UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC)
BarcelonaTech

UNIVERSITAT DE BARCELONA (UB)

UNIVERSITAT ROVIRA i VIRGILI (URV)

MASTER IN ARTIFICIAL INTELLIGENCE

Thesis

Generative Video Face Reenactment by AUs and Gaze Regularization

Advisor
Dr. Meysam Madadi - Department Mathematics and Informatics, UB
Computer Vision Center, UAB

Author
Josep Famadas

Co-advisors
Dr. Sergio Escalera - Department Mathematics and Informatics, UB
Computer Vision Center, UAB
Cristina PALMERO, Universitat de Barcelona

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

FACULTAT DE MATEMÀTIQUES (UB)

ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA (URV)

October 15, 2019

Abstract

In this work, we propose an encoder-decoder-like architecture to perform face reenactment in image sequences. Our goal is to transfer the training subject identity to a given test subject. We regularize face reenactment by facial action unit intensity and 3D gaze vector regression. This way, we enforce the network to transfer subtle facial expressions and eye dynamics, providing a more lifelike result. The proposed encoder-decoder receives as input the previous sequence frame stacked to the current frame image of facial landmarks. Thus, the generated frames benefit from appearance and geometry, while keeping temporal coherence for the generated sequence. At test stage, a new target subject with the facial performance of the source subject and the appearance of the training subject is reenacted. Principal component analysis is applied to project the test subject geometry to the closest training subject geometry before reenactment. Evaluation of our proposal shows faster convergence, and more accurate and realistic results in comparison to other architectures without action units and gaze regularization.

Keywords: Encoder-decoder; Deep Learning; Face Reenactment; Facial Action Units; Gaze Regression

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Face reenactment, a technique that maps the facial performance of a source subject to the appearance of a target subject, has been studied during the last two decades [10, 8], but its popularity has recently increased due to the concept of *deepfakes*. Despite its controversy, this method has countless applications, ranging from video postproduction to virtual reality and virtual try-ons. For instance, in order to increase the feeling of presence in a virtual face-to-face conversation scenario, [17] proposed a method to remove the head-mounted display of the user by combining a 3D face capture of the user's visible face and the eye images captured by the headset. In the cinema industry, it has also been used to synchronize lips movement of an actor to match the dubbed soundtrack [4], and even to decrease movie production and special effects costs.

Existing works in the field of face reenactment are usually divided in two subgroups: facial synthesis by fitting a 3D face model and re-rendering the modified face features, and generative pixel-to-pixel approaches. Most of the existing techniques generate the final image by overlaying the source face, or parts of it, on top of the target's head [16, 14]. Such procedure is commonly performed by applying image transformations based only on face landmarks, resulting in the presence of unrealistic virtual artifacts due to head pose or facial expressions. Indeed, to ensure realistic target face images, face reenactment techniques have to account for coherent physical deformation of the face muscles and eye gaze movements. Furthermore, when applied to videos, such methods have to provide smooth transitions among contiguous frames in terms of facial expressions, facial geometry, head pose and gaze. Extreme head pose variations and varying illumination conditions represent further challenges to this research area.

In this work, we perform face reenactment in videos by combining both geometry and appearance, assuring temporal coherence, while generating realistic reenactments because of the face generation conditioned on action units (AUs) intensity and gaze vectors. To do so, we propose an encoder-decoder network that uses as input the previous generated frame image and the 2D landmarks of the current frame image. Estimated AU intensities and gaze vector of the source image are concatenated to the encoder-decoder's latent vector to improve realism and physical coherence of the target image. Results of the proposed method show faster convergence and more accurate lifelike images in comparison to baseline techniques that do not leverage action units and gaze regularization. A common challenge in deep learning approaches is the lack of generalization to unseen subjects specially when there is a small variability in the training data. To cope with this issue we also propose a face geometry-based data augmentation in training time and data normalization in test time. A qualitative result of our approach is shown in Fig. 1.1.

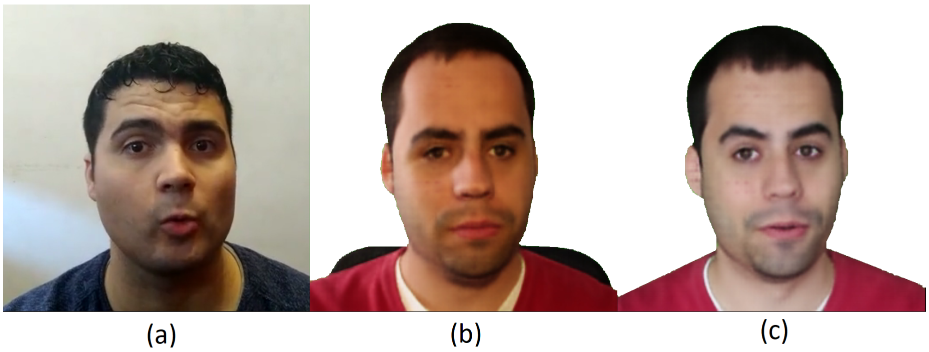


Figure 1.1: Our method generates the lifelike target subject face image (c) by replacing the identity of the source image (a) while maintaining head pose, eye gaze, and facial muscle deformations of the source image (e.g. AU02, AU23 and AU26, see Fig. 3.2). Image (b) is generated just based on input landmarks while image (c) is conditioned on AUs and gaze vectors as well.

The remainder of the thesis is organized as follows. Chapter 2 reviews the state of the art in face reenactment. Chapter 3 describes our generative video face reenactment by AUs and gaze regularization. Qualitative and quantitative results are presented in Chapter 4. Finally, Chapter 5 concludes the work.

1.1 Motivation

The motivation to do this master thesis in the field of video face reenactment is because of the following aspects: 1) the great progress in computer vision deep learning strategies that open the door for new high impact applications; 2) the potential applications of video face reenactment in leisure, such as virtual reality and film industry; 3) the current identified limitations of current state of the art solutions: they use to work in still images and not in an online fashion, generated reenactment results although look promising in terms of ”image quality” criteria es not consider subtly face behavior that can significantly improve the lifelike of the results, such as gaze behavior and fine-grain facial muscle dynamics.

1.2 Goals

The goal of this master thesis is to boost performance of current models for video face reenactment taking into account the limitations of state of the art approaches. This work aims to provide more realistic results in the following terms: smooth and real spatio-temporal reenactment so that generated videos looks natural; constraint the deep learning generation of videos not just in appearance but in face geometry; further regularize the reenactment generation conditioned to facial action units intensities and 3D gaze vectors from still images, so that the results will be more lifelike; finally since the model goal will be to reenact a source subject with a target subject (in order to have target subject identify with source subject face behavior), we further normalize source subject geometry to be approximated by the target subject. Last goal includes both quantitative and qualitative evaluation of the methodology, where for qualitative evaluation we ask the crowd.

Chapter 2

State of the Art

In this section we review in short Convolutional Neural networks and how are they used to build an encoder-decoder, which is the main architecture used in the project. Besides, we provide with a short review of related works in the field of face reenactment.

2.1 Theoretical Background

2.1.1 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are similar to standard Neural Networks in the point that they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and follows it with a non-linearity. The main difference is that the CNNs make the assumption that its inputs are sequence-like, i.e. a value of an input object is partially related to the near values on the same object. A clear example of that are images, where the value of a pixel i highly correlated with the value of its surrounding pixels. With this assumption, CNNs add a new type of layer, the convolutional layer. Each Convolutional Layer consists on a group of trainable filters (known as kernels) which are small spatially (along width and height), but extends through the full depth (Number of channels) of the input volume. The output of a Convolutional Layer is obtained by convoluting each of the kernels with the input volume. CNNs commonly have another type of layer, Pooling layer, which function is to reduce the high and width of the input volume by a factor of n. This allows the network to reduce the amount of output values until a point is reached where the output is enough small to be connected to a fully connected layer, as shown in Fig. 2.1 [1]. CNNs have its ”counterpart”, the Deconvolutional Neural Networks, which are similar to CNNs but instead of reducing the size of the input they increase it. Deconvolutional Neural Networks are commonly used in image generation.

[1] Source https://engmrk.com/convolutional-neural-network-3/

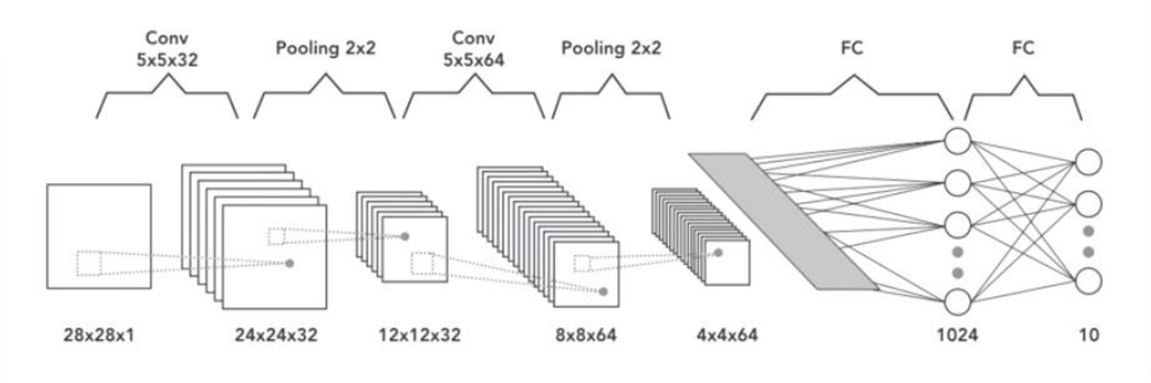


Figure 2.1: Convolutional Neural Network.

2.1.2 Encoder-Decoder

The encoder-decoder architecture is a neural network design pattern. In this architecture, the network is partitioned into two parts, the encoder and the decoder. The job of the encoder is, as its names indicates, to encode its input into some features, e.g. a text encoded as a vector representing its semantic meaning or an image encoded as a vector representing its content. The second part of the network is the decoder, which takes as input the encoder's output and decodes them to generate the output of the Encoder-Decoder, which can be the same as the input (auto-encoder) or not. Since in our case we focus on images, the architecture we will be using is a Convolutional Encoder-Decoder (CED), which consists on a concatenation of a convolutional neural network as the encoder and a deconvolutional neural network as the decoder. This architecture allows to generate an image conditioning it to another input image, as shown in Fig. in Fig. 2.2 [2]

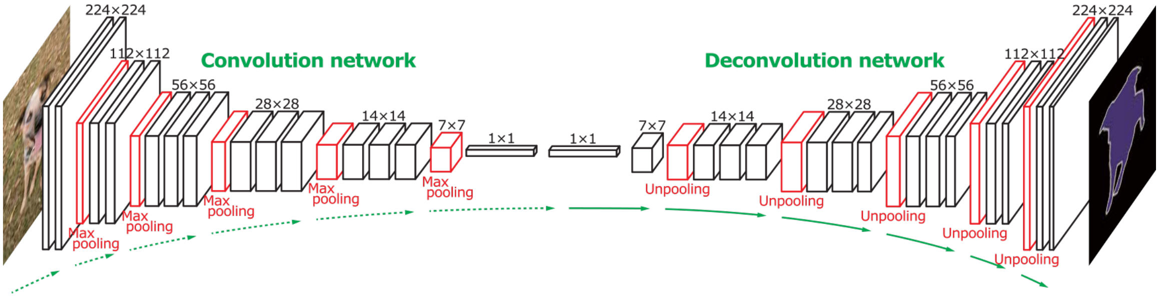


Figure 2.2: Convolutional Encoder-Decoder.

[2] Source https://medium.com/@wilburdes/semantic-segmentation-using-fully-convolutional-neural-networks-86e45336f99b

2.2 Previous work

In the last few years many works have attacked the problem of face reenactment. While this task has been approached in different ways, the state of the art techniques can be mainly grouped into two main categories of generative models: a) iterative parametric models or b) data-driven approaches.

2.2.1 Model based parametric approaches

Model based techniques aim to reconstruct the target's face by extracting independent parameters that describe the subject images, such as identity, position, expressions, etc. The job is achieved by computing the model parameters using the target actor and then deforming the latent sub-spaces that correspond to any of the characteristics of the source actor, to finally synthesize reenactment images. Parameter estimation in model based face reenactment is commonly done by energy minimization [17, 16, 15], where the energy function is defined as a sum of the dissimilarities between model parameters of the current image and the synthesised image.

2.2.2 Data driven approaches

Generative reenactment approaches are mainly referred to data driven approaches where a model is trained to learn a feature space that can be mapped to realistic images. These features can be learned directly from images or by the aim of additional information. They also can be supervised or unsupervised. Thanks to the advances in deep learning techniques and appearance of big datasets, quite realistic images can be generated by these approaches. The majority of these approaches are encoder-decoder networks that receive an input image and are conditioned to the modification parameters, such as facial expressions [13]. An example of generative face reenactment is the work of [5], where an avatar human agent dynamically interacts with users. In this work, facial expressions of a source user are translated to a sequence of affective sketch-GANs, which are used to generate the final photo-realistic interactive avatar.

2.2.3 Model based and data driven Approaches

To the best of our knowledge, the current state of the art in video face reenactment is the model proposed by Hyeongwoo et al. [7] which uses a combination of both model based and data driven approaches. In this model, a sequence of both target and source subjects is taken as input. For each frame of the sequence, a parametric model is fit to encode illumination and identity from the target actor, and pose, expression and eyes from the source actor. These parameters are used to render the synthetic images that are grouped by sequence and fed into a rendering-to-video network. A temporal window of size 11 is used in order to achieve temporal coherence. Finally these rendered videos are input into an encoder-decoder which generates the final output videos. Although generating realistic reenacted faces, the authors encode facial expressions as a coefficient of 64 different expressions. However, a richer representation could be obtained by facial action unit intensity regression, which would map to a higher number of different expressions.

Most of previous works base on global appearance for face reenactment, in some cases guided by facial landmarks. However, they do not consider subtle differences in Action Units intensity nor gaze vectors, which are determinant in order to produce a lifelike reenactment. In this work we propose a generative approach in which the inputs are detected via deep learning models rather than fitting a parametric model. Following a similar idea as the one proposed by Chan et al. in [3], where they teach an encoder-decoder to generate a human body from a set of key points representing its skeleton, we train an encoder-decoder to generate faces from images of the facial landmarks. We also condition our generation not only on the landmarks but also on the gaze and Action Units, obtaining lifelike reenactments.

Chapter 3

Methodology

Our proposed video face reenactment encoder-decoder model takes as input a video of a source person showing facial performances including head and gaze pose, and facial expressions. The goal is to replace the identity of the source person with the target one, while preserving the source facial performances. In this work we do not deal with the scene appearance and lighting of the source video, i.e. we do not force the model to preserve them in the target video. In our approach, the identity of the target person is always fixed as we have just one target person in the training set. Besides, our model can be applied online since the frames are processed individually and there is no need to process the whole video at once. However, to keep temporal coherence we feed to the model the previous generated face along with the current frame. The result is the reenactment sequence R. The model pipeline is shown in Fig. 3.1. Next subsection describes the encode-decoder architecture (Sec. 3.1), AUs and gaze regularization (Sec. 3.2), considered loss function (Sec. 3.3), and face geometry-based data augmentation and test subject normalization procedures (Sec. 3.4).

3.1 Encoder-decoder architecture

We build our model based on pix2pix [6] architecture where the landmarks vector is converted to an image before feeding it to the network. It is shown that decoding the output face from learned latent codes is more effective than decoding it directly from the landmarks vector, and the network can better deal with the facial geometry.

Although pix2pix architecture can deal with face geometry in single frames, it does not guarantee temporal coherence in the sequences of frames. Therefore, we feed the previously generated frame along with the current landmarks image. This is done by concatenating the images on the third dimension and adapting the first layer of the network. This helps to produce smooth transitions of the head pose and facial geometry while keeping lighting coherent among frames.

Architecture details: Input image has 256×256 pixels size. The encoder has 8 Convolutional layers with batch normalization and ReLU. Each layer has a 4×4 kernel which is applied with a stride of 2 after a 0-padding, forcing the output's width and height to be half the input's. Since with this architecture we already reach a middle latent layer of size 1 × 1 we did not add any pooling layer. The number of output channels for each layer is respectively [32, 64, 128, 256, 256, 256, 256, 256]. Decoder has 8 Deconvolutional layers with batch normalization and ReLU. Each layer has also a kernel of size 4 × 4 applied with a stride of 2 and 0-padding so the output's width and height is doubled after each layer. The number of output channels for each layer is respectively [512, 512, 512, 512, 256, 128, 64, 3]. We also add skip connections between corresponding layers of encoder and decoder.

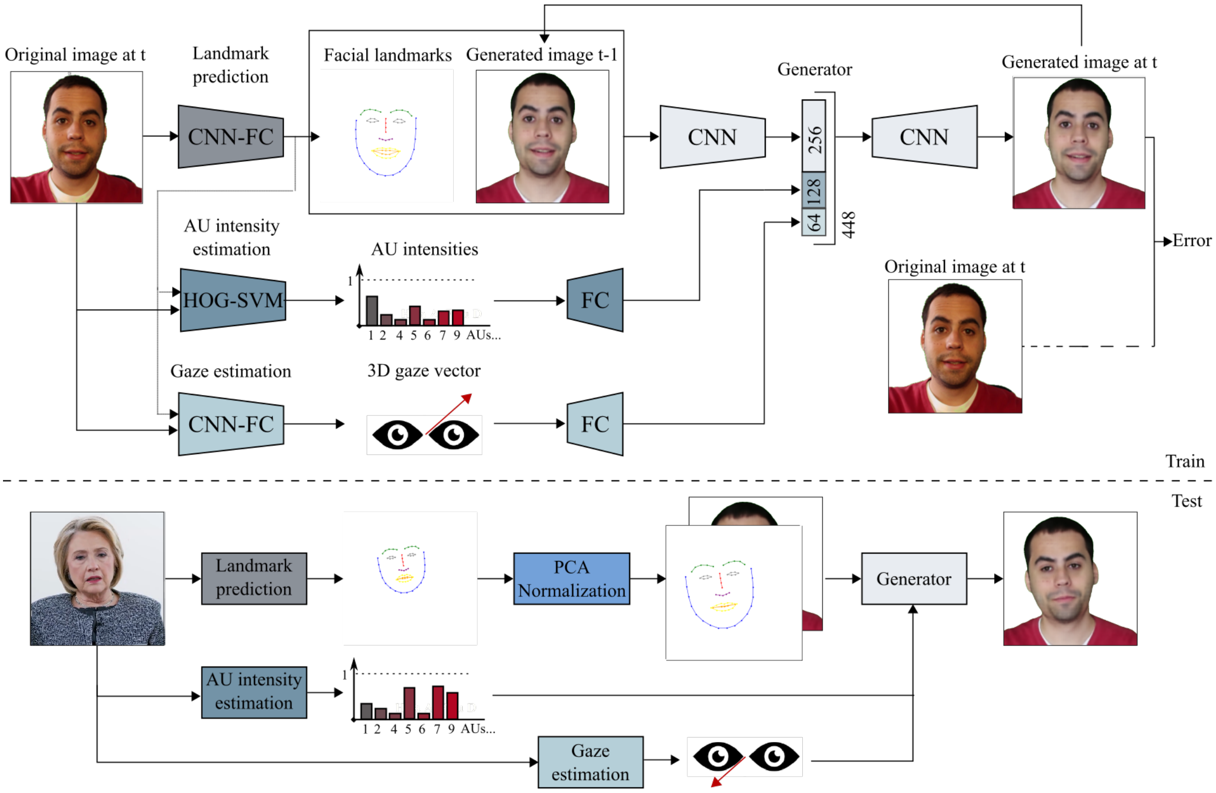


Figure 3.1: Proposed model for video face reenactment. An encoder-decoder CNN architecture receives facial landmarks of the source person stacked with the previous sequence generated frame, and outputs reenacted target face. The encoder-decoder network is conditioned on facial action units and gaze vector in the middle layer of the network. For a given test subject, facial landmarks are approximated to target face using PCA before reenactment generation.

3.2 Conditioning on the action units and gaze

Although landmarks can encode head pose, expressions and facial geometry and location up to a certain degree, they lose much information of the face due to the detection noise, sparsity and projection to 2D plane. To cope with this problem, we condition our model on the facial action units shown in Fig. 3.2[1], coding a diverse set of facial muscles deformation. AUs are base components of facial expressions and provide richer information than high level expression categories like happiness, sadness, etc. We specifically use AU intensities. Therefore we have a continuous space for each AU. To condition on AUs, we feed the estimated AUs vector of the source face to a 2-layer feed-forward neural network (NN) with 64 and 128 nodes per layer and ReLU after each one. We concatenate the output features to the middle latent features of the encoder-decoder network.

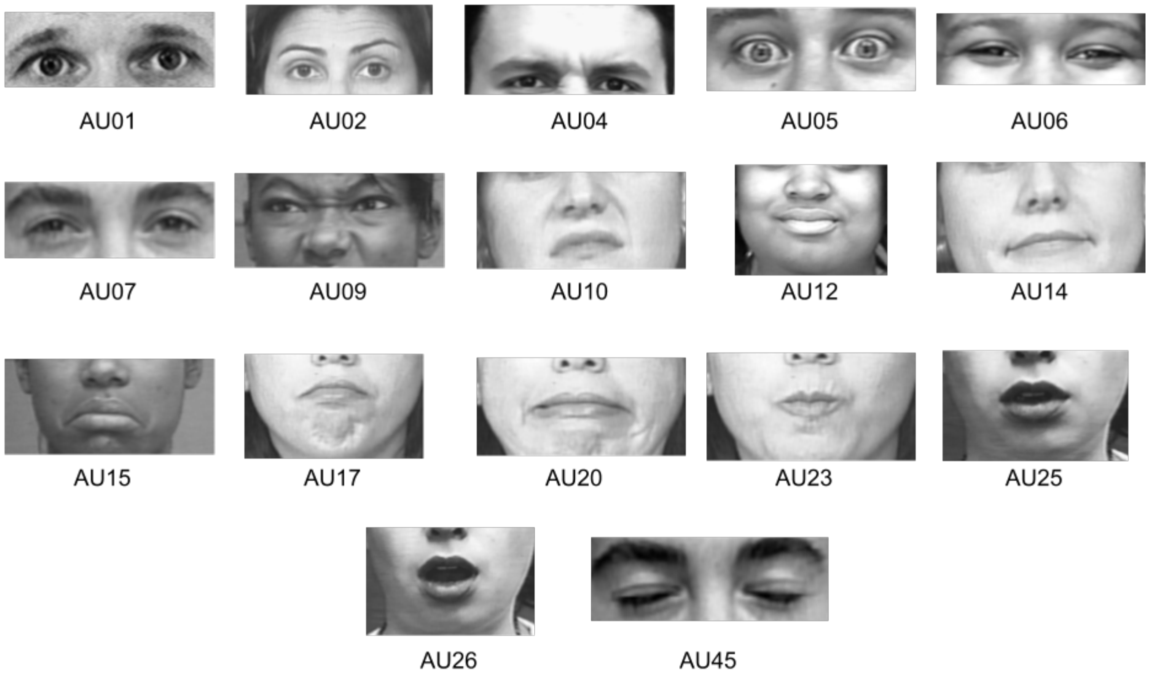


Figure 3.2: Action units used in this work.

Up to now our model can deal with face geometry and facial expressions. However, realism of a generated face is also dependent on the eye gaze.Therefore, similar to AUs, we feed the estimated 3D gaze vector of the source face to a 2-layer NN with 32 and 64 nodes per layer and ReLU after each one. We concatenate its output to the latent layer of the encoder-decoder network. Details of the models used to regress AUs intensity and 3D gaze vector are provide in the experimental section.

[1] Source https://www.cs.cmu.edu/ face/facs.htm

3.3 Loss function

The optimization is performed on a sum of different losses:

$$L = L_I + L_P + L_A + L_G, \tag{3.1}$$

where an equal balancing is applied over the loss functions. L_I is a regular L1 loss of the pixels between generated and ground truth image. L_P is a popular perceptual loss in image generation which is complementary to L_I, as it is known that pixel-wise L1 loss does not necessarily force the network to generate perceptually consistent images. We apply perceptual loss by applying the pre-trained model of [12], which is trained for face recognition task. To compute the loss, generated and groundtruth images are fed to the pre-trained network and a L1 loss is applied between the intermediate activations of the two inputs. L_A and L_G are AU and gaze losses, respectively, which are computed similar to perceptual loss but with the pre-trained networks used to estimate AUs and gaze. By using L_A and L_G, we regularize the network to generate perceptually coherent images w.r.t. AUs and gaze.

3.4 Geometry-based data augmentation and test subject normalization

In our model We benefit from the computed face landmarks to: a) perform a geometry-based data augmentation procedure to enhance model generalization capability; b) project the source subject geometry to the closest target geometry before reenactment generation. This way, we preserve target subject identity.

Geometry-based data augmentation. The training set consists of just one subject, our target person. Training the model with the landmarks of this subject leads to overfitting and may not generalize to unseen subjects with different face geometry. To cope with this issue, we apply an online data augmentation on the landmarks. We apply linear deformations, with a probability of 0.3, to detected 3D face landmarks. Concretely, increasing or decreasing the inter-eye distance, the face height/width ratio, the height of the nose and the size of face regions, i.e. eyes, nose and mouth. Some samples are shown in Fig. 3.3 after projecting augmented 3D landmarks to image plane.

Test subject normalization: During testing, we apply a reverse operation compared to the training augmentation, i.e. we want the landmarks of the source person to be as close as possible to the landmarks of the target person without modifying facial expressions. For this, we base on linear transformation from source to target landmarks. The process involves the following steps. First, landmarks of the target subject from training samples are aligned w.r.t. a reference viewpoint and average face is computed. Second, we compute PCA on this data. Third, the source face landmarks are aligned to the target average face through Procrustes analysis. Then, a subset of PCA components (8 out of 204 in our experiments) are used to project and back-project landmarks from the source to the target. Finally, reconstructed landmarks are transformed back to the source face using the

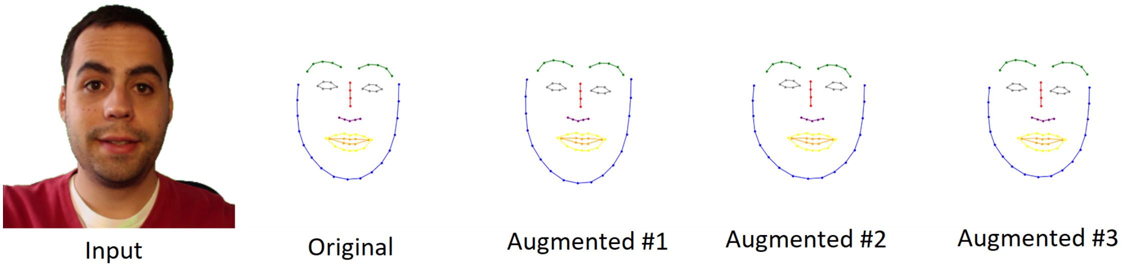


Figure 3.3: Random geometry-based augmentations on original landmarks.

previously computed rotation and translation. However we discard scaling to have the same head size as the target face. These new landmarks are used to generate the reenacted face. A sample of this source-to-target landmark approximation is shown in Fig. 3.4.

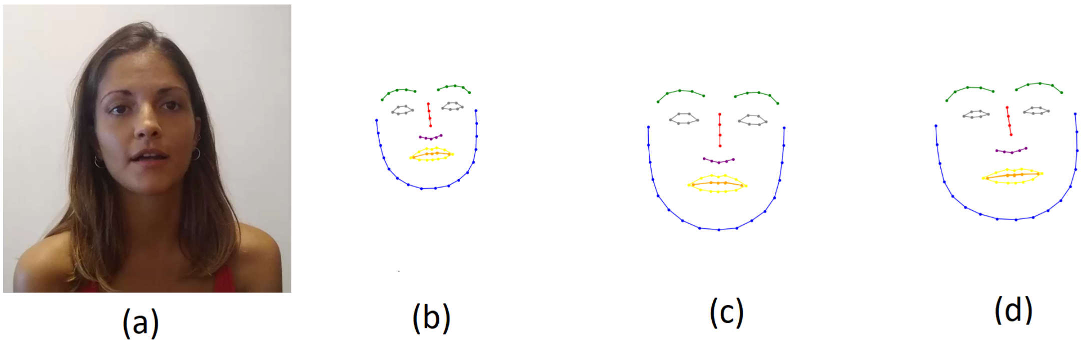


Figure 3.4: For a (a) source subject, (b) raw detected landmarks are (c) aligned to reference training landmarks. The aligned landmarks are (d) projected and back-projected using PCA, and the aligned transformation is undo.

Chapter 4

Experimentation

In order to present the results, first we provide implementation details of the model, the data used for the experiments, and the evaluation protocol. Then we present quantitative and qualitative results, including a Mean Opinion Score (MOS) evaluation.

4.1 Implementation details

We used OpenFace2.0 [1] for landmark detection and AU estimation. Landmark detection is based on Convolutional Experts Constrained Local Model (CE-CLM) [18], that consists of a Point Distribution Model (PDM) which captures landmark shape variations and patch experts, modeling local appearance variations of each landmark. AUs are regressed by a linear Support Vector Machine[2], where HOG features are combined with CNN features of the landmark detection module. Finally, we used a CNN-based model [11] to estimate 3D gaze vectors.

The architecture has been implemented using TensorFlow and trained on a *NVIDIA Quadro P6000* GPU (24 GB RAM). The selected optimizer was Adam, with a learning rate of 2e-4. The training was performed frame by frame, i.e. with a batch size of 1 for every train step. This was needed due to the fact that, for generating a new frame, the model needs the previous generated frame. We applied early stopping and our geometry-based landmark augmentation strategy.

4.2 Data

Training: To train the model, we gathered data from a YouTube channel where a person is talking in front of the camera. Therefore, the training dataset has just one subject. We asked the owner of the channel to perform some additional performance like variations in the head and gaze pose, including different facial expressions. The background is always fixed to white color. The data comes from different sequences, and despite all of them have the same conditions in terms of background and distance from the camera, the illumination varies from one to the other, as can be seen in Fig. 4.1. The used YouTube sessions where recorded in front of a Green screen. Background has been removed in our analysis to focus on the task of face reenactment. The data contains 39781 frames from 6 different sequences, which have been split into train (38000 frames) and validation (1781 frames). Landmarks, AUs and gaze are computed by the methods described in Sec. 4.1.

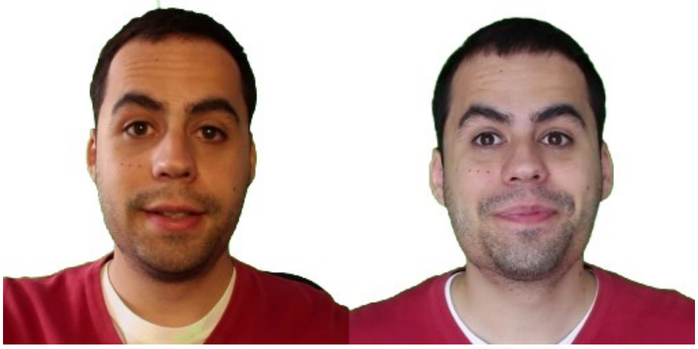


Figure 4.1: Difference in illumination from two different training sequences.

Testing: To test our model, we used sequences from different source actors, varying from gender, age and race, to prove our model can generate the image of the train subject (caucassian male, 24) independently of any of these factors. We show some statistics of the testing dataset in Tab. 4.1.

	Ethnicity	Gender	Age	# frames
Subject 1	Caucasian	male	24	429
Subject 2	Caucasian	female	23	1313
Subject 3	Afro-American	male	58	2883
Subject 4	Caucasian	female	71	1388

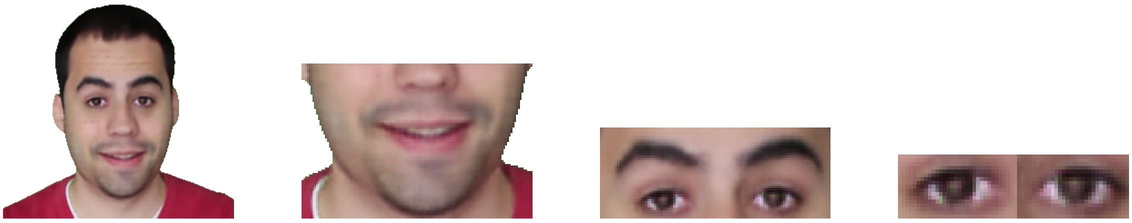
Table 4.1: Statistics of the test set. We recorded subjects 1 and 2. Subjects 3 and 4: Barack Obama and Hillary Clinton, videos gathered from white house weekly address and *NowThis* interview YouTube channels, respectively.

4.3 Evaluation protocol

We use different metrics in validation and test sets to evaluate our model. Given we have groundtruth target face in the validation set, we can evaluate our model quantitatively by the following metrics.

L1: Mean absolute distance in the color space between generated image and ground truth in order to compare the generated image pixel-wise.

DSSIM: Structural similarity index method (SSIM) [19], which measures the perceived similarity (dissimilarity in case of DSSIM) between two images. The index is calculated on



(a) Frame (b) Zone 1 (c) Zone 2 (d) Zone 3

Figure 4.2: Full grame image and the 3 sub-zones

various windows of the images. SSIM between a window x and a window y is measured by:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{4.1}$$

where μ_w is the average of w, σ_w^2 is the variance of w, σ_{xy} is the covariance of x and y, and c_i are variables to stabilize the division based on dynamic range of the pixel values.

Gradient difference - GD: Mean euclidean distance between the gradient of the generated and groundtruth images. This metric mainly focuses on the edges and is less affected by the illumination compared to L1 [9].

We further evaluate each metric in three different regions of the face apart of the whole image: Zone 1 the mouth and the chin, Zone 2 the eyes and eyebrows, and Zone 3 only the eyes, as shown in Fig. 4.2.

Since there is no groundtruth data in the test set, in order to evaluate the performance of our model in comparison to the alternatives without AUs and/or gaze regularization, we performed a Mean Opinion Score (MOS) evaluation. We used the same Youtube channel from the target subject to disseminate for MOS participation. The questionnaire has been answered by 564 people, being 83.5% men and 14.9% female, 92% caucassian and 8% afroamerican, and 40% from Spain and 60% from Latin-America. The generated reenactment sequences for the different models where simultaneously displayed for each test subject, but with a different random layout in order to avoid location voting bias. Ages of the participants were distributed as shown in Fig. 4.3. See Appendix A for MOS details.

4.4 Quantitative results

In order to evaluate our contributions, we trained our model with four different strategies: **BASE** without AUs nor gaze, **B-AU** with AUs combined in the middle layer, **B-GAZE** with gaze, and finally **B-AUGAZE** the full model combining both AUs and gaze.

During training, we observed the first 3 models (**BASE**, **B-AU** and **B-GAZE**) reached the minimum validation error after 40 epochs, while the proposed full model **B-AUGAZE** converged just after 25 epochs.

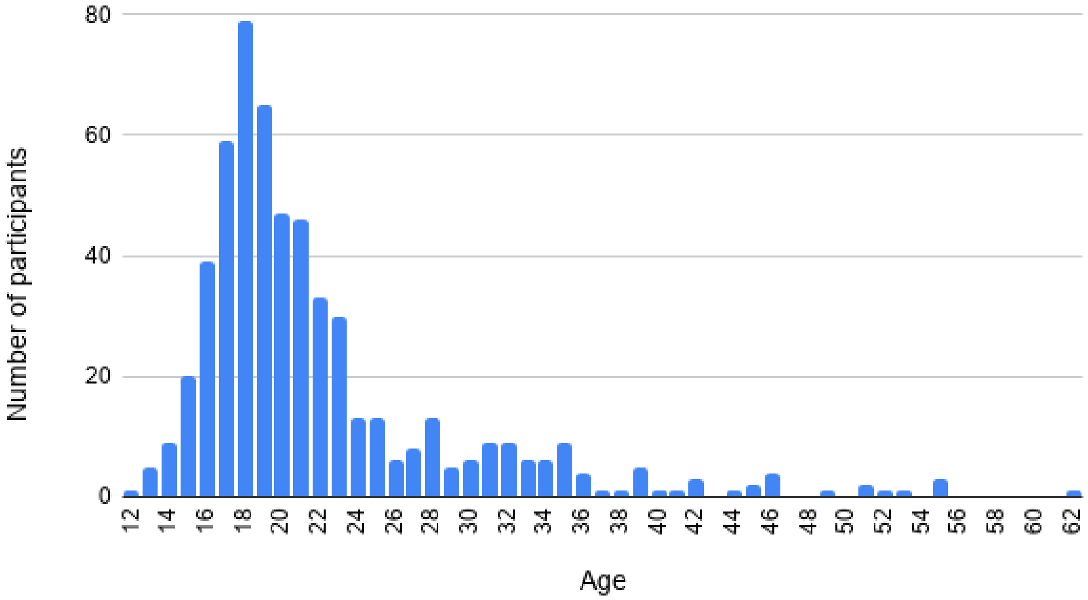


Figure 4.3: Age distribution of people who contributed in the MOS voting. A majority of voters are in the range 16-24.

The results for the three evaluation metrics (L1, DSSIM and GD) on the validation set are shown in Fig. 4.4. As it can be seen, B-GAZE has the highest error among all models and all zones. B-AU shows an improvement over the base model in almost all zones and for all metrics. One of the main reasons can be that the selected AUs covers whole face region, while gaze just impacts in a very small number of pixels. However, the combination of both AUs and gaze provides the best results, showing the benefit of the two complementary regularization sources.

On the L1 error we observe a clear difference between the B-AUGAZE model and the other 3. However, the pixel-wise error is not the best measure to compare images, specially if the generated images can have different illuminations.

In the case of the DSSIM, we can observe that for the bigger zones (full frame and Zone 1), the error gap of B-AUGAZE is smaller comparing to the smaller zones (Zones 2 and 3). This is due to the fact that in the smaller zones the details have much more relevance. This behaviour can be seen in the GD errors as well. One can observe that GD error has the reverse relationship with the size of the zones. This may be because gradients are less influenced by global illumination changes.

4.5 MOS evaluation

We performed a mean opinion score evaluation to ask for the lifelike perception of the image sequences generated by the four different models. We asked participants to sort the 4 generated videos (one video per test subject) from the most realistic one to the least realistic one. From these ranks, we computed the mean rank for every combination Subject-Model. The results are shown in Tab. 4.2. All values are in the range $[1, 4]$, being 1 the best value and 4 the worst. B-AUGAZE achieves the best position for all test subjects. In average (last row in Table 4.2), B-AUGAZE is selected as the first choice, followed by B-AU, B-GAZE, and finally the BASE model.Interestingly, B-GAZE has a better rank among subjects (2.88 in avg.) compared to the BASE model (3.03 in avg.). However, the quantitative results (Fig. 4.4) shows a slightly better performance of the BASE model in comparison to B-GAZE. This disagreement between quantitative and qualitative results reinforce the idea that new automatic metrics of human perception are needed to evaluate lifelike generated images.

	BASE	B-AU	B-GAZE	B-AUGAZE
Subject 1	2.93	2.41	3.13	1.51
Subject 2	2.57	2.76	2.71	1.94
Subject 3	3.33	2.00	3.01	1.63
Subject 4	3.32	2.16	2.66	1.83
Mean	3.03	2.33	2.88	1.73

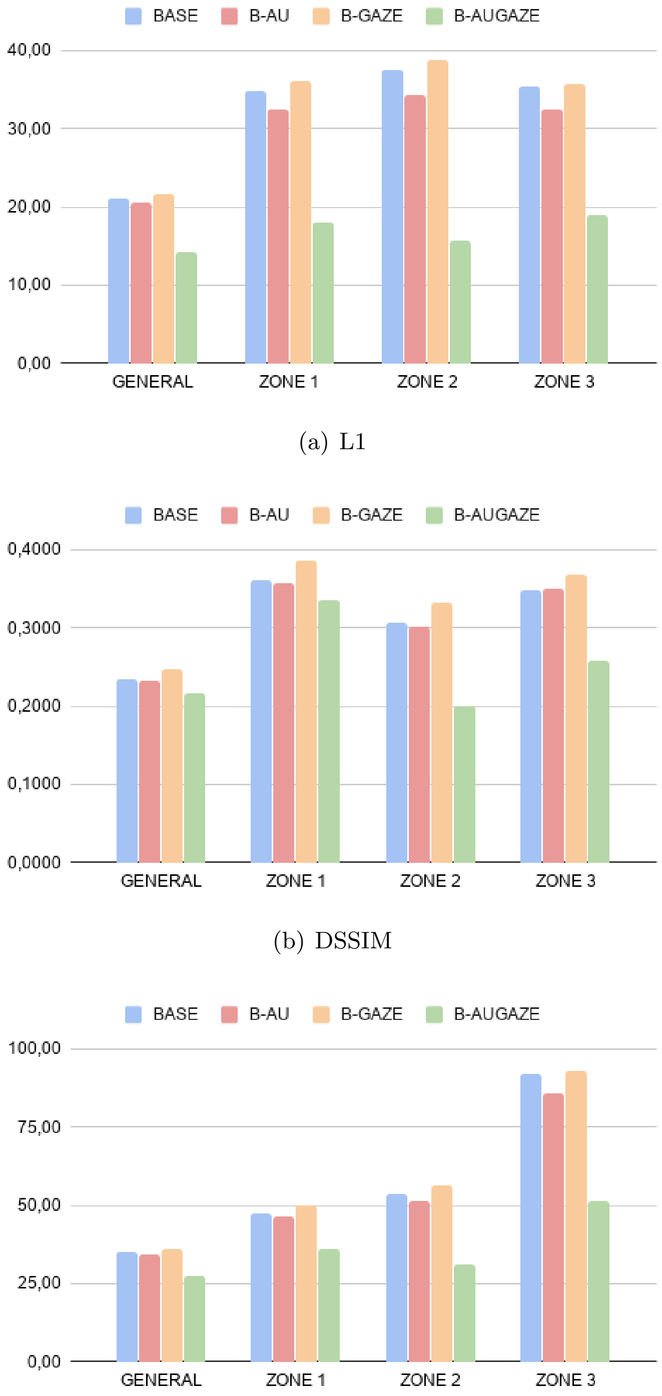
Table 4.2: Average rank for each test subject from MOS evaluation.

4.6 Qualitative results

In this section we analyze the results qualitatively. In Fig. 4.5 one can see that due to the fact that the training data has different illumination sequences, the first 3 models converged to generate darker images than B-AUGAZE. B-AUGAZE results show more realistic faces in terms of global shape, appearance, facial expressions and gaze than the models without including both AU and gaze regularizers.

Results obtained by B-AUGAZE for near-consecutive frames for 2 sample sequences are shown Fig. 4.6. One can see that the head pose and facial expressions of the test subject are accurately translated to the target subject without the appearance of any artifact, producing a natural spatio-temporal transition, and preserving the same illumination along the sequence.

In Fig. 5.1 we show how the PCA normalization, explained in Sec. 3.4, helps when having a subject with different face geometry to the target subject. In the first case, source subject has bigger eyes than the ones of the target subject, which generates a bizarre output image with too big eyes. Our model generates the target subject constrained to the source subject geometry. PCA normalization allows to adapt target-to-source geometry while keeping the expressions of the source image, such as the eyes aperture in this case. In the second case of the figure, the source subject has an eyebrow shape much different from the target's eyebrow shape. If we compare the generated images with and without PCA the difference is clear. Finally, Fig. 5.2 shows the application of our AUs and gaze constrained face reenactment model to 2 avatars with non-ordinary face geometry, different head poses and facial expressions.



(a) L1

(b) DSSIM

(c) GD

Figure 4.4: Validation error for the different models on the different face regions.

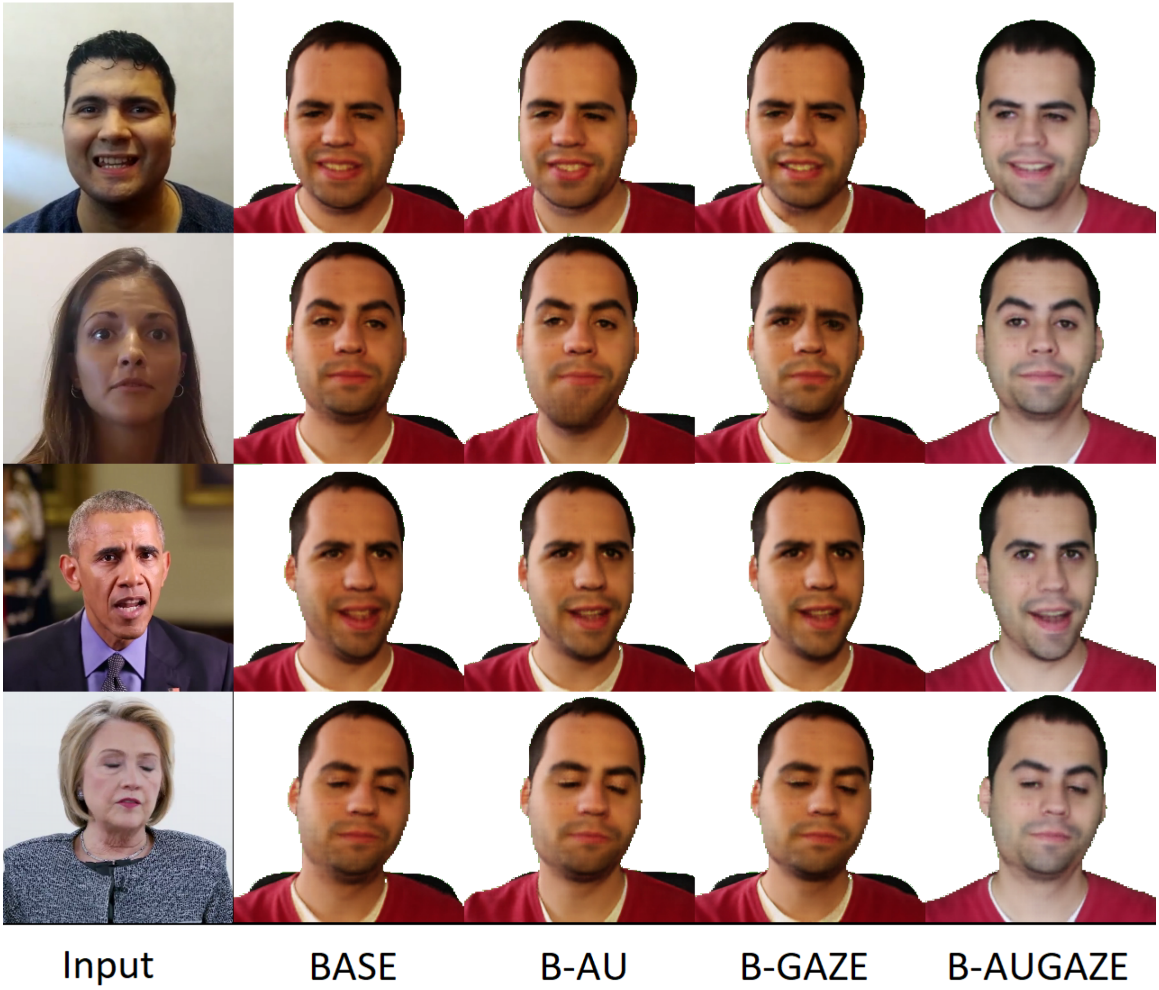


Figure 4.5: Samples generated for the different trained face reenactment models for different test subjects.

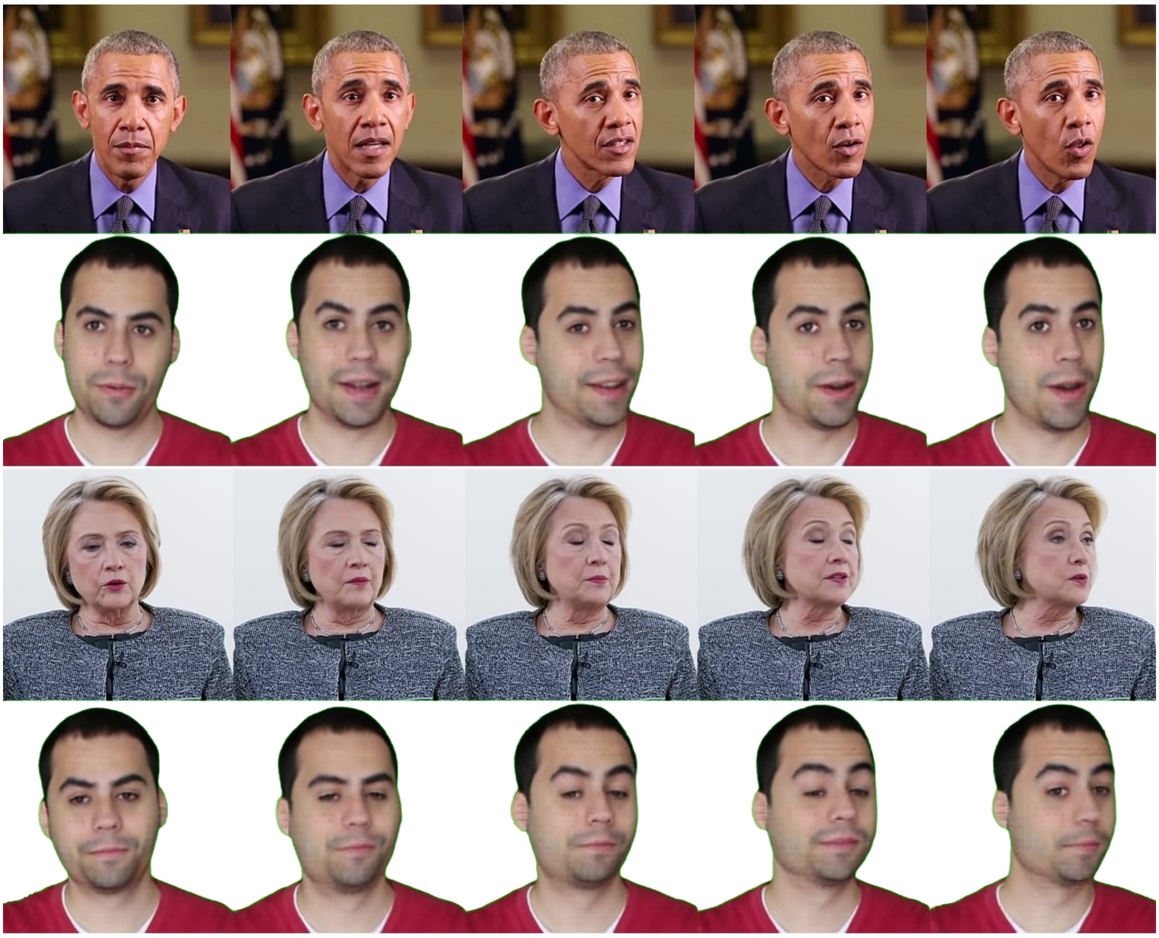


Figure 4.6: Sequences of subject 3 and 4 with a difference of 5 frames (166ms) between images.

Chapter 5

Conclusions

In this work, we proposed a novel face reenactment approach for image sequences that leverages AU intensities and eye gaze information to provide more realistic images. The method is based on an encoder-decoder network that receives as input the previous generated image and the current frame image of facial landmarks stacked together. This way, physical and temporal coherence are maintained while providing a realistic target image.

We collected data from different test subjects, from a Youtuber, and from different celebrities and science fiction characters to evaluate the flexibility and generalization capability of the proposal. Interestingly, the obtained results have shown faster convergence than compared methods (those that does not include facial action units and gaze regularization). In terms of the generated results, and evaluated for different faces regions, we obtained the best scores for the different metrics: distance, perceptual, and gradient. Those metrics where selected to cover different quantitative evaluation aspects: to have a generated frame close in distance to the target frame, to have a high degree of visual realism, and finally to keep shape artifacts. In terms of qualitative evaluation, we disseminated a questionnaire and a set of videos with different evaluated models to the crowd by means of a Youtube channel, obtaining more than 500 opinions. Independently of different characteristics of the raters (in terms of gender and age), overall, the best ranking is achieved by the proposed technique.

Although current proposal generates realistic and lifelike video face reenactment, several issues could be considered for future research: 1) increasing data resolution and making the network to work with more quality data. This, together with generative adversarial networks, will improve the quality of the reenactment results; 2) further fine-grain face analysis and micro-expressions could be considered to include additional regularization terms to enhance realism; 3) the model can be enhanced by performing adaptive face-region loss analysis. This way each face region could be evaluated by a different perceptual metric in order to generate an overall more human-like reenactment.

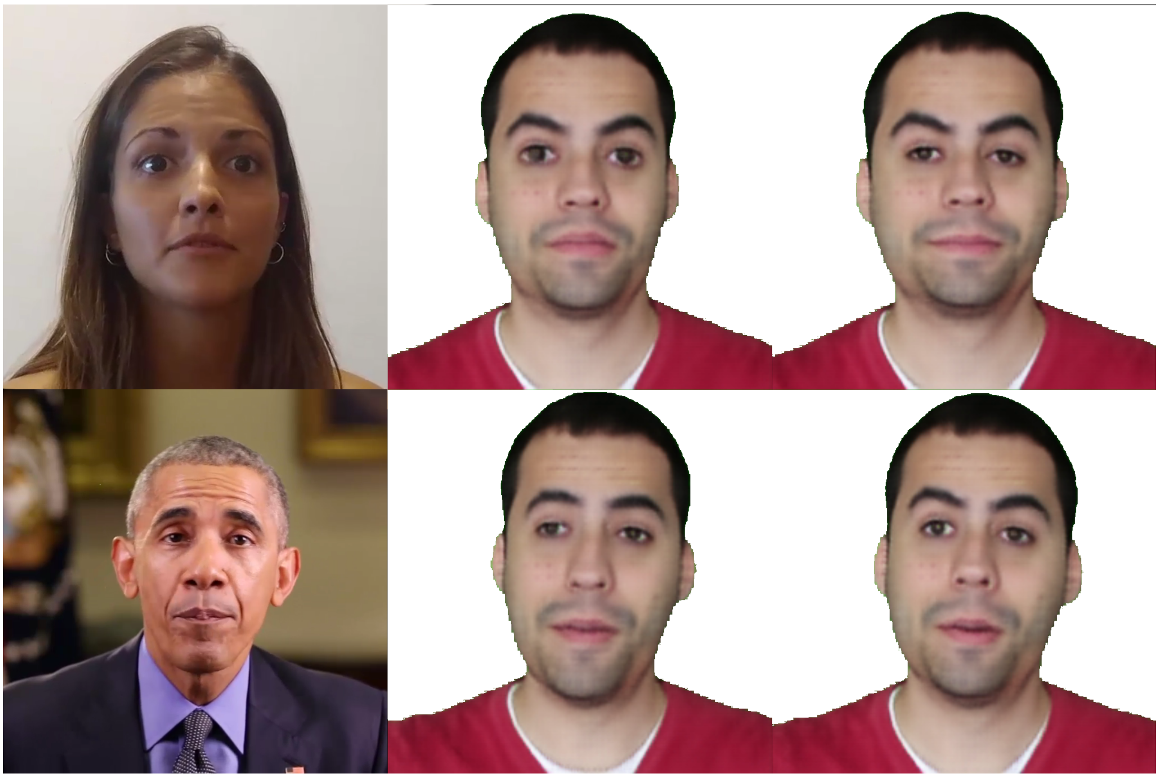


Figure 5.1: Generated images with and without PCA.

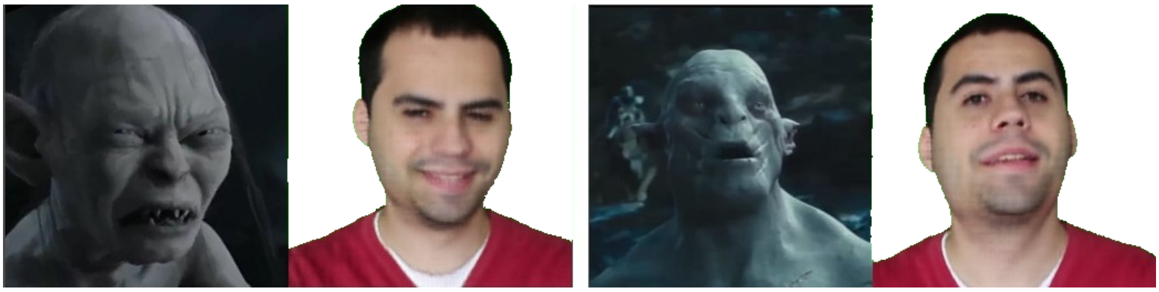


Figure 5.2: Face reenactment with avatars as source.

Appendix A

MOS Details

In this appendix we present the format of the mean observation score videos that had to be ranked in the form. The users had to rank the models in each video without knowing which video correspond to each model, so the models are distributed randomly inside the video and assigned a letter A, B, C or D. We also provide some statistics about the participants that filled the google forms and participated in the MOS. Since the information is retrieved from the google forms and participants where mainly from from Spain and South-America, both the questions and answers are shown in Spanish.

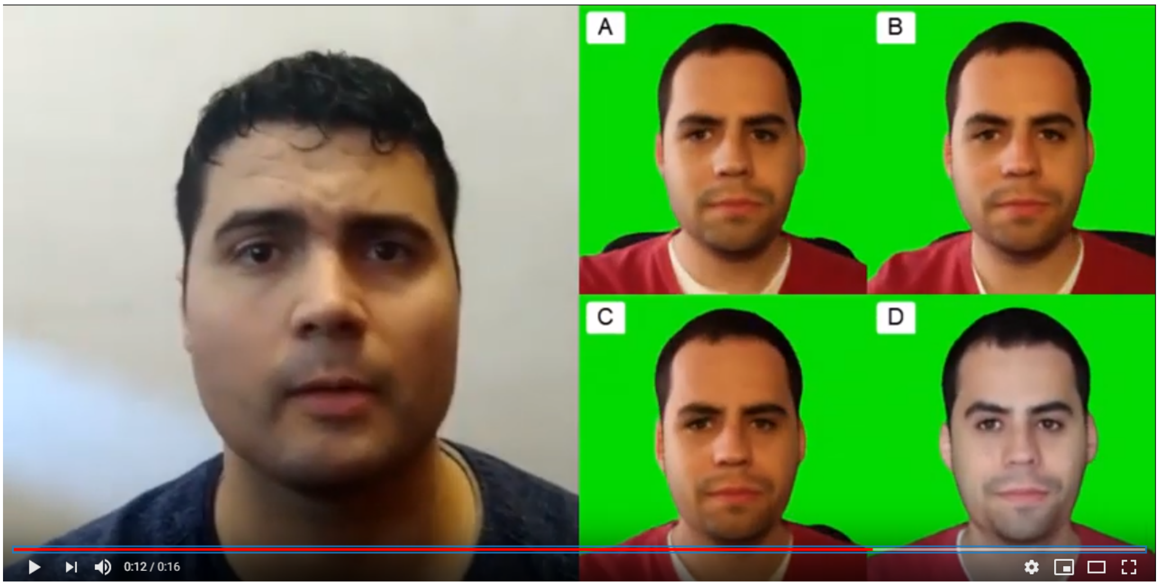


Figure A.1: Subject1: B-GAZE (A), B-AU (B), BASE (C), B-AUGAZE (D).

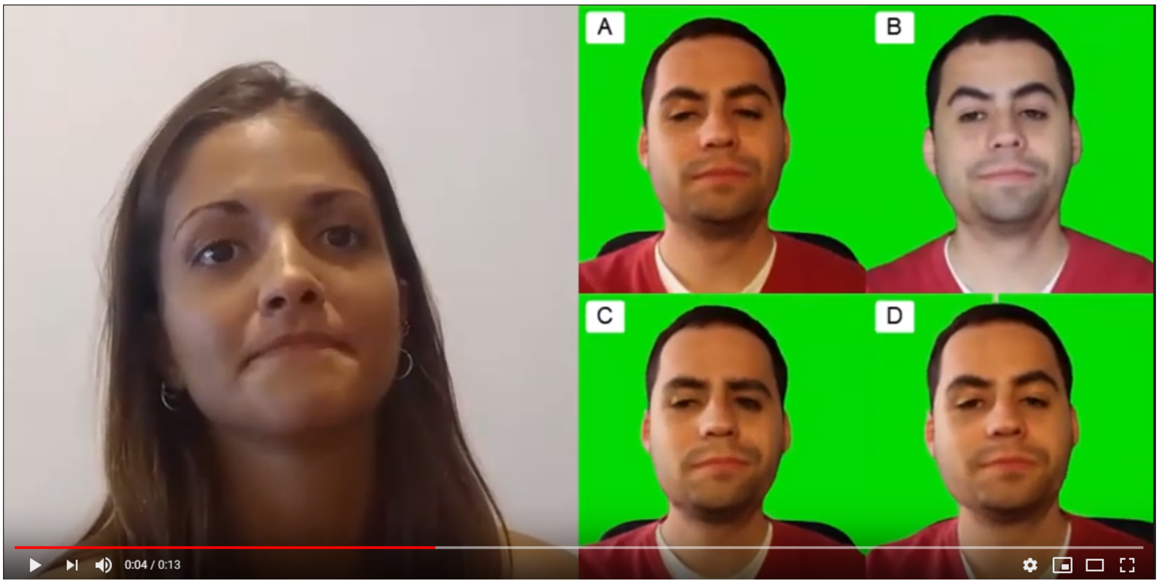


Figure A.2: Subject2: BASE (A), B-AUGAZE (B), B-GAZE (C), B-AU (D).

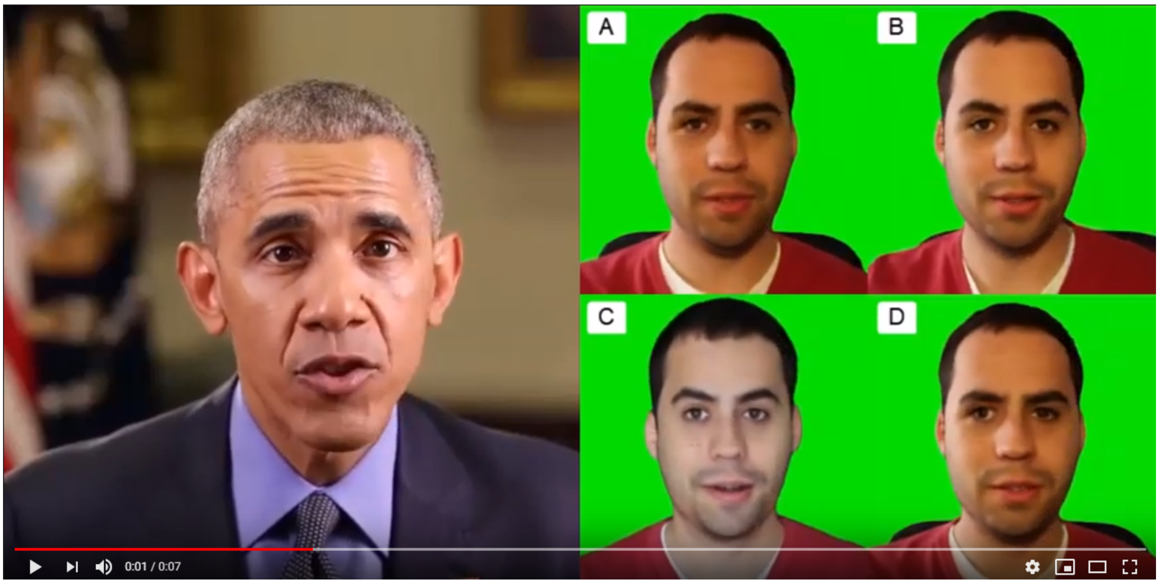


Figure A.3: Subject3: BASE (A), B-AU (B), B-AUGAZE (C), B-GAZE (D).

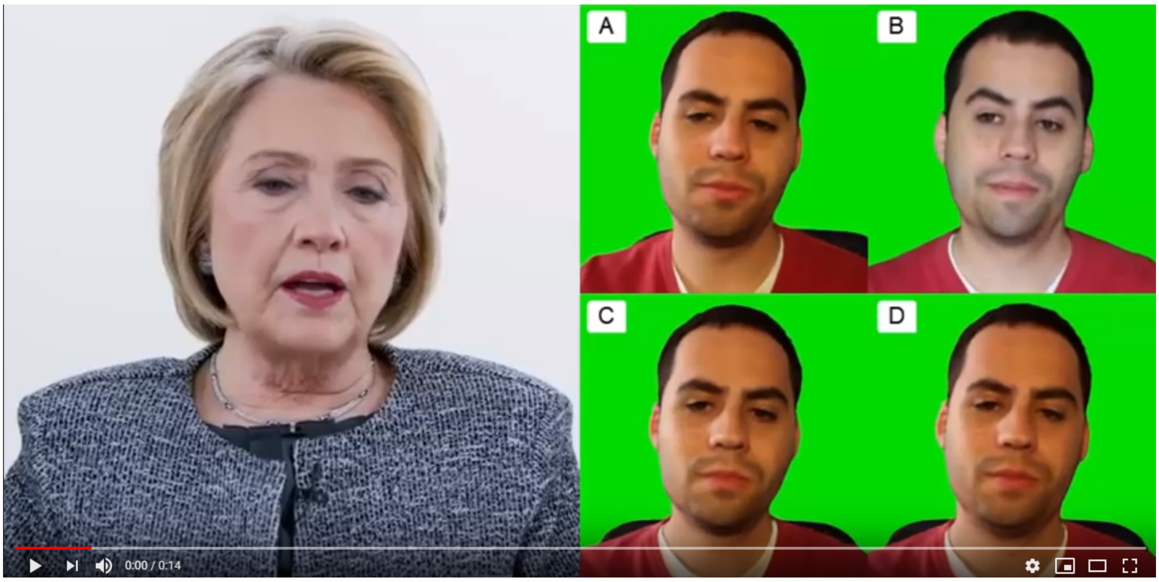


Figure A.4: Subject4: B-AU (A), B-AUGAZE (B), B-GAZE (C), BASE (D).

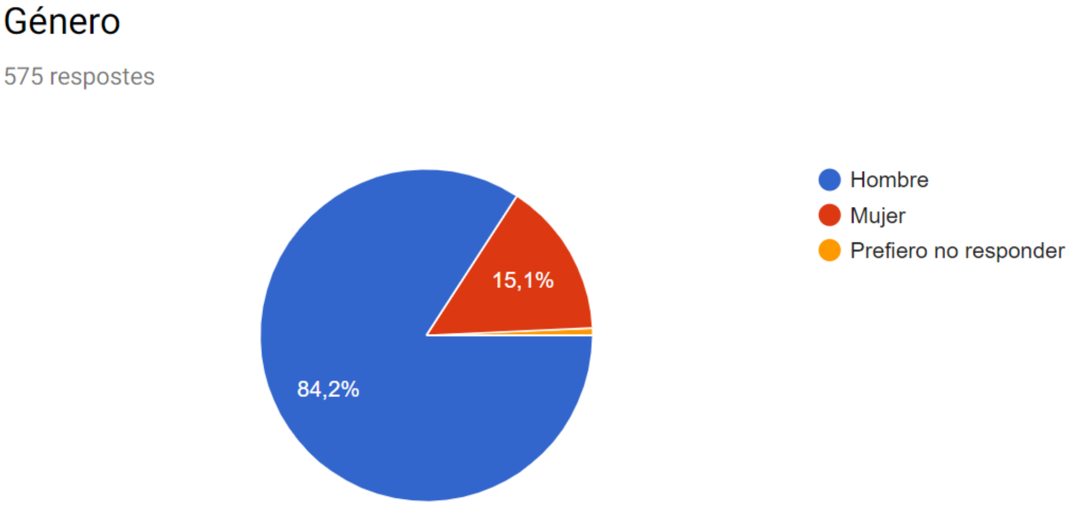


Figure A.5: Gender: Man (Blue), Woman (Red), N/A (Yellow).

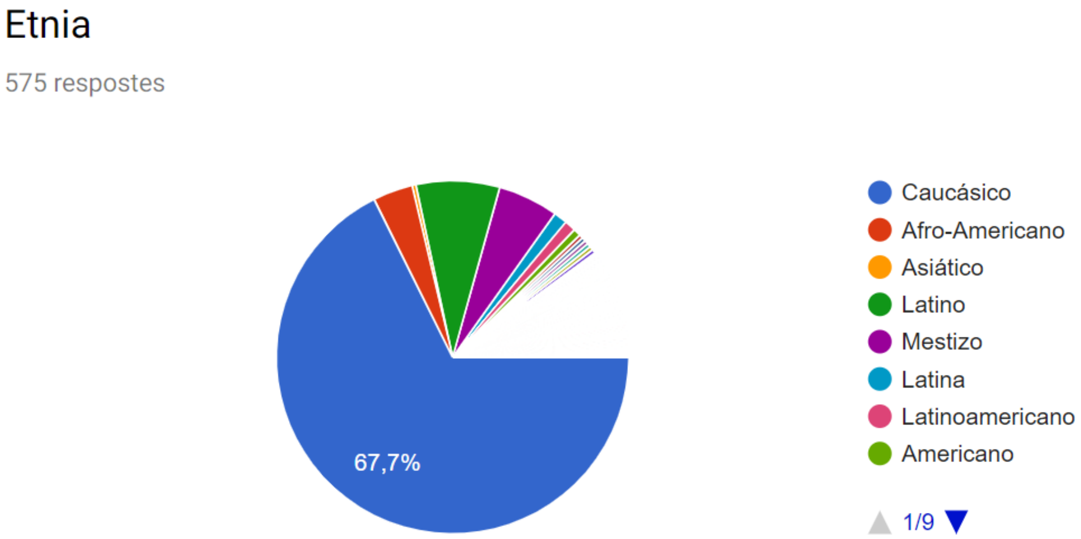


Figure A.6: Etnicity: Caucasian (Blue), Latin American (Green), Half-Blood (Purple), Afro American (Red), Others.

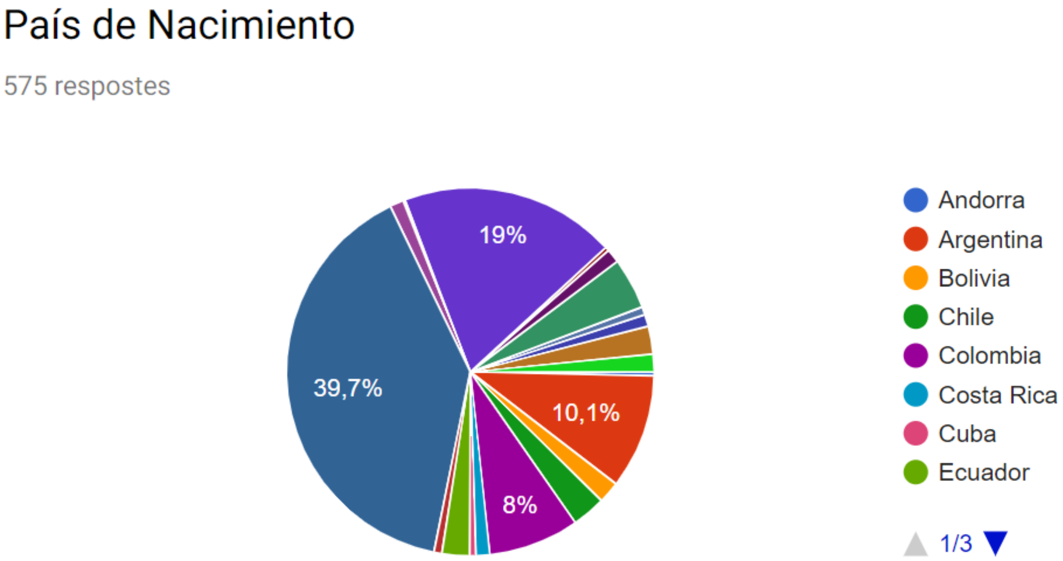


Figure A.7: Native country: Spain (Dark Blue), Mexico (Dark Purple), Argentina (Red), Colombia (Purple), Others.

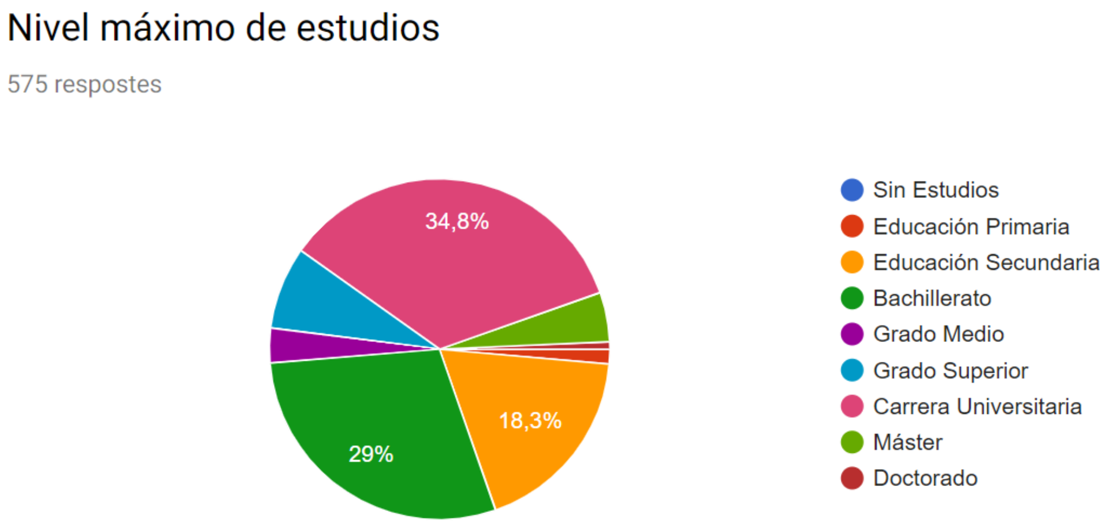


Figure A.8: Level of studies: University degree (Pink), Bachelor (Green), High School (Yellow), Others.

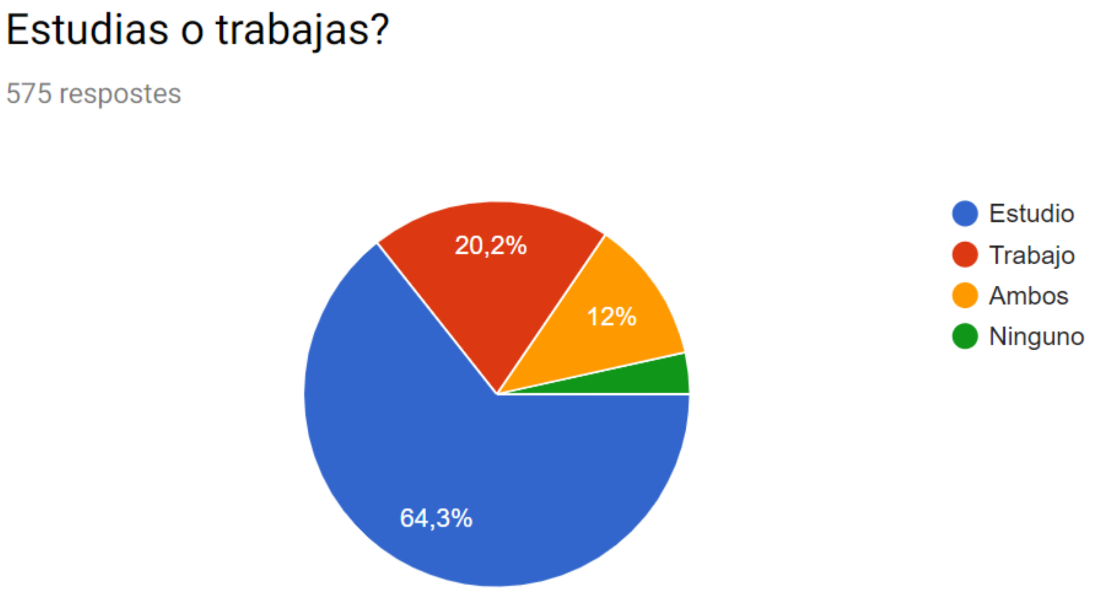


Figure A.9: Study or work: Study (Blue), Work (Red), Both (Yellow), None (Green).

Bibliography

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. pages 59–66, May 2018.

[2] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. 06:1–6, May 2015.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018.

[4] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Perez, and C. Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. 34(2):193–204, 2015.

[5] Y. Huang and S. Khan. A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. pages 437–445, 2018.

[6] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[7] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep Video Portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018.

[8] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. pages 271–276, 2001.

[9] M. Mathieu, C. Couprie, and Y. Lecun. Deep multi-scale video prediction beyond mean square error. 11 2016.

[10] J.-y. Noh and U. Neumann. Expression cloning. pages 277–288, 2001.

[11] C. Palmero, J. Selva, M. Bagheri, and S. Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. 05 2018.

[12] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 1:41.1–41.12, 01 2015.

[13] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. 2019.

[14] I. K.-S. S. Suwajanakorn, S. M. Seitz. What makes tom hanks look like tom hanks. 2015.

[15] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183:1–183:14, Oct. 2015.

[16] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62(1):96–104, Dec. 2018.

[17] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Trans. Graph.*, 37(2):25:1–25:15, June 2018.

[18] A. Zadeh, T. Baltrušaitis, and L. Morency. Convolutional experts constrained local model for facial landmark detection. pages 2051–2059, July 2017.

[19] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.