

Master Thesis



Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Control Engineering

Web application for visual prediction games

Bc. Šimon Pavlín

Supervisor: Ing. Vojtěch Franc, Ph.D.
Field of study: Cybernetics and Robotics
August 2021

I. Personal and study details

Student's name: **Pavlín Šimon** Personal ID number: **470321**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Control Engineering**
Study program: **Cybernetics and Robotics**
Branch of study: **Cybernetics and Robotics**

II. Master's thesis details

Master's thesis title in English:

Web application for visual prediction games

Master's thesis title in Czech:

Webová aplikace pro hraní obrazových predikčních her

Guidelines:

Humans constitute a strong baseline in many computer vision problems like, for example, various prediction tasks emerging in face recognition. However, an exact quantitative evaluation of the human performance is still limited due to lack of databases with sufficient amount of human predictions. The goal of this project is a design and implementation of a web application for playing prediction games which will be used as a mean to collect the missing data via an entertaining form. In a prediction game the task of a human player is to predict a target variable from a presented image. The web page will instantly evaluate the prediction in a form that engages the player's attention and motivates the player to improve his/her performance in the next round. The web page should support easy definition of different prediction games, management of multiple users, management of the collected data and visualization of the results. The framework will be evaluated on prototypical games involving prediction of chronological age from human faces. The collected data will be used to compare human performance with state-of-the-art deep neural networks.

Bibliography / sources:

- [1] Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In IEEE 7th International Conference on Automatic Face and Gesture Recognition, pages 341–345, Southampton, UK, April 2006.
- [2] G. Panis, A. Lanitis, N. Tsapatsoulis, and T.F. Cootes. Overview of research on facial ageing using the fg-net ageing database. IET Biometrics, 5(2), 2016.
- [3] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR-W 2017), Honolulu, Hawaii, June 2017.

Name and workplace of master's thesis supervisor:

Ing. Vojtěch Franc, Ph.D., Machine Learning, FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **28.01.2021** Deadline for master's thesis submission: **13.08.2021**

Assignment valid until:
by the end of winter semester 2022/2023

Ing. Vojtěch Franc, Ph.D.
Supervisor's signature

prof. Ing. Michael Šebek, DrSc.
Head of department's signature

prof. Mgr. Petr Páta, Ph.D.
Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Student's signature

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Ing. Vojtěch Franc, Ph.D., for his valuable guidance and consistent support. I am extremely grateful for all the help and constructive advice he gave me throughout the thesis.

Many thanks also goes to all people who participate in this research. Their efforts allowed to collect important data for our study.

This work would not have been possible to do without the unconditional encouragement and support of my family and friends, who are always there for me.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, 10. August 2021

Bc. Šimon Pavlín

Abstract

This thesis studies a human ability to assess the uncertainty of his/her own predictions. To evaluate this ability objectively, we analyze the human performance in the problem of prediction with the reject option, because the optimal reject option strategy requires the knowledge of the prediction uncertainty. As the test scenario, we considered the problem of predicting the human age from face images. We formulated Age Prediction Games which are special instances of the prediction problems we wanted to study, and which have simple rules understandable to everyone. We implemented a web application where the Age Prediction Games can be played online in an ordinary Internet browser. We used the application to collect responses in the Age Prediction Games on face images from two established age prediction benchmarks, the MORPH and the AgeDB database. We used the collected data to analyze the human performance in age prediction and age prediction with the reject option. We also compare the human performance with the convolution neural network trained on examples. The main finding is that humans are unable to exploit the reject option to improve performance in the game in contrast to simple CNN predictors.

Keywords: age prediction, prediction uncertainty, human versus machine performance

Supervisor: Ing. Vojtěch Franc, Ph.D.

Abstrakt

Tato diplomová práce se věnuje schopnosti člověka odhadnout nejistotu predikce. Pro studování tohoto problému jsme se rozhodli reprezentovat nejistotu predikce jako možnost odmítnout odpovědět ve zdánlivě těžkých případech. Ve studii jsme uvažovali problém odhadu věku člověka na základě poskytnuté fotografie obličeje. Navrhli jsme hry pro odhadování věku, které obsahují pravidla snadno pochopitelná pro člověka a jsou zároveň vhodná pro naši studii. V rámci diplomové práce jsme vytvořili webovou stránku pro hraní predikčních her. Tato aplikace byla následně použita k nasbírání lidských dat pro naši studii na dvou veřejných datasech - MORPH a AgeDB. Nasbíraná data jsme použili pro analýzu rozhodování člověka. Zároveň jsme také porovnali člověka s konvoluční neuronovou sítí. Hlavním poznatkem naší práce je, že lidé, na rozdíl od neuronové sítě, nedokáží vhodně použít možnost odmítnutí odpovědět.

Klíčová slova: odhad věku, nejistota odhadu, člověk versus stroj

Překlad názvu: Webová aplikace pro hraní vizuálních predikčních her

Contents

1 Introduction	1
2 State-of-the-art	5
3 Methodology	9
3.1 Prediction problems and optimal strategies	9
3.1.1 Standard prediction model . . .	9
3.1.2 Prediction with the reject option	10
3.2 Prediction games	11
3.2.1 Type I age guessing game: standard prediction	11
3.2.2 Type II age guessing game: reject option prediction	11
3.2.3 Face datasets used in the study	12
3.3 Machine learning approach	14
3.3.1 The age recognition pipeline .	14
3.3.2 The CNN architecture	15
3.3.3 Training CNN from examples	16
4 Web design and implementation	19
4.1 Developed framework for prediction games	20
4.1.1 Downloading collected data .	20
4.1.2 Measures to prevent adversarial user behaviour	22
4.1.3 Our strategy to attract a large number of players	22
4.1.4 How to enter a new game . . .	23
4.1.5 Query generator	25
4.1.6 User management	26
4.1.7 Implementation details	26
4.2 Age Prediction Games	27
5 Collected data	31
6 Experiments	35
6.1 Training of the CNN based age predictor	35
6.2 Evaluation metrics	35
6.3 Results	36
7 Conclusions	43
Bibliography	45



Chapter 1

Introduction

“To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge.”

Nicolaus Copernicus

Prediction problems occur in many application areas. The problem of prediction involves the estimation of a hidden state based on observations. Humans are trained by evolution to be good in many visual and acoustic prediction problems that have been important for their survival. For example, humans are good at predicting various kinds of information about another human based on his/her face. Later, humans have invented machines which learn to solve prediction problems from data. In the past few years, the field of machine learning has seen a great progress and machines are surpassing humans in a still increasing number of prediction problems. The performance of a predictor, be it human or machine, is most frequently measured in terms of an average discrepancy between the predicted and the true hidden state, and most results comparing humans versus machines are related to this setting. Besides the hidden state, however, it is often equally important to know when we can rely on the prediction and when the prediction is a rather unreliable estimate. The estimate of the prediction uncertainty is important, for example, in safety critical prediction problems, where it is better to refuse predicting than to make a wrong prediction causing a lot of harm. To our knowledge, the human versus machine performances in the setting involving estimation of prediction uncertainty are much less known. We are unaware of any widespread benchmark that would be dedicated to this kind of problems. This thesis aims to fill the gap by defining a prediction problem involving the estimation of the prediction uncertainty and creating benchmark data which can be used to compare the performance of humans and machines.

As a testbed, we consider the problem of predicting the age of another human based on a facial image. This is a problem in which humans are known to be good provided their performance is measured in terms of the average deviation between the predicted and the true age. The question we aim to answer is how humans perform in the scenario, which, besides the age estimate, also requires an estimate of the prediction uncertainty. To answer the question, we evaluate the human performance in the prediction problem with a reject option. It is well known that the optimal reject option

strategy requires the knowledge of the prediction uncertainty. To measure the human performance in the prediction problems, we design two types of Age Prediction Games. The rules of the games are formulated in a simple language and should be easy to understand for the majority of humans without any knowledge of statistics. The game of type I is an instance of the standard prediction model which aims to minimize the average prediction error. The performance in the game of type I is used as a reference point. The game of type II is used to measure the human ability to assess the uncertainty of his/her predictions. Because of it, the game of type II is an instance of the prediction model with the reject option represented by a fixed penalty, which gives the player opportunity to refrain from predicting when if the uncertainty would lead to a higher penalty.

To collect a statistically significant sample of data, we have implemented a web application which allows to play the Age Prediction Games to anyone with access to the internet browser. For our study, we choose two datasets of images for which the human-based data are collected - MORPH dataset and AgeDB dataset. We used the collected data to analyse the human ability to estimate the prediction uncertainty and to compare the human performance with a machine, namely, with a convolution neural network (CNN) trained on examples to predict age from facial images.

The contributions of the thesis are as follows:

1. **Web-application based framework for conducting human studies related to prediction from image** - The purpose of the framework is to collect human-based data by playing prediction games. The framework provides an easy and straightforward way to implement a variety of simple prediction games with minimal coding. We used the framework to implement the web application for playing age prediction games of type I and type II. The public website can be accessed on www.ironbrain.net.
2. **A dataset of human responses in age prediction games** - By using the website, we have collected a significantly large collection of human responses in the age prediction games of types I and II on two established face datasets, MORPH and AgeDB. The data can be primarily used to study human ability to predict age and to assess his/her confidence of the prediction, however, we foresee a much broader utilisation of the data. We intend to make the data public.
3. **Analysis of humans performance in age prediction and in age prediction with the reject option** - Using the collected dataset, we evaluate human performance on the two types of age prediction games. Besides, we compared the humans performance with machine, namely, a CNN trained on examples.

■ The roadmap of the thesis

Chapter II The chapter *State-of-the-art* describes the works related to this thesis. Namely, we mention elementary papers on the reject option prediction and papers comparing human versus machine performance in age prediction. We also list the most established datasets for benchmarking age predictors. Finally, we describe the most related web applications dedicated to the prediction of age.

Chapter III The chapter *Methodology* describes the theoretical background of the thesis. Namely, it describes a formal definition of the standard prediction problem and the problem of prediction with reject option, optimal prediction strategies, age prediction games as special instances of the problems, face datasets used in the study, and, finally, the CNN based age predictor used as a machine player.

Chapter IV In the chapter *Web design and implementation*, we focus on the description of the web application, its implementation, and the main features. The first part of the chapter is dedicated to the framework for the implementation of generic prediction games. The second part presents the implementation of the age prediction games using the framework.

Chapter V The chapter *Collected data* presents the dataset which we collected from the website. It provides basic statistics and properties of the collected data.

Chapter VI The chapter *Experiments* is dedicated to the utilization of the collected data for the evaluation of human and machine performance in age prediction games.

Chapter VII The chapter *Conclusions* gives summary of the thesis, what was done and what are the main findings. We also describe the weaknesses of our approach and how the work could be improved.



Chapter 2

State-of-the-art

Prediction with the reject option. The seminar work from 1970 [Cho70] proposed the model with the cost of rejection. In the paper, the optimal strategy with known distribution of $p(x, y)$ is provided. The article also presents the relations of the error rate with the reject cost and the relation of the reject rate with the reject cost. By these relations, the author proves its monotony. In the paper, the cost of rejection is only in the interval (0,1). The arbitrary reject cost, which we need, was discussed for binary classifiers in [Tor00]. Multi class classification as Bayesian strategy with reject option was described in [SH02]. The authors also point out non-Bayesian tasks and their solutions.

Existing datasets. Table 2.1 shows a list of the most important face datasets which are used in the field of computer vision and machine learning as benchmarks for age prediction algorithms. The datasets differ in many aspects, but most importantly in the type of age annotation. The large datasets, e.g., IMDB or CACD, are annotated by an automated procedure which is cheap, however, the resulting annotation is noisy. The precise age annotation requires a human in the loop and hence it is available for smaller datasets only. In our study, we need the precise annotation of the true age. Hence, we selected the two largest datasets with the precise annotation, namely, MORPH and AgeDB. One output of this thesis is a collection of human responses in the age prediction games for faces from these two datasets. The response of the age prediction game of type I is the apparent age which is already available for other datasets listed in the Table 2.1. However, the responses in the age prediction age of type 2, which involve the rejection option, constitute a novel type of annotation that has not been available before.

Comparison human versus machines. The comparison of human ability to predict age versus performance of a machine predictor trained from examples was studied in [HOJ13]. The authors use a private PSCO dataset composed of police mugshots which are similar to MORPH, and a public FG-Net dataset composed of scanned images from personal albums. The authors collected the human age predictions via the Amazon Mechanical Turk service. Namely, they collected responses for 2,200 faces from PSCO and for 1,002 faces from

Dataset	Year	#Faces	Labels	Environ	
FG-NET	2004	1,002	precise	Personal	[GANT16]
MORPH	2006	55,000	precise	Mugshots	[RT06]
GroupsDataset	2009	28,231	apparent	Flicker	[GC09]
IMFDB	2013	34,512	apparent	Celeb	[SMP ⁺ 13]
OUI-Adience	2014	26,580	apparent	Wild	[EEH14]
CACD	2014	163,446	automated	Celebrity	[CCH14]
IMDB-WIKI	2015	≈1M	automated	Celebrity	[RTG16]
ChaLearn	2016	7,591	apparent	Wild	[ETB ⁺ 16]
AFAD	2016	165,501	automated	Selfie	[NZW ⁺ 16]
AgeDB	2017	16,488	precise	Celebrity	[MPS ⁺ 17]
AppaReal	2017	7,591	precise	Wild	[ATE ⁺ 17]
UTKFace	2017	20,000	automated	Wild	[ZSQ17]

Table 2.1: List of the most established face datasets which have been used in the computer vision/machine learning community as benchmarks for age prediction.

FG-Net. They found that their age predictors, based on hand-crafted features and hierarchically organized SVM classifiers, outperformed humans in terms of MAE on both datasets. The human responses collected in this thesis are much larger, and they are collected for two public-domain datasets. Hence, the collected data can be readily used to compare human performance versus many methods that used MORPH and AgeDB as benchmarks. The combination of machine prediction and human prediction was analyzed in [ATE⁺17]. The data collected in this thesis can be used for this type of experiments, however, they are outside the scope of this thesis.

Existing websites. One of the main outputs of this thesis is a web application for online playing Age Prediction Games. There are many other websites for age prediction from face images. We point out AgeGuess¹ and GuessMyAge² which have similar features as the application developed in this thesis. In some of the other similar applications, the user has only a set of discrete choices from which the guessed age is selected, which is significantly different from our study. The AgeGuess website was developed for science purposes in 2011. The goal of the study was a correlation between the apparent and real age, while we concentrate mainly on the analysis of the reject option prediction. The analysis of the collected data were published in [JNV⁺19]. Besides guessing the age, the player can also give additional information about the image, such as its quality or imperfections. The players also get points for their guesses. The dataset of images on this website is created from images uploaded by users. The application contains a Skip button. In contrast to our study, using the "Skip button" is not penalized or rewarded by points. In contrast to the AgeGuess website, the GuessMyAge website has

¹<https://www.ageguess.org>

²<https://www.guessmyage.net>

been created for entertainment only. The application also allows the user to upload personal images to get the opinion of the others on his/her age. Both GuessMyAge and AgeGuess are designed as infinite guessing games, while our game is composed of 10 guesses with a final summary of his performance. Unlike AgeGuess and our application, the user of GuessMyAge has no other motivation than a personal joy.

Machine age predictors. The current state-of-the-art in age prediction, similarly to computer vision problems, is achieved by using deep convolution models trained on a large set of examples. For example, end-to-end CNN based age prediction models were proposed e.g.in [ABBD16, ASJ16]. Evaluation and interpretation of several CNN architectures for age prediction is presented in [LBM17]. A Ranking-CNN for age prediction, which is composed of a series of binary CNN predictors, was proposed [SCD⁺17]. A loss function tailored for age prediction was studied in [PHSC18]. In this thesis, we use a simple CNN architecture proposed in [FC18] as a baseline machine player for comparison with human players.

Chapter 3

Methodology

This chapter is organized as follows. The mathematical definition of two prediction problems and the corresponding optimal prediction strategies are described in Section 3.1. The first problem describes the standard setting with the goal to minimize the expected prediction loss. The second problem describes the problem of prediction with the reject option. Both prediction problems fall into the Bayesian decision making framework [SH02]. Formulation of the age prediction games, as special instances of the two prediction problems, and a description of the used face datasets are a subject of Section 3.2. Finally, in Section 3.3, we describe the CNN based age predictor and its training from examples, which was used as the machine competitor to humans.

3.1 Prediction problems and optimal strategies

3.1.1 Standard prediction model

Let \mathcal{X} be a set of observations and \mathcal{Y} a finite set of hidden states. We assume that a pair of observation and hidden state (x, y) is generated with probability $p(x, y)$ defined over $\mathcal{X} \times \mathcal{Y}$. Then, a prediction strategy is a function

$$h: \mathcal{X} \rightarrow \mathcal{Y}, \quad (3.1)$$

which outputs an estimate of the hidden state $h(x) \in \mathcal{Y}$ based on the observation $x \in \mathcal{X}$. Let $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a loss function. The quality of a prediction strategy h is measured by the expected risk

$$R(h) = \mathbb{E}_{(x,y) \sim p(x,y)}[\ell(y, h(x))] = \int \ell(y, h(x)) dP(x, y). \quad (3.2)$$

The optimal (Bayesian) prediction strategy h_* is the one which minimizes the expected risk R , written as

$$h_* \in \underset{h: \mathcal{X} \rightarrow \mathcal{Y}}{\text{Argmin}} R(h). \quad (3.3)$$

Based on the observation $x \in \mathcal{X}$, the optimal decision can be found. By equation (3.2) one can write

$$h_*(x) \in \underset{\hat{y} \in \mathcal{Y}}{\text{Argmin}} \sum_{y \in \mathcal{Y}} p(y | x) \ell(\hat{y}, y), \quad (3.4)$$

where $h_*(x)$ is the optimal decision for x .

3.1.2 Prediction with the reject option

In the case, when the observation $x \in \mathcal{X}$ is not sufficiently informative, it can be beneficial to refrain from the prediction instead of providing an uncertain estimate. This prediction scenario is called prediction with the reject option. It is obtained from the standard prediction model by extending the output of the prediction strategy with a special decision - reject, and extending the loss function by the cost of rejection.

Formally, the reject option strategy is a function

$$h^\varepsilon: \mathcal{X} \rightarrow \mathcal{Y} \cup \{\text{reject}\}, \quad (3.5)$$

which, given the observation $x \in \mathcal{X}$, either provides an estimate of the hidden state, $h^\varepsilon(x) \in \mathcal{Y}$, or it refrains from predicting, $h^\varepsilon(x) = \text{reject}$. In the case of prediction, the decision is evaluated by a loss $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ as in the standard model. In the case of refraining from prediction, the decision is evaluated by a fixed penalty $\varepsilon \in \mathbb{R}$. This means, the reject option strategy has a loss function $\ell^\varepsilon: \mathcal{Y} \times (\mathcal{Y} \cup \{\text{reject}\}) \rightarrow \mathbb{R}$ defined as

$$\ell^\varepsilon(y, \hat{y}) = \begin{cases} \ell(y, \hat{y}) & \text{if } \hat{y} \in \mathcal{Y}, \\ \varepsilon & \text{if } \hat{y} = \text{reject}. \end{cases} \quad (3.6)$$

The quality of the reject option strategy h^ε is measured by the expected risk

$$R^\varepsilon(h^\varepsilon) = \mathbb{E}_{(x,y) \sim p(x,y)}[\ell^\varepsilon(y, h^\varepsilon(x))]. \quad (3.7)$$

Similar to the standard prediction problem, the optimal (Bayesian) prediction strategy h_*^ε is defined as the one which minimizes the expected risk R^ε , that is,

$$h_*^\varepsilon \in \underset{h^\varepsilon: \mathcal{X} \rightarrow \mathcal{Y} \cup \{\text{reject}\}}{\text{Argmin}} R^\varepsilon(h^\varepsilon). \quad (3.8)$$

It follows that the optimal decision for observation $x \in \mathcal{X}$ reads

$$h_*^\varepsilon(x) = \begin{cases} h_*(x) & \text{if } r_*(x) \leq \varepsilon, \\ \text{reject} & \text{otherwise} \end{cases}. \quad (3.9)$$

where the expected risk associated with the optimal decision is

$$r_*(x) = \min_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} p(y | x) \ell(y', y). \quad (3.10)$$

The value of $r_*(x)$ can be interpreted as the *prediction uncertainty*. It is seen that to solve the reject option prediction problem successfully, one needs the prediction uncertainty or its reliable estimate.

3.2 Prediction games

We have two types of Age Prediction Games. The games are defined such that the game of the first type matches the standard prediction model described in Section 3.1.1, and the game of the second type matches the prediction with the reject option described in Section 3.1.2). The two types of Age Prediction Games, and a description how they match the formal definition are described below.

3.2.1 Type I age guessing game: standard prediction

In the first game, the player is asked to estimate the real age of 10 randomly selected persons based on their portrait images. Each guess is evaluated by penalty points. The player gets as many penalty points as his/her guess differs from the real age. The objective is to minimize the total number of penalty points received for the 10 guesses.

This prediction game is an instance of the standard prediction model defined in Section 3.1.1. Namely, the set of observations \mathcal{X} contains images of human faces. We consider two populations of human faces: mug shots of criminals and photos of celebrities. The datasets capturing the two populations are described in Section 3.2.3. The set of hidden states contains the age categories, $\mathcal{Y} = \{\text{min_age}, \dots, \text{max_age}\}$. Each of the two face populations (datasets) has a different age range. To measure the prediction accuracy naturally for people, the absolute error was used as the loss function, that is,

$$\ell_{\text{AE}}(y, \hat{y}) = |y - \hat{y}|. \quad (3.11)$$

The expected risk

$$R_{\text{MAE}}(h) = \mathbb{E}_{(x,y) \sim p(x,y)}[\ell_{\text{AE}}(y, h(x))] \quad (3.12)$$

is then the expected absolute error. The optimal player of this prediction game is the Bayesian strategy h_* defined by the equation (3.3). The optimal decision based on observation $x \in \mathcal{X}$ is

$$h_*(x) \in \underset{\hat{y} \in \mathcal{Y}}{\text{Argmin}} \sum_{y \in \mathcal{Y}} p(y | x) \ell_{\text{AE}}(\hat{y}, y). \quad (3.13)$$

3.2.2 Type II age guessing game: reject option prediction

In the second game, the player has an option to reject the guess if he/she feels highly uncertain about age of the person. This prediction game is an instance of the prediction with the reject option formulated in Section 3.1.2. The optimal rejection option strategy (3.9) relies on the expected risk (3.10), which can be interpreted as a negatively taken prediction confidence.

The definition of the observation space \mathcal{X} and the space of hidden states \mathcal{Y} is the same as in the type I game. The reject option is penalized by 6 points,

that is, $\varepsilon = 6$. Otherwise, the decision is penalized by the absolute error as in the type I game. The loss used in this game can be formally written as

$$\ell_{AE}^6(y, \hat{y}) = \begin{cases} |y - \hat{y}| & \text{if } \hat{y} \in \mathcal{Y}, \\ 6 & \text{if } \hat{y} = \text{reject}. \end{cases} \quad (3.14)$$

Note that the penalty has a clear interpretation. Namely, the obtained penalty points correspond to the deviation from the true age in years if the player decides to predict. Otherwise, if the player rejects to predict, he/she receives 6 penalty points which is the same as making the prediction error of 6 years. It follows that the optimal prediction strategy is in line with a common sense approach: if the player feels his/her estimate is worse than 6 years, he/she should reject the prediction. Otherwise, he provides his/her estimate. This strategy fits to the equation (3.9) using ℓ_{AE}^6 as the loss.

3.2.3 Face datasets used in the study

The face images shown in the Age Prediction Games originate from two different datasets. The player has an option to select which of the two face datasets appears in the game. Formally, the face images are samples from the underlying distribution $p(x, y)$, where each sample consists of the image x and the true age y . The datasets contain additional information like the gender or race of the subject, however, these additional attributes are ignored in our study. The two datasets used in the games were obtained by subsampling MORPH dataset [RT06] and AgeDB dataset [MPS⁺17]. There are two reasons for selecting MORPH and AgeDB. Firstly, both datasets are established benchmarks for the evaluation of computer vision-based age recognition systems. Hence, the results of our study are comparable with a large body of literature. Secondly, the age annotation corresponds to the true chronological age of the captured subjects, and the annotation was manually verified by the creators of the dataset. As a result, the age annotation is very reliable and the age is an objectively defined value. Other existing datasets often use an automatically created age annotation containing errors, or the age annotation corresponds to subjective apparent age.

Subsampling

The age distribution of both datasets is highly imbalanced. Most populated are age categories from 20 to 40 years, while the younger and older age categories are represented much less, as can be seen in Figure 3.2a and Figure 3.4a. In our study, we aim to measure the performance in the whole range of ages. For this reason, we subsampled the original datasets such that the resulting age distribution was as uniform as possible. Another reason for subsampling is to reduce the total number of faces to appear in the game to accumulate multiple predictions for some faces. Multiple predictions are needed to study crowd-based prediction, which is outside the scope of this thesis.

The subsampling procedure works as follows. For each dataset, we limited the maximal number of samples in each age category to 100. The 100 faces of each category are selected at random. The resulting age distributions of the subsampled MORPH and subsampled AgeDB are shown in Figure 3.2b and 3.4b, respectively. It is seen that the distribution is not perfectly uniform as the young and old age categories of the original datasets have less than 100 faces. The following sections describe additional relevant information about the used datasets.

■ MORPH

The MORPH dataset contains 55,134 facial images annotated with age, gender, and race of the captured subject. The images are police mugshots of American prisoners from 16 to 77 years old. The resolution of each image is 200 x 240 pixels. The images were captured in a controlled environment, where the background is a wall with solid color and except for a few exceptions, the subjects are looking straight to the camera. Exemplar images of the MORPH dataset are shown in Figure 3.1.



Figure 3.1: Examples of facial images from MORPH dataset along with the age annotation.

Figure 3.2 shows the age distribution in the original dataset and in the subsampled dataset used in the prediction games. The subsampling reduces the size of the dataset to 4,729 samples and makes the age distribution nearly uniform. The age category range of the subsampled dataset is fully uniform from 16 to 58 years.

■ AgeDB

The second dataset we used is called AgeDB. This "in the wild" dataset contains 16,488 images of celebrities. It contains facial images of subjects of age between 1 to 101 years. Besides age and gender, this dataset also provides the names of the subjects. The resolution and background of the individual images are highly variable. Exemplar images can be seen in Figure 3.1.

Figure 3.4 shows the number of samples for each age group. After subsampling, the AgeDB dataset contains 7078 samples and gives a uniform age distribution for ages from 18 to 78.

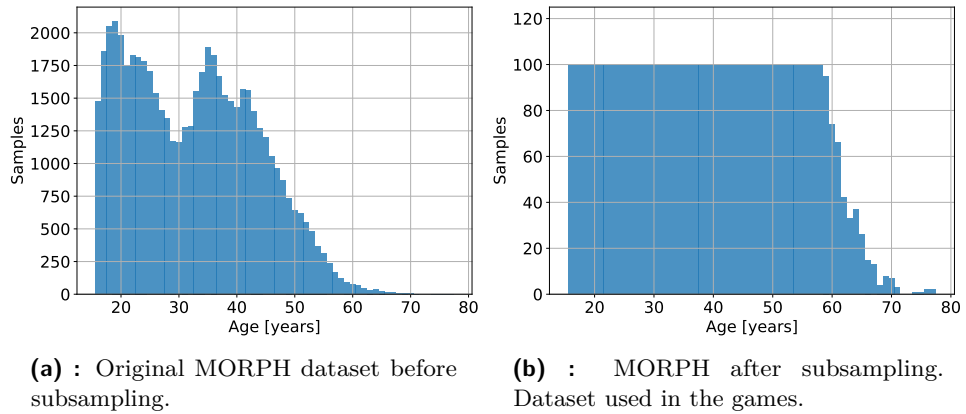


Figure 3.2: Age distribution for MORPH dataset.

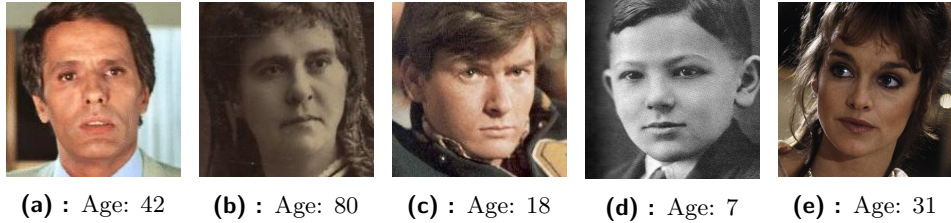


Figure 3.3: Examples of facial images from AgeDB along with the age annotation.

3.3 Machine learning approach

It is highly interesting to compare the performance of humans versus machines. The recent state-of-the-art in visual age prediction, as well as in other computer vision problems, involves deep neural architectures trained from a large number of examples. We implement a computer vision system for automated age prediction based on convolution neural networks (CNN). We treat the age prediction as a multiclass classification problem. The trained CNN outputs an estimate of the posterior probabilities $\hat{p}(y | x)$ of the true age y given a face image x . The age predictor is obtained by plugging-in the distribution $\hat{p}(y | x)$ to formulas defining the optimal Bayes prediction strategy defined by equation (3.4), in case of the standard prediction model, and by (3.9), in case of a model with the reject option

Description of the processing pipeline, the neural network architecture, and the training algorithm is a subject of this section.

3.3.1 The age recognition pipeline

Figure 3.5 shows the processing pipeline of our age prediction system. Before the image is sent to the CNN, the normalization process is applied. The bounding box of the face is founded in the input image and the image is cropped according to the bounding box. The bounding boxes were found by

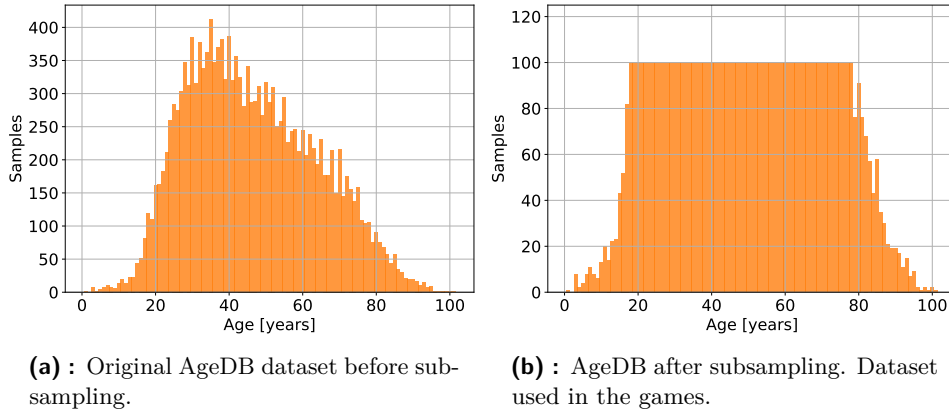


Figure 3.4: Age distribution for AgeDB dataset.

the Adaboost face detector [SM04]. The cropped image is then scaled to the required resolution 100×100 .

Next, the normalized image is processed by CNN, which outputs the probability distribution over ages. The structure of the CNN is described in the following section. The algorithm then determines the optimal prediction from the distribution.

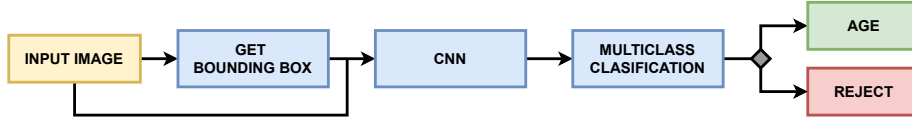


Figure 3.5: Machine learning pipeline

3.3.2 The CNN architecture

We used AgeNet [FC18] which is a CNN composed of 9 layers involving standard processing blocks: convolution, max-pooling, and ReLU. The input is a gray scale facial image of size 100×100 pixels. The exact configuration of the AgeNet is shown in Table 3.1. We treat the age prediction as an ordinary multi-class classification problem with N classes. This means, that the last layer of the network is a linear layer with N outputs, each corresponding to one age category. The number of categories varies for the datasets. In case of MORPH, the number of categories is $N = 62$ corresponding to ages from 16 to 77 years, and in case of AgeDB dataset $N = 101$ which corresponds to ages from 1 to 101. We do not consider the reject option as a separate class. To get the posterior distribution over the age categories, we use the softmax on top of the output of the last layer, that is,

$$\hat{p}(y | x) = \frac{e^{\psi(x)^T w_y}}{\sum_{i=1}^N e^{\psi(x)^T w_i}}, \quad (3.15)$$

where x stands for the input image, $y \in \{1, \dots, N\}$ is a label of the age category, w_y , $y \in \{1, \dots, N\}$, are the weight vectors of the last layer, and,

finally, $\psi(x)$ are features extracted by the CNN from x , i.e., $\psi(x)$ is the output of the penultimate layer.

Layer type	Configuration
Output	distribution over N outputs
Soft-Max	
Convolution	filt: N , k: 1×1 , s: 1, p: 0
ReLU	
Convolution	filt: 2048, k: 1×1 , s: 1, p: 0
ReLU	
Convolution	filt: 2048, k: 5×5 , s: 1, p: 0
ReLU	
Convolution	filt: 128, k: 4×4 , s: 1, p: 0
ReLU	
Convolution	filt: 128, k: 3×3 , s: 1, p: 0
MaxPool	2×2 , s: 2, p: 0
ReLU	
Convolution	filt: 64, k: 3×3 , s: 1, p: 0
MaxPool	2×2 , s: 2, p: 0
ReLU	
Convolution	filt: 64, k: 3×3 , s: 1, p: 0
MaxPool	2×2 , s: 2, p: 0
ReLU	
Convolution	filt: 32, k: 3×3 , s: 1, p: 0
ReLU	
Convolution	filt: 32, k: 3×3 , s: 1, p: 0
Input	100×100 gray-scale image

Table 3.1: Configuration of the AgeNet CNN used to predict age from a facial image. The second column describes the number of filters 'filt', the filter size 'k', stride 's' and padding 'p'.

3.3.3 Training CNN from examples

First, we split the dataset into training, validating, and testing data. As the testing data, we use the same subsampled datasets, described in Section 3.2.3, which are used in the Age Prediction Games. The remaining data are split into training part (90%) and the validating part (10%).

The free parameters of the neural network involve the convolution filters and the parameters of the last fully connected layer. We train the parameters by minimizing the cross-entropy loss in the training part. We use the ADAM optimizer [KB17] with the standard setting of the hyper-parameters. Note that we train a separate network for the MORPH and the AgeDB dataset, because both have very different properties (not only the range of age categories, but

also different resolution of the images and the conditions under which the images were captured).

Chapter 4

Web design and implementation

To collect human-based data, we developed a web application running on a dedicated server. The main advantage of the web solution is an easy access to the application for everyone with an ordinary web browser. The intention was to address a wide population of respondents/players to collect more data. The application can be accessed via a web browser on www.ironbrain.net. The index page of the website is shown in Figure 4.1.

The web application is designed to be applicable for different types of prediction games. The rules of a particular prediction game could be defined with minimal coding. Our hope is that the framework will be found useful by other researchers who need to run a human study of a similar kind.

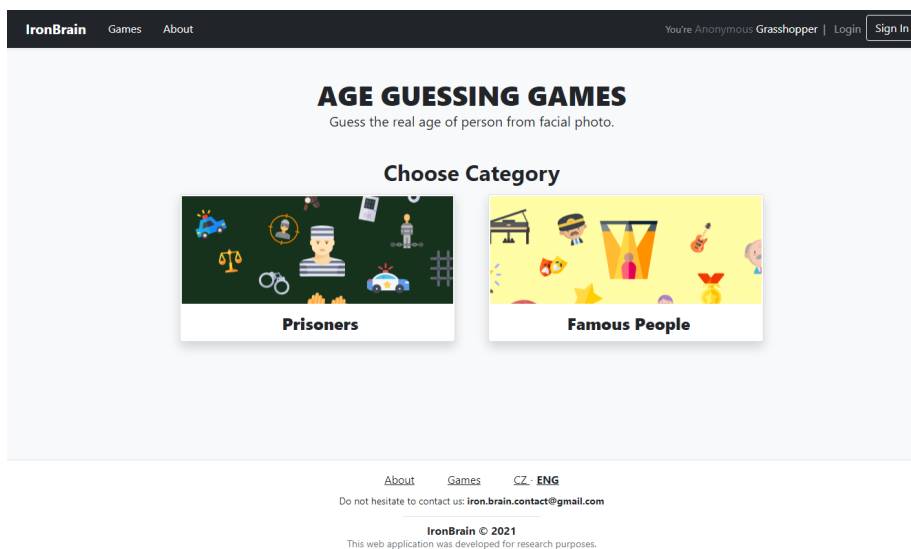


Figure 4.1: Index page of the web application

The generic framework for running prediction games and the way how it can be configured for a particular study is described in Section 4.1. Description of the Age Prediction Games, as a particular instance of the framework, is a subject of Section 4.2.

4.1 Developed framework for prediction games

Each implemented prediction game is composed of a sequence of stages. A diagram of the main life cycle of a prediction game is shown in Figure 4.2. The main stages of a generic prediction game supported by our framework are as follows:

1. The user starts the game.
2. A generator selects a query and presents it to the user. More details on the query generator are in Section 4.1.5.
3. User provides his response to the query.
4. The response is stored to a server database and an immediate evaluation of the response is shown to the user (e.g. the user gets penalty points for mispredicting the age).
5. If the number of queries is less than N , the program continues with the Step 2. In the games described in Section 4.2 we use games with $N = 10$ queries.
6. The game ends and the user gets a summary of his performance in the game.

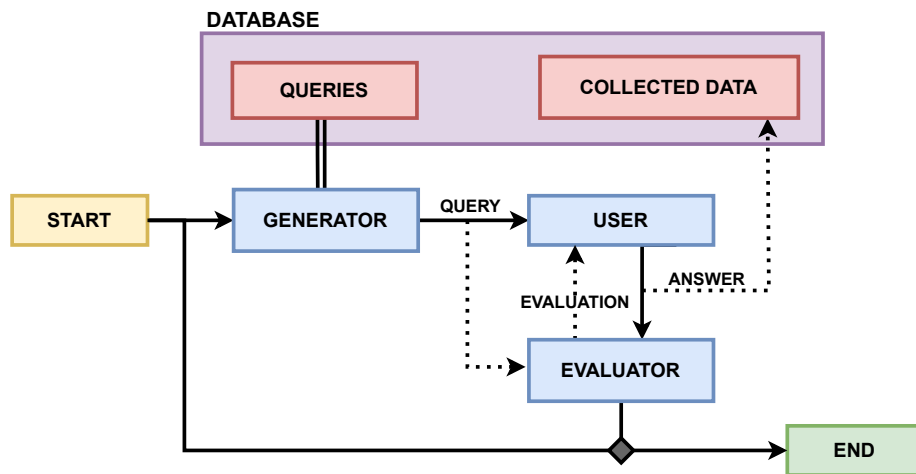


Figure 4.2: Diagram of life cycle of a prediction game

4.1.1 Downloading collected data

The website provides a simple interface for downloading the collected data. This feature is accessible only for admin users. The data are downloaded for each game separately. The data are stored in a simple Coma-Separated-Value (CSV) format. The first line of the CSV file contains description of the columns. The content of each column is the following:

- `answer` - This column contains the user's response.
- `user` - A unique identifier of each user. The identifier is hashed as string for logged-in and anonymous users as well.
- `date` - The exact time of the response.
- `play_id` - A unique identifier of the game. Recall that each game is composed of N responses.
- `task_rank` - The response number within the game.
- `question_id` - A unique identifier of the query in the server database.
- `set_question_id` - A identifier of the query in the provided dataset.
- `question` - Data associated to the query in JSON format.
- `question.xxx` - Data of the query are spread out to multiple columns for easy usage. Examples: `question.age`, `questions.gender` ...

In Figure 4.1, we present an example of the data stored in one row of the file, which can be downloaded by administrators.

Column	Value
<code>answer</code>	62
<code>user</code>	2bbc44d6
<code>date</code>	2021-08-01 22:03:12
<code>play_id</code>	1649
<code>task_rank</code>	5
<code>question_id</code>	63198
<code>set_question_id</code>	8064
<code>question</code>	{ "age":60, "img":63198, "gender":"M", "name":"Bill Murray" }
<code>question.age</code>	60
<code>question.img</code>	63198
<code>question.gender</code>	M
<code>question.name</code>	Bill Murray

Table 4.1: Example of row in the CSV file of the collected data.

■ 4.1.2 Measures to prevent adversarial user behaviour

The player could potentially attempt to determinate the correct answer in an unfair way. In this section we describe the measures implemented to prevent such behaviour.

The correct answer could be determined from the file name of the images, as in many databases the file name reflects the annotation of the image. For example, in the case of MORPH and AgeDB, the file names contain the true age of the person captured in the image. The issue is that this name of each image displayed in a website can be found in the source code of the page. To hide this information, each filename is covered up under a new virtual file name with no information, about the original image.

Theoretically, when the prediction is hard for the player, the player could refresh a page and get a new easier query. To prevent this unwanted behaviour, a new query is not generated if there is already generated query for the user not answered yet. This query is then provided to the user.

As you could see in the GUI of game (Figures 4.6 and 4.7) the interface contains *Quit This Try* button. We added this button to games to avoid worthless answers in the collected data. Imagine a scenario, where the user makes a very bad prediction and is awarded with many penalty points. Then, he/she loss any motivation to properly finish all N guesses. To start over, the player starts answering random guesses as fast as possible. These answers could bring inaccuracies to the collected data. To avoid this, the button to instantly start over was provided.

■ 4.1.3 Our strategy to attract a large number of players

To collect as much data as possible, the application has to be simple and enjoyable to use. To accomplish this approach, several steps were made. In this section, some of the adjustments are described.

As mentioned above, to address a wide audience, the framework is implemented as a web application. Thanks to this, anyone with access to an ordinary web browser can take part in the study. The web application is also available in two languages - Czech and English, which further extends the set of users. To follow the trend and provide a more user-friendly experience, we made the layout of the website responsive, optimal for all sizes of the screen, including smart phones.

Our further intention was to keep the attention of the player and to motivate him to play as many games as possible. To this end, we used two strategies. First, the user can see the best results in recent days. The web shows the TOP 10 leader board for the last 7 days, current month, or overall time. This feature makes the game lively and should invoke a competitiveness among the players to keep them longer on the web. Second, the player obtains a certificate with an evaluation of his performance and comparison with the other players. The certificate is presented to the user that finished at least 5 games of one type. Example of the certificate is shown in Figure 4.3.



Figure 4.3: The certificate given to player

4.1.4 How to enter a new game

Our framework provides a clean solution for implementing simple prediction games. The behavior of the games and the used datasets are not hardwired, but they can be configured with minimal coding. This allows to use the website for a variety of other prediction games.

In the following text, the database is MySQL database containing all important data on the server, such as information about available games on the website, all queries, user information, user answers, or information about single runs of the games. When the administrator is instructed to add data to the database, he/she is encouraged to use the database administration tool as MySQL Workbench¹ or phpMyAdmin².

As the administrator, we mean someone, who can manage and create games in the website. In this section, we describe the procedure that is needed to add and configure a new game to the application.

Registering a new game

Each game is in the server database represented by one row in the table **games**. For the new game, one chooses the *title*, *subtitle*, *description*, *id of the set of queries* and *key name* under which the game could be accessed. In

¹<https://www.mysql.com/products/workbench/>

²<https://www.phpmyadmin.net/>

the *data* column, more attributes of the game are configured. For instance, in the case of age prediction games, it can be an attribute determining game with or without the reject option, or an attribute defining the penalty for rejecting. This column can also contain a description of a game rules as text.

Finally, the most important option to select is the *template*. The template defines the behaviour of the game, the look of the game, and so on. More details on the templates are in the following subsection Defining a game behaviour.

■ Adding a new set of queries

If the admin user needs to use a dataset different from the MORPH and AgeDB, the new queries have to be imported to the server database. In the current implementation, there is no user-friendly tool to fill the database with new data, however the data could be imported directly to the database via a CSV file. Before that, the new dataset of queries has to be registered in the database. This is done by entering a single row to the table `questions_sets`.

To store the queries in the database, one has to use a predefined structure. The CSV file used to import new queries has the following four columns:

- `set_id` - A unique identifier of the set of queries.
- `set_row` - A unique identifier used for matching collected data with the sample in the provided dataset for further analysis.
- `data` - All data about the query. Data are stored in JSON format. For example, in the Age Prediction Game this is a column which stores the true age and an index of the corresponding image.
- `cluster` - A number used in the generator for more advanced query generation.

Each query is represented by one row with the structure described above. We implemented a Python script converting the files with the original annotation of the MORPH and AgeDB datasets into the desired format. In addition to this, the corresponding images have to be presented. The images have to be copied to a folder on the server and each image needs to be represented in the database by a unique identifier. This is done by importing another CSV file with two columns - *id* and *file path to the image*.

■ Defining a game behaviour

As was indicated above, the administrator has a large freedom to customize the game. We created a useful tool which minimizes the number of files that need to be changed when adding a new game.

First, the administrator creates a PHP file, the template, in which behaviour of the new game is defined. The template contains the main information about the game, such as what user interface should be used, what data should

be extracted from the database, or how to evaluate the user responses. The file `age-guess` is an example of the template defining the age prediction games.

Second, the administrator creates a file called `view`, which defines the user interface. This view file is referenced from the template file. The view file is written in HTML. We prepared useful tools for defining what data from the database go where, which elements will be visible during the presentation of the query or the response evaluation, or what data will be sent as the user answer on the given query.

Third, the administrator has to determine which data from the database should be available as the query ("public fields") and what data should be provided to the user only after the user response ("private fields"). For example, in the case of Age Prediction Games, the public field is only the image and the private fields are the age and the name of the subject if available. However, for other games it can be useful to have, e.g., the subject's gender as a private field.

Fourth, the administrator needs to select or implement the evaluation process. The evaluation process is also defined by a single file, where one needs to describe how the user response is processed. This includes the definition of a scoring function, the value of which is provided as a feedback to the user and to update the game score. Besides the scoring function itself, the administrator determines here what data extracted from the user response will be saved to the server database.

4.1.5 Query generator

Every time the user is asked about a query, the program has to decide which query should be chosen. We called the procedure selecting the query as a generator. The properties of the generator are determined by the requirements on the collected data, for example, in the case of the age prediction games, the queries should be sampled uniformly from the database of faces. The generator is configured in the `data` column of a game.

The first functionality of the generator is to avoid unwanted behaviour. To collect valid data, the user has to be asked about each unique query not more than once. This behaviour is hardwired to the generator.

Some studies require to maximize the number of answered queries, while in other studies one prefers multiple responses to a single query at the expense of having less number of unique queries. Our framework allows to customize all such settings.

We also focused on the issues with unbalanced data in the dataset. Image scenario, when a set of queries contains pictures of 10.000 males and 5.000 females. This gives two times more chance for a male to be randomly selected than for a female. To balance this, we added a property called "cluster" to each query in the server database. This property c_q is a number by which the queries can be grouped. For each game, the administrator can define the set of clusters C from which a query should be selected. Before a random query is selected, the server generates a random cluster $c \in C$. The random query

provided to the player is then selected only from queries where $c_q = c$. This ensures the same chance for all queries in C , that the selected query will be in the cluster. Besides balancing the probability distribution over different groups of queries, this feature can also be used to separate the dataset of queries to multiple games, such as a game only with faces of females, or game only with faces of males.

■ 4.1.6 User management

The framework supports both registered and anonymous users. To play a game, the user is not required to be logged in using his/her account. This leads to a more user-friendly approach and, consequently, to collect more data. In the case of anonymous, i.e., not logged in users, a random nickname of the user is generated. This is necessary to distinguish individual users in the database as well as when displaying his/her results in the leader board.

Users can also register an account. During the registration, the users are asked to provide their email address, nickname, and password. This is convenient for the user to keep his identity when playing the games on different devices. After sing-in, the user is also asked to provide additional information, such as age and gender. This information can be used for later analysis of the collected data.

■ 4.1.7 Implementation details

In this section, we share some implementation details about the implemented framework:

- On the front-end side, we used HTML, CSS and Javascript. We used the Bootstrap³ toolkit to get a nicer design with less effort.
- As the back-end language on the server, we used PHP. We used the Laravel framework⁴ for a cleaner and more secure implementation.
- The application uses MySQL database⁵, where all user responses, queries, and games are stored.
- The application is running on Apache server with PHP 7.3. Web servers usually do not use the particular PHP version and do not have the required libraries installed. Upgrading the server can sometimes cause issues for another application running on the same server. For this reason, we used Docker⁶ which packs an image of the server with all requirements. This image can be later run as a subserver on almost any server without a need of configuration.

³<https://getbootstrap.com/>

⁴<https://laravel.com/>

⁵<https://www.mysql.com/>

⁶<https://www.docker.com/>

4.2 Age Prediction Games

The sections above provide a brief description of the generic framework for conducting generic prediction games. This thesis is concentrated around the prediction of human age from facial images. Hence, we applied and tuned the framework for Age Prediction Games described in Section 3.2. Namely, the standard prediction, the game of type I, and the prediction with the reject option, the game of type II. The game of both types can be played on both face datasets, i.e., MORPH and AgeDB, hence we have four unique games in total. Namely, we have the following games:

- **Prisoners** - The standard prediction on MORPH dataset.
- **Prisoners with reject option** - The reject option prediction on MORPH.
- **Famous** - The standard prediction on AgeDB dataset.
- **Famous with reject option** - The reject option prediction on AgeDB.

The age prediction games work as follows. Before playing the game, the player chooses a dataset from which the faces will be generated, i.e., MORPH (a.k.a. "Prisoners") or AgeDB (a.k.a. "Famous People"). The look of the selection page can be seen in Figure 4.1. When the dataset is selected, the player chooses the type of game to play, i.e., the standard age prediction or age prediction with the reject option. The look of the corresponding page is shown in Figures 4.4 and 4.5. Originally we considered a selection page containing all four variants, however, it looked complicated, hence we opted for the sequential selection process described above.

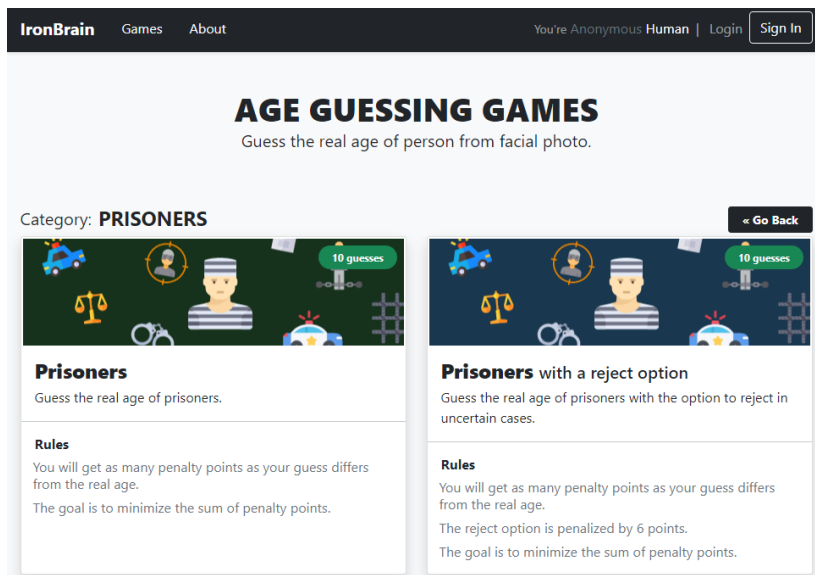


Figure 4.4: Selection page for Prisoners - MORPH dataset

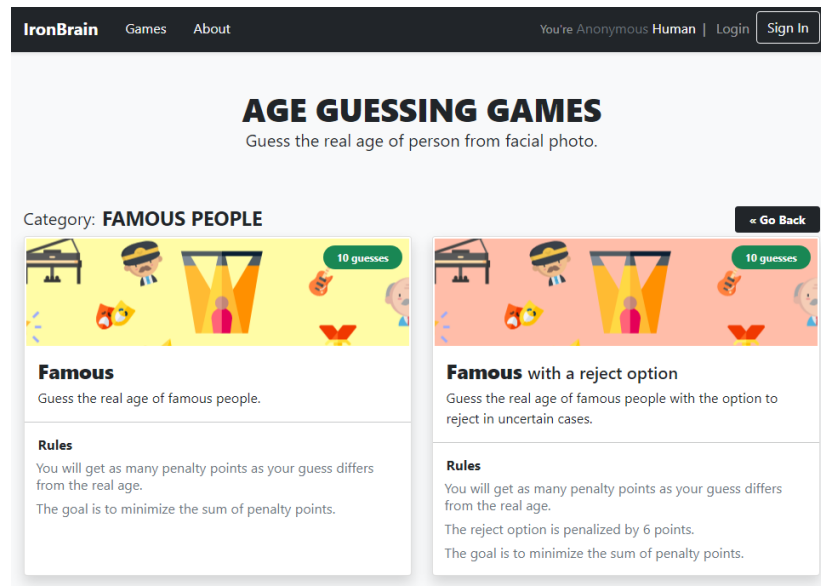


Figure 4.5: Selection page for Famous People - AgeDB dataset

Each age prediction game consists of $N=10$ randomly chosen queries, i.e., facial images, and the game objective is to get the minimal penalty points accumulated for the $N=10$ guesses. In Figures 4.6 and Figure 4.7, one can see an example GUI for both types of prediction games. The GUI displays a short description of the game rules. Visually, the games differ only by the presence or absence of the "Rather Skip" button. The look of the GUI was designed to be intuitive and as simple as possible but still containing all the information that is necessary to play the game like the game rules. To keep the user's attention, the screen also shows the current progress, that is, the penalty points accumulated so far, the corresponding average deviation between the guesses and the ground truth, and the number of already accomplished guesses in the current game.

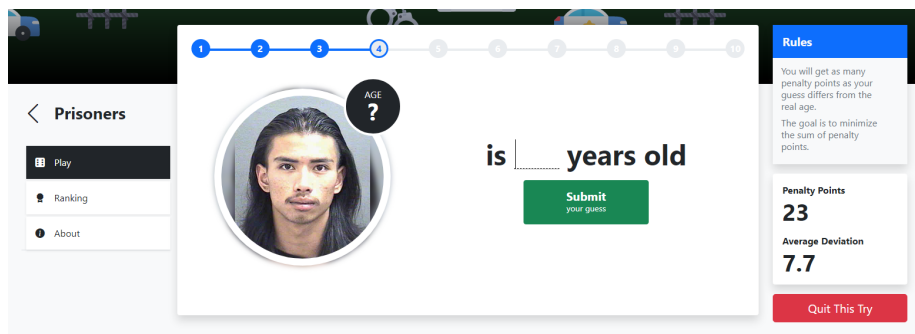


Figure 4.6: User interface for standard Age Prediction Game

The game is concluded with a summary of the achieved results, what can be seen in Figure 4.8. As mentioned earlier, after completing 5 games, the player gets a certificate, an example of which is shown in Figure 4.3.

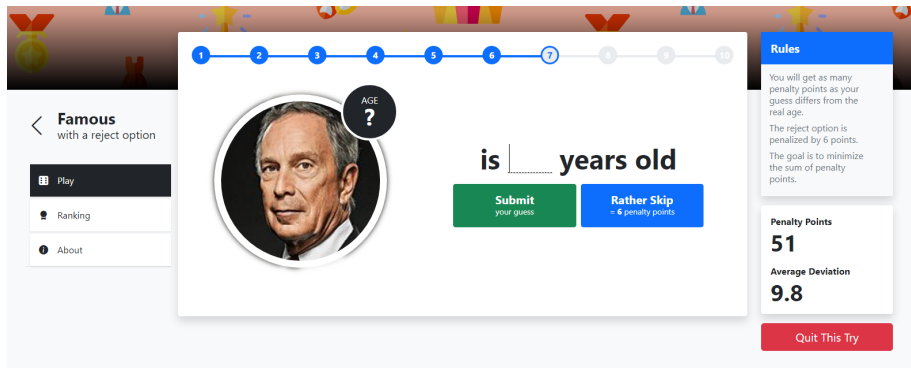


Figure 4.7: User interface for Age Prediction Game with reject option

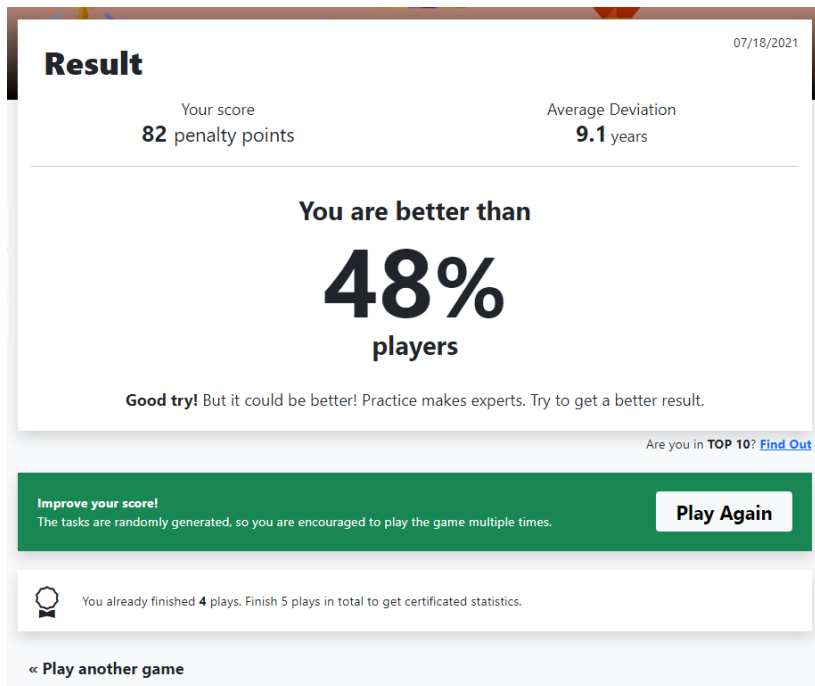


Figure 4.8: Result page shown after the game is completed

Chapter 5

Collected data

We were collecting the data over 4 weeks from 7th of July 2021 to 10th of August 2021. During the period, we collected in summary 16,841 responses. Given a query facial image, a response is either the age guess or the rejection to provide guess. Recall that the data are collected by playing Age Prediction Games, each consisting of 10 face images, for each we collect a response (see Section 4.2). Games were played 2,007 times, out of which 1,510 games were completed, i.e., users submitted a response to all 10 queries. The data were collected from 712 unique users.

Besides the predicted age and the real age, the collected dataset contains additional information about the answer, such as the identifier of the given query, the unique identifier of the user, the date of the prediction or unique identifier of the game. Example of the data is shown in Figure 4.1.

In Figure 5.1a, we can see the number of responses per user. On average, a user submitted 23.7 responses. The median of the number of answers is 15 and the maximal number of responses provided by a single user was 654. As can be seen, there are peaks in the number of responses divisible by 10. This is caused by the fact that a single game contains 10 queries.

The number of completed games per user is shown in Figure 5.1b. On average, each user completed 2.7 games and median of the number of games per user is 2. We can see that the majority of users do not complete more than 2 games. The maximum number of games for a unique user is 61, which corresponds to the user with the maximum number of answers.

The most important statistics of the collected data are summarized in Table 5.1. For MORPH dataset, we collected 7,317 responses, out of which 5,507 responses are for the standard prediction game and 1,810 for the game with reject option. In the figure 5.2, we can see the distribution of answers for MORPH with respect to the age group. For AgeDB dataset, we collected 6,024 responses for the standard prediction game and 3,500 responses for the game with reject option. In total, the application collects 9,524 responses for AgeDB dataset. Figure 5.3 shows the distribution of the responses with respect to age group.

Finally, we present the statistics related to the number of responses per query image. In Figure 5.4, we show the histogram of the number of queries with a given number of responses. To get a better scale of the figure, the

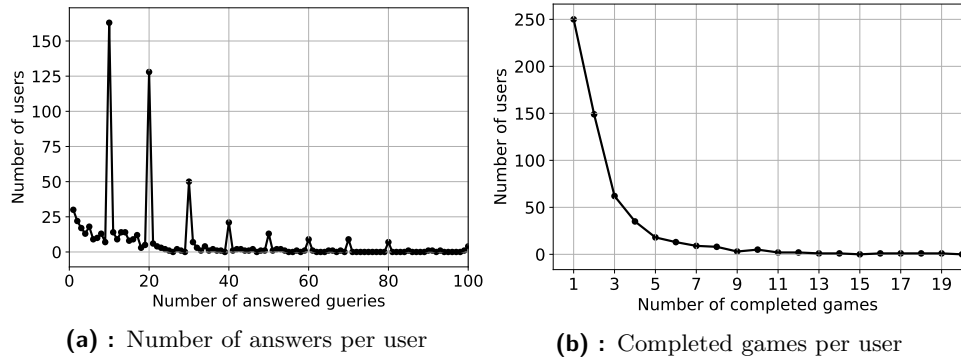


Figure 5.1: The distribution of the number of responses and the number of completed games.

Dataset	Type	Answers	Games	Completed	Users
MORPH	standard	5507	636	510	357
MORPH	with reject	1810	218	161	147
AgeDB	standard	6024	738	528	387
AgeDB	with reject	3500	415	311	198

Table 5.1: The main statistics of the collected data.

queries with no response are not shown. In the case of MORPH, there are 3,237 query images which have at least one age guess for the standard prediction game, and in case of the game with reject option it is 1,533 query images with at least one response. In the case of AgeDB, there are 4,071 query images with at least one response for the standard game, and 2,781 query images with at least one response for the game with reject option.

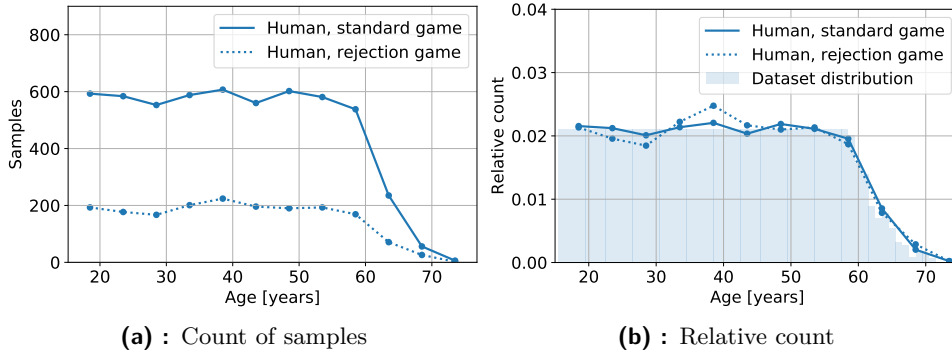


Figure 5.2: Distribution of collected responses for faces from MORPH dataset.

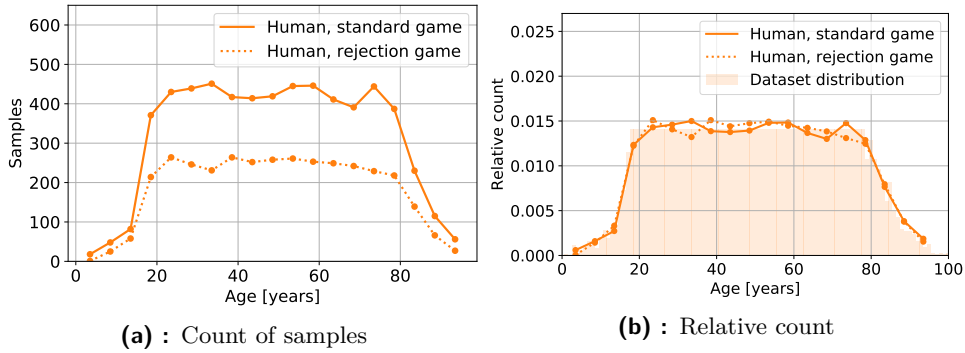


Figure 5.3: Distribution of collected responses for faces from AgeDB dataset.

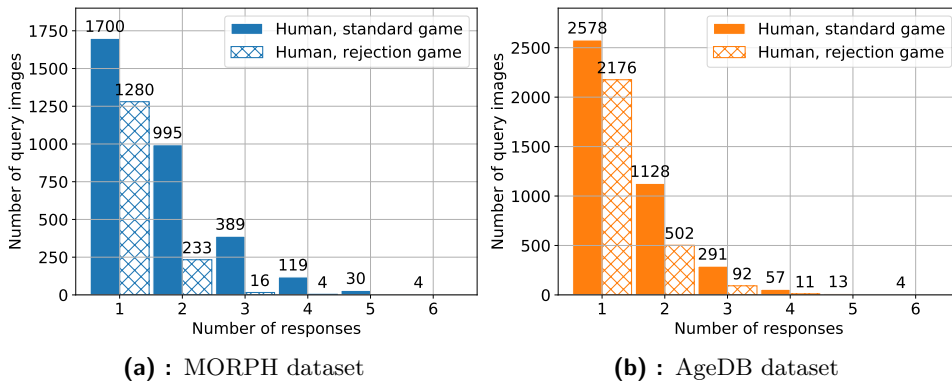


Figure 5.4: The distribution of the number of responses per query image.

Chapter 6

Experiments

In this chapter, we present experiments on the collected data which were described in the previous chapter. Namely, we evaluate the human performance in the age prediction games and compare it with a CNN based predictor. The chapter is organized as follows. In Section 6.1, we describe training of the CNN predictor. The used evaluation metrics are defined in Section 6.2. The observed results and their thorough discussion is subject of Section 6.3.

6.1 Training of the CNN based age predictor

We trained the CNN (AgeNet) age predictor described in Section 3.3.3 separately for MORPH and AgeDB. The test examples used to evaluate the performance of the CNN coincide with the sets used in the Age Prediction games. There is no overlap between the test examples and the examples used for training and validation. In the case of MORPH, we used the remaining 40,293 examples for training and 10,077 examples for validation. In the case of AgeDB, we added to 7,528 training examples from AgeDB another 247,605 examples from IMDB dataset [RTG16]. As the validation set, we used 1,882 examples from AgeDB and 63,480 examples from IMDB. The distribution of the training, validation, and testing examples is shown in Figure 6.1. It is seen that the distribution of the training and testing examples are different. In particular, the very young and very old age categories are not covered in the training set. This implies that the resulting CNN is far from optimal and serves only as a weak baseline.

We used 300 epochs to train the CNN. The best model was obtained based on the MAE computed on the validation set. The training on GPU took approximately 3 hours for MORPH and 7 hours for AgeDB.

6.2 Evaluation metrics

In this section, we define the metrics used to evaluate the performance in age prediction games.

Let Y_i^* be the real (chronological) age, Y_i the guessed (or predicted) age and n the number of testing examples, i.e., the number of faces in the datasets

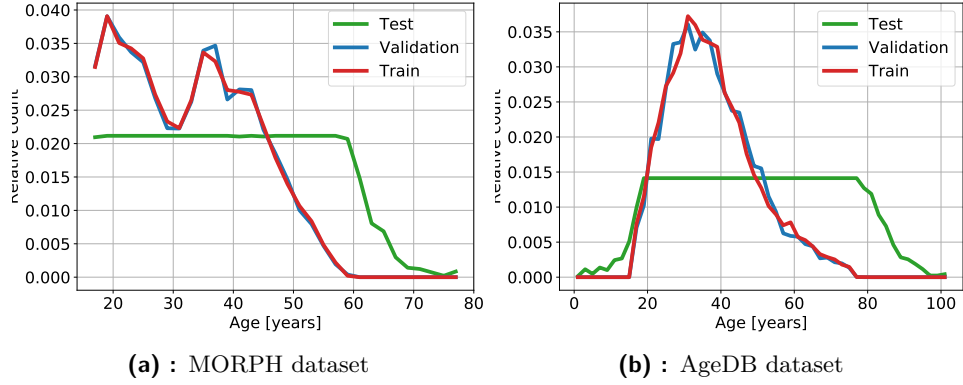


Figure 6.1: The distribution of the training, validation and testing examples used to evaluate performance of the CNN predictor.

used in the games. The mean absolute error is defined as

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - Y_i^*|. \quad (6.1)$$

In addition, we define the mean absolute error for a given age group. Let G be the set of ages in the group, e.g., $G = \{10, \dots, 15\}$. Then, we define

$$MAE_G = \frac{1}{|W_G|} \sum_{i \in W_G} |Y_i - Y_i^*|, \quad W_G = \{i \mid Y_i^* \in G\}. \quad (6.2)$$

Another useful metric is the so-called CS5 score, defined as

$$CS5 = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[|Y_i - Y_i^*| \leq 5], \quad (6.3)$$

which gives the portion of predictions with the absolute deviation not higher than 5 years. Another useful metric is the BIAS, which informs us about the tendency of the player to either underestimate or overestimate the true age. The BIAS is defined as

$$BIAS = \frac{1}{n} \sum_{i=1}^n (Y_i - Y_i^*). \quad (6.4)$$

When evaluating the metrics MAE, CS5, and BIAS, in the reject option setting, we remove the faces for which the guessing was rejected.

6.3 Results

In this section, we compare the performance of a human player and the CNN predictor in the Age Prediction Games of type I standard prediction, and type II prediction with the reject option.

The performance of humans and the CNN is summarized in Table 6.1. Besides the evaluation metrics MAE, CS5, and BIAS, described in Section 6.2,

the table also contains the portion of *Rejects* and the estimate of the expected Risk. The *Rejects* correspond to the portion of queries on which the player rejected to predict, hence, it is provided for the reject option game only. The Risk is computed as the average value of the loss function, the loss ℓ_{AE} in case of the standard prediction game, and the loss ℓ_{AE}^6 in case of the reject option game (c.f. Section 3.2 for the formal definition of the games). The Risk corresponds to the average penalty obtained in the game. If multiplied by 10, it would be the average number of points accumulated in a single game.

Player	Dataset	Type	MAE	[%]		[%]	
				CS5	BIAS	Rejects	Risk
Human	MORPH	standard	6.85	57.2	+0.53	-	6.85
Human	MORPH	reject	6.92	54.7	+0.13	7.02	6.86
Human	AgeDB	standard	9.95	43.5	-0.04	-	9.95
Human	AgeDB	reject	9.15	44.8	+0.46	4.69	9.00
CNN	MORPH	standard	4.20	71.9	-2.31	-	4.20
CNN	MORPH	reject	4.14	72.72	-2.31	2.10	4.18
CNN	AgeDB	standard	8.94	39.7	-3.47	-	8.94
CNN	AgeDB	reject	4.00	50	-4.00	99.97	5.99

Table 6.1: Statistics of prediction in the games

Next, we present the probability distributions of the absolute error of the prediction, i.e., the value of the loss $\ell_{AE}(Y, Y^*) = |Y - Y^*|$. The error distribution for MORPH and AgeDB is shown in Figures 6.2a and 6.3a, respectively. In addition, we show the cumulative histogram of the absolute error in Figures 6.2b and 6.3b. The cumulative histogram shows the portion of predictions which have the absolute error less than or equal to a specific value on the x-axis.

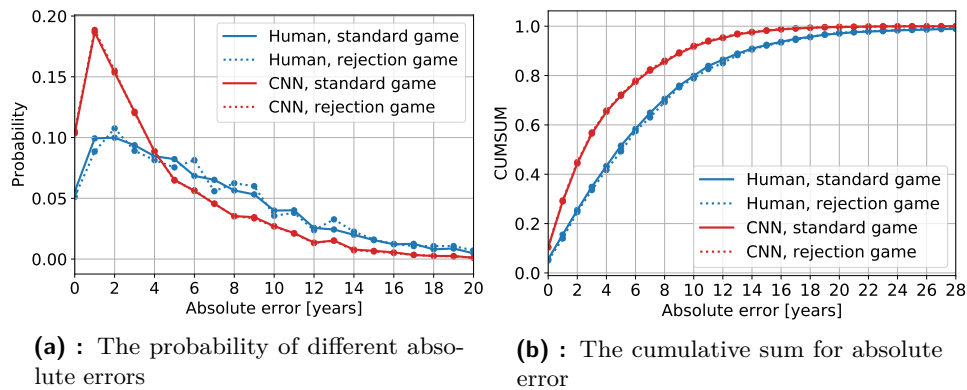


Figure 6.2: Distribution of absolute error for MORPH dataset

In Figures 6.4 and 6.5, we show the the mean absolute error MAE_G as a function of the age category. Note that the very young and very old age categories are less populated in the datasets and hence the results obtained

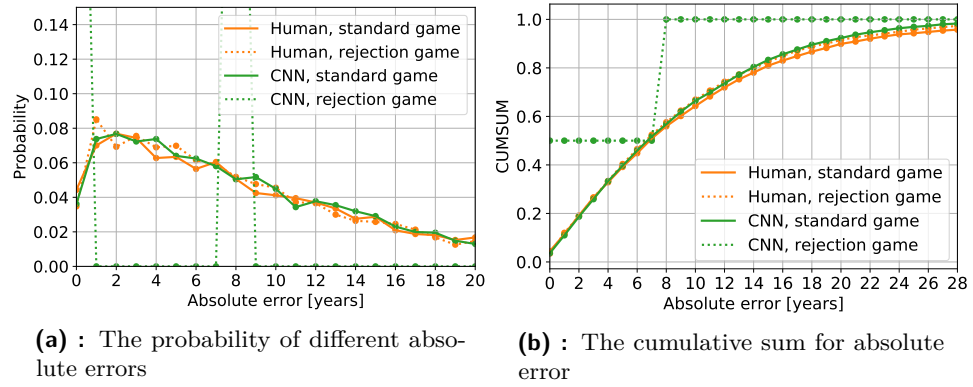


Figure 6.3: Distribution of absolute error for AgeDB dataset

are less reliable. Note that the CNN predictor on AgeDB decided to predict age only in two cases. Hence, in Figure 6.3a we see only two peaks in 0 and 8, while other values of MAEs are not presented at all.

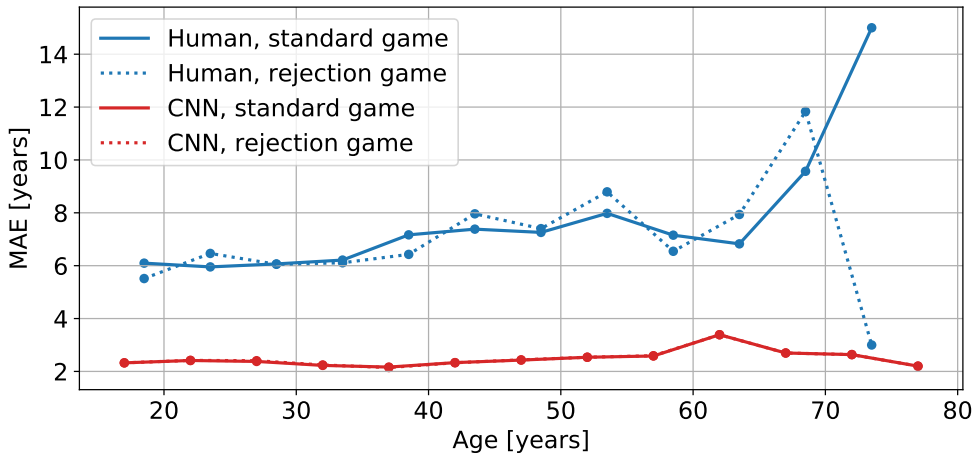


Figure 6.4: Mean absolute error dependence on age for MORPH dataset

Finally, for the games with the reject option, we show in Figure 6.6 the probability of rejection as a function of the age category.

In the remainder of the section, we present a discussion of our main findings:

Performance of humans vs. machines in age prediction. Here we consider the results only in type I game - standard prediction. It is seen that, in the case of MORPH, the CNN is significantly better than humans. In the case of AgeDB, the performance of humans and the CNN is comparable, more precisely the CNN is slightly better, but the difference is statistically not very significant. The performance of the CNN is still surprisingly good taking into account the difference between the training and test distribution of age categories. For example, in the case of MORPH, we see that 50% of predictions have the MAE lower than 3 years.

The expectation was that the mean absolute error will increase with

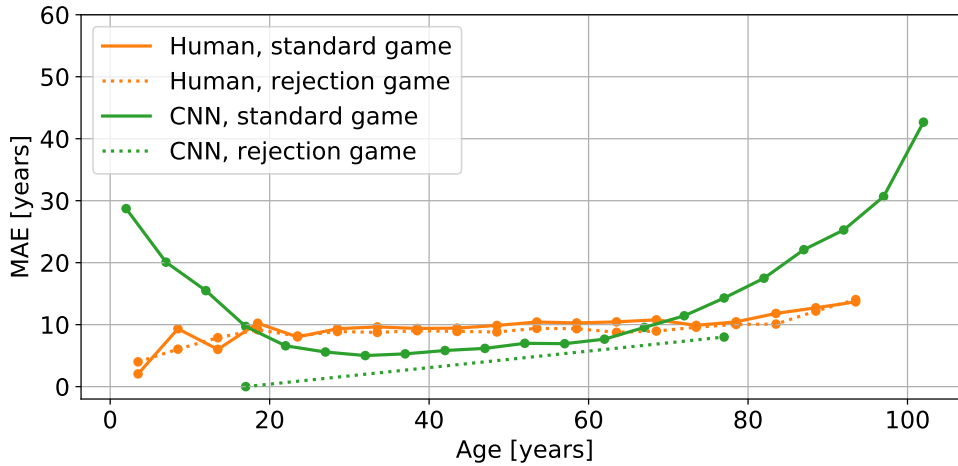
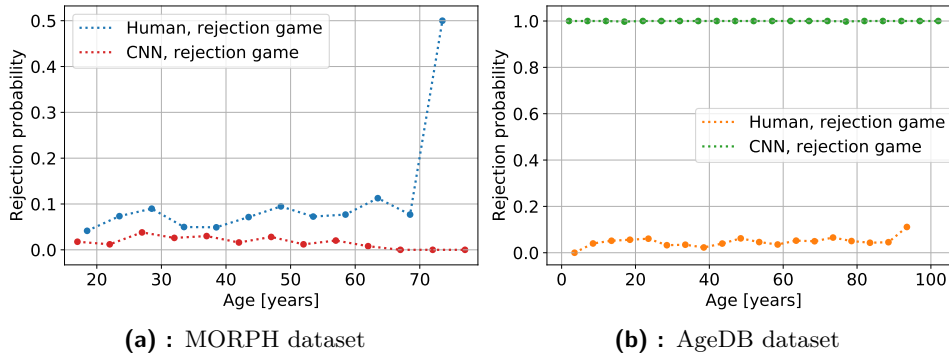


Figure 6.5: Mean absolute error dependence on age for AgeDB dataset



(a) : MORPH dataset

(b) : AgeDB dataset

Figure 6.6: Rejection probability per age category

increasing age. This tendency is clearly observed for human players in all games. The curves for AgeDB shown in Figure 6.5 have a different scale than the same figure for MORPH, and hence the increasing error tendency for humans is not so well visible. In contrast to the expectation, the MAE of the CNN predictor on the MORPH does not have a significant dependence on age. The MAE for the CNN predictor on the AgeDB is the smallest for the age categories around 30, while the performance on young and old categories is very high. As mentioned above, this is due to missing training examples for young/old categories and plenty of examples for middle ages. As a result, the CNN predictor on the AgeDB has better performance than humans only for ages from 20 to 70, while on the MORPH the CNN is consistently better than humans in all age categories.

Performance of humans vs. machines in the reject option games. Here we consider the results only in the prediction game of type II with the reject option. It is seen that except the CNN predictor on AgeDB, the reject option does not show a significant improvement of the performance if compared to the standard prediction as the portion of rejects is very low. In contrast, in the case of the AgeDB, the CNN predictor with the reject option significantly

improves the performance if compared to the standard non-rejection setting.

The reason for the reject option CNN predictor on the MORPH does not provide a significant improvement is the fact that the performance without rejection is already very good enough, the MAE is significantly below 6 for all age categories, hence the predictor has no incentive to reject up to a few cases. On the other hand, the CNN predictor on the AgeDB almost always rejects to predict, namely, it rejects in 99.7% of cases. This is not surprising, because the MAE is for most age categories above 6, hence the rejection is less costly. Recall that the reject option strategy computed from the output of the CNN is far from perfect due to the mentioned problems in the training data. Still, a systematic rejection algorithm, although it is very poor, leads to a significant improvement.

In contrast, humans are unable to exploit the reject option to improve the performance in the game. Note that the MAE for humans is above 6, regardless the dataset, hence a trivial strategy always rejecting the prediction would improve the performance, and in the case of AgeDB, the improvement would be significant. However, on the MORPH dataset, humans rejected only 127 out of 1,810 predictions, i.e., 7.02% responses. In the case of AgeDB, humans rejected 164 predictions out of 3,500, i.e., 4.69% responses. It seems that humans do not have the skill to systematically estimate the prediction confidence, and hence they do not perform well in the reject option setting.

As mentioned above, in the case of humans, the MAE increases with the true age as expected. Hence, our expectation is that the probability of rejection will have the same tendency, i.e., it will increase with age. However, Figure 6.6 shows, that the probability is almost invariant w.r.t. age. The expected tendency is only weakly presented in the data. The same thing was observed for the CNN predictor whose probability of rejection also does not correlate with the MAE. We attribute it to the bat training data on one side. In addition, the CNNs are known to overestimate the posterior probability, which also contributes to the issue. The common remedy is the calibration which, however we did not applied.

Bias of the age prediction. Humans have a very small prediction bias. This is surprising, because the MORPH dataset contains mugshots of Americans with criminal history, whose apparent age seems higher than the real age. We hypothesize that players may have recognized this and adjusted their predictions accordingly.

On the other hand, the CNN predictors underestimate age significantly, i.e., they have a negative bias on both datasets. On average, the CNN predicts the age more than 2 years below the true value. This is caused by the discrepancy between the training and testing distribution of age categories. The training data have the most examples for age categories around 30 years, hence the predictions are biased towards lower ages when predicting from the faces of older subjects. The young categories are much less populated in the test data, hence they do not compensate the bias.

Performance on different datasets. The CNN predictor performs significantly better on the MORPH than on AgeDB. It seems to be caused by the fact that the AgeDB contains "in the wild" images, while the MORPH dataset contains face images captured in a controlled environment, moreover, the subjects were forced to cooperate. On the other hand, the AgeDB contains celebrities who change their appearance, use facial makeup, and their pose varies significantly.

The same behaviour was, not surprisingly, observed for human players. On average, the human MAE is in both games 2.8 years higher on the AgeDB dataset than on the MORPH dataset. The difference is also well seen in Figures 6.2 and 6.3, which show the MAE for individual age categories.

Explanation of bad human performance in reject option games. The small number of rejections in games played by humans is most likely caused by less concentration on the rules of the game. It seems the users most enjoyed being evaluated in the age prediction, and were not sufficiently motivated to minimize the penalty points, although it was the actual goal of the games. This hypothesis can be supported by the observation that the number of games with the reject option, in where the reject option was used at least once is only 23 out of 161 games in the case of the MORPH and 55 out of 311 in the case of AgeDB.

In Figure 6.7, we show the distribution of the prediction errors for the game with the reject option, and a subset of the games with the reject option in which the reject option was used at least once. The intention was to show the results on a subset of games, where the players actually considered the rejection. In the case of the MORPH dataset, we see that the error distribution for the reject games, where the reject was at least once used, shows lower prediction errors. For example, the probability of the 2 year error increased from 0.11 to 0.16. However, in the case of the AgeDB we do not see any significant difference in the standard and reject option games.

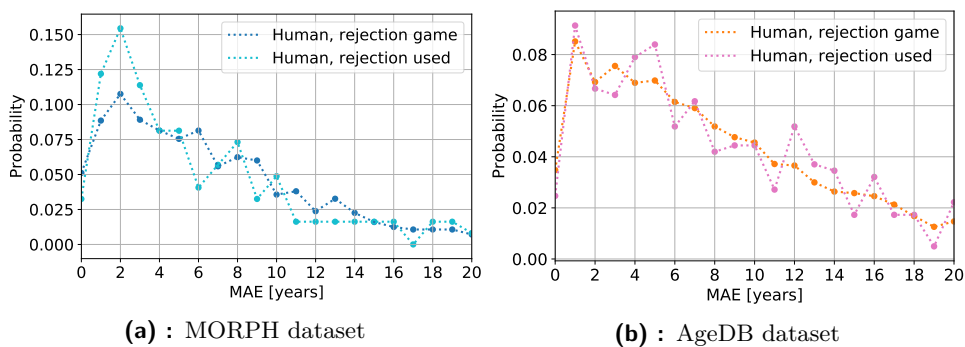



Figure 6.7: Distribution of absolute error for games with reject option. The figures are shown Comparison of all prediction and only predictions in games, where the reject option was used at least once



Chapter 7

Conclusions

In the thesis, we study the human ability to estimate the uncertainty in his/her own predictions. To this end, we have evaluated the human performance in the problem of prediction with the reject option, because the optimal reject option strategy relies on the knowledge of the prediction uncertainty. As the test scenario, we considered the problem of predicting the human age from face images. We have formulated the Age Prediction Games of two types. The game of the first type is a special instance of the standard prediction problem trying to minimize the expected loss, and the performance in this game was used as a reference point. The game of the second type is then a special instance of the cost-based prediction with the reject option. We designed the Age Prediction Games such that they have simple rules understandable for ordinary people without knowledge of statistics. We implemented a web application where the Age Prediction Games can be played online by anyone with the Internet browser. We used the application to collect responses in the two types of prediction games on images from two established age prediction benchmarks, the MORPH and the AgeDB dataset.

We used the collected data to analyze the human performance in age prediction and age prediction with the reject option. We compare the human performance with a CNN based predictor trained on examples. The first finding is that humans are good in age prediction, but even a poorly trained CNN predictor is better. The fact that machines outperform humans in age prediction has been observed before and nowadays it is not very surprising. The second finding is that humans are unable to exploit the reject option to improve the performance in the game. In contrast, the CNN based reject option predictor leads to a significant improvement. Our hypothesis is that the poor performance of humans in the reject option prediction is due to the fact that humans are unable to consistently estimate the prediction uncertainty. To our knowledge, the human performance in the reject option setting has not been studied before.

Our study has two limitations. First, the poor human performance in the reject option prediction may be caused by the player's lack of motivation to do the best in the game. The players may have just enjoyed the age estimation for which they obtain a feedback, instead of trying to minimize the penalty points in the game. This issue could be fixed by increasing the motivation of

the players to minimize the penalty points, e.g., by introducing a financial reward. Another way could be to define a different prediction game, where the user has to provide the confidence of the prediction every time, however, the question is how to define the concept of confidence rigorously. Second, the CNN predictor used as a baseline is trained on examples with significantly different distribution of age categories than has the test data. This deficiency can be fixed by using better training data, or techniques which compensate for the shift in the distribution.

Besides the provided analysis, the thesis has two additional outputs. First, the collected data itself which contain a lot of information about the human performance in age prediction, and which have not been fully exploited in this thesis. Second, a framework for implementation online prediction games which can be used by others to conduct various human studies with a minimal need to code.



Bibliography

- [ABBD16] G. Antipov, M. Baccouche, S.A. Berrani, and J.L. Duglay, *Apparent age estimation from face images combining general and children-specialized deep learning models*, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016.
- [ASJ16] G. Antipov, S.A.Berrani, and J.L.Dugelay, *Minimalistic cnn-based ensemble model for gender prediction from face images*, Pattern Recognition Letters **70** (2016), 59–65.
- [ATE⁺17] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe, *Apparent and real age estimation in still images with deep residual regressors on appa-real database.*, 12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017.
- [CCH14] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu, *Cross-age reference coding for age-invariant face recognition and retrieval*, Proceedings of the European Conference on Computer Vision (ECCV), 2014.
- [Cho70] C. Chow, *On optimum recognition error and reject tradeoff*, IEEE Transactions on Information Theory **16** (1970), no. 1, 41–46.
- [EEH14] Eran Eidinger, Roe Enbar, and Tal Hassner, *Age and gender estimation of unfiltered faces*, Transactions on Information Forensics and Security (IEEE-TIFS), special issue on Facial Biometrics in the Wild (2014).
- [ETB⁺16] S. Escalera, M. Torres, B.Martinez, X. Bar, H.J. Escalante, I.Guyon, M.Oliu, and M.A.Bagheri, *Chalern looking at people and faces of the world: Face analysis workshop and challenge*, In IEEE CVPR Workshops, 2016.
- [FC18] Vojtech Franc and Jan Cech, *Learning cnns from weakly annotated facial images*, Image and Vision Computing (2018).

- [GANT16] G.Panis, A.Lanitis, N.Tsapatsoulis, and T.F.Cootes, *Overview of research on facial ageing using the fg-net ageing database*, IET Biometrics **5** (2016), no. 2.
- [GC09] A. Gallagher and T. Chen, *Understanding images of groups of people*, Proc. CVPR, 2009.
- [HOJ13] Hu Han, Charles Otto, and Anil K. Jain, *Age estimation from face images: Human vs. machine performance*, International Conference on Biometrics (ICB), 2013.
- [JNV⁺19] J.A.B. Jones, U.W. Nash, J. Vieillefont, K. Christensen, and U.K. Misevic, D. Steiner, *The ageguess database, an open online resource on chronological and perceived ages of people aged 5–100*, Scientific Data **6** (2019), no. 1.
- [KB17] Diederik P. Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, 2017.
- [LBM17] S. Lopuschkin, A. Binder, and K.-R. Muller, *Understanding and comparing deep neural networks for age and gender classification*, Proceedings of the ICCV’17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG), 2017.
- [MPS⁺17] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, *Agedb: the first manually collected, in-the-wild age database*, Proceedings of IEEE Int’l Conf. on Computer Vision and Pattern Recognition (CVPR-W 2017) (Honolulu, Hawaii), June 2017.
- [NZW⁺16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, *Ordinal regression with multiple output cnn for age estimation*, In proc of CVPR, 2016.
- [PHSC18] Hongyu Pan, Hu Han, Shiguan Shan, and Xilin Chen, *Mean-variance loss for deep age estimation from a face*, Proceedings of CVPR, 2018.
- [RT06] Karl Ricanek and Tamirat Tesafaye, *Morph: A longitudinal image database of normal adult age-progression*, IEEE 7th International Conference on Automatic Face and Gesture Recognition (Southampton, UK), April 2006, pp. 341–345.
- [RTG16] R. Rothe, R. Timofte, and L.V. Gool, *Deep expectation of real and apparent age from a single image without facial landmarks*, Int. J. Comput. Vis (2016).
- [SCD⁺17] S.Chen, C.Zhang, M. Dong, J. Le, and M. Rao, *Using ranking-cnn for age estimation*, In proc. of CVPR, 2017.

- [SH02] M.I. Schlesinger and V. Hlaváč, *Ten lectures on statistical and structural pattern recognition*, Kluwer Academic Publishers, 2002.
- [SM04] J. Sochman and J. Malas, *Adaboost with totally corrective updates for fast face detection*, Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., 2004, pp. 445–450.
- [SMP⁺13] Setty Shankar, Husain Moula, Beham Parisa, Gudavalli Jyothi, Kandasamy Menaka, Vaddi Radhesya, Hemadri Vidyagouri, J. C. Karure, Raju Raja, Rajan, Kumar Vijay, and C. V. Jawahar, *Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations*, National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Dec 2013.
- [Tor00] F. Tortorella, *An optimal reject rule for binary classifiers*, Advances in Pattern Recognition, Lecture Notes in Computer Science, vol. 1876, Springer, 2000.
- [ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi, *Age progression/regression by conditional adversarial autoencoder*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.