Which parts of the face give out your identity?

Omar Ocegueda Shishir K. Shah Ioannis A. Kakadiaris
Computational Biomedicine Lab, Depts. of Computer Science, Elec. & Comp.
Engineering, and Biomedical Engineering, Univ. of Houston, Houston, TX, USA
{jomaroceguedag@gmail.com, shah@cs.uh.edu, ioannisk@uh.edu}

Abstract

We present a Markov Random Field model for the analysis of lattices (e.g., images or 3D meshes) in terms of the discriminative information of their vertices. The proposed method provides a measure field that estimates the probability of each vertex to be "discriminative" or "non-discriminative". As an application of the proposed framework, we present a method for the selection of compact and robust features for 3D face recognition. The resulting signature consists of 360 coefficients, based on which we are able to build a classifier yielding better recognition rates than currently reported in the literature. The main contribution of this work lies in the development of a novel framework for feature selection in scenarios in which the most discriminative information is known to be concentrated along piece-wise smooth regions of a lattice.

1. Introduction

In this work, we directly address the problem of finding the most discriminative areas of the face for recognition. In our model, we assume that the discriminative information is contained along smooth regions of the face (in our model, it is unlikely that one single isolated vertex contains high discriminative information while its neighbors don't). To the best of our knowledge, none of the existing feature selection techniques exploits the regularity of the features computed from localized regions of images or 3D meshes. Furthermore, none of the existing Face Recognition (FR) techniques directly addresses the problem of finding the areas of the face that contain the most discriminative information.

We formulate the problem as an image/mesh segmentation problem under the framework of Markov Random Fields. The proposed method provides a measure field that estimates the probability of each vertex to be "discriminative" or "non-discriminative". The measure field corresponds to the marginal posterior probability measure field derived from a generation-observation model of the training set. We propose to use the posterior marginal probabilities as a measure of the discriminative information of a vertex. Experimental results show that using the context (in this case using the smoothness prior) may improve the performance of the selected features in this kind of applications. Finally as an application of our framework, we present a simple Forward Feature Selection (FFS) technique that uses the proposed metric for feature scoring.

Our contributions are: (i) the development of a novel framework for feature selection in scenarios in which the most discriminative information is known to be concentrated along piece-wise smooth regions of a lattice; (ii) the definition of a compact signature for 3D FR, which consists of only 360 coefficients (1.2% the size of the signature used in [6]); (iii) the development of a 3D FR system that presents better generalization performance than the state of the art under extreme facial expressions, according to our experiments performed in the publicly available BU-3DFE database [11].

The rest of the paper is organized as follows. In Section 2, we provide an overview of the previous work related to this paper. Section 3 introduces the notation and covers the background. In Section 4, we present the generation-observation model of the training data and formulate our problem as a Markov Random Field estimation problem. In Section 5, we present the proposed MRF estimation algorithm. In Section 6, we present an application and the experimental results of our method for 3D face recognition. We conclude in Section 7 with a discussion and summary of our work.

2. Previous work

Several methods have been proposed to analyze the discriminative power of different facial areas for FR. Etemad *et al.* [2] used LDA on images to extract a small set of features that carry the most relevant information for classification purposes. Instead of directly formulating the problem of finding the most discriminative areas of the face, they analyzed the discriminative power of a horizontal segment of the face that grows in height from top to bottom of the image. In this exploratory experiment, they showed that the most discriminatory information is located around the nose and the mouth. Daniyal *et al.* [1] proposed an algorithm for 3D FR in which the face is represented as a vector of distances between pairs of facial landmarks. They selected the landmarks by brute-forcing the possible combinations of used/unused landmarks, compared the recognition rates, and concluded that the best selection corresponded to the landmarks located around the eyes and the nose. Kakadiaris *et al.* [6] used the wavelet transform of the geometry and normal maps of the face scans to compare pairs of meshes. They annotated the wavelet coefficients according to an Annotated Face Model (forehead, eyes, nose, cheeks, mouth and neck) and then searched the optimal weights for each region to maximize the area under the ROC curve in the verification test. Savvides *et al.* [10] divided the facial regions into three sections: eyes, nose and mouth, and evaluated the performance of their method using each region separately. They reported that the eye-brow region consistently outperformed all other facial regions with a large margin of separation. In their 3D FR system, Faltemier *et al.* [3] defined 38 local regions (some of which overlap) by arbitrarily distributing their centroids along the face. Each region was then defined as the set of points located inside a sphere of a predefined radius centered at its corresponding centroid. Contrary to our approach, this method requires an pre-defined segmentation of the face.

The problem of finding the most discriminative facial areas for recognition may be formulated as the more general problem of feature selection. Huang *et al.* [7] presented an excellent overview of the feature selection problem and the current state of research in that field. They also proposed a unified model of the feature selection process. The feature selection methods may be categorized into three main classes: embedded, wrapper and filter. In the embedded methods, the feature selection is performed as part of the classification model (they are *embedded* into the classification model). Wrapper methods, (e.g. FFS or genetic search) perform a search over the space of all possible subsets of features and evaluate the performance of each subset. However, for large scale problems, wrapper methods are often impractical. In such cases, feature scoring metrics (filter methods) are used independently on each feature and the best k features are then used [4].

3. Background

A lattice $\mathcal{L}$ is a graph $\mathcal{G}(V, E)$ where V is the set of vertices and E is the set of edges. Images and 3D meshes, may be regarded as functions from the set of vertices V to a state space S. For example, if $S = \Re^3$ and to each vertex $v \in V$ we assign a vector x_v according to the position of the vertex in space, then $x = \{x_v | v \in V\}$ is known as the "geometry" of a 3D mesh because the values of the variables x_v define the "shape" of the mesh. The set of edges defines the *topology* of the lattice, since it defines the spatial relationship between the vertices. If instead of directly assigning a value in S to each vertex, we assign a random variable X_v, then x may be regarded as the realization of the random field $X = \{X_v | v \in V\}$. The set x is also called a *configuration* of the random field X. In general, we will assume that X_v is a random vector in $\Re^d$ and we will refer to the function $I : v \mapsto I(v) = x_v$ as x. The value x_v is called a *feature* assigned to vertex v. In the rest of this paper, we will not distinguish between a mesh and an image, as we are only concerned about functions defined on a set of vertices on which we have defined a neighborhood system. We assume that we are given a set of annotated images $\Omega = \left\{(x^i, y^i)\right\}_{i=1}^{N}$ (called the "training set"). The label y^i indicates that the image x^i belongs to the class y^i. For example, in FR, the set of classes represent the set of different individuals from which the images were obtained. Therefore, for a specific k we are given a subset of labeled images $\Omega_k = \left\{(x^i, y^i) | y^i = k\right\}$ which correspond to the same subject k. A segmentation of V is a partition of the vertices in V into a set of disjoint regions $V = \bigcup_{i=1}^{r} V_i$ such that the dataset is uniform, in some sense, over every region. The sense of "uniformity" depends on the segmentation task. We are interested in a binary segmentation of V such that the vertices are classified into two categories: "discriminative" and "non-discriminative", in the sense that the discriminative vertices are the "most important" for distinguishing between different types of images. For example, in FR, the objective is to determine which vertices are the "most important" for distinguishing between different individuals.

4. Generation-Observation Model

The problem of determining the *most discriminative areas* of a lattice $\mathcal{L}$ for a given classification task, may be formulated as a segmentation problem in which a label $q_v \in \{0, 1\}$ is assigned to each vertex $v \in V$, where the property of *being discriminative* is encoded by the label 1 and *being non-discriminative* is encoded by the label 0. The observed set of images can be modeled as:

$$\begin{array}{ll} x_v^i = \mu^k(v) + X_v^k, & y^i = k, v \in V_1 \\ x_v^i = \mu(v) + X_v, & \qquad\quad v \in V_0 \end{array}, \tag{1}$$

where x_v^i represents the value of the variable assigned to vertex v of the ith image, x^i. Specifically, inside the "discriminative" portion V_1 of V, each image x^i is a distorted version of a function $\mu^k(\cdot)$ that depends on the class $y^i = k$ of the image, and inside the "non-discriminative" portion V_0 of V, each image is a distorted version of a common function $\mu(\cdot)$ that does not depend on the class of the image. The random fields $X^k = \left\{X_v^k | v \in V\right\}$ model the deviation

of the data samples x^i of class $y^i = k$ from the k^{th} base shape μ^k, and the random field $X = \{X_v | v \in V\}$ models the deviation of the samples x^i from the common *base shape* μ. We model the "distortion" of the observed data from the smooth functions as white noise, that is X_v and X_u are independent for all $u \neq v$.
Our goal is to estimate the field q defined above, which encodes the segmentation of the lattice. To this end, we model q as a Markov Random Field, and incorporate our prior knowledge about q as a Gibbs distribution:

$$P_q(q) = \frac{1}{Z_q} exp\left(-\beta \sum_{c \in C} W_c(q)\right). \tag{2}$$

The "potential functions" W_c depend only on the values of the field q at the vertices that belong to the corresponding clique $c \in C$, Z_q is a normalization constant and C is the set of "cliques" of the neighborhood system defined in V. We are interested in applications where the features computed for each vertex v are localized, in the sense that the vector x_v represents a localized feature of the image around the vertex v. In such applications, we would like q to be spatially smooth.

This "prior" knowledge about q can be represented using potentials of the form

$$W_{<u,v>}(q) = \left\{ \begin{array}{rl} -1, & q_u = q_v \\ 1, & q_u \neq q_v \end{array} \right. , \tag{3}$$

where u and v are neighboring vertices. The parameter β in Eq. 2 controls the granularity of the field q. Since the distributions f_v^k of X_v^k and f_v of X_v are unknown, we may approximate them using the distributions that maximize the likelihood of the data:

$$\widehat{f_v^k} = \arg\max_{\mathbf{f} \in F} \prod_{x \in \Omega_k} f(x_v), \tag{4}$$

where F is a suitable set of probability density functions. For example, if we choose F to be the set of Gaussian density functions, then $\widehat{f_v^k} = \phi(\hat{\mu}, \hat{\Sigma})$, where ϕ is the Gaussian density function and $\hat{\mu}, \hat{\Sigma}$ are the maximum likelihood estimators of μ and Σ for the subset Ω_k at vertex v. We will approximate f_v accordingly by:

$$\widehat{f_v} = \arg\max_{\mathbf{f} \in F} \prod_{x \in \Omega} f(x_v). \tag{5}$$

The likelihood of the training set is given by:

$$P_{\Omega|q}(q) = \prod_{v \in V} g_v(q_v), \tag{6}$$

where

$$g_v(q_v) = \left\{ \begin{array}{ll} \prod_{i=1}^N \widehat{f_v}(x_v^i), & q_v = 0 \\ \prod_{i=1}^N \widehat{f_v^k}(x_v^i), & q_v = 1, y^i = k \end{array} \right. . \tag{7}$$

The posterior distribution for q is given by:

$$P_{q|\Omega} = \frac{P_q(q) P_{\Omega|q}(q)}{P_g(g)} = \frac{1}{Z} e^{-U(q)}, \tag{8}$$

where Z is a normalizing constant and

$$U(q) = -\sum_{v \in V} log(g_v(q_v)) + \beta \sum_{<u,v>} W_{<u,v>}(q_u, q_v). \tag{9}$$

In a standard image estimation problem, one may be interested in the maximum a posteriori (MAP) estimator for q. However, in this case, $g_v(0)$ is always less than $g_v(1)$ because the maximum likelihood distribution of the model that takes into account the class of each image will always fit better (the data) than the general model that tries to fit the complete data set, regardless of their classification (i.e., the models are *nested*).

5. Gauss-Markov Posterior Marginals

The MAP estimator for q is found by setting $V_1 = V$ (i.e., $q_v = 1, \forall v \in V$). However, the posterior marginal probability at each vertex, $\pi_v(b)$, is much more informative, since it is the probability of each vertex being discriminative ($\pi_v(1)$) or not discriminative ($\pi_v(0)$) under the given prior assumptions:

$$\pi_v(b) = Pr(q_v = b|\Omega) = \sum_{q : q_v = b} P_{q|\Omega}(q), b \in \{0, 1\}. \tag{10}$$

Clearly, the number of possible configurations for q (the number of terms in Eq. 10) makes it impossible to compute the marginal probabilities explicitly. Many algorithms have been proposed to estimate these probabilities [5] [8]. Marroquin *et al.* [8] presented a very efficient and convenient model to estimate the posterior marginal probabilities. They used the Central Limit Theorem to model the distribution of the empirical marginals, p, as a Normal distribution whose mean is the true marginals, π, and whose variance vanishes asymptotically. They show that the empirical marginals form a MRF with the same neighborhood structure as the original, q. By imposing the consistency constraint

$$\lim_{\beta \to 0} p^* = \widehat{g}, \tag{11}$$

where p^* is the optimal estimator for the posterior marginals, they proposed a general form of the energy function for their measure field model (Gauss-Markov Measure Field) which, when $W_{<r,s>}$ is the Ising potential [5], (as in Eq. 3), behaves like a set of decoupled membrane models. Its Gibbsian energy is given by:

$$U(p) = \sum_{v \in V} |p_v - \widehat{g}_v|^2 + \lambda \sum_{<u,v>} |p_v - p_u|^2. \tag{12}$$

The parameter $\lambda \geq 0$ acts as a regularization parameter, which controls the granularity of the final segmentation. The first term will enforce the field to be similar to the normalized likelihood $\widehat{g}_v$, while the second term will enforce the estimated probability field to be smooth. A simple and efficient way to minimize this energy function is to use Gauss-Seidel iterations. After setting the partial derivatives to zero and solving for p_v, we obtain the following relaxation update expression:

$$p_v(b) = \frac{\widehat{g}_v(b) + \lambda \sum_{u \in N_v} p_v(b)}{1 + \lambda \sharp N_v}, \tag{13}$$

where $\sharp$ denotes the *cardinality* operator and N_v denotes the set of neighboring vertices of v ($\sharp N_v$ denotes the *number of neighbors of* v). Note that the vectors p_v have only two elements. Since the membranes are decoupled, we can compute only one of the elements (e.g., $p_v(1)$) of each vector and then take $p_v(0) = 1 - p_v(1)$.

We propose to use the posterior marginal probability as a measure of the discriminative information of a vertex, and will define the "discriminative map" $\Delta : V \to [0, 1]$ as $\Delta(v) = \widehat{p}_v(1)$. We will also refer to these approximated marginal probabilities ($\widehat{p}_v(1)$) under our generation observation model as "Gauss-Markov Posterior Marginals" (GMPM).

6. Experiments

The proposed model requires us to select a family of distribution functions for our input data in order to compute the normalized likelihood $\hat{g}_v$. In this paper, we will show the results of the Gaussian case as we are interested in applying our framework to the wavelet signatures of the system proposed by Kakadiaris *et al.* [6]. A key feature of this system is that after fitting the AFM to the data, we obtain a direct correspondence between the vertices of the fitted meshes, which means that the ith vertex of the AFM corresponds to the same anatomical point of the fitted meshes. The construction of this anatomical correspondence (commonly referred to as "registration") is a standard step in FR and we will assume that this step has already been performed on the images/meshes provided as input to our algorithm (Algorithm 1). Since each wavelet coefficient is a linear combination of the data in a localized region of the face, we may, according to the Central Limit Theorem, (and assuming independence) reasonably approximate the distribution of the wavelet coefficients using a normal distribution.

In the following experiments, we use the Gaussian distribution as the parametric family of distributions to compute the likelihood in our model. The distribution that maximizes the likelihood of the training data $\left\{(x^i, y^i)\right\}_{i=1}^N$ is given by:

$$g_v(q_v) = \begin{cases} \prod_{i=1}^N \phi(x_v^i; \widehat{\mu}_v, \widehat{\Sigma}_v) & q_v = 0 \\ \prod_k \prod_{i:y^i=k} \phi(x_v^i; \widehat{\mu}_v^k, \widehat{\Sigma}_v^k) & q_v = 1 \end{cases}, \tag{14}$$

where $\phi(x, \mu, \Sigma)$ is the normal density function with parameters μ, Σ. The estimators $\widehat{\mu}_v$ and $\widehat{\Sigma}_v$ are the *between-class* maximum likelihood estimators of the mean and variance at vertex v, respectively, while $\widehat{\mu}_v^k$ and $\widehat{\Sigma}_v^k$ are the *within-class* maximum likelihood estimators of the mean and the variance at vertex v, considering only samples of class k.

Algorithm 1 GMPM-Gaussian

```
Require: A set $T = (x_i, y_i)$ of registered images/meshes labeled according to a classification task
Require: $\lambda \geq 0$ {The regularization parameter}
Compute the likelihood using the Gaussian distribution
    • Use the between mean and covariance for $g_v(0)$
    • Use the within means and covariances for $g_v(1)$
for all $v \in V$ do
    $\widehat{g}_v(1) \leftarrow \frac{g_v(1)}{g_v(0)+g_v(1)}$ {Normalize the likelihood}
end for
$t \leftarrow 0$
$p^0 \leftarrow \widehat{g}$ {Initialize the posterior marginals estimator to the normalized likelihood}
repeat
    $p^{t+1}$=$p^t$
    for all $v \in V$ do
        $p_v^{t+1}(1) = \frac{\widehat{g}_v(1)+\lambda \sum_{u \in N_v} p_v^{t+1}(1)}{1+\lambda \sharp N_v}$
    end for
    $t \leftarrow t + 1$
until convergence
for all $v \in V$ do
    $\Delta(v) \leftarrow p_v^t(1)$
end for
return $\Delta$
```

6.1. 3D Face Recognition system

We applied our framework on the 3D FR system presented by Kakadiaris *et al.* [6], since that system won the Face Recognition Vendor Test (FRVT) for 3D FR [9]. The authors call their system "UR3D". The effectiveness and robustness of this system is achieved by an accurate alignment of the meshes to a common reference Annotated Face Model (AFM), followed by the deformation of the AFM to fit each input mesh. The most important property of the AFM is its continuous global UV parametrization, which allows the mapping of vertices from $\Re^3$ to $\Re^2$ and vice versa: each vertex v is assigned 2D coordinates (i, j). The UV parametrization allows to represent the 3D mesh as a three-channel image: when the position x_v of each vertex v is mapped onto its coordinates (i, j) in the image, we obtain the so called "geometry image". From the geometry image, a "normal image" is constructed by computing the normal

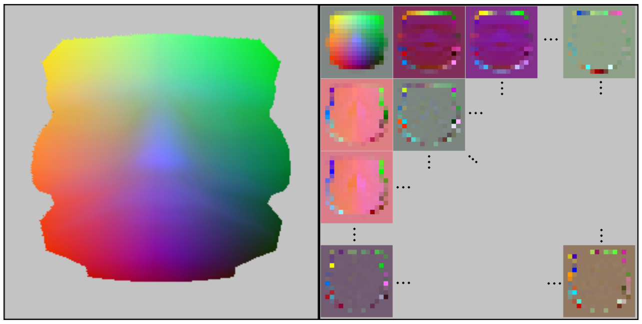


Figure 1: Full Walsh wavelet packet decomposition of the geometry image. After the four channel filter bank is applied, the decomposition is recursively applied to each of the four channels, LL, HL, LH and HH. In the standard UR3D system, the resulting image contains 16x16 packets at the last level of decomposition. The same decomposition is applied to the normal map. Depiction of the (a) geometry image, and (b) full Walsh wavelet packet decomposition.

vector at each vertex. The full Walsh wavelet packet decomposition is extracted from each band (x, y and z) of the geometry and normal images to obtain a total of 512 wavelet packets: 256 corresponding to the geometry image and 256 corresponding to the normal image. Figure 1b illustrates the transformation of the geometry image to a set of 256 wavelet packets. We depict the geometry image as a color image by assigning the (R, G, B) components to the three bands (X, Y, Z), respectively. Only 80 of the above packets define the *signature* of a 3D facial mesh, 40 corresponding to the geometry image and 40 corresponding to the normal image. Since each pixel in the wavelet transform contains information about a specific region of the face, a different weight is assigned to each pixel. In the following sections, we will show an application of our theoretical framework to select packets and weights that define a compact signature using the feature scoring provided by the GMPM model.

6.2. Behavior of GMPM with respect to the hyper-parameters

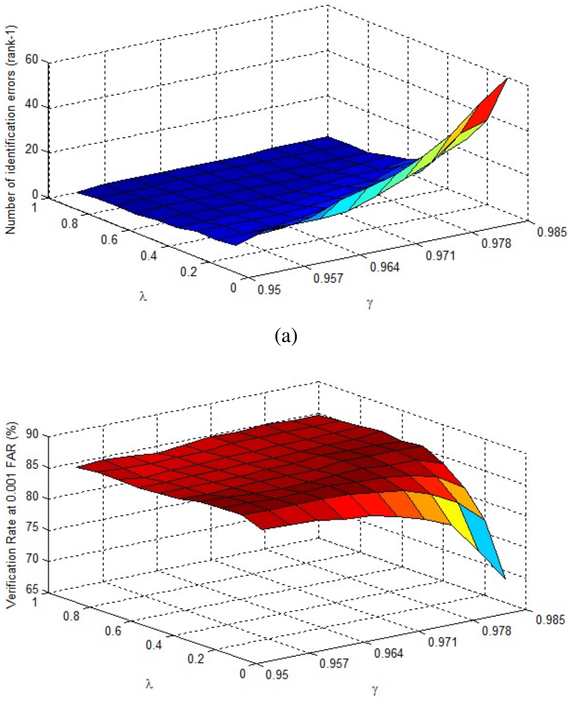
In order to compute the discriminative map of a set of images, we need to specify one hyper-parameter, λ, that controls the granularity of the solution. In order to use the discriminative map for feature selection, we may simply select a threshold τ and retain all the vertices v for which $\Delta(v) > \tau$. An alternative way is to specify the desired proportion $\gamma \in [0,1]$ of coefficients that are to be eliminated. Specifically, given γ we compute $\tau(\gamma)$ as follows:

$$\tau(\gamma) = \min\left\{ x \geq 0 : \gamma \leq \frac{\sharp\{v : \Delta(v) \leq x\}}{\sharp V} \right\}. \quad (15)$$

Note that $\tau(0) = 0$ and $\tau(1) = max\{\Delta(v) : v \in V\}$. Using this threshold, we eliminate a proportion of vertices approximately equal to γ. In the first experiment, we used the full wavelet transforms of the geometry and normal images as input to our method. For each input scan we have one 3-band 256x256 geometry image and one 3-band 256x256 normal image. We computed the discriminative map using 50% of the BU-3DFE dataset [11]. Figure 2(a) shows the number of classification errors in the identification test on BU-3DFE as a function of λ and γ. Notice that for small signatures (high values of γ), the regularization parameter λ has a better impact on the feature selection performance (the graph is decreasing as a function of λ for a fixed γ). Similarly, Fig. 2(b) depicts the verification rate at 0.001 FAR on BU-3DFE as a function of the hyper-parameters. The same observation applies in this case: for high values of γ, the regularization parameter λ has a higher (positive) impact on the feature selection performance (the graph is increasing as a function of λ for a fixed γ). This experiment indirectly compares the performance of the GMPM criterion against the generalized likelihood ratio [12] as a feature scoring metric computed from individual features without considering neighbors. This is because setting the regularization parameter to zero in the GMPM model is equivalent to the log-likelihood ratio criterion. The conclusion of this experiment is that the regularization term in the GMPM model improves the feature selection, and thus the GMPM criterion is superior to its non-regularized counterpart.

6.3. GMPM for compact wavelet coefficient selection

The direct method for feature selection described in Sec. 6.2 illustrates the behavior of the GMPM criterion used as a *filter*-type feature selection method (i.e., selecting the best k features according to the GMPM scoring function). However, it has been shown that wrapper methods often perform better and that the scoring methods can be used to guide the search of *wrapper* methods more effectively [4]. The observed behavior of the GMPM model suggests that we can use a forward-feature selection scheme in combination with the GMPM criterion to efficiently construct a better feature subset. We used a simple FFS algorithm: at each iteration select the packet that, added to the current packet set (starting with an empty set), maximizes the verification rate at 0.001 FAR; the only variation to the standard FFS is that we assign different weights to the pixels using the GMPM as a scoring function. Figure 3 depicts an example of the "discriminative maps" computed in the wavelet domain. Note that each packet defines a different discriminative map of the face. For simplicity, we only show the result for four wavelet packets, the discriminative maps at the top row were computed from the geometry images and the discriminative maps at the bottom row were computed



(a)

(b)

Figure 2: Recognition performance on BU-3DFE as a function of $\lambda \in [0, 1]$ and $\gamma \in [0.95, 0.985]$: (a) number of classification errors, and (b) verification rate at 10^{-3} FAR.

from the normal images. In addition to the wavelet packet selection, our method provides a measure of the discriminative power of the vertices, which we directly use as weights for the wavelet coefficients. At the end, 12 wavelet packets, (6 from the geometry image and 6 from the normal image) containing 360 wavelet coefficients in total, are selected. Figure 4 depicts the ROC curve obtained using this compact signature with the UR3D metric for FR. In summary, we obtain a similar recognition performance (Fig. 4) using a set of coefficients of size 1.2% of the original signature.

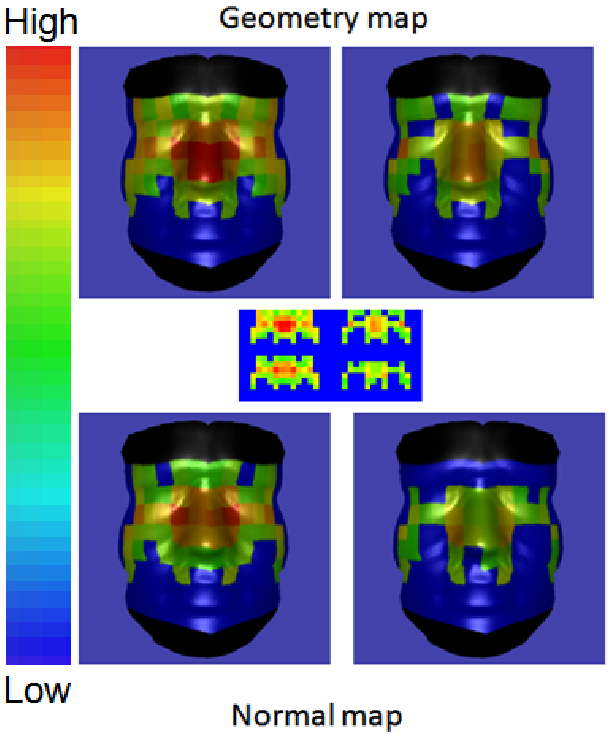


Figure 3: Discriminative map computed in the wavelet domain mapped back to the 3D mesh. The packets at the top row were computed using the geometry images and the packets at the bottom row were computed using the normal images. Different packets define different linear transformations of the face regions which in turn may result in different discriminatory power.

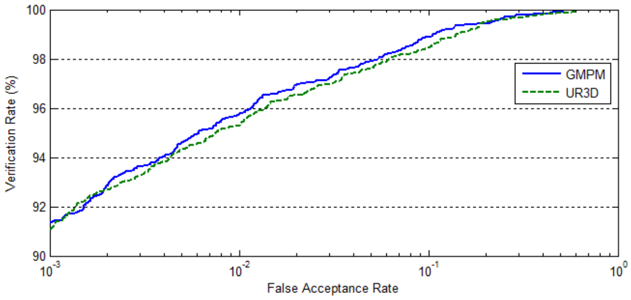


Figure 4: ROC curve on BU-3DFE using the UR3D metric and the 12-packet signature. The new signature contains only 360 coefficients, while the original one contained 30,720 coefficients.

6.4. LDA-based classification

In the previous experiments, the classification model consisted of a direct comparison of the selected coefficients using a weighted L_1 norm (which is the metric used in the UR3D system). However, it is often the case that the information relevant to the separation of the classes is contained in just a few *discriminant directions* [12]. A classical method to find such discriminant directions is Linear Discriminant Analysis (LDA). Given a set of training data $\left\{(x^i, y^i)\right\}_{i=1}^{n}$, $x^i \in \Re^d$, $y^i \in \{1, 2, ..., K\}$ LDA provides a set of discriminant directions $\{\alpha_i\}_{i=1}^{k}$, where $k = min\,\{d, K-1\}$, that maximize the Fisher's criterion

$$\arg\max_{\boldsymbol{\alpha} \in \Re^d} \frac{\boldsymbol{\alpha}^T \boldsymbol{B} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{W} \boldsymbol{\alpha}}, \tag{16}$$

where 16, $\boldsymbol{B}$ is the *between class* scatter matrix and $\boldsymbol{W}$ is the *within class* scatter matrix. These directions are given in decreasing order of importance. It has been shown that using only the leading discriminant directions can often improve the generalization performance of the classifier [12]. After selecting a small number of wavelet coefficients with the procedure described in Sec. 6.3, we applied LDA to the resulting wavelet signatures. We tested the performance using three different training sets:

- Cohort 1 (C_1): datasets from BU-3DFE, which consists of 2,500 3D meshes obtained from 100 individuals. Only 4% of the meshes present a neutral expression
- Cohort 2 (C_2): datasets from FRGC v1, which consists of 943 3D meshes obtained from 275 individuals.
- Cohort 3 (C_3): $C_1 \bigcup C_2$.

For cohorts C_1 and C_2, the following protocol was applied. First, the training data was divided into two disjoint sets containing 70% and 30% of the **individuals**, respectively. The first set was used as *training* set to compute the discriminant directions and the second set was used as *test* set to estimate the out-of-sample error of the classifier. After the best discriminant directions were computed from the training set, the recognition performance of the classifier was computed varying the number of discriminant directions used for recognition. Finally, the minimum number of discriminant directions that maximize the recognition performance in the test set was selected. Once the system was fully defined, we tested its performance on the FRGC v2 data set (no information about FRGC v2 was used during the training phase). For Cohort C_3, we simply combined the two similarity matrices obtained from C_1 and C_2. The recognition performance is defined as the verification rate at 0.001 FAR and the distance between two projected vectors is simply the Euclidean distance. Figure 5 depicts the recognition performance as a function of the number of discriminant directions using BU-3DFE for training. Table 1 summarizes the results obtained using the three different training sets. The significant performance boost of LDA-C_3 is explained by the fact that some of the individuals from which the 3D meshes were obtained are present in both, FRGC v1 and v2. The conclusion is that FRGC v1 actually provides significant information about FRGC v2.

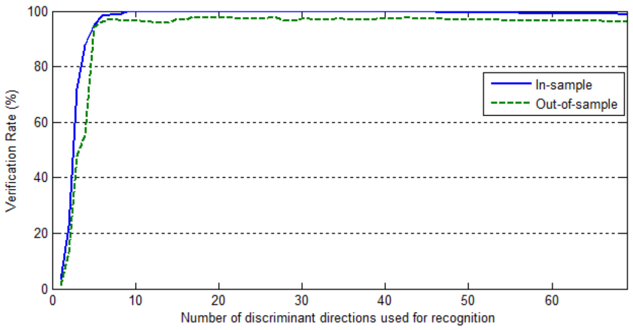


Figure 5: Verification rate at 0.001 FAR as a function of the number of discriminant directions used for recognition. In this case, the training set was BU-3DFE. The minimum directions with the best performance is obtained using 17 directions, with which we obtain 100% verification rate on the training set and 98% verification rate on the testing set.

Experiment	UR3D	LDA C_1	LDA C_2	LDA C_3
ROC I	97.1%	95.2%	96.2%	97.5%
ROC II	96.8%	95.2%	95.7%	97.1%
ROC III	96.7%	95.1%	95.2%	96.8%

Table 1: Verification rate at 0.001 FAR evaluated on FRGC v2 using the 3 different training sets (see text for details). The baseline is the performance obtained by the UR3D system using only the Walsh wavelet packet coefficients.

In FRGC v2, almost 60% of the meshes present a neutral expression. In order to test the generalization performance of the two methods, we used FRGC v2 for training and BU-3DFE for testing, with the same protocol as before. In this case, we used the full UR3D system, which uses a second set of Walsh wavelet packets and pyramid wavelets as additional information to boost its performance [6]. Since BU-3DFE does not provide a set of masks for the verification experiment (as FRGC v2 does), we build two masks with the following characteristics:

- ROC I: uses the neutral expression faces as gallery and the rest as probe.
- ROC II: every pair of meshes is tested.

The second experiment is significantly more challenging than the first one because the gallery may present any facial expression.
Table 2 presents the results of the two experiments described above. In both cases, GMPM+LDA performs significantly better than the UR3D system. This can be explained by the use of a more compact wavelet signature and the *reduced rank* LDA (using a subset of the discriminant directions), which reduces the *curse of dimensionality* and improves the generalization performance of the classifier.

Concerning the computational time of Algorithm 1, the most computationally expensive step is the computation of the normalized likelihood (lines 1 to 6), which involves the computation of a covariance matrix (and its inverse) for each vertex of the lattice, which is an $m \times m$ matrix,

Experiment	UR3D	GMPM+LDA
ROC I	90.5%	96.3%
ROC II	86.6%	94.8%

Table 2: Verification rate at 0.001 FAR evaluated on BU-3DFE using FRGC v2 for training.

where m is the dimension of the feature vector at each vertex. However, in practice, m is reasonably small ($m = 3$ in our experiments). The full discriminative map can be computed in less than 7 minutes using the wavelet coefficients of the full BU-3DFE database (2,500 3-band images of size 256×256). The FFS algorithm proposed in Sec. 6.3 is the most computationally expensive step of the algorithm. It requires the evaluation of the verification rate for each packet, taking into account only the wavelet coefficients selected using the GMPM criterion, which reduces the number of effective coefficients to $26/256$ at each packet. However, the FFS stops immediately if none of the remaining packets improves the recognition performance. In our experiments, the full training for the BU-3DFE database takes less than 4 hours using an AMD Opteron processor at 2.1 GHz (no multithreading).

7. Conclusion

We presented an MRF model for the analysis of lattices in terms of the discriminative information of their vertices. The posterior marginal probabilities of the MRF provide useful information that can be exploited as a feature scoring metric for feature selection. An advantage of using the posterior marginals for feature selection is that the selected features are meaningful in the domain of our problem. For example, in FR, we observed that the nose and the eyes are consistently marked as discriminative regions of the face. We presented an extension of the 3D FR system presented by Kakadiaris *et al.* [6] which significantly improves its generalization performance. We tested the performance of our system using both FRGC v2 and BU-3DFE databases obtaining superior recognition rates. In addition to the gain in recognition accuracy and robustness, the use of a more compact signature describing a 3D facial mesh has a direct impact in the computational efficiency of the recognition system in terms of storage requirements and computational time.

Acknowledgments

This research was funded in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL) and by the University of Houston (UH) Eckhard Pfeiffer Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, the U.S. Government, or UH.

References

[1] F. Daniyal, P. Nair, and A. Cavallaro. Compact signatures for 3D face recognition under varying expressions. In *Proc. Advanced Video and Signal Based Surveilance*, pages 302–307, Genova, Italy, September 2009.

[2] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of the Optical Society of America A*, 14(8):1724–1733, August 1997.

[3] T. Faltemier, K. Bowyer, and P. Flynn. A region ensemble for 3D face recognition. *IEEE Transactions on Information Forensics and Security*, 3(1):62–73, March 2008.

[4] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, March 2003.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[6] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):640–649, April 2007.

[7] H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature selection: An ever evolving frontier in data mining. In *Proc. The Fourth Workshop on Feature Selection in Data Mining*, volume 4, pages 4–13, Hyderabad, India, June 2010.

[8] J. L. Marroquin, F. A. Velasco, M. Rivera, and M. Nakamura. Gauss-Markov measure field models for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):337–348, April 2001.

[9] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831–846, May 2010.

[10] M. Savvides, R. Abiantun, J. Heo, S. Park, C. Xie, and B. Vijayakumar. Partial and holistic face recognition on FRGC-II data using support vector machine kernel correlation feature analysis. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pages 48–53, New York, NY, USA, June 2006.

[11] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *Proc.* 8^{th} *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, Amsterdam, The Netherlands, September 2008.

[12] M. Zhu and T. J. Hastie. Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics*, 12(1):101–120, March 2003.