0.1% FAR for different $\boldsymbol{F}$

Composite representation $\boldsymbol{F}$	FRGCv2	UMB
$[\boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$	94.8	81.2
$[\boldsymbol{X}\ \boldsymbol{Y}\ \boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$	95.8	83.5
$[\boldsymbol{I}_{s}\ \boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$	**95.7**	**84.7**
$[\boldsymbol{X}\ \boldsymbol{Y}\ \boldsymbol{I}_{s}\ \boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$	94.7	83.9
$[\boldsymbol{I}_{lbp}\ \boldsymbol{I}_{s}\ \boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$	93.6	82.0
$[\boldsymbol{I}_{s}\ \boldsymbol{I}_{g}\ \boldsymbol{I}_{m}\ \boldsymbol{I}_{min}\ \boldsymbol{I}_{max}]$	92.3	82.3
$[\boldsymbol{X}\ \boldsymbol{Y}\ \boldsymbol{I}_{s}\ \boldsymbol{I}_{\varphi}\ \boldsymbol{I}_{lbp}]$	78.3	65.6

Table 2: Comparative results

FRGCv2 (VR@01FAR in %)		UMB (ROR in %)	
Drira *et al.* [9]	94.0	Colombo *et al.* [8]	56.6
Inan *et al.* [10]	98.4	Alyuz *et al.* [11]	74.6
Queirolo *et al.* [12]	96.5	Proposed	91.8
Proposed	97.7		

$\boldsymbol{I}_{max}$, surface normal coordinates $\boldsymbol{I}_{nx}$, $\boldsymbol{I}_{ny}$ and $\boldsymbol{I}_{nz}$, local binary patterns $\boldsymbol{I}_{lbp}$ and angle values $\boldsymbol{I}_{\varphi}$ between surface normals and the average facial normal. As can be seen from Table 1, where the results of the experiments are presented in the form of verification rates at the false acceptance rate of 0.1% (VR@01FAR), the best performance is achieved with the following composite representation: $\boldsymbol{F} = [\boldsymbol{I}_{s}\ \boldsymbol{I}_{nx}\ \boldsymbol{I}_{ny}\ \boldsymbol{I}_{nz}]$. This result is particularly interesting as the best performance is not a result of using the composite representation with the largest number of individual representation combined. This indicates that special attention should be given to the selection of appropriate representations.

After optimizing other hyper-parameters of the recognition framework, such as, feature space dimensionality, number of gaussian mixtures, SVM kernel parameters, and alike, it is possible to further improve on the verification rates presented in Table 1. With optimized parameters the proposed framework becomes highly competitive even when compared to state-of-the-art results from the literature as shown in Table 2. Note that the framework achieves comparable performance to state-of-the-art techniques from the literature on the FRGCv2 database, while it significantly outperforms the existing state-of-the-art on the UMB database that contains occluded facial scans (the results for the UMB are presented in term of the rank one recognition rate - ROR).

The robustness of the framework can, of course, be attributed to the use of RCM descriptors and face representations based on differential geometry, which ensure robustness to pose and expression variations. However, the most important part when it comes to robustness to facial occlusions is the modeling procedure. Here, the UBM and the MAP adaptation procedure make it possible to cope with unreliable local features and still construct reliable subject-specific GMMs. To better demonstrate this fact the GMMs are used as generative models. By sampling from the constructed GMMs, it is possible to generate synthetic data in the feature space and subsequently generate face images. Some of these images are shown in Fig. 2. Here, the upper row shows automatically localized faces from the UMB database, while the lower row shows synthetic images generated from the corresponding GMMs. Note how due to the UBM the subject-specific GMMs encode frontal (rotation-corrected), segmented and un-occluded information about the given subjects and carry similar information for the same identities.

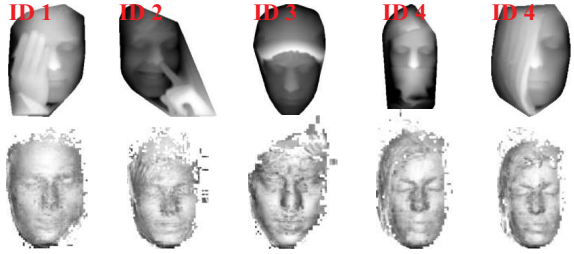

Figure 2: Pre-processed images (top row) and images generated from corresponding GMMs by random sampling(bottom row).

4 Conclusion

The paper introduced a novel framework for 3D face recognition, assessed it on two challenging databases and demonstrated its merits over other techniques from the literature.

References

[1] K. Bowyer, K. I. Chang, and P. J. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *CVIU*, vol. 101, pp. 1–15, 2006.

[2] J. Krizaj, S. Dobrišek, and V. Štruc, "Combining 3D face representations using region covariance descriptors and statistical models," in *IEEE FG Workshops*, 2013.

[3] S. Kluckner, T. Mauthner, and H. Bischof, "A Covariance Approximation on Euclidean Space for Visual Tracking," in *AAPR / ÖAGM*, 2009.

[4] O. Tuzel, F. Porikli, and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," in *ECCV*, 2006, pp. 589–600.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using Adapted Gaussian mixture models," in *Dig. Sig. Proc.*, 2000, pp. 19–41.

[6] H. Bredin, N. Dehak, and G. Chollet, "GMM-based SVM for face recognition," in *ICPR*, 2006, pp. 1111–1114.

[7] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," in *IEEE CVPR*, 2005, pp. 947–954.

[8] A. Colombo, C. Cusano, and R. Schettini, "UMB-DB: A database of partially occluded 3D faces," in *ICCV Workshops*, nov. 2011, pp. 2113–2119.

[9] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3D Face Recognition Under Expressions, Occlusions and Pose Variations," *IEEE TPAMI*, no. 99, 2013.

[10] T. Inan and U. Halici, "3-D Face Recognition With Local Shape Descriptors," *IEEE TIFS*, vol. 7, no. 2, pp. 577–587, april 2012.

[11] N. Alyuz, B. Gokberk, and L. Akarun, "3-D Face Recognition Under Occlusion Using Masked Projection," *IEEE TIFS*, vol. 8, no. 5, pp. 789–802, 2013.

[12] C. Queirolo, L. Silva, O. Bellon, and M. Segundo, "3D Face Recognition Using Simulated Annealing and the Surface Interpenetration Measure," *IEEE TPAMI*, vol. 32, no. 2, pp. 62–73, 2010.