Generating ROC Curve Using Confidence Ratings for Memory Assessment

Student: Nijat Aghayev

Professor: Sen Cheng

Course Name: Mathematical Psychology

1. Abstract

In this report we study generating ROC curves for recognition memory assessment and assessing recognition memory of participants based on these generated ROC curves. Similar experiments have already been done before, but in most of these experiments recognizing numbers or text have been studied. Of course our experiment is a little bit more difficult for studying memory. In discussion section of this report difficulties and drawbacks of the experiment have been highlighted. In methods section of the report the entire experiment has been described and in results section the collected data has been analyzed and evaluated.

There is difference between recognizing an item and remembering an item. Remembering an item is much more difficult and always is preceded by recognizing an item. For example, when we remember somebody, we remember a lot of key features of this person, but when we recognize somebody, we only think that we have already seen this person , but we don’t know his/her name, or other key characteristics. In this experiment we study the recognition part of the memory.

2. Introduction

We generate ROC curves , in order to study recognition memory. ROC curves have been developed in Signal Detection Theory(STD) , in order to measure quality of experiment and SDT was firstly applied in the military , but later this technique became a part of psychological experiments. ROC curves enable us to see the difference among memory capabilities of different individuals.

We rely on the introspective judgements of participants while generating ROC curves , in the experiment participants have to rate their confidence level for every picture. Historically different dependent variables, including response time , latency of heart rate increase, response frequency, and firing rates of individual neurons, have been used for ROC construction. Each of these methods has its drawbacks, including our method of using introspective judgement. For example, when response time is used for ROC construction , several factors can affect response time: Maybe the participant is not comfortable with the specific keyboard layout, or maybe the chosen response keys are not handy to the participant, or maybe the participant does not perform well under time pressure. And drawbacks of introspective judgements could be how to make sure that the participant is not lying about his confidence level.

3. Methods

3.1. Participants

We could construct experiment in three ways:

1) Experiment consists of 1 session . In the first phase of this session 50 pictures are shown to the participants in 1 second time interval , and in the second phase of this session 100 pictures are shown to the participants , whereas 50 of these pictures are "new" and the other half is "old". Participants have to decide

which pictures are new and which pictures have already been shown in the first phase. Then in the third phase the participants are asked to rate their confidence level for each picture and at the end of the experiment ROC curve is constructed based on these confidence levels. There are in general 5 confidence levels for user chosen new pictures , and 5 confidence levels for user chosen old pictures (5 means sure , 1 means not sure, 2 means probably not sure , and so on).

In this experiment we would need multiple participants in order to compare AUC of their ROC curves with one other and answer some scientific questions . E.g. in one scenario ROC curves of 2 men and 2 women, in another scenario ROC curves of 1 teen, 1 adult and 1 elderly, in another scenario ROC curves of 2 Alzheimer's patients and 2 healthy individuals, and so on.

2) Experiment consists of multiple sessions . Each session is exactly like the session described in the first experiment above and different categories are used for different sessions.

In this experiment we would need only one participant. And we could compare different ROC curves , which were generated in different sessions and try to scientifically substantiate differences among these ROC curves.

3) This experiment is hybrid combination of the first and the second experiment , which means multiple participants are involved into the experiment and each experiment consists of multiple sessions.

Because of my location constraints I had to choose the second experiment.

3.2. Detailed description of the chosen experiment

This experiment consisted of 3 sessions . The categories of sessions were human faces for the first session , spiders for the second session and cars for the last session. These 3 categories were chosen on purpose , human faces are important for socializing and survival , spiders can be perceived as threatening creatures (at least for me that is the case) , and cars can be perceived as neutral objects (again at least for me). Here an important point to understand is that categorizing objects and creatures as threatening, neutral or positive are purely subjective , e.g. some people don't have fear of touching spiders , and some people have negative connotation to cars , because they had bad car accident in the past and lost valuable people in their life.

4. Results

4.1. Assessing the data from the first session (category: human faces)

Table 1. Number of responses in the first session of this recognition memory experiment broken down by percentiles of a measure(such as confidence ratings) qualifying the classification response.(The sums (∑) of new and old responses are indicated for the different stimulus classes along with the overall number of targets and lures. The first set of rows(labelled 'RAW') contains counts of the number of responses. The middle set of rows(labelled 'ALL') contains the same data normalized by the respective total numbers of target and lure trials.)

		new						**old**						
percentile		100	80	60	40	20	∑ new	100	80	60	40	20	∑ old	∑
RAW	**lure**	10	9	8	7	5	39	4	2	3	1	1	11	50
	target	2	1	2	3	5	13	2	3	9	10	13	37	50
ALL	**lure**	0.2	0.18	0.16	0.17	0.1	0.78	0.08	0.04	0.06	0.02	0.02	0.22	1
	target	0.04	0.02	0.04	0.06	0.1	0.26	0.03	0.06	0.18	0.2	0.26	0.74	1

Table 2. Cumulative hit and false alarm(FA) rates for different strength criteria inferred from the data in table 1. (The top set of rows(labelled 'RAW') contains the raw frequencies and the bottom set of rows (labelled 'ALL') contains the same data normalized by the total number of targets(for hits) and lures(for FAs).)

		Strength Criterion										
		Liberal			**……………………………………………**							**Conservative**
RAW	**FA**	50	40	31	23	16	11	7	5	2	1	0
	Hit	50	48	47	45	42	37	35	32	23	13	0
ALL	**FA**	1	0.8	0.62	0.46	0.32	0.22	0.14	0.1	0.04	0.02	0
	Hit	1	0.96	0.94	0.9	0.84	0.74	0.7	0.64	0.46	0.26	0

Figure 1. Number of responses (in the first session) in 5 different confidence levels for 2 sets of pictures (user chosen new pictures and user chosen old pictures)

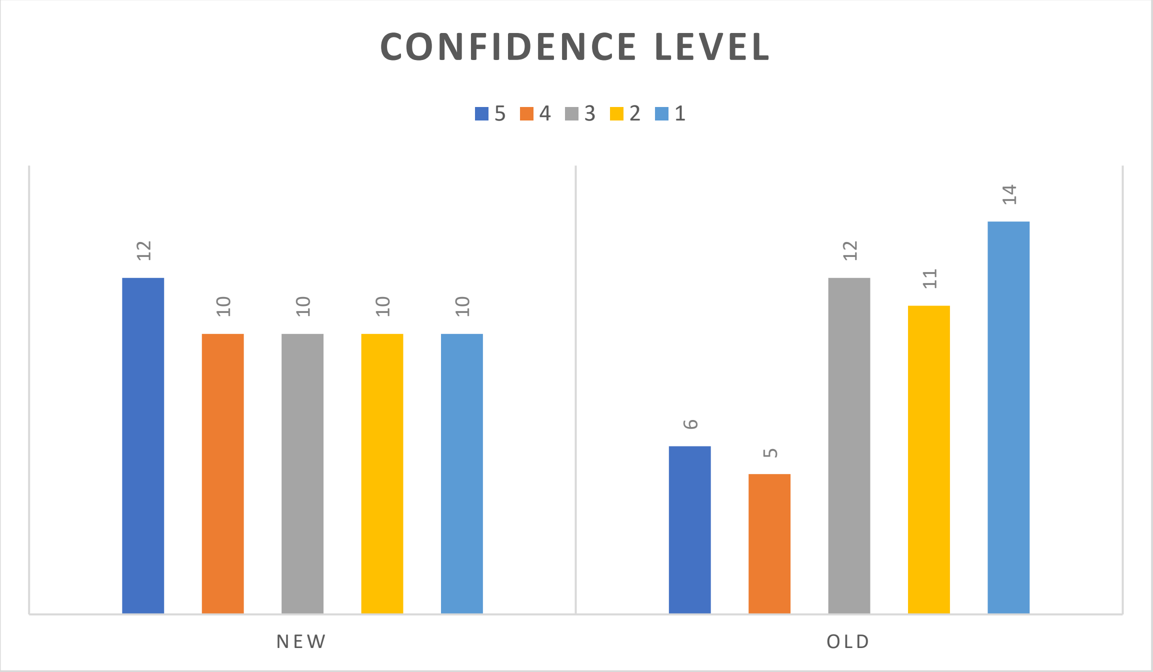
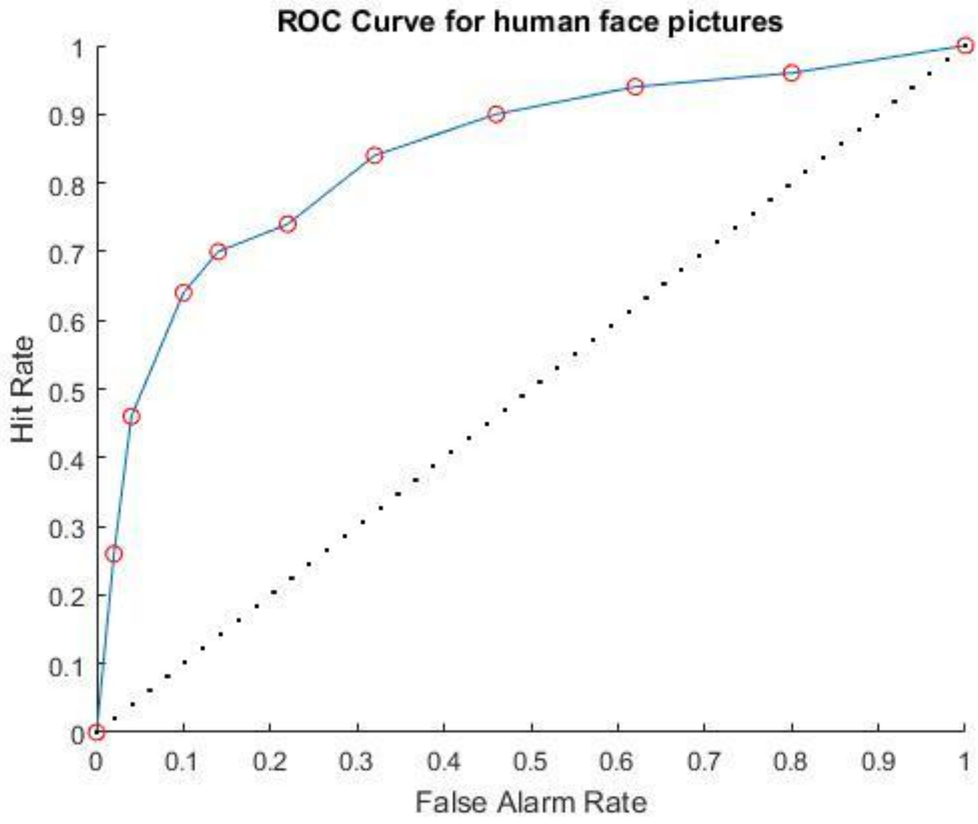


Figure 2. ROC Curve generated based on user responses in the first session. (which corresponds to the human face category.)



4.2. Assessing the data from the second session (category: spiders)

Table 3. Number of responses in the second session of this recognition memory experiment broken down by percentiles of a measure(such as confidence ratings) qualifying the classification response.(The sums (∑) of new and old responses are indicated for the different stimulus classes along with the overall number of targets and lures. The first set of rows(labelled 'RAW') contains counts of the number of responses. The middle set of rows(labelled 'ALL') contains the same data normalized by the respective total numbers of target and lure trials.)

		new						**old**						
percentile		100	80	60	40	20	∑ new	100	80	60	40	20	∑ old	∑
RAW	**lure**	11	10	8	6	5	40	3	2	2	2	1	10	50
	target	2	3	3	4	7	19	5	6	6	7	7	31	50
ALL	**lure**	0.22	0.2	0.16	0.12	0.1	0.8	0.06	0.04	0.04	0.04	0.02	0.2	1
	target	0.04	0.06	0.06	0.08	0.14	0.38	0.1	0.12	0.12	0.14	0.14	0.62	1

Table 4. Cumulative hit and false alarm(FA) rates for different strength criteria inferred from the data in table 3. (The top set of rows(labelled 'RAW') contains the raw frequencies and the bottom set of rows (labelled 'ALL') contains the same data normalized by the total number of targets(for hits) and lures(for FAs).)

		Strength Criterion										
		Liberal				**……………………………………………….**						**Conservative**
RAW	**FA**	50	39	29	21	15	10	7	5	3	1	0
	Hit	50	48	45	42	38	31	26	20	14	7	0
ALL	**FA**	1	0.78	0.58	0.42	0.3	0.2	0.14	0.1	0.06	0.02	0
	Hit	1	0.96	0.9	0.84	0.76	0.62	0.52	0.4	0.28	0.14	0

Figure 3. Number of responses (in the second session) in 5 different confidence levels for 2 sets of pictures (user chosen new pictures and user chosen old pictures)

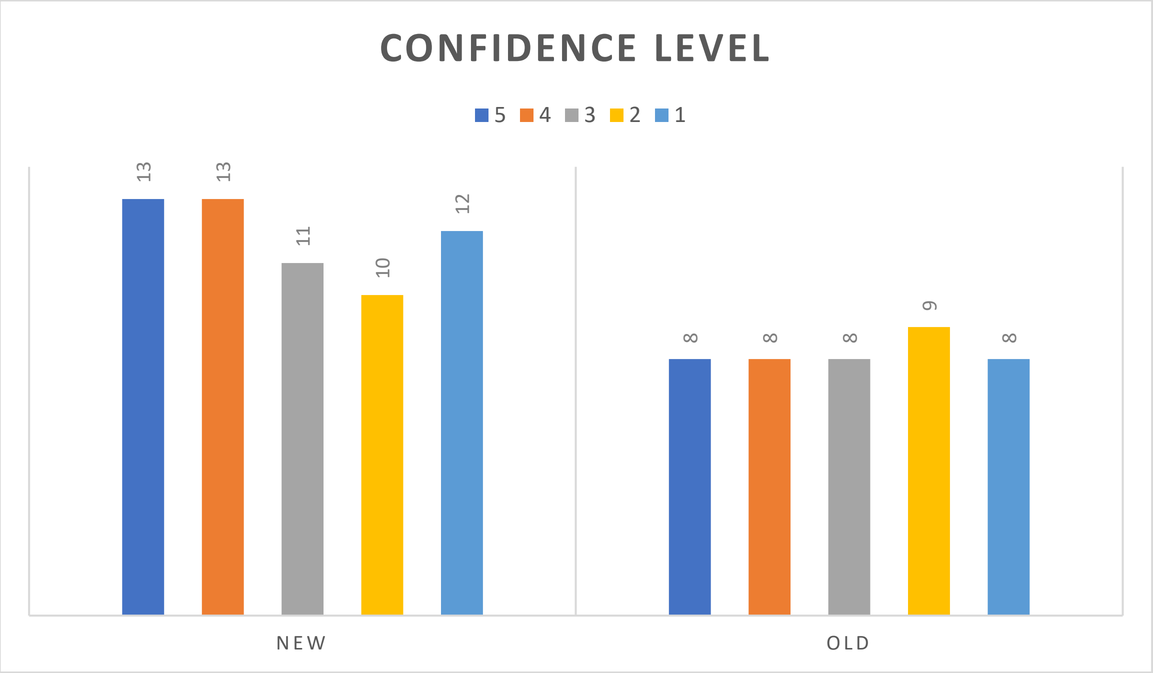
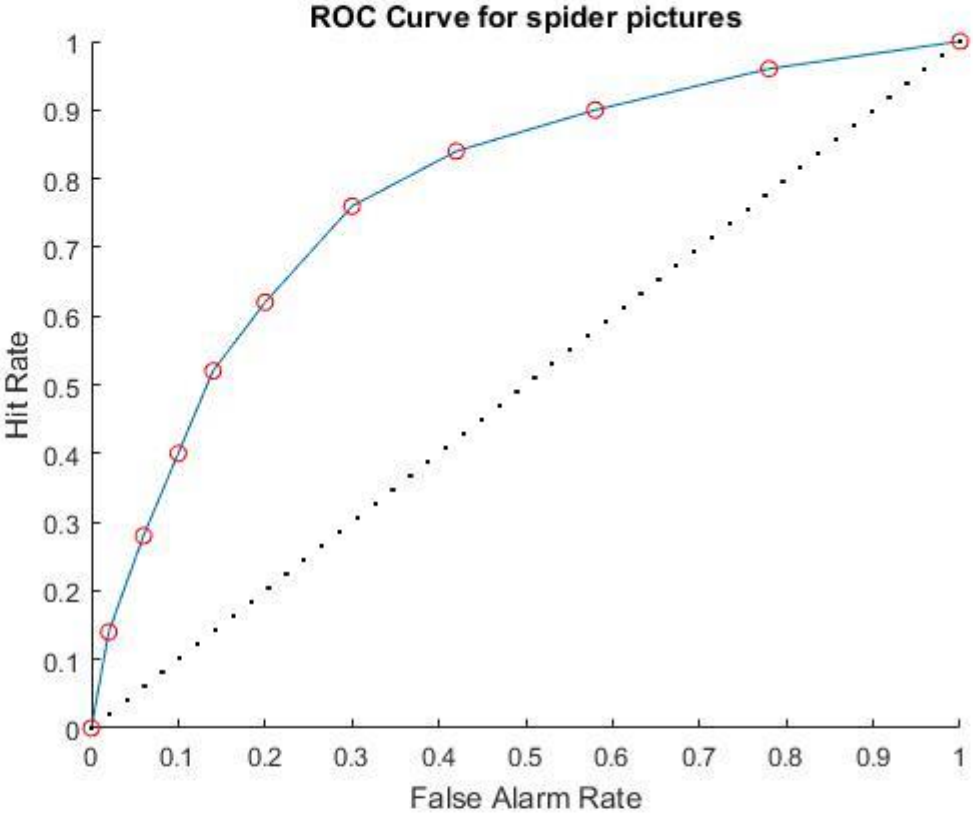


Figure 4. ROC Curve generated based on user responses in the second session. (which corresponds to the spiders category.)



4.3. Assessing the data from the third session (category: cars)

Table 5. Number of responses in the third session of this recognition memory experiment broken down by percentiles of a measure(such as confidence ratings) qualifying the classification response.(The sums (∑) of new and old responses are indicated for the different stimulus classes along with the overall number of targets and lures. The first set of rows(labelled 'RAW') contains counts of the number of responses. The middle set of rows(labelled 'ALL') contains the same data normalized by the respective total numbers of target and lure trials.)

		new						**old**						
percentile		100	80	60	40	20	∑ new	100	80	60	40	20	∑ old	∑
RAW	**lure**	8	7	5	3	4	27	5	6	5	4	3	23	50
	target	3	3	4	4	4	18	10	5	6	7	4	32	50
ALL	**lure**	0.16	0.14	0.1	0.06	0.08	0.54	0.1	0.12	0.1	0.08	0.06	0.46	1
	target	0.06	0.06	0.08	0.08	0.08	0.36	0.2	0.1	0.12	0.14	0.08	0.64	1

Table 6. Cumulative hit and false alarm(FA) rates for different strength criteria inferred from the data in table 5. (The top set of rows(labelled 'RAW') contains the raw frequencies and the bottom set of rows (labelled 'ALL') contains the same data normalized by the total number of targets(for hits) and lures(for FAs).)

		Strength Criterion										
		Liberal		**……………………………………………**								**Conservative**
RAW	**FA**	50	42	35	30	27	23	18	12	7	3	0
	Hit	50	47	44	40	36	32	22	17	11	4	0
ALL	**FA**	1	0.84	0.7	0.6	0.54	0.46	0.36	0.24	0.14	0.06	0
	Hit	1	0.94	0.88	0.8	0.72	0.64	0.44	0.34	0.22	0.08	0

Figure 5. Number of responses (in the third session) in 5 different confidence levels for 2 sets of pictures (user chosen new pictures and user chosen old pictures)

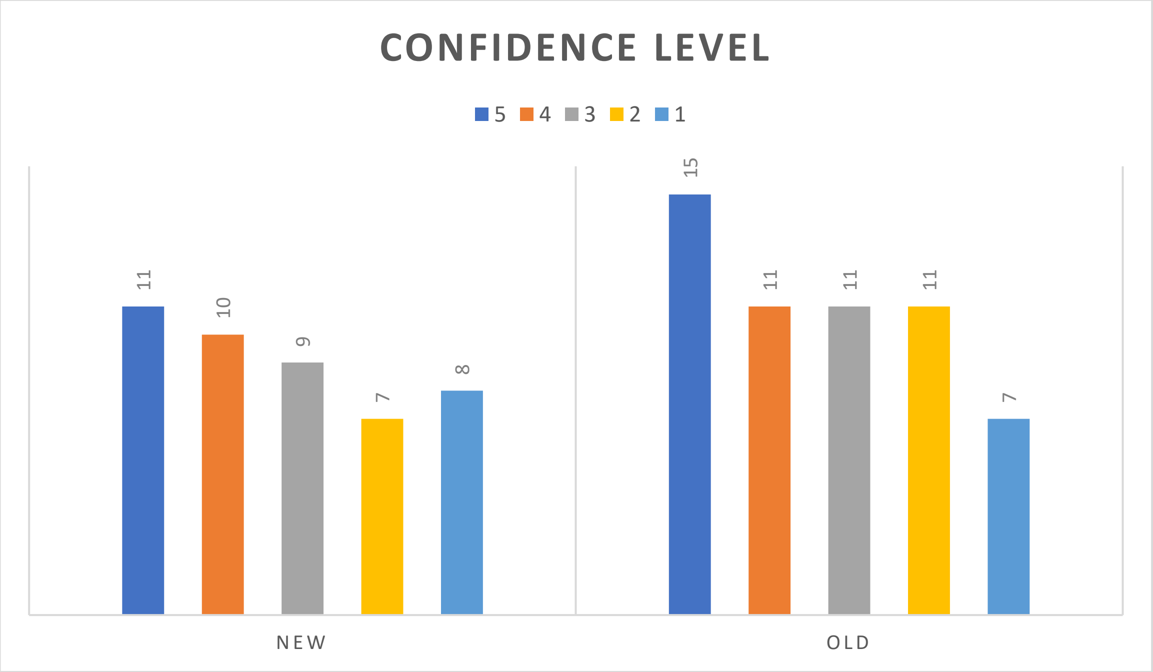
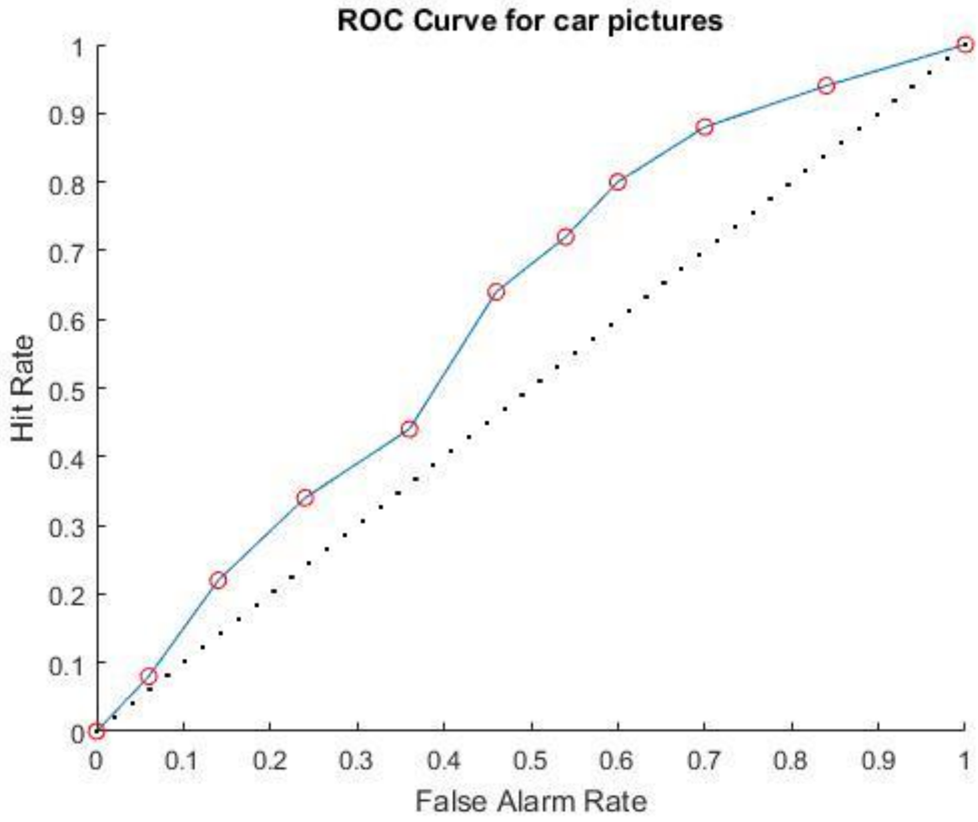


Figure 6. ROC Curve generated based on user responses in the first session. (which corresponds to the cars category.)



4.4. Comparing the ROC curves of these 3 sessions

Figure 7. Comparison of the ROC Curves generated in 3 sessions.

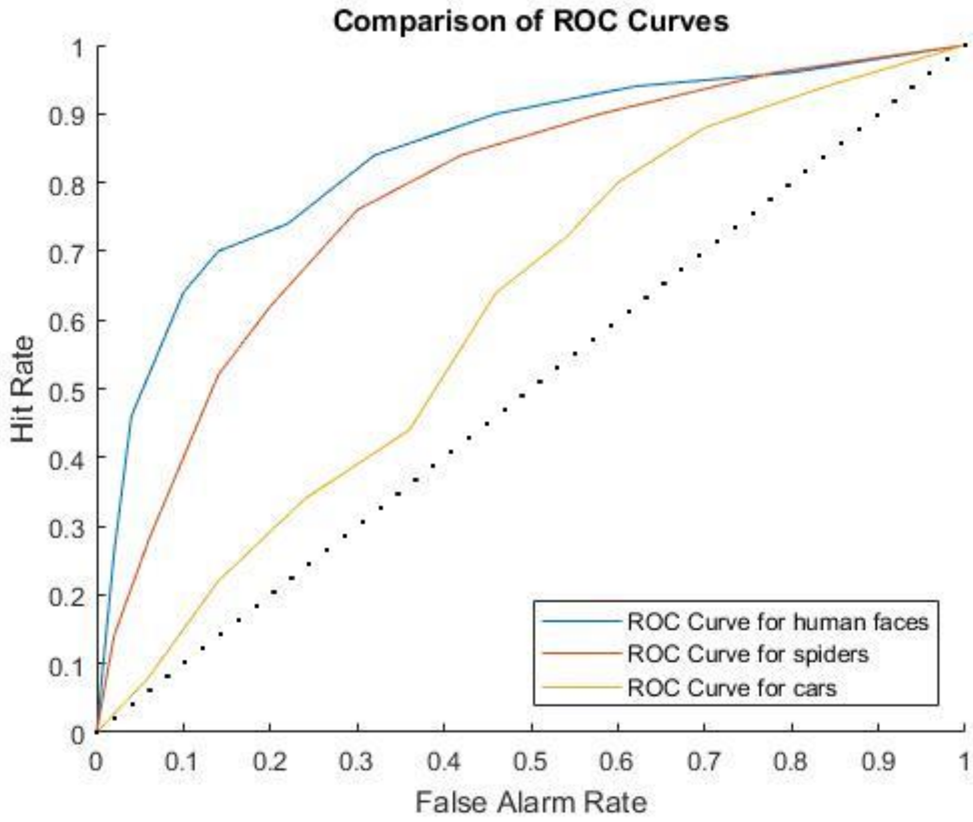
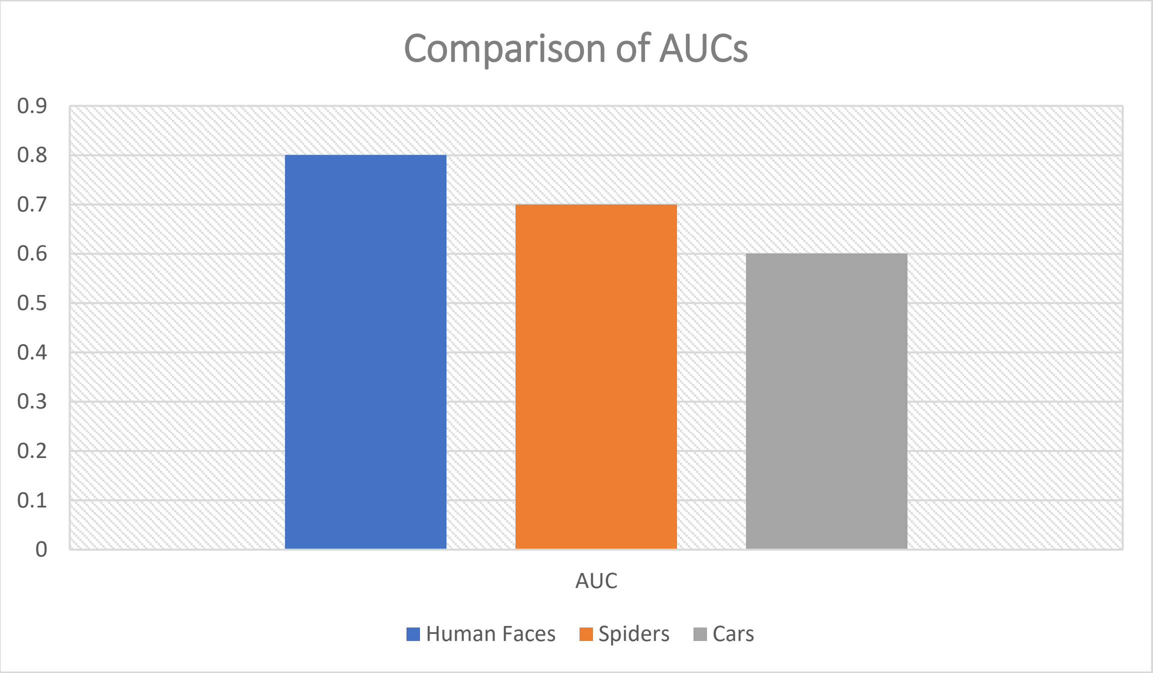


Figure 8. Comparison of Area Ander Curve (AUC) of ROC Curves generated in 3 sessions.



The reason why my result was good at the human faces category can be explained with the fact that we humans are social creatures , indeed in the tribal periods of the mankind, ostracizing individuals from the tribe was a death penalty and that is why they had to socialize with others in order to find food , recreate and protect themselves from other tribes and wild animals. My second best result was at the spiders category, the reason behind this is related to negative emotions that are produced when we encounter with danger, and this negative emotion enhances memory accuracy [3]. My worst result was at cars category , because I perceive cars as neutral objects , and cars are neither important for my survival nor they create negative emotions in me.

5. Discussion

We have used pictures in order to study recognition memory , most experiments of this kind were carried out with numbers or texts. Using pictures brings a lot of challenges.

When using human face pictures we have to make sure that it will not be easy to differentiate one picture from others easily , let's say if in 99 pictures none of human face has eyeglasses , and there are eyeglasses in the 100th picture , then it will be easy to recognize this last picture . Or when 99 pictures belong to males , and 1 belongs to a woman , it will be easy to recognize the picture, which belongs to a woman. Or when 99 pictures belong to white people , and 1 belongs to a person from other ethnicity , it will be easy to recognize that one picture . I have tried to avoid all these scenarios , no human face picture with eyeglasses has been used , human faces from different races and genders have been included , all pictures' pixel size is the same , and all pictures have been presented to the participants as grayscale pictures in order to avoid noises that can be caused by colors.

In the second category spiders from related spider families have been chosen , and again all pictures have been presented as grayscale pictures and have the same pixel size.

In the third category car profiles(from right angle) have been chosen , and backgrounds of these pictures are almost identical. Again all pictures have been presented as grayscale pictures and have the same pixel size.

Despite of the limitations that were described above, this experiment shed some lights on the complex topic : recognition memory . With this experiment we can conclude that not only recognition memory of individuals is different , but recognition memory for different categories is also different . And another conclusion from this experiment would be, when items, creatures that are important to our survival , then our memory tends to work more efficiently , than the case , in which items and creatures are perceived by us as neutral.

References

1. Sofus A. Macskassy, Foster J. Provost, Michael L. Littman. Confidence Bands for ROC Curves.

2. Sofus A. Macskassy, Foster Provost. Confidence Bands for ROC Curves: Methods and an Empirical Study.

3. Elizabeth A. Kensinger. Negative Emotion Enhances Memory Accuracy.Behavioral and Neuroimaging Evidence.

4. Caren M. Rotello. 2016. Signal Detection Theories of Recognition Memory.

5. Weidmann CT, Kahana MJ. 2016. Assessing recognition memory using confidence ratings and response times. R.Soc. openm sci. CTW/0000-0002-4280-2744.

6. Labar, K.S., & Cabeza, R.2006 . Cognitive neuroscience of emotional memory. Nature Reviews Neuroscience

7. Charles E. Metz, Benjamin A. Herman, and Jong-Her Shen, 'Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data', Statistics in Medicine

8. Baranski JV,Petrusic WM.1998 Probing the locus of confidence judgments: experiments on the time to determine confidence.J.Exp.Psychol.Hum.Percept.Perform.24

9. Heeger D. 1997.Signal Detection Theory

10. Zhu W, Zeng N, Wang N. 2010. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations

11. Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27.

12. Tamim H. Receiver Operation Characteristics (ROC) Curve.

13. Cortes C, Mohri M. Confidence Intervals for the Area under the ROC Curve

14. Fawcett T. 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers