This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



# "A Leopard Cannot Change Its Spots": Improving Face Recognition Using 3D-based Caricatures

João Neves, *Student Member, IEEE,* and Hugo Proença, *Senior Member, IEEE*

***Abstract*—Caricatures refer to a representation of a person in which the distinctive features are deliberately exaggerated, with several studies showing that humans perform better at recognizing people from caricatures than using original images. Inspired by this observation, this paper introduces the first fully automated caricature-based face recognition approach capable of working with data acquired in the wild. Our approach leverages the 3D face structure from a single 2D image and compares it to a reference model for obtaining a compact representation of face features deviations. This descriptor is subsequently deformed using a 'measure locally, weight globally' strategy to resemble the caricature drawing process. The deformed deviations are incorporated in the 3D model using the Laplacian mesh deformation algorithm, and the 2D face caricature image is obtained by projecting the deformed model in the original camera-view. To demonstrate the advantages of caricature-based face recognition, we train the VGG-Face network from scratch using either original face images (baseline) or caricatured images, and use these models for extracting face descriptors from the LFW, IJB-A and MegaFace datasets. The experiments show an increase in the recognition accuracy when using caricatures rather than original images. Moreover, our approach achieves competitive results with state-of-the-art face recognition methods, even without explicitly tuning the network for any of the evaluation sets.**

***Index Terms*—Face Recognition, 3D Caricature Generation, Caricature-based Face Recognition, Facial Feature Analysis.**

## I. Introduction

HUMANS have an astonishing capability of recognizing familiar faces in totally unconstrained scenarios. However, this performance decreases significantly in case of unfamiliar faces [1]. The question of how an unfamiliar face becomes a familiar face is not consensual, but there is evidence that this process is carried out in a caricatured manner [2], [3]. According to this theory, familiarization works by analyzing the most significant physical deviations of a face with respect to a mental representation of the average face, followed by the creation of a modified description of the face, where the most distinctive features are exaggerated and average features are oversimplified (similar to drawing a caricature). Moreover, different studies concluded that humans perform better at recognizing individuals from caricatures [4], [5], [6], [7] than veridical faces, supporting the idea that the human brain encodes familiar faces as a caricatured version of the original face.

J. Neves and H. Proença are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilha, Portugal, E-mail: {jcneves,hugomcp}@di.ubi.pt.

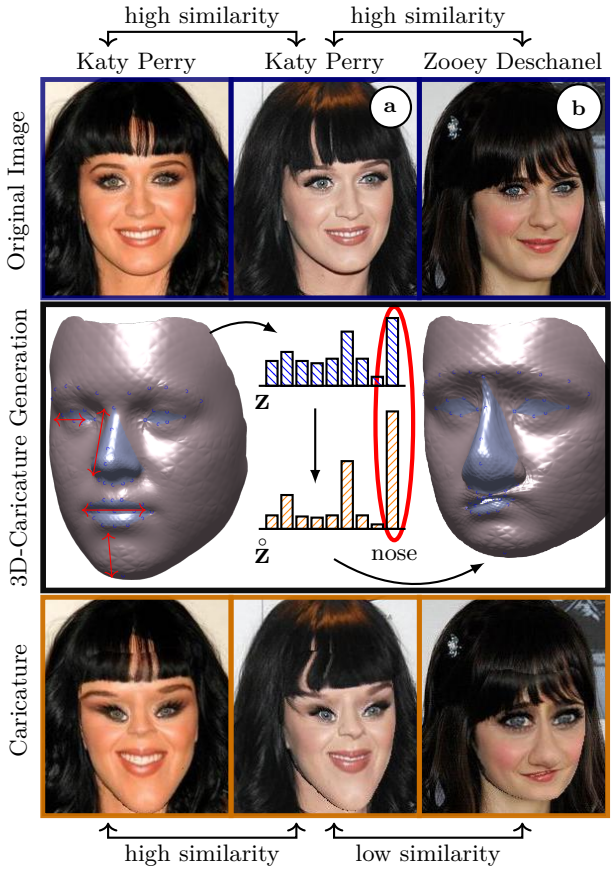


Fig. 1. **Advantages of using caricatures for face recognition.** Both humans and automated systems find difficult to distinguish between visually similar subjects (e.g., Katy Perry and Zooey Deschanel). Familiarized observers overcome this problem by focusing on the most distinguishable features of each face, and several studies suggest that this task is carried out in the brain by creating a caricatured representation of the original image. Our method aims at mimicking this process by analyzing face proportions and exaggerating the most salient ones. The proposed approach is capable of producing 2D caricatures where inter-subject similarity is minimized and intra-subject similarity is preserved.

Inspired by the idea that distinctive feature exaggeration may be the key for the incredible performance of humans on recognizing familiar faces, we introduce a fully automated face recognition approach based on a 3D caricature generation method capable of creating 2D face representations, where likeness is preserved and the inter-class separation is enlarged. The rationale behind our idea is illustrated in Fig. 1, where an unfamiliar observer

perceives incorrectly Fig. 1a) and Fig. 1b) as photos from the same identity. On contrary, it is straightforward to discern between Katy Perry and Zooey Deschanel when observing their caricatures.

For automated caricature generation, the proposed method attempts to mimic the three main stages of the caricature drawing process:

1) **Caricaturists infer 3D face structure from either a single or multiple views of the face.** This phase is replicated by estimating a 3D morphable model from an input image and a set of facial landmarks. The accuracy of the landmarks decreases significantly in unconstrained data, and for that reason, we combine multiple state-of-the-art landmark localization algorithms in an ensemble learning strategy. In addition, we use a model with a reduced number of vertices to account for model stability while maintaining the dominant features of the face.
2) **The caricaturist analyzes facial features for determining the deformation applied to each one.** After inferring the 3D structure, our method compares a set of face regions with a reference 3D model regarding translation, scale and orientation. The region deviations are then normalized and exaggerated using a 'measure locally, weight globally' strategy.
3) **The artist redraws the original face using the deformed proportions.** After determining the positions of the deformed vertices, the mesh is warped with a Laplacian mesh editing technique for preserving local detail and guaranteeing smooth transitions between vertices. The final 2D caricature is obtained by projecting the 3D model in the original camera-view.

In the learning and classification phase, we replicate the strategy introduced in [8] but using caricatures rather than veridical face images. Accordingly, the VGG-Face architecture is trained from scratch on caricatures automatically generated from the VGG dataset, whereas the features produced by the 'fc6' layer are used as face descriptor.

The performance of the proposed face recognition approach is assessed on three state-of-the-art face recognition datasets (LFW [9], IJB-A [10], and MegaFace [11]). To demonstrate the improvements due to the use of caricatures, we measure the relative performance between using caricatures and using original images for network training.

In summary, this paper has two major contributions: 1) a 3D-based caricature generation method for producing 2D caricatures that enhance the performance of face recognition; and 2) the first fully automated caricature-based face recognition approach capable of working in real-time with data acquired in the wild.

The remainder of this paper is organized as follows: Section II summarizes the most important approaches for generating caricatures from 2D images, and the works addressing caricature-based face recognition. Section III provides a detailed description of the proposed method. In Section IV, we discuss the obtained results and the conclusions are given in Section V.

## II. Related Work

The internal process behind recognizing faces has been studied extensively during the last decades [4] and several studies suggest that the brain encodes faces with respect to a general face prototype [12]. Also, for encoding, the brain emphasizes the most deviated physical traits and disregards average features, contributing to increase the inter-class separation while retaining the stability of intra-class separation. These results explain why humans can recognize better caricatures than veridical faces [7], [4] and indicate that, in fact, the brain encodes faces in a caricatured manner [13]. These findings evince that automated face recognition may also benefit from the use of caricatures. However, few works have exploited this idea, and to the best of our knowledge, this paper introduces the first fully automated caricature-based face recognition system. Below, we review the existing approaches for generating caricatures from 2D images, and caricature-based face recognition methods.

### A. Caricature Generation

Creating face models in a caricatured style is a popular topic in computer graphics and can be broadly divided in two families: 1) rule-based approaches; and 2) example-based approaches.

Rule-based approaches amplify the divergence between a probe face and a reference face by modifying the point-to-point distance of a set of fiducial points marked on both images. The first representative work of this family used 165 feature points to control deformations [14]. Liao et al. [15] introduced an automated caricature generation method that detects and analyzes facial features without human assistance. In [16], the normalized deviation from the average model was used to exaggerate the distinctive features, while Tseng et al. [17] used the inter and intra correlations of size, shape and position features for exaggeration. These works are 2D-based and most of them provide semi-automated systems that depend on user input to define the regions to be deformed. With the advent of 3D face databases, 3D-based caricature generation became the most popular approach. Lewiner [18] introduced an innovative 3D caricature generation tool by measuring the face deviations in the harmonic space. Clarke et al. [19] proposed an automatic 3D caricature generator based on a pseudo stress-strain model for representing the deformation characteristics at each feature point. Sela et al. [20] introduced a general approach for deforming surfaces based on the local curvature.

Data-driven approaches learn a mapping between the features of original face images to its corresponding caricature [21]. Liu et al. [22] proposed a machine learning method to map 2D feature points detected in face images to the coefficients of a PCA model learnt from a dataset of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



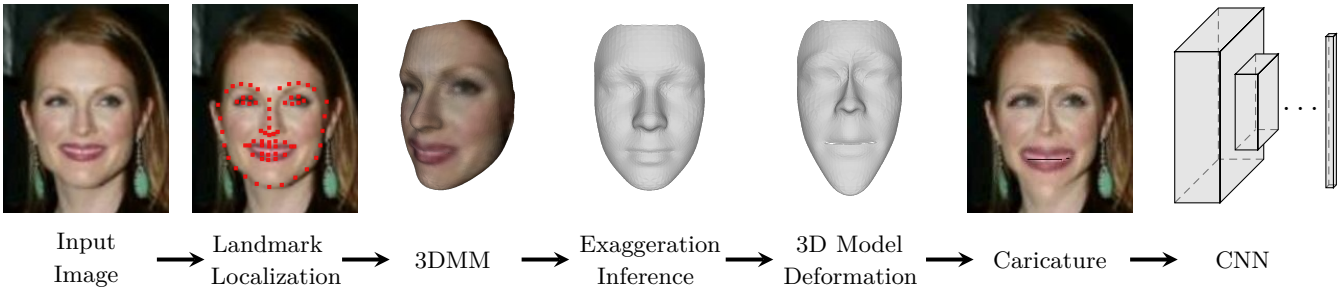


Fig. 2. **Overview of the processing chain of the proposed method.** The 3D face structure of probe images is inferred by a 3DMM method coupled with a set of automatically detected facial landmarks. This three-dimensional model permits the replication of the caricature drawing process by: 1) measuring the deviation of face regions to a reference prototype; and 2) using a 'measure locally, weight globally' strategy for inferring the exaggeration of each region. Using the modified regions as constraints, the original mesh is deformed with the Laplacian mesh editing algorithm, and the 2D caricature is obtained by projecting the deformed model in the original camera-view. At the end, the caricature is passed through a CNN to obtain a caricature-based face descriptor.

200 3D caricature models. Han et al. [23] introduced the first system capable of creating 3D face caricatures from 2D sketches by training a CNN with 2D sketches and the corresponding subspace of 3D shape and expression variations. For a detailed description of automated caricature generation refer to the survey of Sadimon et al. [24].

### *B. Face Recognition*

The performance of face recognition in the wild has significantly increased, mainly due to the advent of deep learning [25]. Nevertheless, the majority of face recognition approaches focused on improving performance via new learning strategies, augmenting training data or learning an embedding in the descriptors space, instead of adjusting the input data to a more suitable representation to address this problem (e.g., using face caricatures). Regarding caricature-based face recognition, there is limited work in the literature. Klare et al. [26] used qualitative features from face images and the corresponding caricatures to train a logistic regression model that predicted the similarity score between a caricature and a photo. However, these features were manually annotated via Amazon's Mechanical Turk, restraining the usability of this approach in a real-world scenario. Abaci and Akgul [27] proposed a method to automatically extract facial attributes from photos, but the attributes of caricatures were manually labeled. On contrary, Ouyang et al. [28] introduced a completely automated approach to match photos with caricatures by using a classifier ensemble for estimating facial attributes in both domains.

## III. Proposed Method

For comprehensibility, we use the following notation: matrices are represented by capitalized bold fonts, vectors appear in bold, and subscripts denote indexes. The proposed method is divided in six main phases, which are depicted in Fig. 2 and define the structure of this section.

### *A. Landmark Localization*

The localization of facial landmarks is a key step in the 3DMM phase of our approach. Besides, spurious landmarks affect significantly the likeness of the caricature, as the inferred 3D face structure does not portray correctly the facial features of the subject. Despite the astonishing increase in performance of landmark localization algorithms, the localization of landmarks in totally unconstrained data remains an open problem. For this reason, we combine $k$ state-of-the-art landmark localization algorithms [29], [30], [31], [32] in an ensemble strategy for predicting the most accurate set of landmarks obtained from these methods. Let $\mathbf{q} = [x_1, y_1, \cdots, x_\nu, y_\nu]^T$ be a vector with the locations of $\nu$ face landmarks in a 2D image, and $\mathbf{Q} = [\mathbf{q}^{(1)}, \cdots, \mathbf{q}^{(k)}]$ the matrix with the locations of the facial landmarks of $k$ distinct landmark localization methods. Assuming that the $k$ landmark localization methods produce uncorrelated outputs, the way they correlate in a particular image may provide insight about the correct set of landmarks, i.e., methods producing landmarks in very close locations are more likely to be correct. Accordingly, the output of the landmark localization algorithms is used to obtain $\mathbf{Q}^{(i)}, i \in \{1, \cdots, N\}$ for $N$ annotated images, and for each image, the vector $\hat{\mathbf{y}}^{(i)} \in \{0,1\}^k$ is determined by:

$$\hat{\mathbf{y}}_j^{(i)} = \begin{cases} 1 & \text{if } \frac{|\mathbf{q}^{(j)} - \mathbf{g}^{(i)}|}{d^{(i)}} < \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathbf{g}^{(i)}$ and $d^{(i)}$ are the manually annotated landmarks, and the inter-ocular distance for the $i^{th}$ image, respectively, whereas $\epsilon$ is a hard threshold controlling the maximum amount of inter-ocular distance that a set of landmarks $\mathbf{q}$ can differ from the ground truth. Each binary vector $\hat{\mathbf{y}}^{(i)}$ denotes the methods that produced the correct landmarks for the $i^{th}$ image, and the vectors of all training images are used to infer the function $\Psi : \mathbb{N}^{k \times n} \mapsto \{0,1\}^k$ by minimizing the following loss function:

$$\sum_{i=1}^{N} \left\| \Psi(\mathbf{Q}^{(i)}; \mathbf{W}) - \hat{\mathbf{y}}^{(i)} \right\|_2, \tag{2}$$

where $\mathbf{W}$ are the weights of the neural network used for inference. Given a probe image and the respective landmarks of the $k$ landmark localization methods, $\mathbf{y} =$

$\Psi(\mathbf{Q};\mathbf{W})$ provides the likelihood of each method being correct, and we choose the landmarks of the method with maximum likelihood.

### B. 3D Morphable Model

Blanz and Vetter introduced the 3D morphable models for the synthesis of 3D faces [33]. The main insight behind this approach is assuming that any face can be constructed using a linear combination of $M$ registered face models. A face is represented by a vector $\mathbf{s}^{(o)} \in \mathbb{R}^{3N}$ and a vector $\mathbf{t}^o \in \mathbb{R}^{3N}$, containing the $x$, $y$ and $z$ components of the shape, and the RGB color information, respectively. $N$ is the number of mesh vertices. Considering the correlation between the components of $\mathbf{s}^o$, each face is actually represented in a more compact version using the principal components (PC) of the shape and texture space, denoted by $\mathbf{s}$ and $\mathbf{t}$, respectively. Given a set of shape exemplars $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_M\}$ and texture exemplars $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_M\}$, a new fitted model ($\mathbf{s}_f$,$\mathbf{t}_f$) is expressed as:

$$\mathbf{s}_f = \sum_{i=1}^{M} \alpha_i.\mathbf{s}_i + \sum_{j=1}^{K} \lambda_j.\mathbf{b}_j \qquad \mathbf{t}_f = \sum_{i=1}^{M} \beta_i.\mathbf{t}_i, \quad (3)$$

where $\alpha$ and $\beta$ are vectors with the weights assigned to each exemplar, whereas $\mathbf{B} = \{\mathbf{b}_1, ..., \mathbf{b}_K\}$ is a set of deviations components of $K$ different facial expressions and $\lambda$ is the vector with the weight of each expression.

Given this, the estimation of the 3D face surface corresponds to the inference of the variables $\alpha$, $\beta$ and $\lambda$. Different strategies have been proposed for this purpose [34], [35], but in our case we opt for using the method of Huber et al. [36] available in the eos library [1]. This methodology works in two steps. The first step infers the camera matrix $\mathbf{P}$ that matches the pose of an average 3D model with the observed pose in the image, whereas the second step adjusts $\alpha$ and $\lambda$ to recover the shape of the 3D model. While $\alpha$ provides information about the face structure, $\lambda$ describes the expression observed in the current face image. The inference is carried out by minimizing the following energy function:

$$E \quad = \sum_{k=1} |\mathbf{q}_k - \mathbf{p}_k|, \quad (4)$$

where $\mathbf{q}$ is a set of 2D landmarks, and $\mathbf{p}_k$ is the projected position of the vertex corresponding to the $k^{th}$ landmark determined by $\mathbf{p}_k = \mathbf{P}.\mathbf{s}_f$. As described in [36], the cost function in (4) can be brought into a standard linear least squares formulation, allowing to recover both $\alpha$ and $\lambda$. Regarding $\mathbf{t}_f$, it is automatically obtained from the original image, making unnecessary the inference of $\beta$. Considering that non-neutral facial expressions affect the exaggeration inference phase and produce distinct caricatures for the same individual, we reset the $\lambda$ vector to guarantee that the fitted model is always in a neutral expression.

[1]https://github.com/patrikhuber/eos

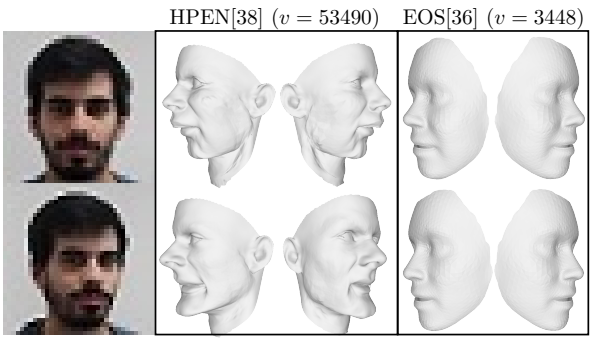


Fig. 3. **Examples of 3D models obtained by different 3DMM methods in low-resolution data.** The use of low-resolution data hinders the process of recovering the latent parameters of the 3D model, particularly when using dense models. The comparison between HPEN (a common 3DMM method coupled with a dense model) and EOS (3DMM particularly adapted for low-resolution data coupled with a sparse model) evidences two major drawbacks of the first approach: 1) the models do not correspond to the face structure of the subject; and 2) they are not consistent in data of the same individual.

In equation (4), the number of landmarks plays an important role in the detail of the recovered 3D model. However, in the case of low-resolution images, the correct localization of some face landmarks may be hard to judge, decreasing the usefulness of using a dense set of landmarks. For this reason, the use of a common subset shared by [29], [30], [31], [32] (68 landmarks) does not significantly impacts the detail of the fitted model.

Regarding the 3D model used, we opted for using the Surrey Face Model [36] (a sparse 3D model with 3,448 vertices) instead of the commonly used Basel Face Model [37] with 53,490 vertices. The rationale for this choice is twofold: 1) images acquired in totally unconstrained scenarios increase the likelihood of spurious landmarks, which in turn may induce aberrations in isolated parts of models with many degrees of freedom (as illustrated in Fig. 3); and 2) the computational cost of the minimization algorithm increases significantly with number of vertices, and the proposed method is intended for real-time applications.

### C. Exaggeration Inference

The correct assessment of which facial features should be exaggerated is the key for drawing recognizable caricatures. This process occurs internally in the human brain and is commonly accepted that is guided by the comparison to a reference model [39] or average model [40], which in our case is the Surrey Face Model.

Let $\boldsymbol{\pi} = [x, y, \theta, s]$ be the attributes of a face region, where $(x, y)$ is the mass center in the frontal model version, $\theta$ is the region orientation in the xy-plane and $s$ the region size. $\mathbf{r} = [\boldsymbol{\pi}^{(1)}, ..., \boldsymbol{\pi}^{(n)}]$ is the vector obtained by concatenating the attributes of $n$ face regions, whereas $\mathbf{r}^{(f)}$ defines the concatenation of the regions of the reference model chosen for our experiments (described in section III-B). The comparison between an exemplar model and the reference model is determined by $\mathbf{r}' = \mathbf{r} - \mathbf{r}^{(f)}$, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



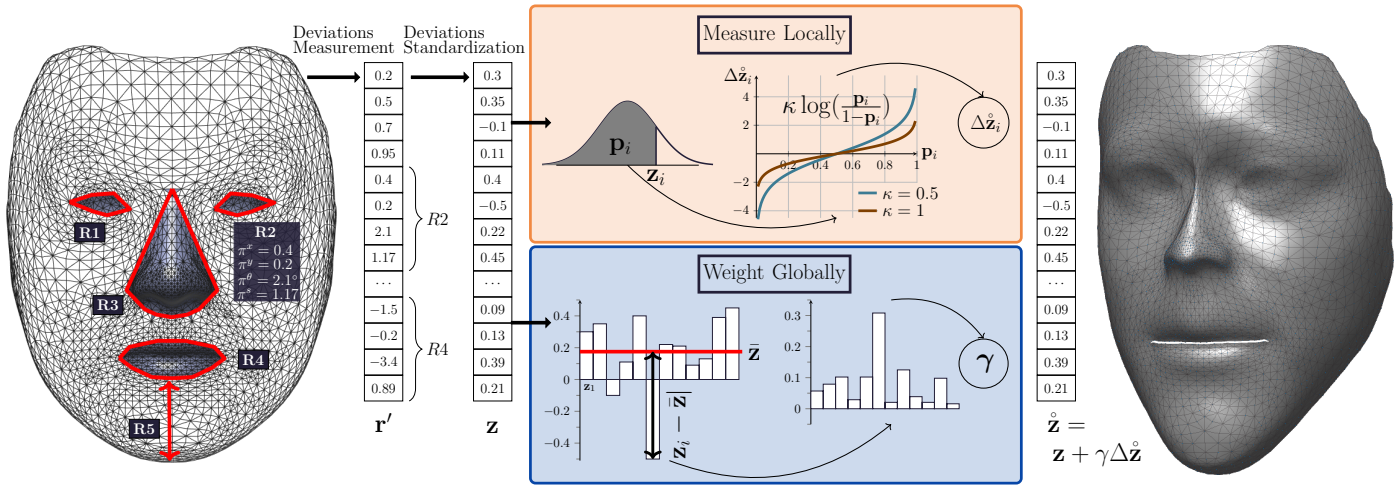


Fig. 4. **Schematic representation of the exaggeration inference phase.** The key for drawing recognizable caricatures is the correct assessment of the exaggeration degree that should be applied to each facial feature. Aiming at replicating the internal brain process that guides caricature drawing, we proceed by measuring the differences between the attributes of the inferred model and a reference model, followed by standardizing these deviations using z-score normalization. The normalized deviations are subsequently deformed using a 'measure locally, weight globally' strategy, allowing to determine the exaggeration degree of each attribute not only by its the individual deviation but also from its global importance in the face context.

element-wise difference between $\mathbf{r}$ and the corresponding regions of the reference model $\mathbf{r}^{(f)}$. The difference operator $\ominus$ between regions is defined as:

$$\boldsymbol{\pi}^{(1)} \ominus \boldsymbol{\pi}^{(2)} = \{x^{(1)} - x^{(2)}, y^{(1)} - y^{(2)}, \theta^{(1)} - \theta^{(2)}, \frac{s^{(1)}}{s^{(2)}}\}. \quad (5)$$

The vector $\mathbf{r}'$ is a compact description of the differences between a face and a reference model. However, each component is derived from attributes with distinct scales and variances. As such, we normalize $\mathbf{r}'$ using the standard score:

$$\mathbf{z}_i = \frac{\mathbf{r}'_i - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i}, \quad (6)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the sample mean and sample standard deviation of the $i^{th}$ attribute (estimated from the training data). This normalization provides a comparable description of how each attribute deviates from the mean.

We believe that a very similar representation is inferred internally by caricaturists [41], and that they exploit it for emphasizing the most distinguishable features of the whole face in a holistic manner, i.e., determine the exaggeration degree of each feature not only by its the individual deviation but also from its global importance in the face context. Inspired by this observation, we introduce a two-step process for inferring the exaggeration degree of the normalized deviation of each attribute.

The proposed inference strategy works in a 'measure locally, weight globally' manner. In the 'measure locally' phase, the exaggeration level of each region is determined without taking into account the exaggeration level of other regions. In the 'weight globally' phase, the exaggeration levels are weighted by their importance in whole face. The formal definition of each phase is provided below.

In the 'measure locally' step, the maximum displacement in the normalized space $\Delta\mathring{\mathbf{z}}$ is individually determined by applying a transfer function to the cumulative probability of $\mathbf{z}_i$ (denoted by $\Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)$):

$$\Delta\mathring{\mathbf{z}}_i = \kappa \log\left(\frac{\Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)}{1 - \Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)}\right) - \mathbf{z}_i, \quad (7)$$

where $\kappa$ is a parameter controlling the level of exaggeration applied to each attribute.

In the 'weight globally' step, the relative importance of each attribute is determined by measuring the absolute distance of $\mathbf{z}_i$ to the mean of the observed attributes ($\bar{\mathbf{z}}_\mathbf{i}$), and the weight of each attribute is given by:

$$\gamma_i = \frac{|\mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i}|}{\sum_{i=1} |\mathbf{z}_i - \bar{\mathbf{z}}_\mathbf{i}|}. \quad (8)$$

Both steps are then combined to produce the deformed deviation in the normalized space:

$$\mathring{\mathbf{z}}_i = \mathbf{z}_i + \boldsymbol{\gamma}_i \Delta\mathring{\mathbf{z}}_i. \quad (9)$$

Fig. 4 provides a summary description of the proposed deformation inference process.

Recovering the regions attributes from the deformed deviations $\mathring{\mathbf{z}}$ is attained by reversing the normalization process:

$$\mathring{\mathbf{r}}_i = (\mathring{\mathbf{z}}_i.\boldsymbol{\sigma}_i + \boldsymbol{\mu}_i) \oplus \mathbf{r}_i^{(f)}, \quad (10)$$

where $\oplus$ is the sum operator between regions. At the end, the 3D position of the region vertices is adjusted to comply with new region properties, i.e., regarding $\pi^{(x)}$, $\pi^{(y)}$, and $\pi^{(\theta)}$ the vertices are simply translated or rotated, whereas for $\pi^{(s)}$ the position of each vertex is adjusted by the vector $\pi^s(\mathbf{v}_i - \mathbf{v}_i^{(f)})$, being $\mathbf{v}_i$ and $\mathbf{v}_i^{(f)}$ the vertices of the $i^{th}$ region in the observed and reference model,

respectively. The updated 3D positions of the vertices of each region are then stored in **u** and used for model deformation.

### D. 3D Model Deformation

Given the updated positions of the face regions vertices, it is necessary to deform the mesh to satisfy these constraints. The deformation applied should, at the same time, comply with the constraints and preserve local details, i.e., produce smooth deformations by varying the position of each vertex with respect to its neighbors. Laplacian mesh editing [42], [43] is a classical algorithm to address this problem. In our approach, we used the implementation of the Laplacian mesh editing algorithm available in the Matlab Mesh Toolkit [2]. The model given in the right side of Fig. 4 depicts a deformed mesh obtained with the Laplacian mesh editing algorithm, where can be observed the smoothness of the deformation.

### E. Caricature Synthesis

The synthesis of the 2D caricature in the original pose is achieved by projecting each vertex with the camera parameters previously determined in the 3DMM phase. For maintaining the original image resolution, we adapt the number of vertices using interpolation.

### F. Feature Encoding and Matching

After generating the caricature from a veridical photo, the VGG-Face architecture is trained to identify individuals from caricatures (refer to Section IV-B for further details). Feature encoding is attained using the learned filters, and caricatures are described by the 4096-dimensional features produced in the 'fc6' layer of the VGG-Face architecture. During the matching phase, the L2 distance between the descriptors is used as dissimilarity score.

## IV. Results and Discussion

For experimentally validating our approach, we had four major goals: 1) evaluation of the performance of the landmark localization phase; 2) measure the running time of the proposed method; 3) determine the impact of spurious landmarks in the final recognition accuracy of our approach; and 4) assess the face recognition accuracy of caricature-based recognition when compared to the accuracy of using original photos. Regarding the first goal, the Annotated Facial Landmarks in the Wild [44] (AFLW) set was used to evaluate the results of the landmark localization phase. The VGG dataset [8] was chosen for its large quantity and diversity of face images (more than 2M images from 2622 celebrities), providing an excellent tool for tuning a CNN to the task of face recognition. Finally, the LFW [9], IJB-A [10] and MegaFace datasets were used for assessing the performance of our approach in data acquired in the wild. All these sets comprise images of celebrities, except for AFLW and Megaface, which contain images of Flickr users. Fig. 5 shows some images from the datasets considered for performance evaluation.

[2]https://www.dgp.toronto.edu/ rms/software/matlabmesh/

Fig. 5. **Examples of the datasets used in the empirical validation of the proposed face recognition method.** The upper row regards the LFW dataset, whereas the bottom rows are from the IJB-A and MegaFace sets.

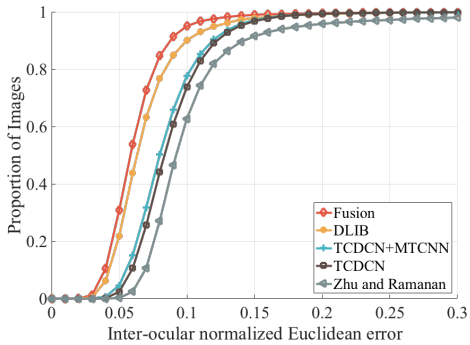


Fig. 6. **Cumulative error distribution curves for a subset of the AFLW dataset.** Four state-of-the-art landmark localization methods (DLIB [30],TCDCN [32],TCDCN+MTCNN [31] and Zhu and Ramanan [29]) and their fusion were evaluated in AFLW for evidencing the advantages of combining their results with an ensemble learning strategy.

### A. Landmark Localization

The AFLW dataset has 25,993 color images, each one annotated with a 21-point markup on visibility. This set was used to compare the performance of the ensemble learning strategy introduced in section IV-A with the individual performance of four state-of-the-art landmark localization methods [29], [30], [31], [32]. These methods are compliant with the popular 68 landmark format [45], while AFLW only provides a maximum of 21 landmarks depending on visibility. For evaluation, we selected a subset of 11 landmarks that share the same semantic positions in the two formats. Also, we considered exclusively samples with pose angles in the intervals yaw $\pm\frac{\pi}{4}$, pitch $\pm\frac{\pi}{2}$ and roll $\pm\frac{\pi}{5}$, according to the plausibility of observing such poses in visual surveillance scenarios. In accordance with the standard evaluation protocol [46], the average point-to-point Euclidean distance normalized by the inter-ocular distance was used as error metric, and the overall accuracy is reported by the cumulative errors distribution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



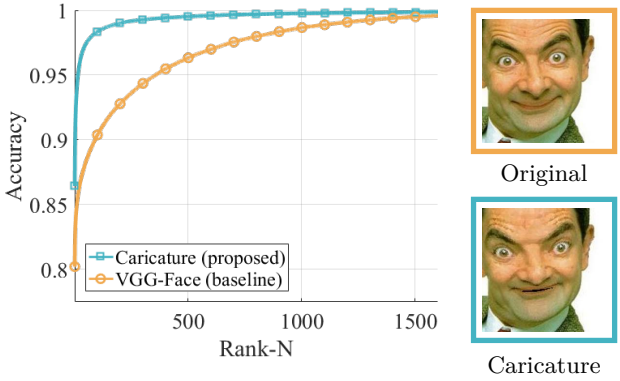


Fig. 7. **Comparison between the performance of the VGG-Face network trained on veridical images and on caricatures.** The improvements in the CMC curve of the network trained on caricatures evidence the benefits of using this representation for automated face recognition. This improvement is justified by the fact that caricatures enhance the distinctive features of subjects, easing the recognition task.

curve in Fig. 6. The comparative performance between the ensemble strategy and the best performing approach shows an increase of 5% in the proportion of images with an inter-ocular normalized error less than 10%. Even though these improvements seem irrelevant, they represent a significant decrease in the number of distorted caricatures caused by spurious landmarks, which in turn ease CNN training and improve the performance of the network.

### *B. CNN Training*

This section details the architecture, the parameters and the training data chosen for optimizing the network to the caricature recognition task. Our goal is to show that, similarly to humans, a CNN attains higher face recognition rates if trained with caricatures than with the original face images. The approach of Parkhi et al. [8] has been found particularly useful for this endeavor because of two major reasons: 1) the authors show that is possible to obtain state-of-the-art face recognition results on different datasets solely by training from scratch a CNN with millions of images automatically retrieved from the web; and 2) the network architecture (the VGG-Face) and the set of images used for training (the VGG dataset) are publicly available, allowing to replicate the experiments of [8], and measure the performance gap between the use of veridical photos and caricatures.

Accordingly, we trained the VGG-Face architecture from scratch on two distinct types of data: 1) original images of the VGG dataset (baseline); and 2) caricature images of the VGG dataset. The original VGG set contained 2.6M images (2622 identities with 1000 images), but at the time of our experiments, only 2.1M images were available on the web. Next, 90% of the images of each subject were randomly selected for training and validation, whereas the remaining were kept aside for performance evaluation. The configuration used and the regularization parameters for model optimization are described in Table I. The level of exaggeration controlling the caricature

TABLE I
TRAINING CONFIGURATION USED FOR ADJUSTING THE WEIGHTS OF THE CNN FROM SCRATCH.

| | |
|---|---|
| • Batch-size | 64 |
| • Momentum | 0.9 |
| • Weight-decay | $5 \times 10^{-4}$ |
| • Dropout Rate | 0.5 |
| • Learning Rate | $10^{-2}$ |
| • Weight Initialization | $X \sim \mathcal{N}(0, 10^{-4})$ |

generation phase ($\kappa$) was optimized using the validation set. For augmenting training data, a 224×224 pixel patch was randomly cropped from the image and horizontal flipping was applied with 50% probability. The model was implemented in the MATLAB toolbox MatConvNet and linked against the NVIDIA CuDNN libraries to accelerate training. All the experiments were carried on a NVIDIA Titan X GPU with 12GB of onboard memory, and each epoch took about 13h to run.

The comparative performance obtained by evaluating the trained models in 10% of the VGG set is depicted in Fig. 7. The results evidence the benefits of using caricatures for automated face recognition, and we argue that this improvement is justified by the fact that caricatures enhance the distinctive features of the subject, easing the recognition task. As an example, Fig. 7 also provides the two representations of an identity of the VGG set, where it is easier to identify the well-known actor Rowan Atkinson by its caricature than by its veridical image.

### *C. Running Time*

The average running time of the caricature generation phase is a crucial variable for two major reasons: 1) evaluating the applicability of the proposed method in a real-time system; and 2) determining the time required for generating the caricatures of the training set, which can be prohibitive in the case of VGG dataset (2.1M images). The extensive processing chain and the use of off-the-shelf implementations affect substantially the processing time, and, as such, some phases of the proposed approach were modified either by using approximations or memoization.

In the 3DMM phase, the maximum number of iterations for inferring the 3D model was changed from 50 to 5, as we noticed marginal differences in the obtained models. Regarding model deformation, the off-the-shelf implementation of the LME algorithm was optimized with memoization, while caricature synthesis was speedup by using triangulation hierarchy during texture rendering. Table II provides a comparison between the original and reduced average running time of each phase on a single core of an i7-4790 CPU, as well as, the total running time per image and the total time required for processing the whole VGG set. The results show that, after the optimization, training images can be generated in few days by distributing the data into multiple computers and exploit all the cores of the CPU.

TABLE II
COMPARISON BETWEEN THE ORIGINAL AND REDUCED RUNNING TIME OF THE PHASES OF THE PROPOSED CARICATURE GENERATION METHOD.

| Phase | Running Time (ms/img) | |
|---|---|---|
| | Original | Reduced |
| Landmark Localization | 160 ±120 | 160 ±120 |
| 3DMM | 800 ±93 | 200 ±52 |
| Face Analysis | 95 ±26 | 95 ±26 |
| 3D Model Deformation | 2 500 ±155 | 500 ±120 |
| Image Synthesis | 740 ±37 | 330 ±16 |
| Total | 4295 ±431 | 1290 ±334 |
| VGG | ≈ 114 days | ≈ 31 days |

TABLE III
**Impact of facial landmark accuracy on the performance of our approach.**

| Inter-ocular normalized Euclidean errors ($\xi$) | Rank-1 face recognition (%) |
|---|---|
| $\xi = 0$ | 86.44 |
| $\xi = 0.1$ | 79.47 |
| $\xi = 0.2$ | 71.63 |
| $\xi = 0.3$ | 55.68 |

### *D. Face Recognition Performance: Impact of Landmark Localization*

Considering that our approach is highly dependent on the performance of the landmark localization phase, we assessed the influence of landmark localization accuracy in the final recognition accuracy of the proposed approach in the test set of the VGG database (refer to section IV-B). For this evaluation, we adopted the strategy of Peng and Yin [47] where the locations of facial landmarks were corrupted by random noise generated from a normal distribution $\mathcal{N}(0, \sigma)$. In order to perceive the impact of landmark accuracy in the face recognition performance, the original set of landmarks used in our experiments was corrupted using different levels of noise. The noise level was controlled by $\sigma$ and adjusted to create four new sets of landmarks with distinct inter-ocular normalized Euclidean errors ($\xi$). The relation between rank-1 face recognition accuracy and landmark accuracy is provided in Table. III.

The relation between $\xi$ and the rank-1 face recognition accuracy evidences that our approach is highly sensible to incorrect landmarks ($\xi = 0.3$), but it can tolerate medium deviations ($\xi = 0.1$) to the correct facial landmark locations. Even though these results suggest that our approach can not operate in real world scenarios, where landmark localization is more likely to fail, it should be noted that only 5% of face landmarks have a $\xi >= 0.2$ (results obtained in a random sample of the VGG manually annotated). For this reason, we can conclude that current performance of landmark localization algorithms does not impact significantly the average recognition rate of our caricature-based face recognition in unconstrained data.

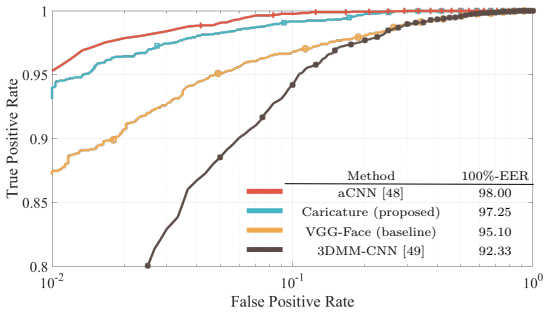


Fig. 8. **Face verification performance for the LFW dataset.** The ROC curves of caricature-based face recognition (our method) and original image face recognition (baseline) show the advantages of using this representation. Also, the results show that the obtained performance is competitive with state-of-the-art algorithms.

### *E. Face Recognition Performance: Improvements due to Caricatures*

To assess the performance of the proposed approach in highly unconstrained data, three state-of-the-art face recognition datasets were used, namely the LFW [9], IJB-A [10], and MegaFace [11]. The rationale for using multiple sets was twofold: 1) ensuring a non-biased evaluation of face recognition in the wild (the particularities of a single set could inadvertently overestimate the recognition rate of the proposed method); and 2) showing that the proposed approach can cope with large variations in data.

During encoding, the metadata of the evaluation sets were used to crop each probe image to a 256x256 sub-image containing the facial region and maintaining aspect ratio. Then, five patches of 224x224 pixels were sampled from each face image (from the four corners and center), and each region was duplicated with horizontal flipping. The ten resulting patches were subsequently input to the network and the obtained descriptors were averaged to produce the face descriptor of the probe image.

*1) Experiments on the LFW dataset:* LFW is a de facto benchmark for evaluating face recognition in the wild, comprising 13,233 images from 5,749 subjects. The evaluation protocol provides 3000 pairs of images organized in 10 splits for assessing the verification performance of face recognition algorithms. Also, each method should report the results under a specific setting with respect to the type of training data used. In our case, even though the descriptor obtained is not tuned in the LFW training data, we should report our results under the 'unrestricted with labeled outside data' setting due to the use of the VGG dataset during model training.

Regarding the comparison with state-of-the-art methods, the works of Tran et al. [49] and Masi et al. [48] were selected for sharing similarities with our approach. In [48], the 3D face structure was inferred from a single 2D image to augment the number of training samples by rendering the original face in a distinct pose, shape and expression. In [49], the authors introduced a regression network for estimating the 3D structure from a single 2D image and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



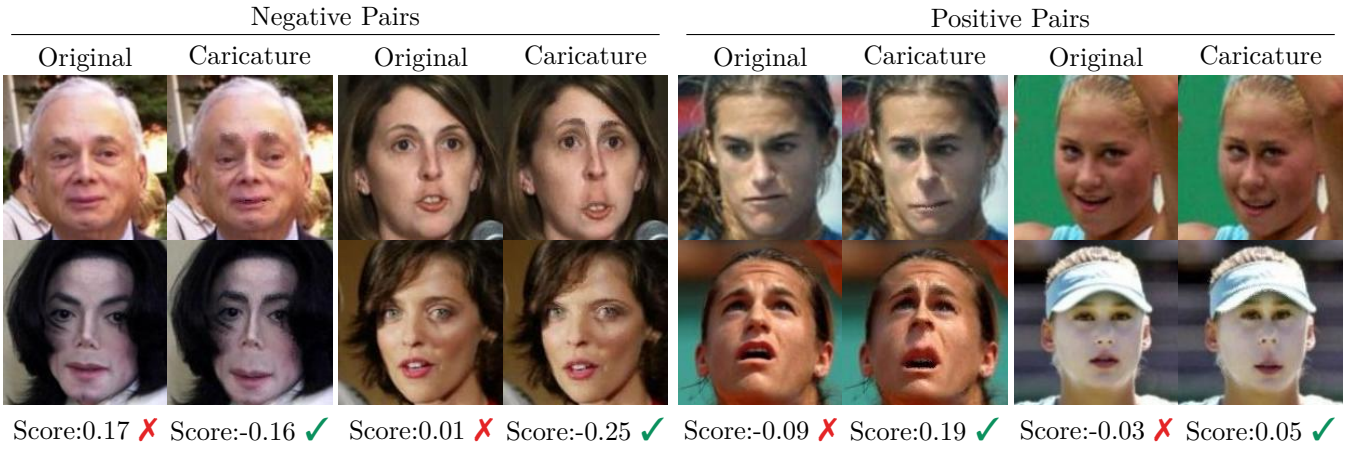


Fig. 9. **Successful cases of the proposed approach.** The advantages of using caricatures for face recognition are represented by four pairs of the LFW and IJB-A sets where our approach produced a correct score, while the use of veridical photos produced an incorrect output.

used this representation for face recognition.

Results are summarized in the Receiver Operating Characteristic (ROC) curves of Fig. 8. When compared to the baseline, our approach achieved a significant decrease in the EER, supporting the claim that automated face recognition benefits from the use of caricatures (refer to Fig. 9 for some examples). Regarding the comparison to similar approaches, our approach performed significantly better than 3DMM-CNN [49], while it produced competing results with respect to aCNN [48]. The results of [49] suggest that texture plays a decisive role in the face recognition task, while the comparison between the performance of our method with [48] indicates that the emphasis of distinctive facial features is as effective as generating multiple views of the original image.

*2) Experiments on the IJB-A dataset:* The IJB-A dataset represents an advance over LFW, by comprising data with a wider range of variations, particularly in pose. It contains 500 subjects with 5,397 images and 2,042 videos split into 20,412 frames, 11.4 images and 4.2 videos per subject. Regarding the evaluation protocol, it differs from LFW by considering template-to-template comparisons rather than image-to-image comparisons, where each template contains a combination of images or frames sampled from multiple image sets or videos of a subject. Algorithms can be evaluated in the verification (1:1 matching) or identification (1:N search) protocol over 10 splits. In the verification protocol, each split contains around 11,700 pairs of templates (15% positive and 85% negative pairs) on average, whereas the identification protocol also consists of 10 splits, each containing about 112 gallery templates and 1763 probe templates. During evaluation, each template is described by the average of image descriptors. Table IV reports the performance of the baseline, the proposed approach, and competing approaches with respect to the standard accuracy metrics of IJB-A.

TABLE IV
SUMMARY OF THE FACE RECOGNITION PERFORMANCE ON IJB-A.

| Method | Trained on IJB-A | Verification | | Identification | |
|---|---|---|---|---|---|
| | | FAR 0.1 | FAR 0.01 | Rank-1 | Rank-5 |
| GOTS [10] | Yes | 62.7 ±1.2 | 40.6 ±1.4 | 44.3 ±2.1 | 59.5 ±2.0 |
| OpenBR [50] | Yes | 43.3 ±0.6 | 23.6 ±0.9 | 24.6 ±1.1 | 37.5 ±0.8 |
| Wang *et* al. [51] | Yes | 89.5 ±1.3 | 73.3 ±3.4 | 82.0 ±2.4 | 92.9 ±1.3 |
| Chen *et* al. [52] | Yes | 96.7 ±0.9 | 83.8 ±4.2 | 90.3 ±1.2 | 96.5 ±0.8 |
| aCNN [48] | Yes | 88.6 | 72.5 | 90.6 | 96.2 |
| VGG-Face [8] | No | 85.4 ±1.2 | 61.1 ±2.3 | 87.6 ±1.6 | 92.8 ±0.9 |
| Caricature | **No** | **86.0 ±1.4** | **63.5 ±2.7** | **88.9 ±1.1** | **94.1 ±0.7** |

Regarding the comparison with the baseline, the improvements of our approach were not statistically significant, contrasting with the performance increase attained in LFW. In our view, the principal cause for this outcome was the failure of landmark localization, rather than the ineffectiveness of caricatures. The particularities of IJB-A (extreme variations in pose, face resolution, and illumination) affect significantly the accuracy of the landmark detector, which in turn distorts the generated caricature (Fig. 10 depicts some examples).

With respect to the comparison to other approaches, our method outperformed the baselines of IJB-A (GOTS and OpenBR), but it fell behind the remaining state-of-the-art face recognition methods. However, it should be stressed that, unlike the other approaches, no particular effort was made to optimize our method for this dataset (e.g., fine-tuning with IJB-A training data), since our major concern is measuring the relative performance between caricatures and original images.

*3) Experiments on the MegaFace dataset:* MegaFace [11] is a recent and very challenging dataset for evaluating face recognition at scale. The gallery set comprises more than 1 million images from 690K different individuals, while the probe set was sampled from the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2018.2846617, IEEE Transactions on Information Forensics and Security



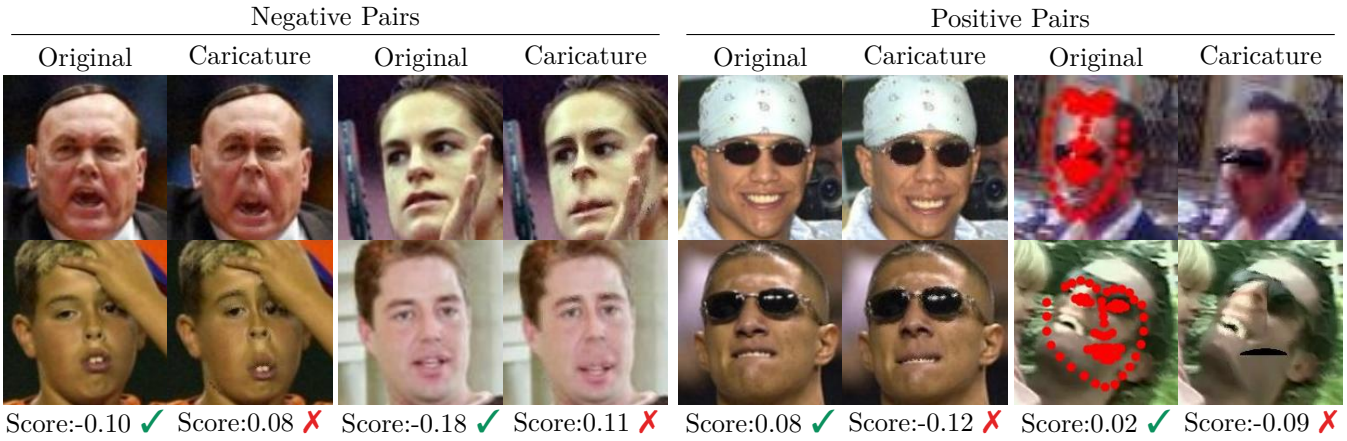


Fig. 10. **Failure cases of the proposed approach.** The major causes of failure in the LFW and IJB-A sets are represented by four pairs where our approach produced an incorrect score, while the use of veridical photos produced a correct output. The failure of landmark localization caused by occlusions, facial expressions and low-resolution data is the major reason for incorrect caricature generation.

FaceScrub dataset. The evaluation protocol provides code for assessing algorithms performance in the verification and identification scenarios.

Table V summarizes the methods performance according to the standard metrics of the dataset. The results show that the proposed approach outperforms the baseline on both recognition settings when using 1M distractors. However, the performance gap is smaller than the one observed in LFW. This can be explained by the extremely large gallery set (1 million identities), which may increase the likelihood of increasing the similarity between mismatch pairs. For example, consider two subjects sharing a salient facial feature (e.g., a big mouth), and diverging in facial features that highly diverge from the average face but represent smaller pats of the face (e.g., height of the eyes). In the case of low-resolution images, our approach would fail to highlight the smaller regions of the face, and it would simply exaggerate the mouth. At end, our approach would incorrectly increase the similarity between the two faces, and this is more likely to occur when a large gallery set is used. Regarding the comparison with commercial approaches and the baseline methods of MegaFace (LBP and Joint Bayes), it is interesting to note that, in the majority of the cases, caricature-based face recognition attained better performance than these systems.

TABLE V
SUMMARY OF THE FACE RECOGNITION PERFORMANCE ON MEGAFACE WITH 1M DISTRACTORS.

| Method | Rank-1 | TAR@FAR=$10^{-6}$ |
|---|---|---|
| Vocord-DeepVo1 | 75.13 | 67.32 |
| NTechLAB-facenx | 73.30 | 85.08 |
| Shanghai Tech | 74.05 | 86.34 |
| Google-FaceNet v8 | 70.50 | 86.47 |
| Beijing FaceAll-Norm-1600 | 64.80 | 67.12 |
| SIAT-MMLAB | 65.23 | 76.72 |
| Barebones FR | 59.36 | 59.04 |
| VGG-Face (baseline) | 75.08 | 74.78 |
| Caricature (proposed) | **75.10** | **76.49** |

### *F. Caricature-based Face Recognition: Additional Insights*

The comparison between training the VGG model with or without caricatures evidenced the benefits of caricature-based face recognition. The obtained results validated experimentally the proposed approach, but they do not provide any insight about the reasons for this improvement. In our opinion, this transformation can be seen as a pre-processing step that simplifies the task of the optimization algorithm, either by allowing a faster convergence or easing the search for minima in the objective function. Considering that deep learning based approaches are supposed to infer the best features/transformations that maximize the performance on a specific task, one may question why the network could not learn directly from the raw data the caricature process. In our view, this transformation is significantly more complex than learning specific features for the problem of face recognition, and for this reason the caricature transform requires much more data to be successfully learned.

Another important topic that should be discussed is the choice of the exaggeration method, which is core of caricature generation. As stated in previous sections, we adopted a 'measure locally, weight globally' because we assume that this rule is the basis of the internal caricaturing process occurring in the brain. However, it should stressed that this assumption may be too simplistic, and for this reason some failure cases of our approach may derive from the proposed exaggeration scheme. Ideally, the exaggeration scheme should be learned automatically from the data. We hope that future works can exploit this idea for proposing new strategies that enable deep learning based approaches to implicitly learn the caricature transformation from raw data.

## V. Conclusion

In this paper, we introduced the first fully automated caricature-based face recognition approach capable of working in real-time with data acquired in the wild. A 3DMM method coupled with a set of automatically detected facial landmarks was used for inferring the 3D face structure of probe images. Next, the inferred model was compared to a reference prototype for determining the divergence between facial regions, and the exaggeration applied to each region was determined by a 'measure locally, weight globally' strategy. The modified regions were given as constraints to a Laplacian mesh editing algorithm for deforming the original mesh, and the 2D caricature was obtained by projecting the deformed model in the original camera-view. During the learning phase, the VGG-Face architecture was trained from scratch on 2.1M caricatures automatically generated from the VGG dataset, whereas classification was performed with the features from the 'fc6' layer.

To assess the advantages of using caricatures for automated face recognition, we used three state-of-the-art face recognition datasets for measuring the relative performance between our approach and the VGG-Face trained on the original images of the VGG dataset. The results revealed significant improvements in the recognition performance when using caricatures rather than veridical images, confirming the usefulness of using caricature-based face recognition. Regarding the comparison with state-of-the-art methods, our approach was capable of obtaining competitive results even without being particularly tuned for any of the evaluation sets. Nevertheless, it should be noted that our goal was not to attain the best performing results on these sets, but measure the performance gap between the use of caricatures and veridical images.

## Acknowledgment

This work is supported by '*FCT - Fundação para a Ciência e Tecnologia*' (Portugal) through the project 'UID/EEA/50008/2013', the research grant 'SFRH/BD/92520/2013', and the funding from 'FEDER - QREN - Type 4.1 - *Formação Avançada*', co-founded by the European Social Fund and by national funds through Portuguese 'MEC - *Ministério da Educação e Ciência*'. Also, we gratefully acknowledge the donation of the NVIDIA Titan X GPU used for this research made by NVIDIA Corporation.

## References

[1] A. Burton, S. Wilson, M. Cowan, and V. Bruce, "Face recognition in poor-quality video: Evidence from security surveillance." *Psychological Science*, vol. 10, no. 3, pp. 243–248, 1999.

[2] J. Kaufmann and S. Schweinberger, "Distortions in the brain erp effects of caricaturing familiar and unfamiliar faces." *Brain Research*, vol. 1228, pp. 177–188, 2008.

[3] M. Itz, S. Schweinberger, and J. Kaufmann, "Effects of caricaturing in shape or color on familiarity decisions for familiar and unfamiliar faces." *PLOS ONE*, vol. 11, no. 2, pp. 1–19, 2016.

[4] G. Rhodes, S. Brennan, and S. Carey, "Identification and ratings of caricatures implications for mental representations of faces." *Cognitive Psychology*, vol. 19, no. 4, pp. 473–497, 1987.

[5] K. Lee and D. Perrett, "Presentation-time measures of the effects of manipulations in colour space on discrimination of famous faces." *Perception*, vol. 26, no. 6, pp. 733–752, 1997.

[6] ——, "Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments." *Perception*, vol. 29, no. 11, pp. 1291–312, 2000.

[7] K. Lee, G. Byatt, and G. Rhodes, "Caricature effects, distinctiveness, and identification: Testing the face-space framework." *Psychological Science*, vol. 11, no. 5, pp. 379–385, 2000.

[8] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *Proceedings of the British Machine Vision Conference*, 2015.

[9] G. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images." in *Proceedings of the International Conference on Computer Vision*, 2007, pp. 1–8.

[10] B. Klare, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a." in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1931–1939.

[11] I. Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[12] V. Bruce, M. Burton, and N. Dench, "What's distinctive about a distinctive face?" *The Quarterly Journal of Experimental Psychology Section A*, vol. 47, no. 1, pp. 119–141, 1994.

[13] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about." *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[14] S. Brennan, "The dynamic exaggeration of faces by computer," *Leonardo*, vol. 18, no. 3, pp. 170–178, 1985.

[15] P. Liao and T. Li, "Automatic caricature generation by analyzing facial features." in *Proceedings of the Asian Conference on Computer Vision*, 2004.

[16] Z. Mo, J. Lewis, and U. Neumann, "Improved automatic caricature by feature normalization and exaggeration." in *ACM SIGGRAPH Sketches*, 2004, p. 57.

[17] C. Tseng and J. Lien, "Synthesis of exaggerative caricature with inter and intra correlations." in *Proceedings of the Asian Conference on Computer Vision*, 2007, pp. 314–323.

[18] T. Lewiner, T. Vieira, D. Martinez, A. Peixoto, V. Mello, and L. Velho, "Interactive 3d caricature from harmonic exaggeration." *Computers and Graphics*, vol. 35, no. 3, pp. 586–595, 2011.

[19] L. Clarke, M. Chen, and B. Mora, "Automatic generation of 3d caricatures based on artistic deformation styles." *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 6, pp. 808–821, 2011.

[20] M. Sela, Y. Aflalo, and R. Kimmel, "Computational caricaturization of surfaces." *Computer Vision and Image Understanding*, vol. 141, pp. 1–17, 2015.

[21] P. Li, Y. Chen, J. Liu, and G. Fu, "3d caricature generation by manifold learning." in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008, pp. 941–944.

[22] J. Liu, Y. Chen, C. Miao, J. Xie, C. Ling, X. Gao, and W. Gao, "Semi-supervised learning in reconstructed manifold space for 3d caricature generation." in *Computer Graphics Forum*, 2009, pp. 2104–2116.

[23] X. Han, C. Gao, and Y. Yu, "Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling." *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[24] S. B. Sadimon, M. S. Sunar, D. Mohamad, and H. Haron, "Computer generated caricature: A survey," in *Proceedings of the International Conference on Cyberworlds*, 2010, pp. 383–390.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[26] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, "Towards automated caricature recognition." in *Proceedings of the International Conference on Biometrics*, 2012, pp. 139–146.

[27] B. Abaci and T. Akgul, "Matching caricatures to photographs." *Signal, Image and Video Processing*, vol. 9, no. 1, pp. 295–303, 2015.

[28] S. Ouyang, T. Hospedales, Y. Song, and X. Li, "Cross-modal face matching: beyond viewed sketches." in *Proceedings of the Asian Conference on Computer Vision*, 2014, pp. 210–225.

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild." in *Proceedings of the IEEE International Conference on Computer Vision*, 2012, pp. 2879–2886.

[30] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees." in *Proceedings of the IEEE International Conference on Computer Vision*, 2014, pp. 1867–1874.

[31] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[32] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 918–930, 2016.

[33] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces." in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[34] S. Romdhani and T. Vetter, "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior." in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 986–993.

[35] G. Hu, P. Mortazavian, J. Kittler, and W. Christmas, "A facial symmetry prior for improved illumination fitting of 3d morphable model." in *Proceedings of the International Conference on Biometrics*, 2013, pp. 1–6.

[36] P. Huber, G. Hu, R. Tena, P. Mortazavian, W. P. Koppen, W. Christmas, M. Ratsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework." in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance*, 2009, pp. 1–8.

[38] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 787–796.

[39] P. Rautek, I. Viola, and M. E. Groller, "Caricaturistic visualization." *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1085–1092, 2006.

[40] M. Roberts, "A unified account of the effects of caricaturing faces." *Visual Cognition*, vol. 6, pp. 1–42, 1999.

[41] J. Van Der Keyl and T. Cannon, *A Caricaturist's Handbook: How to Draw Caricatures and Master Exaggeration.* CreateSpace Independent Publishing Platform, 2010. [Online]. Available: https://books.google.pt/books?id=acgjkgAACAAJ

[42] Y. Lipman, O. Sorkine, D. Cohen-Or, D. Levin, C. Rossl, and H. P. Seidel, "Differential coordinates for interactive mesh editing." in *Proceedings of the International Conference on Shape Modeling and Applications*, 2004, pp. 181–190.

[43] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, and H. P. Seidel, "Laplacian surface editing." in *Proceedings of the Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, 2004, pp. 175–184.

[44] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization." in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2144–2151.

[45] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge." in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[46] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[47] Y. Peng and H. Yin, "Facial expression analysis and expression-invariant face recognition by manifold-based synthesis," *Machine Vision and Applications*, vol. 29, no. 2, pp. 263–284, 2018.

[48] I. Masi, A. Tran, T. Hassner, J. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proceedings of the European Conference on Computer Vision*, 2017, pp. 579–596.

[49] A. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing robust and discriminative 3d morphable models with a very deep neural network." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[50] J. Klontz, B. Klare, S. Klum, E. Taborsky, M. Burge, and A. K. Jain, "Open source biometric recognition." in *Proceedings of the IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[51] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1122–1136, 2017.

[52] J. Chen, V. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features." in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.