UNIVERSITAT DE BARCELONA

Treball de Fi de Grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Audio-Visual Deep Learning Regression of Apparent Personality

---

Alejandro Alfonso Hernández

Directors: Dr. Sergio Escalera Guerrero
Cristina Palmero Cantariño
Dr. Julio Jacques Junior

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 19 de gener de 2020

Abstract

Personality perception is based on the relationship of the human being with the individuals of his surroundings. This kind of perception allows to obtain conclusions based on the analysis and interpretation of the observable, mainly face expressions, tone of voice and other nonverbal signals, allowing the construction of an apparent personality (or first impression) of people. Apparent personality (or first impressions) are subjective, and subjectivity is an inherent property of perception based exclusively on the point of view of each individual. In this project, we approximate such subjectivity using a multi-modal deep neural network with audiovisual signals as input and a late fusion strategy of handcrafted features, achieving accurate results. The aim of this work is to perform an analysis of the influence of automatic prediction for apparent personality (based on the Big-Five model), of the following characteristics: raw audio, visual information (sequence of face images) and high-level features, including Ekman's universal basic emotions, gender and age. To this end, we have defined different modalities, performing combinations of them and determining how much they contribute to the regression of apparent personality traits. The most remarkable results obtained through the experiments performed are as follows: in all modalities, females have a higher average accuracy than men, except in the modality with only audio; for happy emotion, the best accuracy score is found in the Conscientiousness trait; Extraversion and Conscientiousness traits get the highest accuracy scores in almost all emotions; visual information is the one that most positively influences the results; the combination of high-level features chosen slightly improves the accuracy performance for predictions.

La percepció de la personalitat es basa en la relació de l'ésser humà amb els individus del seu entorn. Aquest tipus de percepció permet obtenir conclusions basades en l'anàlisi i interpretació de l'observable, principalment expressions facials, to de veu i altres senyals no verbals, el que permet la construcció d'una personalitat aparent (o primera impressió) de les persones. Les primeres impressions són subjectives, i la subjectivitat és una propietat inherent de la percepció basada exclusivament en el punt de vista de cada individu. En aquest projecte, aproximem aquesta subjectivitat utilitzant una xarxa neuronal profunda multimodal amb senyals audiovisuals com a entrada i una estratègia de fusió tardana de Handcrafted features, aconseguint resultats excel·lents. L'objectiu d'aquest treball és realitzar una anàlisi de la influència de la predicció automàtica de la personalitat aparent (basada en el model Big-Five), de les següents característiques: raw àudio, informació visual (seqüència d'imatges de cares) i high level features, incloses les emocions bàsiques universals d'Ekman, el gènere i l'edat. Amb aquesta finalitat, hem definit diferents modalitats, realitzant combinacions d'elles i determinant quant contribueixen a la regressió dels trets de personalitat aparents. Els resultats més notables obtinguts a través dels experiments realitzats són els següents: en totes les modalitats, les dones tenen una major precisió mitjana que els homes, excepte en la modalitat amb només àudio; per l'emoció feliç, la millor puntuació de precisió es troba en el tret de Consciència; els trets d'Extraversió i Consciència obtenen les puntuacions de precisió més alts en gairebé totes les emocions; la informació visual és la que més influeix positivament en els

resultats; la combinació de high-level features triades, millora lleugerament el rendiment de precisió per a les prediccions.

La percepción de la personalidad se basa en la relación del ser humano con los individuos de su entorno. Este tipo de percepción permite obtener conclusiones basadas en el análisis e interpretación de lo observable, principalmente expresiones faciales, tono de voz y otras señales no verbales, lo que permite la construcción de una personalidad aparente (o primera impresión) de las personas. Las primeras impresiones son subjetivas, y la subjetividad es una propiedad inherente de la percepción basada exclusivamente en el punto de vista de cada individuo. En este proyecto, aproximamos dicha subjetividad utilizando una red neuronal profunda multimodal con señales audiovisuales como entrada y una estrategia de fusión tardía de handcrafted features, logrando resultados excelentes. El objetivo de este trabajo es realizar un análisis de la influencia de la predicción automática de la personalidad aparente (basada en el modelo Big-Five), de las siguientes características: raw audio, información visual (secuencia de imágenes de caras) y high-level features , incluidas las emociones básicas universales de Ekman, el género y la edad. Con este fin, hemos definido diferentes modalidades, realizando combinaciones de ellas y determinando cuánto contribuyen a la regresión de los rasgos de personalidad aparentes. Los resultados más notables obtenidos a través de los experimentos realizados son los siguientes: en todas las modalidades, las mujeres tienen una mayor precisión promedio que los hombres, excepto en la modalidad con solo audio; para la emoción feliz, la mejor puntuación de precisión se encuentra en el rasgo de Conciencia; los rasgos de Extraversión y Conciencia obtienen los puntajes de precisión más altos en casi todas las emociones; la información visual es la que más influye positivamente en los resultados; la combinación de high-level features elegidas, mejora ligeramente el rendimiento de precisión para las predicciones.

Contents

1. Introduction

Personality is the result of the dynamic articulation of the psychological and biological aspects characteristic of each individual that determines their way of thinking and acting in a unique way in their process of adaptation to the environment [17]. On the other hand, personality perception is based on the relationship of the human being with the individuals of his surroundings [46]. This kind of perception allows to obtain conclusions based on the analysis and interpretation of the observable, mainly facial expressions, tone of voice and other nonverbal signals, allowing the construction of an apparent personality (or first impression) of people.

Personality is defined as a series of traits. A trait is a relatively stable over time disposition of the personality that is inferred from the behavior and that in turn determines the behavior [19]. If a person's traits are known, this may allow the person's behavior to be predictable. The Big-Five personality traits (also referred as the Five-Factor model) is a taxonomy for personality traits. The model represents five broad dimensions that have been applied to the description and activity of the human personality and are often represented by the acronym OCEAN which stands for: Openness to Experience (artistic, curious, imaginative, insightful, original, wide interests, etc.), Conscientiousness (responsible, reliable, efficient, planful, organized, etc.), Extraversion (outgoing, energetic, talkative, active, assertive, etc.), Agreeableness (kind, sympathetic, forgiving, generous, appreciative, etc.), and Neuroticism (worrying, self-pitying, unstable, tense, anxious, etc.) [27], [18].

Personality Computing is a currently active research field that studies computational techniques related to human personality. The analysis of automatic personality perception (a branch of this field) is the focus of our work. Apparent personality (or first impressions) is subjective and subjectivity is an inherent property of perception based exclusively on the point of view of each individual. In this project, we approximate that subjectivity using deep learning techniques and achieving promising results. Our motivation is to improve the understanding of the variables that can influence the decision making of intelligent systems (specifically deep neural networks) to regress apparent personality, similar to how the human being would.

Deep learning algorithms require a lot of training data to be able to achieve good results. In the field in which our work is developed, this becomes a more serious problem even due to the extreme shortage of datasets for the realization of studies about human personality, which obviously slows the progress of the same, since there is no great variety of quality datasets that can be used today for this topic. Our work is based on ChaLearn First Impressions (FI) dataset [19], which is in fact, quite recent (2016). The FI dataset is currently the most complete dataset in this field, containing 10,000 short YouTube videos under a creative commons license, with one individual per video talking to a camera. Each video is labeled following the Big-Five personality traits model.

The objective of our work is to perform an analysis on the influence of automatic prediction for apparent personality based on the Five-Factor model, of the following characteristics: facial features, emotions, raw audio, gender and age. To do this, we have defined different modalities,

performing combinations of them and determining how much they contribute to the regression of personality traits.

These modalities are defined as follows:

- **Audio modality**. Modality where we analyze only the impact of the person's voice. To do this, we use a neural network architecture that we train with raw audio as input.
- **Visual modality**. Modality where we analyze the importance of facial features. We use a well-known deep convolutional neural network architecture to extract these features and learn from them to predict personality traits.
- **Audio+Visual modality**. Modality that combines audio signals and face images to see how much the prediction improves. We merge the audio and visual models into a single architecture that is capable of analyzing audiovisual signals.
- **Audio+Visual+handcrafted features modality**. Modality that complements the audio-visual signals with extra information that should improve the results. For this, we have used the Ekman's universal basic emotions (anger, disgust, fear, happy, sadness, surprise), as well as gender and age annotations of the observed people. We use the Audio+Visual modality with a late fusion strategy of handcrafted features.

Furthermore, we perform an analysis of the improvement in accuracy related to the handcrafted features, comparing the results by gender, age-ranges and emotions. Below we mention some of the conclusions we have reached from the experiments, which reflect our main contributions to the realization of this work:

- It is shown that the model with a combination of audio-visual information and high-level features is the one that has the highest accuracy in predicting apparent personality traits, achieving promising results.
- Our studies have shown that visual information is the one that most positively influences the results.
- The combination of high-level features chosen slightly improves the accuracy performance for predictions.
- In all modalities, females obtain a higher average accuracy than males, except in the modality with only audio.
- Audio modality has the worst results when the relation between traits and emotions were analyzed, especially for Neutral emotion and Openness trait. But in spite of everything, it achieved good results for Fear emotion, especially for Neuroticism trait.
- For happy emotion, the highest accuracy score is found in the Conscientiousness trait.
- The Extraversion and Conscientiousness traits get the highest accuracy scores in almost all emotions.

2. Related work

In this section, we discuss some of the most relevant research papers related to the topic of our work: Automatic Personality Perception. We also talk about the different subjective factors that may influence the perception personality.

2.1. Automatic Personality Perception

Personality Computing field addresses three fundamental problems: Automatic Personality Recognition, Automatic Personality Perception and Automatic Personality Synthesis [18]. The first is about the real recognition of personality, how people perceive themselves and is based on self-assessment analysis. On the other hand, Automatic Personality Synthesis tries to recreate the human personality through synthetic speech, artificial agents and robots. However, our study focuses on Automatic Personality Perception, where we analyze the different subjective factors that influence the personality perception that, therefore, are of utmost importance for this topic.

2.3. Facial expressions

In recent years, the fields of computer vision and machine learning have benefited greatly from the incorporation of deep learning techniques. Convolutional Neural Networks (CNNs), one of the most popular networks in deep learning, show a remarkable advantage in automatic visual feature extraction, particularly from human faces. This has had a great impact in the field of Automatic Personality Perception, since there are numerous papers that demonstrate a link between face and apparent personality traits [23], [30], [31], [32].

The face has the highest variability degree in a person [21], it has thousands of combinations of features that make its morphology unique. These characteristics can influence how people relate to each other. For example, we tend to prefer politicians who simply look more competent merely based on their facial appearances [34]. In the task of apparent personality prediction, CNNs mainly analyze eyes, nose and mouth from faces [44], [30]. If two different face images from the same person are observed in different contexts, personality perception can vary radically [39] and may not benefit the task being performed. However, when there is temporal information such as sequences of images from the same clip, this greatly improves the results obtained, since it results in a more consistent and accurate prediction [30].

2.4. Subjective bias in first impressions

There are some papers that analyze whether when audio and extra information (as gender, age, ethnicity, attractiveness, etc.) are also combined with facial expressions, the prediction may be slightly enhanced due to a possible latent bias. In the paper of Guntuku et al. [38], apparent personality traits were predicted by analyzing the eyes and extracting low-level features from images to obtain data such as gender and age. Escalante et al. [25] showed that older men and young women are preferred in job interviews. Latest works combine gender with age as complementary information to audiovisual neural networks [45]. In [35], it has been

observed that the shape of the eyes has a correlation with youthful-attractiveness and a face with rough male characteristics with dominance. Here [40], the authors predict personality impressions from Twitter profile images using a multivariate regression approach with handcrafted features and deep learning with a pretrained VGG19 architecture and using face detection and alignment as pre-processing technique (they also considered background information). Gürpinar et al. [41] created a Kernel Extreme Learning Machine (KELM) apparent personality regressor. To do this, they used a VGG-Face to extract facial expressions, they also made use of ambient information, weighted score level fusion strategy and auditory signals. The paper was based on the same database that we are using in our work: the First Impressions dataset, and they archived state-of-the-art results winning the ChaLearn First Impressions Challenge [42].

In our work, we follow the trend in the field of Automatic Personality Perception with the use of raw inputs, CNNs and handcrafted features. In the following sections, we will analyze different modalities with audiovisual and complementary information (as emotions, gender and age), trying to find a relationship between each modality and the apparent personality.

3. Convolutional Neural Networks

In this section, we discuss the methodological background associated with the subject of this work, based on deep learning techniques and more specifically in convolutional neural networks. We will first make a small introduction to what Artificial Neural Networks are through an analogy with the neural networks of the human brain. Later, we will briefly introduce key questions of how an artificial neural network works and how they get to learn. We will explain from what a Perceptron is to go into detail to Convolutional Neural Networks (CNNs), where we will explain their structures and each type of layer. Finally, we will briefly review the history of the most important architectures, from the first that came out to more complex ones which are currently used by the research community.

3.1. Artificial Neural Networks: a brain analogy

Artificial Neural Networks (ANN) are loosely based on neural networks that make up the nervous system of the human being [1]. The average human brain has approximately 86 billion cells responsible for receiving, processing and transmitting information in the form of nerve impulses, these cells are called neurons and each of them can be connected with up to 10,000 others passing signals to each other through up to 1,000 trillions of synaptic connections [2]. In analogy to the neurons of our brain, a computational model of a neuron receives input signals through its dendrites from the axons of other neurons connected to it by synapses. Subsequently, the neuron sends an output signal to the following neurons in its network depending on the synaptic strength that links them. This strength determines how much the information that travels from one neuron to another can influence. If the synaptic force is strong enough, then the information will be transmitted [3].

3.2. Perceptron and Multi-Layer Perceptron

We can then say that an artificial neuron has inputs x_i (being for example x_0 the input signal of neuron 0) that multiplied by some weights w_i (being for example w_0 the synaptic strength with neuron 0), adding a bias b and applying an activation function f, a bounded output is finally obtained according to f. The formula would be as follows: $f(\sum_i w_i \cdot x_i + b)$ [3]. The single layer perceptron (the simplest type of feed-forward network) in fact uses the formula described above in order to learn a binary classifier whose limitation is precisely just being able to solve linearly separable problems [4]. On the other hand, the multilayer perceptron (MLP) presents in its architecture one or more hidden layers so it allows to solve numerous classification and regression problems, usually considering as many input neurons as labels to be recognized by the task.

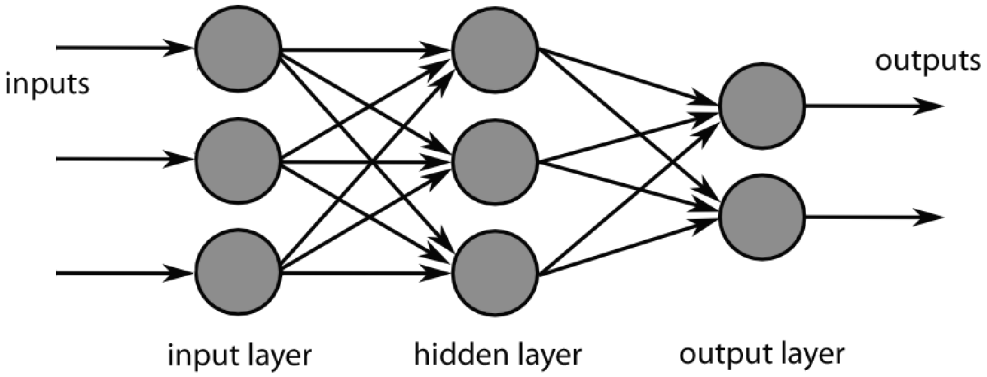


Fig. 1. Example of a Multilayer Perceptron (MLP) with an input, a hidden and an output layer.

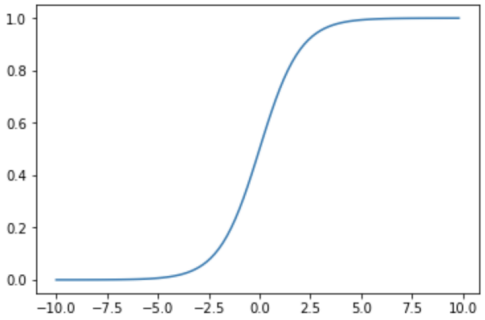
3.3. Activation functions and Optimizers

3.3.1. Activation functions

The activation function takes an input value, performs a certain fixed mathematical operation on it and return the output bounded in a certain range.

There are several activation functions, the most used are:

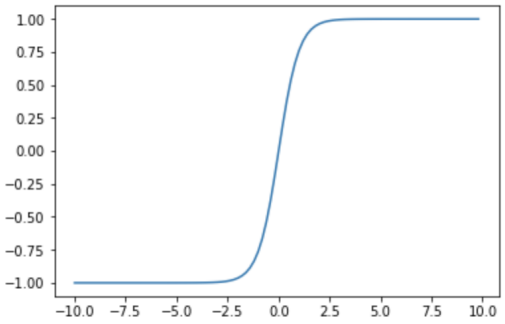
- Sigmoid: it transforms the values entered to a scale (0,1), where the high values are asymptotically to 1 and the very low values tend asymptotically to 0.



$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Fig. 2. Sigmoid function graphical representation.

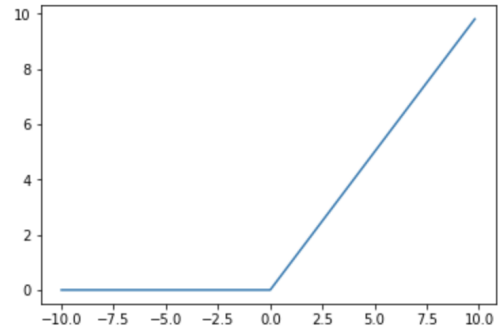
- Tangent Hyperbolic (also referred to as Tanh): this function transforms the values introduced to a scale (-1,1), where the high values are asymptotically at 1 and the very low values tend asymptotically to -1.



$$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

Fig. 3. Tanh function graphical representation

- Rectified Linear Unit (also referred to as ReLU): this function transforms the entered values by canceling the negative values and leaving the positive ones as they enter.



$$f(x) = \max(0, x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases}$$

Fig. 4. Tanh function graphical representation

- Softmax: this function transforms the outputs into a representation in the form of probabilities, such that the sum of all the probabilities of the outputs is 1.

$$f(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

3.3.2. Optimization algorithms

Optimization algorithms are used to optimize a cost function J in order to train the neural network. The cost function is defined as:

$$J(W, b) = \frac{1}{m} \sum_{i=1}^{m} L\,(y'^{i}, y^{i})$$

where L is the loss, y' is the predicted value, y the actual value, m the number of values, W represents the weights and b is the bias. So, the idea is during forward propagation, y' is obtained and then during backpropagation, the values of the cost function J are minimized with the values of W and b.

Next, we will describe some of the most important optimization algorithms:

- Stochastic gradient descent (SGD) [47]

This algorithm, in each iteration, chooses m data points, calculates the average gradient for them and updates the parameters. By periodically applying the gradient descent to the weights, it will eventually arrive at the optimal weights that minimize the loss function and allow the neural network to make better predictions.

Algorithm 8.1 Stochastic gradient descent (SGD) update at training iteration k

Require: Learning rate ϵ_k.
Require: Initial parameter $\boldsymbol{\theta}$
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow +\frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \hat{\boldsymbol{g}}$
end while

Fig. 5. SGD optimizer algorithm [47]

- SGD+Momentum [48]
  It maintains another variable v that would accumulate gradients and uses this variable to update parameters. With this algorithm, by adding the parameter v, when a local minimum is found, depending on v, the data point could exit that local minimum and continue searching for a better option.

Require: Learning rate ϵ, momentum parameter α.
Require: Initial parameter $\boldsymbol{\theta}$, initial velocity $\boldsymbol{v}$.
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient estimate: $\boldsymbol{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
Compute velocity update: $\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \epsilon \boldsymbol{g}$
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}$
end while

Fig. 6. SGD+Momentum optimizer algorithm [48]

- Adam [7]
  Adaptive Moment Estimation (also referred as Adam) is an optimization algorithm that computes an adaptive learning rate for each parameter. It stores an exponentially decaying average of past squared gradients and an exponentially decaying average of past gradients (similar to SGD+Momentum).

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Fig. 7. Adam optimizer algorithm [7]

3.4. Backpropagation

When an MLP has more than one hidden layer it is considered Deep Learning. The objective of these neural networks is to approximate a function f^*.These neural networks are composed of numerous functions using a chain structure. For example, if we have a network with 3 hidden layers, it means we have 3 functions $f^{(1)}$, $f^{(2)}$, $f^{(3)}$ that would form the function $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$ [5]. The output layer produces output values with a certain error coming up from the hidden layers that say how to update the previous weights in order to reduce the loss. For this, the backpropagation algorithm is applied, which efficiently computes the gradient of the loss function with respect to each weight using the chain rule [6]. Each node has its local gradient and as numerical values of gradients coming from upstream are received, the algorithm takes this and multiplies it by the local gradient. The result is sent back to the connected nodes going backwards, only taking into account these immediate surroundings to avoid redundant calculations. Once the derivatives calculated during backpropagation are obtained, an optimization algorithm (such as Adam optimizer [7]) is applied to reach the minimum of the loss function with respect to the parameters.

3.5. How CNNs work and most important architectures

Deep neural networks usually have millions of parameters and this makes learning difficult because the landscape to be navigated by the algorithm becomes high-dimensional [5]. If we talk about images, they have a high dimensionality because each pixel is considered a feature. CNNs [8] (see Figure 8) are a type of feed-forward network that adapts weights as convolutions and solves this problem by trying to figure out and compute the features that are relevant to the problem to be solved. Dimensionality reduction is achieved using filters as sliding windows throughout the image matrix and its task is to multiply its values by the original pixel values. The result of these convolutions are feature maps. These feature maps will be used by the following **convolution layers**, learning increasingly complicated features. A CNN usually includes pooling layers. **Pooling layers** summarize the features present in a region of the feature map generated by a **convolution layer,** reducing considerably the dimensions of the feature map and therefore the number of parameters. There are several types of **pooling layers**, the most common are **max pooling**, that selects the maximum element from each region of the feature map covered by the filter, and **average pooling**, that computes the average from each region instead of the maximum.

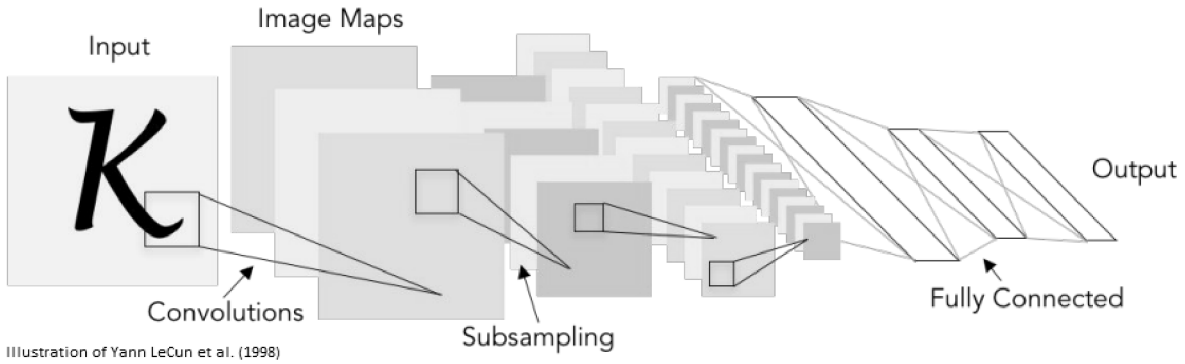


Fig. 8. Standard architecture of a Convolutional Neural Network [8].

The first architecture of a CNN to be able to solve a task such as reading digits was **LeNet-5**, developed by Yann LeCun et al. in 1988 [8]. LeNet-5 became the template network of other more complex architectures that would come after using the idea of stacking convolutional and pooling layers with dense layers at the end. This network has an input of 32x32x1, two convolutional layers of 5x5 dimensions and an average pooling layer of 2x2 after each convolutional layer, finally three dense layers of 120, 84 and 10 neurons each, the last one being the output of the network.

AlexNet, presented in 2012 by Geoffrey E. Hinton et al. [9], was the next architecture that became popular by significantly reducing top 5 error from 26% to 15% in the ImageNet ILSVRC 2012 challenge [10]. AlexNet followed the same idea in the structure of LeNet-5 but deeper with 62.3 million parameters, presenting 11x11, 5x5, 3x3 convolutions, max pooling, dropout, data augmentation, ReLU activations and SGD as optimizer (see Figure 9).

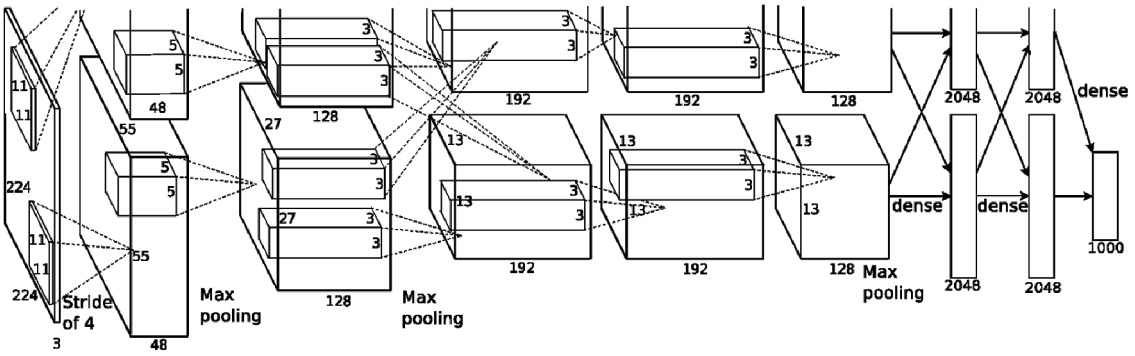


Fig. 9. AlexNet architecture developed by Geoffrey E. Hinton et al. (2012) [9].

The **VGG-19** architecture presented by the Visual Geometry Group from the University of Oxford in 2015 [11] (see Figure 10) was quite similar to AlexNet but with much more filters and it uses additional multi-scale cropping as data augmentation. With this network it was shown that increasing the depth of a network implied an improvement in its performance. This model reaches 138 million parameters and it takes 548 MB of storage space. The first version of this network (known as VGGNet) was the runner-up in ILSVRC 2014 [12].

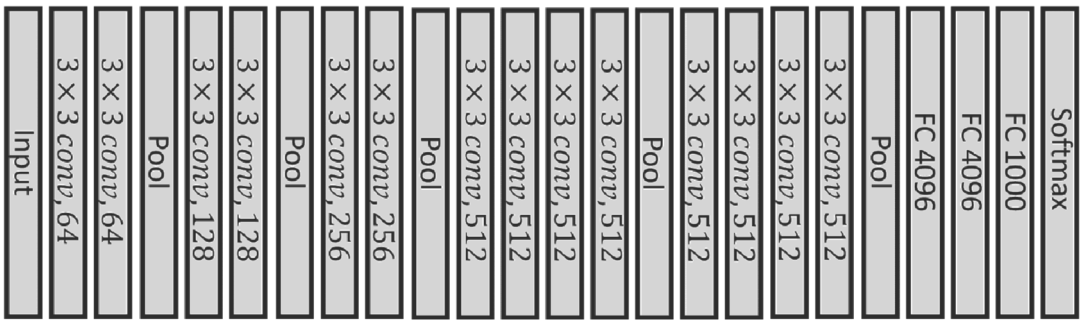


Fig. 10. VGG-19 architecture developed by the University of Oxford (2015) [11].

Previous networks simply added more and more layers and this improves the performance. But, for very deep neural networks with certain activation functions such as the logistic or the hyperbolic tangent, a large change in the input causes a small change in the output, making the derivative small as well. When many small derivatives are multiplied together, the gradient decreases exponentially during the backpropagation process. A small gradient means that the weights of the first layers are being updated very slowly or even nothing at all. This is known as the vanishing gradient problem and can be simply solved by using activation functions that do not cause small derivatives.

Another way to avoid this issue was popularized with the introduction of the **ResNet** architecture [13]. This network features **skip (or residuals) connections** from one layer to another, skipping up to 3 layers in some cases. The idea is to allow gradients to flow through the network directly, without passing through non-linear activation functions. ResNet also was among the first to use batch normalization and does not use dense layers at the end. It won the ILSVRC 2015 achieving a top-5 error rate of 3.57% [14].

Google introduced in 2014 its own CNN called **Inception** (later known as Inception-v1) [15]. This carefully designed network differs radically from the rest. It is made up of the so-called "Inception modules" (see Figure 11) that can reduce the number of parameters, reaching only 5 million. These modules consist of filters of different sizes (1x1, 3x3, 5x5) as well as a max pooling layer of 3x3. It should be noted that before computing the convolutions with the 3x3 and 5x5 filters, it uses an extra 1x1 filter to perform a dimension reduction. The outputs are then concatenated and sent to the next module.

The idea of using different filter sizes is to not lose important features due to the huge variation of the information in images. At the end of the network a global average pooling is performed. Also, to avoid the vanishing gradient problem, the network uses two auxiliary classifiers linked to two different modules in the middle of the network, and each classifier will calculate its auxiliary loss. The total loss would be the weighted sum of these two losses mentioned above and the real loss.

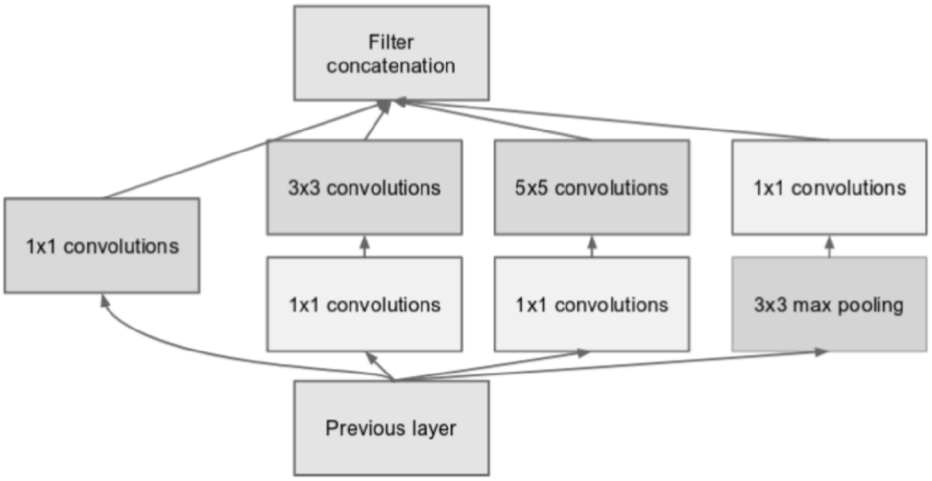


Fig. 11. Inception module from the Inception-v1 architecture developed by Google [15].

Later, other version which optimized the first version were released, until Inception-v4 was presented in 2016, whose main difference with Inception-v3 was the increase in the number of Inception modules, reaching 43 million parameters. In the same paper, Inception-ResNet-v2 was introduced as well which, compared to the Inception v3, includes more modules but in this case replacing them by Residual Inception modules among other details. This architecture has 56 millions parameters [16].

4. First Impressions dataset

In this section, the First Impressions (FI) dataset on which the study of our work is based, is described. We comment on its main characteristics: how the data was obtained and labelled, number of samples, the characteristics of observed people, the techniques used for the labeling of traits, etc.

The FI dataset was released at ECCV 2016 Challenge. It is composed of 10,000 short videos (audio included) with a fifteen seconds average duration extracted from YouTube high-definition videos (720p) under a creative commons license. Each clip contains only one person speaking directly at the camera in at least 80% of the time. The people appearing are of both genders with a wide range of ages (11 to 64 years old), different nationalities and ethnicities, speaking only in English. Amazon Mechanical Turk (AMT) services were used to obtain the ground truth for apparent personality based on the Big Five personality traits model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) in the range of $[0,1]$, labeling each clip through pairwise comparisons between videos (see Figure 12), converting the results into continuous values by fitting a Bradley-Terry-Luce model with maximum likelihood. Age, gender and ethnicity annotations are also included. The FI dataset is split into training, validation and test sets with a $3:1:1$ ratio, respectively [19].

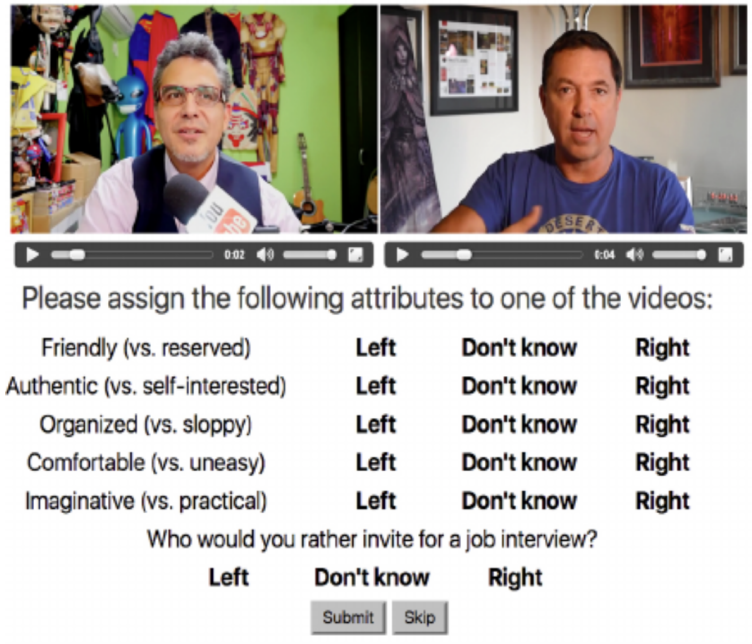


Fig. 12. Example of pairwise comparisons between videos used by AMT annotators. Extracted from [19].

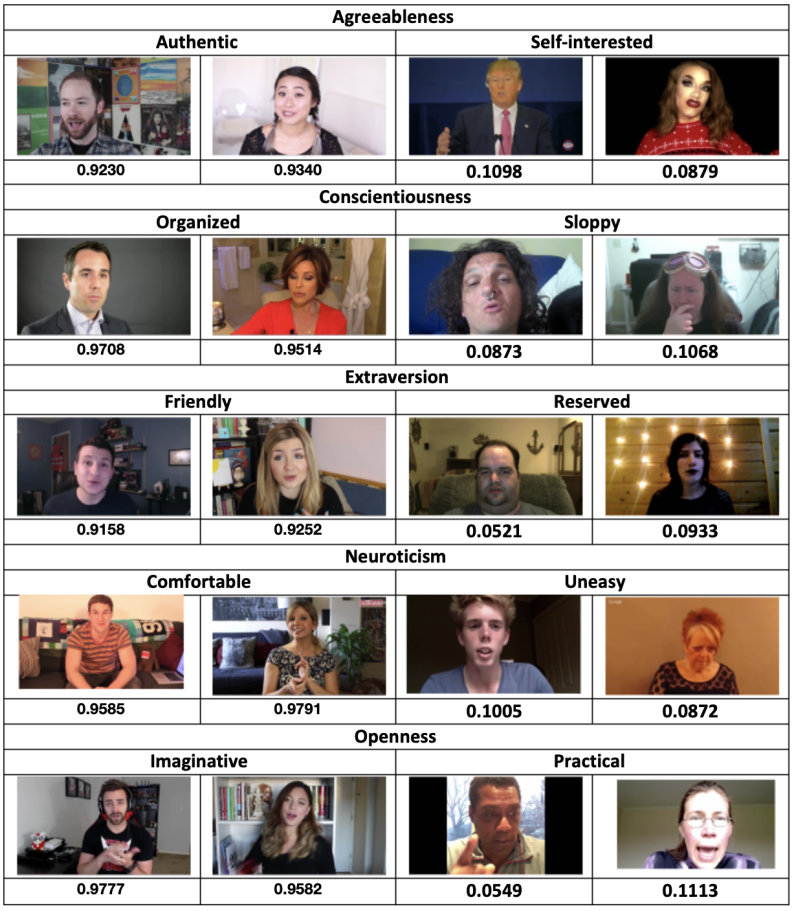


Fig. 13. Example of voted videos where both sides of each trait can be seen clearly. Extracted from [19].

5. Methodology

In this section, we explain first, the pre-processing techniques that we have used on the FI dataset (described in Section 4) to prepare the data with which we feed the different deep neural network models we analyze. Subsequently, we proceed to explain in detail each of the deep learning models proposed by us.

5.1. Data pre-processing and modalities

5.1.1. Extracting face images

In order to extract the faces of the FI dataset videos, the standard approach for “frame selection” has been followed, choosing the frames randomly [24]. In our case, we have chosen ten frames per clip (for those with less than ten, all its frames have been selected). Faces were extracted using a previously trained face detector that uses the classic Histogram of Oriented Gradients (HOG) feature combined with linear classifier, an image pyramid and sliding windows detection scheme [53], [54]. The detected faces are trimmed and aligned (according to the horizontal line between both eyes) with a resolution of 224x224 pixels [19].

5.1.2. Raw inputs and handcrafted features

With the evolution of deep learning over these years, the way of doing feature engineering has also undergone changes. Recently, raw inputs have begun to be used in addition to handcrafted and learned features, due the capability of deep neural networks of automatically learn the most important features by themselves, greatly lightening the preprocessing task.

5.1.2.1. Raw audio

The voice is such a personal pattern, which is currently used to verify identity and allow access to some computer systems. The voice tone presents a series of sound parameters that give meaning, conscious and unconscious, to the message that is being transmitted. Some of them are: the sound intensity, the diction speed, clarity, projection, etc. That is why we added raw audio as input to our model inspired by the paper of Yağmur Güçlütürk et al. [23], where it is shown that audio improves prediction accuracy of personality traits. In our case, we decided to use only the first five seconds per video because it is demonstrated, according to [24], that the first part of a video contains the most relevant audio information to use for personality regression.

5.1.2.2. Age and gender

Escalante et al. [25] studied the influence of age perception for personality first impressions, showing that older men and young women are preferred in job interviews. On the other hand, the studies by Chan et al. [50] showed that in adolescents, the Openness, Extraversion and Neuroticism traits are perceived higher than for older people, while the latter has higher Agreeableness. Apparently, according to the papers mentioned above, age influences apparent personality. Gender also seems to have a certain relationship with personality perception. There

are works that even combine gender and age among other characteristics such as ethnicity [45], [51], [52] showing an improvement in the results, due to the addition of these attributes. The study of Escalante et al. [25] mentioned above, also showed some gender bias in first impressions, where females showed higher Openness and Extraversion than men. For our work, age and gender annotations were used labeled as follows: age as a positive integer and gender as Male $= 1$, Female $= 2$.

5.1.2.3. Emotions

Spontaneous and intuitive conclusions about the personality of others are often based on the expressions we perceive from people's faces. For apparent personality, the face is extremely important because it contains the highest level of variability in a person [21]. Therefore, we have used in this work raw images of faces that allow the deep learning model to learn to associate facial features with the personality perception, just as a human being would naturally do. The human face is composed of 43 muscles, of which 36 are used to express emotions. Paul Ekman, one of the pioneers in the study of emotions, argued that facial expressions associated with emotions are involuntary, unconscious and universal. Ekman defined the following six basic emotions: Anger, Disgust, Fear, Happiness, Sadness, Surprise (see Figure 14) [22]. In our work, we extract Ekman's emotions (also considering the neutral expression) of each frame using an already trained neural network from the study of Rosa et al. [49] which compared AlexNet, VGG16 and ResNet on the task of facial emotions classification. We chose AlexNet architecture over the rest, despite being reported as the one with the lowest performance, the reason is because it offers the best trade-off between accuracy and training speed. Finally, we compute the 5-bin histogram of each emotion obtained from the same sequence of frames. The emotions have been separated into groups, based on the confidence of their prediction. The frequency of frames per emotion has been normalized in the range of [0.0, 1.0] in order to sum 7. Each of these histogram vectors obtained summarizes the emotions of a video. (see Figure 15) forming a vector of 35 values (7 emotions values per each of the 5 confidence ranges). Finally, this vector is concatenated with corresponding gender and age values (its final length is 37).

Fig. 14. Ekman’s universal basic emotions (plus Neutral expression) [22].

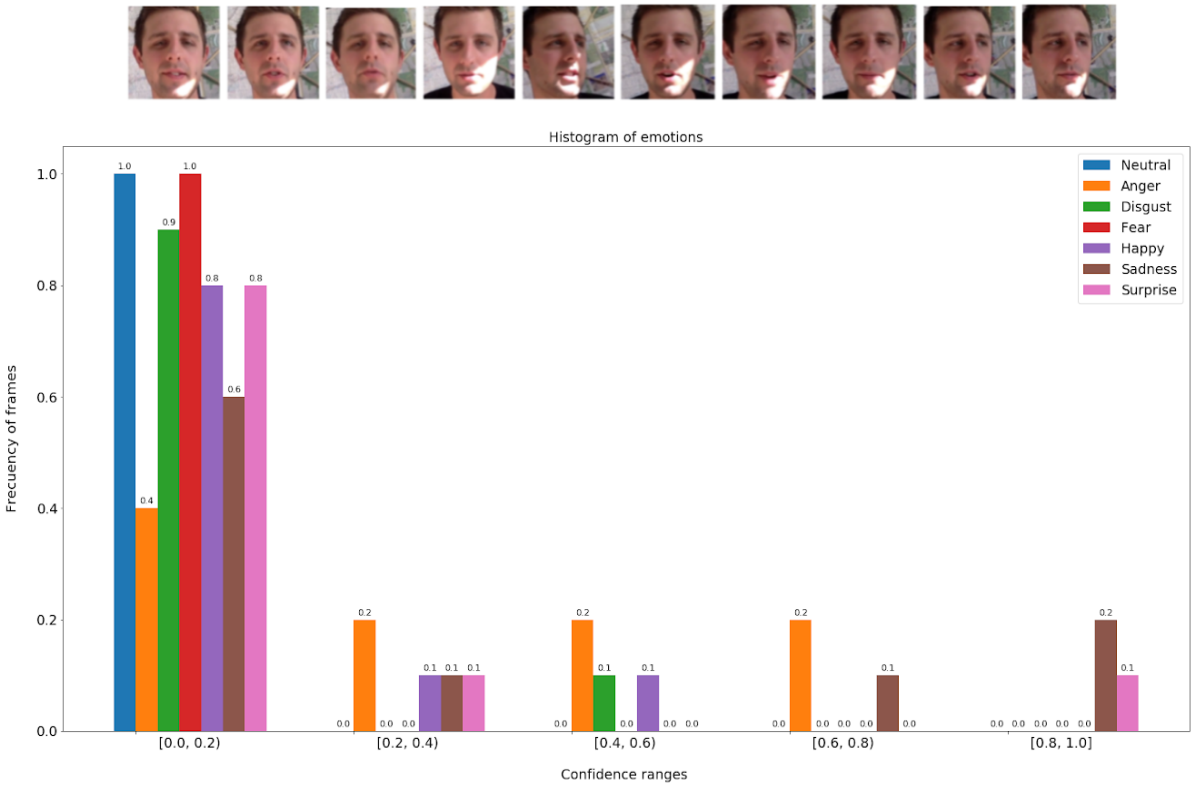


Fig. 15. Histogram of Ekman's universal emotions (plus neutral expression) per confidence range. This figure represents the emotions of one specific sequence of frames shown above the bar chart.

5.2. Proposed models

The deep neural network architectures proposed in this section are aimed at analyzing the influence of raw visual with audio streams with the addition of high-level attributes (gender and age obtained through annotations and emotions of facial expressions through a network already trained for extraction) in the automatic prediction of apparent personality regressing the Big-Five personality traits.

For modalities with visual information, we use the well-known CNN architecture VGG-Face, previously trained for face detection. We use this network because it is widely used in numerous works to perform tasks of the same kind [24], [31], [32], [40], [41], [43], [44]. Our goal with this analysis is not to achieve state-of-the-art-results, but to make a comparison among modalities with combinations of different inputs.

The modalities we analyze and the neural network architectures that we use with each one are as follows:

- **Audio modality**: small CNN with raw audio waveform as input and 3 dense layers plus output.
- **Visual modality**: modified pretrained VGG-Face architecture, with face images as input.
- **Audio+Visual modality**: fusion of both previously mentioned models already trained and 3 dense layers plus output layer.

- **Audio+Visual+handcrafted features modality**: fusion of trained network audio and modified VGG-Face, with a late fusion strategy of the handcrafted features, plus 3 dense layers and the output layer.

Next, we explain in more detail each of the neural network architectures mentioned previously that are used in this work.

5.2.1. Audio modality

The audio model is the simplest architecture in this paper. We use as input the raw waveform of the first five seconds of each video, since according to [24], better results are obtained only by using the first part of the audio. Even so, the input vector still has a length of 224938 values. The model is composed of three dense layers of 512, 128 and 32 neurons respectively, plus the output layer (see Figure 16).

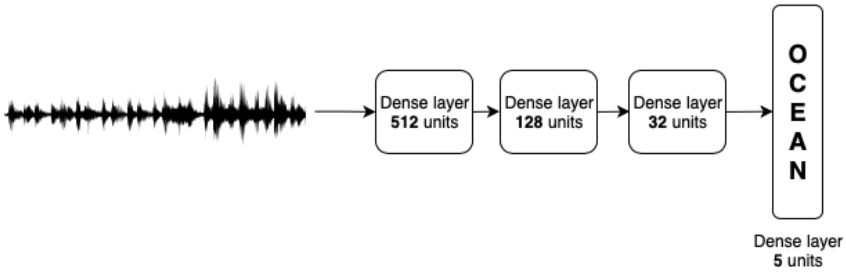


Fig. 16. Audio architecture used in our work.

5.2.2. Visual modality

The VGG-Face architecture has been selected for being a classic and simple model to use as a baseline. The input are face images with a resolution of 224x224 pixels and RGB channels. We performed network surgery to remove the last layer to add an extra convolutional layer with 512 filters and a kernel size of 1 and a max pooling layer with a pool size of 3x3. Then, it presents four dense layers of 1024, 512, 128 and 32 neurons respectively, plus the output layer. Finally, fine-tuning has been applied in two steps, first to just the new layers and afterwards to the whole network in order to improve its performance (see Figure 17).

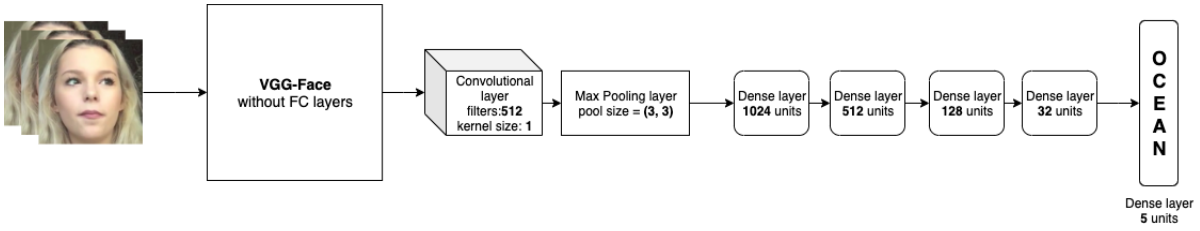


Fig. 17. Modified VGG-Face.

5.2.3. Audio and visual modality fusion

This network is the fusion of the two networks mentioned above, removing only the last layer of both once they have been trained. Three dense layers of 512, 128 and 32 neurons respectively, plus the output layer, are added. Finally, fine-tuning is performed to the whole network (see Figure 18).

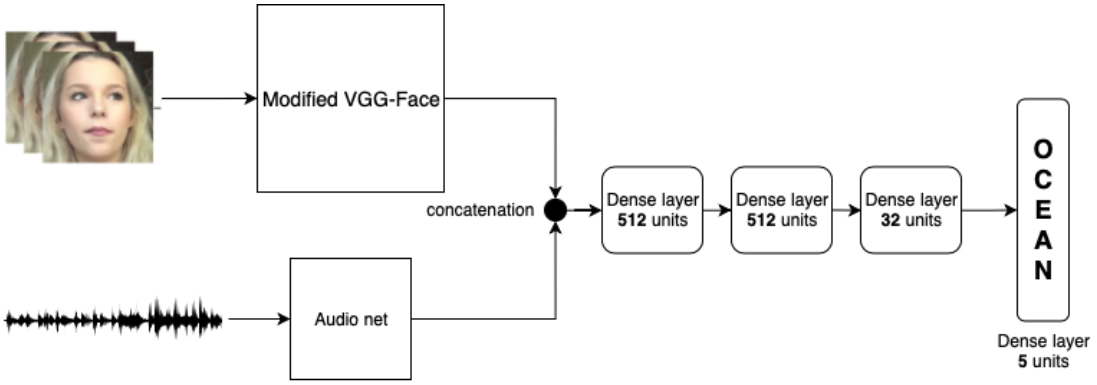


Fig. 18. Architecture created from modified VGG-Face and audio net concatenation.

5.2.4. Addition of high-level attributes

This network is similar to the previous one but with a late fusion strategy of high-level attributes. A vector with 37 values formed by Ekman's emotions histogram, gender and age values followed by two dense layers of 32 and 10 neurons each one, is concatenated to both audio and CNNs. Then, it presents exactly the same dense layers as the previous architecture and fine-tuning is also applied to the entire network (see Figure 19).

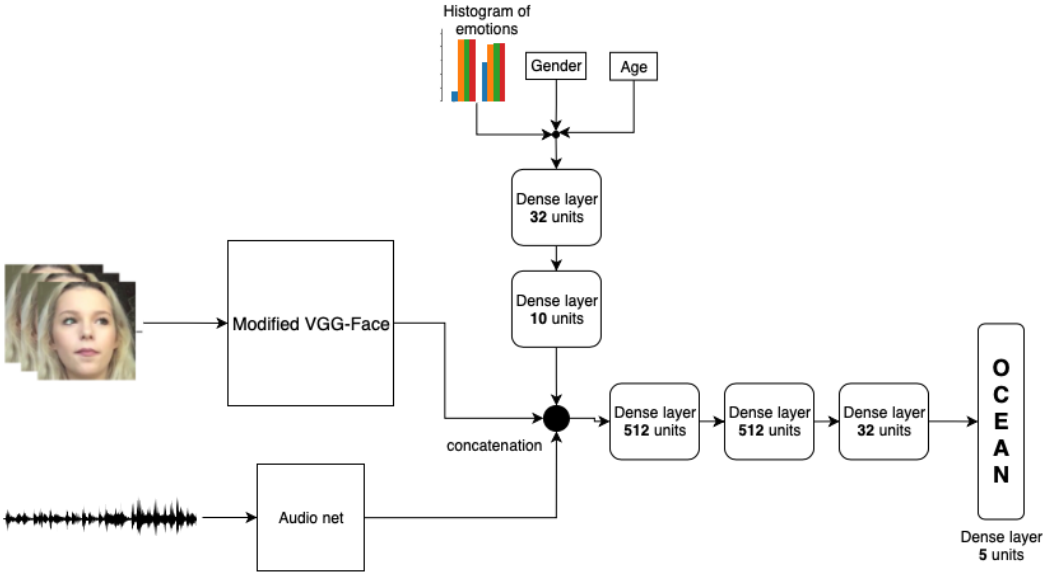


Fig. 19. Architecture created from the concatenation of modified VGG-Face, audio net and a late fusion strategy of handcrafted features.

6. Experimental protocol

In the first part of this section, we comment most relevant implementations details. First, we talk about the training strategy followed for all networks and then, the libraries and hardware used that made possible the implementation of the code. Afterwards, we make a detailed comparison of our four proposed modalities: Audio, Visual, Audio+Visual and Audio+Visual+handcrafted features modalities, through an analysis of the results obtained. With these experiments we demonstrate the influence of audio-visual signals and high-level features (Ekman's universal emotions, gender and age) on apparent personality prediction according to the Big-Five traits model. All experiments were performed using the First Impressions database. We compute the accuracy scores with the following formula:

$$acc_j = 1 - \frac{\sum_{i=1}^{N} |p_{ij} - gt_{ij}|}{N}$$

being p_{ij} the predicted value for frame i with trait j, gt_{ij} their respective ground truth value and N the number of frames in the test set [24].

6.1. Implementation details

6.1.1. Training strategy

The Audio modality network has been trained for 150 epochs and a batch size of 64 samples. For the Visual modality network, we followed a fine-tuning strategy, freezing all dense layers and training remaining layers for 50 epochs with a batch size of 32 samples. Afterwards, we have proceeded to unfreeze these layers and train the entire network for 20 more epochs. Once both previously mentioned architectures have been trained, we have concatenated them to form the Audio+Visual modality network, also adding three dense layers. The final model has been trained for 70 epochs and a batch size of 30 samples. Finally, we have the Audio+Visual+handcrafted features modality network, which follows exactly the same training strategy as the Audio+Visual network, but with a late-fusion strategy of handcrafted features. Afterwards, the network is trained for 100 epochs and a batch size of 32 samples.

All proposed models use Adam optimizer with a learning rate of 1e-05 and Mean Squared Error (MSE) as loss function for the training.

In the Appendix section, charts showing MSE vs. epochs of all models can be found.

6.1.2. Libraries and hardware

The programming language in which all the code has been developed is Python 2.7.3. We have chosen it for its simplicity and elegance and obviously for having an extensive selection of libraries for machine learning.

The library we have used for deep neural networks implementation is Keras 2.1.6 using TensorFlow as backend. We have selected Keras because it is one of the most powerful and easy-to-use open-source neural-network library written in Python and the one currently used by most part of the community for the development of deep learning models.

Below, we show a list of other complementary libraries that we have also used:

- pickle 2.0
- numpy 1.14.5
- zipfile 12.4
- shutil 10.10
- cv2 3.4.1
- math 9.2
- matplotlib 2.2.2
- tensorflow-gpu 1.8

Finally, we list the hardware used to carry out this work.

- MacBook Pro (13-inch display, 2018) with the following technical specifications:
  - Processor: 2,3 GHz Intel Core i5
  - RAM: 8 GB 2133 MHz LPDDR3
  - Graphics: Intel Iris Plus Graphics 655 1536 MB

- 4-GPUs server property of the University of Barcelona, where we have trained all the neural network models. The 4 GPUs model is: NVIDIA GeForce GTX TITAN X with 12GB of memory each one.

6.2. Experiments

6.2.1. Comparison of global accuracy scores per traits and modality

In this experiment, we compute the global accuracy scores per traits for each proposed modality to draw conclusions at a general level of their performance.

Table 1. Global accuracy scores per traits and modalities.

Modalities	O	C	E	A	N	Avg.
Audio	0.87397	0.87333	0.87357	0.87310	0.87362	0.87352
Visual	0.90525	0.91012	0.90783	0.90355	0.90437	0.90623
Audio+Visual	**0.90637**	**0.91447**	0.91171	0.90726	0.90465	0.90889
Audio+Visual+Handcrafted features	0.90623	0.91433	**0.91174**	**0.90753**	**0.90480**	**0.90893**

Table 1 shows the modalities´ global accuracy per personality traits and average. From the two modalities with only a raw input, clearly the winner is the Visual modality, both at the level of traits, as well as the average. This is because the visual signals transmit much more information than the auditory ones and this is noticeable when perceiving the personality from the human face, which offers the highest variability degree in a person [21] and therefore, has thousands of

facial features combinations such as the shape, size and color of lips, eyes, hair, nose, skin, etc.which significantly influence the person's traits prediction.

The Audio+Visual modality shows that the auditory signals, although to a lesser extent than the visual ones, can still improve considerably the results. The combination of both raw inputs increases the accuracy of all traits and the average with respect to the Visual modality; it even surpasses the Audio+Visual+Handcrafted features modality for the Openness and Conscientiousness traits. This last trait along with Extraversion and Agreeableness, are the most benefited from this fusion of inputs.

The modality that gives better results is the Audio+Visual+Handcrafted features. The inclusion of high-level features provides a new range of characteristics, that positively but subtly influence the results of Extraversion, Agreeableness and Neuroticism traits, as well as the average. These results demonstrate the neural network predicts slightly better personality traits with this combination of raw inputs complemented with these handcrafted features that we have chosen for our analysis following the paper [23].

6.2.2. Comparison of modalities´ accuracy scores per trait and gender

Next, we analyze each OCEAN trait by gender and modality. The objective of this experiment is to know if we can find some hints that show whether the proposed modalities have preferences or not in the prediction of the apparent personality according to the gender of the person.

Table 2. Accuracy scores per trait, gender and modality.

Modalities	O	C	E	A	N	Avg.
Audio	M 0.88042 / F 0.85193	M 0.86908 / F 0.86958	M 0.87535 / F 0.87213	M 0.88030 / F 0.87939	M 0.87850 / F 0.86750	M 0.87673 / F 0.86811
Visual	M 0.90456 / F 0.90582	M 0.90900 / F 0.91105	M 0.90819 / F 0.90753	M 0.90301 / F 0.90402	M 0.90401 / F **0.90467**	M 0.90576 / F 0.90662
Audio+Visual	M **0.90650** / F 0.90627	M **0.91362** / F 0.91517	M **0.91097** / F 0.91232	M **0.90700** / F 0.90747	M 0.90514 / F 0.90425	M **0.90865** / F 0.90910
Audio+Visual+handcrafted features	M 0.90563 / F **0.90628**	M 0.91259 / F **0.91546**	M 0.91070 / F **0.91248**	M 0.90686 / F **0.90787**	M **0.90563** / F 0.90461	M 0.90828 / F **0.90934**

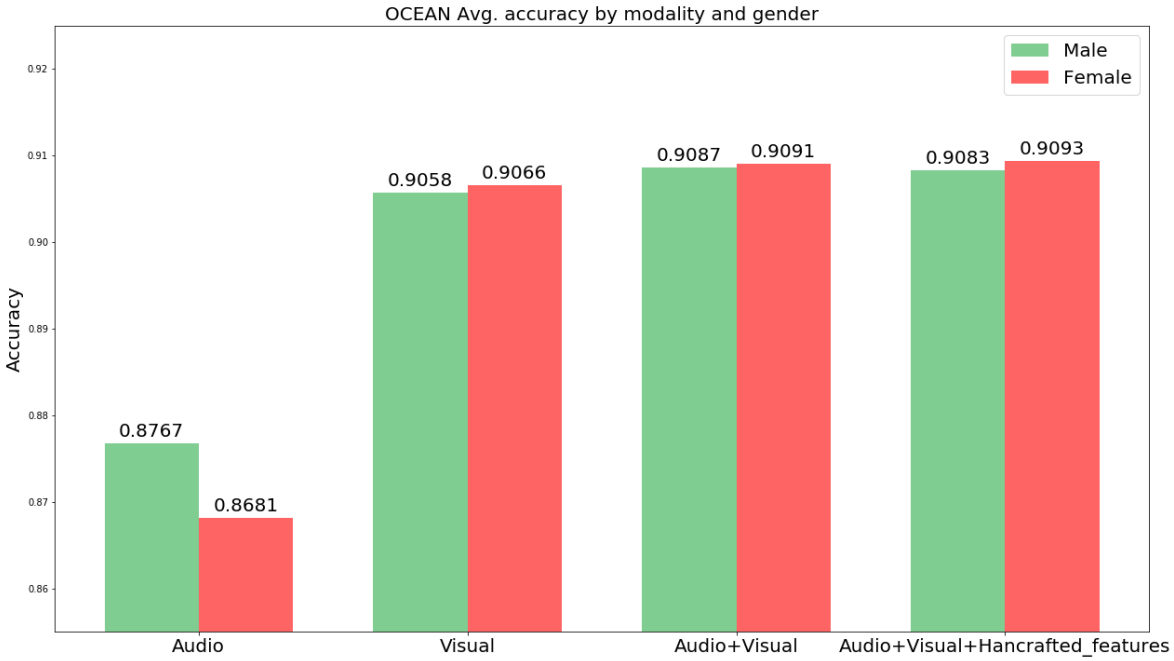


Fig. 20. Average accuracy of modalities per gender.

In Figure 20, the average accuracy per modality/gender is shown. The FI database is composed of 45% males and 55% females. If we look at the train+validation set (composed of 8000 videos in total), we only found 8.75% more females than males. This is a positive point in the analysis, since it shows that the results obtained are not influenced by a gender imbalance in the data. As it can be seen, in all modalities females obtain a higher average accuracy except for Audio modality. This shows that for the architecture with only audio as input, men's voices influence more than women's in the performance of personality predictions. In the rest of modalities, we can clearly see an improvement in the score of both genders as we merge raw inputs and high-level features, reaching a ceiling of 0.908 for males and a 0.909 for females.

Table 2 shows a comparison per trait and gender. We can see that, as we mentioned before for average accuracy, Audio modality at traits level also has better results for the male gender, except for the Conscientiousness trait by a very small difference. For the Visual modality, the results show an improvement in all traits favoring females more, however for the Extraversion trait the males have a slightly higher score. Audio+Visual modality continues and improves the tendency to better predict females in almost all traits. But not so for the Openness and Neuroticism traits. The Audio+Visual+handcrafted features modality stands out in all traits to females except for Neuroticism.

6.2.3. Comparison of OCEAN traits per age ranges

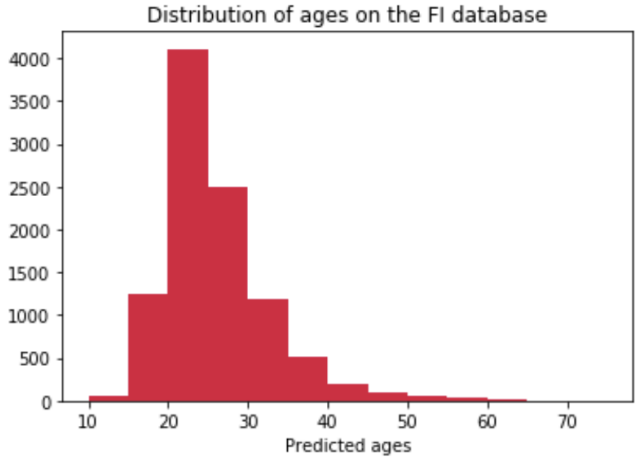


Fig 21. Distribution of ages on the FI dataset.

The FI dataset has an age range of 11-64 years, but if we analyze its distribution (see Figure 21), the ages of the observed people are concentrated mostly from 20-25 years, which is negative for the analysis of age-ranges with fewer samples. For this experiment, we split the test set into six different groups based on the following age ranges: 0-15, 15-25, 25-35 , 35-45, 45-55 and 55-65. The bar chart (see Figure 22), represents the average accuracy per modality, divided by age-ranges. On the other hand, figures 23 and 24 represent the accuracy scores per traits and age-ranges of modalities Audio and Audio+Visual+handcrafted features, respectively. The reason why we have chosen to show the aforementioned charts for only two modalities, is because the Visual and Audio+Visual modalities have a behavior quite similar to the model represented by Figure 24 (Audio+Visual+handcrafted features).

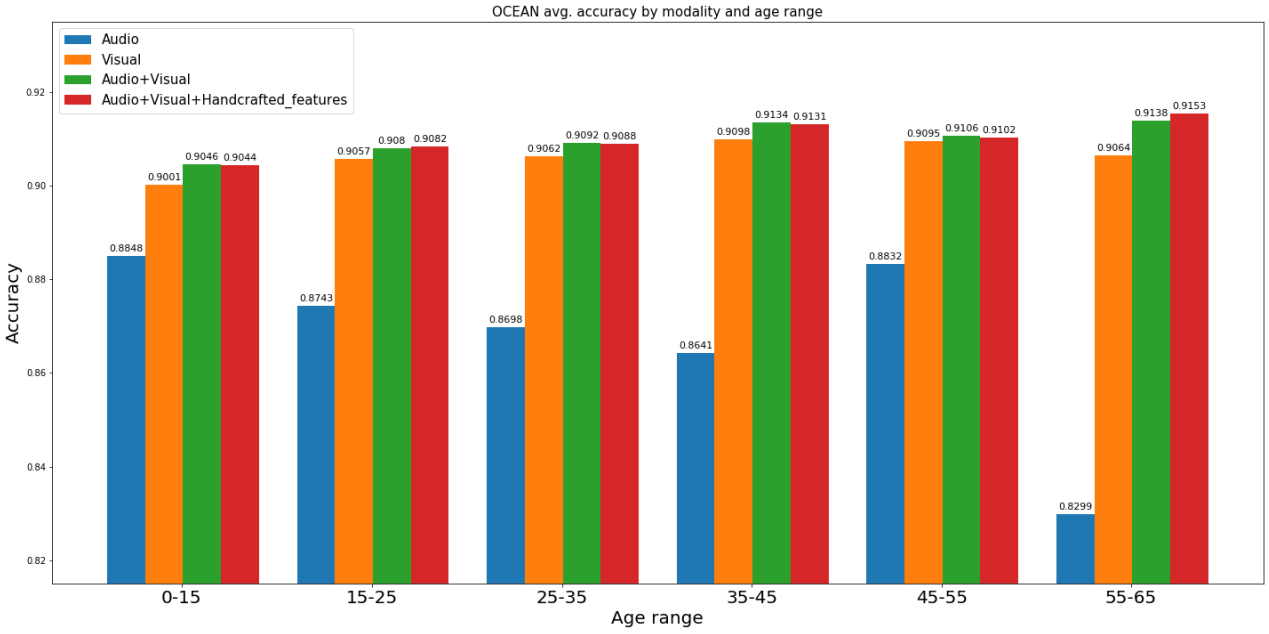


Fig. 22. Average accuracy per age-range and modality.

According to the analysis in Figure 22, Audio modality better predicts personality for the age range of 0-15 with an accuracy of 0.8848, followed by the range of 45-55 years. Where the personality is predicted worse, with a difference of 0.0549 points with respect to its best score, is for the range of 55-65 years, this may be due precisely to what we said before, the database is very poorly balanced with respect to the range of ages and it is clear that this model is the most affected. This might be justified, since a neural network as the one we analyze with only raw audio as input, has its limitations, and it could be more difficult to learn to generalize certain characteristics with less data, since apparently it is more difficult for this model to predict personality as the age increases, except for the age range of 45-55. On the other hand, the Visual modality analyzes face images, from which many more features can be extracted that greatly benefit the task of predicting the personality and therefore, obtains better results than the previous modality, obtaining its highest accuracy in the range of 35-45, followed by the age range of 45-55. The modality that fuses visual and auditory signals improves its accuracy in all age ranges with respect to the Visual modality, obtaining its highest score in the range of 55-65 years with 0.9138. We can also observe that it is subtly better in several of the age ranges than the modality with the handcrafted features, but nothing conclusive. Finally, Audio+Visual+handcrafted features modality obtains the best accuracy of all models in the range of 55-65 with 0.9152. Except for the audio-only model, the rest of the modalities predict personality worse for the range of 0-15 years and where the accuracy scores are more balanced is in the range of 45-55.

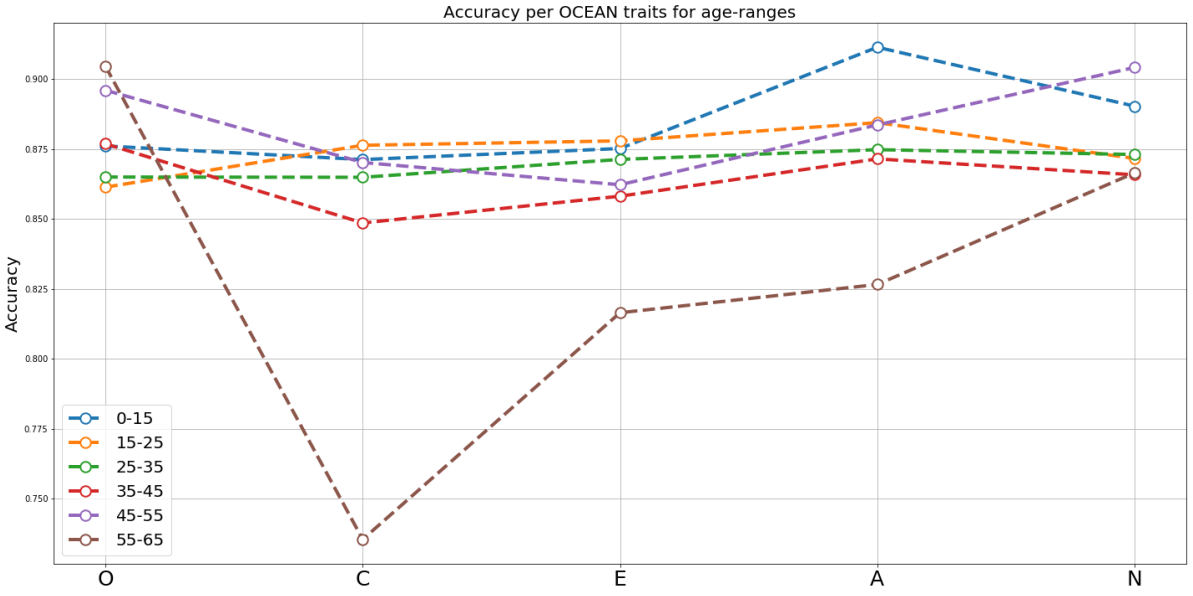


Fig. 23. Accuracy per OCEAN trait and age-range for *Audio modality*.

Looking at Figure 23, we can say that for model with only audio as input, the age-range of 55-65 is the one showing more variability for all traits, presenting the highest score for Openness, but at the same time with the least accuracy score in the rest of traits especially Conscientiousness, which decreases considerably. More accurately Agreeableness is predicted for the range of 0-15 years, which is in turn, is the highest score among all traits. According to the chart, the highest Neuroticism is held by the age-range of 45-55. The rest of the ranges have fewer differences between them in each trait.

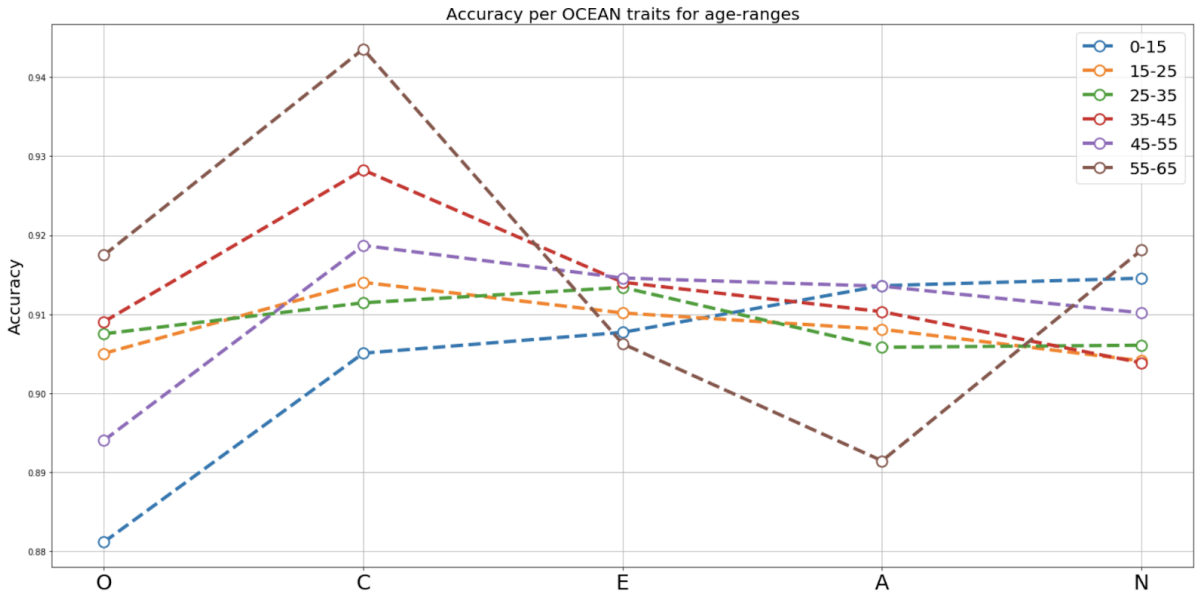


Fig. 24. Accuracy per OCEAN trait and age-range for *Audio+Visual+handcrafted features modality.*

If we analyze Figure 24, we can say this modality presents the highest variability for the age-range of 55-65, as happened for Audio modality, but curiously in this case, for Conscientiousness archive the best accuracy score. It is remarkable how the range of 0-15 presents the lowest score of the entire chart, belonging to the Openness trait. Agreeableness remains the trait in which the range of 55-65 reaches the lowest accuracy, although if we compare it with that in Figure 23, it has improved. Where the smallest difference in traits score is seen is in Extraversion, reaching the range of 45-55 the highest accuracy. In general, we can observe that for all traits, with respect to Figure 23, this modality achieves higher accuracy scores in all age ranges.

6.2.4. Comparison of OCEAN traits with Ekman's universal emotions

To perform this experiment, we do as follows:

1. First, we select from each frame the emotion with highest confidence.
2. Then, we separate the frames into seven groups (one for each emotion with highest confidence).
3. Each frame has its traits prediction, so we compute the mean accuracy for each of the seven groups per trait.

The result is five bar charts (one chart per trait), where we will compare the accuracy per trait and emotions for each of the four modalities.
We will proceed with the analysis in two steps. First, individually per chart and then, establishing a relationship among them.

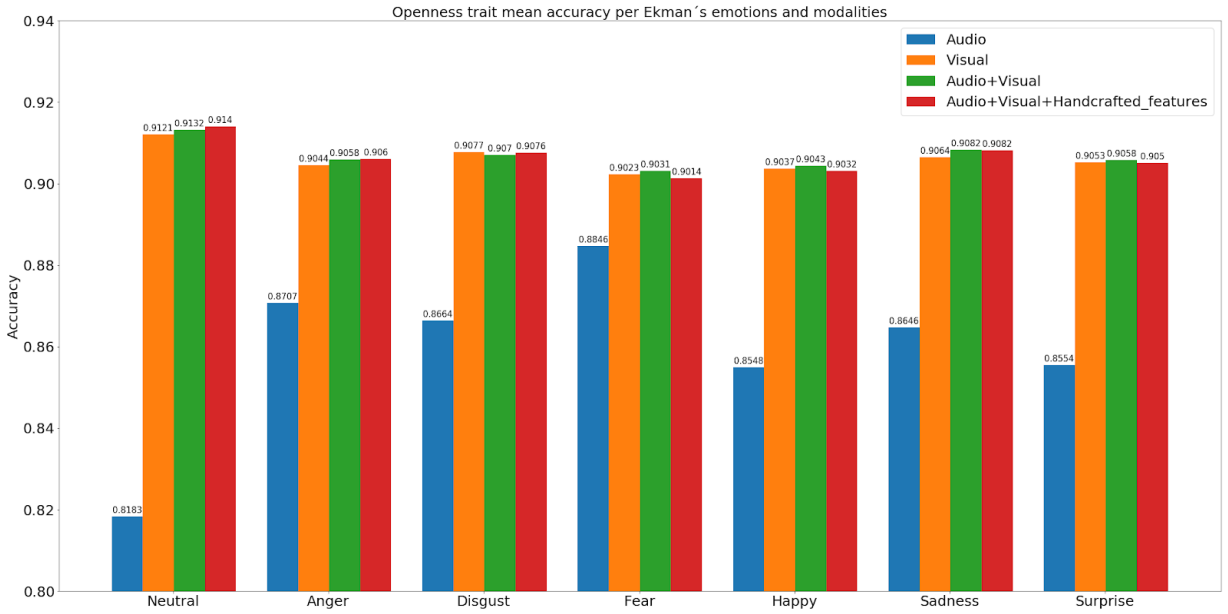


Figure 25. *Openness* trait mean accuracy per Ekman's universal emotions and modalities.

In Figure 25, we observe that for the Openness trait, the Neutral emotion is the one that presents the best results in all modalities that include visual information in this chart. In fact, for this emotion, an improvement is observed as the modalities merge to reach a ceiling of 0.914 accuracy. However, for the audio-only model, it is the worst result with 0.8183 accuracy and not only that, it is the worst overall for this trait. This makes some sense, since the person being neutral usually transmits less significant auditory signals than when expressing another emotion, therefore there is less audio information that allows for better prediction. Also, we can interpret that for the network, knowing that the person is in a neutral state is beneficial because it allows the model to make Openness prediction more accuracy. For the Fear emotion, the Audio modality has a greater accuracy, it may be because that emotion is noticed more than others in the voice and is especially characteristic to better predict this trait. After Neutral, the most prominent emotions for this trait are Disgust and Sadness. In the first one, the Visual modality stands out slightly with an accuracy of 0.9077 and in the other one, the modalities Audio+Visual and Audio+Visual+handcrafted features are tied with a score of 0.9082.

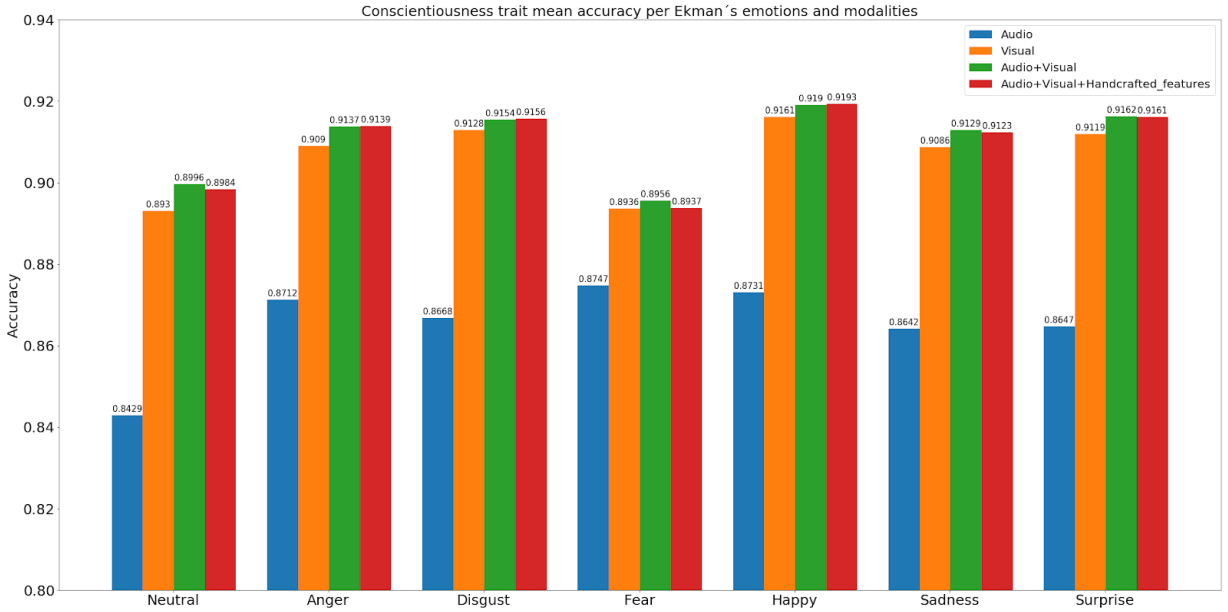


Fig. 26. *Conscientiousness* trait mean accuracy per Ekman's universal emotions and modalities.

In Figure 26, we observe that for Conscientiousness, the emotion that stands out is Happy, obtaining better results in all modalities with visual information, reaching with the Audio+Visual+handcrafted features an accuracy of 0.9193. The relationship between this emotion and the Conscientiousness trait can be interpreted as that the people analyzed are usually concentrated, with the ideas to express very clear, but expressing joy while looking at the camera. The worst result for this emotion is obtained by the Audio modality and this makes sense, since a person with a high capacity for concentration who is happy before the camera, does not usually alter his/her voice enough to make it understood by only audio, just be smiling for example. The emotion with the worst results for this trait is Fear. People with high Conscientiousness, presents a level of self-control greater than the rest, therefore it can be understandable that it costs them to exalt and express fear clearly.

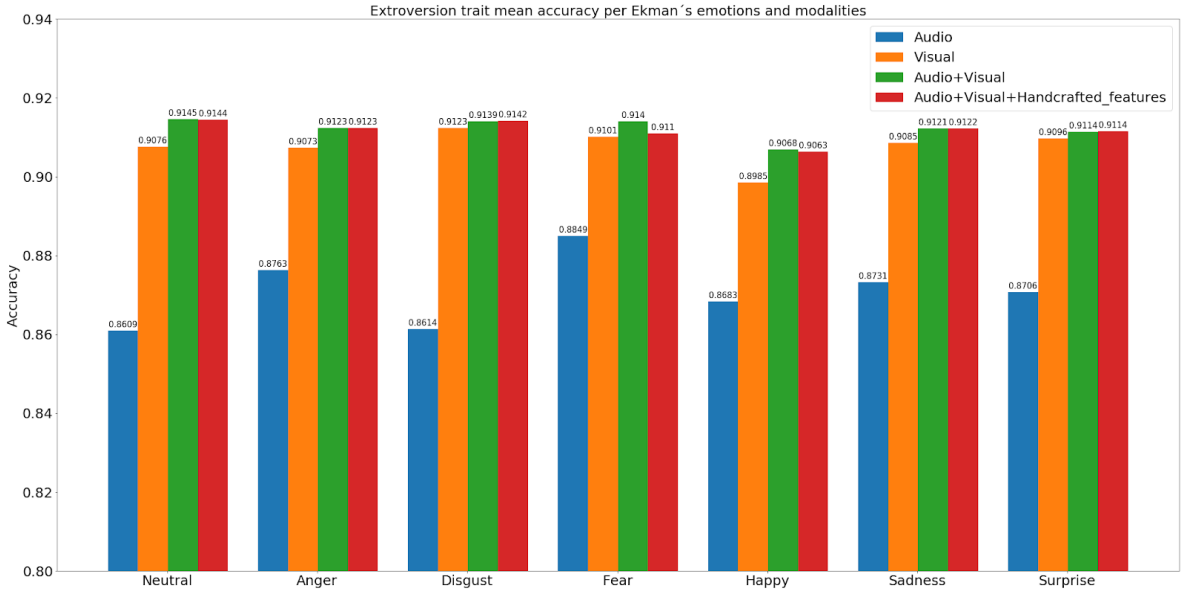


Fig. 27. *Extraversion* trait mean accuracy per Ekman's universal emotions and modalities.

In Figure 27, we observe that for the Extraversion trait, the emotion that stands out is Neutral, followed by Disgust and Fear. For the first one, Audio+Visual modality reaches 0.9145 accuracy and for the rest, the modality that adds handcrafted features with 0.9142 stands out. The relationship among emotions Neutral, Disgust and Fear with this trait could mean that the most common is that the people analyzed, are a large part of the clip, with a neutral expression facing the camera. They are open and feel comfortable expressing themselves to other people, in this case with some disgust or fear about the topic they are dealing with in the video. The emotion with the worst results for this trait is Happy, the Visual modality being the most affected among those who use visual information with a 0.8985 accuracy. However, the Audio modality is still the worst globally in this chart, with 0.8609 accuracy in the Neutral expression, possibly due to the same reason we commented with Figure 25, the poor auditory information in people with such facial expression.

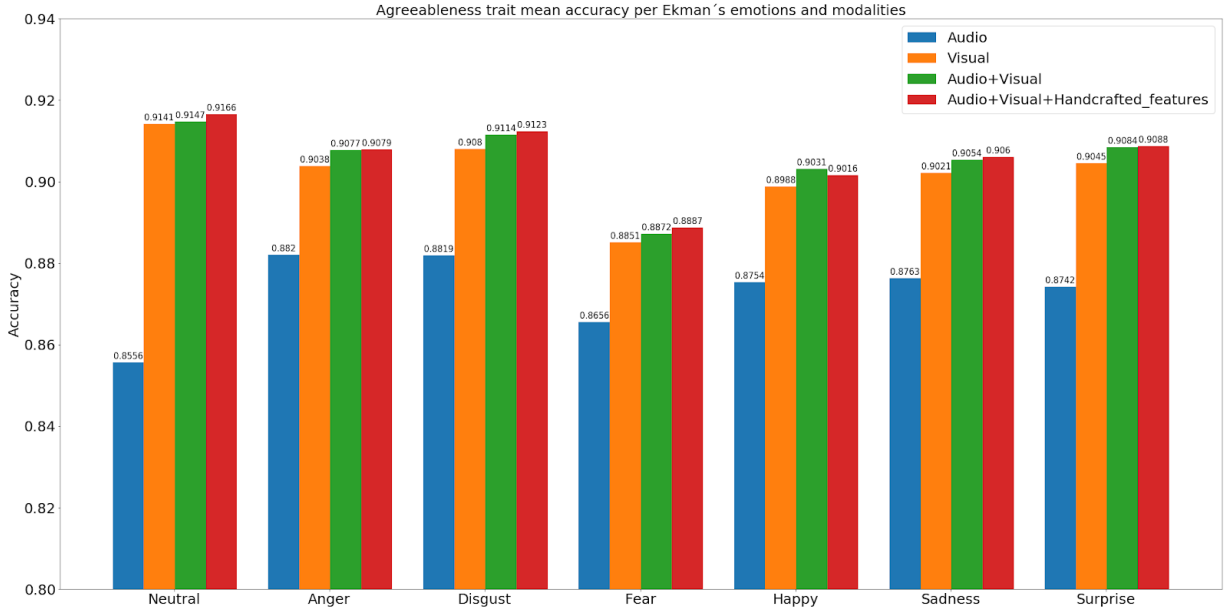


Fig. 28. *Agreeableness* trait mean accuracy per Ekman's universal emotions and modalities.

In Figure 28, we observe that for the Agreeableness trait, the emotion that stands out is Neutral, followed by Disgust and Surprise. For the first one, a 0.9166 accuracy is achieved with the modality that adds handcrafted features. While 0.9123 and 0.9088 are obtained for the second and third emotions respectively. People who have a high level of Agreeableness, are usually tolerant and they are very understandable with the problems and emotions of others, so it is possible that they feel a certain disgust or surprise caused by their level of sympathy. We can observe that the Audio modality had a greater influence on Anger and Disgust with respect to the rest of emotions.

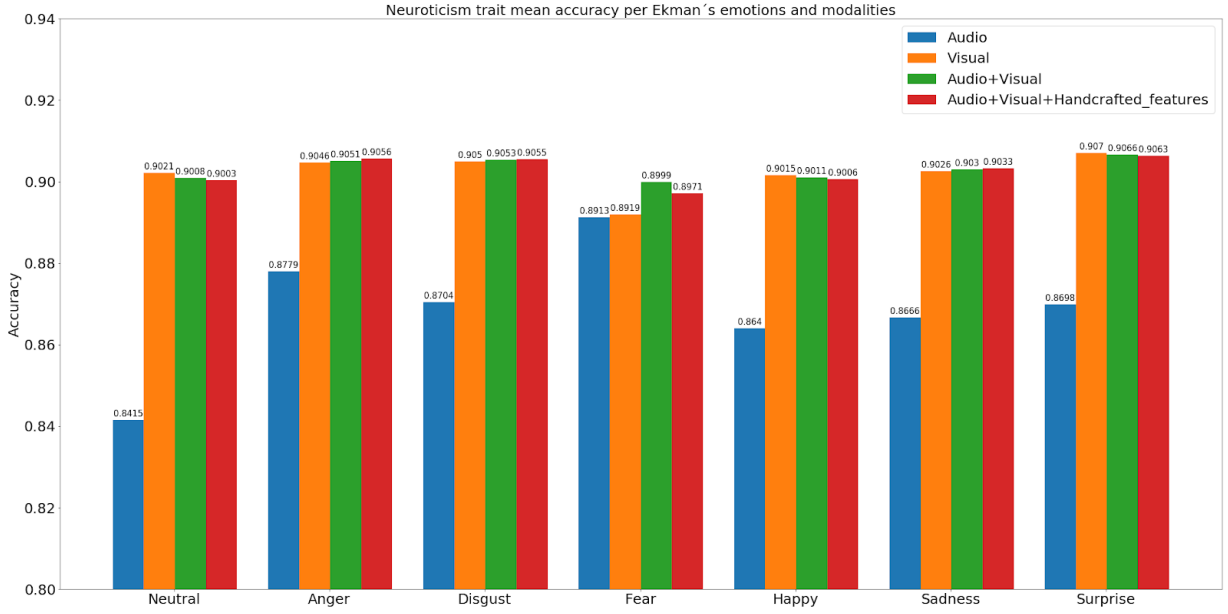


Fig. 29. *Neuroticism* trait mean accuracy per Ekman's universal emotions and modalities.

In Figure 29, we observe that for the Neuroticism trait, the emotion that stands out the most is Surprise with the Visual modality, obtaining 0.9070 accuracy. The next emotion with better results would be Anger, followed by Disgust with 0.9056 and 0.9055 accuracy respectively, both obtained with the Audio+Visual+handcrafted features modality. To justify the reason for these emotions linked to this trait, we can argue that a person who has high scores of this trait is usually emotionally unstable, vulnerable to stress and interpret ordinary situations as threatening. That is why they tend to experience negative emotions and be easily surprised at any situation. It is curious to observe, as the Audio modality practically ties with the Visual modality in the Fear emotion (0.8913 versus 0.8919). It seems that auditory cues greatly influence the relationship of this emotion with the Neuroticism trait. This may be due to the fact that very nervous and unstable people tend to express themself verbally when they feel fear, raising their tone of voice and emitting louder sounds, enriching the information of the analyzed audio.

Once the bar charts have been analyzed separately, we will proceed with a global analysis of them.

Table 3. Table extracted from the analysis of the previous graphs. It shows the traits that work best and worst for each emotion as well as the modalities responsible for those accuracy scores.
Note: we have also included the traits that have the worst results if Audio modality is not considered, since it gives the worst results for all emotions so we found interesting to do the analysis without taking it into account as well.

Emotion	Works better for	Works worst for
Neutral	**Agreeableness** (0.9166 with Audio+Visual+handcrafted features modality) **Extraversion** (0.9145 with Audio+Visual modality)	**Openness** (0.8183 with Audio modality) --- *(If Audio modality is NOT considered)* **Conscientiousness** (0.893 with Visual modality)
Anger	**Conscientiousness** (0.9139 with Audio+Visual+handcrafted features modality) **Extraversion** (0.9123 with Audio+Visual+handcrafted features modality)	**Openness** (0.8707 with Audio modality) --- *(If Audio modality is NOT considered)* **Agreeableness** (0.9038 with Visual modality)
Disgust	**Conscientiousness** (0.9156 with Audio+Visual+handcrafted features modality) **Extraversion** (0.9142 with Audio+Visual+handcrafted features modality)	**Extraversion** (0.8614 with Audio modality) --- *(If Audio modality is NOT considered)* **Neuroticism** (0.9050 with Visual modality)

Fear	**Extraversion** (0.9140 with Audio+Visual modality) **Neuroticism** (0.8999 with Audio+Visual+handcrafted features modality)	**Agreeableness** (0.8656 with Audio modality) --- *(If Audio modality is NOT considered)* **Agreeableness** (0.8851 with Visual modality)
Happy	**Conscientiousness** (0.9193 with Audio+Visual+handcrafted features modality) **Extraversion** (0.9068 with Audio+Visual modality)	**Openness** (0.8548 with Audio modality) --- *(If Audio modality is NOT considered)* **Extraversion** (0.8985 with Visual modality)
Sadness	**Conscientiousness** (0.9129 with Audio+Visual modality) **Extraversion** (0.9122 with Audio+Visual+handcrafted features modality)	**Conscientiousness** (0.8642 with Audio modality) --- *(If Audio modality is NOT considered)* **Agreeableness** (0.9021 with Visual modality)
Surprise	**Conscientiousness** (0.9162 with Audio+Visual modality) **Extraversion** (0.9114 with Audio+Visual+handcrafted features modality)	**Openness** (0.8554 with Audio modality) --- *(If Audio modality is NOT considered)* **Agreeableness** (0.9045 with Visual modality)

From the Table 3, we can conclude the following:

- The Extraversion and Conscientiousness traits get the highest accuracy scores in almost all emotions, this might be due to a bias from the annotators side.
- The highest accuracy score is for the Conscientiousness trait with 0.9193 and is associated with the Happy emotion.
- Audio modality got the worst results in all emotions, especially for Neutral emotion and Openness trait, because there is possibly not enough audio information when the person remains neutral and therefore results in a prediction with low confidence (see Openness chart analysis for more information). However, it better predicts the Fear emotion for almost all traits, almost matching the result of the Visual modality in Neuroticism as we have mentioned before (see Neuroticism chart analysis for more information), except for Agreeableness where it is the lowest, since a person who feels fear, is uneasy and unkind.
- It is shown that when using visual information to predict personality, the accuracy of traits linked to emotions, increases considerably, and the fusion of Audio+Visual information and high-level features slightly improves the results.

7. Conclusions

In this work we have studied the influence and relationship of different characteristics of observed subjects in an automatic personality perception setup, being able to partially explain some of them. To do this, we have performed a comparative analysis of four proposed deep neural networks trained on the FI database. These 4 models provide different combinations of audiovisual information and high-level features (emotions of facial expressions, age and gender) to regress apparent personality traits scores based on the Big-Five model.

Experiments showed that with the fusion of raw audio, sequences of face images and handcrafted features, the best results are obtained. There are a variety of possible biases linked to apparent personality perception. Future work may include the analysis of other complementary sources of information such as background and clothing understanding, upper-body gestures, heart rate, audio transcription, other camera angles, as well as other attributes such as attractiveness, ethnicity and nationality. In addition, we could also regress real personality on the same data and try to establish a link between apparent and real personality. We could go even further and extend the study to the analysis of the relationship between two or more people in the same scene.

8. References

[1] Towards Data Science. *Introduction to Artificial Neural Networks (ANN)* https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9

[2] The human memory. *Brain neurons and synapses* https://human-memory.net/brain-neurons-synapses/

[3] CS231n Convolutional Neural Networks for Visual Recognition, Stanford University. *Biological motivation and connections* http://cs231n.github.io/neural-networks-1/#intro

[4] Marvin Minsky and Seymour Papert (1969). *Perceptrons: an introduction to computational geometry.*

[5] Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016). *Deep Learning.*

[6] Terence Parr and Jeremy Howard (2018). *The Matrix Calculus You Need For Deep Learning.*

[7] Diederik P Kingma and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization.*

[8] Yann LeCun, León Bottou, Yoshua Bengio and Patrick Haffner (1998). *Gradient-Based Learning Applied to Document Recognition.*

[9] Geoffrey E. Hinton, Ilya Sutskever and Alex Krizhevsky (2012). *ImageNet Classification with Deep Convolutional Neural Networks.*

[10] ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012) *http://image-net.org/challenges/LSVRC/2012/results.html*

[11] Karen Simonyan and Andrew Zisserman from the Visual Geometry Group, University of Oxford (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition.*

[12] ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC 2014) *http://image-net.org/challenges/LSVRC/2014/results*

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun (2015). *Deep Residual Learning for Image Recognition.*

[14] ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC 2015) *http://image-net.org/challenges/LSVRC/2015/results*

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich (2014). *Going Deeper with Convolutions.*

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi (2016). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*

[17] Gordon W. Allport (1937). *Personality: a psychological interpretation.*

[18] Alessandro Vinciarelli Member, IEEE, and Gelareh Mohammadi (2014). *A Survey of Personality Computing*.

[19] V.Ponce-López,B.Chen,M.Oliu,C.Corneanu,A.Clapés,I.Guyon, X. Baró, H. J. Escalante and S. Escalera (2016). *Chalearn lap 2016: First round challenge on first impressions-dataset and results.*

[20] O. M. Parkhi, A. Vedaldi and A. Zisserman from the Visual Geometry Group, University of Oxford (2015). *Deep Face Recognition.*

[21] C.A.Sutherland, A.W.Young,and G.Rhodes (2017). *Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties*.

[22] Paul Ekman (1970). *Universal Facial Expressions of Emotion*.

[23] Yağmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven and Rob van Lier (2016). *Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition*.

[24] Ricardo Darío Pérez Principi, Cristina Palmero, Julio C. S. Jacques Junior, and Sergio Escalera (2019). *On the Effect of Observed Subject Biases in Apparent Personality Analysis from Audio-visual Signals*.

[25] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. Jacques Junior, M. Madadi, S. Ay- ache, E. Viegas, F. Gürpinar, A. S. Wicaksana, C. C. S. Liem, M. A. J. van Gerven, and R. van Lier (2018). *Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos.*

[26] T. Yeo (2010). *Modeling personality influences on YouTube usage.*

[27] L. Qiu, H. Lin, J. Ramsay, and F. Yang (2012). *You are what you tweet: Personality expression and perception on twitter*.

[28] D. Quercia, D. Las Casas, J. Pesce, D. Stillwell, M. Kosinski, V. Almeida, and J. Crowcroft (2012). *Facebook and privacy: The balancing act of personality, gender, and relationship currency*.

[29] R. R. McCrae and O. P. John (1992). *An introduction to the five-factor model and its applications.*

[30] Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier and Sergio Escalera (2018). *First Impressions: A Survey on Vision-based Apparent Personality Trait Analysis.*

[31] J. I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez (2012). *Facetube: predicting personality from facial expressions of emotion in online conversational video.*

[32] F. Gürpinar, H. Kaya, and A. A. Salah (2016). Combining deep facial and ambient features for first impression estimation.

[33] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli (2014). *Face-based automatic personality perception*.

[34] C. C. Ballew and A. Todorov (2007). *Predicting political elections from rapid and unreflective face judgments.*

[35] R. J. Vernon, C. A. Sutherland, A. W. Young and T. Hartley (2014). *Modeling first impressions from highly variable facial images*.

[36] L. Batrinca, N. Mana, B. Lepri, F. Pianesi and N. Sebe (2011). *Please, tell me about yourself: automatic personality assessment using short self-presentations*.

[37] S. Okada, O. Aran and D. Gatica-Perez (2015). *Personality trait classification via co-occurrent multiparty multimodal event discovery.*

[38] S. C. Guntuku, L. Qiu, S. Roy, W. Lin and V. Jakhetiya (2015). *Do others perceive you as you want them to? Modeling personality based on selfies.*

[39] A. Todorov and J. Porter (2014). *Misleading first impressions different for different facial images of the same person*.

[40] A. Dhall and J. Hoey (2016). *First impressions - predicting user personality from twitter profile images.*

[41] F. Gürpinar, H. Kaya, and A. A. Salah (2016). *Multimodal fusion of audio, scene, and face features for first impression estimation*.

[42] H. J. Escalante, V. Ponce, J. Wan., M. Riegler, C. B., A. Clapes, S. Escalera, I. Guyon, X. Baro, P. Halvorsen, H. Müller and M. Larson (2016). *Chalearn joint contest on multimedia challenges beyond visual analysis: An overview.*

[43] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu (2016). *Deep bimodal regression for apparent personality analysis.*

[44] C. Ventura, D. Masip, and A. Lapedriza (2017). *Interpreting CNN models for apparent personality trait regression.*

[45] G. Levi and T. Hassner (2015). *Age and gender classification using convolutional neural networks.*

[46] Bargh, John & Gollwitzer, Peter & Oettingen, Gabriele. (2010). *The Handbook of Social Psychology.*

[47] Herbert Robbins and Sutton Monro (1951). *A Stochastic Approximation Method.*

[48] Rumelhart, David E.; Hinton, Geoffrey E.; Williams and Ronald J. (1986). *Learning representations by back-propagating errors.*

[49] J. L. Rosa Ramos, A. Cencerrado and S. Escalera (2018) *. Deep Learning for Universal Emotion Recognition in Still images.*

[50] W. Chan, R. R. McCrae, F. De Fruyt, L. Jussim, C. E. Löckenhoff, M. De Bolle, P. T. Costa, A. R. Sutin, A. Realo, J. Allik et al (2012). *Stereotypes of age differences in personality traits: Universal and Accurate?*

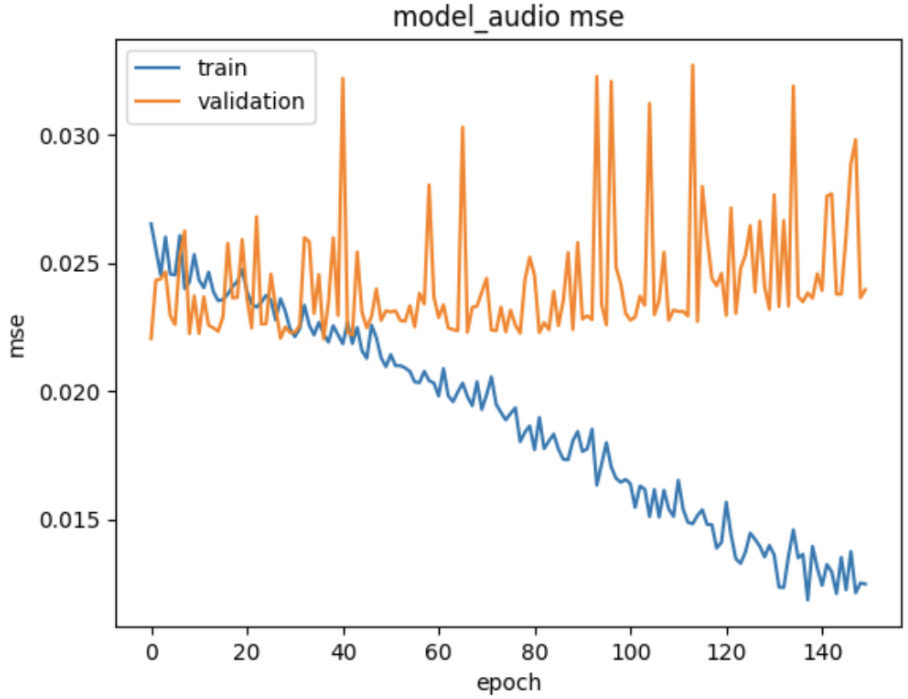
[51] G. Guo and G. Mu (2014). *A framework for joint estimation of age, gender and ethnicity on a large database.*

[52] H. Han, A. K. Jain, F. Wang, S. Shan and X. Chen (2018). *Heterogeneous face attribute estimation: A deep multi-task learning approach.*

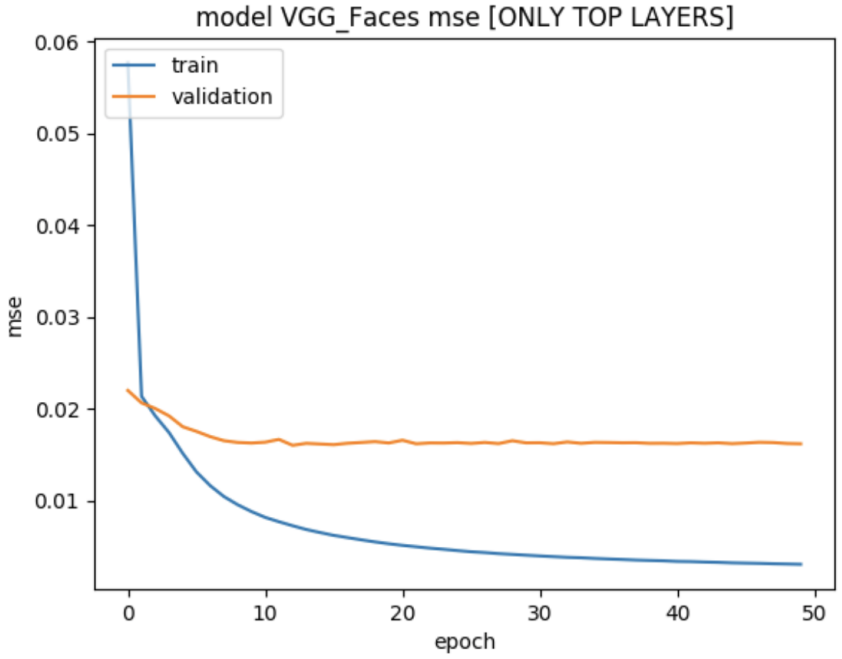
[53] Vahid Kazemi and Josephine Sullivan (2014). *One Millisecond Face Alignment with an Ensemble of Regression Trees.*

[54] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou and M. Pantic (2016). *Special Issue on Facial Landmark Localisation "In-The-Wild"*.

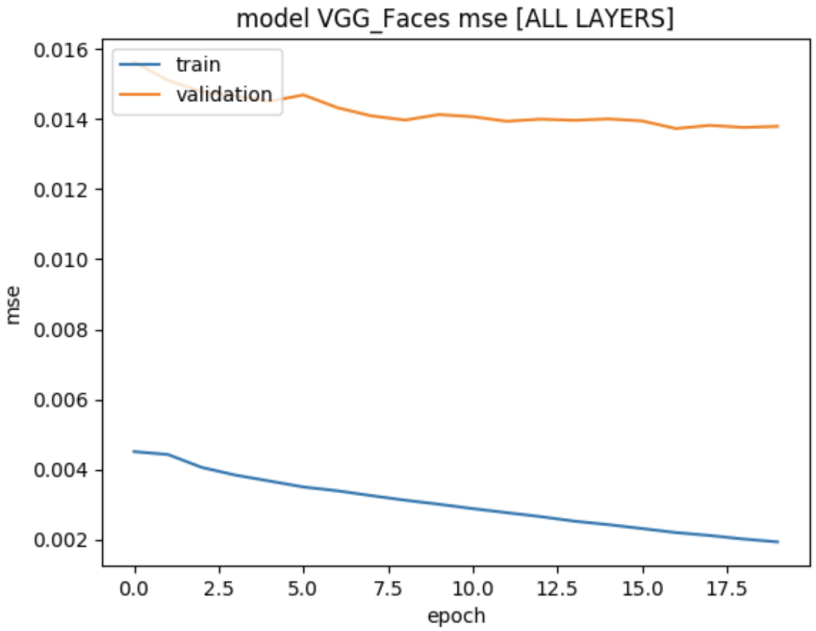
9. Appendix



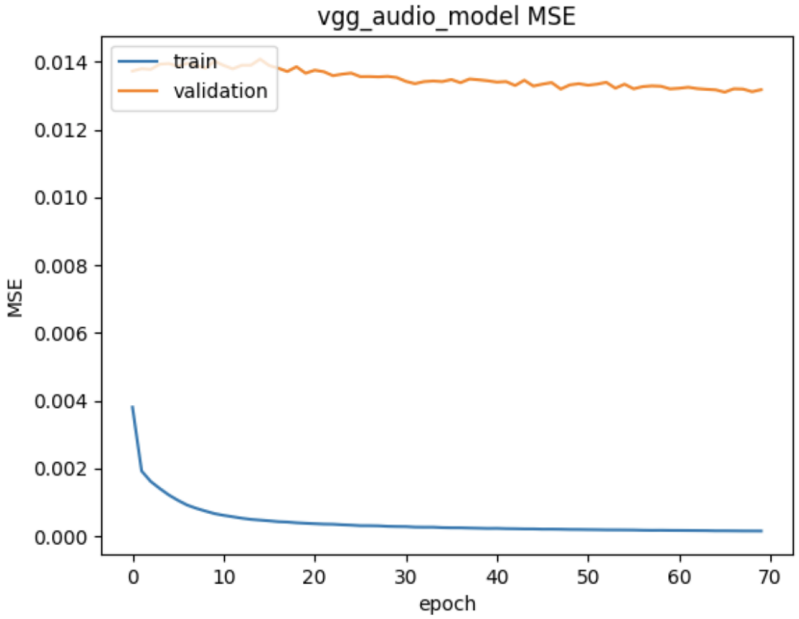
Appendix A. MSE vs. epochs. Training **Audio architecture** for 150 epochs.



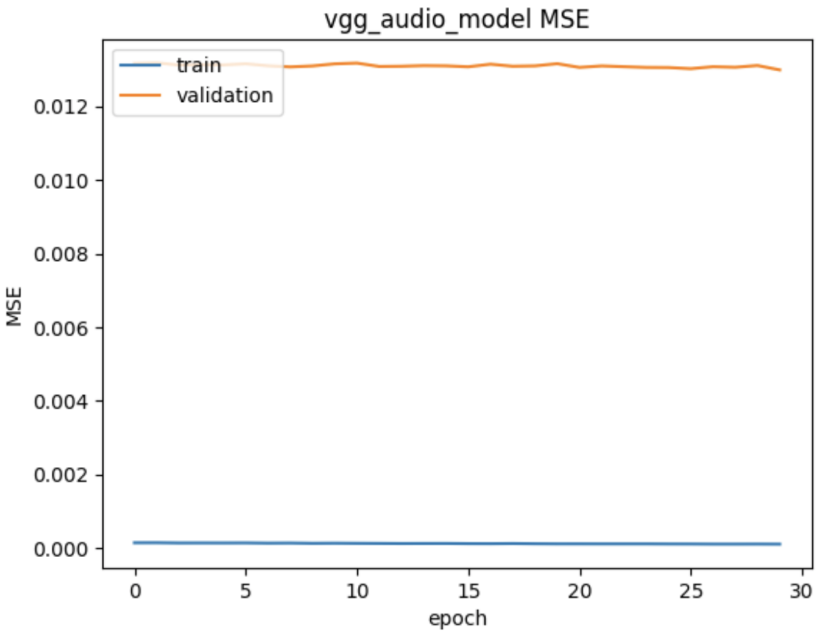
Appendix B. MSE vs. epochs. Training **Modified VGG-Face architecture** (only top-layers**)** for 50 epochs.



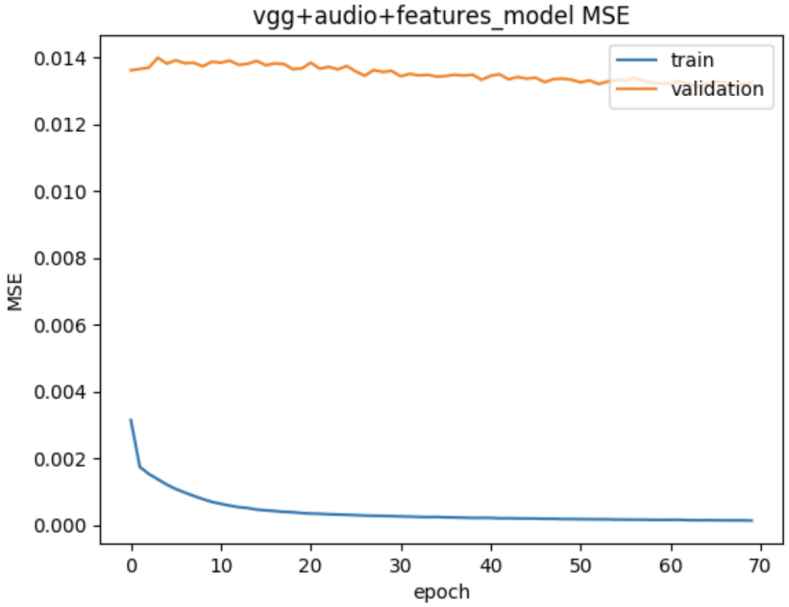
Appendix C. MSE vs. epochs. Training **Modified VGG-Face architecture** (all layers and after its top-layers were trained) for 20 epochs.



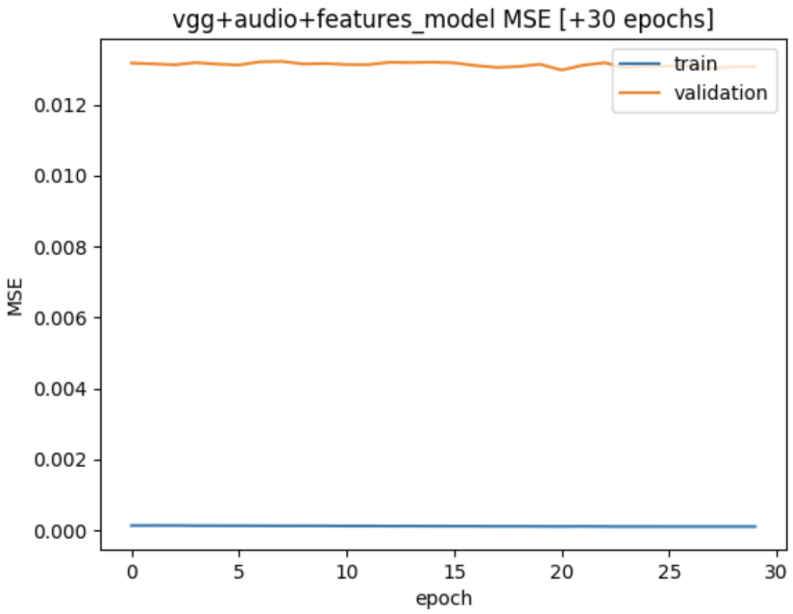
Appendix D. MSE vs. epochs. Training **Modified VGG-Face+Audio architecture** (once modified VGG-Face and audio net were trained separately) for 70 epochs.



Appendix E. MSE vs. epochs. Training **Modified VGG-Face+Audio architecture** for 30 epochs more.



Appendix F. MSE vs. epochs. Training **Modified VGG-Face+Audio+handcrafted features architecture** (once modified VGG-Face and audio net were trained separately) for 70 epochs.



Appendix G. MSE vs. epochs. Training **Modified VGG-Face+Audio+handcrafted features architecture** (after being trained 70 epochs) for 30 epochs more.