



# **Behaviour-based detection of Visual Interaction Obstacles with 1D and 2D Convolutional Neural Networks**

Bachelor Thesis at the Cognitive Systems Lab  
Prof. Dr.-Ing. Tanja Schultz  
Faculty 3: Mathematics and Computer Science  
University of Bremen

by

**Anthony Mendil**

Supervisor:

Mazen Salouz

Examiner:

Felix Putze

Unknown

Day of registration: 1. Mustermanat 1851

Day of submission: 3. Mustermanat 1963



---

I hereby declare that I am writing this work independently and have not used any sources or resources other than those specified.

Bremen, the 3. Mustermonat 1963



## **Zusammenfassung**

... deutsch ...

Der deutsche Abstract wird in jedem Fall benötigt.



## **Abstract**

... english ...

Der englische Abstract wird nur benötigt, wenn die Arbeit in englischer Sprache verfasst wird.





# Contents

<b>1</b>	<b>Notes</b>	<b>1</b>
1.1	Anmerkungen . . . . .	3
<b>2</b>	<b>Introduction</b>	<b>5</b>
<b>3</b>	<b>Memory Game</b>	<b>7</b>
<b>4</b>	<b>CMM-based Cognitive User Simulation</b>	<b>9</b>
<b>5</b>	<b>Collected Data</b>	<b>11</b>
5.1	Behavioural data . . . . .	11
5.2	Brain activity . . . . .	13
<b>6</b>	<b>Data Preparation</b>	<b>15</b>
6.1	Similarity matrix creation . . . . .	15
6.1.1	Conceptualization . . . . .	15
6.1.2	Screenshot extraction . . . . .	19
6.1.3	Implementation . . . . .	20
6.1.4	Testing . . . . .	21
6.1.5	Final Matrix . . . . .	21
6.2	Adaptation and correction of logs . . . . .	23
6.3	Removal of invalid logs . . . . .	23
6.4	Simulation of user behaviour . . . . .	24
6.5	Sorting logs by quality . . . . .	24
6.6	Evaluation of simulation . . . . .	24
6.7	Feature generation . . . . .	30
6.7.1	1D CNN features . . . . .	30
6.7.2	2D CNN features . . . . .	30
<b>7</b>	<b>Convolutional neural networks</b>	<b>35</b>
<b>8</b>	<b>Classification</b>	<b>37</b>
8.1	Hardware resources . . . . .	37
8.2	Models . . . . .	37
8.2.1	1D CNN . . . . .	38
8.2.2	2D CNN . . . . .	39
8.3	Training setup . . . . .	40
8.4	Training results and evaluation . . . . .	41
<b>9</b>	<b>Voting based system</b>	<b>45</b>

<b>10 Conclusion</b>	<b>47</b>
10.1 Prospect . . . . .	47
<b>11 Appendix</b>	<b>49</b>
<b>Bibliography</b>	<b>51</b>
<b>Index</b>	<b>53</b>

# List of Figures

1.1	Bild kurz . . . . .	4
3.1	Bild kurz . . . . .	8
3.2	Bild kurz . . . . .	8
6.1	Bild kurz . . . . .	16
6.2	Bild kurz . . . . .	17
6.3	Bild kurz . . . . .	22
6.4	Bild kurz . . . . .	25
6.5	Bild kurz . . . . .	25
6.6	Bild kurz . . . . .	26
6.7	Bild kurz . . . . .	26
6.8	Bild kurz . . . . .	28
6.9	Bild kurz . . . . .	28
6.10	Bild kurz . . . . .	31
6.11	Bild kurz . . . . .	31
6.12	Bild kurz . . . . .	32
6.13	Bild kurz . . . . .	33
8.1	Bild kurz . . . . .	38
8.2	Bild kurz . . . . .	39



# List of Tables

1.1	Tabelle mit kurzer Unterschrift . . . . .	3
6.1	add caption. . . . .	22
6.2	p values in paired t-test for different comparissons of matching pairs per round. All 1000 simulated games per real game were used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . .	27
6.3	p values in paired t-test for different comparissons of penalties per round. All 1000 simulated games per real game were used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . .	27
6.4	p values in paired t-test for different comparissons of matching pairs per round. All 1000 simulated games per real game were used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . .	27
6.5	p values in paired t-test for different comparissons of penalties per round. All 1000 simulated games per real game were used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . .	27
6.6	p values in paired t-test for different comparissons of matching pairs per round. For the simualted data only the best 10 simulated logs from each real log are used. Only the first 10 rounds are used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . . .	28
6.7	p values in paired t-test for different comparissons of penalties per round. For the simualted data only the best 10 simulated logs from each real log are used. Only the first 10 rounds are used. The following abreviatrions are used: real glare effect - r_g, real no obstacle - r_n, simulated glare effect - s_g, simulated no obstacle - s_n . . . . .	28
6.8	mean values of matching pairs for real and simulated games. sd10x. for the first 10 rounds . . . . .	29
6.9	add caption. Upper and lower boundaries are inclusive. . . . .	32
8.1	add caption . . . . .	37
8.2	add caption . . . . .	37
8.3	1D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps. Before the changes to the simulator. . . . .	41
8.4	2D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps. Before the changes to the simulator. . . . .	42

8.5	1D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps. After the changes to the simulator. . . . .	42
8.6	2D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps. After the changes to the simulator. . . . .	42

# 1. Notes

- fragen an mazen:
- demographic
- original mapping of cards to colours
- nachfragen ob static mapping so ist wie ich denke
- sind whickser nur standardabweichung oder was ist das?
- ist es ok dass ich meine eigene tabelle gemacht habe für delta e interpretation? im interent stand es gibt keine allgemeine..hab auch keine von cielab gefunden
- ich brauche mll knecht

- Einleitung: Was es für verschiedene Obstacles gibt, was schon gemacht wurde, was ich mache, wieso keine eeg daten obwohl sie da sind (sind interaktion obstacle zu erkennen und nutzung zu verbessern ist nicht da wenn man eeg maske tragen muss)
- memory game kurz erklären
- libraries und so die ich benutz habe? pandas keras..

- **TODO: vizualizing of intermediate activations reinnehmen? (ist auskommentiert!)**

- **simulation:**

- simulatin damit man mehr daten hat
- generell simulator erklären und sinn von similarity matrix
- similaity matrix erstellung und gedanken (plus anpassungen an memory game, anpassungen am simulator damit similarity matrix benutzt werden kann)
- es gab invalid logs und hab code geschrieben der das überprüft damit man die rausnehmen kann. vor simulation (validLogsCollecot)
- similarity matrix mapper, waren vorher nicht gmapped und mapper macht das und ersetzt alte matrix durch die neue für glare effect
- erklären wie ich korrektheit der matrix überprüft habe
- erklären wie ich korrektheit der farbextraktion und unterschied berechnung überprüft habe (online rechner und anfangs matrix für no obst erstellt und mit alten verglichen aber gibg nicht weil struktur nicht zu erkennen war bei alter matrxx von vorherigen arbeiten)
- initially simulationergebnisse mit plots für qualität (plus erklärung)
- sagen was noch nicht optimal und, was am simulator dafür geändert wurde (2 sachen: random decay ab 10 zügen plus eine andere sache) mit begründung und wie die ergebnisse am ende aussahen (zwischenschritte eher niht glaube ich. nur kurz erwähnen)
- darüber reden wie es im realfall ist: wir benutzen nicht alle simualtionen sondern nur die besten, plots zeigen mi nur den besten (vor und nach änderung vielleicht, oder nur nach änderung (aber dann sieht man nicht das änderung sinnvoll war))
- paried t-test kram um signifikante unterschiede /keine unterschiede zu zeigen für verschiedene bedingenen (siehe mazens nachricht)

- **modelle:**

- feature engeniering: also was die komponenten sind und so, wie sie berechnet wurden
- für 1d cnn wuren die so übernommen
- erklärung komponenten von 1d cnn
- (villeicht auch mal nicht mit allen featires probieren, aber da hatte ich problme mit dem cnn. man müsste glaube ich die struktur ändern)
- struktur 1 d cnn mit begründung
- 2d cnn komponenten erklärunen (snythetic image erklären und zeigen wie berechnet wurde und wie es aussieht)



dies	ist	eine	Tabelle
mit		zwei	Zeilen

Table 1.1: Tabelle mit einer langen Unterschrift

- idee von 2d cnn für diesen fall
- 2d cnn struktur erklären und begründen
- **Training und analyse:**
  - vielleicht gucken welche falsch erkannt werden und woran es liegt, also wenn die zum beispiel echt schlecht oder gut sind obwohl es nicht so sein sollte (rausnehmen und gucken wie ergebnisse sind, vielleicht nur bei bestem modell)
  - in gleichen tabellen ergebnisse vor änderung am simulator und nach änderung betrachten und vergleichen
  - training auf welchem rechner/n, was von: cpu oder gpu oder beides, hardware kurz erwähnen, vpn (fernzugriff)
  - train test splt, leave one out k fold (+ begründung mit deep learning ..reliable results etc, randomness)
  - kurz erklären wieso weniger züge besser sind bei dieser hci erkennung
  - (vielleicht kram zu adaptive learning rate ändern und gucken wie es so ist, ansonsten begründen wieso ich das nicht brauche)
  - 20 steps und 40 steps für beide trainieren mit sd0x bis sd10x plus sd20x
  - (entweder so dass 1d cnn mit 20 steps auch konvergence erreicht oder danach nochmal anpassung von 20 steps 1d cnn damit es konvergiert)
  - darüber schreiben dass letzten n züge nicht mehr so gut simuliert sind (zeigen) und deshalb 40 steps nicht so viel besser ist
  - mit maximalen guten steps (glaube 16 züge oder so) trainieren und vergleichen
  - statistical tests um signifikante unterschiede /keine unterschiede zu zeigen für verschiedene modelle etc (siehe mazens nachricht)
- in anhang alle skripte aufzählen wie markus
- **TODO: tabellen unterschift muss glaube ich dadrunter und nicht drüber!**

## 1.1 Anmerkungen

Zitationen [Rab89] sind keine Wörter sondern Referenzen und stellen somit keinen Teil des Satzes dar. In anderen Worten: Der Satz muss auch noch funktionieren, wenn die Zitation einfach entfernt wird.

Index-Einträge

Wir haben Tabellen 1.1 und Bilder 1.1.



Figure 1.1: Bildunterschrift

## 2. Introduction

- papers von denen lesen. da sind richtig gute einleitungen an denen man sich orientieren kann.

- 

- relevance of topic - glare detection has been done before with for example a light detector but not with behavioural data (volatile?) -

Neural networks greatly benefit from lots of data. As there are only 40 **TODO: richtige zahl herausfinden** overall games from participants available, more data can potentially improve the classification results. The cognitive systems lab already implemented a system that takes the original game logs and simulates user behaviour in order to create new logs. As a result it is possible to create any number of games out of a single original one, from which some may be better than others. However, in order to simulate glare effect games multiple additional steps and changes are necessary.

- general approach, simulation, matrix creation, often used 1d cnn, novel approach of creating synthetic images and using 2d cnns shows in many cases improved classification results. Finally four voting based systems one each for 5, 10, 15 and 20 rounds are implemented, combining the prediction of 400 models each to decide the final predictions.

- vielleicht begründung: referenz zu denen. weil es zeigt dass man unter Verwendung simulierter spiele eine deutliche Verbesserung der ergebnisse auf den echten spielen sehen kann...



### 3. Memory Game

The game used for the data collection is a matching pairs game written in Kotlin for Android devices. The original version was developed by Rafael Miranda **TODO: ref** in the context of a bachelor thesis. Originally, the game consisted of cards picturing animals and was only used to model general poor eyesight. Later it was expanded by adding a coloured card set to also model colour blindness. In the following the current functionalities of the game will be explained, only covering those related to the coloured cards. One game consists of 14 cards and each round consists of two cards being turned face up. If the cards match, they will be removed from the field. Otherwise they are turned face down again. The player wins once all 7 pairs have been discovered.

The game differentiates between two groups of obstacles: memory obstacles and visual obstacles. Additionally there is a classic mode providing the possibility to play without any obstacles which uses the red-green card set, shown in fig **TODO: ref**. The memory obstacle mode consists of the same card set and an additional side task: Every time a card is flipped a random number between 0 and 10 is called out and the player must add all these numbers. After the final pair was found the sum of all numbers is requested. The game contains modes for two types of visual obstacles: Colour blindness and the glare effect. Colour blindness, more precisely a red-green weakness, is simulated by replacing the red-green card set with a brown shifted card set. The glare effect describes a scenario in which a light source shines onto the display, reflects from it and makes it more difficult to differentiate the colours of the cards. This effect can be seen in figure **TODO: ref**.

To help in case of interaction obstacles the game provides multiple ways of assistance. Although they are not relevant for this thesis as only data is used where probands had no assistance, they will be explained briefly. There are two types of assistance: memory and visual assistance. Memory assistance is provided by flipping the cards from the previous turn once two non matching cards are flipped. After a short period of time they are turned face down again. If the game is played with visual assistance, each pair gets assigned a letter that is called out once a card is flipped. This introduces a new way of orientation that does not rely on the sense of sight, but instead on the sense of hearing. There is a game mode for each combination of obstacle

(no obstacle, memory obstacle, colour blindness, glare effect) and assistance (no assistance, memory assistance, visual assistance). There were also other assistances implemented such as increasing the size of cards if they are revealed, but they were only used for tests. The glare effect no assistance and the no obstacle no assistance modes are the only ones relevant in this thesis. Figure **TODO: ref** and **TODO: ref** show screenshots in the two modes.

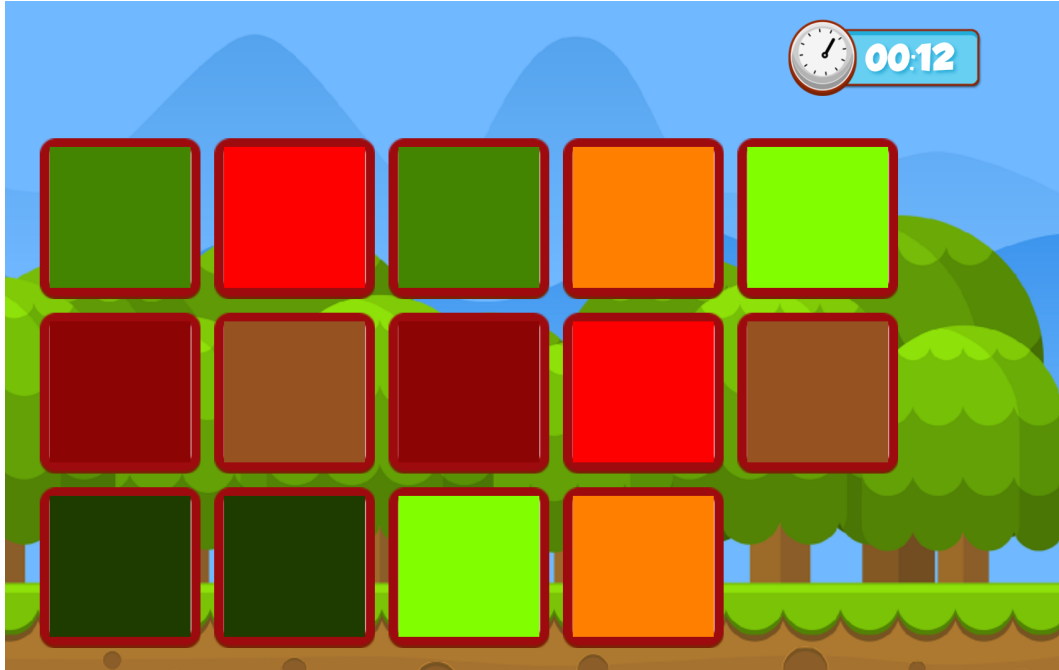


Figure 3.1: Screenshot of the game without obstacles. All cards are turned face up.



Figure 3.2: Screenshot of the game field with simulated sunlight. Sunbeams are especially visible on the right hand side. All cards are turned face up.

## 4. CMM-based Cognitive User Simulation

The cognitive system lab developed the computer program cognitive memory model (CMM). It utilizes concepts of the act-r-theory to model the cognitive human memory. ACT-R is a cognitive architecture that aims to describe and integrate central basic mechanisms of human cognition. **TODO: ref:** <https://www.google.com/search?q=intimates+deutsch&oq=imimates+&aqs=chrome.1.69i57j0l7.2543j0j15&sourceid=chrome&ie=UTF-8> In the context of the matching pairs game the CMM was used to implement a generative model. By modelling memorized and forgotten revealed cards and deciding in each turn whether to explore unknown cards or to exploit the gathered knowledge to reveal matching pairs, the course of a matching pairs game can be simulated taking the capability of human memory into account. Furthermore it is possible to define a similarity matrix to simulate similarities between cards in matching pairs games. This matrix includes similarity values between 0 and 1 for each colour combination of cards. Such a similarity matrix can be used in the context of visual interaction obstacles to emulate human confusion. In related research it was used to simulate the confusion caused by colour blindness and the usage of a red-green card set. For this thesis the only two modes that are relevant are the glare effect (no assistance) and the no obstacle (no assistance) modes. In order to simulate no obstacle games no similarity matrix is needed, but for glare effect games a new similarity matrix is created in section **TODO: ref.**

To accurately simulate the performance of human memory multiple parameters are randomly initialized and then repeatedly optimized using a genetic optimization algorithm. These parameters can be further extended. The genetic algorithm works by repeatedly simulating game sessions and updating the parameters by applying mutation and selection operations. To select the best parameter values two performance measurements are defined and used to compare the simulated game sessions and a real game as reference. These performance measurements are the number of matching pairs and the number of penalties per round. Penalties are given if a card is revealed whose partner has already been seen before but the pair was not picked up. The optimization of the parameter population is repeated over many generations so that the performance in the simulated games best fits that in the real game. As

example one of the parameters optimized by the genetic optimization algorithm is the similarity decay. It reduces the similarity effects during the course of the game, since the user may learn to adapt to the visual interaction obstacle and therefore be less effected by it. Furthermore the similarity effect is continually reduced as there are less cards in the game after pairs were discovered.

As input the simulator takes a number of game logs and simulated a specified amount of new sessions using the genetic optimization algorithm. For instance 20 game logs can be passed to simulate 1000 new sessions out of each game log, resulting in 20000 simulated games. The data contained in the game logs is shown in section **TODO: ref**. In the context of simulating no obstacle games changes to the simulator were made that are explained in section **TODO: ref**. In section **TODO: ref** can be seen whether these changes resulted in improvements of the actual classification result.



## 5. Collected Data

While the probands played the memory game, behavioral data and brain activity data was collected. Relevant for this work is only the behavioural data recorded in glare effect and no obstacle games. The data was collected from 22 probands. **TODO: grobe demographic herausfinden, mazen fragen.** Each participant except one recorded 2 no obstacle and 1 glare effect game. The one exception did not record any glare effect games. As a result there are 44 no obstacle and 21 glare effect games. The data was already provided and was not collected during this bachelor thesis.

### 5.1 Behavioural data

In order to record behavioural data, each card was initially given a fixed card code and it was recorded which pairs of cards were turned face up in each round. The card code consists of two numbers separated by a dot. The first number specifies the colour of the card (between 1 and 7 as there are 7 colours) and second number specifies if it is the first or second card with that color in the order of the cards on the field **TODO: nachfragen ob das so ist, aber müsste weil am anfang zum beispiel 5.2.** The behavioral data was saved in logs, consisting of the card codes of each round followed by the unix timestamps of when the cards were flipped. Data of 20 rounds and therefore at maximum 40 flipped cards was recorded. If the game was completed in less than 20 turns, the remaining card code entries were filled with 0.0 and the remaining timestamp entries were filled with 0. An example of the sequence of card codes recorded during a game can be seen below. The timestamps are irrelevant for this work and therefore not shown.

```
5.2,2.2,4.2,4.1,6.1,7.2,6.2,6.1,1.2,1.1,  
7.1,2.2,2.1,7.1,5.2,5.1,3.1,2.1,7.1,3.1,  
7.2,7.1,3.2,2.2,2.1,2.2,3.1,3.2,0.0,0.0,  
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
```

**TODO: gucken ob in original logs wirklich nichts anderes drin ist, also die 4 andeen statistical features TODO: initial mapping herausfinden** With

the assignments of for instance of 5 for red and 2 for green, the first two entries in the log file mean that first the second red card and then the first green card was turned face up. This mapping is static as colours have the same numbers across all recorded games. However, the simulator requires a dynamic mapping of the card codes. Additionally each of the dynamically mapped card codes has to receives a similarity value for the simulation mentioned in chapter **TODO: ref**. Each card codes describes the two cards that were flipped in a turn and the similarity value describes how similar the colours of the two cards are. These similarity assignments for each combination of colours needed to be added to the logs. These assignments had to be made according to the dynamic mapping of card codes.

By dynamic mapping of the card codes is meant, that the colours are not assigned to fixed numbers but that the numbers are assigned in the reveal order specific to each game. This becomes clear, by looking the card code sequence from above, but dynamically mapped.

```
1.1,2.1,3.1,3.2,4.1,5.1,4.2,4.1,6.1,6.2,
5.2,2.1,2.2,5.2,1.1,1.2,7.1,2.2,5.2,7.1,
5.1,5.2,7.2,2.1,2.2,2.1,7.1,7.2,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
```

**TODO: sobald ich originales mapping weiß, weiß ich auch welche farben die karten hatten und kann die folgenden farben im text ersetzen** In this dynamic mapping of the card codes, each entry likewise consists of 2 numbers that are seperated by a dot. However, the numbers are differently interpreted. The first number specifies what number of colour it is to be revealed and the second number specifies if it was the first or the second card in that colour. In the example above the colour green is the first to be revealed and it was the first green crad. Therefore the first entry is 1.1. Then a red card is turned face up, meaning the second entry is 2.1 since is was the first red card and so on. Once the other red card is dicovered, the entry will be 2.2. This pattern continues throughout the whole sequence. This means that the mapping is specific to each game and depends on the reveal order of the cards.

Logs with remapped card codes and added similarity values were already provided, but there were two changes regarding the glare effect logs that had to be made before they could be used. The first one being that the similarity matrix for the glare effect game was just a placehoulder and did not correclty portray the color differneces under the influence of sunlight. To successfully simulate glare effect games the simulator requires such an similarity matrix. Therefore a new one is created in section ref. The other neccessary change derived from the fact that although the card codes were dynamically mapped, the similarity assignments were made according to the static mapping of the card codes inestead of the dynamic mapping. This meant that the assignment of simialrity values to numbers for the cards compared were identical in all games, but the numbers of the cards stood for different colours in different games. To solve this issue the newly created similarity matrix for the glare effect was used to determine the dynamically mapped similarity values for each game. The old values in the glare effect logs are then replaced with the new values and correct mapping in section ref. The reason why these issues only apply to glare effect logs and not to no obstacle logs is that the simulator does not use any similarity values when simulating

no obstacle games, meaning that these values are ignored anyway. Therefore there is no need to replace them. It should also be noted that both the glare effect and the no obstacle logs additionally include four statistical features for the last turn. These consist of the number of remaining cards, the number of never revealed cards, the highest number of times the same card was revealed and the number of rounds since all pairs were found. These four values are also ignored, as they are newly calculated for every turn in section ref. Last but not least all logs end with a label, describing in which game mode the log was created.

## 5.2 Brain activity

Eeg data collected from all participants during the game. However, there is one problem with using the eeg data: The overarching context of this work is to find interaction obstacles and ultimately improve the user interaction. This means that the way of discovering such an interaction obstacle should not worsen the interaction experience. If all people had to wear eeg masks when interacting with software just for the purpose of noticing interaction obstacles, the interaction experience itself would suffer. A method of recording brain activity without a deterioration of the interaction experience has not been discovered yet. Additionally it is not cost efficient for every user to use an eeg sensor. As a result the decision was made not to use eeg data and instead only use the behavioural data that is collected directly through the interaction and stays unnoticed by the user. As eeg data is not used in this work it will not be further explained.



## 6. Data Preparation

To prepare the data for the training many steps were necessary. One big step is to increase the available data by simulating user behaviour. Therefore the simulator from chapter **TODO: ref** is used. For simulating games with no obstacle there is no need for a similarity matrix. However, in order to successfully simulate glare effect games, a special similarity matrix is required. While without any obstacles the color differences are independent from the position of the cards, this is not the case in glare effect games since the intensity of the light is not the same across the whole field. As a result the first step is to create a similarity matrix that describes the differences of colours under the influence of the glare effect. Once completed, it replaces the old similarity matrix in the glare effect logs. Additionally all logs need to be remapped since they were mapped statically instead of dynamically. As not all logs are valid and can be used in the simulator, the invalid log must be removed. These valid logs are then used to simulate user behaviour and thereby create new logs. These simulated logs are sorted by their quality and the simulation is evaluated. If the performance in the simulated games is similar to that in the original games it is preceded with the generation of the features for the training. Otherwise changes to the simulator are made and the simulation, the sorting and the evaluation is repeated. In case of very bad simulations of glare effect logs it would have also been possible to change the approach of creating the similarity matrix or find a new approach. However, the simulations of glare effect were very good, which can be seen in **TODO: ref**, which is why the approach was kept.

### 6.1 Similarity matrix creation

#### 6.1.1 Conceptualization

The aim is to create a similarity matrix that describes the differences between colours under the influence of the simulated sun light. The difficulty hereby is that the intensity of the light is not spread evenly across the whole field. Instead, as it would be if a real sun would shine onto the display, a certain area has the highest intensity and the intensity is lower the further away from that area. Additionally there are wide lines of higher intensity coming from the brightest area, that simulate sunbeams. This influence of the simulated sunlight can be seen in figure 6.1.



Figure 6.1: Screenshot of the game field with simulated sunlight. Sunbeams are especially visible on the right hand side. All cards are turned face up.

**TODO: vielleicht bild raus nehmen weil es oben schon einmal ist und referenzen nach oben**

As a result simply extracting the rgb values for one pixel of each card in one game and calculating the differences has two major problems: Firstly, the differences are highly influenced by the position of the specific cards. If the matrix is calculated using a single glare effect game the differences are only representative for exactly that game and may be completely different if the cards are positioned differently. Secondly, extracting single pixels for each card in an image could lead to color values that do not represent the overall observed colour due to varying brightnesses on different points of a card. This behaviour is undesired, as the final similarity matrix is supposed to represent the differences of the observed colours of the cards. **TODO: vielleicht quelle suchen die sagt das menschen farben im ion bereichen mitteln oder so** In order to solve these problems a more complex approach than simply extracting single pixels out of a game was needed.

The problem of the positional influence can be solved by creating similarity matrices for more than one game and calculating the mean color differences of those matrices. The idea is that by creating using so many games that all or most combinations of positions and colours are included and calculating the mean of all comparisons for , the result will not be influenced by the position of the cards. However, first needs to be determined which number of games satisfies this condition. Each game consists of 14 cards, with each two cards having the same colours. When determining the difference between two colours there are two different cases:

- 1. The colours are not the same: In each game there are exactly 4 combinations of positions for those two colours, because there are two cards for each colour.

- 2. The colours are the same: In each game there is only one combination of positions for those colours.

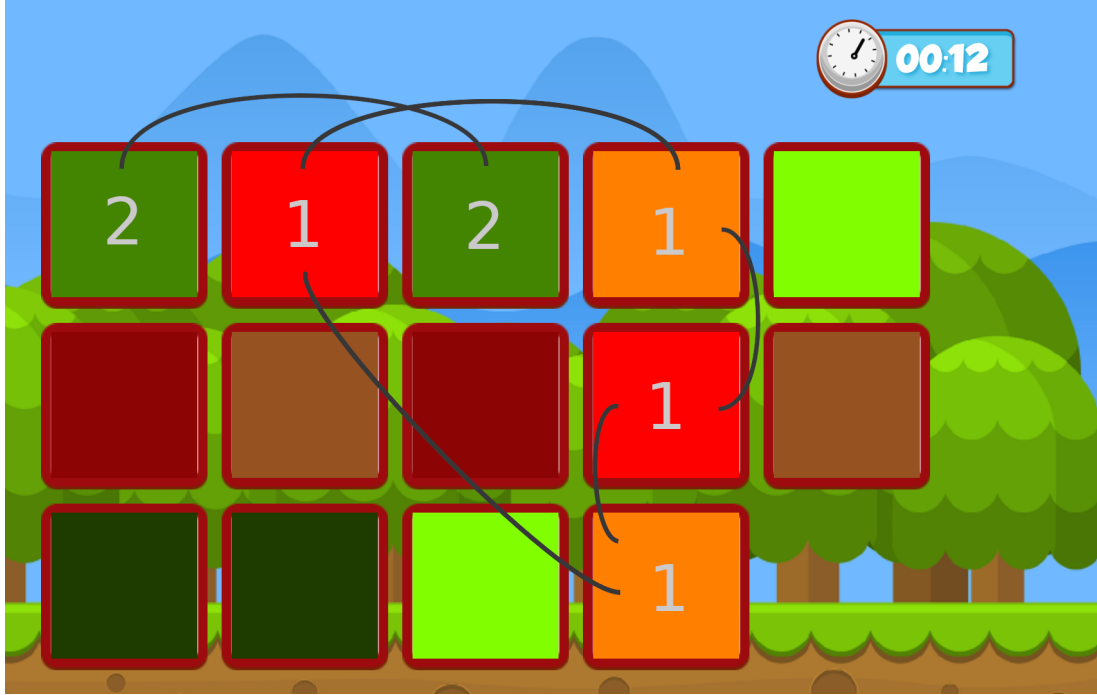


Figure 6.2: Exemplary showing the number of different comparisons for case 1 and 2. The numbers on the cards represent the case and the edges indicate unique comparisons. All cards are turned face up. A game without any obstacles is used only for the purpose of clarifying the different comparisons. All screenshots used for the actual calculation are taken from glare effect games.

First it is calculated how many screenshots are necessary for the first case. For two different coloured arbitrary but fixed cards there can be

$$C_1 = 14 \cdot 13 = 182$$

combinations of positions are possible. The reason that it is not  $14 \cdot 14$  is that 14 comparisons of cards with themselves would be included. We formulate the condition that enough screenshots need to be taken so that an arbitrary but fixed combination of positions for two different colours is included with a probability of 95%.

As there are 4 such combinations of positions in each game the probability  $P(A)$  to get a specific combination in a game is

$$P(A) = \frac{4}{182}$$

The counter probability  $P(\neg A)$  is

$$P(\neg A) = 1 - \frac{4}{182} = \frac{178}{182}$$

The number of necessary screenshots to include a specific combination with 95% probability is calculated by solving

$$\begin{aligned}
 1 - P(\neg A)^n &\geq 0.95 \\
 1 - \left(\frac{178}{182}\right)^n &\geq 0.95 \\
 1 &\geq 0.95 + \left(\frac{178}{182}\right)^n \\
 \left(\frac{178}{182}\right)^n &\leq 0.05 \\
 n &\geq \log_{\left(\frac{178}{182}\right)}(0.05) \\
 n &\geq 134.8024
 \end{aligned}$$

This means that about 135 Screenshots of games are needed in the first case. Now we perform the same calculation for the second case where there is only one combination of positions for the colours. However, there are less combinations of positions that are relevant than in the first case. For example for two green cards at index 0 and 1 the 182 comparisons would include a comparison of the first green card and the second green card as well as a comparison of the second green card and the first green card. This goes for every two cards with the same colour, meaning there are only half as many different comparisons in the second case than in the first case. As a result the number of possible combinations in the second case is

$$C_2 = \frac{C_1}{2} = \frac{14 \cdot 13}{2} = 91.$$

Now we calculate how many screenshots must be taken to include a specific combination of positions for two equal colours with a probability of 95%.

As there is one such combinations of positions in each game the probability  $P(A)$  to get a specific combination in a game is

$$P(A) = \frac{1}{91}$$

The counter probability  $P(\neg A)$  is

$$P(\neg A) = 1 - \frac{1}{91} = \frac{90}{91}.$$

The number of necessary screenshots to include a specific combination with 95% probability is calculated by solving

$$\begin{aligned}
 1 - P(\neg A)^n &\geq 0.95 \\
 1 - \left(\frac{90}{91}\right)^n &\geq 0.95 \\
 1 &\geq 0.95 + \left(\frac{90}{91}\right)^n \\
 \left(\frac{90}{91}\right)^n &\leq 0.05 \\
 n &\geq \log_{\left(\frac{90}{91}\right)}(0.05) \\
 n &\geq 271.111
 \end{aligned}$$



This means that about 272 screenshots of games are needed in the second case. As there are more screenshots needed for the second case, the first case is also included. This inclusion is caused by the fact that each screenshot includes comparisons for the first as well as the second case. To have a buffer the decision was made to take screenshots of 300 games. In theory these screenshots include any arbitrary but fixed combinations of positions and colours with a probability of over 95%, which should eliminate the problem of the positional influence.

Nonetheless, this does not solve the second problem of extracting single pixels that can potentially be directly in the area of a sunbeam. This can be fixed, by extracting all pixels in a certain area of the card and averaging their rgb values.

Last but not least it needs to be clarified how the colour differences are calculated. To capture how colour differences are observed by humans, the rgb colour scale is not suitable. Using the rgb colour scale similarly strong perceived color differences do not necessarily have the same euclidean distance. Contrary to that the CIELAB colour space aims to do exactly that. Although not perfect, it more accurately describes human colour perception than the rgb colour space. Therefore the colour differences described in the similarity matrix are calculated after converting the colours into the CIELAB colour space. The colour distance is called Delta E score. The lower this score is the more similar appear the colours to human eyes. In related work, when creating the similarity matrix for the effect of colour blindness, the CIE1976 colour model was used for the calculations. However, the formula for calculating the colour difference has since then been improved multiple times, resulting in the CIE2000 colour model. Through multiple modifications of the formula it got closer to a visual equidistance, meaning similarly strong perceived color differences have more similar Delta E scores than before. Therefore when creating the similarity matrix for the glare effect the CIE2000 colour model is used instead of the old one.

To sum it up, the chosen approach is to take 300 screenshots of glare effect games with all cards turned face up, extracting areas of rgb values for each card, converting the average rgb values of areas into CIELAB colours, calculating a similarity matrix for each game that contains the delta E scores and finally averaging the delta E scores across all similarity matrices. To achieve values between 0 and 1, the colour differences additionally need to be scaled down in the end.

### 6.1.2 Screenshot extraction

To play the memory game and take screenshots an emulator for the pixel 2 in android studio was used. In order to take the screenshots in the needed way and collect additionally necessary information some changes to the memory game were made. In the way the game is intended there are always only maximum two cards turned face up. However, for the screenshots all cards need to be turned around so that the colour differences can be calculated. Therefore the memory game was adjusted so that all cards are turned face up once the player turns a card. Additionally it was changed so that cards stay face up once they get turned around for the first time. This makes it possible to take screenshots with the colours of all cards visible. Furthermore, for the creation of the similarity matrix it needs to be known what the original colours without the influence of the simulated sun are. To collect the screenshots and the according colours of the cards in each game a semi-automatic

approach was chosen. Once a game is started and a card is flipped, resulting in all cards being turned face up, the colours of all cards are saved in a list. Then a screenshot is manually taken and afterwards the game is exited to enter the main menu. From there on the process is repeated 300 times. As a result there are 300 hundred screenshots of games and a two dimensional list that includes the colours of the cards in the 300 games. The first dimension specifies the game and the second dimension the index of the card. The values of this list are stored in a text file before the game is completely closed. To assure that no mistakes were made when taking the screenshots, the number of collected screenshots is observed during the whole process and afterwards all screenshots are manually checked so that all cards are turned face up.

### 6.1.3 Implementation

This section will show and explain the implementation of the similarity matrix generation. Code is only included if it helps to further understand what is being explained. First of all the screenshots and the information about the original colours of the cards are loaded. Before the actual calculation of a similarity matrix for each image, the average rgb values of the cards on the field need to be determined.

```

1 def determine_glare_rgb_values(image):
2     '''
3     Calculates the rgb average rbg values for each card in an image.
4     :param image: The screenshot of the memory game with all cards
5     turned face up.
6     :return: The average rbg values in the specified 150 x 150 pixel
7     areas for each card.
8     '''
9     glare_rgb_values = []
10    for corner in card_corners:
11        x_border = corner[0] + 150
12        y_border = corner[1] + 150
13        card_values = []
14        for x in range(corner[0], x_border, 1):
15            for y in range(corner[1], y_border, 1):
16                coordinates = x, y
17                pixel_values = image.getpixel(coordinates)
18                card_values.append(pixel_values[: -1])
19        card_r = int(round(np.mean([color[0] for color in card_values])))
20        card_g = int(round(np.mean([color[1] for color in card_values])))
21        card_b = int(round(np.mean([color[2] for color in card_values])))
22        glare_rgb_values.append((card_r, card_g, card_b))
23    return glare_rgb_values

```

Listing 6.1: Add caption

The cards are each  $250 \cdot 250$  pixels big and the rgb values are extracted from squared  $150 \cdot 150$  pixel areas. The coordinates the pixels are extracted from are manually selected in gimp. As a result the areas are only correct for the used resolution of 1080p. Once the mean rgb values in those areas are calculated and converted into LAB colours the similarity matrix can be created. LAB colours are necessary to calculate the colour difference using the CIE2000 colour model. Each of the 28 cells in the similarity matrix falls into one of the two cases explained in **TODO: ref**, meaning there are either 4 or only 1 unique combination for the calculation of each

cell value. For every combination the lab colour distance is determined and the values for each cell are averaged. This completes the steps for a single image, resulting in a similarity matrix with unscaled values. Once a similarity matrix with unscaled values for every image is created, the average of all matrices is calculated. During the whole calculation of the single matrices it is being kept track of the highest occurring colour difference. The last step necessary to complete the final similarity matrix is to divide all values in the matrix by use the highest colour difference to scale them down to be between 0 and 1.

#### 6.1.4 Testing

To verify the correctness of the color extraction and the calculation of the similarity matrix, the main functionalities are manually tested. The colour extraction was tested by extracting colours of a screenshot with the script and comparing the rgb values to the actual values in the images using gimp. Furthermore the calculation of the delta e score was tested, by comparing results of the script with those of an online calculator.

To assure that the 300 games actually include most of the combinations and therefore eliminate the positional influence of cards, the coverage of combinations can be additionally calculated. Therefore the number of unique combinations included in the 300 screenshots must be divided by the overall number of possible combinations. The number of unique combinations can be counted during the process of creating the similarity matrix, but the number of possible combinations must separately be determined. The similarity matrix for glare effect has 28 entries, from which 7 are comparisons between same and 21 between different colours. With  $C_1$  and  $C_2$  being the possible combinations for two arbitrary but fixed coloured cards in the two cases mentioned above, the number of overall possible combinations  $C$  is

$$\begin{aligned} C &= 21 \cdot C_1 + 7 \cdot C_2 \\ &= 21 \cdot 182 + 7 \cdot 91 \\ &= 4459. \end{aligned}$$

The 300 games used for creating this matrix include 99.57% of all possible combinations. Furthermore it was verified that certain combinations are not included significantly more often than others and in that sense have stronger influence on the result. **TODO: code ergänzen und nachprüfen wie oft die combinationen vorkamen..histogramm?.**

#### 6.1.5 Final Matrix

In figure 6.3 the final similarity matrix can be seen, displaying how colour differences are observed under the influence of the simulated sunlight, averaged over 300 games. As mentioned before, the lower the delta e score the more similar the cards appear to the human eye. The maximum delta E score that occurred during the calculation was 8.861. All values are divided by that number to achieve scores between 0 and 1.

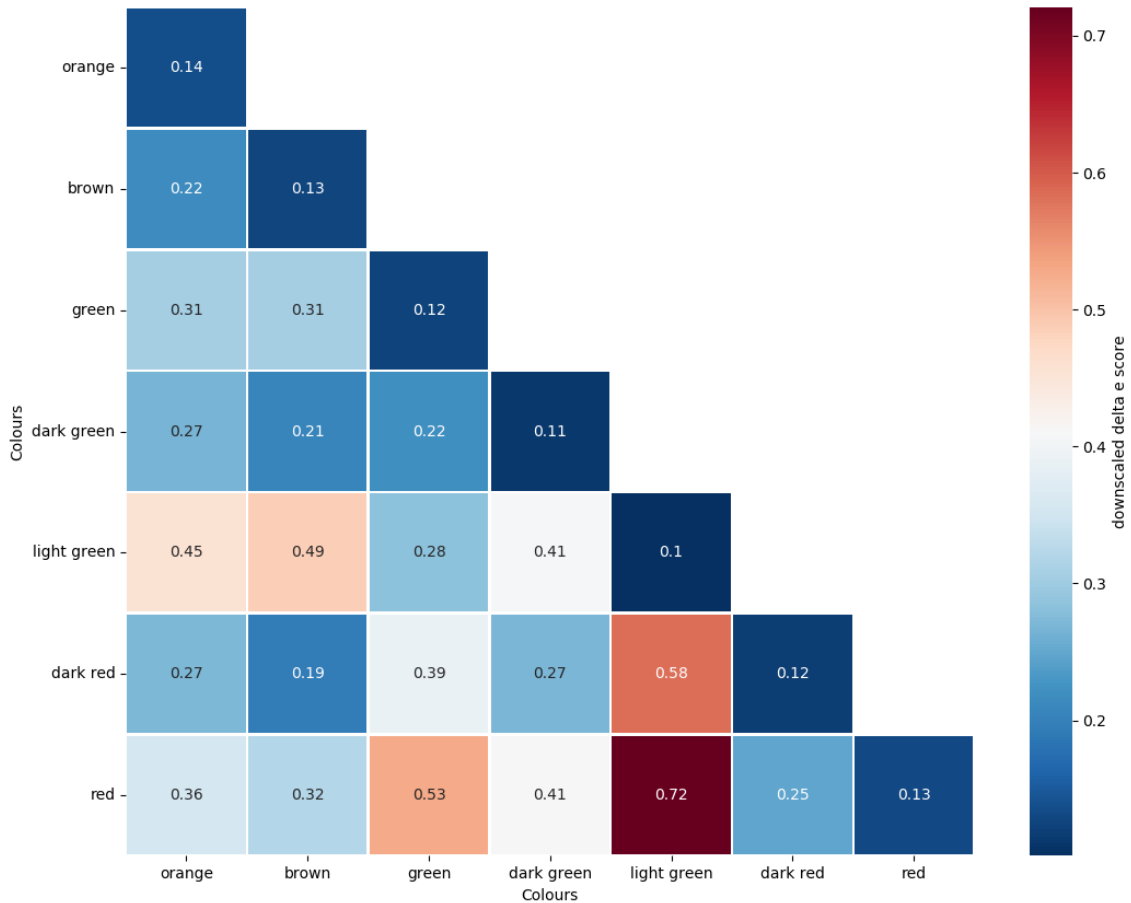


Figure 6.3: Add caption

The delta E score is less a definitive answer, and instead a helpful metric to apply to a specific use case. Although there are tendencies regarding the interpretation of colour differences, there is no definitive table that describes what the different scores mean. One reason for this is that the perceived colour difference may vary in different situations and circumstances. For instance, is the colour difference perceived differently depending on how long the colours are exposed to the human eye, as humans can over time adapt to the colours. This means that in a setup where all colours are always visible the colour difference will likely be perceived weaker than in the case of the memory game, as the cards are only flipped momentarily, leaving little time for adaptation. Table ref was created through own observation and in that sense is highly influenced by the strength of the sense of sight from the creator of this work. Therefore it should not be taken definitive but only serve the purpose of clarifying how the values in the matrix can roughly be interpreted.

Table 6.1: add caption.

Delta E	Downscaled values	Perception of difference
$\leq 1.33$	$\leq 0.15$	Not perceptible by human eyes
$1.33 - 2.66$	$0.15 - 0.3$	Only perceptible through close observation
$2.66 - 4.43$	$0.3 - 0.5$	Perceptible colour difference
$\geq 4.43$	$\geq 0.5$	Major colour difference

Before actually using the constructed similarity matrix in the simulator it is unknown

if the chosen approach results in high quality simulations. If this is not the case it is possible to change the concept or try other approaches.

## 6.2 Adaptation and correction of logs

One problem with the logs is, that the mapping was done statically when they were created. However, the simulator requires dynamic mapping. The difference is explained in chapter **TODO: ref**. To do so there was already a script available, that was used after replacing some deprecated functions from libraries with supported ones. Furthermore the code was extended so that the old placeholder similarity matrices in the logs are replaced by the new one. As the code was initially written for similarity matrices with 21 entries and the one for glare effect has 28 entries, the code for remapping the logs had to be adapted. Finally all remapped logs are saved in new files. To check whether the remapping is still done correctly, the resulting logs were compared to logs that were remapped by the original script. The only thing that was not original were the replaced deprecated functions as this was necessary to execute the script. Besides different similarity values due to different matrices as well as the fact that the logs now also contained entries for same coloured cards, the mapping was identical.

## 6.3 Removal of invalid logs

For the simulation to work, the game log that is used for the simulation must have had all cards turned around at least once. If this is not the case the log is classified as invalid. Additionally to removing invalid logs, the aim is to create a balanced data set for training. Therefore the number of games in each game mode has to be equal. Using twice as many no obstacle games for training than glare effect games can lead to games being more often classified to have no obstacle only because they appeared more often during training. As a result the model would be less capable of detecting actual visual interaction obstacles. Furthermore the decision was made that the number of games from each participant in each game mode should be equal, too. This means that the data should not contain more no obstacle games from a participant than glare effect games, and vice versa.

As stated in chapter **TODO: ref** from the 22 participants there are 44 no obstacle and 21 glare effect game logs. To collect the data used for the simulation, a script was written that collects one no obstacle and one glare effect log from each participant. Half of the available no obstacle logs are not collected, because there are not as many glare effect logs. Once the logs are collected, they are validated and the valid ones are saved in new files. If at least one of the two logs from a participant from different game modes is invalid, both are removed. Otherwise the training data would be inconsistent in that sense that the logs from different game modes could be from different participants. From the 22 no obstacle and 21 glare effect logs collected by the script, one glare effect log was invalid. This resulted in 20 no obstacle and 20 glare effect logs being used for simulation. These 40 real logs combined with those simulated form the data used for training.

```

1 def validate_log(log):
2     '''
3     Validates logs.
```

```

4  :param log: The log to validate.
5  :return: If the log is valid.
6  '''
7  needed_entries = [ '1.1,', '2.1,', '3.1,', '4.1,', '5.1,', '6.1,', '
    7.1,', '1.2,', '2.2,', '3.2,', '4.2,', '5.2,', '6.2,', '7.2,' ]
8  for needed_entry in needed_entries:
9      if needed_entry not in log:
10         return False
11  return True

```

Listing 6.2: Add caption

## 6.4 Simulation of user behaviour

The whole process of simulating user behaviour is divided into two parts: First configuration files are created, using a generic optimization algorithm, that contain the optimized parameter values and secondly these configuration files are used to simulate games.

Multiple additions were made to the simulator. In total four classes were added. Two are for generating configuration files for the simulation of game with and without the glare effect obstacle (NoObst\_ConfigGenerator\_dataCollection2020.java and GlareEffect\_ConfigGenerator\_dataCollection2020.java) and the other two for using those files to simulate new games based on the user behaviour in the original games (NoObstWinStrategyTrainingDataGenerator\_dataCollection2020.java and GlareEffectWinStrategyTrainingDataGenerator\_dataCollection2020.java). These classes utilize the functionalities already implemented in the simulator. **TODO: grob erklären wie code funktioniert? momentan glaube ich eher nicht** However, as the simulator only worked with similarity matrices that have 21 entries, instead of the 28 of the matrix for glare effect games, the simulator was adjusted so that it can also handle matrices with 28 entries. After all changes and additions were completed, each log was used to simulate 1000 games, resulting in 40000 logs. From these logs 20000 are no obstacle games and the other 20000 are glare effect games. The 40000 logs contain 1000 no obstacle and 1000 glare effect logs for each of the 20 probants.

## 6.5 Sorting logs by quality

The simulated logs were sorted by their quality, using the root mean squared error between the performances....**TODO: klären genau was da gemacht wird mit mazen, ich glaube rmse für alle performance werte in einem game und dem originalen game** as measurement. This is important because it enables to only use the best n simulated logs for training instead of using all of them or a random subset.

## 6.6 Evaluation of simulation

To evaluate whether the performance in the simulated logs is similar to the one in the original logs two performance measurements are used: The matching pairs in each round and the penalties in each round. Penalties are given if a card is revealed whose partner has already been seen before but the pair was not picked up. The

Initial simulation results can be seen in figure 6.4 and 6.5. **TODO: erklären was whiskers sind. ich glaube einfach nur std abweichung?**

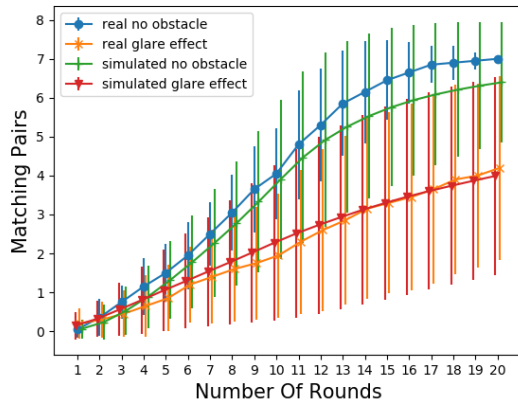


Figure 6.4: Add caption

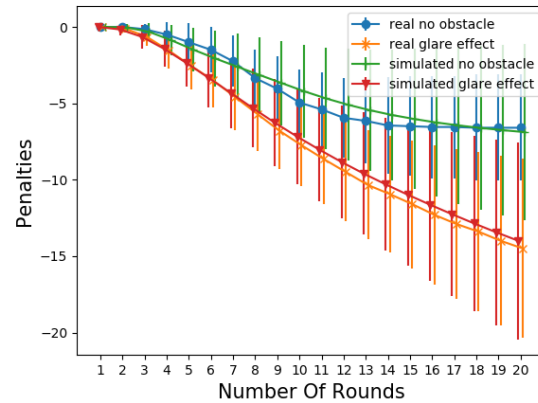


Figure 6.5: Add caption

It can be seen that the simulations of glare effect games are very similar to the real games, meaning that the constructed similarity matrix creates good simulation results. As a result there is no need to find a different approach for creating the similarity matrix, and no changes to the simulator regarding the simulations of glare effect games are made. However, the simulation of no obstacle games is not quite as accurate. Especially the number of matching pairs after the tenth round in the simulated games is lower than in the real games. Furthermore, however less noticeable, is the penalties in the simulated games between round 8 and 18 lower than in the real games.

The fact that matching pairs per round for the no obstacle simulations are noticeably lower than in the real games inspired two changes to the simulator. The first one was to introduce a new parameter for the optimization algorithm to optimize. It was called randomizing decay and is a value between 0 and 1, and addresses the problem of worse performance in the last turn compared to real games. The boundary limit, describing..., is multiplied with the randomizing decay each round beginning with the 20th round. This reduces the ... and therefore results in better performance in the last 10 steps. The second change was to reduce the probability of revealing a random card (instead of pursuing a win strategy) by 0.2 (randomrevealprobability is a random value between 0 and 1). This was done to increase the overall performance in the simulations by reducing the randomness. The results of these changes can be seen in figure 6.6 and 6.7. **TODO: herausfinden und fertig beschreiben was die änderungen bewirken. Frage: wieso ist in ersten 10 zügen keine änderung obwohl die eine änderung eigentlich generell performance verbessert**

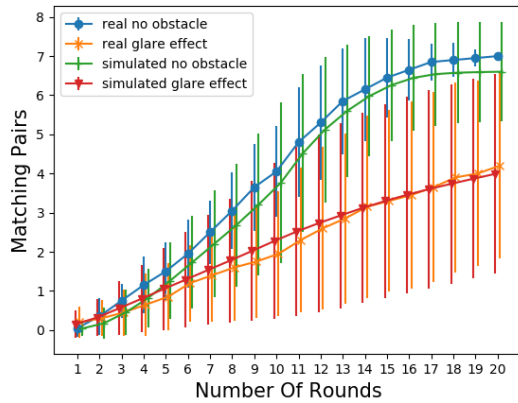


Figure 6.6: Add caption

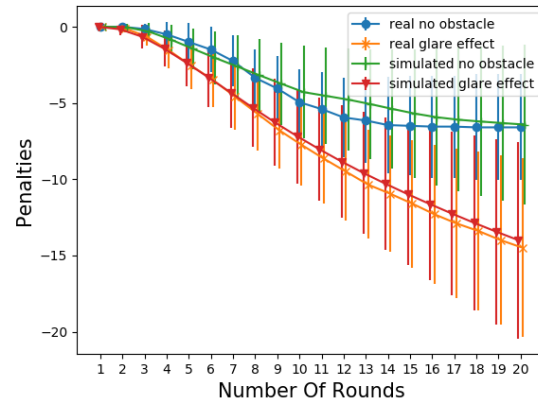


Figure 6.7: Add caption

It is noticeable that the number of matching pairs per round in the simulated no obstacle games is closer to the real games. However, this reduction of randomness also resulted in less penalties. It can be seen that the penalties per round are less close to the real games than before the changes. It seems that the improvements regarding the matching pairs overweight the deteriorations of the penalties, but if these changes actually improve the overall quality of the simulations will be seen when comparing the classification results before and after the changes. This is done in chapter **TODO: ref**. Due to the relatively small changes it is likely that the classification results give no clear answer on whether the changes were beneficial for the quality of the simulations.

Although major tendencies of similarities and differences are visually noticable in figure **TODO: ref** and figure **TODO: ref**, a more accurate analysis is needed to to make reliable statements. Therefore multiple paired samples t-test are performed, comparing real no obstacle, simulated no obstacle, real glare effect and simulated glare effect games with each other. For each combination two tests are performed. One comparing the mean penalties per round and the other one comparing the mean numbers of matching pairs per round. The resulting p values can be seen in table **TODO: ref** and table **TODO: ref**. These values express whether the compared lists of mean values are significantly different or not. Values higher than 0.05 indicate no significant difference, while values below 0.05 indicate a significant difference. All paired sample t-tests were performed after the changes to the simulator. It would be optimal if all comparisons between different game mode show significant difference, while comparisons between real and simulated games of the same type show no significant difference.

**TODO: die tabellen kann man mit minipages gut nebeneinander tun**



Table 6.2: p values in paired t-test for different comparisons of penalties per round. All 1000 simulated games per real game were used. The following abbreviations are used: real glare effect - r\_g, real no obstacle - r\_n, simulated glare effect - s\_g, simulated no obstacle - s\_n

mp	s_n	s_g	r_n	pen	s_n	s_g	r_n
r_g	$5.7e-06$	0.0688	$9.3e-07$	r_g	$4.8e-06$	$1.1e-06$	$3.2e-06$
r_n	$1.8e-10$	$2.7e-06$		r_n	0.0195	$5.6e-06$	
s_g	$1.7e-05$			s_g	$7.2e-06$		

Table 6.4: p values in paired t-test for different comparisons of matching pairs per round. All 1000 simulated games per real game were used. The following abbreviations are used: real glare effect - r\_g, real no obstacle - r\_n, simulated glare effect - s\_g, simulated no obstacle - s\_n

	s_n	s_g	r_n
r_g	$5.7e-06$	0.0688	$9.3e-07$
r_n	$1.8e-10$	$2.7e-06$	
s_g	$1.7e-05$		

Table 6.5: p values in paired t-test for different comparisons of penalties per round. All 1000 simulated games per real game were used. The following abbreviations are used: real glare effect - r\_g, real no obstacle - r\_n, simulated glare effect - s\_g, simulated no obstacle - s\_n

	s_n	s_g	r_n
r_g	$4.8e-06$	$1.1e-06$	$3.2e-06$
r_n	0.0195	$5.6e-06$	
s_g	$7.2e-06$		

It can be seen that all tests comparing different game mods show significant difference in penalties per round as well as matching pairs per round, which is desired. However significant difference is not desired in the four test comparing real and simulated games of the same mode, which are highlighted by colours. The only test showing no significant difference is the one between matching pairs per round in real and simulated glare effect games, highlighted in green. The other three tests express undesired significant differences between real games and their simulations. As the simulated games are sorted by their quality in section **TODO: ref**, it is possible to only use a certain number of best simulations for the comparisons, and see whether this improves the undesired properties mentioned above. Furthermore the number of rounds in the comparison can be changed, so that not all 20 rounds are included. The consideration of including less rounds is of importance, because it can reveal which parts of the game are simulated accurately and which are not. The search for a better configuration of those two settings was manually done by varying the number of steps and the ratio between simulated and real data. The aim was to include as many simulations and steps as possible while achieving significant differences only in tests where it is desired. A configuration that was better in the sense of achieving this aim consisted of using only the best 10 simulated games per game

mode from each participant and only including the first 10 rounds of each game. The results can be seen in table **TODO: ref** and **TODO: ref**. The two performance measurements whn only using the best 10 simulated games per game mode from each participant are additionally portrayed in figure **TODO: ref** and **TODO: ref**.

**TODO: die tabellen kann man mit minipages gut nebeneinander tun**

Table 6.6: p values in paired t-test for different comparissons of matching pairs per round. For the simualted data only the best 10 simulated logs from each real log are used. Only the first 10 rounds are used. The following abreviatrions are used: real glare effect - r\_g, real no obstacle - r\_n, simulated glare effect - s\_g, simulated no obstacle - s\_n

	s_n	s_g	r_n
r_g	0.0064	$1.5e-05$	0.0058
r_n	0.0882	0.0039	
s_g	0.0044		

Table 6.7: p values in paired t-test for different comparissons of penalties per round. For the simualted data only the best 10 simulated logs from each real log are used. Only the first 10 rounds are used. The following abreviatrions are used: real glare effect - r\_g, real no obstacle - r\_n, simulated glare effect - s\_g, simulated no obstacle - s\_n

	s_n	s_g	r_n
r_g	0.0020	0.0564	0.0014
r_n	0.9005	0.0012	
s_g	0.0017		

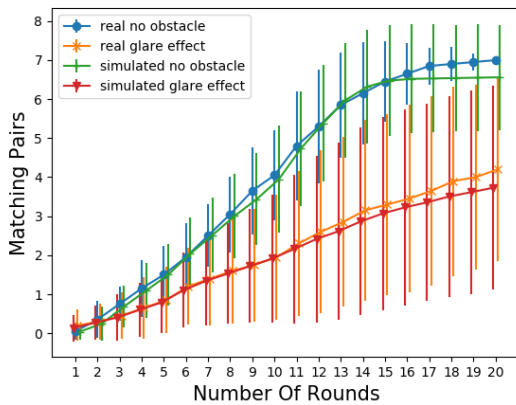


Figure 6.8: Add caption

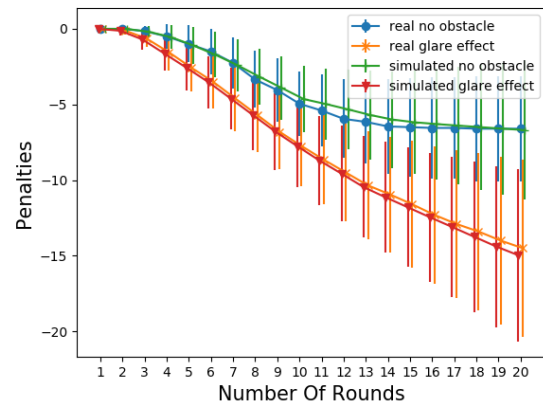


Figure 6.9: Add caption

A configuration that only results in desired outcomes of statistical tests was not found. If less simulations are used, the three formerly significant differences become insignificant, but the difference in matching pairs bewtween simulated and real glare effect games becomes significant. Therefore using the 10 best simulations and the first 10 steps is a good compromise. It should be empghazised that this does not

mean that the simulation of the glare effect games is bad regarding the number of matching pairs per round. This only means that the paired sample t-test finds a significant difference in that comparison. If the two lists of mean values that supposedly are significantly different are manually compared it can be seen that in reality the difference is very small. These values can be seen in table **TODO: ref.**

Table 6.8: mean values of matching pairs for real and simulated games. sd10x. for the first 10 rounds

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10
real	0.2	0.3	0.45	0.65	0.85	1.2	1.4	1.6	1.75	1.95
simulated	0.135	0.27	0.405	0.6	0.79	1.105	1.34	1.53	1.72	1.915

The highest mean difference is in round 6 with a difference of 0.095 matching pairs in that round. This shows that it does not mean that all glare effect simulations used are bad regarding the matching pairs, only because the statistical test finds a significant difference. The fact that the paired sample t-test concludes a significant difference could be related to the fact that the range of the values is very small ranging only from .. to .. meaning a small difference is much according to the t-test. **TODO: checken ob das logisch ist, also gucken was der ttest überhaupt macht und das begründen** (auschnitt beschrieben wenn ich weniger nehme.). This would also fit with the fact that using more steps and therefore having a bigger range of values for matching pairs in glare effect games, the difference is not significant (see example where all data is used, ref).

Although no clear boundary was found in which all simulations have no significant difference to the according real games was not found it can be said that the simulator is capable to simulate the first 10 rounds realistic enough to not be significantly different according to the paired sample t-test, given only the 10 best simulations are used. Having said that, this does not mean that the simulator is not capable of realistically simulating the other rounds, but only that a state of no significant difference between simulations and real games (except one comparison) is achievable when using the first ten steps instead of all 20 steps. This analysis also indicates, that it is not ideal to use all simulated data for training as that results in a significant difference between the simulated and the real games. **TODO: das ist schlecht, weil ich das zu diesem Zeitpunkt eigentlich noch nicht weiß. Besser: eher Theorie aufstellen** As it can be seen in chapter results, using a certain amount of simulations increases the overall performance compared to only using real games. This observation paired with the knowledge that using too many simulations will result in a significant difference between the simulated and the real games, leads to the theory that there should be a sweet spot regarding how many simulations to use. By using different amount of steps and simulated data this sweet spot is tried to be discovered in chapter **TODO: ref.**

Nonetheless the aim of no significant difference between real and simulated games is set very high. Overall the simulations and real games are very similar when comparing them with the two performance measurements. If only the best simulations are used the performance becomes even more similar.

## 6.7 Feature generation

### 6.7.1 1D CNN features

For the 1D CNN five statistical features for each step are used: The card codes, the number of cards left, the number of never revealed cards, the highest number of times the same card was revealed and the number of steps since all pairs were found. They are calculated using the game logs from **TODO: ref zu dem data chapter**. These features can be directly fed into the 1d CNN explained in chapter **..TODO: ref**.

The code for calculating the statistical features out of game logs was already given. A script was written that incorporates this functionality in order to calculate the features for all logs. Additionally, after creating the necessary directory structure the script saves the files for the splits of the raw data and the features. The reason for saving the features is that they do not have to be calculated before every training and instead can be loaded from files. The reason for also saving the raw data even though this work does not need them anymore is, that the working group that was collaborated with also trained other models and does not directly load the features but instead calculates them before each training.

### 6.7.2 2D CNN features

For the second approach of using a 2D CNN further steps are taken. Therefore synthetic images in gray scale colours are created using the features mentioned above. As the images are in gray scale colours, only one color channel is needed. One can visually think of this approach as creating graphs for each feature that display their values in each timestep, taking a photograph of each graph and stacking them on top of each other. This can be clarified by looking at the code that produces the synthetic images.

```

1 def create_image(game, components=[True, True, True, True, True]):
2     '''
3     Creates synthetic images out of the five statistical features.
4     :param game: The statistical features in each step.
5     :param components: Five values describing which of the features
6     should be used to create the image.
7     :return: The synthetic image.
8     '''
9     card_codes = np.zeros((7, steps))
10    cards_left = np.zeros((8, steps))
11    never_revealed_cards = np.zeros((14, steps))
12    max_same_card_reveals = np.zeros((20, steps))
13    rounds_since_done = np.zeros((27, steps))
14
15    x_position = 0
16    for step in game:
17        card_code = math.floor(step[0])
18        first_or_second = int(round((step[0] % 1) * 10))
19        if card_code != 0:
20            card_codes[card_code - 1][x_position] = first_or_second
21        pairs_left[int(step[1] / 2)][x_position] = 1
22        never_revealed_cards[int(step[2])][x_position] = 1
23        max_same_card_reveals[int(step[3])][x_position] = 1
24        rounds_since_done[int(step[4])][x_position] = 1

```

```
25     x_position += 1
27     image = np.zeros((0, steps))
28     if components[0]:
29         image = np.vstack((image, card_codes))
30     if components[1]:
31         image = np.vstack((image, max_same_card_reveals))
32     if components[2]:
33         image = np.vstack((image, rounds_since_done))
34     if components[3]:
35         image = np.vstack((image, cards_left))
36     if components[4]:
37         image = np.vstack((image, never_revealed_cards))
39     return image
```

Listing 6.3: Add caption

By using pseudo colors, the images can be displayed with colours, like in figure 6.10. These  $76 \cdot 40 \cdot 1$  (height  $\cdot$  width  $\cdot$  colour channels) images are directly fed to the 2D CNN explained in chapter .. **TODO: rerf**. By setting the origin of the image to the lower instead of the upper left corner a more naural looking image is created and can be seen in figure 6.11.

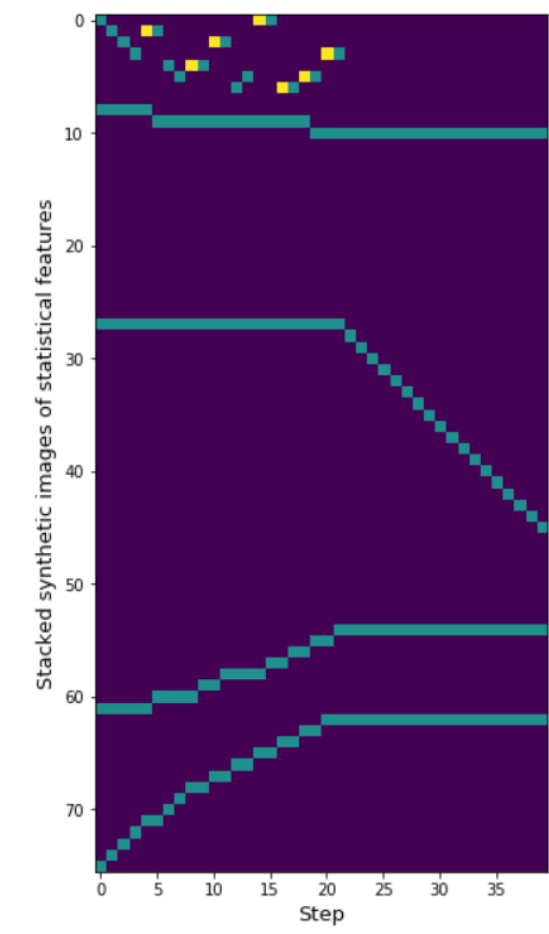


Figure 6.10: Add caption

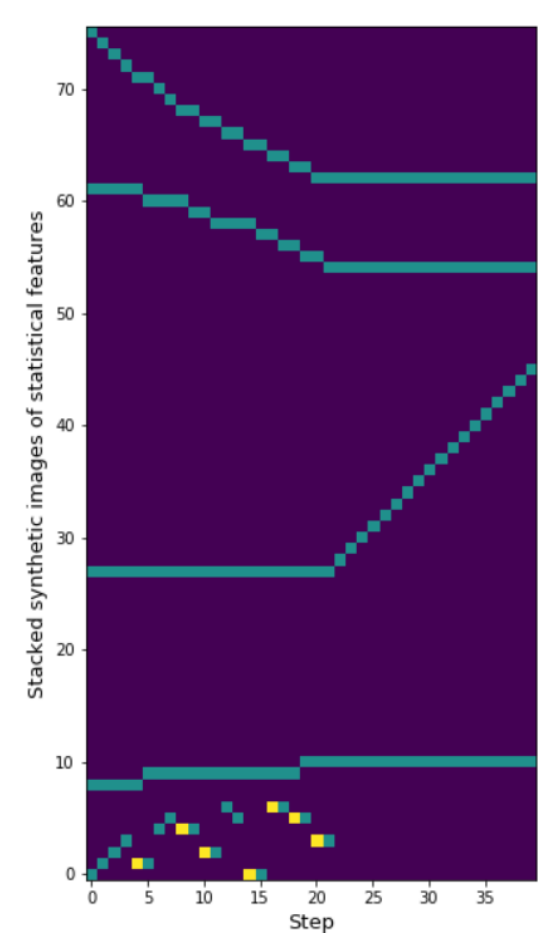


Figure 6.11: Add caption

The width of the image is 40 because that the number of steps recorded. Table

**TODO: ref** shows which of the areas in the image describe which statical feature. The different vertical spaces given to statistical features were purposefully chosen. The aim was to use as much space as neccessary but as little as possible. The card codes have 7 pixels of vertical space since there are 7 different colours in the game. As each round consists of two cards being turned face up and the impossibility to chose the same card twice in one turn, a card can be at maximum revealed 20 times during 20 rounds. Considering that this value is at minimum 1 because it is not possible to flip no cards, the range of 20 values is sufficient for this value. Furthermore 14 steps are at minimum needed to complete the game, since there are 14 cards. As a result this value can range from 0 to 26, meaning a vertical space of 27 is needed. In order to save space, the statistical feature of the number of cards left was converted to the number of pairs left. By dividing by 2 the vertical space neccessary to visualize this feature is halved, without information being lost. Instead of representing the numbers left on the fiel, the value now represents the number of pairs left. Last but not least the number of never revealed cards can range from 0 to 13. The value 14 is not possible because two cards have to be chosen each turn. The stacking order was chosen so that the coloured pixels of different statistical features rarely touch each other.

Table 6.9: add caption. Upper and lower boundaries are inclusive.

Statistical feature	Range	Vertical space in the image
Card codes	1-7	0-6
Maximum of same card reveals	1-20	7-26
Steps since game ended	0-26	27-53
Pairs left	0-7	54-61
Never revealed cards	0-13	62-75

Regarding the visual representatoin of he card codes two two major decisions were made: Using different colours for the first and the second card of a colour and combining the information about the 14 cards in an vertical area of size 7.

The idea behind using different colours for the first and the second card of a colour is that this vizualizes important behavioural characteristics that help deciding whether the visual obstacle being the sunlight is involved or not. By comparing the the card code sections of two images from which one is created using a glare effect and the other one with a no obstacle game, these visual differneces become clear.

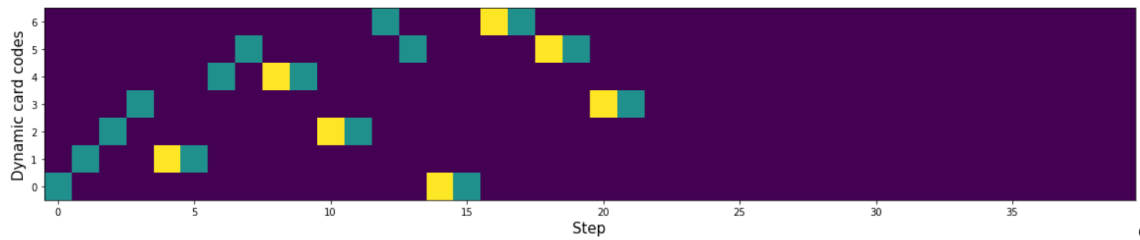


Figure 6.12: Add caption

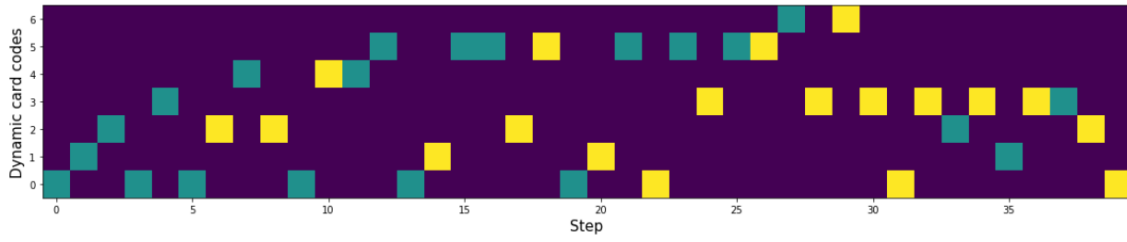


Figure 6.13: Add caption

In 6.12, showing the image for an no obstacle game, there are yellow pixels directly left of green ones, meaning that the probant flipped a card, knew that he had already seen the matching card and directly flipped it. However, this happens less often in figure 6.13 which shows the image for an glare effect game. This is not the case in every game, but the theory is that separations of different coloured cells in the image are more likely in glare effect games than in no obstacle games. Therefore it could be beneficial for the model to also learn these characteristics. If the same colour was used for all card codes, there would be no visual difference between flipping the same card in consecutive turns and flipping two different cards with the same colour. This of course would also mean that the model could not differentiate between those two cases.

Instead of combining the information about the 14 cards in an vertical area of size 7, it would have been also possible to use double the vertical space so that the first and the second card of a colour each have separate areas. The first cards of a and the second cards of a colour would each have individual vertical spaces with sizes of 7. However, when considering how convolutions in 2d cnns are performed, this approach seems less promising. During the convolutions, each filter has a fixed size and step by step walks over the image, extracting features from areas to create a feature map **TODO: ref to cnn chapter**. This means that the extracted features during a convolution and therefore the gained knowledge of the model depends on the information found in areas with the size of the filter. Regarding the topic at hand, choosing individual areas for the card codes would create a significant distance between the pixels of the first and second card of a colour. In order to extract features during the convolution that capture the connection between those distanced pixels, the filter size would need to be at least as big as the distance. This results in large areas that include many different pixels overshadowing the importance of the two specific pixels describing the first and the second card of a color and thereby making it harder for the model to learn the connection between them. Hence, the approach of doubling the vertical space results in high unlikelyness of the model learning this characteristic. Contrary to that, using a combined space and different colours results in spatial proximity of the pixels describing the desired characteristic, which can therefore be captured by smaller filter sizes. In conclusion, choosing a representation for the card codes that contains local features and using a small filter size, results in a higher likelihood of capturing the desired characteristics.





## 7. Convolutional neural networks

As convolutional layers are complex and have many variations for different use cases, only concepts and functionalities that are relevant for this work will be covered. This also means that 3 dimensional cnns will not be explained, but in general cnns function the same way whether they have 1, 2 or 3 dimensions. The main difference lies in how the filter, also known as kernel or feature detector, moves across the data. This is later explained in more detail. (**TODO: ich glaube input data ist keine regel sondern nur meist so..zum beispiel kann man auch 1d convolution bei 1d, 2d, oder 3d daten machen**).

First of all the core concepts of cnns will be explained and afterwards the models used in this work will be described.

One of the core concepts of convolutional neural networks are the convolutions.

Intuitively, a 2d cnn must not necessarily have 2 dimensional data as input.

Notes for models:

- stride = wie viele schritte auf einmal, default 1 deshalb vermute ich bei mir 1
- maxpooling: pool size =  $2 \times 2$ , und bei keras wird stride dann default auch zu  $2 \times 2$  (=pool size)
- input size =  $85 \times 40$  glaube ich
- filter weights are randomly initialized, so that they don't all initially learn the same features. To briefly explain the behaviour of filters a separation between two cases can be made: The first case is that not all features of high quality have been learned yet, high quality meaning that learning them would lower the cost function. In that case it is highly unlikely that each filter would begin to resemble other filters, as that would most certainly result in an increase of the cost function and therefore no gradient descent algorithm would head in that direction. The other case is that all features of high quality have already been learned by a subset of the available filters. In this case the cost function would not be increased if the remaining filters learn similar features than the other filters.
- in einem der schritte wird eine zeile ausgelassen, aber ich trainiere nicht nochmal

alles neu (oder?)

- 1d and 2d convolution - reduction of dimension - relu - number of filter, random initialization, etc - kernel and kernel size - loss function and optimizer, stochastic gradient descent - Dropout layers - Pooling, max pooling - flatten - dense relu, softmax - dropout layer, pooling ändert nichts an der anzahl der feature maps - **TODO: bessere quelle finden, buch oder so** - stride (ganz kurz, beschreibt verhalten wenn etwas über bleibt, zero padding oder weglassen) - back propagation - wie schaffen die es bspw. 64 feature maps auf 32 zu reduzieren? also wie funktionieren die filter in dem fall?

# 8. Classification

## 8.1 Hardware resources

Multiple servers, provided by the cognitive system lab, were used. One of the provided servers is optimized for storing large amounts of data and was therefore used to store the histories and the results from the models during training while the other servers focus on computing power and were used to execute that training and evaluation. Mainly three servers were used for the training. Two have the hardware specification shown in table **TODO: ref** and the other one those shown in table **TODO: ref**.

Table 8.1: add caption

Component	Model	Additional details
Processing Unit	Intel® Xeon® Gold Prozessor 5218	64 logical cores, turbo boosts to 3,90 GHz
System Memory	-	377 GB
Grafics	2 x NVIDIA RTX 2080 ti	11 GB graphics memory

Table 8.2: add caption

Component	Model	Additional details
Processing Unit	Intel® Xeon® Prozessor E5-2650 v4	24 logical cores, turbo boosts to 2,90 GHz
System Memory	-	252 GB
Grafics	NVIDIA GTX 1080 ti	11 GB graphics memory

**TODO: specs nochmal checken sobald ich ins cartesium kann und da nachfragen. vor allem grafikkarte und hesteller des rams**

## 8.2 Models

For determining whether probands were blinded, one dimensional and two dimensional convolutional neural networks were utilized. One dimensional convolutional neural

networks have become popular for time series classifications **TODO: ref**, while two dimensional convolutional neural networks are mostly used for image processing. However, a possibility is to create synthetic images from the data and use these images in a two dimensional convolutional network as done in section **TODO: ref**. The dimensions of the data and the feature maps in the visualizations are calculated for the case that all 20 rounds are used. If a different number of rounds is used the values differ. The CNN models are intentionally created not very complex for two reasons: Firstly, a lot of training needs to be done with different configurations regarding the steps included and the ratios of simulated to real data. As mentioned in **TODO: ref**, 76800 models need to be trained. As the time this work is created in is limited, training that many complex models would take too long. Especially the training of the 2D CNNs requires a lot of time. Secondly, to assure that the structures are complex enough, more complex models were exemplary trained and they showed either similar or worse performance. The results of the more complex models can be seen in the appendix **TODO: ref?**.

### 8.2.1 1D CNN

The structure of the 1D CNN used is shown in figure 8.1.

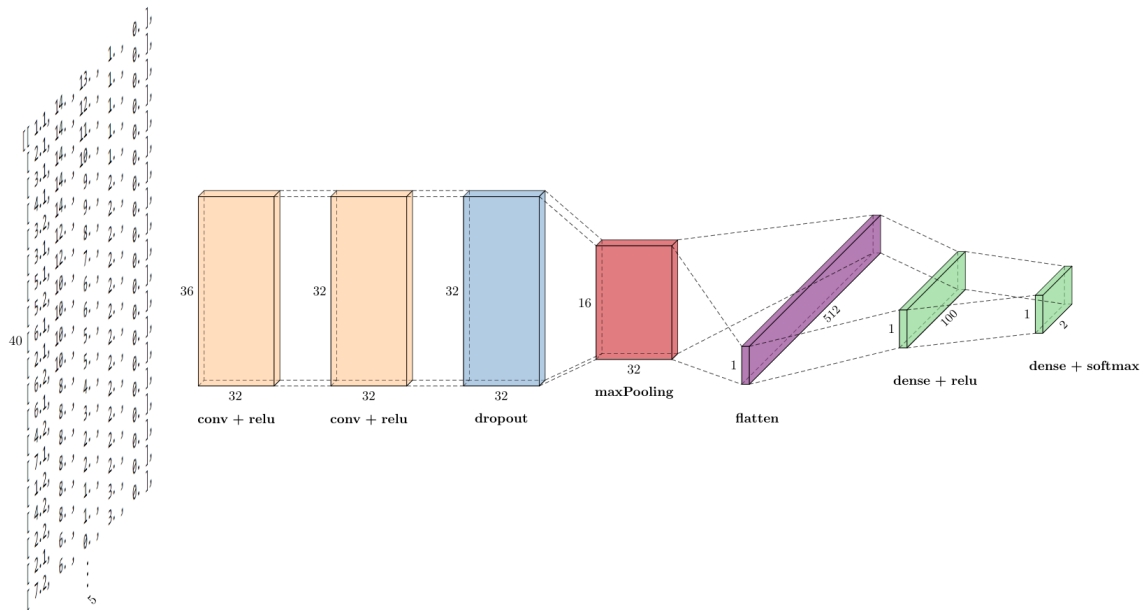


Figure 8.1: The width describes the number of feature maps and the height and depth describe the dimensions of the feature maps. As example after the first convolutional layer the data consists of 32 one dimensional feature maps with the length of 36.

The input data consist of the 5 statical features, explained in section **TODO: ref**, for each of the 40 steps (20 rounds), resulting in the input dimensions  $40 \cdot 5$ . Initially the data is passed to the first convolutional layer of the network. The first layer has 32 filters and the kernel has a size of 5, meaning that every step from the kernel includes all data from 5 steps in the game. Since the input data is two dimensional and the convolution is one dimensional, the created feature maps are one dimensional. Therefore the first convolutional layer creates 32 feature maps each with a size of

36. These feature maps are passed to the second convolutional layer with the same number of filters and the same kernel size as the first layer. This again results in 32 feature maps with a decreased size of 32. Afterwards a dropout layer with a rate of 0.5 randomly sets input units to 0 and by that helps to prevent overfitting. Then a one dimensional max pooling layer with a pool size of 2 reduces the size of the 32 feature maps to 16. And finally the network is completed with a flatten layer and two dense layers. **TODO: vielleicht genauer erklären was dense macht, aber eigentlich ja oben auch schon, nur einen groben satz hier.** Both convolutional layers, and the first dense layer use a rectified linear unit as activation function. The last dense layer uses softmax normalization to bring the data down two two values for the two classes. The elements of the output vector are between 0 and 1 and sum up to 1 and specify the result of the classification. In total 57486 parameters are trained in this model.

### 8.2.2 2D CNN

The structure of the 1D CNN used is shown in figure 8.2.

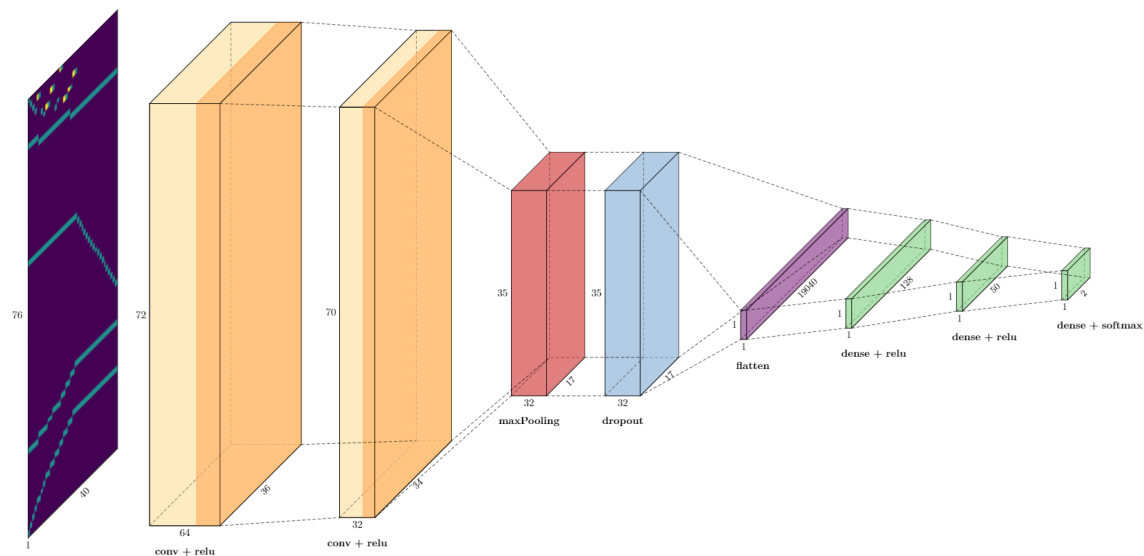


Figure 8.2: The width describes the number of feature maps and the height and depth describe the dimensions of the feature maps. As example after the first convolutional layer the data consists of 64 two dimensional feature maps with the size of  $72 \cdot 36$ .

**TODO: vielleicht darstellung anpassen, damit die gleichen diemnsionen das gleiche bedeutet wie bei 2d cnn, mit feature maps und so** The input data consist of the synthetic images generated in .. **TODO: ref** using the 5 statistical features, resulting in the input dimensions  $76 \cdot 40 \cdot 1$ . Initially the data is passed to the first convolutional layer of the network. The first layer has 64 filters and the kernel has a size of  $5 \cdot 5$ , meaning that each step from the kernel includes a  $5 \cdot 5$  sized area of the synthetic image. This convolution results in 64 feature maps each with dimensions of  $72 \cdot 36$ . These feature maps are passed to the second convolutional layer with 32 filters and a kernel size of  $3 \cdot 3$ . This results in 32 feature maps and a decreased dimensionality of  $70 \cdot 34$ . Then a two dimensional max pooling layer with a pool size of  $2 \cdot 2$  reduces the dimensinalty to  $35 \cdot 17$ . Afterwards a dropout layer with

a rate of 0.2 randomly sets input units to 0 and by that helps to prevent overfitting. And finally the network is completed with a flatten layer and three dense layers **TODO: vielleicht genauer erklären was dense macht, aber eigentlich ja oben auch schon, nur einen groben satz hier**. Both convolutional layers, and the first two dense layers use a rectified linear unit as activation function. The last dense layer uses softmax normalization to bring the data down to two values for the two classes. The elements of the output vector are between 0 and 1 and sum up to 1 and specify the result of the classification. In none of the different operations zero padding was required in order to prevent information loss. In total 2463928 parameters are trained in this model.

### 8.3 Training setup

There is one script each for training the 1D cnn and the 2d cnn. The only differences are in the structure of the cnns, the number of training epochs and that before training the 2d cnn the loaded statistical features must first be used to create the synthetic images. The 1d cnns are trained for .. epochs while the 2d cnns are trained for ... epochs **TODO: epochs nachgucken**. Both scripts load the same statistical features, already divided into 20 splits. As there are statistical features for 20 real and 20000 simulated no obstacles games as well as the same amount for glare effects games, this results in each split containing data from 19 real and 19000 simulated games for each of the two classes, resulting in 38 real and 38000 simulated games available for the training. From 2 real and 2000 simulated games not used for training, only the data from the real games is used for testing. This is done because it is not desirable that the models adapt to the simulated data as in reality they will be only used in real games. Furthermore it can be chosen for which ratios between simulated and real games models should be trained. It was chosen to use ratios of one real game to 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 20 simulated games. Moreover it can be chosen how many turns in the game should be used (a turn consists of flipping 2 cards). In real world scenarios, the less the faster the decision is made and assistance can be applied, but if waited the longer the decision will likely be correct more often. To analyse this, models for different amounts of turns are trained. The turns chosen are the first 5, 10, 15 and 20. More than 20 turns were not recorded. **TODO: Vielleicht rein: Training models for ranges of turns in the middle, for instance from turn 10 to 15, was left out, because this is not very interesting when considering real world usage**. All models are trained twice. Once on the simulated data created before modifying the simulator and once afterwards. This can potentially show whether the changes improved the classification results.

The structure and hyper parameters of the models for all 1d cnns are identical. Same goes for all 2d models. The only difference is the input shape of the models since it depends on the number of steps used in the training. Identical structure for models of the same type are used so that different training results for different amount of steps can be reliably compared. It should be noted the structure and hyper parameters are not the same between the 1d and 2d cnns.

Once one of the scripts is executed, required directory structures are created and training processes for the different ratios of real to simulated games are performed

successively. The training for one ratio consists of multithreaded **TODO: multithreaded or multiprocessing?** training of 20 splits and the training for each split is repeated 20 times. As a result for each ratio 20 models are created for every split. This means that 400 models are trained for every ratio. As there are 12 different ratios trained, this results in 4800 models trained for each execution of one of the scripts. As mentioned above, the training is done for 4 different amounts of steps for the 1d cnn and the 2d cnn. This adds up to 38400 models being trained. Furthermore this is done twice: Once before the changes to the simulator and once afterwards. All in all this results in 76800 models. For each model the training history is saved in a file. Saving and loading the history is necessary due to the parallel training of multiple models through multithreading **TODO: or multiprocessing**. Once the training of the 4800 models of one script is completed, the training histories are loaded from the files. The results from the 400 training histories for each ratio of real to simulated games are averaged, plotted and the final figures are saved for the analysis. The reason for averaging the results over many models is, that deep neural networks by default initialize their weights randomly, resulting in potentially different outcomes of each training. By averaging the results, the values and therefore the interpretation becomes more reliable and meaningful. All models are trained on the cpu. The fact that 76800 models are trained lead to the decision, not to save all models but instead first evaluate the results and then retrain and save the models that create the best results.

## 8.4 Training results and evaluation

The best accuracies for each number of turns in each table are highlighted in green. If multiple ratios of real to simulated game for a certain number of turns produce the same accuracy only the one with the smallest amount of simulations is highlighted.

Table 8.3: 1D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps. Before the changes to the simulator.

	5 r.		10 r.		15 r.		20 r.	
	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss
sd0x	65.12%	0.6584	81.12%	0.5116	78.37%	0.5377	79.62%	0.492
sd1x	71.25%	0.6385	80.88%	0.4777	77.25%	0.5355	80.12%	0.4999
sd2x	68.5%	0.6476	77.38%	0.5283	78.25%	0.5275	81.25%	0.4887
sd3x	69.38%	0.6494	75.5%	0.5252	76.13%	0.5329	79.75%	0.4999
sd4x	67.5%	0.6436	75.88%	0.538	77.25%	0.5322	81.25%	0.4975
sd5x	72.25%	0.6212	76.5%	0.5252	77.75%	0.5338	81.25%	0.4985
sd6x	71.87%	0.6346	74.75%	0.547	76.25%	0.5376	79.38%	0.5075
sd7x	70.25%	0.6252	73.62%	0.5395	76.0%	0.5326	80.0%	0.5078
sd8x	74.25%	0.6121	72.5%	0.5509	77.0%	0.535	79.75%	0.4993
sd9x	74.88%	0.5826	74.63%	0.5483	76.0%	0.542	81.75%	0.4987
sd10x	74.87%	0.5971	74.5%	0.5558	77.12%	0.5394	80.87%	0.5036
sd20x	73.25%	0.5831	73.5%	0.5493	75.63%	0.5355	80.75%	0.5013

- interessant: 1d cnn 5 rounds gucken ob sich der trend weiter führt und sogar noch besser wird

Table 8.4: 2D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps.  
Before the changes to the simulator.

	5 r.		10 r.		15 r.		20 r.	
	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss
sd0x	63.5%	0.6471	74.0%	0.4991	79.62%	0.4726	79.0%	0.4778
sd1x	68.12%	0.619	76.62%	0.5171	79.5%	0.4984	81.75%	0.4735
sd2x	68.25%	0.6382	80.0%	0.5243	80.38%	0.4851	84.12%	0.4573
sd3x	69.5%	0.6392	78.88%	0.5518	80.25%	0.5038	84.62%	0.4718
sd4x	67.63%	0.6415	76.38%	0.5685	79.62%	0.5106	85.0%	0.4726
sd5x	70.88%	0.6339	74.63%	0.5668	80.0%	0.51	84.88%	0.474
sd6x	70.62%	0.638	77.0%	0.5623	80.12%	0.5077	84.88%	0.4776
sd7x	69.75%	0.6543	75.38%	0.5762	79.5%	0.5176	85.0%	0.4839
sd8x	70.0%	0.6467	74.25%	0.5775	80.38%	0.5164	85.0%	0.484
sd9x	71.25%	0.6395	78.75%	0.5696	81.5%	0.5064	85.0%	0.4814
sd10x	68.75%	0.6296	77.75%	0.5674	80.12%	0.5098	85.0%	0.4802
sd20x	73.0%	0.5833	76.63%	0.5486	79.75%	0.5064	84.88%	0.4704

Table 8.5: 1D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps.  
After the changes to the simulator.

	5 r.		10 r.		15 r.		20 r.	
	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss
sd0x	63.87%	0.6698	81.88%	0.5014	78.5%	0.5401	78.62%	0.5107
sd1x	67.75%	0.6648	79.0%	0.5095	75.75%	0.5305	79.38%	0.5101
sd2x	66.0%	0.6506	79.25%	0.4845	76.0%	0.5325	78.37%	0.514
sd3x	64.38%	0.6585	77.38%	0.5037	77.62%	0.5261	79.25%	0.5113
sd4x	67.0%	0.6377	75.38%	0.5289	77.62%	0.5331	80.25%	0.5027
sd5x	67.0%	0.6452	76.75%	0.5146	78.75%	0.5226	79.0	0.5042
sd6x	70.25%	0.6427	75.25%	0.5335	78.75%	0.5208	79.38%	0.5049
sd7x	70.25%	0.6449	76.88%	0.5114	78.0%	0.5129	78.75%	0.5115
sd8x	71%	0.628	76.25%	0.523	77.25%	0.5261	79.12%	0.5156
sd9x	70%	0.6276	75.37%	0.5226	77.62%	0.5281	78.63%	0.522
sd10x	71.5%	0.6242	76.25%	0.5261	76.38%	0.5331	78.62%	0.4986
sd20x	71.88%	0.6246	72.38%	0.5617	76.25%	0.5416	79.12%	0.5

Table 8.6: 2D CNN. Best accuracies and losses for 5, 10, 15 and 20 steps.  
After the changes to the simulator.

	5 r.		10 r.		15 r.		20 r.	
	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss
sd0x	63.62%	0.6461	73.87%	0.4929	80.0%	0.4557	78.5%	0.476
sd1x	67.88%	0.6446	76.12%	0.531	79.5%	0.4911	79.75%	0.4874
sd2x	69.75%	0.6463	79.62%	0.5243	80.87%	0.4905	80.0%	0.4859
sd3x	69.5%	0.6449	77.5%	0.5507	82.25%	0.4994	80.0%	0.4865
sd4x	69.0%	0.6347	72.75%	0.5684	80.12%	0.5156	80.0%	0.495
sd5x	69.38%	0.6342	73.62%	0.5634	80.5%	0.5122	0	0
sd6x	71.0%	0.6326	74.5%	0.5586	80.62%	0.5038	0	0
sd7x	70.0%	0.6328	71.62%	0.555	80.0%	0.5061	0	0
sd8x	69.88%	0.6287	72.75%	0.5567	80.25%	0.5057	0	0
sd9x	69.12%	0.6238	72.38%	0.5532	79.5%	0.5067	0	0
sd10x	69.5%	0.6203	73.25%	0.5536	0	0	0	0
sd20x	69.88%	0.5935	73.0%	0.5494	0	0	0	0

- It should be clarified that the chosen models to be trained don't .. keinen anspruch haben auf vollständigkeit und den idealen punkt zu finden... as training models takes time and the time of this work was limited. The chosen training cycles however to a good job at finding very good results. They might not be the best results possible but they show the capability of cnns in classifying this visual interaction obstacle.

It would be possible to investigate more in potentially promising direction, but this was not possible due to time constraints and must be left open for future work.



In the results show that using a certain number of simulated games for the training can often significantly improve the classification results. While too many simulations can worsen the results. This is evident when comparing for instance the 79% without any simulations and the 85% with 4 simulations per real game of 2d cnns for 20 rounds before the changes to the simulator. **TODO: besseren vergleich findend wo man steigung, höherpunkt und verschlechterung sieht** ... show plots for both ... In the plots can also be seen that this difference in performance is not based on the fact that for one of the configurations the accuracies reached convergence while not for the other one.

beobachtung: bei 2d sind simulationen länger besser. Analysieren ob es daran liegt dass die anderen keine convergence erreichen. Wenn nicht dann ist es coole entdeckung. Weil es ein anzeichen dafür ist, dass die 2d cnn besser mit simulationen lernen kann als 1d cnn. This also fits to the observation, that the results from 1d cnn with more than 10 rounds don't get much better if not worse. This can be caused by the fact that the simulations are less accurate after the first 10 steps, and therefore the 1d cnn does not improve by including them. However, the 2d cnn does improve using the last steps. This is an additional indicator, that the 2d cnn can better handle simulated games. A possibility is, that the difference between simulated and real games in the synthetic images created using the feature is not as significant as when directly using the statistical features in the 1d cnn. **TODO: noch mit zahlen aus den tabellen begründen. und sagen dass die alle convergence erreichen, weshalb man das sagen kann**

- easy to believe that in some cases more the optimal ratio has not been reached yet, if only looking at the tables. However there are more factors that influence the result. This becomes evident when plotting the training histories for accuracy and loss. ... plott wo man sieht das divergence nicht erreicht wurde... The fact that the trainings with a lower ratio result in worse accuracy don't necessarily are caused by the fact that more simulated games would be better, but could also be influenced by the fact that the divergence has not been reached yet, meaning that either a different learning rate or more epochs could increase the accuracy. Therefore it can not be said with certainty that more simulated games would in that case lead to the optimal result, as the optimal result could also be with less simulated but with changed hyper parameters. However, the hyper parameters were intentionally chosen the same so that the results could be compared with each other.

- All models used to calculate the best mean entries in the two tables for the results before the changes to the simulator are retrained and saved. This is only done if they are the best results regarding the number of steps. This means that for the 1d cnns the models for 5 rounds and sd9x and for 10 rounds and sd0x are saved. For the 2d cnn the models for 15 rounds and sd9x and for 20 rounds for sd4x are saved. This results in  $400 \cdot 4$  models. These 1600 models are used in the voting based system explained in chapter **TODO: ref**.

- it was considered to perform statistical tests comparing different models and ratios to reveal significant or insignificant differences. However, the decision was made that this does not really give reliable results, because the values in the tables **TODO: ref** don't tell the whole story. As mentioned above not all configurations reach convergence meaning the value in the table don't always represent the best accuracy achievable with that configuration. If parameters such as the learning rate or the

number of epochs are not chosen equal across all models of one table (meaning all models for 1d cnns, etc) and instead would be chosen more individually to optimize the models specifically for each number of rounds and ratio between simulated and real games, the accuracy could be significantly different. Therefore the results of statistical tests performed on the achieved accuracies in the tables, would not justify any statements regarding whether 1d cnns or 2d cnns perform significantly better in different configurations. Such statistical test would only show significant differences or indifferences in results, but could not be interpreted for any valuable findings, as some models are used to their full potential converge while others are not and therefore don't converge.

- ..table showing best result for all categories before and after training (16 values) It can be seen that the results after the changes to the simulator are noticeably worse than before. The reason for this is unknown and can be highly complex, due to the many factors involved. A theory is, that reducing the randomness of the simulator and therefore making the simulated games more similar to the real games, may have resulted in decreased generalisation in the models and therefore less robustness on unknown data. However, this theory can hardly be verified. The decreased performance after the changes leads to the decision to only use the models trained on the simulated data created before the changes to the simulator.

.. table for best results for the different number of steps and what their config was (4 values).. It can be seen in table **TODO: ref** that the 1d cnns trained, perform better for 5 and 10 rounds than 2d cnns while 2d cnns perform better for 15 and 12 rounds. Depending on the number of rounds it can therefore be beneficial to use either 1d cnns or 2d cnns.

- kurz erklären wieso weniger züge besser sind bei dieser hci erkenntung
- statistical tests um signifikante unterschiede /keine unterschiede zu zeigen für verschiedene modelle etc (siehe mazens nachricht) ?

## 9. Voting based system

The accuracies shown in section **TODO: ref** are not representative of how the models would be used in real life scenarios. Since the accuracy is the mean accuracy of 20 repetitions for each split there is no single model that produces that accuracy. For using the models in real life scenarios a voting based system is deployed. It consists of multiple models classifying the input data and then voting for the final classification. The accuracy of such a system more accurately describes the real world performance. Therefore a voting based system is created each for the best CNNs and ratios of simulated to real data for each number of step. As mentioned in section **TODO: ref** the 1d CNN creates the best accuracy for 5 and 10 rounds while the 2d CNN creates the best accuracies for 15 and 20 rounds. Hence, in total 4 voting based systems are created, one for each of the 4 number of rounds. Each of the 20 repetitions of a split is tested as before, only with the two real games (one no obstacle and one glare) that were not included in the training. However instead of directly calculating the accuracy of each repetition of that split, the output labels are saved for all repetitions and used to perform a voting for the final label. Meaning that if for instance 15 out of 20 repetitions classified the game as a glare effect game, while the other 5 classified it as a no obstacle game, the final classification will be that of a glare effect game. This is done for the two test games. Using the final labels created by the voting, the accuracy is calculated. In the example above, if assumed that both final labels would be correct, the accuracy of that split would be 100%. This procedure is repeated for each of the 20 splits and the accuracies are averaged.

.. table of accuracies of the 4 voting based systems ...

The voting based system produces noticeably increased accuracies. ... analyse...

It would be possible to use these four systems during a game. Depending on the number of step a different system is used. In reality when using these systems, the accuracy produced by the voting could be even more improved. The reason for that is that the voting can be done by all 400 models instead of only 20 for each split. When testing the voting was done for each split independently because the other for each split being tested, the test data was used in the other 19 splits for training. This is not the case in real world scenarios when the game is being played.



# 10. Conclusion

85 % etc.

## 10.1 Prospect

- es wurde zwar implementiert nicht alle features zu verwenden, aber es wurde nie ausprobiert. Hätte man testweise machen können.
- Vielleicht hätte man ein weiteres statistical feature für penalties machen können, so wie es beim qualitätscheck der simulationen berechnet wird.
- statistical tests could be used to validate whether the changes to the simulator improved the quality of the simulations. However, as shown by the classification results even though they seemed to improve the quality or at least match it, the classification results were noticeably worse when using the simulated games created by the modified simulator. There are many variables and factors that influence the final result, which makes it hard to validate whether changes to the simulator improved the overall results, even when using statistical tests.
- a user study could be done to validate the voting based systems in real scenarios. As mentioned in chapter **TODO: ref** the results shown in table **TODO: ref zu table mit voting ergebnissen** could be even more improved in real scenarios because more models are able to participate in the voting.
- man hätte ausgehend von initialen trainingsergebnissen noch mehr trainieren können in die richtigen, aber wegen zeitgründen nicht gemacht



# 11. Appendix

gucken wie man appendix richtig macht in latex





# Bibliography

- [Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.



# Index

example intro, [3](#)