Combining 3D Face Representations using Region Covariance Descriptors and Statistical Models

Janez Križaj, Vitomir Štruc and Simon Dobrišek
Faculty of Electrical Engineering, University of Ljubljana,
Tržaška 25, 1000 Ljubljana, Slovenia
Email: {janez.krizaj, vitomir.struc, simon.dobrisek}@fe.uni-lj.si

***Abstract*—The paper introduces a novel framework for 3D face recognition that capitalizes on region covariance descriptors and Gaussian mixture models. The framework presents an elegant and coherent way of combining multiple facial representations, while simultaneously examining all computed representations at various levels of locality. The framework first computes a number of region covariance matrices/descriptors from different sized regions of several image representations and then adopts the unscented transform to derive low-dimensional feature vectors from the computed descriptors. By doing so, it enables computations in the Euclidean space, and makes Gaussian mixture modeling feasible. In the last step a support vector machine classification scheme is used to make a decision regarding the identity of the modeled input 3D face image. The proposed framework exhibits several desirable characteristics, such as an inherent mechanism for data fusion/integration (through the region covariance matrices), the ability to examine the facial images at different levels of locality, and the ability to integrate domain-specific prior knowledge into the modeling procedure. We assess the feasibility of the proposed framework on the Face Recognition Grand Challenge version 2 (FRGCv2) database with highly encouraging results.**

I. Introduction

Biometric recognition based on 3D facial imagery is becoming increasingly popular and is attracting more and more research groups from around the world each year. The interest in the technology can mainly be attributed to its potential market value and desirable characteristics of 3D data, such as inherent robustness to illumination or pose changes. Nevertheless, next to problems related to the data-acquisition devices there are still a number of open issues of the recognition technology, as emphasized by various surveys, e.g., [1]. These issues pertain mainly to recognition in the presence of varying facial expressions and the overall reliability of the recognition procedure.

In this paper we introduce a novel framework for 3D face recognition, which capitalizes on region covariance matrices and Gaussian mixture models (GMMs). 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 provide for a highly discriminative description of the 3D face images and form the foundation for our modeling procedure. Unfortunately, the matrices do not reside in an Euclidean space and cannot be subjected directly to the GMM construction step. We, therefore, adopt the unscented transform [2] to produce a number of feature vectors from each of the region covariance matrices and use these vectors as input to our modeling stage. Once the model is constructed a SVM classification scheme is used to make a decision regarding the identity of 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/integration (through the region covariance matrices), the ability to examine the facial images at different levels of locality, etc.

The rest of the paper is structured as follows: in Section II we elaborate on the proposed framework, its procedural parts and characteristics and present a detailed evaluation of the proposed approach in Section III. We conclude the paper in Section IV with some final comments and directions for future work.

The work presented in this paper was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the postdoctoral project BAMBI (ARRS ID Z2-4214), the joint Bulgarian research project entitled "Fast and reliable 3D face recognition" with ARRS ID Bi-Bg/11-12-007 and European Union's Seventh Framework Programme (FP7-SEC-2010-1) under grant agreement number 261727 (SMART).

II. Our Methodology

This section describes the basic characteristics of the 3D face recognition framework proposed in this paper. It commences by presenting a brief description of the entire framework, and then proceeds by explaining in detail the preprocessing, data representation, feature extraction, modeling and classification stages of the proposed approach. The section concludes by elaborating on the characteristics and merits of the proposed methodology.

A. Overview

In Fig. 1, a block diagram of the proposed 3D face recognition framework is presented. 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 and any potential holes and spikes in the data 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. This composite representation is nothing more than different representations of the 3D facial data stacked one after another (see Fig. 1 for details). The composite representation is then analyzed block-by-block and a region covariance matrix (RCM) is extracted from each examined block. Note that differently from most other feature extraction techniques, RCM descriptors can be extracted from regions of variable sizes, thus, allowing to inspect the data from local as

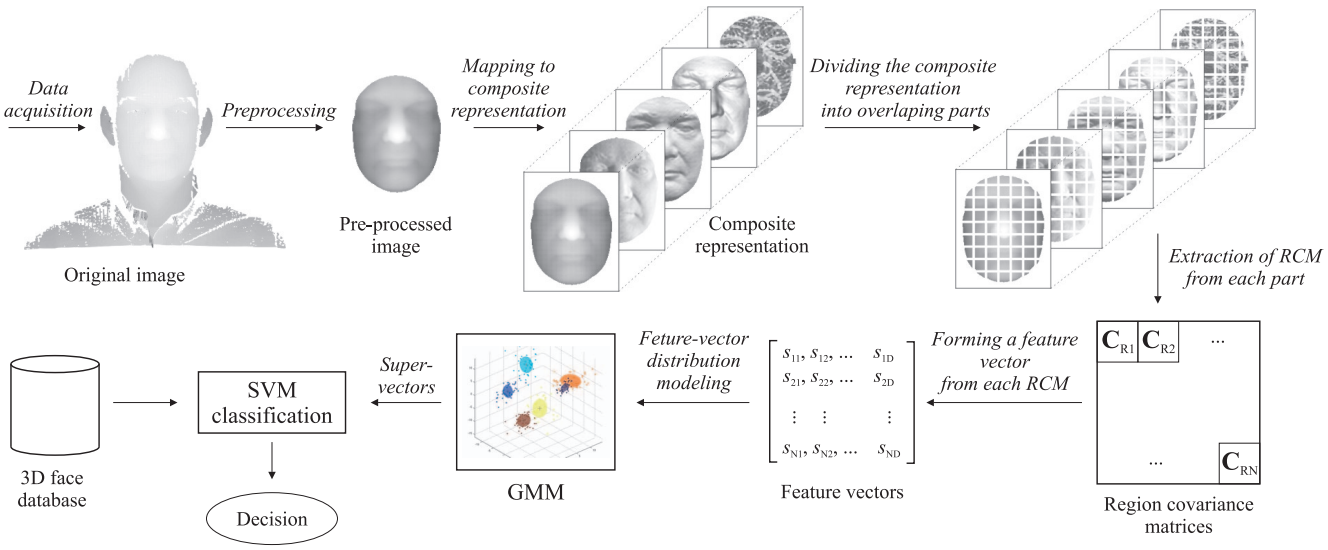


Fig. 1. Conceptual diagram of the proposed system

well as holistic points of views. 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 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 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 and discuss their importance.

B. 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 on a grid of 0.5 mm resolution on the (x, y) plane. Finally, the depth data is smoothed with a mean filter. Automatic localization of the face is performed based on the detection of the nose-tip and the points outside a sphere with the nose-tip at its origin and the radius of 100 mm are discarded. The entire procedure is very similar to the one presented in [3]. It is important to stress at this point that the employed localization procedure is not as precise as commonly employed Iterative Closest Point procedures. However, it is computationally extremely simple and due to the robust nature of the proposed system, more than enough to ensure satisfactory recognition results.

C. Data representation

Let $\boldsymbol{I}$ represent a preprocessed and localized face depth image of size $w \times h$. We then construct a $w \times h \times d$ dimensional composite representation $\boldsymbol{F}$ from the given 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 extracts 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 constructed 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. To summarize, 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. A conceptual representation of the composite representation is shown in Fig. 1.

Note that there is no rule on how many or which 3D data representations to combine into $\boldsymbol{F}$ for the optimal face recognition performance. This issue has to be resolved experimentally, and for our system will be addressed in the experimental section.

D. Region Covariance Matrix

Once we compute our composite representation $\boldsymbol{F}$ from the given 3D face image, we can compute RCM descriptors from it and use them to derive feature vectors for our 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]

$$\boldsymbol{C}_R = \frac{1}{n-1} \sum_{k=1}^{n} (\boldsymbol{f}_k - \boldsymbol{\mu}_r)(\boldsymbol{f}_k - \boldsymbol{\mu}_r)^T, \tag{2}$$

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 [4], [5] - $\boldsymbol{C}_R$ does not capture the ordering of the incorporated vector $\boldsymbol{f}$ in the image grid nor

does it capture the information related to the size of the region from which it was extracted.

E. The Unscented Transform

Covariance matrices do not lie on Euclidean space (e.g. the space is not closed under multiplication with negative scalars). Since the majority of standard machine learning techniques are defined on Euclidean spaces, they are not directly applicable to work with covariance matrices. Non-linear mappings to Riemannian manifolds [6] or the Lie algebra [7] are, therefore, traditionally used to obtain vector spaces, in which the metrics for machine learning methods are defined. This concept is also utilized in the Förstner metric [8], which approximates covariance dissimilarity measurement through log-manifold mapping and was originally proposed to measure the similarity of two RCM descriptors [6]. Since we plan to use the computed RCM descriptors as input for our GMM modeling procedure, we cannot adopt the Förstner metric for our computations. Instead, we consider a different approach in this paper and exploit the Unscented Transform (UT) [2], [8].

The UT approximates a distribution by specified sampling. 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 concept of 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:

$$\begin{aligned} \boldsymbol{w}_0 &= \boldsymbol{\mu}, \\ \boldsymbol{w}_i &= \boldsymbol{\mu} + (\sqrt{\alpha \boldsymbol{C}_R})_i, \\ \boldsymbol{w}_{i+d} &= \boldsymbol{\mu} - (\sqrt{\alpha \boldsymbol{C}_R})_i, \end{aligned} \tag{3}$$

where $i = 1 \dots d$ and $(\sqrt{\alpha \boldsymbol{C}_R})_i$ defines the i-th column of the square root of $\boldsymbol{C}_R$. The scalar α denotes a weighting factor for the elements in the covariance matrix and is set to $\alpha = 2$ in the case of the Gaussian distribution. To demonstrate the equivalence of the initial distribution and the approximated one, we can compute the approximated sample mean vector $\boldsymbol{\mu}'$ and covariance matrix $\boldsymbol{C}'_R$:

$$\boldsymbol{\mu}' = \frac{1}{2d+1} \sum_{i=0}^{2d} \boldsymbol{w}_i \approx \boldsymbol{\mu}, \tag{4}$$

$$\boldsymbol{C}'_R = \frac{1}{2d} \sum_{i=0}^{2d} (\boldsymbol{w}_i - \boldsymbol{\mu}')(\boldsymbol{w}_i - \boldsymbol{\mu}')^T \approx \boldsymbol{C}_R. \tag{5}$$

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 that is eventually used as input to our modeling procedure:

$$\boldsymbol{s} = [\boldsymbol{w}_0^T \boldsymbol{w}_1^T ... \boldsymbol{w}_{2d+1}^T]^T. \tag{6}$$

F. 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 and the unscented transform 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)

$$p(\boldsymbol{s}|\lambda) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{s}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{7}$$

where the k-th Gaussian PDF $p(\boldsymbol{s}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is defined as

$$p(\boldsymbol{s}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{N/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{s} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{s} - \boldsymbol{\mu}_k) \right\}, \tag{8}$$

and $\{\pi_k\}_{k=1}^K$ 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

$$\log p(\boldsymbol{\Psi}|\lambda) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\boldsymbol{s}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{9}$$

Maximum likelihood (ML) solutions for the model parameters are found via the Expectation-Maximization (EM) algorithm [9] in our case initialized using K-means clustering.

When building user-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 user-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 (9) on all available training data. Once the UBM is build, user-specific GMM are computed via maximum a posteriori (MAP) adaptation [10], where only the mean vectors $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are adapted, by iterative evaluation of

$$\hat{\boldsymbol{\mu}}_k = (1 - \alpha)\boldsymbol{\mu}_k + \alpha \boldsymbol{\mu}_k^{EM}. \tag{10}$$

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 user, it can be used to train a classifier for this specific user. During the enrollment phase, when the user is first presented to the system, a SVM [11] classifier is trained for that user and a decision hyperplane between the super-vector of the "enrollee" and the super-vectors from some training 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, which is used as the similarity score.

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

G. Characteristics of the Proposed Approach

In this paper we propose a novel framework for 3D face recognition, which combines RCM-based local features and GMMs. The framework exhibits some useful characteristics that ensure robust and effective recognition performance as evidenced by the results presented in the next section:

i) RCM descriptors are capable of elegantly combining various face representations into a single coherent descriptor; they can be treated as an efficient data fusion/integration scheme;

ii) RCM descriptors do not encode information relating to the ordering or number of feature vectors in the region from which they were computed and can, therefore, be made scale and rotation invariant to some extent, but only if appropriate feature representations are selected for constructing $\boldsymbol{F}$ (see e.g., [4], [5]);

iii) since RCM descriptors are computable regardless of the number of feature vectors used for their computation, they can handle missing data (i.e., even in the presence of holes in the face scans or at regions near the borders of the face scans the RCM descriptor is still computable) in the feature extraction step; note that this is not the case for other local features commonly used with GMMs, such as 2D DCT features, which require that all elements of a rectangular image-block are present;

iv) 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; thus, RCM-based feature vectors allow for a *multi-aspect analysis*[2] of the 3D face scans;

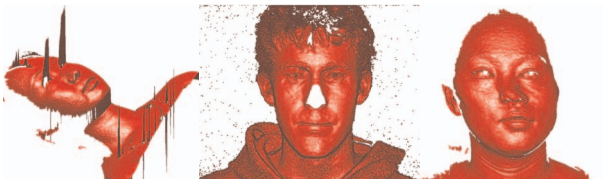
v) GMM-based systems treat data (i.e., feature vectors) as independent and identically distributed (i.i.d.) observations and, hence, present 3D facial images in the form of a number of orderless blocks; this characteristic is reflected in good robustness to imperfect face alignment, moderate pose changes and expression variations, as demonstrated by various researchers, e.g., [12], [13];

vi) the probabilistic nature of the GMMs makes it easy to include domain-specific prior knowledge into the modeling procedure, e.g., by relying on the universal background model (UBM).

III. Experiments

A. Database and Experimental Protocol

In our experiments, we use the Face Recognition Grand Challenge version 2 (FRGCv2) data set [14]. The images in the FRGCv2 data set have frontal view, exhibit minor pose- and major expression-variations. The images often contain shape artifacts such as deformed areas caused by subjects moving during the scanning procedure, nose absence, holes, small protrusions and/or impulse noise (see Fig. 2). The FRGCv2 experimental protocol provides a set of standard verification experiments and defines three data sets - the training set, the gallery set and the probe set. The training set is used to build global face models. The gallery set contains images with known identities (intended for enrollment), while the probe set contains images with unknown identities presented to the system for recognition (note that for the FRGCv2 data set the probe and gallery sets are actually identical). Our experiments use the entire FRGCv2 data set of 4007 depth images belonging to 466 subjects. We use only one enrollment image per client, so each image pair from the gallery and probe sets is considered independently. The employed experimental configuration (i.e., often referred to as the *all vs. all* configuration) results in more than 16 million comparisons. Note that we use 512 GMM components in the modeling phase of our procedure, construct a SVM for each depth image from the gallery set (using the FRGCv2 training set for the negative examples) and employ ZT-score normalization [13] as a standard procedure for all techniques when generating our results.



(a) (b) (c)

Fig. 2. Sample images with artifacts from the FRGCv2 data set: (a) impulse noise around eyes, (b) nose absence, (c) deformed image

B. Experiment 1: Constructing the composite representation

We have emphasized in Section II-C that there is no rule on how to select the face representations that should constitute the composite representation $\boldsymbol{F}$ for optimal recognition performance. In the first series of our recognition experiments we, therefore, aim at selecting a number of appropriate representations that could be used for this purpose. Specifically, we experiment with different feature types, such as pixel coordinates (x, y), depth values $\boldsymbol{I}$, shape index values $\boldsymbol{S}$, Gaussian curvature values $\boldsymbol{K}$, mean curvature values $\boldsymbol{H}$, minimum curvature values $\boldsymbol{P}_{min}$, maximum curvature values $\boldsymbol{P}_{max}$, surface normal coordinates $\boldsymbol{N}_x$, $\boldsymbol{N}_y$ and $\boldsymbol{N}_z$, and angle values φ between surface normals and the average facial normal. Note that the RCM descriptors are extracted from regions of two different sizes in all of our experiments, namely, from regions of 30×30 and regions of 40×40 image points. The regions are sampled from the images with a step size of 5 pixels[3].

If we look at Table I, where the results of our experiments are presented in the form of true accept rates (TAR) at the false accept rate (FAR) of 0.1%, the first thing to notice is that different combinations of image representations result in significantly different recognition results. Somehow unexpectedly,

[2]With the term *multi-aspect analysis* we refer to the fact that the face can be examined at different levels of locality all the way up to the holistic level.

[3]Note that more region sizes could be used in our experiments. However, due to the limited amount of space available for the paper, a more detailed analysis of the impact of the region size and number of region sizes adopted for the experiments will be presented elsewhere.

TABLE I. VERIFICATION PERFORMANCE (TAR @ 0.1% FAR) FOR DIFFERENT FEATURE VECTORS DEFINING $\boldsymbol{F}$

$\phi(\boldsymbol{I}, x, y)$	TAR
$[x\ y\ \boldsymbol{S}(x,y)\ \boldsymbol{N}_x(x,y)\ \boldsymbol{N}_y(x,y)\ \boldsymbol{N}_z(x,y)]$	**94.4%**
$[\boldsymbol{S}(x,y)\ \boldsymbol{N}_x(x,y)\ \boldsymbol{N}_y(x,y)\ \boldsymbol{N}_z(x,y)]$	93.4%
$[\boldsymbol{I}(x,y)\ \boldsymbol{S}(x,y)\ \boldsymbol{N}_x(x,y)\ \boldsymbol{N}_y(x,y)\ \boldsymbol{N}_z(x,y)]$	93.2%
$[\boldsymbol{N}_x(x,y)\ \boldsymbol{N}_y(x,y)\ \boldsymbol{N}_z(x,y)]$	86.3%
$[\boldsymbol{S}(x,y)\ \boldsymbol{\varphi}(x,y)]$	72.4%
$[\boldsymbol{I}(x,y)\ \boldsymbol{\varphi}(x,y)]$	65.6%
$[\boldsymbol{S}(x,y)\ \boldsymbol{K}(x,y)\ \boldsymbol{H}(x,y)\ \boldsymbol{P}_{min}(x,y)\ \boldsymbol{P}_{max}(x,y)]$	58.9%

composite representations with more face representations do not always outperform composite representations with fewer face representations. This fact suggests that complementary information needs to be included in the composite representation for better recognition performance. Among the assessed combinations, the highest true accept rate is achieved by the following 6-dimensional feature vector defining $\boldsymbol{F}$:

$$\boldsymbol{f} = [x\ y\ \boldsymbol{S}(x,y)\ \boldsymbol{N}_x(x,y)\ \boldsymbol{N}_y(x,y)\ \boldsymbol{N}_z(x,y)], \qquad (11)$$

and is, therefore, used in all of our subsequent experiments.

C. Experiment 2: Delta features

In the second series of experiments we examine how the introduction of delta features impacts the verification performance. Delta coefficients describe the dynamics between features of neighboring image blocks and introduce spatial dependencies into the feature vectors. When delta features are added to the feature vectors, each feature vector is implicitly linked to its (spatially) neighboring feature vectors, resulting in a chain of dependencies. Consider two d-dimensional local-feature vectors extracted from two (vertically or horizontally) neighboring blocks, i.e. $\boldsymbol{s}_i = [s_j^{(i)}]_{j=0}^{d}$ and $\boldsymbol{s}_{i+1} = [s_j^{(i+1)}]_{j=0}^{d}$ [12]. The j-th delta coefficient can then be defined as

$$\Delta s_j = s_j^{(i+1)} - s_j^{(i)}. \qquad (12)$$

In the case when the feature vectors $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i+1}$ are extracted from vertically neighboring blocks, we obtain vertical-delta coefficients and, similarly, if they are extracted from horizontally neighboring blocks we obtain horizontal-delta coefficients. Adding vertical and horizontal delta coefficients to the local feature vectors prior to GMM modeling is a common step in other fields, such as speaker [15] or face recognition [13], where, however, DCT-based features are used instead of RCM-based ones. In this paper we additionally introduce depth-delta coefficients, the concept of which is presented in Fig. 3. From the results of this series of experiments, presented in the form of ROC curves in Fig. 4, we can see that delta coefficients efficiently encode the spatial structure of the face and consecutively improve the verification performance. In fact, we can see that all types of delta features, including the newly introduced depth-delta features improve the recognition performance of our approach.

D. Experiment 3: Comparative assessment

In our last series of experiments we compare the performance of the proposed framework to the performance of some popular local-feature-based methods. The first method included in our comparison is based on the Scale-invariant

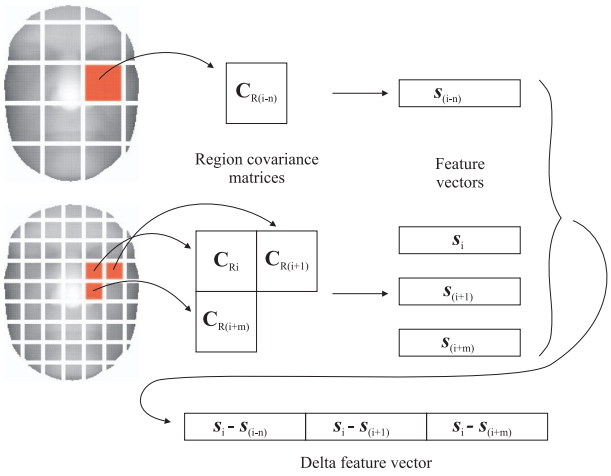


Fig. 3. Conceptual diagram of delta descriptor extraction

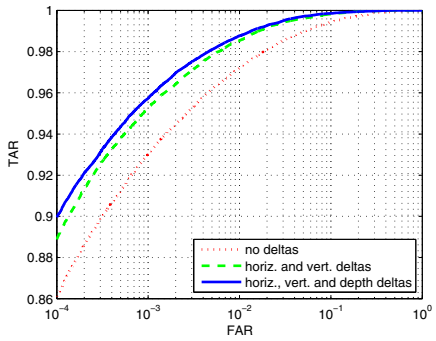


Fig. 4. ROC curves for different delta features (*all vs. all*)

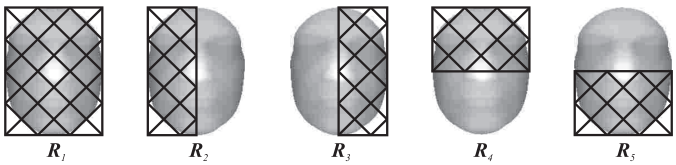


Fig. 6. Five RCMs are constructed from overlapping regions of the composite representation image in the RCMF approach

feature transform (or SIFT) [16], [17]. Here, the matching procedure is based on the number of successfully matched feature vectors from two images. If the number of successfully matched feature vectors is above some predefined threshold, the two images are declared a match, otherwise, they are declared a non-match. It should be emphasized here, that the feature vectors of the two images being compared are matched based on the procedure proposed in [17]. Thus, a feature vector $\boldsymbol{x}_1$ is successfully matched with the feature vector $\boldsymbol{x}_2$ only if the Euclidean distance $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ multiplied by some threshold value is not greater than the distance of $\boldsymbol{x}_1$ to all other descriptors. The default value of this threshold is 1.5 and is also used in all of our subsequent experiments. In addition to histogram-based features, as used in the original SIFT technique [17], we also assess the use of the RCM-based-local features in conjunction with the SIFT matching procedure (this approach is denoted as SIFT_RCM in Fig. 7). Furthermore, we compare the proposed framework with the method (denoted here as RCMF), which is often used in

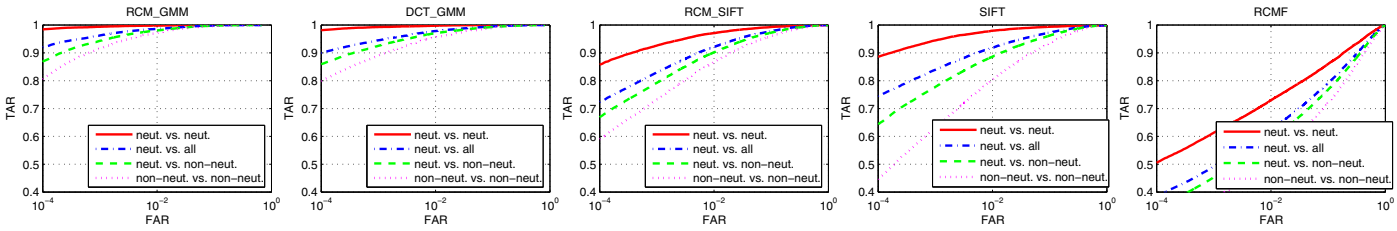


Fig. 5. Robustness to facial expression variations (from left to right): RCM_GMM, DCT_GMM, RCM_SIFT, SIFT, RCMF

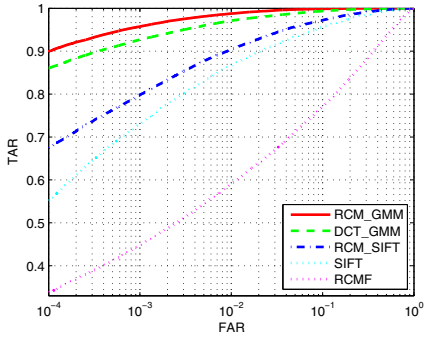


Fig. 7. ROC curves of compared methods

conjunction with RCM-based local features [4], [5], [8]. This method extracts the RCM descriptors from five regions of the face image as shown in Fig. 6. The similarity measure between the probe image $\boldsymbol{I}_p$ and the gallery image $\boldsymbol{I}_g$ is then defined as

$$\rho(\boldsymbol{I}_p, \boldsymbol{I}_g) = \sum_{j=1}^{5} \rho(\boldsymbol{C}_i^p, \boldsymbol{C}_i^g) - \max_j[\rho(\boldsymbol{C}_j^p, \boldsymbol{C}_j^g)], \qquad (13)$$

where $\boldsymbol{C}_i^p$ and $\boldsymbol{C}_i^g$ are RCMs from $\boldsymbol{I}_p$ and $\boldsymbol{I}_g$, respectively, and ρ is the Föerstner metric as defined in [6]. Finally, the last technique to be included in our comparison is a technique, where the RCM-based feature vectors are substituted with feature vectors based on Discrete Cosine Transform (DCT), while all other steps of the recognition framework are kept unchanged. This method is denoted as DCT_GMM in Fig. 7.

The verification performance (in the *all vs. all* configuration) of all the assessed methods is presented in Fig. 7. Note that among all tested techniques the proposed framework performs the best, achieving the TAR of 95.8% at the FAR of 0.1%, and followed in order by the DCT_GMM, RCM_SIFT, SIFT and RCMF techniques. The results suggest that both RCM features as well as the GMM modeling step contribute to the recognition performance. The worst performance is exhibited by the RCMF technique, where five descriptors computed from the 3D face image are clearly not enough to appropriately capture the identity information contained in the image.

In the second part of our comparative assessment, we examine how the evaluated techniques handle expression variations. To this end, we match all images marked as neutral from the probe set against all images marked as non-neutral in the gallery set (denoted as *neut. vs. neut.*). We also conduct experiments with other combinations of neutral and non-neutral images from the probe and gallery sets (note again that these two sets are actually identical). The hardest setting in this part of our assessment is the setting, where the non-neutral probe images are matched against the non-neutral gallery images (denoted as *non-neut. vs. non-neut.*). The results of this part of our assessment are shown in Table II and Fig. 5. We can see that among all tested methods, the proposed framework again results in the best performance on all four configurations, indicating that it exhibits some extent of robustness to facial expression changes. All techniques perform the best on the easiest configuration, where neutrally-labeled images are matched against each other. Here, the proposed framework results in the TAR of 99.4% at the FAR of 0.1%, while all other techniques perform worse. On the non-neutrally-labeled images, the proposed framework is the only assessed technique achieving a TAR of more than 90% at the FAR of 0.1%. However, as we can see, there is still plenty of room for improvement.

TABLE II. ROBUSTNESS OF THE ASSESSED TECHNIQUES TO THE FACIAL EXPRESSION VARIATIONS PRESENTED AS TAR @ 0.1% FAR

	neut. vs. neut.	*neut. vs. all*	*neut. vs. non-neut.*	*non-neut. vs. non-neut.*
RCM_GMM	99.4%	96.3%	94.4%	91.6%
DCT_GMM	99.3%	94.5%	92.5%	89.1%
RCM_SIFT	92.7%	83.0%	79.3%	73.2%
SIFT	94.2%	83.7%	77.8%	63.1%
RCMF	61.4%	49.2%	45.3%	37.6%

In the last part of our assessment we evaluate the time needed for the framework to classify an image and compare it to the time needed by the the remaining techniques. Note here that we used an Intel Xeon CPU @ 2.67 GHz and 6 GB of RAM. All techniques were implemented using Matlab and could, therefore, be significantly sped up if implemented with a compiled language like C/C++. The results of this part of our assessment are shown in Figs. 8 and 9. Note that the RCM_GMM framework requires the most time for the feature extraction step and, therefore, ranks somewhere in the middle when compared to the remaining techniques. The fastest among the tested methods is the RCMF technique. However, this comes at the expense of the lowest recognition performance.

E. Experiment 4: Comparison with the state-of-the-art

Last but not least, we compare the performance of the proposed framework to the performance of state-of-the-art techniques from the literature. The results of this comparison are shown in Table III. Note here that similar to all previous

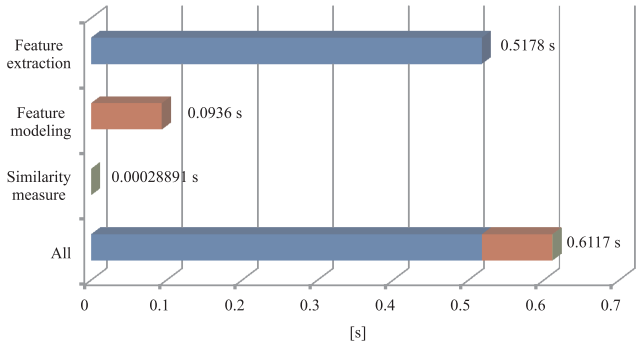


Fig. 8. Average time for classification of one image using RCM_GMM

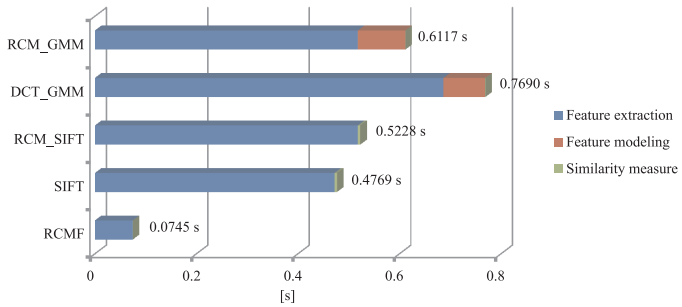


Fig. 9. Comparison of average times needed by all assessed techniques

TABLE III. COMPARISON OF THE VERIFICATION PERFORMANCE (TAR @ 0.1% FAR) WITH STATE-OF-THE-ART TECHNIQUES FORM THE LITERATURE

Method (*all vs. all*)	TAR
Mian et al. [18]	86.8%
Maurer et al. [19]	87.0%
Cook et al. [20]	92.3%
Faltemier et al. [21]	93.2%
Queirolo et al. [22]	96.5%
Li et al. [23]	95.3%
Our framework	95.8%

experiments the so-called *all vs. all* FRGCv2 experimental configuration is used for generating the results. We can see that the framework proposed in this paper is highly competitive and ranks among the top performers of the assessment. Another major point that we need to make here, is the fact that the proposed framework relies on no meta-data at all. All steps are completely automated, including the face localization/alignment/registration procedure, while this is not necessarily true for all techniques from Table III. In fact, as presented in Section II, our registration technique only crops the raw face scan in a circular manner around the nose tip and performs no further (more precise) alignment.

IV. CONCLUSION

We have presented a novel framework for 3D face recognition that relies on region covariance matrices and Gaussian mixture models. We have demonstrated the feasibility of the framework on the FRGCv2 database and shown its competitiveness when compared to state-of-the-art techniques form the literature. For our future work, we plan to more thoroughly assess the impact of various hyper-parameters of the framework on its performance, devise more elaborate schemes for selecting the face representations to combine into the composite representation and incorporating score normalization techniques into the framework to further improve its performance.

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," *Comp. Vis. and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[2] S. Julier and J. K. Uhlmann, "A General Method for Approx. Nonlinear Transformations of Probability Distributions," Tech. Rep., 1996.

[3] M. Segundo, C. Queirolo, O. Bellon, and L. Silva, "Automatic 3D face segment. and landmark detection," in *ICIAP*, 2007, pp. 431–436.

[4] O. Tuzel, F. Porikli, and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," in *ECCV*, vol. 3952, 2006, pp. 589–600.

[5] Y. Pang, Y. Yuan, and X. Li, "Gabor-Based Region Covariance Matrices for Face Recognition," *TCSVT*, vol. 18, pp. 989–993, 2008.

[6] O. Tuzel, F. Porikli, and P. Meer, "Human Detection via Classification on Riemannian Manifolds," in *IEEE CVPR*, 2007, pp. 1–8.

[7] F. Porikli, O. Tuzel, and P. Meer, "Covariance Tracking using Model Update Based on Lie Algebra," in *CVPR*, vol. 1, 2006, pp. 728–735.

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

[9] T. Moon, "The Expectation-Maximization Algorithm," *Sig. Proc. Magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

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

[11] H. Bredin, N. Dehak, and G. Chollet, "GMM-based SVM for face recognition," in *IEEE ICPR*, vol. 3, 2006, pp. 1111–1114.

[12] J. Križaj, V. Štruc, and S. Dobrišek, "Towards Robust 3D Face Verification using Gaussian Mixture Models," *International Journal of Advanced Robotics Systems*, vol. 9, pp. 1–11, 2012.

[13] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of normalization techniques from speaker to face authentication using GMMs," *IEEE TIFS*, vol. 7, no. 2, pp. 553–562, 2012.

[14] 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 *CVPR*, 2005, pp. 947–954.

[15] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE TASLP*, vol. 16, no. 5, pp. 980–988, 2008.

[16] T. R. Lo and J. P. Siebert, "Local feature extraction and matching on range images: 2.5D SIFT," *Comput. Vis. Image Underst.*, vol. 113, pp. 1235–1250, 2009.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[18] A. Mian, M. Bennamoun, and R. Owens, "An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition," *IEEE TPAMI*, vol. 29, no. 11, pp. 1927–1943, 2007.

[19] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, and G. Medioni, "Performance of Geometrix ActiveID 3D Face Recognition Engine on the FRGC Data," in *IEEE CVPR*, 2005, pp. 154–154.

[20] J. Cook, C. McCool, V. Chandran, and S. Sridharan, "Combined 2D/3D Face Recognition Using Log-Gabor Templates," in *IEEE Int. Conf. Video and Sig. Based Surveillance*, 2006, pp. 83–88.

[21] T. Faltemier, K. Bowyer, and P. Flynn, "A Region Ensemble for 3D Face Recognition," *IEEE TIFS*, vol. 3, no. 1, pp. 62–73, 2008.

[22] 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.

[23] X. Li and F. Da, "Efficient 3Dfacerecognition handling facial expression and hair occlusion," *Image Vision and Computing*, vol. 30, no. 9, pp. 668–679, 2012.