UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze
Dipartimento di Ingegneria dell'Informazione (DINFO)
Corso di Dottorato in Ingegneria dell'Informazione

Curriculum: Informatica - (SSD ING-INF/05)

Representing faces: local and holistic approaches with application to recognition

Candidate
Claudio Ferrari

Supervisors
Prof. Alberto Del Bimbo

Dr. Giuseppe Lisanti

PhD Coordinator
Prof. Luigi Chisci

ciclo XXX, 2014-2017

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

Dedicated to all the adversities and misfortunes,
and the beauty of our own relentless will.

Acknowledgments

I would like to acknowledge the efforts and valuable input of my supervisor, Prof. Alberto Del Bimbo, and all my colleagues and friends of the Media Integration and Communication Center (MICC). In particular, my sincere thanks go to Dr. Giuseppe Lisanti and Prof. Stefano Berretti, precious advisers who collaborated on the main parts of my research work.

My loving gratitude is also dedicated to my family, Loretta, Michele, Giulia, Nada, Marco, Guido and Maria; to my dearest friends Alessandro, Andrea B., Andrea P., Lorenzo, Marco and Simone; to Bianca, my most precious gift.

Thank you, really.

Abstract

Face analysis from 2D images and videos is a central task in many computer vision applications. Methods developed to this end perform either face recognition or facial expression recognition, and in both cases results are negatively influenced by variations in pose, illumination and resolution of the face. Such variations have a lower impact on 3D face data, which has given the way to the idea of using a 3D Morphable Model as an intermediate tool to enhance face analysis on 2D data. In the first part of this thesis, a new approach for constructing a 3D Morphable Shape Model (called DL-3DMM) is proposed. It is shown that this solution can reach the accuracy of deformation required in applications where fine details of the face are concerned. The DL-3DMM is then exploited to develop a new and effective frontalization algorithm, which can produce a frontal facing view of unconstrained face images. The rendered frontal views result artifact-free and pixelwise aligned, so that matching consistency between local descriptors is enhanced. Results obtained with this approach are comparable with the state-of-the-art. Lately, in contrast to local descriptors based approaches, methods grounded on deep learning algorithms proved to be dramatically effective for face recognition in the wild. It has been extensively demonstrated that methods exploiting Deep Convolutional Neural Networks (DCNN) are powerful enough to overcome to a great extent many problems that negatively affected computer vision algorithms based on hand-crafted features. The DCNNs excellent discriminative power comes from the fact that they learn low- and high-level representations directly from the raw image data. Considering this, it can be assumed that the performance of a DCNN are influenced by the characteristics of the raw image data that are fed to the network. In the final part of this thesis, the effects of different raw data characteristics on face recognition using well known DCNN architectures are presented.

Contents

Chapter 1

Introduction

The automatic analysis of the human face has always attracted much attention in the computer vision and machine learning fields. In the last few years, thanks to the technological advances that have opened the way to the effective use of deep learning algorithms, we have witnessed substantial changes in how computer vision problems are addressed, including the ones concerning the analysis of the human face. Classic computer vision methods used to exploit hand crafted feature descriptors and human engineered methods to devise a suitable representation of an image and extract semantic information from it. With deep learning methods based on *Convolutional Neural Networks* (CNN) the design of this representation is delegated to the network itself. Low and high level features are learned by the networks directly from the raw image data. As a result, approaches based on the latter have demonstrated to outperform classic computer vision methods in many applications by a large margin. The natural consequence is that most of the research conducted up to this point is largely focused on deep learning.

However, in both cases, a great challenge summarily consists in finding the best "transformation" to convert the raw pixels of the image to a compact and discriminative representation that condense the semantic information carried by the image.

1.1 The objective

The main focus of this thesis is the problem of face recognition "in the wild" from images and video sequences. Face recognition trivially consists in associating a label that represents the subject's identity to a face image. Generally speaking, most of the face-related tasks consist in associating a label to a face image; this can be the emotional state of the subject captured, its expression, the gender and so on. All these problems are related by the fact that a suitable representation of the face image must be derived in order for a machine to extract discriminant information from it and perform the labeling task. It is clear in this sense that, depending on the task and the algorithms used to perform the recognition, different representations are likely to be more or less effective. Thus, it is reasonable to argue that, even though recognition pipelines usually consist of many modules, the chosen representation plays an important role for the final performance of the system.

In this context the main challenge for face recognition systems that operate in unconstrained scenarios is that the appearance changes in the face images due to real world conditions are usually very large. The semantic information associated to the identity that can be extracted is likely to be concealed by other factors. These can concern the subject captured, for example changes in pose, expression and aging but can also be related to the images themselves like illumination changes, different resolutions, occlusions and context. An example is shown in Fig. 1.1; the reader can easily deduce that the problem is considerably hard. The goal of this research work consisted in finding an effective face representation that is invariant, at least to some extent, to the above mentioned issues. The problem has been addressed both from a computer vision and a deep learning point of view. The main difference between the two consists in how the image representation is devised.

Computer vision approaches are generally based on computing *local* feature descriptors; these are human engineered descriptors built in order to capture recurring patterns in the image pixels e.g. Local Binary Patterns [1], or some related characteristics e.g. gradient orientations [50]. All these descriptors have been developed in the attempt of capturing discriminant characteristics of objects. Even though different descriptors can result more or less effective depending on the object to be described, they are rather generic so that they can be used in different applications e.g. object detection, face recognition, image stitching, 3D reconstruction. In all these cases, such descriptors are usually computed in many different locations across the image and concatenated so as to devise a representation of the whole object, faces in our case. However, in spite of their versatility and fast computation, in real world scenarios, where appearance changes are due to a large number of factors, they do not suffice to capture the details that allow to decide if two instances of the same face are actually the same individual. Stated differently, their generic nature makes them weak in effectively describe the semantics of the image content and thus to infer the appearance changes that are ascribable to different identities rather than external factors.

Deep learning approaches based on Convolutional Neural Networks instead draw their powerfulness by learning a representation directly from both the raw image data and the task. This is a fundamental peculiarity that can be illustrated with an example: assume to have two face images of two persons. One is smiling and the other not. It is likely in this context that what makes us decide that the two images represent two subjects is not the same visual content that makes us say that one is smiling and the other is not. More precisely, the visual content is actually the same but the information that we extract from it, it is probably not. Although this problem could still be resolved with local feature based approaches, their performance tend to drop when either *(i)* the external conditions induce very large appearance changes or *(ii)* the number of different instances to be classified increases a lot e.g. recognize thousands of different individuals. On the other hand CNNs have two main drawbacks: the first is that, for each task, a network must be trained to perform that task in order to extract a meaningful representation. Secondly, the amount of images needed to train effectively the network is massive, likewise the number of parameters to be learned.

For what concerns face recognition approaches based on local features, we mainly focused on the problem of pose variations i.e. the orientation of the head in the 3D space. In the context of face recognition "in the wild", among all, pose variations are surely the ones that can make a recognition system fail the most. This can be ascribed to the fact that engineered descriptors are conceived so as to be invariant to rather smooth illumination changes [59] and reasonable spatial transformations [25, 50]. While extreme cases occur rarely in real conditions, natural pose variations, in a non cooperative context, appear frequently. Moreover, these variations induce *(i)* a misalignment in the spatial location of the image content, *(ii)* self occlusions i.e. loss of information and *(iii)* a drastic change in the pixels pattern of the same object. For these reasons, since pose variations occur in 3D space, we argued that the use of 3D data could be convenient. To this aim, a statistical 3D

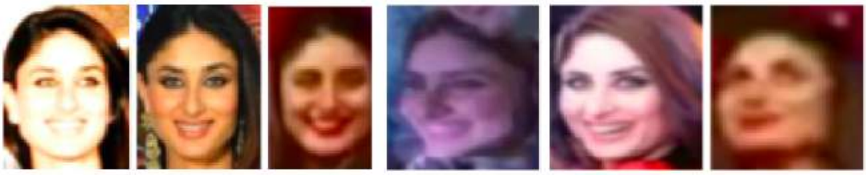


Figure 1.1: Images of a single person captured “in the wild”.

modeling technique has been developed. It grounds on a 3D Morphable Model [9] which is adapted to the face images and used to render a frontal facing view.

Referring to approaches based on CNNs instead, it has been extensively demonstrated that such networks are able to retain the discriminative information of an image even in challenging conditions and perform recognition accurately [75, 79, 81]. Despite this, not that much effort has been put in understanding how much the images themselves impact on the performance of a deep network. To mention an example, most of the works in the field apply a similarity transformation to the images so as to align them to a common reference before giving them as input to the network; is this actually needed or even useful? Are the networks able to account for visual information other than the face i.e. the background? We try to answer these questions through an extensive analysis of the data characteristics and preprocessing operations that can be applied on them.

1.2 Organization of the thesis

In chapter 2 the proposed statistical 3D modeling technique is presented. It is based on a modification on the classic 3D Morphable Model (3DMM); instead of applying PCA to model the 3D shapes variabilities, a dictionary learning (DL) technique is exploited. It will be shown that the proposed DL-3DMM solution can more accurately reconstruct the shape of a face given a 2D image. Chapter 3 presents a “Frontalization” technique based on the DL-3DMM described in the previous chapters. This new image representation is used instead of the original images to extract local descriptors. We experimented this solution in three different tasks, namely face recognition, emotion recognition and action units detection. The transition from local approaches based on hand crafted features to deep learning is presented in chapter 4, where we expound an extensive analysis of the effect that different image characteristics and preprocessing operations have on the performance of a CNN. Finally conclusions and future direction of research are reported in chapter 5.

1.3 Contributions

The main contributions of this theses are reported in the following:

Chapter 2: In this chapter a new statistical 3D modeling technique is presented. First, we describe a method to establish a dense correspondence between scans even in the case of expressions that include topological variations such as open/closed mouth. This allows us to include expressive scans in the training set, enlarging the generative capabilities of the 3DMM. Such training data is used to build a new approach to capture the statistical variability that, instead of exploiting standard PCA, learns a dictionary of deformations from the deviations between each 3D scan and a generic model computed on the vertex positions of the densely aligned training scans. We refer to this new model composed by the average model and the learned basis of deviations as DL-3DMM. Finally, the DL-3DMM is used to build an efficient fitting method that only relies on the correspondence between 2D and 3D landmarks of the face, which avoids a costly iterative optimization by estimating the model parameters through a closed form solution.

Chapter 3: In this chapter an effective face frontalization approach is presented. In particular, we can show that performing a frontal rendering of an unconstrained face image using the proposed technique and a properly constructed 3DMM, capable of effectively adapting to faces with varying expression, ethnicity and gender, achieves results in line with the state of the art even using baseline descriptors and learning solutions. The proposed face representation is experimented in three different tasks, namely face recognition, emotion recognition ad action units detection.

Chapter 4: The main contributions and outcomes of this chapter are: *(i)* a thorough experimentation on face data in the "wild" that evaluates the effect on recognition results of bounding box dimensions, alignment, positioning and data source; *(ii)* the evidence that deep architectures do not benefit from preprocessing operations that normalize input data both at train and test time and *(iii)* the proof that different distance measures lead to very diverse results and can be used as indicators of the effectiveness of the face representation learned by a CNN.

Datasets: Frontalized faces in the wild The technique presented in chapter 3 has been used to frontalize the face images contained in the dataset Labeled Faces in The Wild [38].

Face images collector software: Finally, a web application that collects huge amounts of face images provided a list of names has been developed. It gets a list of names as input and queries three different search engines. The urls returned by the engines are collected and the images are downloaded. It performs a filtering of possibly wrong images for each subject by extracting CNN descriptors from a subset of the downloaded images and training a SVM classifier. It grounds on the hypothesis that the number of wrong images is less than the correct. The trained classifier is used to discard images that are classified as not belonging to the considered identity. A web tool allows users to check and refine the set of collected images.

Chapter 2

Dictionary Learning Based 3D Morphable Shape Model

In this chapter we present a new approach for constructing a 3D Morphable Shape Model (called DL-3DMM) and show our solution can reach the accuracy of deformation required in applications where fine details of the face are concerned. For constructing the model, we start from a set of 3D face scans with large variability in terms of ethnicity and expressions. Across these training scans, we compute a point-to-point dense alignment, which is accurate also in the presence of topological variations of the face. The DL-3DMM is constructed by learning a dictionary of basis components on the aligned scans. The model is then fit to 2D target faces using an efficient regularized ridge-regression guided by 2D/3D facial landmark correspondences. Comparison between the DL-3DMM and the standard PCA-based 3DMM demonstrates that in general a lower reconstruction error can be obtained with our solution.

2.1 Introduction and related work

In recent years, the analysis of human faces has become increasingly relevant, with a variety of potential computer vision and multimedia applications. Examples include human identification based on face [24,48,90], emotional state detection [87, 101], enhanced human-computer interaction using facial pose and expression [71,82,86,100,103], facial expression detection for medical assistance or investigation [18,65], prediction of drivers cognitive load [66,99], just to cite some of the most studied. All these applications share the difficulty of dealing with problems such as variations in pose and expression, illumination and resolution of the face, which are mostly related to 2D data. In this framework, a potentially interesting idea is that of using hybrid 2D/3D solutions in the attempt of overcoming to some extent the afore mentioned limitations. To this end, being the acquisition of high-quality 3D data expensive and difficult, learning a generic 3D face model capable of generating new face instances with plausible shape and appearance can be convenient. This can be done by capturing the face variability in a training set of 3D scans and constructing a statistical face model that includes an average component and a set of learned principal components of deformation. Such a model can be derived with a reasonable amount of 3D data and would allow either to generate new face instances, or deform and fit to 2D or 3D target faces.

In their seminal work, Blanz and Vetter [9] first proposed to create a 3D morphable model (3DMM) from a set of exemplar 3D faces and showed its potential and versatility. They showed how to derive a 3DMM by transforming the shape and texture from a training set of 3D face scans into a vector space representation based on PCA. A gradient-based optical flow algorithm was used to establish dense correspondence between pairs of 3D scans taking into account for texture and shape values simultaneously. A reference scan was then used to transfer correspondences across scans. However, the training dataset had limited face variability (200 neutral scans of young Caucasian individuals were included), thus reducing the capability of the model to generalize to different ethnicity and non-neutral expressions. Despite these limitations, the 3DMM has proved its effectiveness in image face analysis, also inspiring most of the subsequent work, with applications to computer graphics for face inverse lighting [76,102] and reanimation [8], craniofacial surgery [78], 3D shape estimation from 2D image face data [106], 3D face recognition [2], pose robust face recognition [10,55], etc.

The 3DMM was further refined into the Basel Face Model by Paysan et al. [62]. This offered higher shape and texture accuracy thanks to a better scanning device, and a lower number of correspondence artifacts using an improved registration algorithm based on the non-rigid iterative closest point (ICP) [3]. However, since non-rigid ICP cannot handle large missing regions and topological variations, expressions were not accounted for in the training data also in this case. In addition, both the optical flow used in [9] and the non-rigid ICP method used in [2, 62] were applied by transferring the vertex index from a reference model to all the scans. As a consequence, the choice of the reference face can affect the quality of the detected correspondences, and ultimately the final 3DMM. The work by Booth et al. [11], introduced a pipeline for 3DMM construction. Initially, dense correspondence was estimated applying the non-rigid ICP to a template model. Then, the so called LSFM-3DMM was constructed using PCA to derive the deformation basis on a dataset of 9,663 scans with a wide variety of age, gender, and ethnicity. Though the LSFM-3DMM was built from the largest dataset compared to the current state-of-the-art, the face shapes still were in neutral expression.

Following a different approach, Patel and Smith [61] showed that Thin-Plate Splines (TPS) and Procrustes analysis can be used to construct a 3DMM. Procrustes analysis was used to establish correspondence between a set of 104 manually labeled landmarks of the face, and the mean coordinates of these landmarks were used as anchor points. A complete deformable model was then constructed by warping the landmarks of each sample to the anchor points and interpolating the regions between landmarks using TPS. Finally, consistent resampling was performed across all faces, but using the estimated surface between landmarks rather than the real one. In [20], Cosker et al. described a framework for building a dynamic 3DMM, which extended static 3DMM construction by incorporating dynamic data. This was obtained by proposing an approach based on Active Appearance Model and TPS for non-rigid 3D mesh registration and correspondence. Results showed this method overcomes optical flow based solutions that are prone to temporal drift. Brunton et al. [12], instead, proposed a statistical model for 3D human faces in varying expression. The approach decomposed the face using a wavelet transform, and learned many localized, decorrelated multilinear models on the resulting coefficients. In [53], Lüthi et al. presented a Gaussian Process Morphable Model (GPMM), which generalizes PCA-based Statistical Shape Models (SSM). GPMM was defined by a Gaussian process, which makes it inherently continuous. Further, it can be specified using arbitrary positive definite kernels, which makes it possible to build shape priors, even in the case where many examples to learn an SSM are not available.

In this chapter, a new approach to the construction of a 3D Morphable

Shape Model is expounded (note that, even though we consider the sole shape component, throughout the chapter we will use the term 3DMM for our solution). The proposed model is capable of capturing much of the large variability of human faces, and it is grounded in three distinct contributions:

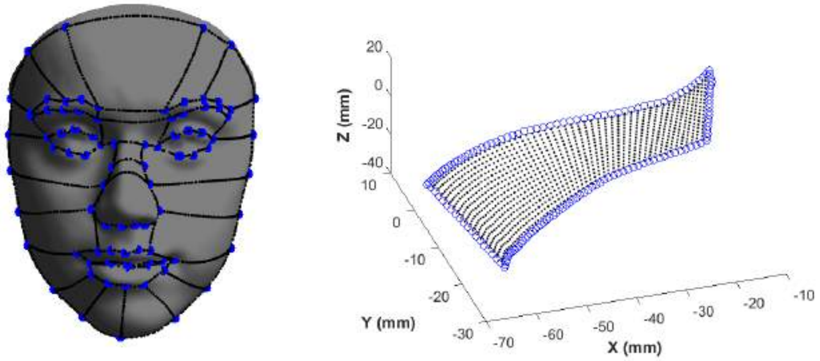
1. a new method to establish a dense correspondence between scans even in the case of expressions that include topological variations such as open/closed mouth. This allows us to include expressive scans in the training set, enlarging the generative capabilities of the 3DMM.
2. a new approach to capturing the statistical variability in training data that, instead of exploiting standard PCA, learns a dictionary of deformations from the deviations between each 3D scan and the average model computed on the vertex positions of the densely aligned training scans. We refer to this new model composed by the average model component and the learned basis of deviations as DL-3DMM;
3. an efficient fitting method that only relies on the correspondence between 2D and 3D landmarks of the face, and avoids a costly iterative optimization by estimating the model parameters through a closed form solution.

In the experiments, we demonstrate the DL-3DMM compares favorably with respect to the standard PCA-based 3DMM in terms of reconstruction error.

The rest of the chapter is organized as follows: in Sect. 2.2, we present the method for determining dense correspondence between the 3D scans of a training set with a large spectrum of face variations; the DL-3DMM construction using dictionary learning is proposed in Sect. 2.3; in Sect. 2.4, we present the 3DMM fitting method; in Sect. 4.4, we compare the DL-3DMM to the PCA-3DMM, and present their results in terms of reconstruction error; finally, discussion and conclusions are reported in Sect. 2.6.

2.2 Finding 3D Dense Correspondence

Given a training set, finding a dense point-to-point correspondence between the vertices of 3D scans can be seen as a sort of mesh re-parametrization where corresponding points must have the same anatomical reference. The limited number of facial points detectable with sufficient accuracy, and the presence of large regions with strong photometric variations, self-occlusions, facial expressions and changes in the topology of the face surface (as in the case of mouth-closed / mouth-open), make this problem highly complex.



(a) (b)

Figure 2.1: (a) A face scan of the BU-3DFE with the 87 landmarks (in blue), and the geodesic paths used to connect some of them (in black). These paths partition the face into a set of non-overlapping regions. (b) Geodesic contour of the *cheek / zygoma* region on the right side of the face. The geodesic contour is resampled so that points on it (circles in the plot) are at the same geodesic distance from each other. The interior of the region is also resampled using linear paths on the surface (dots in the plot), which connect corresponding points on opposite sides of the contour.

In our approach, similarly to Patel and Smith [61], we initially rely on a set of landmarks to establish a correspondence between salient points of the face (see Fig. 2.1(a)). However, differently from [61], where warping and TPS interpolation is applied between the average landmarks, we interpolate and sample the scan surface, region-by-region, while maintaining a dense correspondence. We first partition the face into a set of regions using geodesic paths between facial landmarks, applying the variant of the Fast Marching algorithm on triangular mesh manifolds of [43], and resample the geodesics with a predefined number of points posed at equal geodesic distance. As an example, Fig. 2.1(b) shows (with circles), the sampled points of the geodesic contour delimiting the cheek/zygoma region comprised between the nose and the face boundary on the right. Hence, we sample the surface of the face regions so that points of homologous regions are in dense correspondence across all the training scans. This is obtained by using the geodesic contour of the region to guide the dense resampling of its interior surface. The idea here is to connect pairs of sampling points on opposite side of a geodesic contour with a linear path on the surface [51]. This line is then sampled at the desired resolution, as illustrated in Fig. 2.1(b). Being based on the annotated landmarks and their connections, this approach proved to be robust to facial expressions. In particular, the presence of landmarks which delimit the internal and external border of the lips, makes it possible to maintain such region correspondence also across faces with mouth-closed/mouth-open expressions. While the method of [61] is only able to estimate the real surface, in our case, we are able to interpolate and sample the true surface of the face scans, region-by-region, maintaining a dense correspondence and do not require an average model as in [61]. With respect to the solutions in [3, 9, 62] our approach does not require a reference face model, that could request a new face parametrization. It only requires that training faces are labeled with a set of landmarks, that is easily obtained with good accuracy using available detectors both in 2D [42, 93] and 3D [63].

Learning a 3DMM requires a training set of 3D face scans with high variability in terms of gender, age and ethnicity. Since we aim to generalize to expressive data, including scans with facial expressions is also important. To this end, we used the publicly available Binghamton University 3D Facial Expression dataset (BU-3DFE) [97] as training set. This dataset includes a balanced sample of human face variability and facial expressions and has been largely employed for 3D expression/face recognition. In particular, the BU-3DFE contains scans of 44 females and 56 males, with age ranging from 18 to 70 years old, acquired in a neutral plus six different expressions: anger, disgust, fear, happiness, sadness, and surprise. Apart from neutral, all the other facial expressions were acquired at four levels of intensity, from low to exaggerated (2500 scans in total). The subjects are distributed across different ethnic groups or racial ancestries, including *White*, *Black*, *Indian*, *East-Asian*, *Middle East Asian*, and *Hispanic-Latino*. The 83 facial landmarks annotated and released with the BU-3DFE provide correspondence across the training faces for a limited set of anchor points in correspondence to the distinguishing traits of the face.

Four additional landmarks located in the forehead have been derived from this initial set using anthropometric considerations on face proportions [26]. The overall set of 87 landmarks is shown with blue spheres on the face scan in Fig. 2.1(a). It is evident that these landmarks delimit salient parts of the face: the eyebrows, the eyes, the upper and lower lips, the nose, and the face boundary. By connecting selected pairs of landmarks through geodesic paths, we identified 12 regions in each side of the face (comprising the superorbitali, eyebrow, eye, cheek, jaw and chin), plus 9 regions covering the middle part of the face (including the lips, the region between the upper lip and the nose, the nose, the region between the eyes, and the forehead). As a result, each face was partitioned into 33 regions, each delimited by a closed geodesic contour passing through a set of landmarks, as shown in Fig. 2.1(a).

2.3 DL-3DMM Construction

Once a dense correspondence is established across the training data, we build our DL-3DMM by learning a dictionary of deformation components exploiting the *Online Dictionary Learning for Sparse Coding* technique [54]. Learning is performed in an unsupervised way, without exploiting any knowledge about the data (e.g., identity or expression labels).

Let N be the set of training scans, as obtained in Sect. 2.2, each with m vertices. Each scan is represented as a column vector $\mathbf{f}_i \in \mathbb{R}^{3m}$, whose elements are the linearized X, Y, Z coordinates of all the vertices, that is:

$$\mathbf{f}_i = [\ X_{i,1}\ Y_{i,1}\ Z_{i,1}\ \ldots\ X_{i,m}\ Y_{i,m}\ Z_{i,m}\ ]^T \in \mathbb{R}^{3m}\ .$$

The average model $\mathbf{m}$ of the training scans is computed as:

$$\mathbf{m} = \frac{1}{|N|}\sum_{i=1}^{|N|}\mathbf{f}_i\ . \tag{2.1}$$

Then, for each training scan $\mathbf{f}_i$, we compute the field of deviations $\mathbf{v}_i$ with respect to the average model $\mathbf{m}$:

$$\mathbf{v}_i \leftarrow \mathbf{f}_i - \mathbf{m}\ , \quad \forall\ \mathbf{f}_i \in N\ . \tag{2.2}$$

In the classic 3DMM framework [9], new 3D shapes are generated by deforming the average model $\mathbf{m}$ with a linear combination of the principal components. In this work, instead, we propose to learn a set of deformation components through dictionary learning. In particular, the dictionary atoms are learnt from the field of deviations $\mathbf{v}_i$. Then, we morph the average model exploiting a linear combination of the dictionary atoms. Note that the PCA model is also constructed on the training set $\mathbf{v}_i$.

Dictionary learning is usually cast as an ℓ_1-regularized least squares problem [54]. However, since the learnt directions are used to deform the average model, the sparsity induced by the ℓ_1 penalty can lead to a noisy or, in the worst case, a discontinuous or punctured model. We thus decided to formulate the dictionary learning as an *Elastic-Net* regression. The Elastic-Net is a type of regression method that linearly combines the sparsity-inducing ℓ_1 penalty and the ℓ_2 regularization. The ℓ_1 norm is known to act as a shrinkage operator, reducing the number of non-zero elements of the dictionary, while the ℓ_2 norm avoids uncontrolled growth of the elements magnitude, while forcing smoothness. By defining $\ell_{1,2}(\mathbf{w}_i) = \lambda_1 \left\|\mathbf{w}_i\right\|_1 + \lambda_2 \left\|\mathbf{w}_i\right\|_2$, where λ_1 and λ_2 are, respectively, the sparsity and regularization parameters, we can formulate the problem as:

$$\min_{\mathbf{w}_i,\, \mathbf{D}} \frac{1}{|N|} \sum_{i=1}^{|N|} \Big(\left\|\mathbf{v}_i - \mathbf{D}\mathbf{w}_i\right\|_2^2 + \ell_{1,2}(\mathbf{w}_i) \Big) \ , \tag{2.3}$$

where the columns of the dictionary $\mathbf{D} \in \mathbb{R}^{3m \times k}$ are the basis components, $\mathbf{w}_i \in \mathbb{R}^k$ are the coefficients of the dictionary learning, and k is the number of basis components of the dictionary. The number of components (dictionary atoms) must be defined a priori. Instead, the set of coefficients $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{k \times k}$ is obtained as the cumulated sum of the coefficients at each iteration of the dictionary learning. The coefficients of the matrix $\mathbf{W}$ are in general concentrated on the diagonal [54], and represent the contribution of the k-th basis element in reconstructing the training vectors.

The above minimization can be rewritten as a joint optimization problem with respect to the dictionary $\mathbf{D}$ and the coefficients $\mathbf{W}$, and solved by alternating between the two variables, minimizing over one while keeping the other one fixed [54]. The average model $\mathbf{m}$, the dictionary $\mathbf{D}$ and the diagonal elements of the matrix $\mathbf{W}$, namely the vector $\hat{\mathbf{w}} \in \mathbb{R}^k$, constitute our *Dictionary Learning based* 3DMM (DL-3DMM).

2.4 Efficiently Fitting the DL-3DMM

Fitting a 3DMM to a 2D face image allows a coarse 3D reconstruction of the face. To this end, estimating the 3D pose of the face, and the correspondence between 3D and 2D landmarks are prerequisites. In the following, both the average model and the basis components of the learned dictionary will be represented in $\mathbb{R}^{3 \times m}$, rather than in $\mathbb{R}^{3m}$, and we refer to them as $\hat{\mathbf{m}}$ and

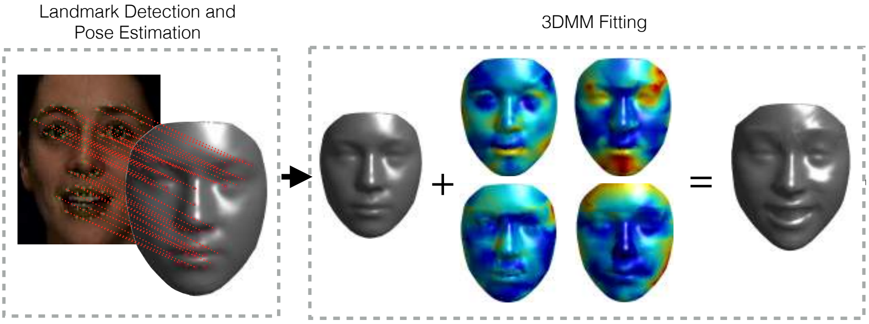


Figure 2.2: The proposed 3DMM fitting and frontal face rendering: (left) the 3D head pose is estimated from the correspondence of 2D and 3D landmarks; (right) the average 3D model is deformed using the basis components.

$\hat{\mathbf{D}}$, respectively. In order to estimate the pose, we detect a set of 49 facial landmarks $\mathbf{l} \in \mathbb{R}^{2\times 49}$ on the 2D face image using the technique proposed in [42] (see Fig. 2.2 on the left). An equivalent set of vertices $\mathbf{L} = \hat{\mathbf{m}}(\mathbf{I}_v) \in \mathbb{R}^{3\times 49}$ is manually annotated on the average 3D model, where $\mathbf{I}_v$ is the set of indices of the vertices corresponding to the landmark locations. Under an affine camera model [55], the relation between $\mathbf{L}$ and $\mathbf{l}$ is:

$$\mathbf{l} = \mathbf{A} \cdot \mathbf{L} + \mathbf{T} \ , \tag{2.4}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 3}$ contains the affine camera parameters, and $\mathbf{T} \in \mathbb{R}^{2\times 49}$ is the translation on the image. To recover these parameters, firstly, we subtract the mean from each set of points and recover the affine matrix $\mathbf{A}$ solving the following least squares problem:

$$\arg\min_{\mathbf{A}} \|\mathbf{l} - \mathbf{A} \cdot \mathbf{L}\|_2^2 \ , \tag{2.5}$$

for which the solution is given by $\mathbf{A} = \mathbf{l} \cdot \mathbf{L}^+$, where $\mathbf{L}^+$ is the pseudo-inverse matrix of $\mathbf{L}$. We can estimate the affine matrix with a direct least squares solution since, by construction, facial landmark detectors assume a consistent structure of the 3D face parts so they do not permit outliers or unreasonable arrangement of the face parts (e.g., nose landmarks cannot stay above the eyes). Finally, the 2D translation can be estimated as $\mathbf{T} = \mathbf{l} - \mathbf{A} \cdot \mathbf{L}$. Thus, the estimated pose $\mathbf{P}$ is represented as $[\mathbf{A}, \mathbf{T}]$ and used to map each vertex of the 3DMM onto the image.

Using the learned dictionary $\hat{\mathbf{D}} = [\hat{\mathbf{d}}_1, \ldots, \hat{\mathbf{d}}_k]$, we find the coding that non-rigidly transforms the average model $\hat{\mathbf{m}}$ such that the projection minimizes the error in correspondence to the landmarks. The coding is formulated as the solution of a regularized *Ridge-Regression* problem:

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k} \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \circ \hat{\mathbf{w}}^{-1} \right\|_2 , \tag{2.6}$$

where $\circ$ is the Hadamard product. Since the pose $\mathbf{P}$, the basis components $\hat{\mathbf{d}}_i$, the landmarks $\mathbf{l}$, and $\hat{\mathbf{m}}(\mathbf{I}_v)$ are known, we can define $\hat{\mathbf{X}} = \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v)$ and $\hat{\mathbf{y}}_i = \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)$. By considering their linearized versions[1] $\mathbf{X} \in \mathbb{R}^{98}$ and $\mathbf{y}_i \in \mathbb{R}^{98}$ with $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$, we can finally estimate the non-rigid coefficients which minimize the cost of Eq. 2.6, in closed form as follows:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda \cdot \mathrm{diag}(\hat{\mathbf{w}}^{-1}) \right)^{-1} \mathbf{Y}^T\mathbf{X} , \tag{2.7}$$

where $\mathrm{diag}(\hat{\mathbf{w}}^{-1})$ denotes the diagonal matrix with vector $\hat{\mathbf{w}}^{-1}$ on its diagonal. The term $\hat{\mathbf{w}}^{-1}$ is used to associate a reduced cost to the deformation induced by the most relevant components. Indeed, weighting the deformation parameters $\boldsymbol{\alpha}$ with the inverse of the coefficients $\hat{\mathbf{w}}$, reduces the cost of the deformation induced by components $\hat{\mathbf{d}}_i$ with a large coefficient $\hat{\mathbf{w}}_i$, while the contribution of unstable and noisy components is bounded. In the classic PCA model, the same principle applies, but in this case the deformation components $\hat{\mathbf{d}}_i$ are represented by the PC, while the vector $\hat{\mathbf{w}}$ corresponds to the eigenvalues associated to the PC.

Figure 2.3 shows the dictionary coefficients compared to the well known behavior of the PCA eigenvalues. The DL-3DMM coefficients contain the energies used by the dictionary atoms to reconstruct the training signals; though all the atoms contribute to the reconstruction, the actual contribution of an atom is quantified by the related coefficient. In this sense, the weighting $\hat{\mathbf{w}}^{-1}$ privileges the more contributing atoms.

Note that the pose estimation and fitting steps are alternated; we experimentally found that cleaner reconstructions can be obtained by repeating the process while keeping a high λ. This is motivated by the fact that the initial 3D and 2D landmark layouts are likely to be very different due to the presence of expressions, and the pose can be coarsely estimated. In this scenario, the non-rigid deformation which fits the landmark locations is likely to

[1] The dimension 98 results from the concatenation of the coordinates of the 49 landmarks.

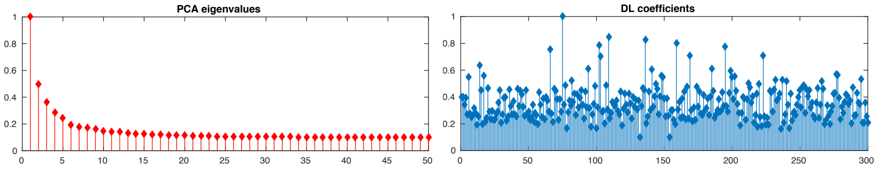


Figure 2.3: Comparison between the DL coefficients and the PCA eigenvalues.

excessively deform the model in the attempt of compensating also the error introduced by the pose. On the contrary, a high λ avoids to some extent this behavior and permits refinement of both the pose and the non-rigid deformation in the next step. Thus, a balance is required between the number of steps and the value of λ. We empirically found that the best configuration is repeating the process 2 times, with λ ranging from 0.0001 to 0.05. More than 2 repetitions do not produce appreciable improvement in the fitting.

A fitting example obtained using this solution is shown in Fig. 2.2. As a result, the 3D model is deformed according to the target face image.

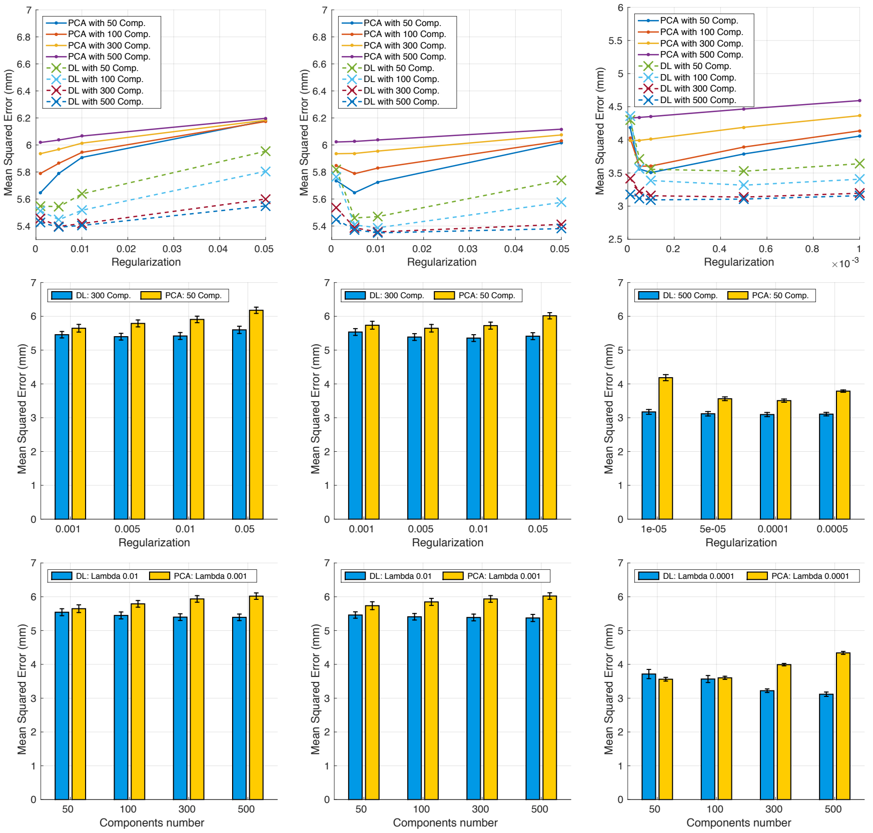
2.5 Experimental Results

The proposed DL-3DMM has been evaluated in two sets of experiments. First, we investigate the modeling ability of the DL-3DMM compared with its PCA-based counterpart in terms of 3D to 2D fitting, and direct 3D to 3D fitting on the BU-3DFE. Then, we evaluate a cross-dataset fitting between the BU-3DFE and the Face Recognition Grand Challenge (FRGC v2.0) [64] dataset, by training on one dataset and testing on the other one, and vice versa. In both these experiments, two reference projection matrices are defined: $\mathbf{P}^{f}_{ref}$ simulates a subject facing the camera (*front view*); $\mathbf{P}^{s}_{ref}$ has been taken simulating a pose with approximately 45° in yaw (*side view*). The 3DMM is fit following the approach of Sect. 2.4. For the direct 3D fitting, instead, we remove the projection $\mathbf{P}$ from Eq. (2.6) so as to perform the fitting directly in the original 3D space.

2.5.1 3D Shape Reconstruction

We comparatively evaluate how the DL-3DMM and PCA-3DMM fit to a set of test images. Experiments were performed on the BU-3DFE, processed as illustrated in Sect. 2.2 so that scans are densely aligned with the same number of vertices. To train and test the 3DMMs, we split the scans into two halves based on subject identity (so that train and test identities are completely separated): one half of the scans is used to construct the average model $\hat{\mathbf{m}}$, the deformation components $\hat{\mathbf{d}}_i$, and the weights $\hat{\mathbf{w}}$ for both the DL-3DMM and the PCA-3DMM; the other half is used for test. This process is repeated 10 times on each train/test partition, and results are averaged across the trials. To perform the 3D to 2D fitting, for each test scan we select the set of landmarks through the indices $\mathbf{I}_v$ and project them onto the 2D plane. These landmarks are used as a surrogate for the landmarks detected on a face image and allow both avoiding inaccuracies induced by detection and a misleading source of error not directly ascribable to the fitting. Since the 2D landmarks are generated from the 3D scans, the original 3D data can be used as ground-truth of the model resulting from the fitting process. Based on this, we computed the 3D *reconstruction* error by accumulating the vertex-to-vertex Euclidean distance between the ground-truth scan and the deformed 3DMM. This measure exploits the knowledge of the exact correspondence between all the vertices of the 3D scans given by the dense alignment. Thus, the errors can be calculated by considering the distance between vertices with the same index in the meshes, without requiring any nearest vertex search. This is important, since in the presence of strong topological changes as determined by expressive scans, finding meaningful corresponding points for computing the errors is a complex task.

Reconstruction errors for three fitting conditions, namely, 3D-2D *front view*, 3D-2D *side view*, and 3D-3D are reported in Fig. 2.4(a), (b) and (c), respectively. The plots in the first row of the Figure compare the results obtained with the DL-3DMM and the PCA-3DMM as a function of the regularization parameter λ of Eq. (2.6) and for different number of components. The bar graph in the middle row shows the effect of varying the regularization parameter λ when the number of components is fixed at its best performing number, while in the bottom row it is shown the opposite, i.e., the effect of varying the number of components at the best regularization value. Results show that our DL-3DMM performs generally better than the PCA-3DMM. In particular, the two methods show a quite different behavior regarding the number of components used. For PCA-3DMM, we observe that increasing the number of components degrades the performance. This fact can be explained considering that 3D scans are noisy regardless of ac-



(a) 3D-2D Fitting - Front (b) 3D-2D Fitting - Side (c) 3D-3D fitting

Figure 2.4: Reconstruction error on the BU-3DFE dataset: (a) 3D-2D fitting with front view; (b) 3D-2D fitting with side view; and (c) direct 3D-3D fitting. Each plot in the first row reports the errors for both DL- and PCA-based 3DMM as a function of the regularization parameter λ and for different number of components. The second row reports, for the best number of components, the effect of varying λ, while in the third row the effect of varying the number of components for the best λ value is shown. Standard deviation is also reported for each bar.

quisition accuracy, and the alignment process can mitigate such nuisances only to some extent. Furthermore, it is likely that some PCs reflect less significant characteristics of the data. These facts eventually cause a drop

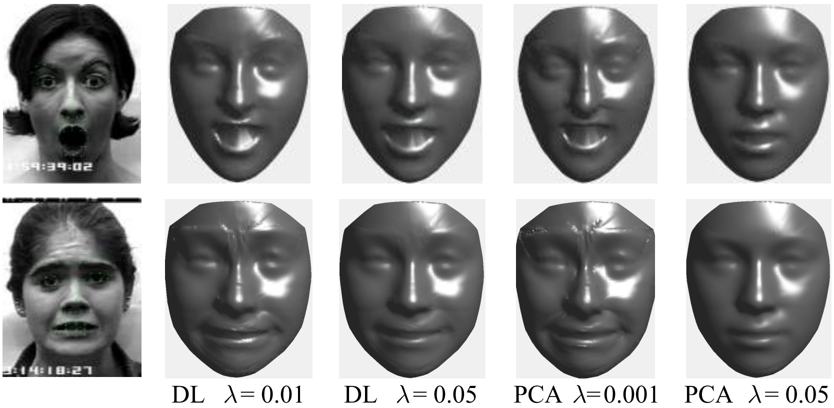


Figure 2.5: 3DMM fitting examples with both DL- and PCA-based 3DMM for optimal or high regularization values. It is appreciable how our DL-3DMM both introduces less noise in the 3D models and retains its modeling ability even for high regularization values (face images from the CK+ dataset [52]).

of fitting accuracy due to the introduction of noisy and ineffective components, although regularized by their eigenvalues. This behavior is consistent with the concept of *compactness* of a model (i.e., the ability of explaining the most and significant variability of the data with the fewest number of components). On the opposite, the DL-3DMM improves its modeling ability with a larger number of components. This behavior is related to the fact that larger dictionaries allow more combinations of the atoms thus covering a wider range of possible deformations.

Results show that an optimal value of λ is about 0.01 and 0.001 for the DL and PCA methods, respectively. We point out here that despite producing the minimum error, using low regularization values to fit the 3DMM can occasionally result in noisy models; it is desirable instead to generate a model which is as smooth as possible. It can be observed from Fig. 2.4 that the reconstruction error is more stable across increasing λ values for the DL-3DMM rather than for the PCA-3DMM. It is then possible to choose a larger regularization value to ensure a smooth model, without renouncing modeling precision. This behavior is accentuated for increasing number of DL components. Apart from the increased accuracy, since the fitting is

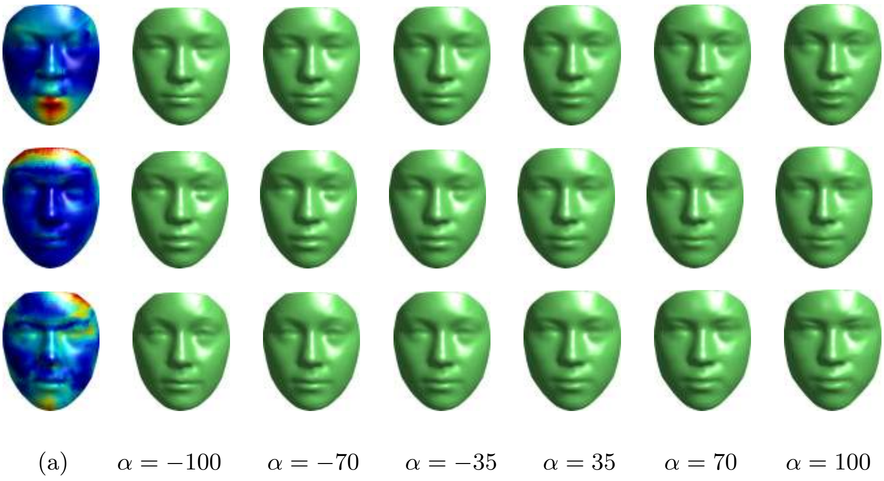


Figure 2.6: Example of the deformation obtained using single dictionary atoms. In column (a), the deformation heat-maps are reported; the models generated by applying different deformation magnitudes are shown in the other columns.

quickly performed in closed form, we also note that the computational time still is acceptable even for a large number of components. We experimentally found that 2 repetitions of the whole fitting process of Sect. 2.4 take $17, 31, 103$ and $185ms$ for, respectively, $50, 100, 300, 500$ components for both DL- and PCA-based 3DMM. We also found that after model deformation, the pose estimate is improved of about 0.5 degrees, with a final mean error of $5.0, 2.4, 4.1$ degrees, respectively, for pitch, yaw and roll angles. In Fig. 2.5 some examples of the 3DMM fitting, obtained using all the components, are shown. Both the DL-3DMM and the PCA-3DMM are able to model expressive faces but, nonetheless, our model has some advantages: 1) using the optimal λ value it introduces less noise in the resulting 3D model with respect to the PCA one; 2) if a smoother model is desired, the regularization value can be increased without sacrificing modeling ability. The PCA-3DMM, on the other hand, is not able to fit the expression properly in this case.

In Fig. 2.6 we show some examples of the deformation obtained using single dictionary atoms. Observe that DL components result in localized variations of the model, with a remarkable gap between different face parts.

Moreover, by varying the magnitude of the deformation applied to the average model it is possible to generate new meaningful models.

Additional experiments have been conducted in order to assess the robustness of the proposed model and fitting technique. First, we compare the modeling ability of the DL-3DMM and the PCA-3DMM when some landmarks are missing. Then, we show the importance of regularizing the fitting procedure by first removing the inverse weighting of the deformation parameters $\boldsymbol{\alpha}$, then by totally removing the regularization.

Missing data To evaluate the modeling ability in the case of missing landmarks, we report in Fig. 2.7 the reconstruction error obtained subsampling the landmarks with step 2 and 4 so as to perform the fitting with 24 (top row) and 12 (bottom row) landmarks, respectively. The experiment has been conducted both for faces in front view (left column) and side view (right column). We observe a predictable slight increase in the overall error (not higher than $0.2 - 0.3$mm), both for the DL-3DMM and PCA-3DMM in all the cases. Nonetheless, the general behavior of the error for different number of components and λ values is consistent with the case where all the landmarks are used. This suggests that both the models and the fitting procedure are rather robust to missing data.

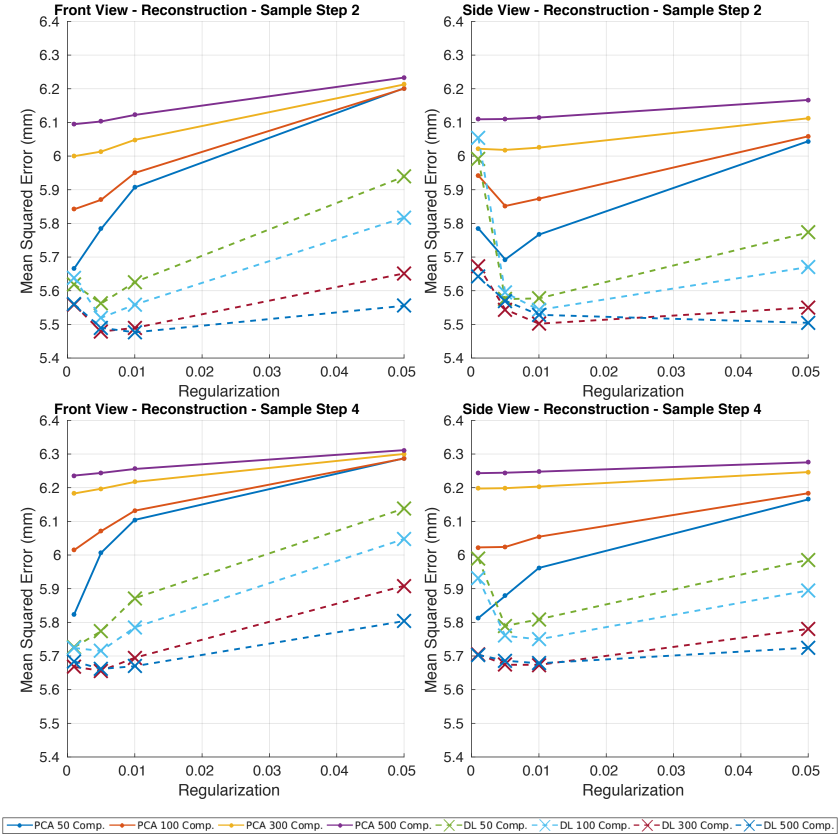
Removing the inverse weighting In our fitting procedure, the deformation parameters $\boldsymbol{\alpha}$ are weighted by the vector $\hat{\mathbf{w}}^{-1}$. Depending on the considered model, this vector contains the coefficients of the dictionary learning, for DL-3DMM, or the eigenvalues for PCA-3DMM. Without the weighting term, the minimization problem is rewritten as:

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k} \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_2 \ . \tag{2.8}$$

Considering the definitions of Sect.2.4 the solution then becomes:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T \mathbf{Y} + \lambda \cdot \mathcal{I} \right)^{-1} \mathbf{Y}^T \mathbf{X} \ , \tag{2.9}$$

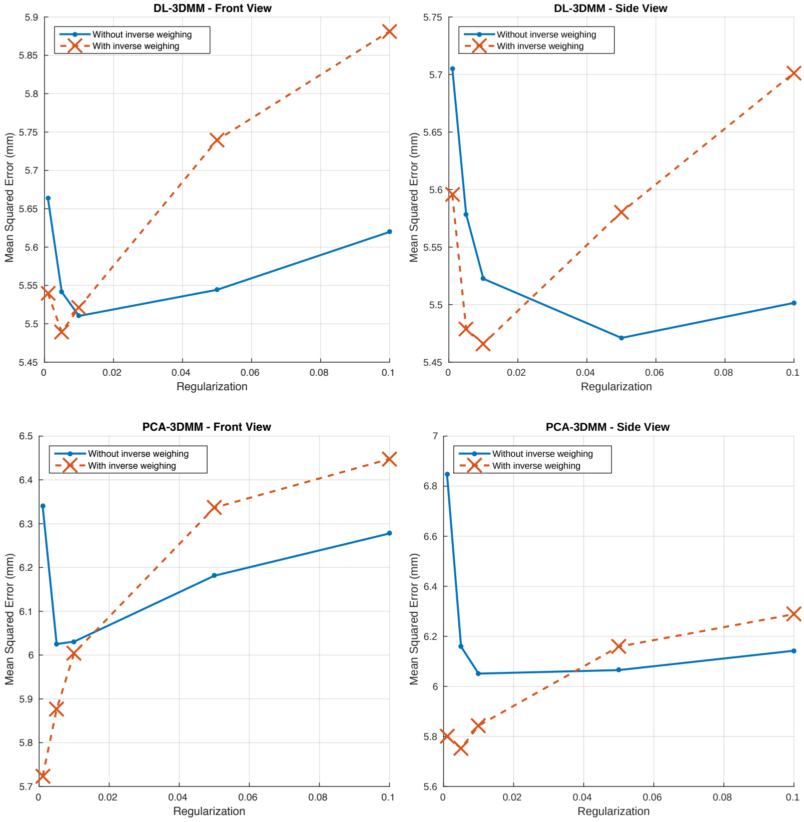
where $\mathcal{I}$ is the identity matrix. Experiments have been performed considering the best configuration both for the DL-3DMM and the PCA-3DMM, i.e., with 300 and 50 components, respectively. Results are reported in Fig. 2.8. While the impact of removing the weighting is not crucial for the DL-3DMM (even though the minimum error is obtained with the weighting), the results



(a) front view (b) side view

Figure 2.7: Each plot reports the reconstruction error for the DL- and PCA-based 3DMM as a function of the regularization parameter λ and the number of components. The 3DMM fitting is obtained using 24 landmarks (*top*) and 12 landmarks (*bottom*).

for the PCA-3DMM drop. This behavior does not surprise much: the eigenvalues of the PCA quantify the amount of variance retained by the related eigenvectors, and most of the total variance lies on the first few principal components. It is reasonable to suppose that some eigenvectors will, to some extent, represent the variance induced by the noise or less significant characteristics of the data. The term $\hat{\mathbf{w}}^{-1}$ then limits the contribution of



(a) front view (b) side view

Figure 2.8: Effect of removing the inverse weighting $\hat{\mathbf{w}}^{-1}$ from the 3DMM fitting: DL-3DMM (top); PCA-3DMM (bottom).

such vectors increasing their cost in the minimization, while reducing the cost associated to the most relevant ones. Similarly, in the DL-3DMM, we aim at giving the priority to components that contributed more in reconstructing the training set.

Removing the regularization To assess the importance of the regularization term in the fitting procedure, we carried out an experiment setting the λ value to 0, i.e., removing the regularization. Without the regularization

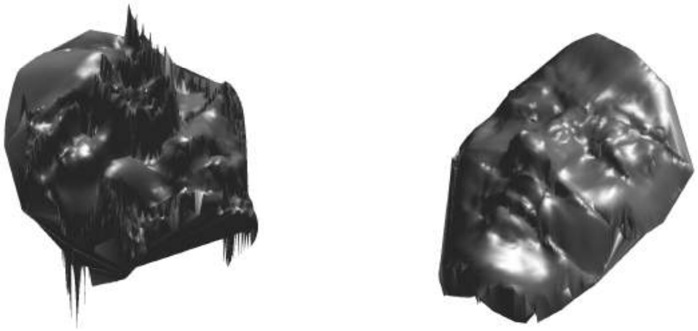


Figure 2.9: Examples of deformed models when the regularization is removed. The uncontrolled growth of the deformation coefficients $\boldsymbol{\alpha}$ leads to an excessive deformation of the 3D model.

term on the deformation coefficients, the problem becomes:

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k} \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 . \tag{2.10}$$

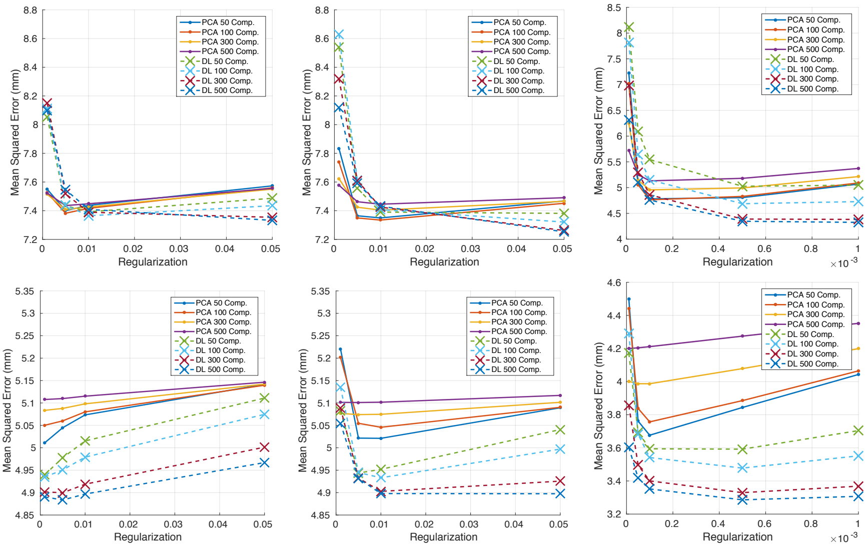
Considering again the definitions of Sect 2.4, we retrieve the coefficients as:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1} \mathbf{Y}^T\mathbf{X} . \tag{2.11}$$

As expected, the effect of removing the regularization is dramatic: the uncontrolled deformation coefficients lead to an excessive deformation of the model that barely resembles a human face, as shown in Fig. 2.9.

2.5.2 Cross-dataset 3D Shape Reconstruction

We performed a cross-dataset fitting experiment using the FRGC dataset in addition to the BU-3DFE. The FRGC v2.0 includes 4,007 scans of 466 subjects acquired with frontal view from the shoulder level, with very small pose variations. About 60% of the faces have neutral expression, while the others show spontaneous expressions of disgust, happiness, sadness, and surprise. Scans are given as matrices of 3D points of size 480 × 640, with a binary mask indicating the valid points of the face. 2D RGB images of the face are also available and aligned with the matrix of 3D points. Ground-truth landmarks are not available in this dataset. To apply our alignment procedure, we first run the landmark detector in [42] to extract 68 points from the 2D images



(a) 3D-2D Fitting - Front (b) 3D-2D Fitting - Side (c) 3D-3D fitting

Figure 2.10: Cross-dataset reconstruction errors obtained using FRGC for train and BU-3DFE for test (*top*) or vice versa (*bottom*). (a) 3D-2D fitting with frontal camera; (b) 3D-2D fitting with side camera; and (c) direct 3D-3D fitting. Each plot reports the errors for both DL- and PCA-based 3DMM as a function of λ, and for different number of components.

(the detection failed on just 6 images). Since 2D images and matrices of 3D points are aligned pixel-wise, the 2D landmarks position, plus 6 landmarks in the forehead of the face, can be transferred to 3D scans. Then, the alignment process described in Sect 2.2 is applied. In order to have a meaningful alignment between the two datasets, the same partitioning described above has been applied to the BU-3DFE considering a subset of 68 out of the 83 landmarks available as ground truth and re-aligning the whole dataset. In this experiment, the whole FRGC dataset was used to construct the average model $\hat{\mathbf{m}}$, the deformation components $\hat{\mathbf{d}}_i$, and the weights $\hat{\mathbf{w}}$, while all the models of the BU-3DFE have been used for test. The same experiment was performed considering the BU-3DFE as train and the FRGC for test. Reconstruction errors obtained for both DL- and PCA-based 3DMM shape fitting are reported in Fig. 2.10. It is possible to appreciate that when the FRGC is used for train, the reconstruction error is higher for both DL- and

PCA-based 3DMM. A possible motivation for this is that, though the FRGC dataset contains about four times the number of identities of the BU-3DFE, it includes less intense expressions. Comparing the results of the DL- and PCA-based 3DMM, they are very close, even though DL obtains a slightly smaller error. On the other hand, when the BU-3DFE is used to learn the deformation components, the error decreases of about 2mm. We explain this with the fact that adding more heterogeneous expression variations in the training permits the model to have a larger spectrum of deformations that ultimately result in more accurate reconstructions.

2.6 Conclusions

This chapter presented a dictionary learning based method for constructing a 3DMM, and its effectiveness compared to traditional methods for 3DMM construction based on PCA was shown. The proposed solution has the advantage of permitting more localized variations of the 3DMM that can better adapt to expressive faces. This capability to account for fine face deformations also depends on the inclusion in the training data of faces with large expression variability. This required us to develop a new method to establish a dense, point-to-point, correspondence between training faces. An approach to effectively deforming the 3DMM has been also proposed, which includes pose estimation and regularized ridge-regression fitting. The comparative evaluation of the DL- with the PCA-based 3DMM shows a clear advantage of the DL based solution in terms of 3D reconstruction error.

A potential drawback of a 3DMM that includes expressive scans is the difficulty in discriminating between components modeling identity traits and components modeling facial movements. Further investigation would be useful to determine: *1)* if more accurate vertex correspondences can be found by using different landmark detectors that induce more uniform partitioning of faces (which would also improve visual appearance of our models); *2)* if an extended solution can be found that balances the tradeoff between the efficiency of fitting against greater precision; and *3)* if deviations beyond shape can be accounted for in an extended 3DMM (for example by applying DL also to the texture component of faces).

In the next chapter it is described a *frontalization* technique based on the DL-3DMM. This technique exploits the 3D information to compensate the out of plane rotation of the face in a 2D image and render a frontal view.

Chapter 3

Effective 3D Based Frontalization for Face Analysis

This chapter describes a new and effective frontalization algorithm for frontal rendering of unconstrained face images. Initially, the DL-3DMM presented in the previous chapter is fit to a 2D face image, then an interpolating function is used to map each pixel inside the face region to the 3D model's space. In this way we can render a frontal view without introducing artifacts in the final image thanks to the exact correspondence between each pixel and the 3D coordinates of the model. The alignment induced by the 3D model allows to extract image descriptors on the frontalized images in repeatable positions across different images, enhancing the matching consistency. The effectiveness of this solution is experimented in the three different tasks, namely face recognition, emotion/expression recognition and Action Units (AU) detection.

3.1 Introduction and related work

Face recognition and facial expression recognition represent the core of many biometric techniques. A clear advantage of approaches based on face analysis with respect to other biometric signatures resides in its non-intrusiveness, that allows deployment also in unconstrained scenarios, without user cooperation. This latter capability is one of the main reasons for the increasing demand for surveillance systems that can operate in real contexts, even under strong variations in the face pose, expression, illumination, resolution, etc. In a broad sense, face recognition performs *coarse* grained (*inter-class*) face analysis, where face variations that separate different identities are accounted for. Conversely, in applications that recognize facial expressions or Action Units (AU), *fine* grained (*intra-class*) face analysis is required, where subtle and local variations of the face occur under the action of groups or individual facial muscles. In both the cases, 2D data manifest evident limitations and performing face analysis in 3D can be convenient [16, 73, 105]. Nevertheless, in real scenarios the availability of 3D data cannot be granted. For these reasons, using 2D/3D solutions becomes a suitable alternative.

In the last few years, an impressive development has been registered in this research area, with results which have substantially closed the gap with the human-level performance, also thanks to the introduction of the *Deep Neural Net* (DNN) architectures and learning methods [80]. Despite these recent advancements, there are some aspects in the conventional face recognition pipeline (including *detection*, *alignment*, *representation*, *classification*) that require further investigation. In particular, the alignment step is of fundamental importance for the subsequent stages, as for many other face analysis applications [23]. The alignment involves, amongst other things, the compensation for *in-plane* and *out-of-plane* rotations of the head. In most of the cases, this also demands for precise detectors of face landmarks, which is, by itself, a difficult problem, particularly in the presence of face occlusions due to pose variations. In unconstrained face recognition, compensating out of plane rotations is one important issue. Since head rotations occur in the 3D space, pose normalization (also known as face *frontalization*) solutions require that some 3D information of the face is inferred. Methods that address this problem are usually classified as 2D or 3D. In general, effective results have been obtained with methods in both categories, but since pose variations occur in the 3D space, 3D methods are more promising in perspective [104].

2D methods usually cope with the lack of explicit depth information by relying on a training image database, which includes images with different pose (and thus different 3D views). Some 2D transformations (*e.g.*, piecewise affine, thin plate splines) are often used to approximate the 3D transformation, while the error is compensated by some statistical learning strategy. Following this general idea, Berg and Belhumeur [7] presented a face verification built upon a large and diverse collection of linear classifiers that distinguish between two people. Authors propose an identity-preserving alignment procedure based on the detection of 95 face parts that enforces a fairly strict correspondence across images. This alignment procedure uses a reference dataset to distinguish geometry differences due to pose and expression from those that pertain to identity. Ho and Chellappa [33] proposed a method for reconstructing the virtual frontal view of a non-frontal image by using Markov Random Field (MRF), and a variant of the belief propagation algorithm. In this approach, the input face image is divided into a grid of overlapping patches and a set of possible warps for each patch is obtained by aligning it with images from a training database of frontal faces. A statistical approach to face frontalization is also proposed by Sagonas et al. [72]. The key observation of this work is that, for the facial images lying in a linear space, the rank of a frontal facial image, due to the approximate structure of human face, is much smaller than the rank of facial images in other poses. Based on this, a unified method is proposed for joint face frontalization (pose correction), landmark localization, and pose-invariant face recognition using a small set of frontal images only.

3D methods are based on a 3D face model, either *deformable* or *not-deformable*, used to precisely estimate the 3D face. In one of the first examples, Blanz and Vetter [10] used their 3DMM to simulate the process of image formation in 3D space, and estimated 3D shape and texture of faces from single images for face recognition. Later, Romdhani and Vetter [70] used the 3DMM for face recognition by enhancing the deformation algorithm with the inclusion of various image features. To build a pose robust face recognition system, Yi et al. [95] used a 3DMM, but performing the transformation in the filter space. Differently from the other 3DMM based methods, this solution proposes a “Pose Adaptive Filter” method, which transforms the filters according to the pose and shape of face image retrieved by fitting a 3DMM to the face image, and then uses the pose adapted Gabor filters for feature extraction. Later on, Juefei-Xu et al. [41] proposed the *Spartans* framework, which uses a 3D Generic Elastic Model (3D-GEN) to generate virtual face images with various poses for the gallery, and then match the probe to the virtual face images. In particular, the 3D-GEN is used to derive the depth information from a single frontal image per subject of the training set. The high-dimensional Walsh LBP descriptor is uniformly sampled on periocular regions of facial images with robustness toward alignment. During the learning stage, subject-dependent correlation filters are learned for pose-tolerant non-linear subspace modeling in kernel feature space followed by a coupled max-pooling mechanism. Zhu et al. [107] presented a 3DMM based pose and expression normalization method to recover the canonical-view, expression-free image, preserving the face appearance with little artifact and information loss. Variants of 3D methods use a single, unmodified 3D reference model to estimate a rough approximation of the 3D face surface, and use this surface to generate the new views [30, 56, 80].

Recently, this idea has been followed by Hassner et al. [31]. First, a face is detected, cropped and rescaled to a standard coordinate system. Then, facial feature points are localized [93] in the query image, and used to align it to the feature points on a *reference* face photo.From the 2D coordinates on the query image and their corresponding 3D coordinates on the model, a projection matrix is estimated. An initial frontalized face is obtained by back-projecting the appearance (colors) of the query image to the reference coordinate system using the 3D surface as a proxy. A final result is produced by borrowing appearances from corresponding symmetric sides of the face wherever facial features are poorly visible due to the pose of the query.

In all these cases, the 3DMM was used mainly to compensate for the pose of the face, with some examples that performed also illumination normalization. Expressions were typically not considered. Indeed, the difficulty in making 3DMM work properly in fine face analysis applications is confirmed by the almost complete absence of methods that use 3DMM for expression recognition. Among the few examples, Ramanathan et al. [67] constructed a 3D Morphable Expression Model incorporating emotion-dependent face variations in terms of morphing parameters that were used for recognizing four emotions. Ujir and Spann [84] combined the 3DMM with Modular PCA and Facial Animation Parameters (FAP) for facial expression recognition, but the model deformation was due more to the action of FAP than to the learned components. In [21], Cosker et al. used a dynamic 3DMM [19] to explore the effect of linear and non-linear facial movement on expression recognition through a test where users evaluated animated frames. Huber et al. [40] proposed a cascaded-regressor based face tracking and a 3DMM shape fitting for fully automatic real-time semi dense 3D face reconstruction from monocular in-the-wild videos.

In this chapter an effective face frontalization approach is presented. In particular, we can show that performing a frontal rendering of an unconstrained face image using the proposed technique and a properly constructed 3DMM capable of effectively adapting to faces with varying expression, ethnicity and gender, achieves results in line with the state of the art, even using baseline descriptors and learning solutions.

The rest of the chapter is organized as follows: In Sect. 3.2, the proposed frontalization technique is presented. In Sect. 3.3 we describe the feature extraction and unsupervised learning process. A comparative evaluation of the proposed approach with respect to other frontalization methods in terms of face recognition is reported in Sect. 3.4, together with the application of the proposed frontalization to face recognition, expression recognition ad AU detection in comparison with state-of-the-art methods. Finally, discussion and future work are reported in Sect. 3.5.

3.2 3D based Face Frontalization

Our face frontalization grounds on two steps: *(i)* 3D head pose estimation and 3DMM fitting; *(ii)* estimation of the transformation used to back-project the image texture to the 3D model's space and render the frontal image. The pose estimation and model fitting have been described in the previous chapter. For more details see 2.4.

Once the 3D model is fitted and projected onto the image, a straightforward way to perform an image rendering consists in associating to each projected vertex of the model the RGB value of the pixel it falls onto, as in [55]. In this manner, we get a full correspondence between 3D vertices and RGB values. Even though we are now able to build a rendering at arbitrary poses, the original coordinate frame of the 3D model is constructed such that the model faces the z axis; thus, we can easily build a frontal view by just dropping the z value and construct the image by defining a dense regular grid and by putting the RGB values in correspondence of the (x, y) coordinates of the model on the grid. Points in the grid where no vertices fall are interpolated. This approach is easy, but the quality of the rendering is not optimal and many artifacts are introduced in the final image. This can happen for many reasons; for instance, depending on the 3D rotation, some vertices can fall on the same pixel once projected onto the image plane or, on the contrary, pixels can be missed resulting in additional interpolations. The same happens also depending on the scale factor induced by the image resolution; in low resolution images many vertices will be projected onto

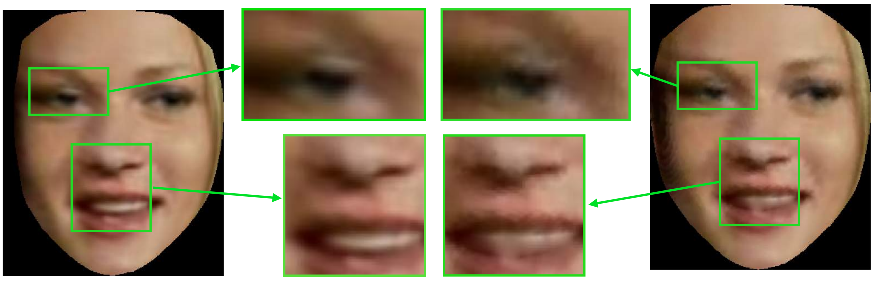


Figure 3.1: Difference between the proposed frontalization approach (left) and the one used in [55] (right). It can be appreciated how the rendering artifacts are removed. Note that the original image size is 250 × 250, but the face bounding box is approximately 90 × 90. Better seen in digital.

the same pixel, while the opposite happens in high resolution images. The proposed frontalization approach overcomes such issues by exploiting the prior knowledge of the face 3D shape. Basically, instead of interpolating the RGB values of pixels associated to any or multiple vertices, we interpolate the 3D position of each image coordinate inside the region $\mho$ defined by the convex hull of the projected 3D model. This can be practically done since for each vertex i in the 3D shape, we know the 2D position on the image, $P_i = (X_i, Y_i, Z_i) \mapsto (x_i, y_i) = p_i$. We can use these correspondences to fit a surface of the form $P = F(x, y)$, being P_i a vertex in the 3D model and p_i the corresponding projection on the image plane, *i.e.*, a pixel coordinate. We can then evaluate the surface values for each pixel (u, v) inside the face region $\mho$. In doing so, a new 3D shape can be built:

$$\forall (u, v) \in \mho,\ P_{(u,v)} = F(x_{(u,v)}, y_{(u,v)}) \ . \tag{3.1}$$

The resulting 3D model's vertices perfectly fall on each pixel of the image regardless the resolution or the viewpoint. We can now use the new 3D model to sample the RGB values and build the frontalized image in the same way as in [55], but in a clean and more accurate way.

An issue arising here is that out-of-plane rotations will eventually make some points to be self occluded; once projected onto the image, self-occluded points will have (approximately) the same (x, y) coordinates visible ones, but a much different z coordinate in 3D. The interpolating function will then estimate ambiguous values and fail. To overcome this problem, first an estimate

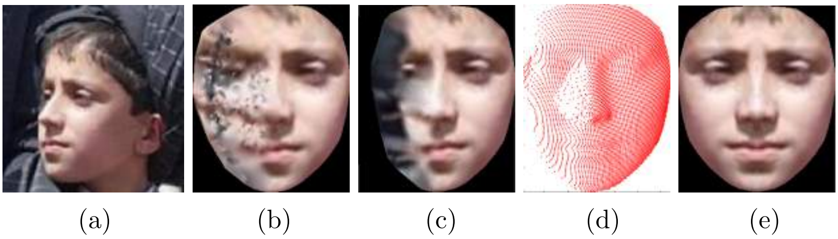


Figure 3.2: Frontalization result for heavy out-of-plane rotation (> 40 degrees). In (a) the original image. In (b), frontalization without self occlusion handling; many errors due to ambiguities in the 3D coordinates interpolation arise. Excluding the occluded points eliminates the errors and introduces black areas (c). In (d), the 3D model resulting from Eq. (3.1) is reported. Final result, obtained by symmetrizing the visible part is shown in (e).

of the visible 3D vertices given the 3D rotation is obtained, then the surface $F(x, y)$ is computed considering those visible points only. Evidently, the resulting 3D model and, accordingly, the rendered image will have some missing values, substituted by black pixels. Nevertheless, they correspond to points where the information is actually missing. These missing areas, generated by self-occlusions, can be filled with the symmetric visible part. An illustrative example is shown in Fig. 3.2. It is worth to notice in Fig. 3.2(d) how the reconstructed 3D model is not uniform, but shows a sort of parametrization imposed by the actual appearance of the particular image. In the example of Fig. 3.2, the face image undergoes a yaw rotation of $\approx -45°$; even if the density of the 3D model is higher in the visible part of the face, we can notice that it depends also on the orientation of the point's normals with respect to the image plane: the more the projected surface patches normals show orientation parallel to the image plane, the less dense the surface will be.

3.3 Face representation

Usual face analysis approaches perform the interested task e.g. face verification, expression recognition, by computing local image descriptors on the whole sub-image defined by the face bounding box, process them in some way, and feed the resulting early fusion to some classifier. Other than building an accurate frontal rendering, we exploit our frontalization method to precisely localize the image coordinates, where feature descriptors will be computed. As described in Sect. 3.2, the rendered image is built based upon the 3D model; thanks of this, we can easily back-project the 3D points in the frontal image. Such points define the coordinates in the image where descriptors are going to be extracted, as done in [55]. This has a two fold advantage: (1) We can choose where to extract our descriptors; it is well known that some facial areas are more discriminative than others; (2) Thanks to the 3DMM, we get an intrinsic alignment between single descriptors across the different images. The final face descriptors will always have the same length, regardless of the image size.

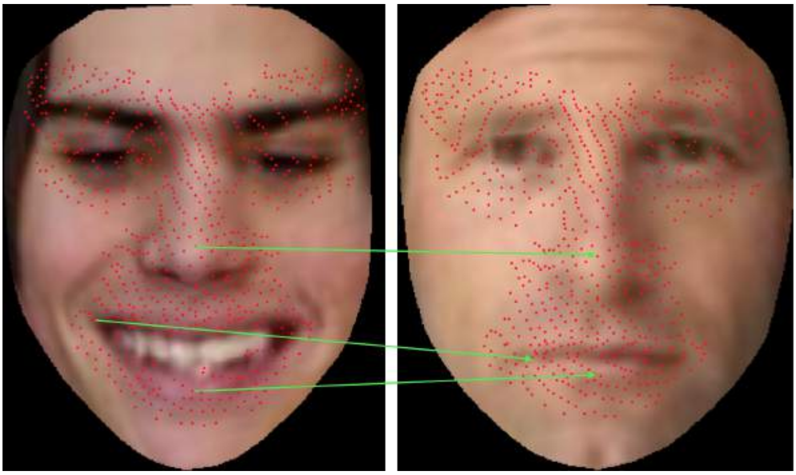


Figure 3.3: Localizing the feature descriptors. Our localization strategy permits us to more accurately match descriptors with the same semantic meaning in spite of the location on the image (green arrows).

Two different strategies were implemented depending on the task; for face recognition/verification the feature descriptors were computed selecting a subset of vertices of the 3D model located around the landmarks. We argued that for the specific task, where *inter-personam* variations are of interest, the most of the information was retained in such areas (namely, eyes, eyebrows, nose, mouth); moreover those are the parts less prone to be self occluded. The latter points have been chosen pretty densely so as to generate redundancy based on the result in [13], where it is demonstrated that face recognition benefits from high dimensional feature vectors. Differently, for expression recognition and AU detection, where learning *intra-personam* variations is crucial and the information from the whole face is significant, feature descriptors were computed on a subset of the vertices of the whole 3D model.

Since the proposed 3DMM is capable of moving the vertices to fit the face image, we get a more accurate alignment between descriptors. As shown in Fig. 3.3, this allows us to precisely match descriptors related to points with the same semantic meaning (see in particular the mouth area). For each task, we concatenate the descriptors extracted from a face so as to form a unique descriptor, and reduce its dimensionality by applying PCA with a number of PCs that retain at least the 95% of variance. The metric used for matching is the *cosine* distance.

3.4 Experimental Results

In this section, we report experiments on our approach. The proposed face representation has been evaluated in three different tasks, namely face recognition, emotion recognition and AU detection. For all the tasks a common pipeline has been defined, which follows a standard classification structure and consists of: 1) image alignment *i.e.* frontalization in our case; 2) feature extraction; 3) classifier training 4) matching. There will be subtle differences in the various pipelines depending on the task; details will be given in the related sections.

For face recognition, first we report the comparison with other face frontalization algorithms using the standard feature extraction approach, which is performed by computing image descriptors over a regular grid on the image. Then, the proposed face representation is compared to state-of-the-art methods on two benchmark datasets.

The evaluation on the tasks of emotion recognition and AU detection is conducted comparing the developed representation against other baseline feature extraction methods and the state-of-the-art on two benchmark datasets; details are given in Section 3.4.2.

3.4.1 Face Recognition

Tests have been performed on the Labeled Faces in the Wild (LFW) benchmark [38], and on the IARPA Janus Benchmark A (IJB-A) dataset [45]. The

LFW dataset represents a challenging benchmark for face verification algorithms including about 13000 face images of 5749 subjects taken under spontaneous conditions, with variabilities in terms of expressions, occlusions and partial pose variations. The recent IJB-A dataset pushes these challenges to the limit including images and videos taken under extreme conditions of illumination, resolution and including full pose variations (i.e., full profiles).

For the LFW dataset, we designed our solution following the *View*-1 protocol defined in [38] and used the *View*-2 protocol to produce our final results. *View*-2 provides 10 sets of 600 image pairs, each set including 300 pairs of the same subject and 300 pairs of different subjects. Ten-fold cross validation is used. We followed the "Unsupervised" protocol and report the results in terms of *Area under the ROC curve* (AUC). More details on the above mentioned protocols can be found in [36].

The IJB-A dataset [45] provides for two types of protocols, namely, *search* and *compare*. The *search* protocol is intended to measure the accuracy of search among N gallery templates, each of which including one or more images of a subject, in terms of the *true acceptance rate* (TAR) at various *false acceptance rates* (FAR). The *compare* protocol, instead, aims at evaluating the verification accuracy between two templates. The metrics used are the TAR corresponding to FAR equal to .1 and .01, and the rank-1 and rank-5 accuracy. The IJB-A contains 10 splits of data. A detailed descriptions of the protocols and metrics for the evaluation can be found in [45].

Comparison with Other Frontalization Algorithms We compared our frontalizations with the ones obtained with the *funneling* [35] and *deep-funneling* [37] algorithms, and with the solution proposed by Hassner et al. [31]. For this experiment, which aims at evaluating the quality of the frontalizations, we considered a slightly different version of our pipeline, identical for all the above mentioned methods; instead of localizing the descriptors exploiting the re-projected 3D model, we densely sampled LBP features on a regular grid with cells of size 10×10 on the whole image. The face images obtained with funneling, deep-funneling and with the Hassner's technique have size around 100×100 pixels, while our solution generates bigger images ($\approx 200 \times 200$). For a fair comparison, we also report results obtained with a rescaled version of our frontalizations (*Our-resized*), in order to approximately match the size of the others. Matching is performed using a simple NN-classifier. Results reported in Fig. 3.1 and Table 3.2 show

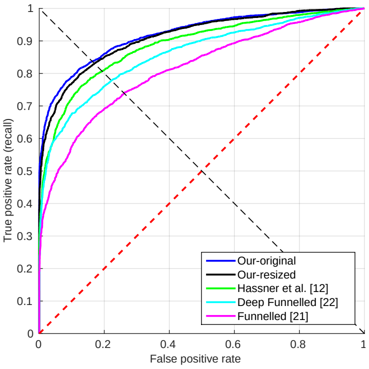


Table 3.1: ROC curves on LFW using dense sampling on the frontalized face images.

Method	**AUC**	**EER**
Funneled [35]	81.36	26.07
Deep-funneled [37]	85.91	22.20
Hassner et al. [31]	88.69	19.30
Our-resized	91.28	16.97
Our-original	**92.00**	**16.27**

Table 3.2: Area Under Curve (AUC) and Equal Error Rate (EER) values on LFW using dense sampling on the frontalized images. Values are in percentage

that our method produces a more effective frontal rendering inasmuch as the same verification algorithm is used. Even halving the size of the images does not significantly undermine the performance.

Comparison with State of the Art In the following, we report the results obtained on the LFW dataset and the IJB-A dataset using our full pipeline, and compare them with the state-of-the-art. We report the results on the LFW dataset obtained following the "Unsupervised" protocol in comparison with the four best performing state of the art techniques[1], namely: MRF-MLBP [5], Spartans [41], MRF-fusion-CSKDA [4] and Pose Adaptive Filter (PAF) [95]. It is possible to appreciate from Fig. 3.3 and Table 3.4 that we obtain comparable performance with respect to the state of the art. It is also worth to note that our method and the PAF technique, both based on fitting a 3DMM, show the same trend in Fig. 3.3, while the other methods are based on different algorithms.

In Table 3.5, we report the results on the IJB-A dataset in comparison with two baselines: a *government-off-the-shelf* (GOTS) algorithm, and the open source face recognition algorithm OpenBR [46]. We obtain higher performance with respect to the two baselines. However, it is difficult to analyze our performance in comparison to the GOTS algorithm due to miss-

[1] We do not report the curve for MRF-Fusion-CSKDA [4] since the relative data are not available.

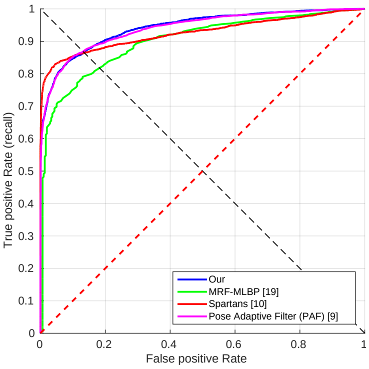


Table 3.3: ROC curves: comparison with state of the art on LFW.

Method	**AUC**
MRF-MLBP [5]	89.94
Spartans [41]	94.28
PAF [95]	94.05
MRF-fusion-CSKDA [4]	**98.94**
Our	94.29

Table 3.4: Area Under the Curve (AUC) values for our method and the state of the art on LFW. Values are in percentage

ing details about this solution. For OpenBR, instead, despite the similar pipeline, which comprises a step of dimensionality reduction via PCA followed by the application of Linear Discriminant Analysis (LDA), we show largely improved performance. We can argue that our solution generates a frontal rendering of the face image, which is more effective than unprocessed images if applied to recognition.

Metric	**GOTS**	**OpenBR** [46]	**Our**
	1:N (Search Protocol)		
TAR@FAR=0.01	.406 ± .014	.236 ± .009	**.609** ± **.015**
TAR@FAR=0.10	.627 ± .012	.433 ± .006	**.801** ± **.013**
RANK@1	.443 ± .021	.246 ± .011	**.608** ± **.023**
RANK@5	.595 ± .020	.375 ± .008	**.767** ± **.014**

Table 3.5: Results on the IJB-A dataset

3.4.2 AU Detection and Emotion Recognition

To the best of our knowledge, 3DMMs have not been used for the analysis of facial expressions; this can be reasonably ascribed to the difficulty of including expressive scans in the training data, which limits the capability of deforming a 3D model accurately in the presence of facial expressions, as shown in Fig. 3.4.

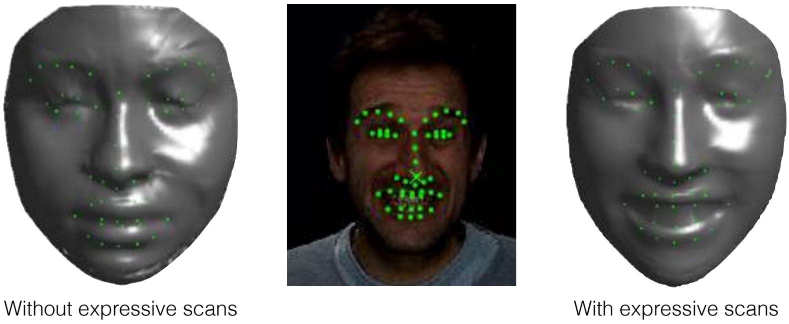


Figure 3.4: Example of fitting an expressive face with a 3DMM. The importance of including expressive scans in the training set can be appreciated: a 3DMM built without expressive scans fails in fitting the expressive face.

Facial expression analysis can be conducted mainly at two different levels: a finer one, i.e., *Action Unit (AU) detection*, which aims at detecting subtle movements of small parts of the face; and a more holistic one, which tries to classify the emotional state of the subject based on the whole face appearance, i.e., *Emotion recognition*.

Facial AUs are defined by the Facial Action Coding System (FACS) [32], which categorizes human facial movements based on the face appearance changes induced by the activity of the underlying muscles. The activation of an AU can thus be inferred from the observation of a face image. The AU detection task consists in deciding whether a particular AU is active or not in a given face image. Using this definition, in the literature, facial expressions have been systematically defined as the simultaneous activation of different AUs [52]. Facial expressions share common characteristics in the resulting face appearance and are also related to the emotional state of the subject showing the expression. Despite the precise definition, it is common that experts manually label face images referring to a set of standard discrete emotions, e.g., anger, fear, disgust, joy, happiness, relief, contempt, sadness and surprise.

To perform AU detection and Emotion Recognition the pipeline expounded in Section 3.4 has been followed. The frontalized face images are described using LBP features [59], that are concatenated and projected to a lower dimensional space by PCA. Finally, classification/detection is performed using linear SVM classifiers, trained separately for each AU or emotion. The choice of using baseline image descriptors (LBP) and classifiers (LinearSVM) is motivated by the fact that our final goal is to assess the improvement that can be obtained using the 3DMM to enhance image description.

State of the art methods for AU detection and emotion recognition [17, 23, 29, 92] have been evaluated and compared mainly on the *Extended Cohn-Kanade* (CK+) [52] and the *Facial Expression Recognition and Analysis* (FERA) [85] datasets. The CK+ dataset contains image sequences of posed and non-posed spontaneous expressions of 123 subjects (593 sequences in total). Each sequence has an average duration of about 20 frames, with the initial neutral expression varying up to a peak. The peak frame is AU-labeled, while an emotion label is associated to the entire sequence. The FERA dataset contains video sequences of 7 trained actors portraying 5 emotions. As in [23, 29], we used the training subset, which includes 87 videos ranging between 40 and 110 frames in length. Each frame is AU-labeled, while there is a single emotion label for the entire sequence. In both the datasets, the head pose is frontal in most of the sequences.

In the experiments, face images are described by LBP features [59], with a radius of 10px, following four different configurations:

- *Dense grid*, DeGr: First, the face image is cropped. Then, eyes position is retrieved from landmark detection, and used to align the image to a common reference. In this phase, in-plane rotations are compensated. Finally, the image is resized to 200 × 200 pixels, and LBP descriptors are computed over 20 × 20 non overlapping patches;

- *Landmarks*, LM: LBP descriptors are computed over patches centered in correspondence to 49 landmarks detected on the original image using the method in [42];

- DL-(O) or PCA-(O): LBP descriptors are computed over patches localized by a subset of the vertices of the 3DMM, projected onto the original image;

- DL-(F) or PCA-(F): LBP descriptors are computed over patches localized by a subset of the vertices of the 3DMM, projected onto the frontalized image.

The first two solutions do not use the 3DMM; the third and fourth, instead, perform local image description exploiting the localization provided by the

3DMM vertices. We experimentally found that a uniform subsampling of the vertices with step of 7 is the best balance between the face descriptor dimension and the patches overlap ratio. In fact, it is known that high dimensional face descriptors and a large overlapping ratio between patches improve the effectiveness of the face description [13]. For each modality, we concatenate the LBP extracted from a face so as to form a unique descriptor, and reduce the descriptor dimensionality by applying PCA with a number of PCs that retain at least the 95% of variance.

AU Detection – According to the experimental setup suggested in [52, 85], both for the CK+ and FERA datasets a *leave-one-subject-out* cross validation has been performed. For the CK+, only the neutral (first frame) and peak frames of each video sequence were used (the peak frame is the only one labeled). On the contrary, the FERA dataset comes with AUs labeled for each frame. However, not all the frames of a sequence have been used in the training phase since AUs are characterized mainly by an onset, a peak, and an offset phase. As suggested in [85], for each sequence, we consider the set of consecutive frames labeled with the peak label, and take its middle frame as corresponding to the peak phase.

Since the effect of each AU is limited to a portion of the face, accordingly to [85], AUs have been divided into *upper* and *lower* AUs corresponding to the upper half and lower half of the face, respectively. To train the SVMs, we used only the descriptors computed on points in the lower or upper part of the face, depending on which AU is considered. Each SVM is also trained independently, without accounting for the semantic relationships between different AUs (e.g., if the AU associated to the eyebrows raising is active, the AU associated to the eyebrows lowering cannot be active).

In Tables 3.7, 3.6, 3.8 and 3.9, we report the AU detection results for the CK+ and FERA datasets, respectively. Detection performance is measured in terms of $F1$-score (i.e., the harmonic mean of *precision* and *recall*) and *Area Under the ROC Curve* (*AUC*). Three main facts emerge evidently: First, localizing the descriptors with either DL-3DMM or PCA-3DMM, rather than using the regular dense grid improves the results, since the alignment is more significant; Secondly, the greater number of points provided by the projected mesh allows the computation of more descriptors, which improves the performance; Lastly, the alignment and consistency of the image representation provided by our frontalization improves the discriminating power resulting in higher overall results. This behavior is more evident for the FERA dataset, which is more challenging than the CK+. Indeed, the continuous and spontaneous nature of the sequences included in the FERA dataset induces strong nuisances in the resulting feature descriptors. The alignment and consistency obtained with our representation, however, proved to be effective in reducing the complexity to be learned by the classifier, increasing the overall results on both CK+ and FERA.

AU	*AUC*					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	95.4	95.4	98.4	98.0	98.6	98.2
2	96.7	94.2	97.6	97.3	97.5	96.9
4	92.9	91.9	95.9	96.3	96.6	97.0
5	95.7	98.0	96.6	96.1	97.8	97.5
6	94.8	94.0	95.2	95.0	95.6	95.6
7	87.7	91.8	90.1	91.4	90.1	91.9
9	99.4	99.5	99.6	99.6	99.6	99.6
11	91.2	89.1	92.8	92.6	92.9	94.5
12	98.5	98.4	98.8	98.5	98.9	98.5
15	94.7	94.9	95.5	95.4	96.2	96.5
17	95.9	94.2	97.1	97.5	97.9	97.9
20	97.7	96.5	98.4	98.0	98.6	98.5
23	91.2	95.0	94.7	95.4	94.7	94.8
24	88.6	92.8	91.9	93.0	93.0	93.7
25	97.9	98.8	99.0	98.8	99.0	98.9
26	91.6	89.0	89.6	88.8	89.7	90.2
27	99.6	99.8	99.8	99.8	99.8	99.8
Avg.	95.3	95.5	96.7	96.7	97.0	**97.1**

Table 3.6: AU detection on CK+. Comparison of different feature extraction modalities. Results are reported in terms of *AUC*. The average is weighted with respect to the number of positive instances, as indicated in [52]

For the comparison between using DL or PCA for 3DMM shape fitting, on the CK+ results are very close and this is in some way expected. In this dataset, for each sequence, we have that only the peak frame is AU labeled. Furthermore, the expressions shown are also rather exaggerated, as appreciable in the examples of Fig. 2.5. This makes the separation between

AU	$F1$-score					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	77.6	75.8	84.8	81.2	84.9	83.2
2	81.2	79.2	81.4	79.7	79.4	77.2
4	71.0	67.0	77.7	79.9	80.6	79.3
5	72.5	81.3	78.2	78.7	79.6	79.1
6	68.4	66.4	67.7	72.2	70.5	68.8
7	58.2	60.7	60.0	64.8	65.1	64.9
9	85.9	91.7	88.9	90.7	90.3	92.1
11	45.1	36.1	30.8	32.3	41.7	40.6
12	85.2	81.5	85.1	84.4	85.9	84.1
15	71.3	60.2	74.0	73.2	77.6	76.1
17	80.5	73.8	83.1	84.7	83.0	82.8
20	74.7	76.0	81.5	81.7	85.4	83.6
23	52.8	69.9	58.5	64.1	69.3	65.1
24	58.3	58.8	62.7	64.5	59.7	62.8
25	88.3	92.8	92.3	91.1	92.6	91.1
26	41.5	37.6	33.7	35.5	38.1	30.8
27	89.1	90.9	89.5	89.4	90.7	91.9
Avg.	75.3	75.1	78.2	78.9	**80.0**	78.8

Table 3.7: AU detection on CK+. Comparison of different feature extraction modalities. Results are reported in terms of $F1$-score. The average is weighted with respect to the number of positive instances, as indicated in [52]

the activation of different AUs somewhat easy and localizing the descriptors with sufficient precision becomes not crucial. This is proved by the fact that results on this dataset tend generally to saturate towards the maximum, with a rather small gap between baseline methods (DeGr and LM) and the 3DMM. The FERA dataset is instead much more challenging. The continuous and spontaneous nature of the sequences makes the gap between baseline methods (DeGr and LM) and the 3DMM increase significantly, supporting the usefulness of the latter. Finally, results show that DL performs better than PCA-3DMM on this dataset; this is mainly motivated by the fact that the face variations are more subtle and smooth and thus a better modeling improves the classification performance.

AU	$F1$-score					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	47.7	55.9	64.8	63.8	65.3	70.2
2	56.2	54.7	62.0	62.6	61.3	65.6
4	17.4	32.2	25.8	20.0	26.1	29.5
6	55.5	52.6	60.8	57.0	66.7	66.3
7	48.3	55.8	45.5	47.7	52.9	52.0
12	39.2	55.1	55.2	55.9	58.0	59.3
15	68.5	65.0	77.2	77.1	79.7	80.4
17	26.4	25.8	36.9	42.6	31.1	33.1
Avg.	44.9	49.6	53.5	53.4	55.1	**57.1**

Table 3.8: AU detection on FERA. Comparison of different feature extraction. Results are reported in terms of $F1$-score.

AU	AUC					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	77.7	78.8	83.0	81.9	85.1	83.9
2	63.5	71.0	80.7	79.1	79.2	85.8
4	48.2	53.1	46.5	51.8	52.1	54.7
6	73.0	77.3	76.3	72.9	81.0	80.0
7	71.1	66.8	57.5	57.0	62.1	64.9
12	66.5	62.9	64.8	66.5	63.9	64.9
15	73.8	81.5	84.6	82.7	85.8	87.5
17	60.5	66.9	65.3	69.9	58.8	61.7
Avg.	66.8	69.8	69.8	70.2	71.0	**72.9**

Table 3.9: AU detection on FERA. Comparison of different feature extraction. Results are reported in terms of $F1$-score and AUC

In Table 3.10 and 3.11 we provide a comparison with the state of the art in terms of average $F1$-score and AUC values. For the sake of completeness, results for the CjCRF method [92] on FERA are also reported, though they have been obtained by testing only on 260 frames out of the about 5000 total frames. Our method obtains comparable performance with respect to the state of the art on both datasets. Lower performance on the FERA dataset is likely due to the fact that our solution uses off-the-shelf descriptors and

Method	$F1$-score	AUC
IF [23]	76.6	91.3
Wang *et al.* [88]	**82.4**	96.7
CjCRF [92]	80.7	94.9
PCA-(F)	80.0	97.0
DL-(F)	78.8	**97.1**

Table 3.10: AU detection on CK+. Comparison with the state of the art. Results are reported in terms of $F1$-Score and AUC

Method	$F1$-score	AUC
Wang *et al.* [88]	52.3	-
Data-Free [49]	52.6	-
IF [23]	59.0	**74.5**
DICA [29]	59.1	-
CjCRF [92]*	**59.6**	-
PCA-(F)	55.1	71.0
DL-(F)	57.1	72.9

Table 3.11: AU detection on FERA. Comparison with the state of the art. Results are reported in terms of $F1$-Score and AUC

classifiers, and does not compensate directly for the influence of the identity in the training as is explicitly done in [23, 29]. We believe that in this sense still there is enough room for improvements.

Emotion Recognition – Data used for emotion recognition have some particular characteristics: as in the AU case, in the CK+ dataset each sequence has only two labels, one for the neutral and one for the peak frame; in the FERA dataset instead, each sequence is marked with a single label, representing the emotion of the entire sequence. For the CK+ dataset, emotion recognition is performed by considering the peak frames of each sequence in both the train and test sets; for FERA, we subsample each sequence and consider only 1 frame every 10.

In Table 3.12 and Table 3.13, we report emotion recognition results obtained using the four feature extraction methods presented in Sect 3.4.2. Consistent with the AU detection case, the results on CK+ are saturated with a small gap between the solutions that include the 3DMM and the

Emotion	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
Anger	97.6	99.0	98.6	98.8	98.9	99.4
Contempt	99.8	99.6	99.9	99.7	99.8	99.9
Disgust	99.2	97.3	97.3	93.9	99.6	99.7
Fear	99.9	99.9	99.9	99.9	99.9	99.9
Happiness	98.2	99.9	99.7	99.2	98.6	99.0
Sadness	98.8	98.9	99.2	99.0	98.8	98.8
Surprise	98.1	99.6	99.4	99.3	97.6	99.4
Avg.	98.8	99.1	99.2	98.6	99.1	**99.5**

Table 3.12: Emotion recognition on CK+. Comparison of different feature extraction modalities. Results are reported in terms of *AUC*

Emotion	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
Anger	56.4	66.7	64.4	63.0	67.7	70.5
Fear	85.8	73.7	77.4	73.0	81.9	88.4
Joy	93.0	91.4	90.9	91.9	92.1	91.5
Relief	80.2	76.4	77.4	75.6	79.5	79.0
Sadness	81.1	78.0	81.0	80.7	86.2	81.5
Avg.	79.3	77.2	78.2	76.8	81.5	**82.2**

Table 3.13: Emotion recognition on FERA. Comparison with the state of the art. Results are reported in terms of *AUC*

others. However, the ones that exploit 3DMM and frontalization are the best performing. Results on FERA, instead, show that there is actually a tangible advantage in using the 3DMM for emotion recognition. From Table 3.13 we can see that DL-(F) and PCA-(F) are, respectively, the best and the second best performing solutions, but DeGr performs better than DL-(O) and PCA-(O). This behavior can be explained considering that emotion recognition is based on the observation of the whole face appearance. In this case, localizing the descriptors precisely seems to become less important than having a consistent and pixel-wise aligned image representation.

In Table 3.14 and Table 3.15, we report our results in terms of AUC in comparison with state of the art solutions, respectively, for the CK+ and FERA datasets. We observe that our solution outperforms the state of the art on the CK+ dataset, but scores lower performance than [23] on FERA.

Emotion	IF [23]	PCA-(F)	DL-(F)
Anger	96.4	98.9	**99.4**
Contempt	96.9	99.8	**99.9**
Disgust	96.0	99.6	**99.7**
Fear	95.5	**99.9**	**99.9**
Happiness	98.9	98.6	**99.0**
Sadness	93.3	**98.8**	**98.8**
Surprise	97.6	97.6	**99.4**
Avg.	96.4	99.1	**99.5**

Table 3.14: Emotion recognition on CK+. Comparison with the state of the art. Results are reported in terms of *AUC*

Emotion	IF [23]	PCA-(F)	DL-(F)
Anger	**78.6**	67.7	70.5
Fear	85.5	81.9	**88.4**
Joy	**95.0**	92.1	91.5
Relief	**88.4**	79.5	79.0
Sadness	84.8	**86.2**	81.5
Avg.	**86.5**	81.5	82.2

Table 3.15: Emotion recognition on FERA. Comparison with the state of the art. Results are reported in terms of *AUC*

As for AU detection, this deficit of performance can be safely ascribed to the fact that differently from [23], we do not compensate the identity influence in the training.

3.5 Conclusions

In this chapter it has been described an effective algorithm able to generate an artifact-free frontal rendering of unconstrained face images based on fitting the DL-3DMM. The fitted 3D model is then used to locate the coordinates where to extract local feature descriptors on the frontalized images. This strategy enhances the consistency and alignment between descriptors, leading to improved results. The method has been tested in three different tasks, namely face recognition, emotion recognition and AU detection. It has been demonstrated that with the proposed face representation results comparable with the state-of-the-art can be achieved, even using baseline descriptors and learning strategies.

However, the method is not exempt from limitations. First of all, it heavily relies on the accuracy of the landmark detector. Moreover, the 3DMM fitting, besides being lightly affected by the accuracy of the landmark detection as well, is conditioned by the image resolution since it indirectly determines the magnitude of the deformation applied to the 3DMM. Finally, for extreme poses ($>\approx 60^\circ$ in yaw rotation), the method introduces some artifacts in the final image due to a wrong estimation of the projected model's convex hull. Some future developments will regard finding solutions to the latter issues.

Chapter 4

Face recognition with DCNN representation

Deep learning based approaches proved to be dramatically effective to address many computer vision applications, including "face recognition in the wild". It has been extensively demonstrated that methods exploiting Deep Convolutional Neural Networks (DCNN) are powerful enough to overcome to a great extent many problems that negatively affected computer vision algorithms based on hand-crafted features. These problems include variations in illumination, pose, expression and occlusion, to mention some. The DCNNs excellent discriminative power comes from the fact that they learn low- and high-level representations directly from the raw image data. Considering this, it can be assumed that the performance of a DCNN are influenced by the characteristics of the raw image data that are fed to the network. In this chapter, we evaluate the effect of different bounding box dimensions, alignment, positioning and data source on face recognition using DCNNs. A thorough evaluation of different distance measures is also presented exploiting two well known, public DCNN architectures.

4.1 Introduction and Related Work

In Computer Vision, the human face has been studied for long time either for understanding emotional states from expressions or as biometric feature for recognizing subjects' identity. Face recognition, in particular, compared to other biometric modalities is attractive since it does not require the contact with any sensor and can be performed at a distance in an uncooperative way. However, recognition based on faces suffers from several factors that can potentially impair the accuracy of the results. Many of these factors are not directly related to the natural variability of human faces due to sex, ethnicity, age. Some of them depend on variations of the face induced by facial expressions, beard, face occlusions due to hair or accessories like glasses, scarves, etc. We refer to these factors as *intrinsic*, since the variations associated to them directly affect the face surface. On the other hand, other factors that make face recognition a difficult task are due to the *extrinsic* conditions under which the face is captured. These include ambient illumination, pose, distance, resolution of the captured images, availability of single or multiple images or videos. Three-dimensional acquisitions of the face are also possible.

Most of the research work on face recognition tried to define and extract hand-crafted features capable of capturing the traits of the face that can better discriminate from subject to subject. For many years, this has been done on images acquired in cooperative contexts. The shift from cooperative to uncooperative datasets, acquired in the wild without subjects cooperation [39], contributed to substantially advance the research in this field orienting it towards more realistic solutions. Indeed, the last few years have seen the increasing success in applying deep learning based solutions to face recognition [60, 75, 79, 81]. One substantial innovation of deep convolutional neural networks (DCNNs) is the idea of letting the deep architecture to automatically discover low-level and high-level representations from labeled (or/and unlabeled) training data, which can then be used for detecting, and/or classifying the underlying patterns. However, this implies an extremely costly training phase, where millions of parameters must be optimized, thus requiring a huge number of example images. This problem can be smoothed by learning on one dataset and then reusing such learned features in different contexts using transfer learning [83] or fine tuning [98].

The literature on face recognition has been dominated for long-time by the definition and use of hand-crafted features such as Local Binary Patterns

(LBP) [1], Histogram of Gradients (HOG) [25] or Scale Invariant Feature Transform (SIFT) [28]. These features were extracted from face images and subsequently used for training classifiers like SVM [77]. The trained classifiers were ultimately used to recognize the identities. In the last few years, the scenario has been drastically changed by the combined availability of increasing computational resources and of very large datasets that made possible the effective training of neural networks with deep architecture. These learning tools showed impressive recognition results in several visual tasks, including face recognition. In the following, we revise some recent works that use DCNN architectures for face recognition.

In [81], Taigman et al. proposed DeepFace, a nine-layer deep neural network architecture for face recognition. DeepFace comprised more than 120 million parameters using several locally connected layers without weight sharing, rather than the standard convolutional layers. This network was trained on an identity labeled dataset of four million facial images belonging to more than 4,000 identities. Explicit 3D face modeling was used to align the images using a piecewise affine transformation. The learned representations coupling the accurate model-based alignment with the large facial database generalized well to faces in unconstrained environments, even with a simple classifier.

In [79], Sun et al. proposed to learn a set of high-level feature representations through deep learning for face verification. These features, referred to as Deep hidden IDentity features (DeepID), were learned through multi-class face identification tasks, whilst they can be generalized to other tasks (such as verification) and new identities unseen in the training set. DeepID features were taken from the last hidden layer neuron activations of DCNN. When learned as classifiers to recognize about 10,000 face identities in the training set and configured to keep reducing the neuron numbers along the feature extraction hierarchy, these DCNNs gradually form compact identity-related features in the top layers with only a small number of hidden neurons. These features were extracted from various face regions to form complementary and over-complete representations.

The FaceNet system proposed in [75] by Schroff et al., learned a mapping from face images to a compact Euclidean space, where distances directly correspond to a measure of face similarity. Once this space is obtained, tasks such as face recognition, verification and clustering were implemented using standard techniques with FaceNet embedding as feature vectors. A

DCNN was trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer as in previous deep learning approaches. Triplets of roughly aligned matching / non-matching face patches generated using an online triplet mining method were used for training, with the main benefit of a better representation efficiency. State-of-the-art face recognition performance was obtained using only 128-bytes per face.

In the work of Parkhi et al. [60], a much simpler and yet effective network architecture achieving near state-of-the-art results on all popular image and video face recognition benchmarks was proposed. On the one hand, they showed how a very large scale dataset (2.6M images of over 2.6K people) can be assembled by a combination of automation and human in the loop, and discussed the trade off between data purity and time. On the other, they traversed through the complexities of deep network training and face recognition to present methods and procedures to achieve comparable state of the art results.

The work of Masi et al. in [57], addressed unconstrained face recognition in the wild focusing on the problem of extreme pose variations. As opposed to other techniques that either expect a single model to learn pose invariance through massive amounts of training data, or normalize images to a single frontal pose, this method explicitly tackled pose variation by using multiple pose specific models and rendered face images. DCNNs were used to learn discriminative representations, called Pose-Aware Models (PAMs) using 500K images from the CASIA WebFace dataset [96]. In a comparative evaluation, PAMs achieved better performance than commercial products also outperforming methods that are specifically fine-tuned on the target dataset.

Unsupervised joint alignment of images has been demonstrated to improve performance on face recognition. The alignment reduces undesired variability due to factors such as pose, while only requiring weak supervision in the form of poorly aligned examples. Following this idea, Huang et al. [34] proposed Deep funneling as a novel combination of unsupervised joint alignment with unsupervised feature learning. Specifically, they incorporated deep learning into the alignment framework. In addition, the learning algorithm was modified for the restricted Boltzmann machine by incorporating a group sparsity penalty, leading to a topographic organization of the learned filters and improving subsequent alignment results. The method was applied to the LFW database. Using the aligned images produced by this unsupervised algorithm, higher accuracy in face verification was achieved compared to prior work in both unsupervised and supervised alignment.

In [58], a comprehensive study was presented that evaluates the performance of deep learning based face representation under several conditions, including the varying head pose angles, upper and lower face occlusion, changing illumination of different strengths, and misalignment due to erroneous facial feature localization. Face representations were extracted using two successful and publicly available deep learning models, namely, VggFace [60] and Lightened CNN [91]. Images acquired in controlled conditions were used in the experiments. The obtained results showed that although deep learning provides a powerful representation for face recognition, it can still benefit from preprocessing, for example, for pose and illumination normalization. In particular, from this study it emerged that if variations included in test images were not included in the dataset used to train the deep learning model, the role of preprocessing became more important. Experimental results also showed that deep learning based representation is robust to misalignment and can tolerate facial feature localization errors up to 10% of the inter-ocular distance.

Though the proliferation of deep learning based solutions for face recognition, there are several aspects of their behavior that remain not completely understood or that have not been investigated at all. In addition, the effect on the final recognition accuracy of intrinsic or extrinsic factors has been evaluated only in a limited set of cases under controlled conditions [58].

In this chapter, it is presented a thorough study on the effect that different bounding boxes, alignment and positioning variations have on deep learning based face recognition. In addition, we also experiment how different data sources (still images of video frames) weigh on the effectiveness of the representations learned through DCNNs. To this end, we first identified two publicly available and effective DCNN architectures, namely, AlexNet [47] and vgg-vd16 [60]. We trained these networks on face data in the "wild" taken from [60] and tested them on the publicly available IARPA Janus Benchmark-A (IJB-A) [44] and YouTube Faces [89] datasets.

We compared the results obtained by using the images/frames included in the original dataset, with respect to the case where these images have been normalized in a preprocessing phase. In summary, the main contributions and outcomes of this work are: *(i)* a thorough experimentation on face data in the "wild" that evaluates the effect on recognition results of bounding box dimensions, alignment, positioning and data source; *(ii)* the evidence that deep architectures do not benefit from preprocessing operations that normalize input data both at train and test time and and *(iii)* the proof that different distance measures lead to very diverse results and can be used as indicators of the effectiveness of the face representation learned by a CNN.

4.2 Face Images Preprocessing for DCNN

The effectiveness of a face recognition system based on CNN architectures depends on some main aspects. First, the network architecture and learning strategy: depending on the task, different networks and learning methodologies can be more or less effective, for instance in face recognition it has been demonstrated that deeper architectures obtain better results [60]. Second, the image content: the effect that variations in illumination, pose, expression, resolution and others have on the final performance is a crucial aspect that indeed has been extensively studied in controlled conditions [58]. Third, the data preprocessing: this includes, first of all, the detection and the clipping of the interested area, *i.e.*, the face, the compensation of nuisances such as in-plane or out-of-plane rotations, misalignments and scale differences. Finally, the source of the data, *i.e.*, whether video frames or still images are considered.

Our aim consists in evaluating the impact that different factors have on the performance of a face recognition system based on CNN representation, rather than the image content itself. To this aim, we consider the following aspects: *(i)* bounding box dimension; *(ii)* alignment and *(iii)* positioning.

4.2.1 Bounding Boxes Dimension

The dimension of the bounding box that contains the face is relevant inasmuch as it works as a trade-off between the amount of useful information, *i.e.*, the face and non-useful information, *i.e.*, background that will be fed to the network. Tighter bounding boxes will reduce the amount of background included but, on the other hand, will eventually reduce the amount of facial information and vice versa. In this sense, it can be beneficial to understand how size differences can impact on the representation obtained through the CNN. Since many different face detection algorithms exist, inferring a general rule to simulate their response and define a bounding box might result

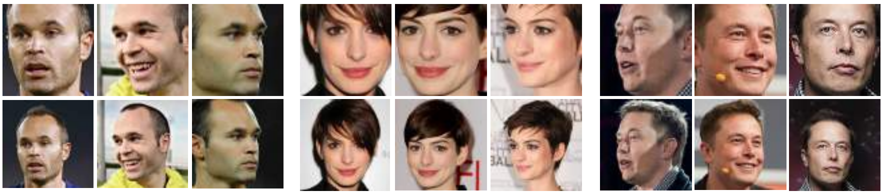


Figure 4.1: Examples of different bounding box dimensions: *(top)* tight bounding boxes; *(bottom)* large bounding boxes.

tricky. Alternatively, one could have tried to define such boxes based on any arbitrary consideration on fiducial parts of the face *e.g.* eyes distance. However, the bottleneck of this solution is that face shapes can differ a lot, yielding to subject-specific boxes that ultimately could impair the generality of the conclusions. Thus, we considered the output of a generic face detector to devise two different bounding box sizes:

- *Tight*: these bounding boxes consider a square that goes from the chin to just above the eyebrows. They resemble the output of most of the available face detectors, which are designed to minimize the amount of background. See examples in Fig. 4.1 (top row);

- *Large*: these bounding boxes are taken so as to include the whole head, thus the amount of background is variable depending on the head position. To roughly obtain such boxes, the *tight* ones have been enlarged by 15% on each side, see Fig. 4.1 (bottom row).

4.2.2 Alignment

The alignment process consists in bringing all the faces in the same relative position inside the crops so as to enhance the description semantics. Although the usefulness of the alignment step is well founded for engineered computer vision methods based on hand-crafted features, it has not been fully investigated if the effort made to perform the alignment is worth when using CNN representations. To this end, we applied two different strategies on the images used both to test and train the networks:

- *Similarity Transformation*: it is performed using the eyes position, identified by either manual annotation (if available) or exploiting a

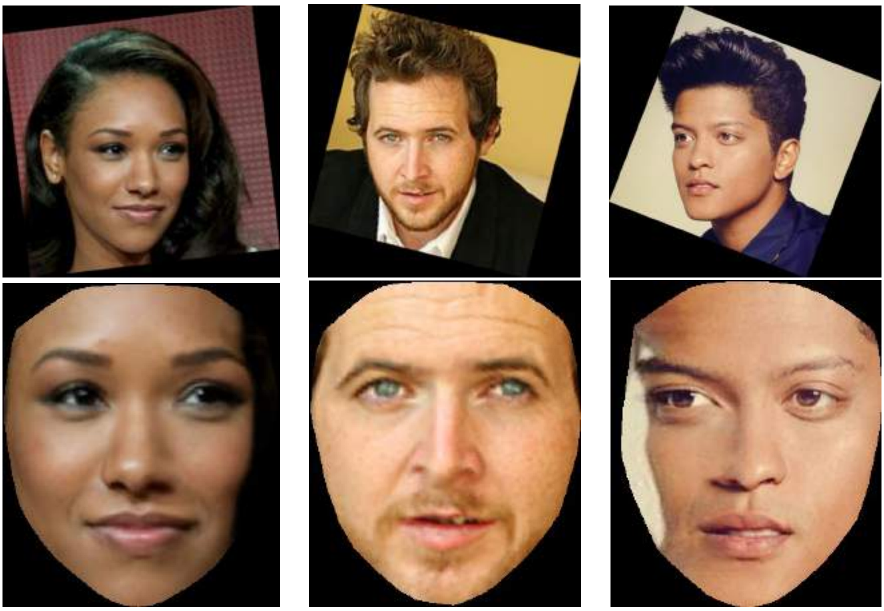


Figure 4.2: Faces aligned with a similarity transformation (top row) and frontalization (bottom row).

landmark detector [42]. Following a standard procedure, the image is warped so that the line connecting the eyes is horizontal and the distance between them is 100px. Their relative position inside the image is kept fixed. Some examples of aligned faces are shown in Fig. 4.2 (top row);

- *Frontalization*: with the term "frontalization" we refer to the process of bringing a generic face image in a frontal pose. This implies the compensation of out-of-plane rotations of the head and the rendering of a virtual frontal face image. To perform the frontalization, the method in [27] has been used. It exploits the 3D information provided by a 3D Morphable Model (3DMM). Through 2D and 3D landmark correspondences, the method estimates the 3D pose of the head and fits the 3D model to the face image. It then samples and associates the face pixel values to the 3D model vertices and finally renders a frontal face image. The rendered image is pixelwise aligned by construction.

Some examples are shown in Fig. 4.2 (bottom row).

4.2.3 Positioning

If the alignment is not applied to the images, the relative position of the face inside the bounding box can vary, with more pronounced variations for larger bounding boxes. Assuming that different face detectors can produce different outputs and that we cannot exclude detection errors, the goal here is to evaluate if and how much this behavior affects the recognition. To this aim, we consider the larger bounding boxes (we can assume that it is always possible to enlarge a bounding box if it is too tight) and take random or fixed crops out of it. In doing so, we also have the chance to understand if there are some face parts that retain more discriminative information than others.

4.3 Face Representation with DCNN

We used the data collected in [60] to train two DCNN architectures, namely AlexNet [47] and vgg-vd16 [60]. Different versions of these two architectures have been trained varying the preprocessing applied to the training face images. In particular, we considered different bounding boxes dimensions and alignments, as described in Sect. 4.2.1 and Sect. 4.2.2, respectively.

These networks have been trained as face classifiers considering $N = 2,622$ unique individuals. For each individual, an average of 1000 face images have been used during training, for a total of $2,622,000$ images. The final fully-connected layer containing N linear predictors, one per identity, along with the empirical softmax log-loss are used to train the classifier.

4.3.1 AlexNet

The architecture of this network takes a 227×227 image and is made up of 8 layers, 5 convolutional (Conv) and 3 fully connected (FC), each one followed by a rectification layer (ReLU). Max pooling is applied after the second and the fifth Conv layers. Three FC layers follow, and the output of the final FC is fed to a $2,623$-way softmax, which produces a distribution over the classes.

We trained five configurations of this architecture. As input for the training of two of these networks, we considered the original training images with two different bounding boxes dimensions (tight or large), but without alignment. Then, we trained two other configurations applying the similarity transformation described in Sect. 4.2.2 to both tight and large bounding boxes. For these four configurations, augmentation based on both random flip and the choice of a random crop have been used during training.

Finally we trained a network considering the frontalized images. In this case the training set comprises about $1,800,000$ images; this is due to landmarks detection failures for the remaining $800,000$ images. Data augmentation is not applied since frontalized faces are pixel-wise aligned and thus selecting a random crop would only result in a loss of information.

4.3.2 Vgg-vd16

We also considered the VggFace pre-trained network built upon the Vgg-vd16 architecture that has been released by the authors of [60]. This network takes a 224×224 input image and has 8 convolutional blocks, each one followed by a ReLU. Max pooling is applied every 2 Conv layers until layer 10, then every 3. The last 3 blocks are FC layers and, similarly to AlexNet, the output of the last FC is fed to the softmax layer. No alignment has been applied to the face images used for training. Augmentation based on both randomly flipping and cropping the images was applied during training.

In this work we exploit the publicly available pre-trained model and, for a more thorough comparison, we also trained a Vgg-vd16 network using the frontalized images and the settings described in Sect. 4.3.1 for the AlexNet-frontalized.

4.4 Experimental Results

We evaluate the performance of the different DCNNs in a set of experiments that involve the preprocessing operation presented in Sect. 4.2. First, we evaluate face identification and verification accuracies both for different combinations of train and test data normalization, *i.e.*, whether alignment or frontalization are applied or not, and in function of the bounding box dimension. Regarding the latter, we also conduct an experiment aimed at finding its optimal size. Then, an evaluation of which face part carry the most valuable and discriminative information is performed. These analyses have been carried out considering mainly the AlexNet architecture. Finally, a specific protocol in which gallery and probe images are divided in terms of the data source (i.e., still images or frames) is devised, so as to figure out how much this aspect influence a DCNN recognition accuracy.

Experiments have been carried out on the recently released IARPA Janus Benchmark-A (IJB-A) [44] and the YouTube Faces (YTF) [89] datasets. Both are divided in ten splits for experimental evaluation; for each trial, we use 1 split as test set and the other 9 splits as training set.

IJB-A: Released by IARPA, this dataset is specifically designed to push the challenges of face recognition to the extreme, including face imagery coming both as still images or video frames captured under severe variations of imaging conditions, focusing on the extreme cases. The dataset comprises a total of 25800 images and video frames of 500 subjects. There are two main protocols defined: face identification (1:N) and face verification (1:1); in both the protocols the identities to be matched or retrieved are expressed by means of templates, *i.e.*, sets of images/frames of the same subject. This setting is sometimes referred in literature as *template based face recognition*. Specifically, in the identification protocol, identities in the *probe* set have to be retrieved among the ones in the *gallery* set. In the gallery, each template corresponds to a single identity while in the probe set a single identity can have more than one template.

YouTube Faces: The YTF dataset collects videos from YouTube and it is specifically designed to study the problem of face verification in videos. The dataset contains 3425 videos (the average video length is 181 frames) of 1595 subjects, and the task is to decide whether two video sequences contain the same subject.

4.4.1 Recognition Pipeline

In order to assess the role that different image preprocessing procedures have on the final performance, we followed a standard recognition pipeline, exploiting the trained DCNNs as feature extractors and applying the preprocessing methods described in Sect. 4.2 to the test images. For all the tests, the output of the last fully connected layer is used as 4096-dimensional face descriptor. The latter is extracted from the images and their horizontally flipped version; the final descriptor is obtained as the average of the two. The descriptors of the training set are used to compute a PCA projection matrix to perform dimensionality reduction on the test set. Finally, we perform the matching, though in a slightly different manner for the IJB-A and YTF;

specifically, for the IJB-A, the cosine distance between each image included in each template is computed and the sum of the minimum of the distances and their average is taken as final measure. We found that including the average improves the results since it attenuates the effect of possible outliers in the templates. For YTF instead, for each video sequence, the average descriptor is coarsely obtained from all the frames and used as final descriptor for the subject in the sequence. The verification is then performed by computing the cosine distance between pairs of descriptors.

4.4.2 Preprocessing Analysis

Results for the identification and verification protocols on the IJB-A are reported in Table 4.1 and 4.2, respectively. The second and third columns indicate the type of preprocessing; *Large*(L) or *Tight*(T) refer to the bounding box dimension while *Original*(O) or *Aligned*(A) refer to whether a similarity transformation is applied to the images or not. All the possible combinations of train and test data have been experimented and for each training data type, the best configuration is reported in bold (excluding the configurations that use the frontalized version of the images).

From the results, we can first observe that there is a clear advantage in using larger bounding boxes both in training and testing the networks. This suggests that the networks are able to separate between useful (face) and non useful (background) content themselves while training, taking advantage from the larger amount of available information. If larger bounding boxes are used in the training phase, however, the performance loss using tighter boxes in the test phase is evidently less than the opposite case. This is somewhat not surprising, since it is evident that the networks cannot recognize visual information unseen during the training. A more surprising fact instead is that, for each testing configuration, better results are achieved when using larger boxes with non aligned data to train the networks. This evidence suggests that the networks are able to account for and be somewhat invariant to similarity transformations. This capability is beneficial also if aligned data is being tested (note that the original VggFace architecture used in this work exploits non aligned data for training). The fact that this is not true when using tighter boxes can be ascribed to the lack of meaningful visual information. We can reasonably suppose from the experimental evidence that the available visual content is not sufficient to make the network fully extrapolate the features that carry the identity information. Finally, it is worth to stress that the consistency between training and testing data is of fundamental importance; for all the different training configurations, the best performance are obtained with testing data that is consistent with the training one.

			Identification 1:N				Verification 1:1	
Net	Train	Test	TAR@0.01FAR	TAR@0.001FAR	Rank@1	Rank@10	TAR@0.01FAR	TAR@0.001FAR
AlexNet	A-L	A-L	**0.873 ± 0.012**	**0.728 ± 0.029**	**0.861 ± 0.014**	0.967 ± 0.004	**0.850 ± 0.018**	**0.731 ± 0.028**
AlexNet	A-L	A-T	0.806 ± 0.014	0.603 ± 0.022	0.797 ± 0.011	0.947 ± 0.007	0.795 ± 0.019	0.651 ± 0.031
AlexNet	A-L	O-L	0.870 ± 0.008	0.712 ± 0.018	0.857 ± 0.011	**0.971 ± 0.003**	0.845 ± 0.017	0.709 ± 0.030
AlexNet	A-L	O-T	0.832 ± 0.008	0.638 ± 0.026	0.819 ± 0.008	0.956 ± 0.005	0.833 ± 0.020	0.693 ± 0.035
AlexNet	O-L	A-L	0.887 ± 0.010	0.738 ± 0.020	0.872 ± 0.008	0.971 ± 0.004	0.854 ± 0.018	**0.732 ± 0.033**
AlexNet	O-L	A-T	0.825 ± 0.010	0.633 ± 0.018	0.811 ± 0.014	0.955 ± 0.008	0.807 ± 0.022	0.668 ± 0.029
AlexNet	O-L	O-L	**0.894 ± 0.010**	**0.753 ± 0.022**	**0.886 ± 0.010**	**0.977 ± 0.003**	**0.862 ± 0.020**	0.731 ± 0.025
AlexNet	O-L	O-T	0.867 ± 0.009	0.697 ± 0.016	0.857 ± 0.007	0.968 ± 0.004	0.857 ± 0.021	0.720 ± 0.040
AlexNet	A-T	A-L	0.728 ± 0.025	0.516 ± 0.025	0.724 ± 0.023	0.919 ± 0.009	0.742 ± 0.026	0.606 ± 0.037
AlexNet	A-T	A-T	**0.827 ± 0.013**	**0.666 ± 0.031**	**0.817 ± 0.016**	0.939 ± 0.006	0.808 ± 0.024	**0.687 ± 0.038**
AlexNet	A-T	O-L	0.754 ± 0.019	0.541 ± 0.027	0.749 ± 0.017	0.932 ± 0.008	0.754 ± 0.027	0.616 ± 0.031
AlexNet	A-T	O-T	0.816 ± 0.013	0.632 ± 0.024	0.807 ± 0.019	**0.946 ± 0.005**	**0.819 ± 0.017**	0.682 ± 0.050
AlexNet	O-T	A-L	0.596 ± 0.024	0.330 ± 0.023	0.582 ± 0.022	0.859 ± 0.018	0.651 ± 0.020	0.515 ± 0.025
AlexNet	O-T	A-T	0.717 ± 0.023	0.497 ± 0.024	0.717 ± 0.020	0.911 ± 0.011	0.731 ± 0.019	0.582 ± 0.040
AlexNet	O-T	O-L	0.653 ± 0.019	0.384 ± 0.031	0.642 ± 0.022	0.896 ± 0.013	0.690 ± 0.025	0.539 ± 0.024
AlexNet	O-T	O-T	**0.749 ± 0.020**	**0.507 ± 0.050**	**0.750 ± 0.021**	**0.924 ± 0.009**	**0.779 ± 0.024**	**0.604 ± 0.079**
AlexNet	F	F	0.839 ± 0.014	0.698 ± 0.032	0.832 ± 0.019	0.952 ± 0.006	0.817 ± 0.021	0.563 ± 0.125

Table 4.1: Results on the IJB-A dataset using AlexNet architecture with different train and test data preprocessing methods.

			Identification 1:N				Verification 1:1	
Net	Train	Test	TAR@0.01FAR	TAR@0.001FAR	Rank@1	Rank@10	TAR@0.01FAR	TAR@0.001FAR
VggFace	-	A-L	0.903 ± 0.010	0.760 ± 0.028	0.890 ± 0.011	0.975 ± 0.004	0.883 ± 0.017	0.749 ± 0.030
VggFace	-	A-T	0.880 ± 0.015	0.712 ± 0.027	0.867 ± 0.013	0.967 ± 0.006	0.853 ± 0.017	0.707 ± 0.039
VggFace	-	O-L	**0.926 ± 0.011**	**0.804 ± 0.022**	**0.910 ± 0.014**	**0.983 ± 0.003**	**0.896 ± 0.016**	**0.759 ± 0.041**
VggFace	-	O-T	0.914 ± 0.011	0.746 ± 0.032	0.894 ± 0.011	0.979 ± 0.003	0.888 ± 0.017	0.735 ± 0.052
Vgg-vd-16	F	F	0.852 ± 0.010	0.725 ± 0.022	0.849 ± 0.008	0.938 ± 0.006	0.824 ± 0.021	0.574 ± 0.122

Table 4.2: Results on the IJB-A dataset using the VggFace architecture with different train and test data preprocessing methods.

Acknowledged that larger bounding boxes lead to improved representations, an analysis on the optimal dimension is conducted. The DCNN used in this experiment is the AlexNet architecture trained on large non aligned images. Fig. 4.3 reports results obtained enlarging and reducing the bounding box of a certain percentage starting from a base dimension, that is the one that precisely contains the whole head. The results evidence that the latter is the optimal dimension. One could have instead expected that, since the network has been trained on larger boxes, the performance could have benefit from an enlargement. However, we observe that, being equal the percentage, the accuracy drop is relative when enlarging the box while being more significant when reducing its dimension. This suggests us that the DCNN indeed takes advantage from all the available useful information and suffers more when that information is missing rather than when more background is included.

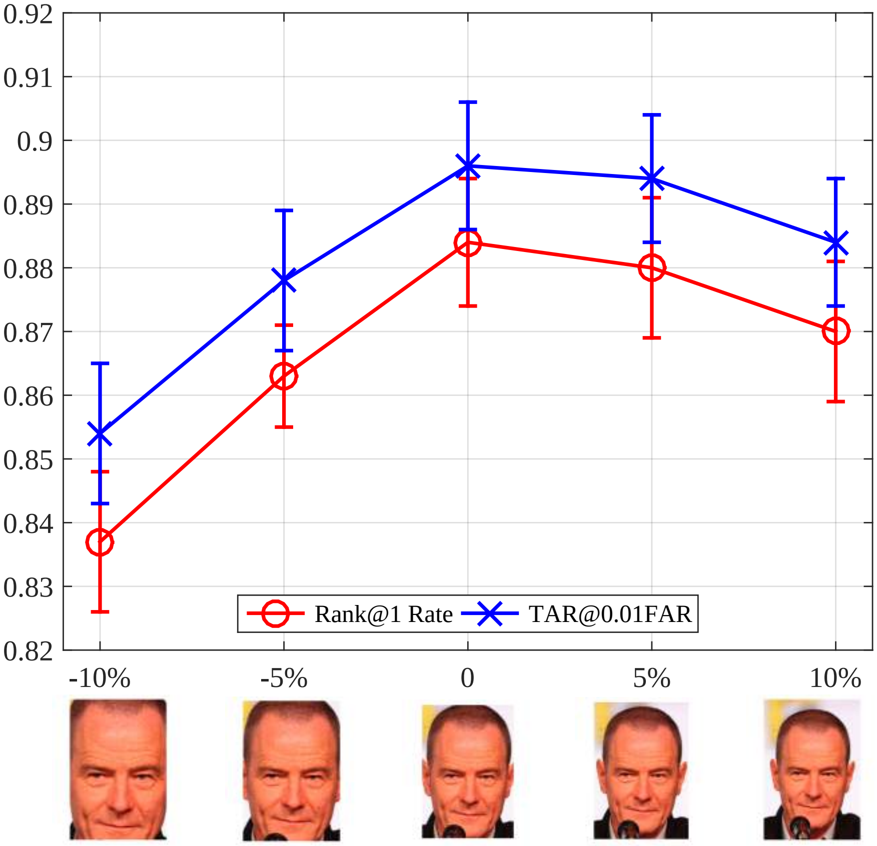
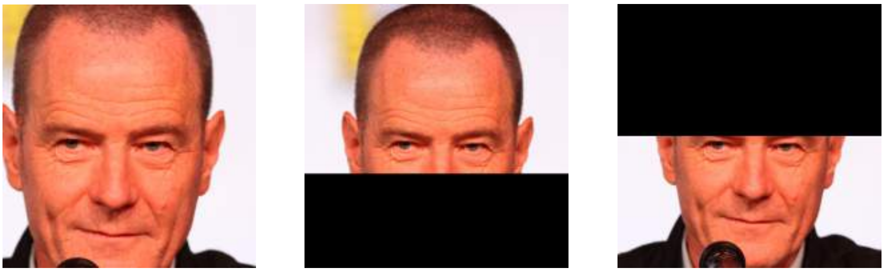


Figure 4.3: Recognition performance as a function of the dimension of the bounding box.

Table 4.3 reports results obtained simulating different shifts in the bounding box position that can occur due to detection errors. As shown in Fig. 4.4, we considered 3 cases: the first case (Fig. 4.4 (a)) simulates slight errors in the detection; the images are resized to 256 × 256 and random 224 × 224 crops are selected. Fig. 4.4 (b) and (c) instead refer to more extreme cases, where respectively only the upper or the lower halves of the face are visible. We here aim at assessing which face regions carry the most of the identity



(a) Random crop (b) Upper half (c) Lower half

Figure 4.4: Examples of: *(a)* random bounding box of the face; *(b)* upper part of the face visible; *(c)* lower part of the face visible.

information. A similar analysis regarding the occlusion of face parts is also conducted in [58], where subjects wearing sunglasses (eyes region occlusion) and scarfs (mouth-nose region occlusion) are considered. In [58] the authors show that occlusions of the eyes region dramatically worsen the recognition, while occlusions of the lower area do not influence much the results. Differently, in our experiments, we included the eyes region in both the cases (Fig. 4.4(b)-(c)) so as to deepen which of the two regions carries more discriminative information. Considering the asymmetric vertical position of the eyes, in order to retain approximately the same amount of visual information, we cut out a slightly smaller region for the upper half case (Fig. 4.4 (b)). The sizes of the crops are the 28% and 35% of the image height for the upper and lower halves respectively. Consistently with the finding of [58], Table 4.3 shows that removing the upper half of the face leads to a more significant drop of performance than excluding the lower half. Nonetheless, we can conclude that, since the eyes region is present in both, the eyebrows and forehead parts are of greater importance for the final representation. Randomly shifting the position of the bounding box, and so removing a small portion of the content, is instead not that crucial as data augmentation is applied in training the network.

4.4.3 Data Source Analysis

As data coming from both video sequences and still images is available in the IJB-A dataset, we devised a protocol to evaluate the impact of the data

	AlexNet		
	TAR@FAR 0.01	Rank 1	Rank 10
Lower-Half	0.612 ± 0.023	0.600 ± 0.018	0.881 ± 0.011
Upper-Half	0.724 ± 0.016	0.707 ± 0.015	0.924 ± 0.008
Random-Crop	0.886 ± 0.011	0.869 ± 0.011	0.974 ± 0.003
Best Configuration	**0.894 ± 0.010**	**0.886 ± 0.010**	**0.977 ± 0.003**
	VggFace		
Lower-Half	0.684 ± 0.022	0.700 ± 0.018	0.921 ± 0.011
Upper-Half	0.745 ± 0.017	0.743 ± 0.014	0.936 ± 0.005
Random-Crop	0.918 ± 0.010	0.899 ± 0.011	0.981 ± 0.003
Best Configuration	**0.926 ± 0.011**	**0.910 ± 0.014**	**0.983 ± 0.003**

Table 4.3: Comparison of different bounding box positioning for the best configuration of train and test data.

source. In this protocol four setups in which gallery and probe sets contain exclusively still images or frames are considered. To this end, we select the subset of the IJB-A identities that have at least one still image and one frame. Since in the original protocol identities in the probe set can be missing in the gallery set, this selection is made only for the gallery, so as to maintain the same set across all the setups. It resulted that, for each split, 95 out of the total 112 gallery identities are retained in all the setups. For the probe set instead, images are filtered out depending on whether still images or frames are used.

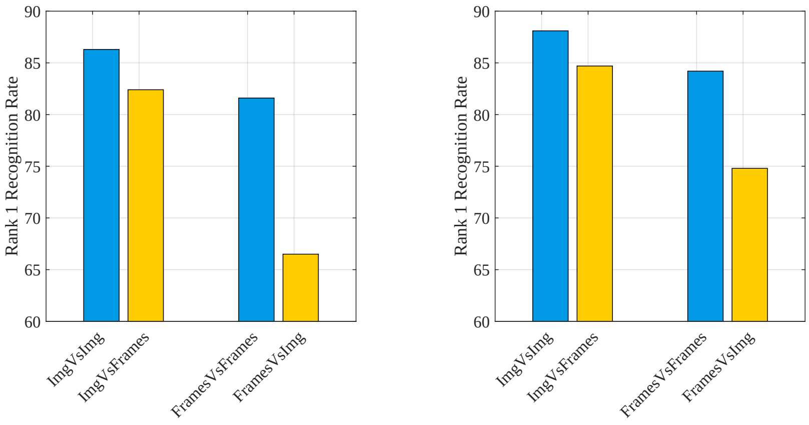


Figure 4.5: Rank-1 accuracy using different sources for gallery and probe; *(left)* AlexNet; *(right)* VggFace.

In Fig. 4.5 is shown that actually the data source does influence the accuracy. For both the DCNN architectures a performance drop is observed when gallery and probe data come from different sources, with a much more significant loss when the gallery is composed of video frames. Being aware that generally video frames have a lower resolution than still images, we believe that the different capturing formats still lead to changes in the image content and so in the extracted representation. This suggests us that can be useful to include video frames in the training set of a DCNN.

In order to confirm this assumption and better understand how different training data influence the learning procedure, a dataset containing annotated video frames has been collected, the UMD-Faces dataset [6]. It comprises 367,888 still images for 8,277 subjects and over 3.7 million annotated video frames from over 22,000 videos of 3100 subjects. Some of the 3100 identities appearing in the video frames set overlap with the identities of the YTF dataset; for this reason we selected 2927 identities out of the 3100. Following the guidelines regarding the bounding boxes, the ones provided in the dataset have been enlarged so as to approximately match the optimal size derived in Sect. 4.2.1. No alignment has been applied to the cropped face images.

The dataset has been used to finetune/retrain both AlexNet and VggFace and in particular, the following operations have been performed:

- Retrained AlexNet from scratch. The images have been subsampled in order to approximately match the number of images included in the data collected in [60]. Out of the $\approx$ 2,500,000 images, $\approx$ 150,000 are still images while the rest are video frames. Similarly to the original one, we trained the network for 20 epochs. We will refer to this solution as “AlexNet-umd”;

- Fine-tuned VggFace selecting $\approx$ 175,000 images and video frames. A new fully connected layer is stacked upon and trained with the softmax supervision to classify the new identities. For what concerns the underlying layers, two different strategies have been used :

    1. All the layers have been kept frozen except for the last fully connected layer i.e. the most semantic layer. We will refer to this solution as “Vgg-fc7”;
    2. All the layers have been fine-tuned. We will refer to this solution as “Vgg-all”;

	VggFace	Vgg-fc7	Vgg-all
ImgVsImg	0.886	0.877	**0.895**
ImgVsFrames	0.842	0.834	**0.859**
FramesVsFrames	0.846	0.830	**0.864**
FramesVsImg	0.761	0.759	**0.786**

Table 4.4: Rank@1 recognition accuracy using different sources for gallery and probe for different configurations of VggFace.

	AlexNet	AlexNet-umd
ImgVsImg	**0.863**	0.816
ImgVsFrames	**0.817**	0.750
FramesVsFrames	**0.824**	0.758
FramesVsImg	**0.665**	0.583

Table 4.5: Rank@1 recognition accuracy using different sources for gallery and probe for different configurations of AlexNet.

Both these two solutions have been trained for 10 epochs.

The experiments on the new protocols have been repeated with the new networks and results are reported in Tables 4.4 and 4.5. Results show that, first, different data sources and generally the training data has a certain impact on the final capabilities of a CNN. Secondly that there is a clear difference between the behaviors of the two networks. For what concerns the VggFace architecture, we can see that the accuracy drops if all but the last fully connected layer are blocked. We argued that the slightly different characteristics of still images and video frames could have impaired the learned representation while attempting to classify the new identities. This is demonstrated by the fact that, if all the layers are fine-tuned, the accuracy increases. On the contrary, the performance of the AlexNet architecture drop considerably if trained with the UMD dataset. Such a behavior could have many reasons that need further investigation. Among them, one could be that the AlexNet architecture is too shallow or simple to model the wider variabilities induced by the inclusion of video frames. Another reason could be the imbalance between frames and still images as well. Other than that, we wondered if the better performance obtained with VggFace could be ascribable to the fact that the original dataset contained only still images. An attempt of fine-tuning AlexNet using the same strategy and data as for VggFace has been then made, but it failed to converge, even changing various configurations.

Moreover, the new networks have been tested also on the full IJB-A dataset and results are reported in Table 4.6. These outcomes prove that the different training data types do impact on the results and indeed it can be useful to include video frames in the training data. On the other hand, it has also been shown that, depending on the architecture, the additional complexity carried by the different datasets can impair the learning capabilities and make the general training procedure harder.

	Identification 1:N				Verification 1:1	
	TAR@0.01FAR	TAR@0.001FAR	Rank@1	Rank@10	TAR@0.01FAR	TAR@0.001FAR
VggFace	0.926	0.804	0.910	0.983	0.897	0.760
VggFace-fc7	0.911	0.773	0.896	0.979	0.881	0.721
VggFace-All	**0.937**	**0.825**	**0.917**	**0.984**	**0.906**	**0.784**
AlexNet	**0.894**	**0.753**	**0.886**	**0.977**	**0.860**	**0.729**
AlexNet-umd	0.844	0.642	0.829	0.968	0.811	0.671

Table 4.6: Results on the IJB-A dataset for the different configuration of VggFace and AlexNet.

Even though these conclusions could be somewhat expected, in the attempt of better understanding the implications on the learned representation, in the following we expound an analysis regarding the final matching procedure. As stated in Sect. 4.4.1, different distance measures can be used depending on the scenario. To better understand the implications of this choice, we might want to consider the following fact: the softmax-loss used to train a network and classify identities tries to maximize the conditional probability of all the examples in the training mini-batches. In so doing, it tends to fit well to high quality faces, while difficult ones are ignored so that their uncertainty weighs as little as possible in the final cost. As a result, descriptors associated to hard examples eventually share a very low $L2$-norm, while good examples for which the classifier is confident, have high $L2$-norm [68]. For this reason, in the descriptors space, hard examples tend to be randomly displaced, usually in a common "uncertainty" area far away from the centroid of the belonging distribution, *i.e.*, identity, as shown in Fig. 4.6. Hence, if a template includes such examples, correctly or wrongly matching two templates considering the minimum distance, as with a nearest neighbor classifier, ends up to be a matter of chance. On the contrary, if the network learns effective descriptors, the majority of them should be located close to each other and, on average, farther from descriptors of different classes. Inspired by this, we conducted a series of experiments on

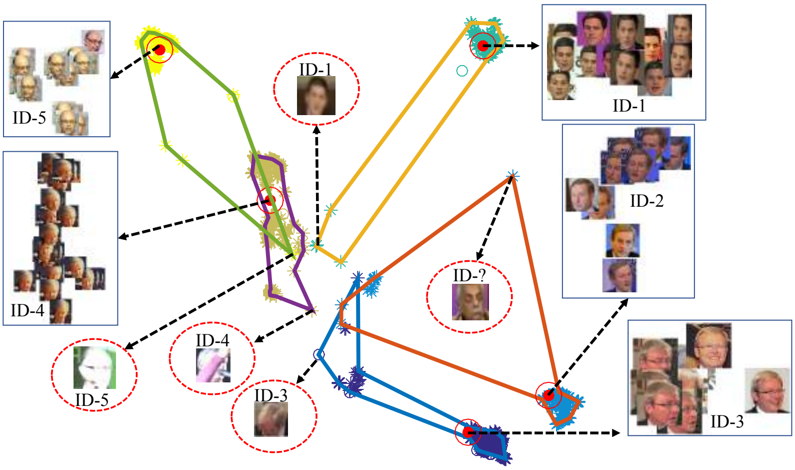
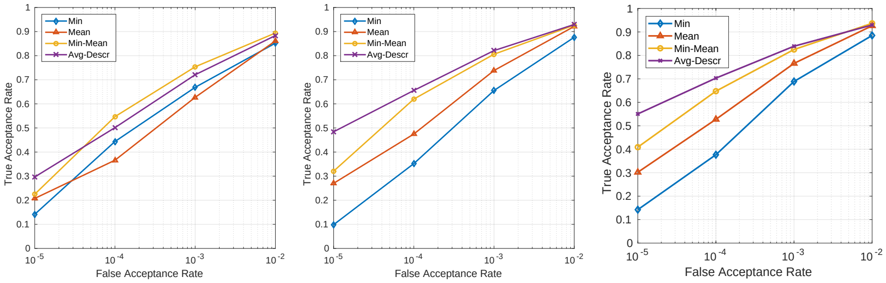


Figure 4.6: T-sne plot of descriptors for 5 different identities from the IJB-A. Different colors indicate different identities while the red circled dots represent the centroid of each set of descriptors.

the various distance measures used to match the templates, in particular we considered:

- MIN: the minimum of the distances between templates;
- MEAN: the average distance between templates;
- MIN+MEAN: the minimum plus the average of the distances, as depicted in Section 4.4.1.
- MEAN-DESCR: the average descriptor is computed from all the template descriptors.

The last strategy has a particular meaning: in fact, it gives some clues about both the goodness of the learned representation and the templates themselves. Referring to Fig. 4.6, we can see that the centroids (red circled dots) of the different identities are well separated, while some outliers make the regions (polygons) intersect. In this sense, the accuracy gap between the different distances can help in understanding the goodness of the face description. Nevertheless it should be noted that, in Fig. 4.6, all the descriptors of the 5 identities are considered. In the matching phase, the templates are

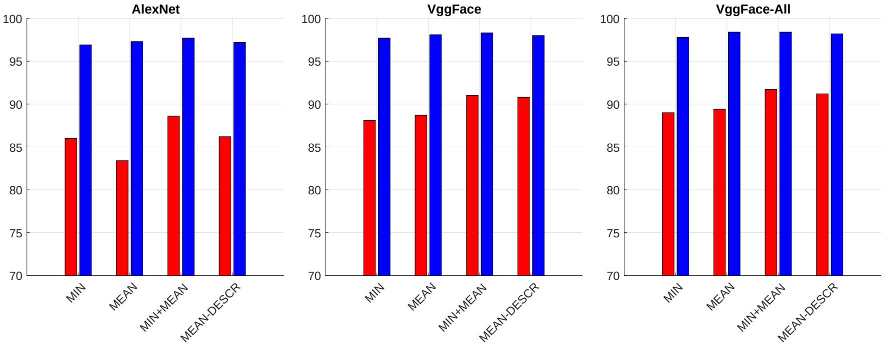


(a) (b) (c)

Figure 4.7: True acceptance rate for different false acceptance rates on the IJB-A dataset for (a) AlexNet, (b) VggFace, (c) VggFace-All in function of the different distance measures.

actually composed by a subset of them. Depending on both the number of images per template and which of them are selected, we can get many useful clues about which cases make the recognition fail or which images produce uncertain descriptions. In any case, intuitively, good results obtained using the average descriptor imply that the most of the descriptors of each identity are close to each other and the amount of outliers is reduced. Thus, while selecting "random" subsets, the probability of having centroids that are close to the belonging distribution is higher. On the contrary we argue that the minimum distance as matching measure is not totally faithful if the objective is to build an effective face representation where similar images end up close to each other in the descriptors space. Consider the situation in which some images, for whatever reason, are projected near the centroid of another class; we can take as example the pale blue points (ID-1) in Fig. 4.6 located near the centroid of the ocher class(ID-4). If these points are enough close to each other, the minimum distance will classify them correctly. On the other hand, this means that the related descriptors havebeen generated as if they belonged to a different class (the ocher class); in other words, the network did not correctly model the distribution of the data and the images have been misclassified.

In Figs. 4.7 and 4.8 results on the IJB-A for different distance measures are reported. Observing the results we can gather some considerations. First, the behavior of the two networks is different; while the VggFace architecture benefits from computing the mean descriptor or the average distance, the



(a) (b) (c)

Figure 4.8: Rank@1 and Rank@10 on the IJB-A dataset for for (a) AlexNet, (b) VggFace, (c) VggFace-All in function of the different distance measures.

AlexNet architecture does not. This can be reasonably attributed to the minor modeling power of the AlexNet network, which can also be the cause of the lower accuracy obtained with the second version of the network trained on the UMD dataset, as reported in Table 4.6. Secondly, referring to Fig. 4.7 results show that, more evidently for the VggFace network, we get a higher performance boost for lower false acceptance rates; for a false acceptance rate of 10^{-5}, we get an increase of approximately 10% over each different distance measure. Computing the mean descriptor results in a noticeable 55.1 TAR@10^{-5}FAR.

However, as stated previously, in the optimal case we expect the results of different measures to converge to a similar value. From the gaps between these values, some questions arise: one may wonder for instance if larger gaps are due to an improved or worse description capability or to minor invariance to nuisances such as pose variations. For the particular cases taken into account, the AlexNet architectures show worse results but rather small gaps between them, while the VggFace networks, on the opposite, are much more accurate but also much more sensible. Nevertheless, results in terms of retrieval i.e. Rank@1/10 in Fig. 4.8 show a stronger stability for the VggFace architectures.

	Identification 1:N				Verification 1:1	
Method	TAR@0.01FAR	TAR@0.001FAR	Rank@1	Rank@10	TAR@0.01FAR	TAR@0.001FAR
UMD (DCNN+metric) [15]	–	–	0.852 ± 0.018	0.954 ± 0.007	0.787 ± 0.043	–
UMD (DCNN$_{fusion}$) [14]	–	–	0.903 ± 0.012	0.977 ± 0.007	0.838 ± 0.042	–
PAMs [57]	–	–	0.840 ± 0.012	0.946 ± 0.007	0.826 ± 0.018	0.652 ± 0.037
Template Adaptation [22]	0.774 ± 0.050	–	0.928 ± 0.010	0.986 ± 0.003	$\underline{0.939 \pm 0.013}$	$\underline{0.836 \pm 0.027}$
TPE [74]	$\underline{0.932 \pm 0.010}$	0.753 ± 0.030	0.932 ± 0.010	0.977 ± 0.005	0.900 ± 0.010	0.813 ± 0.020
All-In-One CNN + TPE [69]	0.792 ± 0.020	–	$\underline{0.947 \pm 0.008}$	$\mathbf{0.988 \pm 0.003}$	0.922 ± 0.010	0.823 ± 0.020
NAN [94]	0.817 ± 0.041	–	$\mathbf{0.958 \pm 0.005}$	$\underline{0.986 \pm 0.003}$	$\mathbf{0.941 \pm 0.008}$	$\mathbf{0.881 \pm 0.011}$
AlexNet*	0.894 ± 0.010	0.753 ± 0.022	0.886 ± 0.010	0.977 ± 0.003	0.862 ± 0.020	0.731 ± 0.025
VggFace*	0.926 ± 0.011	$\underline{0.804 \pm 0.022}$	0.910 ± 0.014	0.983 ± 0.003	0.896 ± 0.016	0.759 ± 0.041
VggFace-All*	$\mathbf{0.937 \pm 0.008}$	$\mathbf{0.825 \pm 0.018}$	0.917 ± 0.008	0.984 ± 0.001	0.905 ± 0.008	0.784 ± 0.028

Table 4.7: State of the art results on the IJB-A dataset. *Best configurations for both AlexNet and VggFace have been selected. Best results are reported in bold and second best are underlined.

4.4.4 Comparison with State of the Art

For the sake of completeness, we compare the best configuration for both the two considered DCNNs with state of the art methods on the IJB-A and YTF datasets. Results are reported in Table 4.7 and Fig. 4.9, respectively. Results on the IJB-A show that our best configurations get very competitive results.

For what concerns the YTF dataset, we considered the original frames (without any preprocessing) to extract the DCNN descriptors. As for the bounding boxes, the provided annotations define a crop that resembles the *tight* one shown in Fig. 4.1. As we found that the best option is to have a large bounding box, we conducted the experiment two times, using both the original annotations and an enlarged version. The bounding boxes have been enlarged of 15% so as to approximately match the optimal dimension in the latter case. The ROC curves in Fig. 4.9 show that the best performance is obtained with the VggFace-All network, finetuned on the UMD, which outperforms the state of the art. We remark here that the overlapping identities between the two datasets have been removed from the training data. The results of the other architectures are consistent with the conclusions expounded so far and follow the same trend as for the IJB-A dataset.

4.5 Conclusions

In this chapter, nuisance factors that can influence face recognition performance have been investigated. We focused on the images preprocessing steps, for both training and testing. From the experimental evidence we can

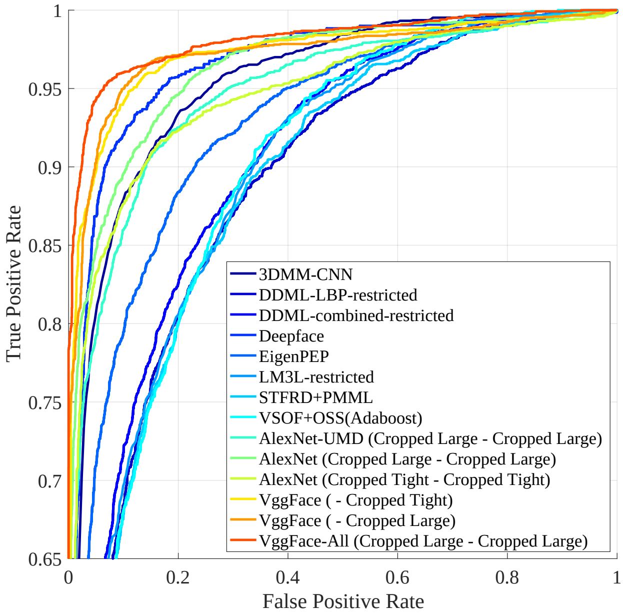


Figure 4.9: ROC curves on the YouTube Faces database for the trained architectures and the state of the art.

mainly conclude that there exist a strong dependency between train and test data and that actually the image representation derived from the DCNNs does not benefit from image normalization operations. Moreover, we evidenced that the data source combinations (images or frames) have a certain impact on the final performance, and along with them the different distance measures employed.

From this analysis some useful insights have also been derived that can help to lighten the effort in developing new solutions for face recognition in the wild exploiting DCNN.

Chapter 5

Conclusion

In the following sections, the contributions of this thesis work are presented and possible directions for future research are expounded.

5.1 Summary of contribution

The first part of the thesis dealt with the problem of face recognition "in the wild" based on local, hand-crafted features. The main challenges that comes with this problem are basically induced by the heavy changes in the face image appearance, which can be due to pose, illumination and expression variations, to mention some. The assumption on which this work is grounded on is that we can overcome some of the challenges by constructing a face representation in which variations are less relevant. This is achieved by means of a deformable 3D model.

In chapter 2, we described an effective 3D Morphable Shape Model based on Dictionary Learning which is able to fit a 3D shape to a single 2D face image in an accurate way relying solely on some facial landmarks correspondences. The model is constructed exploiting a dataset of 3D scans, which have to be registered together so as meaningful statistical information on the shapes can be derived. As a contribution, it is proposed a dense alignment method that can effectively register the set of 3D scans even in the presence of large topological changes, which are mainly due to expressive scans. We then build the morphable model by applying a dictionary learning (DL) technique on the aligned scans. We demonstrate that in presence of expressive models, which lead to a greater statistical variability in the data, DL better retains the information and can deform the model more accurately than a classic PCA based solution.

In chapter 3 we exploit the afore mentioned 3DMM to fit to target face images and generate a frontalized i.e. frontal facing view image that we prove to be more effective if used for face recognition. Briefly, the 3DMM is projected on the image plane exploiting 2D/3D landmark correspondences and fit to the face image. Then the RGB values of the face region are sampled so as to build a complete 3D model comprising shape and texture. The model is built computing an inverse projection that brings each single pixel in the face region to the 3D space. The complete model can then be rotated to a frontal pose and a new image is generated. In doing so, the model can be again back projected onto the new image; this allows us to define a set of points where to extract local feature descriptors. With respect to dividing the image into non overlapping cells and extracting the features in correspondence of the landmarks, the proposed solution get much better results. We tested the approach in three different tasks, namely face recognition, emotion recognition and Action Units detection. In all the cases the proposed face representation and feature extraction technique performed better than baseline approaches; moreover, results are comparable to state-of-the-art solutions, even using basic descriptors and learning solutions.

Nonetheless, simultaneously, the interest and the development of deep learning based solutions for face recognition improved drastically the average performance in the field. It has become clear that older techniques struggled to endure. In chapter 4 we present a thorough analysis on the behavior of deep networks trained for face recognition. We focused the analysis on the image representation, having being the latter a key step so far. In the chapter it is shown that, even though the focus is often directed on other aspects like the network architecture or the training algorithm, we can still get large benefits if the correct image representation is used.

In summary, the main findings of this work are the following:

- Dictionary Learning techniques can be effectively applied to capture the variability of the shape of human faces and used to construct a deformable model (DL-3DMM) able to fit unseen faces accurately, even in the presence of expressions.

- The DL-3DMM can be exploited to produce a frontal facing view of a generic face image. This enhances the representation by alleviating the effect of various nuisances like pose variations. An effective technique based on the 3D model to localize points where to extract local features is also proposed. This technique improved the description semantics by inducing a more accurate alignment between descriptors, which finally resulted in better recognition performance.

- The preprocessing operations applied to the images and the image representation itself are relevant also when deep learning algorithms are concerned. Depending on the characteristics of the data, its consistency in the training and testing phases and the different preprocessing operations, results can consistently vary. Taking the right choice can help to improve the final results and understand the critical aspects of such algorithms.

5.2 Directions for future work

Besides the huge evolution of computer vision and machine learning algorithms, there is still a great room for improvements and discoveries. Together with the renovation of the techniques adopted to solve the problems, the problems themselves are also evolving, in the attempt of pushing to the limit the requirements and the capabilities of systems. While results on many benchmark datasets are saturating, many other more challenging ones are being released, slightly changing also the problem itself. In the field of face recognition, this is happening in the sense that:

- Datasets include always more and more identities and images, so that the size of the problem is increasing. This obviously has an impact both on the effectiveness and efficiency requirements;

- Extreme cases are the ones of most interest; these include people turned the other way from the camera, very low resolutions, almost full occlusions and so on. This suggests that there is an interest in studying the inclusion of other information other than the face itself to solve the problem;

- Deep learning solutions are being applied also to 3D data, so that the opportunity of building hybrid methods seems feasible

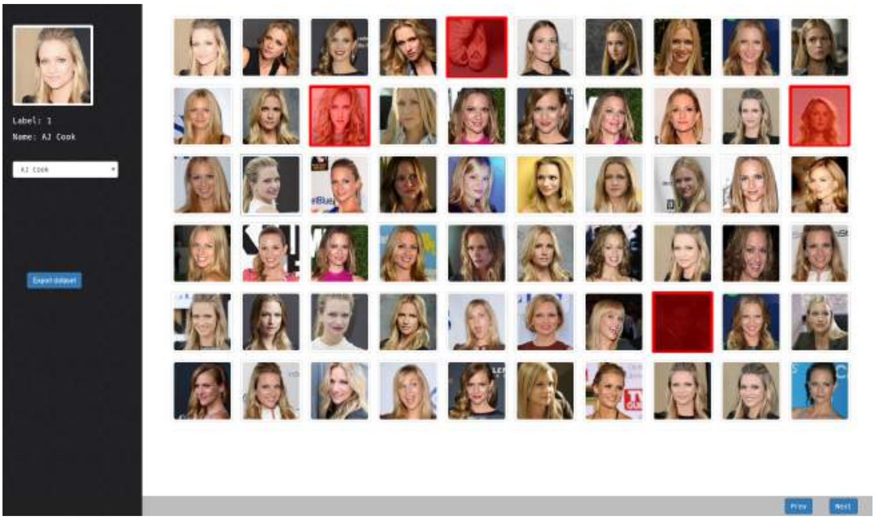
Appendix A

Automatic face image collector based on CNN descriptors

In this appendix a web application to automatically collect huge amounts of face images provided the name of the subjects is presented. The application works as follows:

- Through the list of names it queries three differente search engines. The urls returned by the engines are collected and the images are downloaded;
- A face detector is run on each image and the faces are automatically cropped;
- A filtering of the possibly wrong images is performed i.e. images of subjects different from the requested one; for each subject it extracts CNN descriptors from a subset of the downloaded images and trains a SVM classifier. This step grounds on the hypothesis that the amount of wrong images is lower than the amount of correct ones. This step is necessary also because many images contain more than one face e.g. group pictures.
- The trained classifier is finally used to discard images that are classified as not belonging to the considered identity.

In addition, a tool to manually check the downloaded images and the applied filtering has been implemented. The tool allows to check the whole set of



Label: 1

Name: AJ Cook

Figure A.1: Screenshot of the tool developed to manually check the results of the filtering.

images for each subject. The images that have been filtered out from the set are marked in red (Fig. A.1). The user can manually remove other images by selecting them or restore the discarded images if a wrong filtering has been applied. Finally the dataset can be exported considering only the images that have been selected as genuine. It is also possible to exclude a whole identity by clicking on the thumbnail of the subject (the image in the top left corner of Fig. A.1).

Appendix B

Frontalized Faces in The Wild

In this appendix we present a dataset created using the technique expounded in Sect. 3. The dataset considered is the "Labeled Faces in the Wild" [38],a database of face photographs designed for studying the problem of unconstrained face recognition. It contains 13233 images of 5749 people. [1]

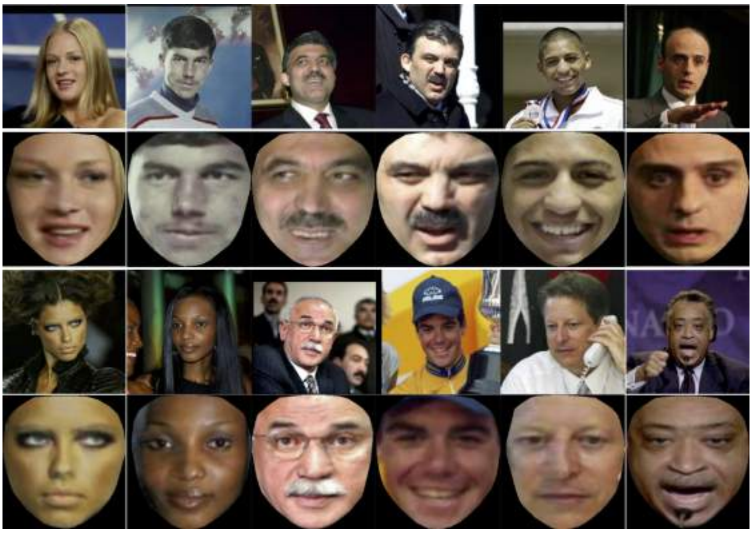


Figure B.1: Images from the frontalized faces in the wild dataset built.

[1] Available at https://www.micc.unifi.it/resources/datasets/frontalized-faces-in-the-wild/.

Appendix C

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.[1]

International Journals

1. **Claudio Ferrari**, G. Lisanti, S. Berretti, A. Del Bimbo. "A dictionary Learning Based 3D Morphable Shape Model", *IEEE Transactions on Multimedia*, vol. PP in press, May 2017.[DOI: 10.1109/TMM.2017.2707341] *1 citation*

International Conferences and Workshops

1. **Claudio Ferrari**, G. Lisanti, S. Berretti, A. Del Bimbo. "Investigating Nuisance Factors in Face Recognition with DCNN Representation", in *Proc. of IEEE International Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honululu, (Hawaii), 2017. **Best Paper Award**
2. **Claudio Ferrari**, G. Lisanti, S. Berretti, A. Del Bimbo. "Effective 3D based Frontalization for Unconstrained Face Recognition", in *Proc. of IEEE International Conference on Pattern Recognition (ICPR)*, Cancun (Mexico), 2016. *6 citations*
3. **Claudio Ferrari**, G. Lisanti, S. Berretti, A. Del Bimbo. "Dictionary Learning Based 3D Morphable Model Construction for Face Recognition with Varying Expression and Pose", in *Proc. of IEEE International Conference on 3D Vision (3DV)*, Lyon (France), 2015. *10 citations*

[1] The author's bibliometric indices are the following: H-index = 3, total number of citations = 25 (source: Google Scholar on January, 2018).

4. I. Masi, **Claudio Ferrari**, A. Del Bimbo, G. Medioni. “Pose Independent Face Recognition by Localizing Local Binary Patterns via Deformation Components”, in *Proc. of IEEE International Conference on Pattern Recognition (ICPR)*, Stockholm (Sweden), 2014. *8 citations*

Bibliography

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[2] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[3] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[4] R. Arashloo, Shervin and J. Kittler, "Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 12, pp. 2100–2109, Dec. 2014.

[5] S. R. Arashloo and J. Kittler, "Efficient processing of MRFs for unconstrained-pose face recognition," in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, 2013.

[6] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks."

[7] T. Berg and P. N. Belhumeur, "Tom-vs-Pete classifiers and identity-preserving alignment for face verification," in *British Machine Vision Conference*, 2012, pp. 1–11.

[8] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Computer Graphics Forum*, vol. 22, no. 3, pp. 641–650, 2003.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *ACM Conf. on Computer Graphics and Interactive Techniques*, 1999.

[10] ——, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[11] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahand, and D. Dunaway, “A 3D morphable model learnt from 10,000 faces,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 5543–5552.

[12] A. Brunton, T. Bolkart, and S. Wuhrer, “Multilinear wavelets: A statistical shape space for human faces,” in *European Conf. on Computer Vision*, 2014.

[13] D. Chen, X. Cao, F. Wen, and J. Sun, “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[14] J.-C. Chen, V. M. Patel, and R. Chellappa, “Unconstrained face verification using deep cnn features,” in *IEEE Winter Conf, on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[15] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa, “An end-to-end system for unconstrained face verification with deep convolutional neural networks,” in *IEEE Int. Conf. on Computer Vision Workshops*, 2015, pp. 118–126.

[16] B. Chu, S. Romdhani, and L. Chen, “3d-aided face recognition robust to expression and pose variations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1899–1906.

[17] W.-S. Chu, F. De la Torre, and J. F. Cohn, “Selective transfer machine for personalized facial action unit detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[18] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, “Detecting depression from facial actions and vocal prosody,” in *Int. Conf. on Affective Computing and Intelligent Interaction and Workshops*, 2009.

[19] D. Cosker, E. Krumhuber, and A. Hilton, “Perception of linear and nonlinear motion properties using a FACS validated 3D facial model,” in *ACM Applied Perception in Graphics and Vision*, 2010.

[20] ——, “A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling,” in *International Conference on Computer Vision*, 2011.

[21] D. Cosker, E. Krumhuber, and A. Hilton, “Perceived emotionality of linear and non-linear aus synthesised using a 3D dynamic morphable facial model,” in *Proc. of the Facial Analysis and Animation*, ser. FAA’15. ACM, 2015, pp. 7:1–7:1.

[22] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman, “Template adaptation for face verification and identification,” *arXiv preprint arXiv:1603.03958*, 2016.

[23] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, “Intraface,” in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2015.

[24] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *IEEE Trans. on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[25] O. DÃ©niz, G. Bueno, J. Salido, and F. D. la Torre, “Face recognition using histograms of oriented gradients,” *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598–1603, 2011.

[26] L. G. Farkas, *Anthropometry of the Head and Face.* Raven Press, 1994.

[27] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, “Effective 3D based frontalization for unconstrained face recognition,” in *Int. Conf. on Pattern Recognition*, 2016.

[28] C. Geng and X. Jiang, “Sift features for face recognition,” in *IEEE Int. Conf. on Computer Science and Information Technology*, 2009, pp. 598–602.

[29] C. Georgakis, Y. Panagakis, and M. Pantic, “Discriminant incoherent component analysis,” *IEEE Trans. on Image Processing*, vol. 25, no. 5, pp. 2021–2034, 2016.

[30] T. Hassner, “Viewing real-world faces in 3D,” in *Int. Conf. on Computer Vision*, 2013.

[31] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[32] C.-H. Hjortsjö, *Man’s face and mimic language.* Studen litteratur, 1969.

[33] H. T. Ho and R. Chellappa, “Pose-invariant face recognition using Markov Random Fields,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 1573–1584, Apr. 2013.

[34] G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller, “Learning to align from scratch,” in *Int. Conf. on Neural Information Processing Systems (NIPS)*, 2012, pp. 764–772.

[35] G. B. Huang, V. Jain, and E. Learned-Miller, “Unsupervised joint alignment of complex images,” in *IEEE Int. Conf. on Computer Vision*, 2007.

[36] G. B. Huang and E. Learned-Miller, “Labeled faces in the wild: Updates and new reporting procedures,” University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003, May 2014.

[37] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, “Learning to align from scratch,” in *Advances in Neural Information Processing Systems*, 2012.

[38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[39] ——, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[40] P. Huber, P. Kopp, M. Rätsch, W. J. Christmas, and J. Kittler, “3d face tracking and texture fusion in the wild,” *CoRR*, vol. abs/1605.06764, 2016. [Online]. Available: http://arxiv.org/abs/1605.06764

[41] F. Juefei-Xu, K. Luu, and M. Savvides, “Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios,” *IEEE Trans. on Image Processing*, vol. 24, no. 12, pp. 4780–4795, Dec. 2015.

[42] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[43] R. Kimmel and J. A. Sethian, “Computing geodesic paths on manifolds,” *Proc. of the National Academy of Science*, vol. 95, no. 15, pp. 8431–8435, 1998.

[44] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1931–1939.

[45] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[46] J. Klontz, B. Klare, S. Klum, A. Jain, and M. Burge, “Open source biometric recognition,” in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, 2013.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[48] H. Li and G. Hua, “Hierarchical-pep model for real-world face recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[49] Y. Li, J. Chen, Y. Zhao, and Q. Ji, “Data-free prior model for facial action unit recognition,” *IEEE Trans. on Affective Computing*, vol. 4, no. 2, pp. 127–141, 2013.

[50] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[51] X. Lu and A. K. Jain, “Deformation modeling for robust 3D face matching,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1346–1357, 2008.

[52] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in *IEEE Conf. on Computer Vision and Pattern Recognition-Workshops*, 2010.

[53] M. Lüthi, C. Jud, T. Gerig, and T. Vetter, “Gaussian process morphable models,” *CoRR*, vol. abs/1603.07254, 2016. [Online]. Available: http://arxiv.org/abs/1603.07254

[54] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Int. Conf. on Machine Learning*, 2009.

[55] I. Masi, C. Ferrari, A. Del Bimbo, and G. Medioni, “Pose independent face recognition by localizing local binary patterns via deformation components,” in *Int. Conf. on Pattern Recognition*, 2014.

[56] I. Masi, G. Lisanti, A. D. Bagdanov, P. Pala, and A. D. Bimbo, “Using 3d models to recognize 2d faces in the wild,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[57] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, “Pose-aware face recognition in the wild,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[58] M. Mehdipour Ghazi and H. Kemal Ekenel, “A comprehensive analysis of deep learning based representation for face recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 34–41.

[59] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[60] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *British Machine Vision Conf. (BMVC)*, vol. 1, no. 3, 2015, p. 6.

[61] A. Patel and W. A. P. Smith, “3D morphable face models revisited,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[62] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D face model for pose and illumination invariant face recognition,” in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, 2009.

[63] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, “3D facial landmark detection under large yaw and expression variations,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1552–1564, 2013.

[64] 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 Workshop Face Recognition Grand Challenge Experiments*, 2005.

[65] E. Piątkowska and J. Martyna, “Spontaneous facial expression recognition: Automatic aggression detection,” in *Int. Conf. on Hybrid Artificial Intelligent Systems*, 2012.

[66] E. Pontikakis, C. Nass, J. N. Bailenson, L. Takayama, and M. E. Jabon, “Facial expression analysis for predicting unsafe driving behavior,” *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 84–95, 2011.

[67] S. Ramanathan, A. Kassim, Y. V. Venkatesh, and W. S. Wah, “Human facial expression recognition using a 3D morphable model,” in *Int. Conf. on Image Processing*, 2006.

[68] R. Ranjan, C. D. Castillo, and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” *arXiv preprint arXiv:1703.09507*, 2017.

[69] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, “An all-in-one convolutional neural network for face analysis,” *arXiv preprint arXiv:1611.00851*, 2016.

[70] S. Romdhani and T. Vetter, “Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[71] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh, “An approach for automatic pain detection through facial expression,” in *Int. Conf. on Intelligent Human Computer Interaction*, vol. 84, 2015.

[72] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, “Robust statistical face frontalization,” in *IEEE Int. Conf. on Computer Vision*, 2015.

[73] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, “Static and dynamic 3D facial expression recognition: A comprehensive survey,” *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[74] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, “Triplet probabilistic embedding for face verification and clustering,” in *IEEE Int.*

Conf. on Biometrics Theory, Applications and Systems (BTAS), 2016, pp. 1–8.

[75] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[76] D. Shahlaei and V. Blanz, "Realistic inverse lighting from a single 2D image of a face, taken under unknown and complex lighting," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2015.

[77] J. Sivic, M. Everingham, and A. Zisserman, "Who are you? learning person specific classifiers from video," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1145–1152.

[78] F. C. Staal, A. J. Ponniah, F. Angullia, C. Ruff, M. J. Koudstaal, and D. Dunaway, "Describing Crouzon and Pfeiffer syndrome based on principal component analysis," *Journal of Cranio-Maxillofacial Surgery*, vol. 43, no. 4, pp. 528–536, 2015.

[79] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1891–1898.

[80] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 1701–1708.

[81] ——, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[82] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. on Multimedia*, vol. 15, no. 7, pp. 1543–1552, 2013.

[83] L. Torrey and J. Shavlik, "Transfer learning," *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, vol. 1, p. 242, 2009.

[84] H. Ujir and M. Spann, *Facial Expression Recognition Using FAPs-Based 3DMM*, ser. Lecture Notes in Computer Science. Springer Netherlands, 2013, pp. 33–47.

[85] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 4, pp. 966–979, 2012.

[86] D. J. Walger, T. P. Breckon, A. Gaszczak, and T. Popham, "A comparison of features for regression-based driver head pose estimation under varying

illumination conditions," in *IEEE Int. Work. on Computational Intelligence for Multimedia Understanding*, 2014.

[87] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[88] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *IEEE Int. Conf. on Computer Vision*, 2013.

[89] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 529–534.

[90] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[91] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation," *CoRR*, 2015.

[92] Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[93] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[94] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," *arXiv preprint arXiv:1603.05474*, 2016.

[95] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[96] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. arXiv:1411.7923, 2014.

[97] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006.

[98] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[99] A. Yuce, H. Gao, G. Cuendet, and J. P. Thiran, "Action units and their cross-correlations for prediction of cognitive load during driving," *IEEE Trans. on Affective Computing*, vol. to appear, 2016.

[100] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, “Learning personalized models for facial expression analysis and gesture recognition,” *IEEE Trans. on Multimedia*, vol. 18, no. 4, pp. 775–788, 2016.

[101] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[102] L. Zhang, Y. Wang, S. Wang, D. Samaras, S. Zhang, and P. Huang, “Image-driven re-targeting and relighting of facial expressions,” in *Computer Graphics Int.*, 2005.

[103] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, “A deep neural network driven feature learning method for multi-view facial expression recognition,” *IEEE Trans. on Multimedia*, vol. to appear, 2016.

[104] X. Zhao, W. Zhang, G. Evangelopoulos, D. Huang, S. K. Shah, Y. Wang, I. A. Kakadiaris, and L. Chen, “Benchmarking asymmetric 3d-2d face recognition systems,” in *IEEE Int. Conf. and Work. on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[105] Q. Zhen, D. Huang, Y. Wang, and L. Chen, “Muscular movement model-based automatic 3D/4D facial expression recognition,” *IEEE Trans. on Multimedia*, vol. 18, no. 7, pp. 1438–1450, 2016.

[106] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li, “Discriminative 3D morphable model fitting,” in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2015.

[107] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.