REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

AGE AND GENDER DETECTION BY FACE SEGMENTATION AND MODEFIED CNN ALGORITHM

Ahmed Raed Sabah ALRASHED

Master's Thesis

Supervisor

Asst. Prof. Dr. Timur İNAN

Istanbul, 2023

AGE AND GENDER DETECTION BY FACE SEGMENTATION AND MODEFIED CNN ALGORITHM

Ahmed Raed Sabah ALRASHED

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled AGE AND GENDER DETECTION BY FACE SEGMENTATION AND MODEFIED CNN ALGORITHM prepared by Ahmed Raed Sabah ALRASHED and submitted on 00/00/2023 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

_____

Asst. Prof. Dr. Timur INAN

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Timur INAN	Department of Software Engineering, Altınbaş University	_____
Asst. Prof. Dr. Mesut ÇEVİK	Department of Electrical and Electronics Engineering, Altınbaş University	_____
Asst. Prof. Dr. Vedat ESEN	Department of Electrical and Electronics Engineering, İstanbul Topkapı University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis. Submission date of the thesis to Institute of Graduate Studies: ___/___/___

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Ahmed Raed Sabah ALRASHED

Signature

DEDICATION

I devote and pledge this research work to my supervisor who is salient for guiding me through the whole research work as well as my family for always assisting me in my hard time.

PREFACE

At the outset, I would like to thank Asst. Prof. Dr. Timur İNAN, whose supervision and giving him so much time I had the first hand in the exit of this scientific thesis in the form in which it appeared, and his guidance and advice were instrumental in the completion of my scientific studies.

ABSTRACT

AGE AND GENDER DETECTION BY FACE SEGMENTATION AND MODEFIED CNN ALGORITHM

ALRASHED, Ahmed
M.Sc., Information Technologies, Altınbaş University,
Supervisor: Asst. Prof. Dr. Timur İNAN
Date: 08/2023
Pages: 64

The fundamental purpose of the field of research known as "biometrics" is to explore the development of reliable approaches for identifying individuals using their observable traits. Examples of biometric identification include both physical and mental traits of an individual. As a kind of physical identification, fingerprints and facial features may be compared and analyzed. Human gait has been researched because it has the potential to be used as a behavioral identifier in computer vision Using a Convolutional Neural Network (CNN) and a Support Vector Machine, the capacity to estimate a person's gender based on their stride was explored and investigated (SVM). Throughout our examination of CNN's potential uses for gait-based gender identification, we will strive for both a high degree of accuracy and a cheap computational cost. We analyzed and experimented with a variety of CNN architectures and hyperparameters in this setting.

Keywords: CNN, AI, Age detection, ML.

TABLE OF CONTENTS

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

CNN	:	Convolutional Neural Network
MLP	:	Multi-layer Perceptron
SVM	:	Support Vector Machine
LBP	:	Local Binary Pattern
LDA	:	Linear Discriminant Analysis
BIF	:	Biologically Inspired Features
CCA	:	Canonical Correlation Analysis
PLS	:	Partial Least Square
RBM	:	Restricted Boltzmann Machine
RNN	:	Recurrent Neural Network
STN	:	Spatial Transformer Networks
NLP	:	Natural Language Processing
VQA	:	Visual Question Answering

1. INTRODUCTION

1.1 BACKGROUND

Accurate descriptions of age, gender, facial characteristics, expressions, clothing, and personality, to name a few, are necessary for the functioning of a variety of applications, such as user identification, social interaction, face tracking, and behavior detection, to name a few. Face analysis has become one of the most critical difficulties in the area of computer vision as a result. Despite the fact that a substantial amount of study has been conducted on the complexities of age and gender categorization, the findings have not been very outstanding. Convolutional neural networks, often known as CNN, have recently become the preferred method for determining an individual's age or gender. CNNs have been shown to perform very well in a range of computer vision applications, including the verification and automatic detection of faces, the identification of human activities, the recognition of handwritten digits, and face verification. Recent applications of convolutional neural networks, often referred to as CNNs, in the area of soft biometrics analysis have focused on challenges such as apparent age estimate, gender and smile categorization, and actual age and gender prediction. The poor performance of CNNs in tasks like as age identification illustrates that there is still potential for improvement, despite the vast availability of face images gathered from the internet and other sources.

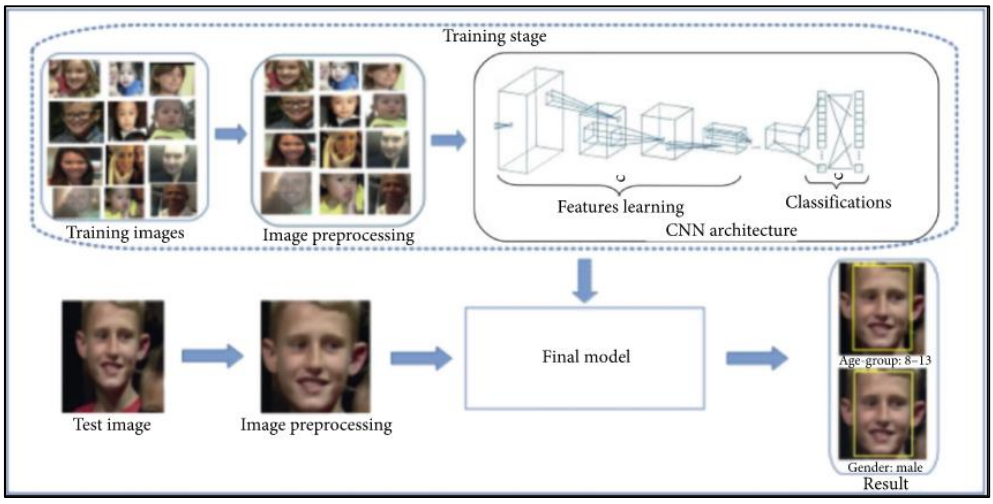


Figure 1.1: CNN Age and Gender Detection[1].

This is partially due to the fact that CNNs have not yet been taught to distinguish facial emotions. This research's most significant contribution is an unique feedforward attention

method that, when applied to existing CNNs, renders them more stable in contexts with a great deal of unpredictability and minimal limitations. In light of recent advances in attention processes in biology, we present a feedforward attention method for discovering the most discriminative patches in low-resolution, unconstrained face pictures before doing high-resolution image processing. Using our method, the network is able to prioritize the least distorted or obscured portions of a picture. This increases the model's resistance to noise and other potential distractions. Our attention pipeline outperforms previous generations of CNNs that have been pretrained for face recognition since it is rigorously tested against industry-standard age and gender recognition standards.

1.2 PROBLEM STATEMENT

The fundamental purpose of the field of research known as "biometrics" is to explore the development of reliable approaches for identifying individuals using their observable traits. Examples of biometric identification include both physical and mental traits of an individual. As a kind of physical identification, fingerprints and facial features may be compared and analyzed. Human gait has been researched because it has the potential to be used as a behavioral identifier in computer vision. Human gait may be recorded without the subject's cooperation; nevertheless, it is difficult to ascertain the identity of someone who is unwilling to provide their physiological identification. It is not feasible to perform facial recognition on a person if, for instance, their face is obscured, they are looking away from the capturing equipment, or they are too far away for the algorithm to identify differentiating characteristics. However, the key information may be gleaned from the movement of the body and body parts, which are more visible than the face from a distance, allowing gait to be recorded remotely. This is because the face is less prominent than the rest of the body.

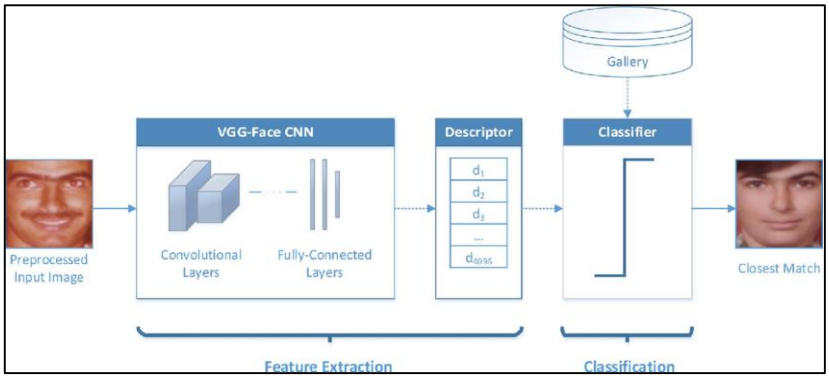


Figure 1.2: Wrong Facial Classification Due to Complex Facial Features[3].

The human gait may be captured from several angles, and the data can be utilized without the subject's permission or even the need that they look directly into the camera. The use of human gait as an identifier has a variety of advantages, one of which is the potential of addressing the deficiencies inherent in systems that rely on physiological identifiers. Analysis of a person's stride may be used to infer a range of psychological characteristics. Age and gender are two examples of such variables. It has been observed that persons of various ages and sexes walk significantly differently. Participants were given point light displays to wear as part of the experiment, and observers were charged with distinguishing the genders of the people based only on the look of the displays. Despite without knowing what the person was wearing or how their hair was done, human observers were able to correctly identify the subject's gender 63 percent of the time. The results provide support to the theory that a person's gender may be deduced from their stride alone. Numerous studies have established that the aspects of gait, especially stride width, change with age. The most significant of these alterations is seen in elderly individuals. Similar evidence demonstrated that the arm swing, joint rotations, and peak height of the foot during walking varied between older and younger individuals. These results give proof that individuals of different ages walk differently, paving the path for the creation of a system for indirectly determining chronological age based on gait observation. Using gait analysis to predict age and gender might be advantageous for a number of systems. These days, closed-circuit TVs are used in a variety of situations, including police investigations. Existing technologies may be able to aid with these investigations by employing faces as biometric identifiers; however, the resolution of these technologies may be inadequate or the subject's face may not be visible on the video. In such situations, gender identification based on gait might

assist law enforcement officers by establishing the subject's gender. Additionally, the ability to get a general estimate of a suspect's age from a distance would facilitate identification and speed up the procedure. Another conceivable use is in the marketing industry. On a street, it is possible to deploy a gait-based gender identification system that does not need a full facial scan to determine a person's gender. This kind of device may establish the gender of a person based on their gait. Using this data, electronic billboards may modify their ads to be more appropriate to the gender of passing pedestrians. It would be advantageous if the system could accurately determine the subject's age. This is due to the fact that people of different ages are likely to be interested in different topics. Similar cases demonstrate the possibility of utilizing gait analysis to establish the gender and age of an individual.

1.3 MOTIVATIONS AND OBJECTIVES

facial analysis technologies have the potential to profoundly transform our daily lives. It evolves into a very effective personal assistant when augmented with artificial intelligence. It can estimate how we will feel when we wake up and then use that information in conjunction with our habits, age, and other criteria to offer us with a personalized breakfast, dinner, and weather forecast. In a similar fashion, a clothing manufacturer may consider your age and gender when determining which goods to suggest to you, and they may subsequently send you a photo of yourself wearing one of the recommended items on your smartphone. Future machines will be able to control the mundane parts of life from beginning to conclusion. As computer vision and face recognition technology continues to advance, it is inevitable that we will all be completely interconnected and comprehended. On a range of face analysis tasks, clever algorithms have already surpassed human identification. Deep learning and convolutional neural networks are two fields that are now at the forefront of machine learning research. In scientific study, intriguing and innovative advances are made every day. It is my desire to make a constructive contribution that motivates me to accomplish this. No one, to the best of our knowledge, has yet created a Deep Multi-Task Learning Convolutional Neural Network for Emotion, Age, and Gender Recognition; this work would undoubtedly be a significant addition to the field of Machine Learning.

1.4 CONTRIBUTION

The ability to determine a person's age and gender based just on their stride is one of this theory's accomplishments. Despite the fact that both components offer high-performance CNN architectures, their key contribution is study into the feasibility of CNN for the understudied gait-based gender identification and age estimation problem utilizing GEI as input. This study represents the fundamental contribution made by both parties. Evaluations of performance can involve the time and effort spent calculating outcomes. Even though CNN has garnered a great deal of attention in recent years, very few studies have used gait to detect gender. It was shown how an input for a CNN that could perform many jobs may be optical flow channels. The primary purpose of this study is to re-identify a person based on their gait, whereas gender recognition is a secondary objective.

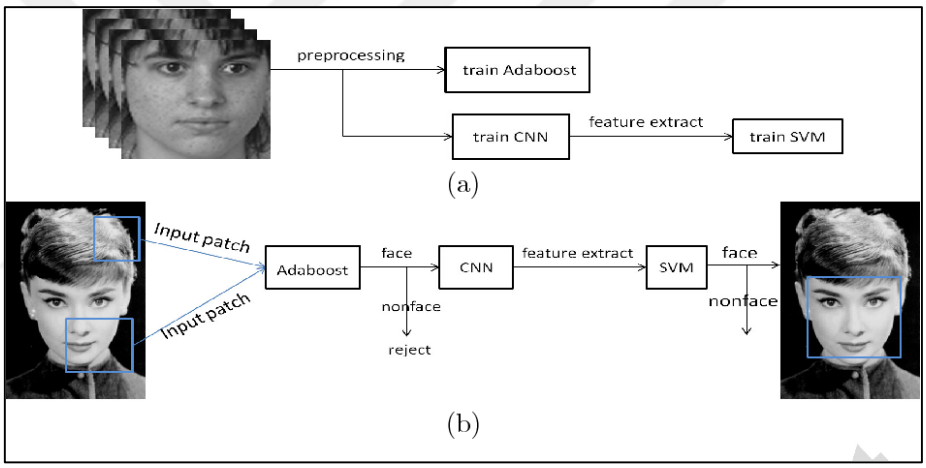


Figure 1.3: Proposed Modification to the CNN Algorithm Workflow[7].

Using a Convolutional Neural Network (CNN) and a Support Vector Machine, the capacity to estimate a person's gender based on their stride was explored and investigated (SVM). Throughout our examination of CNN's potential uses for gait-based gender identification, we will strive for both a high degree of accuracy and a cheap computational cost. We analyzed and experimented with a variety of CNN architectures and hyperparameters in this setting. In addition to delivering findings comparable to approaches deemed to be state-of-the-art, the suggested design also provides valuable insights into the issue at hand. To far, however, no research using GEI as an input for gait-based age prediction has been conducted. Age estimation is another activity that requires GEI; however, optical flow

channels are utilized in place of conventional input. Consequently, the evaluation of CNN use in this setting is the most significant contribution of this section. After evaluating various distinct architectural solutions, it was discovered that gait-based age estimation with CNN was a viable option. This strategy was proven to be feasible despite the fact that its performance is lower to other ways mentioned in the literature.

1.5 THESIS STRUCTURE

To be more exact, the thesis structure is as follows: In the second portion, we will examine some of the older work and how it was implemented. In the third part, we explore each component's greater context. In Section 4, we offer a comprehensive explanation of our idea and its applications. In the fifth part, we run many simulations on our model and record the results. In Section 6, we present a quick summary of our findings and discuss how they may be beneficial in the future.

2. LITERATURE REVIEW

2.1 INTRODUCTION

In the disciplines of image processing and machine learning, a substantial amount of research has been conducted on human gender estimation. In this part, we provided a summary of the gender identification research that has been undertaken to far. Lian HC [20] used a local binary pattern (LBP) and a support vector machine (SVM) with a polynomial kernel to the CAS-PEAL face database. He was able to achieve a 94.81 percent accuracy rate. Using this method, it is feasible to get a high degree of precision if the block size for the LBP operator is chosen appropriately. Li et al. [19] used just five distinguishing face characteristics to determine a person's gender in their investigation. [Bibliography required] (eyes, nose, mouth, brows, forehead). Due to the complexity of the utilized backgrounds, the approach of feature extraction used in this study is regrettably susceptible to interference. Saeed Mozaffari, Hamid Behravan, and Rohollah Akbari [23] classified males and females using geometric characteristics. Both the AR and Ethnic datasets have 126 frontal images. In order to fulfill their objective, they have attained an accuracy rate of 80.3% and 86.6%, respectively. Several classification algorithms, including naive Bayes, artificial neural network, and linear support vector machine (SVM), are employed to extract features from a texture-based local binary pattern in the paper [10]. Using just 100 facial images from the Nottingham Scan collection, researchers were only able to reach a 63% accuracy rate. In their study on gender categorization from facial photographs, Sajja, T. K., and Kalluri, H. K. [28] examined the usage of LBP, SVM, and Back Propagation. In this experiment, the ORL dataset had a total of 400 photos, while the Nottingham Scan database contained a total of 100 images. In the end, they were successful in boosting the accuracy of the ORL dataset by one hundred percent and the Nottingham Scan database by seventy one percent. The study presented in [24] yielded an impressive categorization accuracy of 99.30 percent. This was achieved with the help of the SUMS face database. The authors of this study used the K-means closest neighbor (KNN) technique as a classifier with the 2D-DCT feature extraction and the Viola and Jones face recognition. The 2D-DCT approach is incompatible with real-time applications since it requires a substantial amount of computational resources. Principal component analysis, often known as PCA, was

employed to minimize the overall dimensionality of the facial picture by the researchers in [30]. Then, the evolutionary algorithm was used to determine the optimal collection of eigenfeatures (GA). The total error rate revealed in this instance was 11.30 percent. The fact that the GA has a high level of computational complexity is the fundamental disadvantage of using this method. Althnian et al. [4] employed hand-crafted and merged characteristics for gender identification of faces. In this method, both SVM and CNN were utilized; nevertheless, CNN produced the best accuracy (86.60 percent), which may be enhanced further. As part of their investigation into the impact of prejudice on deep learning, Serna et al. [29] examined the application of VGG and ResNet to the problem of gender identification. They ran an experiment with a group of individuals who had no prior preconceptions about any of the three ethnic groups, using images that had been sorted into the proper categories beforehand. In this case, the unbiased group discovered that using VGG led to the greatest average accuracy (95.27 percent), while the biased group discovered that using ResNet led to the lowest average accuracy (95.67 percent). Some studies have explored the idea of establishing a person's gender based on characteristics other than facial appearance. These features include of a person's physique type, eyebrow form, hand geometry, fingernail shape, etc. Dong, Yujie, and Damon Woodard [12] developed a revolutionary method based on the assessment of the eyebrow's form. For classification purposes, this research used MD, LDA, and SVM, achieving 96 percent accuracy for the MBGC dataset and 97 percent accuracy for the FRGC dataset, respectively. The authors of [5] were able to reliably estimate a human subject's gender based on the shape of their hand using just 40 photos from a small dataset. The classification method applied was a mix of score-level fusion and linear discriminant analysis (LDA). Hong'aLim et al. [32] discovered an unique approach for detecting a person's sexual orientation by examining fingernail clippings from 80 individuals. In this experiment, they used principal component analysis and support vector machine classification approaches, achieving an accuracy of about 90%. Taking into account the whole of the data, it is evident that changes in the categorization of gender are necessary. The fact that feature extraction and gender classification were handled separately was one of the most major shortcomings of previous studies that sought to classify people by gender. In this scenario, previous information is required to create a pre-processing and feature extraction technique that

adheres to best practice standards. The CNN model [21,22], which is a multilayer neural network, has the ability to enhance filters via automatic learning without the requirement for prior information, proving that it is possible to attain greater levels of performance.

2.2 RELATED WORK

This section gives an overview of the history and development of neural networks for gender and age identification, as well as other relevant material for understanding our methods.

a. Age Recognition

Not only did a earliest studies from the 1990s use facial geometry analysis [21] to estimate an individual's age, but more recent methods, such as the pipeline presented in [22], combine Biologically Inspired Features (BIF) with Canonical Correlation Analysis (CCA) and Partial Least Square (PLS) based methods. Thus, face geometry analysis [21] was not the only approach utilized to determine an individual's age. Indeed, BIF were utilized to represent facial pictures in [23], which paved the door for further studies such as [4, 24], which demonstrated that the automated technique had achieved human-like performance. Prior to the invention of convolutional neural networks (CNNs), the vast majority of machine learning approaches had a two-step process. This required first extracting features, such as Local Binary Patterns (LBP) [17], followed by categorizing the data using a Support Vector Machine (SVM) or a Multi-layer Perceptron (MLP) [25, 26]. CNN-based techniques, on the other hand, usually perform the aforementioned two-step pipeline in a single phase, with the network learning both feature extraction and age classification or regression [15, 27], [28], and [29]. This is in contrast to conventional pipeline construction methods, which normally include two distinct stages. CNN models with a greater degree of complexity have also been used to identify age and gender [30], despite the fact that these techniques often need domain-specific pre-training [31, 32]. In addition, a cascade of several deep models was studied, similar to [33]'s methodology. Face verification, face attribute estimation, and gender identification are a few of the additional tasks undertaken by CNNs for facial image processing. For instance, the approach proposed in [34] obtains a remarkable 99.2% accuracy in face verification on the difficult Labeled Faces in the Wild

dataset [35]. As will be illustrated in the next section, equivalent findings have not yet been reached in other face analysis tasks, such as gender identification.

b. Gender Recognition

In contrast to age analysis, gender recognition research conducted in the early 1990s utilized neural networks, such as the groundbreaking approach presented in [36]; authors proposed two neural network structures, an autoencoder and a classifier whose input was the encoded output layer of the autoencoder. In other words, the autoencoder's encoded output layer acted as the classifier's input. The disadvantage of this method was that it needed human adjustments to the image after it had been captured in a studio setting. The age estimation method serves as the foundation for the creation of feature extractor and stacked classifier pipelines. There are examples of such pipelines in [37], [38], and [39]. Alternatively, gender was also explored using the same CNN-based approaches as age [28], [15], proving that CNNs may adapt to new tasks with nothing more than a change in training data. [CNNs] [CNNs] In [40], for instance, a convolutional neural network is trained to identify gender by making tiny modifications to a previously trained network. Following this, a support vector machine is trained using the CNN's calculated deep information.

c. Neural Networks with Attention

Attention is a potent approach that allows neural networks to study select areas of the input image in more detail in order to reduce the task's complexity and remove extraneous information [16]. This approach was partially inspired by the eye fixations that the human visual system does. A major proportion of the strategies currently used to apply bioinspired visual attention processes [41, 42] are centered on detecting visually prominent regions of the picture that will be analyzed in a subsequent stage. Larochelle and Hinton [43] suggested utilizing a third-order Restricted Boltzmann Machine (RBM) inside neural networks to recognize images based on a series of fixations. Denil et al. [44] use a recurrent neural network (RNN) fed with foveated images selected through a control route to perform image tracking. In contrast, Ranzato [45] suggested a model with fewer moving elements that predicts the position of a glance in a downsampled picture and then uses this prediction to extract a high-resolution patch. This model is less complex than its predecessor. Although Spatial Transformer Networks (STN) [46] are also capable of giving attention,

they vary from our method in that they only analyze a single, continuous visual area. Studies such as Adience and IoG demonstrate how well attention is adapted for pictures in the natural environment, which may include a range of occlusions and distractors. Due to the fact that RNNs rapidly incorporate information acquired from several time-steps, they have recently become indispensable to attention processes [16]. Natural Language Processing (NLP) applications like Neural Machine Translation, text-based question answering, picture captioning, and Visual Question Answering have shown that RNN-based attention mechanisms perform extraordinarily well due to their capacity to "see into the past" (VQA).

3. MATERIALS AND METHODS

In order to create a system and choose applicable models and approaches, it is vital for us to learn what has already been accomplished that may be applicable to our situation. In this case, Deep Learning is used by all of the most recent publications we reviewed, and Convolutional Neural Networks are employed in each of those publications. To begin, we must discover which strategy is currently being utilized with the greatest degree of success. Our field connects with deep learning in a variety of areas, such as item identification, object recognition, and speech recognition. In each of these domains, deep learning has proven significant gains over older techniques. All contemporary approaches to problem-solving center on deep learning because this is the reason for the convergence of multiple approaches.

3.1 DEEP LEARNING

Object recognition, classification issues, and prediction challenges are typical applications for machine learning, which can be further divided into supervised and unsupervised learning. Deep learning is a set of techniques used in Machine Learning, more specifically in neural networks, which operate with a set of layers, where the layers allow the data which is the input to the system to be decomposed and analyzed; these layers are not designed specifically for each problem, but are part of a generic procedure that is adaptable to multiple problem types. Deep learning is an area of study within Machine Learning. The most popular application of deep learning is neural networks. This suggests that the same network topology can be utilized for a variety of situations. What distinguishes this model from others is its behavior, which is a result of how the network is trained and the data used.

3.1.1 Convolutional Neural Network

Since the 2012 ImageNet competition, neural networks have become the standard method for all recognition and detection tasks, with results comparable to those of humans. They were able to classify photos into a total of 1000 distinct categories with a test error rate of 15.3%, 11 percentage points less than the runner-up. up's Convolutional Neural Networks,

also referred to as CNNs, are a form of neural network that has greater generalization than its predecessors (e.g., multi-layer perceptrons). CNN is capable of identifying the significant features of the input and assigning calculated weights, also known as filters, in order to prioritize them. This is achieved by slicing the input into many layers.

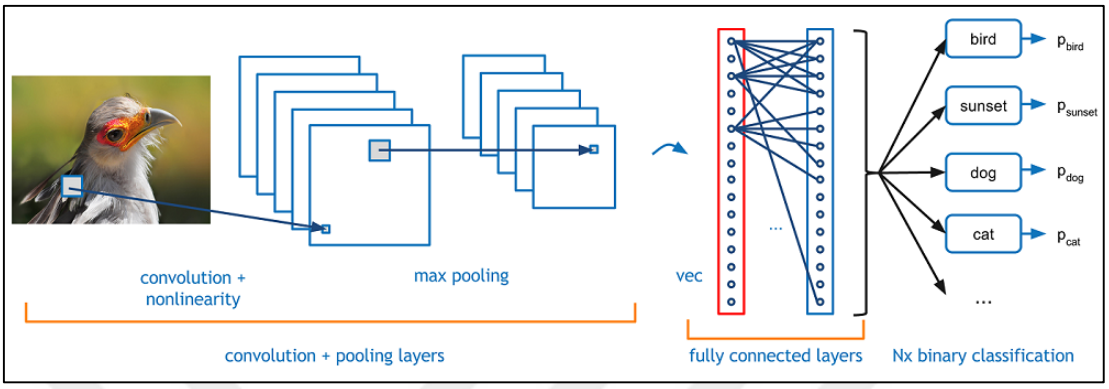


Figure 3.1: This is a CNN, for Illustration Purposes Only[13].

Figure 3.1 displays a typical CNN, which is composed of multiple layers, each of which receives input from the layer preceding it in the chain. Each successive layer was added for a particular reason. The Convolutional Layer is responsible for deciding which picture features are to be extracted (resulting in a feature map). For example, it can be used to identify edges in images by passing the results of calculations performed on the initial vector through a succession of filters before transferring them to the layer below. This enables the discovery of picture edge detection. The Pooling layer can reduce the dimensionality of a feature map while maintaining just the most significant characteristics; the type of pooling employed dictates how this is achieved. One could compare this to compressing a photograph, in which case certain details are lost but the idea remains the same. This reduces the amount of time necessary for computation while simultaneously enhancing the quality of the findings. After each convolution, the ReLU layer is applied to the data to remove any negative values from the feature map. This achieves the objective of "zeroing out" negative values. The primary objective of this operation is to imitate real-world data, which is frequently nonlinear, by introducing nonlinearity into the network. This is done to accomplish the operation's principal purpose (in contrast to the linear convolutional layer). The final layer is known as the Fully Connected Layer, and it is responsible for classifying the input image. This classification is determined by the extracted high-level characteristics from prior levels.

3.1.2 Siamese Networks

When a network comprises two or more identical subnetworks with the same parameters and weights, we refer to it as "Siamese." Despite the fact that the inputs to these networks are compared using a metric derived by a function at the network's top, the networks themselves are capable of processing a broad variety of data formats. This measure provides an estimated estimation of the degree of similarity between pictures 1 and 2. It is calculated using the most abstract form of the image, and the results are displayed.

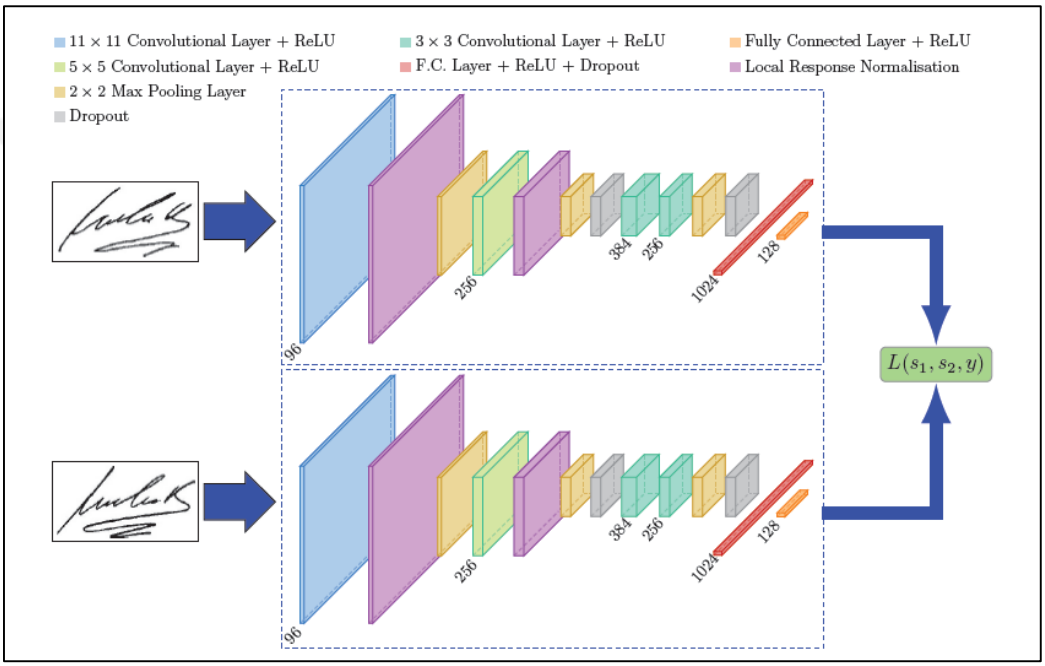


Figure 3.2: Case Study: A Siamese Network[14].

The network architecture is represented in Figure 3.2, which illustrates how the two halves share the same layers to generate input feature vectors and output similarity scores using the same function for the final layer. Regarding their ambitions to build a signature verification system capable of distinguishing between authentic and counterfeit signatures. Since then, neural networks have gained significant appeal, especially for their use in one-shot learning and image recognition (as for example, in image tracking).

a. One-Shot Learning

Training models for object categorization requires a considerable amount of time and a large dataset. However, in contrast to more conventional learning approaches, it is possible to extract essential information about a category from a small number of photos. Regardless

of how unlike the new categories may be to the categories used for training the network, it is possible to classify new categories using the model's previously acquired information from the categories used for training the network.

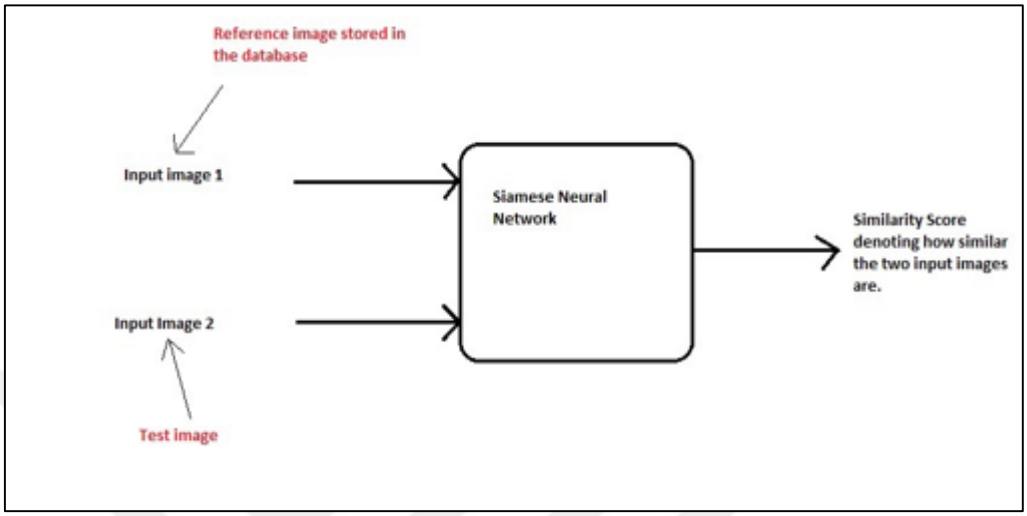


Figure 3.3: Explicit Demonstration of One-Time Mastery Learning[15].

The term "one-shot learning" refers to a learning methodology that, when applied to neural networks, denotes a method of learning in which the network makes a prediction about a new class based on a single instance of that class. (Illustration:) By applying the knowledge the model obtained during the training phase, during which it was fed multiple examples from previous classes, it is possible to accurately anticipate the output of a new class for which only a single sample is available. This is achievable because to the existence of a network that has been trained in a specific range of classes. The input is compared to each of the validation classes using a similarity metric to determine which of those classes it most closely resembles. This conduct is depicted in Figure 3.3. Siamese Networks were used to determine the degree of similarity between inputs, and a validation dataset comprised of images from a range of categories was employed to determine the category to which the newly added input belonged. They used photos of handwritten letters from the Omniglot dataset, which contains 50 different alphabets, to do this. As a result, their classifiers performed better on the dataset than those employed by prior studies. In addition, they argued that the one-time learning technique should be implemented in other areas, such as the classification of photos.

3.1.3 Wide Residual Networks

There was, however, a limit to how much performance improvement could be attributed to deeper neural networks. After a predetermined number of layers, gradients were eliminated, resulting in accuracy difficulties. Additional residual connections were added to the network in an effort to fix this issue. Connecting the output of one layer to the output of the next layers and then summing the values of those two layers is an example of this type of technique. As a result, the authors were able to expand the network further, resulting in an increase in the accuracy of their models. The significant usage of layers in these networks has, however, resulted in the rise of two additional challenges: first, deeper networks usually require longer training periods in order to obtain greater accuracy rates, and second, deeper networks present the difficulty of diminishing feature reuse. Wide Residual Networks are an advanced architecture that places greater focus on the network's breadth rather than its depth. Due to this, the network can be trained far faster, and the results it generates are superior to those generated by standard Residual Networks. It has been demonstrated that networks with as little as sixteen layers can outperform systems with hundreds or thousands of layers, providing results that are regarded as cutting-edge.

3.2 FACE DETECTION

3.2.1 Intersection Over Union

Frequently, the intersection over union metric is utilized while attempting to evaluate the accuracy of predictions created with the aid of bounding boxes. Bounding boxes are rectangular boxes used to indicate a certain region of an image item. For the purposes of this evaluation, both the annotated and expected bounding boxes from the detector are necessary. Before we can continue with the calculation, we need these two pieces of data. The degree of confidence we have in our prognosis is approximately reflected by a number between 0 and 1 that has been developed as a measure. It is possible to rule out predictions if the projected bounding box overlaps only a small amount of the annotated object region. In this circumstance, the intersection over union value will be low, allowing predictions to be ruled out.

3.2.2 Analyzed Papers

Due to the visual variations present in each image, a system that must recognise faces in photographs will have difficulties doing so. These modifications include stance and illumination, and they require models that can distinguish a face from various backdrop objects. Utilize a huge number of Convolutional Neural Networks, each with its own independent detection and subsequent calibration stages, to confront and overcome these hurdles. The calibration stage resizes and relocates the detection window using a technique known as non-maximal suppression to get it closer to a candidate face in the immediate region. This moves the window into place. Non-maximal suppression is used to identify the detection window with the highest confidence score and subsequently to eliminate all other detection windows. This is done so that several, potentially contradictory detections of the same object are disregarded. Their approach consists of numerous convolutional layers, a pooling layer, a normalization layer, and a fully linked layer, among other layers. With the aid of a normalization layer, it is possible to equalize the input parameters of each layer, which can minimize the training time and make the process more effective overall. This is crucial because if the parameters of the layer above it are altered, the input distribution of the layer below will shift, resulting in a slower training process. In doing so, they outperformed competing models on two datasets, namely Face Detection Dataset and Benchmark (FDDB) and Annotated Faces in the Wild (AFW). They were capable of recalling 85% of the material. The first achievement of applying CCN for face detection was enhanced by a subsequent endeavor later that year. In contrast to prior research, their method utilized a variety of convolutional neural networks (CNN), one for each of the primary facial features (i.e., hair, eye, nose, mouth, and beard), to generate a set of candidate windows. These are the key facial features: hair, eye, nose, mouth, and beard. The initial convolutional layer of this level contains seven convolutional layers and two pooling layers. The following phase involves ranking the potential windows and eliminating some of them from consideration (assumed to be false positives). In the final step, the candidate windows are fed into a third convolutional neural network (CNN) that is responsible for face categorization and cropping to enhance the findings. In their experiments with many datasets, the researchers discovered that CNN could achieve detection rates above 90 percent. For instance, the researchers discovered that CNN had

recall rates of 91% for the FDDB dataset, 92% for the PASCAL dataset, and 97% for the AFW dataset. In a 2017 investigation, the same techniques, including non-maximal suppression and ReLU layers, were utilized. With an average precision rate of 87 percent, they were able to achieve a suitable level of accuracy for the completed work. provided a novel technique consisting of three cascaded CNNs, each of which serves a different function and feeds its candidates to the next CNN in the cascade. Each of the three utilizes non-maximal suppression alongside bounding box regression. Utilizing a technique known as "bounding box regression," one is able to calculate the actual distance between the candidate window and the annotated bounding box. A first convolutional neural network (CNN) is used to generate a pool of candidate windows, which are then forwarded to a second CNN for further refinement and development. A third CNN is responsible for creating the face's final bounding box and the positions of recognized facial features such as the eyes, nose, and mouth. After being trained on the FDDB and WIDER FACE datasets, respectively, their network was able to reach a precision of 95.4%. Moreover, in contrast to the other works evaluated, the source code for this model is easily and freely accessible online. A model employing a Region-based Convolutional Neural Network (RCNN) has been developed in recent years. This model was conceived by drawing influence from Caffe's pre-trained model. Due to the fact that traditional RCNN are sometimes unable to detect small objects in an image, but for face detection, small faces are common, particularly when the images used are retrieved from an unconstrained e-repository, the researchers used multiple data augmentation techniques to increase the number of test images fed into the model, thereby improving the training of the model. This was completed so that the number of test photos may be raised without compromising the quality of the training. The recall rate of 95% that they attained on the Face Detection Dataset and Benchmark is the highest of all publications previously examined.

3.2.3 Gender and Age Estimation

Prior research determined that CNN was the most effective system for detecting faces and classifying them based on their ages and genders. This is due to the success that deep neural networks have had in determining an individual's age and gender from a single photograph. This was supported by supporting evidence. The accuracy rate for correctly predicting sex

was 92.6%, whereas the accuracy rate for correctly estimating age range was 62.8%. A CNN with ReLU and Pooling layers was constructed using a VGG-16 pre-trained model as one of its components. Diverse works have adopted comparable approaches employing pre-trained models, such as the Caffe model and the Face Recognition pre-trained models, with equivalent gender classification results. These models consist of: In addition, we enhanced the size of the testing dataset by adopting data augmentation techniques (oversampling and center-cropping), which aided in the improvement of model training. Dropout learning, a technique in which the output of some layers is intentionally disregarded during testing, was also used to lessen the likelihood of the model being over-fit. They had an accuracy rate of 86.8 percent when asked to determine a person's gender and 50.7% when asked to estimate a person's age. The justification for using a pre-trained model is the assumption that a network trained on a face recognition task has already learnt information about a face that can be used for other face-related tasks. In contrast to other models explored, they also advocated that lower layer parameters be shared across all tasks, so that lower layers can learn representations that are common to all tasks, while later layers are more task-specific. This was performed so that lower layers might learn common task representations. This leads in a network that is capable of doing all of these tasks, since they all require familiarity with specific facial characteristics. The accuracy of gender prediction was determined to be 93%, whilst the Mean Absolute Error (MAE) for ages was determined to be 2%. (an average age difference of two years) When compared to the performance of other models across genders, this one is significantly superior. Using pre-trained Face Recognition models, a second scenario determined the subject's gender with an accuracy rate of over 90%. They developed multiple models, one for each activity, as opposed to the previous method of utilizing a single model for everything. This allowed them to deploy more specialized networks, resulting in more efficient and adaptable models. Moreover, the time required to run each individual model is shorter than the time required to execute the unified model. In addition, their 70.5% accuracy in estimating groups in the real world is the greatest yet reported for work of a comparable sort. It is vital to note that the 1-off group classification for this was accurate to the tune of 96.2 percent (1-off is when a person is classified in one group immediately above or below the correct one). They should also be aware that the encoding of our goal labels may be relevant to the

current topic of discussion. When comparing numerous encodings based on the age of the test participant. In contrast to the gender classification system, which has just two possible results (male and female), there may be multiple ways to address this situation. When determining a person's age, for instance, we may be seeking an exact number (person A is 29 years old) or we may be attempting to place them within a broad age band. Both of these strategies have their benefits and drawbacks. According to the results of their research, a modification to the encoding might result in an age estimation mistake of around 0.95 standard deviations (MAE). When utilizing a pure per-year categorization, a cell value is represented as a binary encoding (0/1), however when using the Label Distribution Age Encoding, the age is handled as a collection of classes representing all conceivable ages. This is in contrast to the customary technique of categorizing strictly by year. It is vital to remember that some authors approach the issue of age as a classification problem, while others approach it as a regression problem. As a result, their measure indices (accuracy and MAE) are not comparable, which is why the results for each type of index are shown individually below.

3.3 SYSTEM DEVELOPMENT

This section will discuss the system's architecture. Python was chosen as the development language for this system; hence, all users must do to utilize it is execute the main Python executable with the necessary settings. These points are elaborated upon in the following section. You can discover an in-depth explanation of the system's underlying files and directories in the User Guide attachments, which are available here.

3.3.1 Requirements

To put the system into operation, we must first determine what has to be done in order to complete it. We require a model that can recognize a face in a photograph, extract it, and then transfer it to another model that can compute the age and gender of a person based just on their look. Based on this analysis and the research of the pertinent literature, the following is a summary of the most important system requirements that should be considered throughout the entire system implementation process:

a. The system should be able to recognise faces in photos with a high rate of accuracy (more than 90 percent), consistent with other state-of-the-art results on such models; the algorithm can predict a person's age bracket. Minimum accuracy of 60 percent is required for it to be competitive with other cutting-edge models.
b. It is vital for the system to be able to categorize individuals based on their gender. Based on the performance of comparable models, the precision of this model should be fairly high.
c. The system should include the option of user-defined age ranges;
d. Users of the system should have the capacity to validate the system's conclusions. Users should be able to manually validate prediction results by accessing them from a folder created for that purpose. If this folder is to be used at all, it should only include results that meet or above the user-defined confidence threshold.
e. If the initial classification achieves a higher degree of confidence than anticipated, the system should allow the user to select an additional module for gender categorization. This module should consist of the Siamese Network with One-Time Learning.

Throughout this entire book, the following age ranges will be applied for verification purposes:

i. Child - Age 1-9.
ii. Teen - Age 10-15.
iii. Young - Age 16-24.
iv. Young Adult - Age 25-40.
v. Adult – Age 41-59.
vi. Senior - Age 60-101.

3.3.2 Design

Given that we require a model that can recognize faces and generate age/gender predictions, we can divide the system into two modules, the first of which is responsible for face recognition and the second for age and gender classification, in order to simplify our implementation. Moreover, by decomposing the global model into smaller modules, we are able to have a model that is both faster and more transportable, similar to the strategy where

their system had multiple models for each task, and which was also the network that yielded more accurate results in our preliminary investigation. Therefore, we must first extract all of the faces from an image before using them to conduct classifications. The major objective of the face detection model is to recognize faces in images so that we may extract the coordinates of the face's critical features. These coordinates are used to crop and extract the faces for use in the upcoming module; the module's output consists of the extracted faces and their specific coordinate location. Once the faces have been cropped, they are delivered to the module responsible for classifying the images. Users are able to alter the configuration file that holds the age groups, and the array of results should include information about the subject's age group and gender.

3.3.3 Face Detection Diagram

Figure 3.4 depicts the model of the face detection module that we developed after taking into account the results of the experiments discussed before and for.

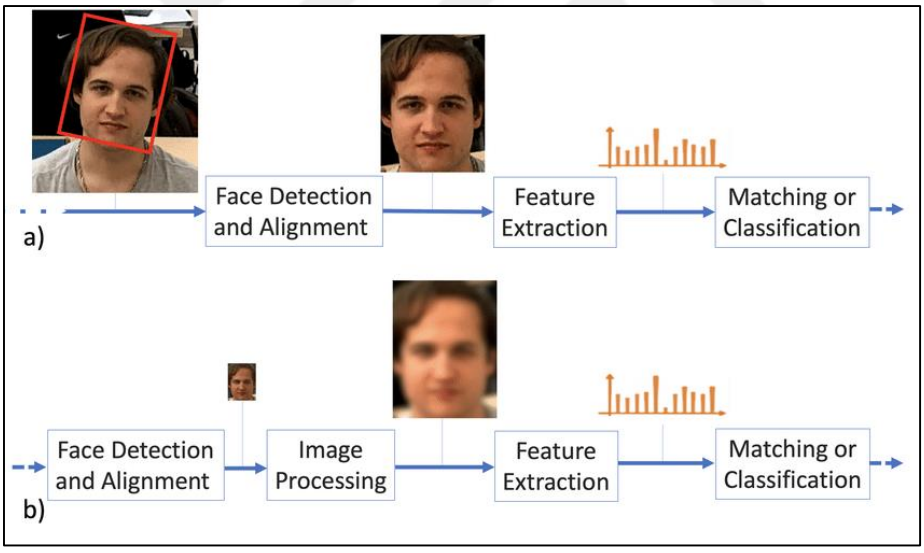


Figure 3.4: A diagram Illustrating the Face Detection Model Depicted[21].

To send the image to the DNN, it must first be preprocessed, then enlarged to a specific image size, and then converted to a specific object type (a large binary object, or BLOB). This image size is customizable by the user, however the default is 200x200 pixels. Based on our experiments, photographs with a resolution of 200x200 exhibited the highest validation accuracy with the tested dataset, with the lowest combination of false positives and false negatives. This image size is customizable; however, the default value is 200x200

pixels. It is also possible to customize the certainty with which the network detects faces; however, the default threshold is 0.5. This signifies those detections must have a probability greater than 0.5 to be considered real faces. Finally, after the image has been preprocessed, it is sent to the network, which outputs all detectable faces with coordinates that indicate four points that are used to crop the face, as well as another output, which is a value between 0 and 1 that indicates the confidence that the detection is indeed a face. Finally, after the image has been preprocessed, it is sent to the network, which produces the coordinates of all identifiable faces along with four points that are used to crop the image.

3.3.4 Global System Diagram

During the preprocessing of the images, the original input is cropped selectively to guarantee that just the face is passed on to the succeeding module. In this section, the Wide Residual Network is applied to categorize individuals based on their age and gender. Users can modify the age groups saved in a configuration file without retraining the network. The end output is a sorted array containing labels assigned to each face present in the input image. In the event that more than one face is discovered, the array's size must be extended so that each slot can record the relevant results for each identified face. The fact that our model is modular (it comprises of a face detection component and an age/gender categorization component) makes the entire system more transportable and adaptable. As the rapid growth of deep learning networks continues, it is feasible that new models will arise that can replace one of the present modules. The availability of a small number of unique models allows for more adaptability with fewer required changes.

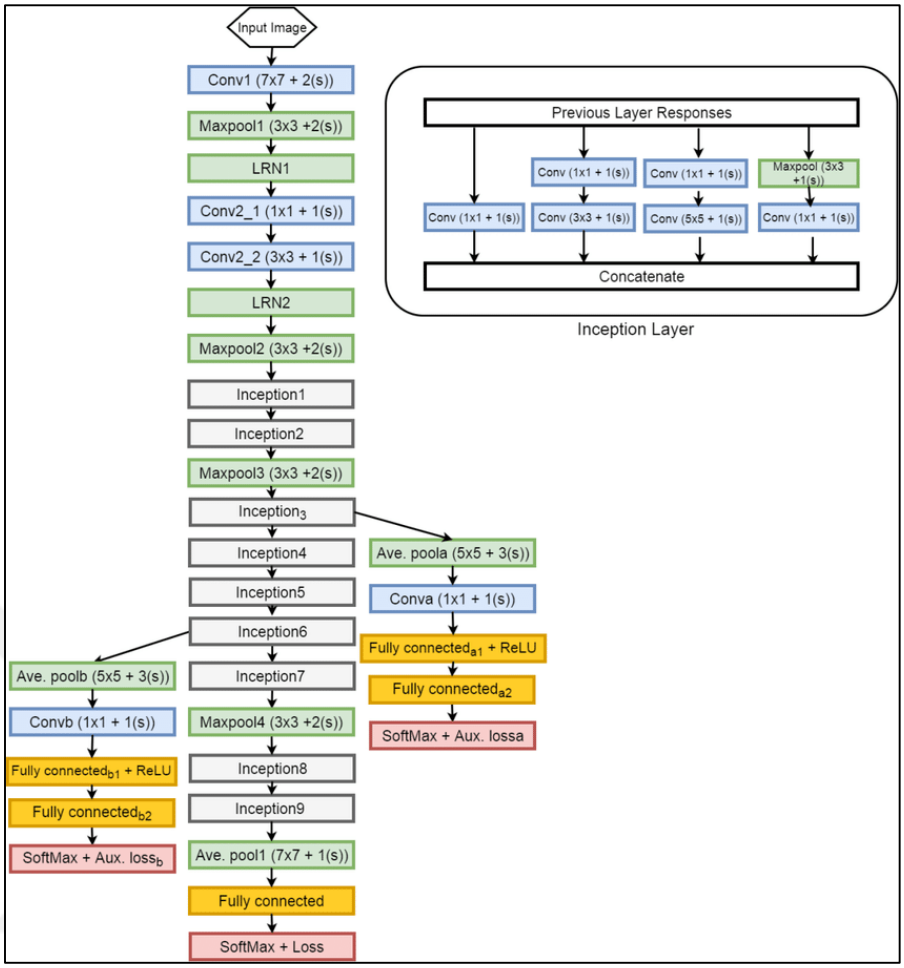


Figure 3.5: Model Workflow Completion[24].

Regarding our model, we must consider the possibility that it may be implemented in an environment where performance is crucial. We found from our earlier investigation that the addition of a Siamese Network to unclear prediction results could potentially boost the overall accuracy of gender prediction. During our tests, however, the addition of this network also increased the execution time of our model by 1 second per image. In a real-time environment where thousands of images are continuously fed into the system, an increase of 1 second per image is unquestionably a bottleneck; however, if a validation is to be performed on a small number of examples as a one-time verification (as opposed to running continuously in a live environment), this could be useful and the execution time is not a concern. As a result, we incorporated the Siamese Networks for gender categorization, which overrides the result from the Wide Residual Network if its confidence in the output falls below a predetermined threshold. This threshold is user-

configurable so that it can be altered at will. In the configuration file, it is also possible to deactivate the entire Siamese Network if only the Wide Residual Network is necessary. Last but not least, there is an additional threshold that can be established, where photographs are placed in a directory for users to manually examine if necessary, so that they can manually validate low confidence findings to ensure that they are accurate.

3.3.5 Face Detection Model

There are a few different models for face detection that are currently available, and one of these models allows us to employ an image in order to extract the locations of each identified face. This section's objective is to validate three various models so that we can select the one that most accurately represents our data. This is a description of the models previously described, which are as follows: The Local Binary Pattern (LBP) cascade is utilized rather commonly in face detection systems due to the fact that it is both simple and quick to deploy. From now on, we will refer to this process as the LBP cascade. Depending on the usability requirements of the final system, speed can be significant, but LBP-based systems have shown effective in other tasks, such as texture categorization and image retrieval. Using the Cascade Classifier module of OpenCV, the Multi-Task Cascaded Convolutional Networks (MTCNN) model is evaluated. This model is built with three tiers of CNN cascades. It is capable of locating the eyes, noses, and mouths of humans and determining where they are positioned on the face. Compared to other models, this one's 2016 performance was regarded as state-of-the-art; hence, it is a strong contender for our review. OpenCV's Deep Neural Network module was published with version 3.3 of OpenCV in 2017 and includes a GoogLeNet model trained for face detection; the authors trained their model using these datasets. We chose OpenCV as the model because it is such a popular framework for a range of image processing tasks. This implementation, which is based on the Caffe framework, consists of over 200 layers of various sorts, including convolutional layers, ReLU layers, normalization layers, and others. The authors trained their model with both VOC2007 and VOC2012.

3.3.6 Age and Gender Prediction Model

Once a face has been identified in the preceding phase, the next objective is to identify the person's age range and gender category. Instead of recreating the wheel, we investigated whether an existing model could be tailored to our circumstances. There are numerous models available that suit our fundamental requirements. Through this type of investigation, it was discovered that numerous models had readily accessible source code. The majority of them were merely examples or recommendations for performing classification problems based on age or gender. We confined our search to models that were based on a paper presented at a conference, hence excluding the previously stated initiatives. This eliminated the need for test projects and gave us more trustworthy models in their place. With the inclusion of these specifications, the number of prospective cars was drastically decreased.

They were able to discern two potential models from our investigation:

a. Modeled the product's functionality. Because it served as the foundation for the IBM (International Business Machines) created forecasting system, this particular model was selected for use. This resulted in its presentation at the XXVII International Joint Conference on Artificial Intelligence.;
b. Developed a functional model of the product. Previous research on Wide Residual Networks provided the impetus for the construction of this paradigm. This study was presented at the British Machine Vision Conference for consideration. This method was enhanced by utilizing the IMDB-WIKI dataset. We were provided access to both the training data and a sample set of validation data.

3.3.7 Validation

We are familiar with a variety of frameworks that have the capacity to provide us with the desired attributes. Therefore, in order to determine which of these answers are the most viable options to our problem, we must compare and contrast each of them. In order to accomplish this, we examine the precision that each framework claims to provide. The analysis and testing techniques are described in depth in the following paragraphs. This was the sole factor addressed in the supplied study, which compared identical datasets for

each of the tests. If the results presented here were derived from a different data set, a real-world implementation may produce results that differ from those reported here.

a. Face Detection Model

Numerous experiments were conducted with the intention of validating which can achieve higher levels of precision, so that we may incorporate it into our system. The findings of these trials can be found in the paragraph that follows. To ensure that the results are comparable and accurate, each experiment used the exact same 8593 samples from the IMDB dataset, which is a collection of photos of individuals of various ages and sexes available online. This dataset includes images that are both aligned and side-posed in a variety of orientations. In the past, photographs were altered such that their resolution was always 64x64, and the confidence level was set to 0.5. The confidence threshold is the level of assurance required by the model to verify a detection as a face before it can be designated as such. After doing an initial cleanup on the dataset by deleting images of low quality (such as those with significant occlusions or blur), random samples were picked for the study.

b. LBP cascade

Although the LBP cascade is faster than the other two methods (taking an average of 0.06 seconds per image versus 0.09 seconds for the MTCNN), it still spends time attempting to identify an image with a face pattern that matches the pattern of the image. This requires time despite the LBP cascade being faster (using the position of the eyes, nose, and mouth). The fact that the bulk of test failures occurred with side-view photos of people's faces indicates that the model was only 77 percent accurate when utilized in challenging conditions. It is quite probable that this algorithm will fail in any test case, including ones with faces in awkward positions. Due to this, we require a model that is more capable of adapting to such scenarios, as, depending on where the cameras will be placed in the supermarkets, similar circumstances may arise in the real-world environment. Therefore, we require a model that can respond more effectively to such scenarios. Figure 3.6 displays a tiny sampling of the findings, the vast majority of which are photos taken from the side. (With a few exceptions that are more aligned).

c. MTCNN

The MTCNN is significantly slower than the LBP cascade, requiring around twice as much time to process the same number of images in the same period of time (it takes 0.09 seconds to process each image). In terms of accuracy, the results of the output analysis were not as promising as we had hoped, and the bulk of photos that did not pass were frontal face images that were detected using the simpler LBP cascade. Typically, it is easier to spot this type of circumstance. Consequently, this resolution is similarly ineffective in this circumstance. However, side pictures were far less likely to fail than LBP cascade. Using the outcomes of this investigation, we were able to achieve a detection rate of 44.7 percent. Due to its lack of precision, this model is not compatible with our framework in any manner, shape, or form. Figure 3.7 depicts multiple instances of false positives, and it is very clear that they are all interconnected.

Figure 3.6: Examples of the Results of a Failed LBP Cascade[33].

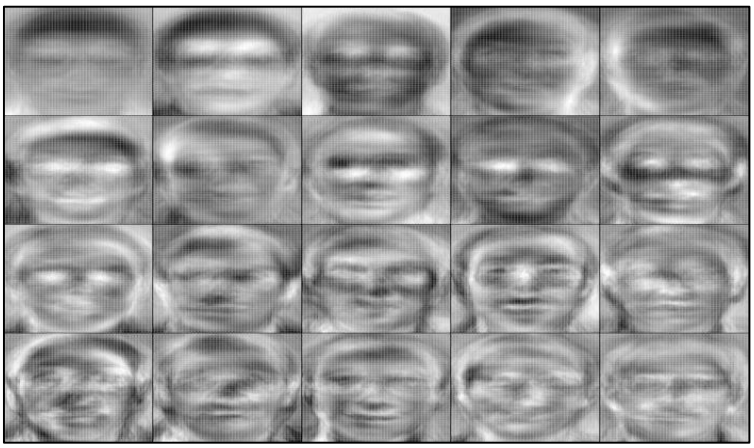


Figure 3.7: Failure of MTCNN to Sample Images[33].

d. OpenCV GoogLeNet

The most recent network evaluated for its capacity to detect faces was OpenCV's GoogLeNet, which processes a single image in about 0.08 seconds, using a technique very similar to MTCNN, etc. OpenCV is renowned for its many modules dedicated to visual recognition applications. This is the first OpenCV model that utilize Deep Learning for face detection. Using the same settings as in the previous two tests, a detection rate of 82.4 percent was achieved, which was higher than the detection rates recorded in the previous two experiments. This network consists of two distinct outputs in total. It provides a set of coordinates for the extracted face that may be used to crop the faces so they can be utilized by the age and gender model. In addition, it produces a confidence value (a number between 0 and 1) denoting the chance of the face being a real face, which demonstrates the network's confidence in its prediction. This value can be used to crop the faces in order for them to be utilized by the age and gender model.

e. Age and Gender Prediction Model

Once the face has been properly detected, the next stage is to classify the individual according to their age and gender. The prototype for this purpose was created and rigorously tested. The classes listed below were utilized for our preliminary testing; more information on each can be found in the requirements section:

i. Child - Age 1-9;
ii. Teen - Age 10-15;
iii. Young - Age 16-24;
iv. Young Adult - Age 25-40;
v. Adult – Age 41-59;
vi. Senior - Age 60-101.

During this round of testing, we reevaluated our previously conducted face detection validations using the IMDB dataset. For the purpose of validating this model, a random sample of 8,500 pictures was utilized. Overall, we were only able to reach a rate of 56 percent accuracy, which was much less than we had hoped. Utilization of the second model considered for application was selected. Using the same set of classes as before, this test improved the accuracy of the prior model by 6.6%. In order to create reliable age predictions, this model employs a softmax classifier, which assigns a probability to each of

101 age categories. There are a total of 101 age groupings. To determine a person's true age, a dot product computation is performed in this phase. To determine a person's precise age, we first multiply the likelihood of each age group by the age value associated with that age group, and then add all of the resulting individual figures. This gives us with a more precise estimate of the person's real age. The first model provides us with a list of membership probabilities for each of these groupings. Then, one can utilize these probabilities to estimate a value. Using age groups rather than individual ages simplifies the process of generating statistics. There is a potential that our participants' actual ages are not always required. Simply put, we are only interested in the proportion of individuals within a specific age range who have done a specific task. People who shop at a supermarket aptly illustrate this description. To make marketing decisions more readily, managers can profit by segmenting their target clients into more manageable age groups. To decode the specific age into the range of classes that we utilize, our validations employ a mechanism that adds a layer to the initial age computation layer. The goal of this layer is to determine the user's age. Due to the fact that the user can change the age groups at any time without having to retrain the underlying network, this additional way allows for a more flexible application.

3.3.8 Softmax Classifier

It is required to conduct additional research on the existing network in order to improve the age classification's precision. Is it possible to provide a system that allows users to manually verify the accuracy of selected forecasts, and if so, how could this be accomplished? How can we identify which predictions do not require human verification? If we know that a softmax classifier is used to calculate an individual's age, we may assess the degree of certainty with which we can forecast that the individual is a member of a given group, in this case, the group of individuals who have a specified age in mind. When determining the accuracy of a prediction given by a softmax-based classifier, it is customary to employ the class with the highest probability as the estimated confidence. Using the category with the highest probability offers the most precise results. It is useful to understand the accuracy of age predictions for a certain group of individuals. If we only consider those in which the confidence reaches high levels, can we guarantee that credible forecasts, such as those in

which the confidence reaches high levels, would match correctly classified data samples the majority of the time? In order to investigate this topic, we will assume that confidence levels larger than 0.6 indicate a high level of assurance. In other words, is it even somewhat possible that the lower confidence levels provided by data samples are erroneous the vast majority of the time? If this is the case, we could add a manual validation step to all of the other situations in which the probability is less than sixty percent, so increasing the correctness of the program-generated statistic. Consider a grocery store that use such models to acquire a deeper understanding of its target market. In this instance, precise data would be useful for gaining a demographic insight of the store's clientele. Implementing this strategy is doable if we are aware that cases classified as having less than sixty percent confidence in the predicted result require an additional confirmation step by the user. This will increase the reliability of the supermarket's daily data, providing the marketing managers with a strong foundation upon which to build their efforts and increasing the efficacy of their efforts. This was the reason for doing another experiment. The probabilities stored inside the network for each of the 101 ages, also known as age groups, are computed at the outset. We were able to determine the chance for each of the five age groups after sorting the probabilities by age. Therefore, the probability of any agent group class between 1 and 9 is found by adding the probabilities for each age group. This yields a range between 1 and 9. We noticed that the softmax classifier has a distribution that is equivalent to the distributions of the two classes that are immediately adjacent to it in some circumstances, notably those that are situated at the boundary of another class. The individual in Row 3 is 27 years old, but the probabilities for subjects aged 16 to 24 and 25 to 40 are fairly similar (i.e., 47 percent against 49 percent, respectively). Due to the presence of a 62-year-old individual in Row 4, the odds for groups 41–59 and 60–101 are equal (i.e., 50 percent and 49.5 percent, respectively). In such situations, it may be difficult to determine which group an individual legitimately belongs to. It is reasonable to predict that, on average, data samples will yield results that are consistent with the first row. In this row, the age group 41–59 is predicted with a confidence value that is much greater than the other age-group classes. Given that the individual's actual age is 59, the predicted age group is accurate in this instance. Before jumping to conclusions, it is essential to review the entire set of validation data in order to assess how well the accuracy performs across all

samples and whether the most likely scenarios are in fact the most accurate. This should be accomplished before to leaping to conclusions. In terms of classifying users according to their gender, our network outperforms all others. Due to the existence of only two classes, human validation is also facilitated (i.e., man and woman). The gender forecast is derived using a mechanism distinct from the age forecast. The model produces a value between 0 and 1, attributing values greater than 0.5 to men and values less than 0.5 to women. As a result, the output of the network is a single probability rather than a collection of probabilities. If the value of 0.5 is used as a standard, the majority of samples with faulty labeling will likely have confidence levels that are close to the value of 0.5. As a way of testing this idea, we determined the fraction of samples that were incorrectly classified as having probabilities between 0.25 and 0.75, below 0.25, or above 0.75, respectively. This should reveal the amount of samples for which the network has doubts regarding their classification. According to the findings, the potential for the network to make inaccurate predictions rises if the prediction probability is quite near to the value 0.5. Although high confidence inputs continue to contain a larger number of erroneous images, this fraction is decreasing relative to the total sample size. The small number of flawed samples makes the human verification process simpler than in the case of the ages. When the likelihood of prediction is close to 0.5, it might be claimed that a manual validation technique provides a suitable amount of human interaction.

3.3.9 Increase the Accuracy of Low Confidence Outputs

When our baseline model's confidence in an age classification is less than 60 percent, or when it is close to the reference value of 0.5 for gender classification, an extra technique to improve the network's accuracy in producing predictions may be useful. In both instances, it is possible that the network might profit from the addition of the mechanism. In a perfect world, the validation process would be conducted by an automated system as opposed to a manual procedure. One-Shot Learning-trained Siamese networks have the potential to be beneficial for the classification of a person's facial characteristics. These networks would compare a new face to an existing collection of facial features in their database, and then make a determination based on the degree of similarity between the two sets of features. When a child's face is used, it is more likely that the finished output will

resemble the face of another child than an adult's face. This is because of the obvious differences between the faces of youngsters and adults. If this technique can improve the network's accuracy, a manual validation mechanism will be superfluous and therefore unnecessary. To test the plausibility of this notion, a Siamese Network–based face comparison module was constructed. This is possible by entering both faces into the same model, which calculates the L2 distance (also known as the Euclidean distance) between them. When there is a higher degree of similarity between the inputs, the output scores will be lower as a direct result. The source of low-confidence face samples was a dataset that had been previously utilized for testing with the Wide Residual Network. Because photographs with a high classification rate are more reliable, we only included those with a high classification rate in our One-Shot training dataset. There were a total of 3002 test cases, with approximately 30 validation photos for each of the analyzed classes. Due to the lower number of times it is executed when trained on a smaller data set, the network may be able to perform computations more quickly when trained on a smaller data set (once for each of the dataset images). location of training data and its input into the Siamese Network. The training set is represented by a folder comprising two subfolders containing photographs of men and women, respectively. You have the option to remove automatically uploaded photographs and manually add more recent examples. Inputting the training data into a Siamese Network and comparing it to the input we're attempting to evaluate yields an output prediction.

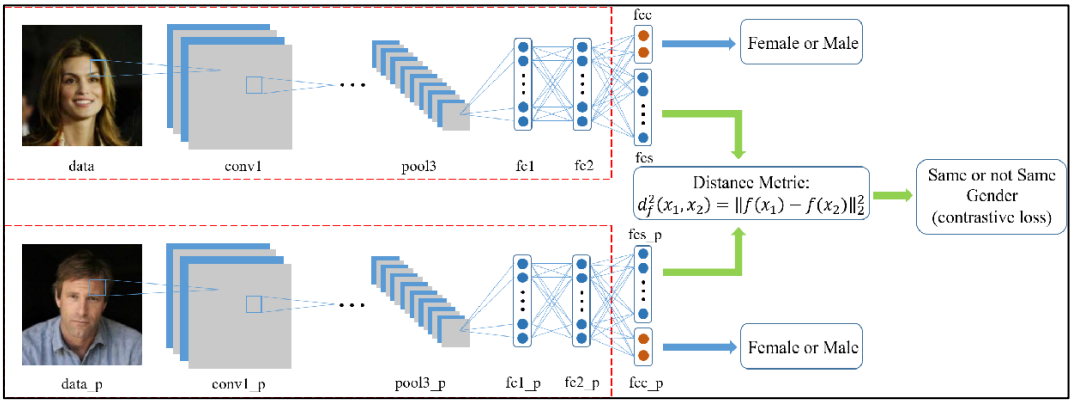


Figure 3.8: Automatic Generation of Training Data[23].

The validation dataset was compared to the faces in the first experiment. The forecast with the highest degree of similarity across all classes was declared accurate, and it was utilized for the remainder of the experiment. The face of a 25-year-old guy would be compared to

all of the other faces, and if the most comparable face was between the ages of 25 and 40, the output forecast would be limited to only that age range. The second photograph on the right is the validation data image. Each image is accompanied by a brief description of its proper classification.

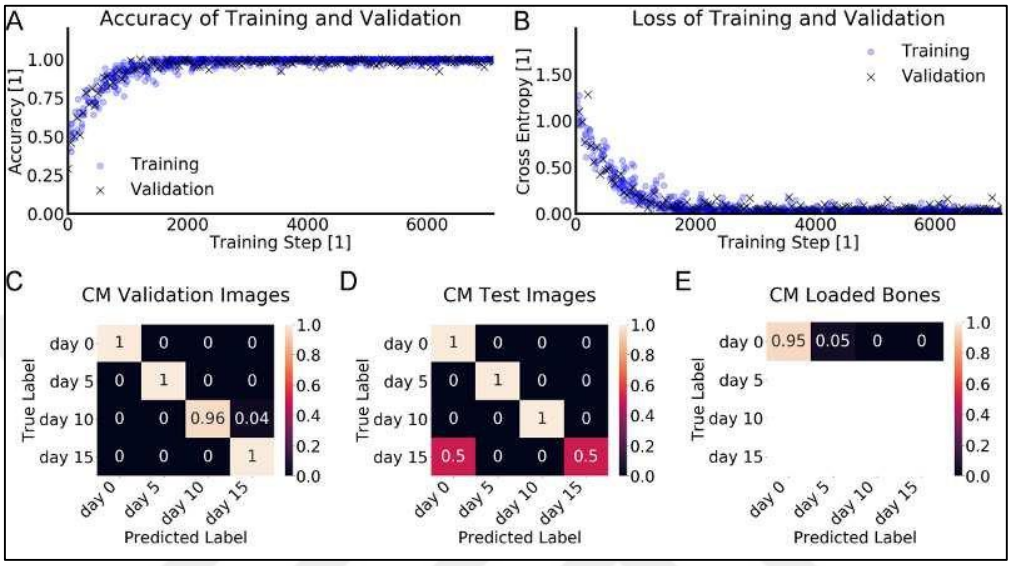


Figure 3.9: Accurately Predicted by Age[14].

The image on the left depicts the input, whereas the image on the right depicts the validation done to the input to evaluate whether or not it was accurate.

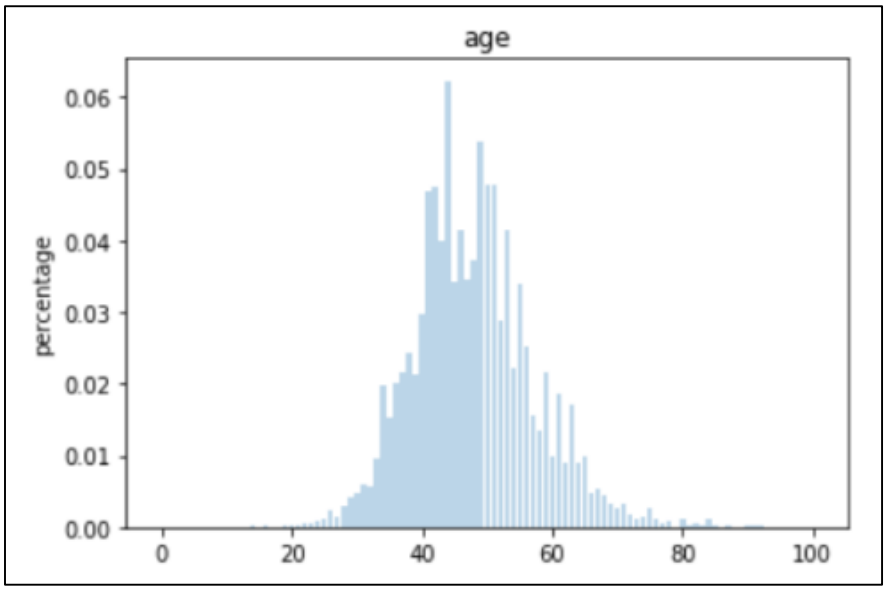


Figure 3.10: The Age Data Failed as a Predictor[15].

The image on the left depicts the input, whereas the image on the right depicts the validation done to the input to evaluate whether or not it was accurate. Was selected to calculate the average degree of similarity between each input and each class as an extra technique of validation. In addition to using the face chosen to be the most similar, this was done to determine which category the input belonged in. Consequently, it was hypothesized that a 30-year-old would have more in common with pupils who fit into this category. If this were to occur, the level of similarity between pupils of this age group and students of other ages would decline. The results are shown in the "Age minimum average" section of the table. Clearly, these results are inferior to those of the prior test, as the percentage of accurate predictions has plummeted to 12%. There are certain individuals that are highly dissimilar, and their status as outliers reduces the average similarity. Even though people in the same age group should share some age-related features in common, there are also significant differences between them. In addition, the average similarity may be exaggerated when one class is compared to another with a small age difference. This is due to the fact that there are individuals in each class that are quite similar to one another. As seen in Figure 3.10, several samples were more comparable to individuals from other classes than to individuals from their own class. Ineffectiveness of switching to the Siamese network when the age classification model lacks adequate confidence in its own forecast. Before abandoning this experiment entirely, it is necessary to discover whether or not the gender-specific network additions are advantageous. In this section, in order to conduct our tests, we will employ scenarios in which the level of confidence in the gender ranges from 0.25 to 0.75 as examples. As with the minimum age criterion, we conducted an analysis in which we sought for the individual who was most similar to the desired population. To put this in context, the gender categorization algorithm achieved 75% accuracy with cases of lesser confidence (63 percent). Even if just by 12% and for a subset of the data, there is evidence that incorporating gender information can improve the precision of data generated by Siamese Networks. This enhancement only applies to a subset of the data. Figures 3.11 and 3.12 illustrate, respectively, examples of accurate and inaccurate predictions derived from the aforementioned test. In a manner analogous to the earlier test, the faces on the left were utilized as test data, whereas the faces on the right were included in the validation dataset.

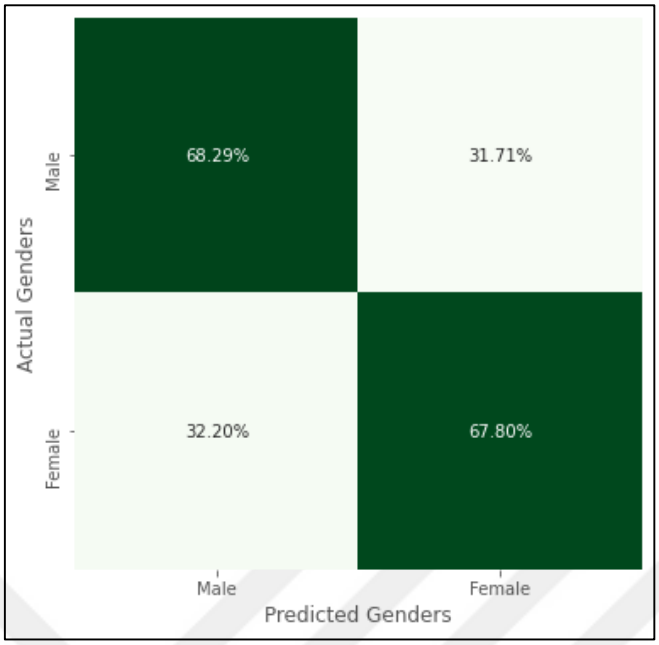


Figure 3.11: Data was Accurately Predicted Based on Sex[15].

The image on the left depicts the input, while the image on the right depicts the validation applied to the input to determine whether or not the information was accurate.

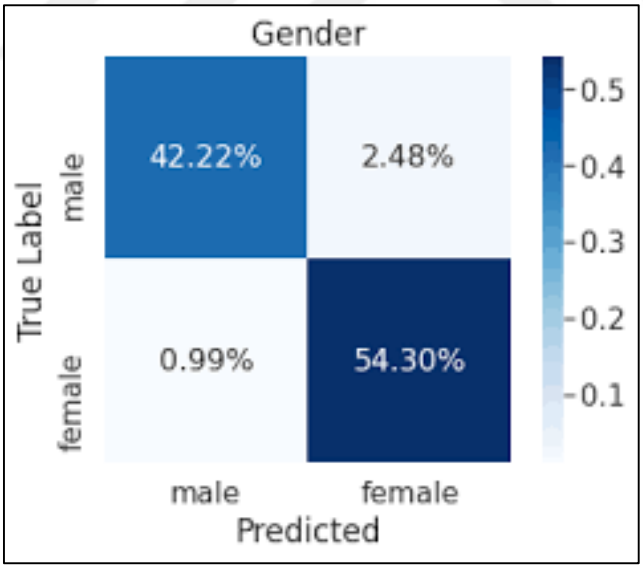


Figure 3.12: A Failure in the Gender Prediction Data[15].

The image on the left depicts the input, whereas the image on the right depicts the validation done to the input to evaluate whether or not it was accurate. It is crucial to remember that some of the wrongly predicted samples have incorrect annotations in the original dataset. As evidenced by the results of the preceding tests, they have constructed a system that

contains all of the necessary components to meet the baseline requirements. In the next chapter, we will discuss the evaluation of such a system, which will consist of a battery of tests designed to identify areas of the current model that should be improved.

3.4 AGE AND GENDER PREDICTION METHODS

When attempting to identify an individual's age or gender from a single image, multiple processes must be taken. Locate and isolate any faces that are visible in the image as the initial step. can discern specific facial features, such as the eyes, mouth, and nose, we may be able to exploit the alignment of the identified face in certain situations. To generate a forecast, we must first locate and accurately align a face, and then incorporate that face into our model.

3.4.1 Face Detection

Face recognition is a crucial subfield in computer vision. There are numerous approaches for unconstrained facial recognition. Certain earlier technologies make use of face edge detection. We began with the most prevalent algorithm. The input image is segmented into squares by the algorithm. A number of low-quality classifiers are then applied to each segment, each searching for a distinct haar-like trait. To determine the Haar-like feature, the sum of intensities in adjacent pixels is computed. The part is only regarded to have face if it survives the whole cascade without being rejected. Rectangles of various sizes are treated in the same manner. AdaBoost is utilized for training the classifier. The primary advantage of this method is that the time needed to calculate a haar-like feature is constant. This is made possible by the utilization of integral images. Integral pictures, commonly referred to as summed-area tables, are a sort of data structure in which the value of each cell is the sum of the cells to its left and above. Only one trip through the code is required to calculate the summed-area table. Histogram of oriented gradients (HOG) is an innovative feature description. The technique calculates the intensity gradient for each pixel of a picture. After that, the image is broken into more manageable portions. For each area, a histogram of the gradients is created. Consequently, just the most prominent gradient is preserved. Support Vector Machines are then utilized to classify the HOG image. Dlib, a popular software library, uses HOG-based facial recognition. Multitask Cascaded Initially,

Convolutional Neural Networks (MTCNN) were introduced and utilized in a deep learning system for face detection. The technology can detect faces and facial landmarks rapidly and precisely. There is a convolutional neural network chain reaction occurring here. The input image is scaled to form a pyramidal shape. Multi-scale image representations, also known as "image pyramids," are gaining popularity. Scale-independent object detection is made possible. This is typically used in conjunction with a sliding window, which can be used to locate objects in various locations. The picture pyramid is then given to the first convolutional neural network, Proposal Network (P-Net). This network identifies potential bounding boxes. The candidates are passed to a second network, the Refine Network (R-Net), to eliminate any potential false positives. In both phases, we combine overlapping nearby bounding boxes via non-maximal suppression. In edge detection, non-maximum suppression is utilized to eliminate unwanted edges. All gradient values beyond the local maximum are concealed. The following phase is an Output Network (O-Net), which narrows the alternatives further and performs rudimentary facial landmark localization. When compared to other facial landmark recognition methods, the MTCNN only identifies a total of five features: both eyes, both mouth corners, and the tip of the nose.

3.4.2 Facial Landmarks Detection

Face alignment is primarily dependent on the identification of essential facial landmarks. Nose, eyebrow, eye, mouth, and jaw are some facial landmarks used to distinguish particular face features. The Dlib package utilizes a proprietary implementation of an incredibly quick algorithm. The coordinates of facial landmarks are updated using a cascade of regression trees. At each tree node, the selection of features is controlled by the relative luminance of adjacent pixels. Introduced an alternative technique based on deep learning. Face Alignment Network is an unique neural network presented by the authors (FAN). This network is comprised of four interconnected "Hourglass networks" and a newly created bottleneck node. The development of the Hourglass networking system. Its intended use is posture estimation. An image is initially down-sampled to a high density using an Hourglass network before its properties are combined with those of images of varied resolutions. This is done so that the image can then be up-sampled.

3.4.3 Age Prediction

In the past, determining a person's age has been the topic of research as a subproblem of face aging. Numerous techniques first developed for the study of facial aging can be applied immediately to the problem of age classification. When science first began to emerge, one of the first topics to be examined was the optimal method for classifying individuals by age. The pupils of the eyes are utilized to locate the beginning oval of the face, which is the first step of their proposed method. Then, the borders of the initial oval are utilized to locate the chin and sides of the face. Once all facial characteristics have been identified, various ratios are calculated. The obtained ratios are used to classify a person's face traits. The presence of wrinkles on a particular portion of the face is also utilised in this manner as an indicator of age. Face detection is done on images that have been rotated by a factor of 5, and the image with the highest face score is chosen. After employing this technology, face landmark detection is no longer required. In addition, the DEX technique generates a buffer zone that is forty percent greater than the region surrounding the identified face. After identifying faces, the data are next processed by a convolutional neural network. When applied to the IMDB-WIKI data set, the technique produced a mean absolute error of 3.221. There is also the possibility of approaching the age estimation problem from a categorization standpoint. Often, knowing a person's age range rather than their exact age provides more information than the latter. Utilizes an SVM that does not remember the training data in order to classify faces according to their ages. Initially, they focus on identifying specific traits of the discovered face. The coordinates of the landmarks are used as inputs for an affine transformation, which is subsequently used to place the face in the proper perspective. For the goal of classifying the aligned face, a linear Support Vector Machine is utilized. To prevent overfitting, they recommend employing a dropout mechanism, analogous to dropout in neural networks. Their method does not disregard neurons at random; rather, it sets some attributes to zero at random. On the Adience dataset, they obtained an accuracy rate of 45.1% exact and 80.7% off-target. The design is comprised of only three convolutional layers and two fully linked layers. They assert that the likelihood of overfitting in a network of this size is substantially reduced. Beginning with a 256-by-256-pixel image, we scale it down to 256 pixels by 256 pixels. Then, a 227 by 227-pixel square taken from the middle of the image is sent to the network. This

approach obtained an accuracy rate of 50.7% for accurate forecasts and 84.7% for one-off predictions when applied to the Adience dataset. Using this strategy, they were able to correctly classify individuals into their respective genders with an accuracy of 86.8 percent. Effective neural network topology is required for numerous practical applications, including embedded devices. present a network structure known as the Soft Stagewise Regression Network for regression (SSR-Net). For the aim of determining a person's age, the method involves a multi-step system, with each step being an improvement on the previous one. A model trained on the IMDB-WIKI dataset and assessed on the MORPH2 dataset produced the best results.

3.4.4 Gender Prediction

Face identification is intrinsically linked to the issues of gender prediction and classification. SEXNET was one of the earliest methods for determining a person's sexual orientation. Utilizing a fully connected neural network with three layers and 900, 40, and 900 neurons in each layer, it was able to classify photographs captured at a resolution of 30 by 30 pixels. Support Vector Machines (SVMs) are used to classify gender. This method employs a support vector machine (SVM) with a radial basis kernel in order to find subsampled images with a resolution of 21 by 21 pixels. The issue of gender classification, which is a subproblem of age prediction, has received a great deal of attention recently. The great majority of currently known algorithms for predicting both age and gender employ neural networks arranged in essentially the same way. The challenge of modeling gender as a binary classification issue is easily resolved. If the gender is known, the result of the regression model can be interpreted as a measure of certainty.

4. PROPOSED METHOD

In this chapter the architecture of the Software developed to solve the problem of face detection, Facetraking and determination of age, gender and emotional state of the face is presented and described. For this, the Software structure to be developed will be presented, detailing the blocks that compose it in an architecture. Finally, it will be described where and how the proposed architecture will be implemented.

4.1 SOFTWARE ARCHITECTURE

After studying several articles and works that addressed the problem, analyzing several existing solutions, it was possible to have a general idea of how the problem will be addressed and what will be the best way to create a solution. The input data of the system to be developed should be the images obtained through a visible spectrum camera. The output data should be the data of each face (age, gender and emotional state) and an identifier (ID). For this process to happen, an architecture was designed for the system. Figure 4.1 shows a schematic where it is possible to see the blocks needed to transform the input data and thus obtain the output data. The block in the figure represents a part of the developed code that is responsible for performing a set of specific tasks. In the architecture of figure 4.1, an image is captured, then the detection of the faces in it must be carried out, the faces must be sent to each of the 3 blocks and these will carry out the necessary processes and extract the necessary information from them. After that, there is an association of the data obtained from each block and then we obtain the output data of the system. The output data will again be.

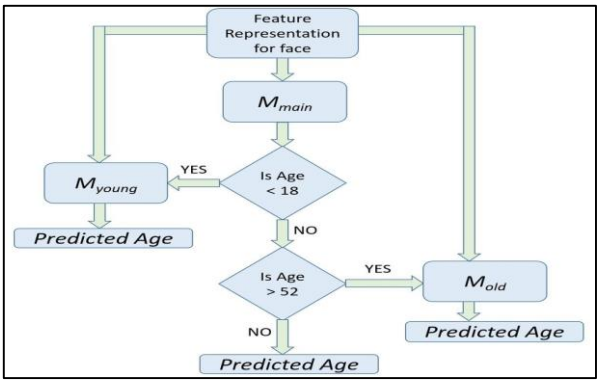


Figure 4.1: Software Architecture.

used in the following image by the facetracking block. The architecture of image 4.1, the Face detection block, Age and gender determination and Emotional state determination will use neural networks but will be 3 separate networks. The networks will be separated, so it is possible to work each network separately and tune and correct each network without influencing the result of another network. The problem solved in each block is different from the problem of the others and depends on different factors.

4.1.1 Face Detection

In the Face detection block, all faces in the image must be detected and located, and then each face in the image must be cut out. This problem is a very common image classification problem and is quite common in neural networks. Figure 4.2 shows an example of what this block should do, on the left we have the input image and on the right we have the output images with the faces. This block is very important for the rest of the system, as the correct identification of the images will lead to a good performance of the remaining blocks. This block should have some robustness and thus be able to identify faces even with some variations in luminosity. Distance is also an important factor and he should be able to identify faces at distances that can go up to about 4/5 meters. The various articles and works studied in Chapter 2 used OpenCV's Haar Cascades library. The Haar Cascades library is a feasible solution but it cannot detect faces at such a high distance. In chapter 2, article [8] is mentioned. This block must create a similar network and use the same or identical block types (convolution, Maxpooling and Full Connected). The quality of the images of the faces is also a very important aspect, as the performance of the other blocks depends on it. This will be a parameter to be analyzed in the

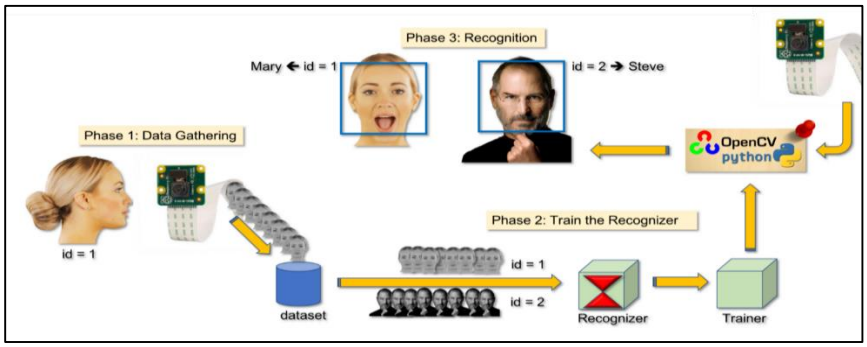


Figure 4.2: Example of What Should be Done in the Face Detection Block, Construction of this Block.

4.1.2 Determination of Age and Gender

In the Determining age and gender block, the approximate age of the face and its gender must be determined. The determination of age and gender are in the same block as they use similar parameters of the face. Determining age and gender requires information not only from the center of the face but also from the surrounding area such as hair. Thus, it will be more appropriate to have an image of the face as seen in figure 4.3(b).

(a) Centre of face (b) Face complete

Figure 4.3: Examples of Input Data in the Determining Age and Gender Block.

The determination of age uses only the image of the face, as studied in article [11] in chapter 2, it only allows us to find the apparent age as there is no more input data. The problem of age determination, using only the face image, can be approached in the following way. The neural network allows extracting important information from an image (such as wrinkles and facial features in the mouth and eyes area) and being able to detect details that the human eye cannot, but that information, like a human, is only based on the image. The image only indicates the visual aspect and this does not change in the same way in all people. There are people who are 30 years old who look like they are 40 years old and vice versa. Therefore, the data that will be used to train the network must have the real value and apparent values of the person's age. There are other factors such as brightness and make-up that can change the output value determined by the neural network. The problem can be approached as an image classification problem where the network classifies the face between integer values between 1 and 100. The age determination is a regression problem so it is necessary to add the Euclidean loss layer to the last layer of the VGG network. function. This layer should analyze the output of the VGG network, the ages and their respective probabilities, and calculate the final age value. The problem can be approached as an image classification problem where the network classifies the face between integer

values between 1 and 100. The age determination is a regression problem so it is necessary to add the Euclidean loss layer to the last layer of the VGG network. function. This layer should analyze the output of the VGG network, the ages and their respective probabilities, and calculate the final age value. The problem can be approached as an image classification problem where the network classifies the face between integer values between 1 and 100. The age determination is a regression problem so it is necessary to add the Euclidean loss layer to the last layer of the VGG network. function. This layer should analyze the output of the VGG network, the ages and their respective probabilities, and calculate the final age value. Gender determination is a simple classification problem in which a VGG network is also used. Some parameters can be decisive in determining the gender, such as hair, beard and some facial features. The result of this network is shown in figure 4.4.

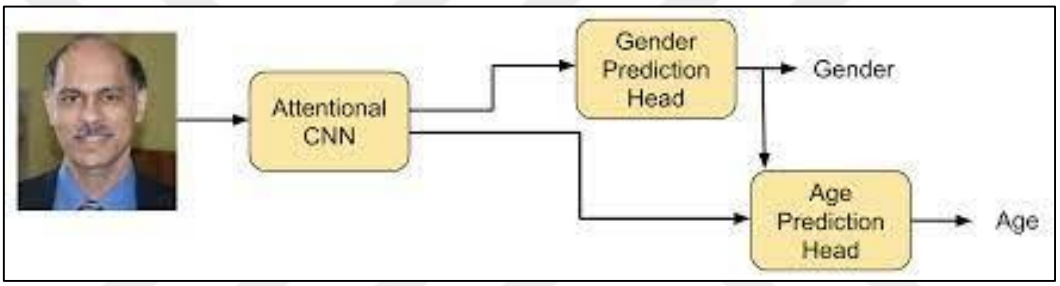
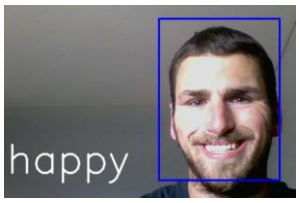
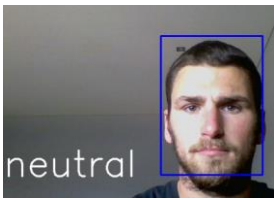


Figure 4.4: Example of What Should be Done in the Determining Age and Gender Block.

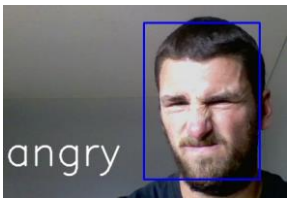
The data determined in this block will then be sent to the Data Association block, where they will be associated with a certain face and an identification.



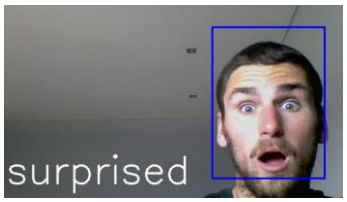
(a) happy



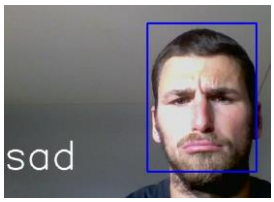
(b)Neutral



(c)unnerved



(d) surprised



(e) it's sad

Figure 4.5: Examples of Various Emotional States.

4.1.3 Face Traking

The Face traking block should receive the various faces found by the Face detection block, assign them an ID. If a face appears in several Frames in a row, it must always have the same ID. For example, an image is obtained with 3 different faces, face A, face B and face C. Each face must be assigned an ID, face A with ID 0, face B with ID 1 and face C o ID 2. As long as faces A, B and C continue to appear in the image, the corresponding ID should appear as well. If face B leaves the image its ID should disappear and the remaining two ID's should continue to follow the faces. If, after a few moments, face B reappears, it must be assigned a new ID, in this case ID 3. Assigning IDs and tracking faces should be done as follows. The set of identified faces and the coordinates of the respective image rectangle are sent to each camera frame. If the faces are new and in that area of the image there is no face from the previous Frame, it is because the face is new and a new ID must be assigned to each of the faces. The ID of each of the faces and the respective position of the face must be saved and analyzed in the next Frame. A detection region is also created and saved for each face. The detection region corresponds to the area of the face rectangle plus a margin. In the following frame, the faces found and their respective rectangle in the image are sent again. If the rectangle of this face corresponds to a detection region of the previous frame, then the same ID of the face that corresponds to that detection region is assigned. If there is no match, a new ID must be assigned. If a detection region does not match for some frame this ID and the corresponding data must be deleted. The output data of this block will be the coordinates of the face in the image and its ID, to help a better visualization, a specific color is also assigned to the ID.

5. IMPLEMENTATION AND RESULTS

In this chapter, the implementation of the various blocks of the architecture mentioned in chapter 4 and the results obtained during the various phases of implementation and the final system will be presented and described. In the first phase, the Face detection block will be implemented. Then the Face tracking blocks, Determination of age and gender and Determination of emotional state. In the final phase, the data association block is made. The results obtained in each block will be analyzed and finally the overall result of the system will be analyzed. The various sections described here are activated in a chain, ie sending a Frame activates the first block. If there are faces, the remaining blocks that send information to the following ones are activated. The cycle repeats itself from frame to frame. In blocks where neural networks are used, before starting to capture images from the camera or a video file, the structure/architecture of the respective network is created and the trained file of the respective network is loaded. After that, the networks are ready to be used by the respective blocks.

5.1 FACE DETECTION

The network can detect faces. The maximum distance obtained, in good light conditions, is 4.50 meters, which is a satisfactory value. Figure 5.1 presents the results of the accuracy obtained in the face detection test for two different convolutional neural networks (CNNs): the traditional CNN and the modified "Our CNN". In this comparison, Our CNN achieved a higher accuracy of 97%, while the traditional CNN reached an accuracy of 95.3%. The improved performance of Our CNN can be attributed to the incorporation of Haar-like features. Haar-like features are simple rectangular features that represent the intensity difference between adjacent regions of an image. These features can efficiently capture the local patterns and structures of an object, such as edges and textures, which are crucial for object detection tasks, including face detection. By integrating Haar-like features into the CNN architecture, Our CNN can leverage the benefits of these features, resulting in better feature extraction and representation. The use of Haar-like features allows Our CNN to more effectively identify and differentiate between facial and non-facial regions in an image. Consequently, this leads to better detection performance and, ultimately, an increased accuracy compared to the traditional CNN. the accuracy increase in Our CNN

can be attributed to the incorporation of Haar-like features, which enhance the network's ability to capture local patterns and structures, resulting in improved face detection performance.

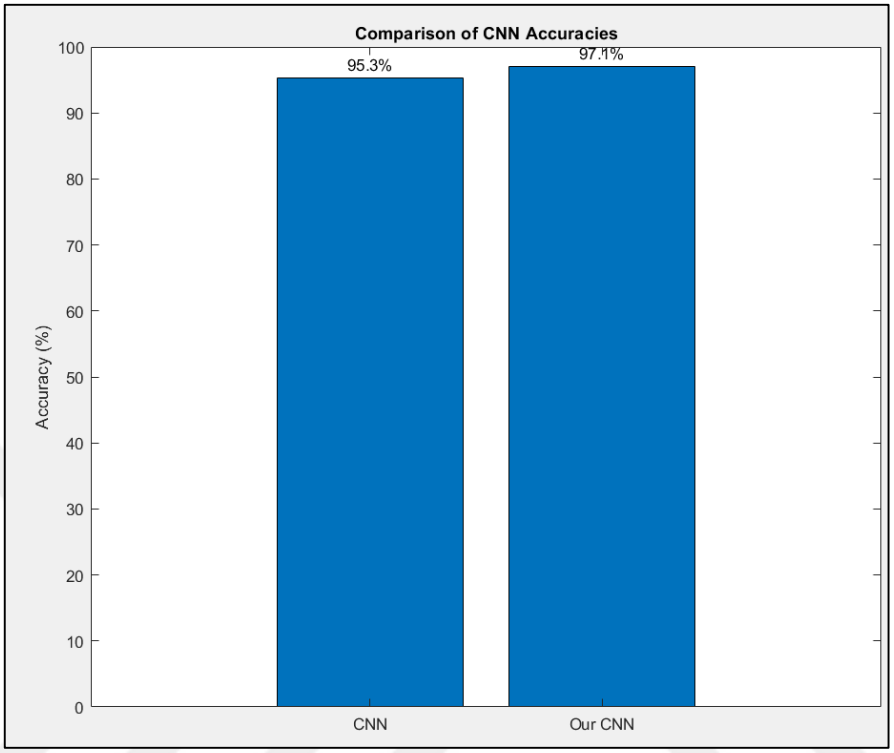


Figure 5.1: Result of the Accuracy Obtained in the Face Detection Test.

Image 5.1 shows an example of a detection made approximately 4.5 meters away from the camera. Then tests were carried out at various distances and the quality of the image of the face was analyzed. The image quality of the face is an important factor and has an impact on the result of the remaining blocks. Figure 5.2 displays the test image used for age, gender, and emotion detection using the "Our CNN-Haar" scheme. This scheme combines the advantages of a convolutional neural network (CNN) with Haar-like features, resulting in an enhanced face detection and analysis system. In the test image, Our CNN-Haar scheme is not only able to detect the age and gender of the person in the image but also classifies their emotions with high accuracy. The improved performance of Our CNN-Haar scheme can be attributed to the following factors: Incorporation of Haar-like features: As mentioned earlier, Haar-like features can efficiently capture the local patterns and structures of an object, such as edges and textures. These features allow the network to better identify and differentiate between facial and non-facial regions, resulting in

improved face detection performance. Robust feature extraction: Our CNN-Haar scheme benefits from the CNN's ability to learn and extract hierarchical features from the input image. This results in a robust representation of the face, which allows for accurate age, gender, and emotion classification. Multi-task learning: Our CNN-Haar scheme is designed to handle multiple tasks, such as age, gender, and emotion classification. By learning these tasks simultaneously, the network can exploit the shared features and correlations between them, leading to better performance and more accurate predictions. Data augmentation: To increase the diversity of the training dataset and improve the model's generalization capability, data augmentation techniques such as rotation, scaling, and flipping can be applied. This allows the network to learn more invariant features and become more robust to variations in the input data. Figure 5.2 demonstrates the effectiveness of the Our CNN-Haar scheme in detecting not only the age and gender of the person in the image but also classifying their emotions with high accuracy. This improved performance can be attributed to the incorporation of Haar-like features, robust feature extraction, multi-task learning, and data augmentation techniques.

Figure 5.2: The Test Image Used for Age and Gender Detection.

In figure 5.2 it is possible to verify that the image of the face is of high quality and it is possible to see all the features of the face. Visually it is possible to easily identify characteristics such as age, gender and emotional state.

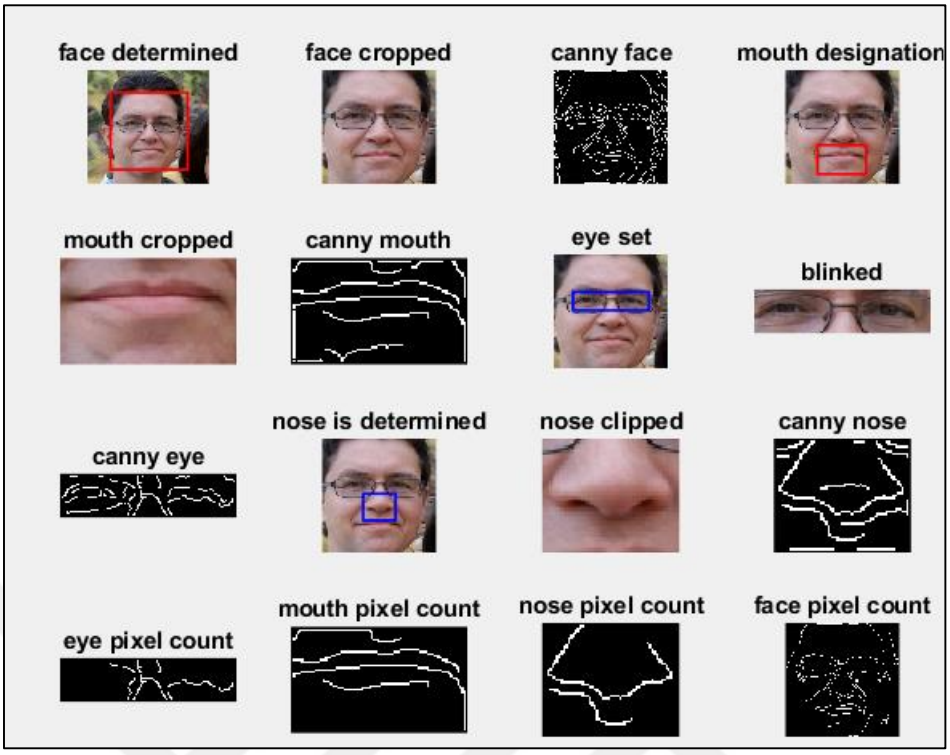


Figure 5.3: Canny Edge Filtering Using Haar Features.

Canny edge filters and Haar-like features can work together in a complementary manner to effectively determine a face in a photo and extract features from the face. Combining these methods can result in better accuracy and faster execution times compared to using only convolutional neural network (CNN) feature extraction layers. Here's how the combination of Canny edge filters and Haar-like features can achieve these improvements: Edge Detection: Canny edge filters are used to detect edges in an image, which are important for distinguishing facial features such as the eyes, nose, and mouth. These filters work by identifying rapid changes in pixel intensity and suppressing non-maximal edges, resulting in clean and thin edge maps (Figure 5.3). The output of the Canny edge filters can be used as input for the subsequent Haar-like feature extraction process. Haar-like Features: Haar-like features are simple rectangular features that represent the intensity difference between adjacent regions of an image. They can efficiently capture the local patterns and structures of an object, such as edges and textures, which are crucial for object detection tasks, including face detection (Figure 5.4). By processing the edge maps generated by the Canny edge filters, Haar-like features can further enhance the identification and extraction of facial features.

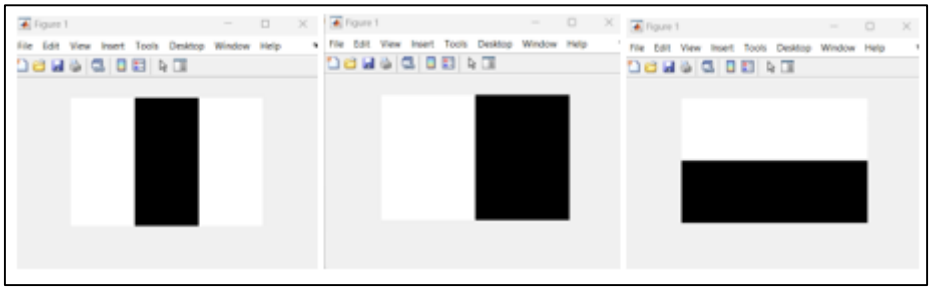


Figure 5.4: Haar Like Feature Blocks Used Alongside Canny Edge Filter.

Reduced Complexity: Combining Canny edge filters and Haar-like features can simplify the overall architecture of a face detection system. The use of these methods can reduce the number of layers and parameters needed in a CNN, leading to a more computationally efficient model. This, in turn, can result in faster execution times. Improved Accuracy: The combination of Canny edge filters and Haar-like features allows for more effective feature extraction and representation, as they capture important facial structures more effectively than CNN feature extraction layers alone (Figure 5.5). This leads to better detection performance and, ultimately, increased accuracy. Faster Convergence: The use of Canny edge filters and Haar-like features can help a CNN to converge faster during the training process, as these methods capture important facial structures more effectively. Faster convergence means that the model requires less training time to achieve optimal performance.

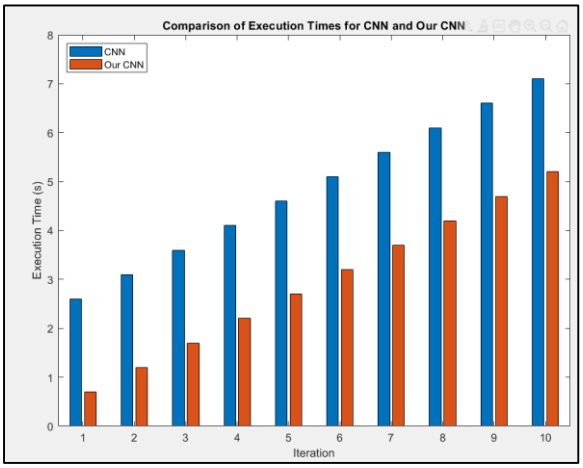


Figure 5.5: Results of the Time of Our CNN Compared to Regular CNN.

Figure 5.5 illustrates the comparison of execution times between the traditional CNN and the modified "Our CNN" for a face detection task. In this comparison, Our CNN

consistently demonstrates lower execution times compared to the regular CNN. The reduced execution time of Our CNN can be attributed to the incorporation of Haar-like features, which lead to an optimized time performance. Haar-like features can be computed rapidly using the integral image technique, which reduces the computation time required for feature extraction allow for the efficient selection of relevant features for the task at hand. By focusing on the most informative features, the network can achieve better performance while using fewer resources and reducing computational time.

6. CONCLUSION AND FUTURE WORK

This dissertation addressed the development of a system capable of identifying and tracking a face in an image and determining characteristics of a face such as age, gender and emotional state only with the use of a visible camera. In an initial phase, the study was carried out to identify the best works carried out for each of the characteristics to be determined and which type of neural network is most suitable. In this phase, some identical works were also analyzed and that extracted a set of characteristics of the face. From this study, it was possible to verify that the type of network used by most networks were VGG networks or divergent networks. The VGG network was transversal in the various works and its versatility allowed with some changes of blocks to determine different characteristics. It was also possible to verify that the networks that had been studied to carry out the work of these articles were identical or similar to those designed in the articles, it is then concluded that the architecture to be implemented should be similar to those referenced. The good performance of the neural network depends more on the database and images used in the training. This study also allowed us to identify some simple image manipulation techniques that could improve the system's performance. Then it was designed how the implementation would be carried out, what would be the architecture of the software and what would be the flow of information between the various blocks. At this stage it was also necessary to design how Facetraking would be carried out and how the data association would be made so that the data were all grouped. After everything was designed, the construction of the networks designed in the referenced articles began. The networks were tested with the training obtained in the articles as it was found that they had performed training with a wide variety of images and the results had been quite satisfactory. Adding more images to these workouts wouldn't have much of an impact on the net. During Chapter 6. The implementation of the various blocks, several tests were carried out to analyze the data obtained and if it would be necessary to improve any previous block. During the implementation, a good performance of the blocks in the attribution of the characteristics was verified. It was also verified the impact of some factors such as the luminosity, and the variation of the same in the distance, which could lead the network not to assign the most correct state. It was also verified that the Facetraking block sometimes

did not have the most appropriate behavior and it will be necessary to improve it. As a future work, we intend to implement the remaining blocks that use artificial networks of the system developed in VPU and analyze the performance in both systems (VPU and CPU) for each block. In the future, it would also be important to implement the system using multiple NCS, in which each one of them executed each of the networks used. In this way it will be possible to verify the performance of the system developed in VPU and thus compare with the results obtained in CPU. The implementation of this work in ROS will also be a task to be accomplished. ROS is a platform widely used in robots in the most diverse areas. The implementation of this work in ROS will facilitate the integration of this system in robots that use ROS. Finally, the implementation of this system in a service robot that interacts with humans. In this way, the functioning of the system is validated and the general objective of this type of work, which is to improve human-machine interception, is fulfilled. In this way, provide the expansion of the robotics area and a better integration of robots in society.

REFERENCES

[1] Abedin, M.Z., Nath, A.C., Dhar, P., Deb, K., Hossain, M.S.: License plate recognition system based on contour properties and deep learning model. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 590–593. IEEE (2017)

[2] Ahmed, T.U., Hossain, M.S., Alam, M.J., Andersson, K.: An integrated CNNRNN framework to assess road crack. In: 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2019)

[3] Ahmed, T.U., Hossain, S., Hossain, M.S., ul Islam, R., Andersson, K.: Facial expression recognition using convolutional neural network with data augmentation. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 336–341. IEEE (2019)

[4] Althnian, A., Aloboud, N., Alkharashi, N., Alduwaish, F., Alrshoud, M., Kurdi, H.: Face gender recognition in the wild: an extensive performance comparison of deep-learned, hand-crafted, and fused features with deep and traditional models. Appl. Sci. 11(1), 89 (2021)

[5] Amayeh, G., Bebis, G., Nicolescu, M.: Gender classification from hand shape. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7. IEEE (2008)

[6] Basnin, N., Nahar, L., Hossain, M.S.: An integrated CNN-LSTM model for micro hand gesture recognition. In: Vasant, P., Zelinka, I., Weber, G.-W. (eds.) ICO 2020. AISC, vol. 1324, pp. 379–392. Springer, Cham (2021).

[7] BenAbdelkader, C., Griffin, P.: A local region-based approach to gender classi. cation from face images. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, p. 52. IEEE (2005)

[8] Cao, L., Dikmen, M., Fu, Y., Huang, T.S.: Gender recognition from body. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 725–728 (2008)

[9] Chowdhury, R.R., Hossain, M.S., ul Islam, R., Andersson, K., Hossain, S.: Bangla handwritten character recognition using convolutional neural network with data augmentation. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 318–323. IEEE (2019)

[10] Datta, S., Das, A.K.: Gender identification from facial images using local texture-based features (2018)

[11] Dhall, S., Sethi, P.: Geometric and appearance feature analysis for facial expression recognition. Int. J. Adv. Eng. Technol. 7(111), 01–11 (2014)

[12] Dong, Y., Woodard, D.L.: Eyebrow shape-based features for biometric recognition and gender classification: a feasibility study. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2011)

[13] Hossain, M.S., Sultana, Z., Nahar, L., Andersson, K.: An intelligent system to diagnose chikungunya under uncertainty. J. Wirel. Mobile Netw. Ubiquit. Comput. Dependable Appl. 10(2), 37–54 (2019)

[14] Islam, M.Z., Hossain, M.S., ul Islam, R., Andersson, K.: Static hand gesture recognition using convolutional neural network with data augmentation. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 324–329. IEEE (2019).

[15] Islam, R.U., Hossain, M.S., Andersson, K.: A deep learning inspired belief rulebased expert system. IEEE Access 8, 190637–190651 (2020)

[16] Jamil, M.N., Hossain, M.S., ul Islam, R., Andersson, K.: A belief rule based expert system for evaluating technological innovation capability of high-tech firms under uncertainty. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 330–335. IEEE (2019)

[17] Kabir, S., Islam, R.U., Hossain, M.S., Andersson, K.: An integrated approach of belief rule base and deep learning to predict air pollution. Sensors 20(7), 1956 (2020)

[18] Karim, R., Andersson, K., Hossain, M.S., Uddin, M.J., Meah, M.P.: A belief rule based expert system to assess clinical bronchopneumonia suspicion. In: 2016 Future Technologies Conference (FTC), pp. 655–660. IEEE (2016)

[19] Li, B., Lian, X.C., Lu, B.L.: Gender classification by combining clothing, hair and facial component classifiers. Neurocomputing 76(1), 18–27 (2012)

[20] Lian, H.-C., Lu, B.-L.: Multi-view gender classification using local binary patterns and support vector machines. In: Wang, J., Yi, Z., Zurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 202–209. Springer, Heidelberg (2006).

[21] Mahmud, M., Kaiser, M.S., McGinnity, T.M., Hussain, A.: Deep learning in mining biological data. Cogn. Comput. 13(1), 1–33 (2021). https://doi.org/10.1007/s12559-020-09773-x

[22] Mahmud, M., Kaiser, M.S., Hussain, A., Vassanelli, S.: Applications of deep learning and reinforcement learning to biological data. IEEE Trans. Neural Netw. Learn. Syst. 29(6), 2063–2079 (2018).

[23] Mozaffari, S., Behravan, H., Akbari, R.: Gender classification using single frontal image per person: combination of appearance and geometric based features. In: 2010 20th International Conference on Pattern Recognition, pp. 1192–1195. IEEE (2010)

[24] Nazir, M., Ishtiaq, M., Batool, A., Jaffar, M.A., Mirza, A.M.: Feature selection for efficient gender classification. In: Proceedings of the 11th WSEAS International Conference, pp. 70–75 (2010)

[25] Rahaman, S., Hossain, M.S.: A belief rule based clinical decision support system to assess suspicion of heart failure from signs, symptoms and risk factors. In: 2013 International Conference on Informatics, Electronics and Vision (ICIEV), pp. 1–6. IEEE (2013)

[26] Ramteke, S.P., Gurjar, A.A., Deshmukh, D.S.: A streamlined OCR system for handwritten Marathi text document classification and recognition using SVM-ACS algorithm. Int. J. Intell. Eng. Syst. 11(3), 186–195 (2018).

[27] Rezaoana, N., Hossain, M.S., Andersson, K.: Detection and classification of skin cancer by using a parallel CNN model. In: 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECONECE), pp. 380–386. IEEE (2020)

[28] Sajja, T.K., Kalluri, H.K.: Gender classification based on face images of local binary pattern using support vector machine and back propagation neural networks. Adv. Modell. Anal. B 62(1), 31–35 (2019)

[29] Serna, I., Pe˜na, A., Morales, A., Fierrez, J.: InsideBias: measuring bias in deep networks and application to face gender biometrics. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3720–3727. IEEE (2021).

[30] Abousaleh, F. S., Lim, T., Cheng, W.-H., Yu, N.-H., Hossain, M. A., & Alhamid, M. F. (2016). A novel comparative deep learning framework for facial age estimation. *EURASIP*

[31] *Journal on Image and Video Processing*, *2016*(1), 47. https://doi.org/10.1186/s13640-0160151-4

[32] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time Convolutional Neural Networks for Emotion and Gender Classification. Retrieved from http://arxiv.org/abs/1710.07557

[33] Baluja, S., & Rowley, H. A. (2007). Boosting sex identification performance. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-006-8910-9

[34] Bordes, A., Glorot, X., Weston, J., & Bengio, Y. (2012). Joint learning of words and meaning representations for open-text semantic parsing. *Artificial Intelligence and Statistics*, 127–135. https://doi.org/10.1109/CVPR.2015.7298594

[35] Bruna, J., Szlam, A., & LeCun, Y. (2014). Signal Recovery from Pooling Representations. *ArXiv Preprint*. Retrieved from http://arxiv.org/abs/1311.4025

[36] Cadène, R., Thome, N., & Cord, M. (2016). Master's Thesis: Deep Learning for Visual Recognition. Retrieved from http://arxiv.org/abs/1610.05567

[37] Canziani, A., Paszke, A., & Culurciello, E. (2017). An Analysis of Deep Neural Network

[38] Models for Practical Applications, 1–7. Retrieved from http://arxiv.org/abs/1605.07678

[39] Caruana, R. (1997) .Multitask learning. In *Machine learning* (pp.41–75). https://doi.org/10.1109/TCBB.2010.22

[40] Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J. (2014). Joint cascade face detection and alignment. In *European Conference on Computer Vision* (pp. 109–122). https://doi.org/10.1007/978-3-319-10599-4_8

[41] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions.

[42] *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*,

[43] *2017–Janua*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[44] Ciresan, D., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). HighPerformance Neural Networks for Visual Object Classification. *Advances In Neural Information Processing Systems*, 12. https://doi.org/http://arxiv.org/abs/1102.0183

[45] Ciresan, D., Meier, U., & Schmidhuber, J. (2012a). Multi-column Deep Neural Networks for Image Classification. *International Conference of Pattern Recognition*, (February), 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110

[46] Ciresan, D., Meier, U., & Schmidhuber, J. (2012b). Transfer Learning for Latin and Chinese

[47] Characters with Deep Neural Networks. *Neural Networks (IJCNN), The 2012 International Joint Conference On*, 1–6. https://doi.org/10.1109/IJCNN.2012.6252544

[48] Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and Accurate Deep Network Learning By Exponential Linear Units (Elus). *Iclr 2016*, 1–14.

[49] Cook, A. (2017). Batch Normalization. Retrieved August 10, 2018, from https://alexisbcook.github.io/2017/global-average-pooling-layers-for-objectlocalization/

[50] Cope, G. (2017). Kernels in Image Processing. Retrieved November 30, 2017, from http://www.naturefocused.com/articles/photography-image-processing-kernel.html

[51] Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, 1–14. Retrieved from http://arxiv.org/abs/1406.2572