



A Study for Correlation Identification in Human-Computer Interface Based on HSB Color Model

Yikang Dai, Chengqi Xue^(✉), and Qi Guo

School of Mechanical Engineering, Southeast University,
211189 Nanjing, China
ipd_xcq@seu.edu.cn

Abstract. In recent years, visual perception has been paid more attention by many researchers in the field of data visualization. The study of visual perception has become another research hot spot in the research of visualization. As an important tool for visualization, this paper focuses on the scatterplots. Series of scatterplot were generated by programming and then used in the experiment. The results of experiments indicate that the influence of the color, amount and correlation of the interference points on the reaction time is significant under the white background and suitable combination is found which is important for designing the scatterplots.

Keywords: Visual perception · Scatterplots · Correlation identification

1 Introduction

Human-computer interface which is used to transfer and exchange information is the medium and dialogue interface between human and computers. It is an important part of the human-computer system which converts the internal form of information into an acceptable one to users. Status of the system can be understood by users so that they can monitor the output of it while running and adjust it at any time due to the human-computer interface.

With the development of computer and network technology, the amount of information in human-computer interface increases sharply. Massive amounts of data are generated by various kinds of devices, sensors, electronic websites and social networks every day, triggering an explosive growth in data size. As a result, the concept of ‘Big data’ appears. Lately, the term ‘Big Data’ tends to refer to the use of predictive analytics, user behavior analytics, or other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Big Data has 5 features (named 5 V): Volume, Variety, Velocity, Veracity and Value of high density, it brings new opportunities and challenges for human. Nowadays, big data has become a hot topic in academic research and is considered to be another revolutionary information technology after cloud computing and the Internet of Things.

2 Background

Scatterplot which has also been called scatter diagram, scatter gram, and scatter graph was first introduced in the 18th century, during the boom of statistics graphics [1]. There are several variations and extensions, including animated scatterplot, scatterplot matrix [2], and glyph-based scatterplot [3]. Scatterplots can be used for observing correlation, clusters, and outliers [4, 5].

As one of the visualization tool, scatterplots have many advantages in terms of displaying data sets in high dimension: It can quickly preview and analyze data sets, a variety of data points are applicable. Both continuous data set and discrete data points can be visualized by using it; In scatterplots, users can observe the distribution of data points and obtain the overall information through the way of dimensionality reduction of data. Moreover, they can also estimate the correlation among data plots to make judgment in order to provide better service; Scatterplots can easily visualize the data displayed to get useful information. This visualization technology can quickly and easily help users distinguish the abnormal points from all plots. Also, scatterplots play an important role in ensuring the correctness of data. In summary, scatterplots visualization is a valuable tool for data visualization.

Cartesian coordinates are used in scatterplots to present multi-dimensional data in a two-dimensional plane. Information which implied in the data can be clearly revealed in this way. Linear association in the scatter-plots is the most basic and simple relationship between variables, and at the same time, statistical analysis uses correlation to quantify the strength and direction of such bivariate linear associations [9]. Throughout the method above, not only can the relationship between variables display visually, but users can observe the overall information of the data. As a result, using scatterplots has become an effective way to visualize data.

If the graphical representation is designed well, data can be analyzed rapidly, accurately, and precisely. In such situations, the way of analyst's visual system perceiving structure in a dataset is the same way as it perceiving structure in the real world. Therefore, the perception of such graphical representations has considerable potential to help researchers investigate various aspects of our visual intelligence [6–8].

In recent years, visual perception has been paid more attention by many researchers in the field of data visualization. The study of visual perception has become another research hot spot on visualization.

In statistical analysis, R (Pearson's product-moment coefficient, PPMC) is commonly used to define the correlation between variable [9]. In a specific data set, R equals to

$$R = \frac{S_{xy}}{S_x S_y} \quad (1)$$

For x and y are the sample data for the two variables in this data set. Where S_x and S_y are the sample standard deviations of x and y , S_{xy} is the sample covariance between x and y . As shown above, it can be seen that r ranges from 0 which means the perfect negative correlation to 1 which means perfect positive correlation [10]. Figure 1(a) and

(b) show the sampling distributions of R is 0.3 and 0.5, respectively, Fig. 1(c) and (d) show the sampling distributions of R is 0.7 and 0.9, respectively.

There are a number of cognitive influences which can affect the human perception of correlation. These may include gestalt laws of grouping, shape interpretation, learned knowledge about statistical measures such as Pearson's product-moment correlation coefficient (PPMCC). In recent years, there has been growing interest in studying the underlying models of human perception of correlation.

Sher et al. [11] found that for different PPMCC values (R value), data distribution had an impact on the average offsets which meant the differences between the estimated and actual PPMCC, and result showed that only large variations in density caused a statistically significant impact.

Li [12] studied the symbol size perception of scatterplots in the context of analytic tasks which required size discrimination. They conducted an experiment in three visual analysis tasks represented by a circle, divided into three groups with 8 linearly varying radii. 24 subjects were participated in this experiment. The result showed that approximate uniformity of size perception existed in complex tasks and could be described by power-law transformation with an index of 0.4.

Li et al. [9] observed correlation as a function of the sample correlation under different visualization methods, sample sizes and observation time. In the study, they introduced a discriminating index to characterize performance under different conditions. Furthermore, they came to a conclusion that users could reliably differentiate two different degrees of relevance when using scatterplots and using PCP.

Micallef et al. [13] in order to automatically set parameters to enhance the visual quality of the scatterplot, they studied the use of perceptual models and quality metrics. They took the construction of cost function as the key consideration to capture several relevant aspects of the human visual system and they used different experiments to test different data analysis task in a scatterplot design.

Gleicher et al. [14] used a realistic task which evaluated the difference in class means in a scatterplot. And in the end, they explored the assemble judgment in visualizations using.

Stenholt et al. [15] explored the visualization of 3D scatterplots in an immersive virtual environment. The results showed that CVA glyphs did better understand shapes in 3D scatterplots than regular perspective glyphs, especially when there were large numbers of clutter. In addition, their assessment showed that the perception of structure in the 3D scatterplot was affected by the volumetric density of the glyphs in the figure significantly.

Bertini et al. [16] in order to reduce cluttering in the scatterplots, they presented a strategy which relied on a combination of non-uniform sampling and pixel displacement. The paper mentioned that it could also be used to define the precise quality measurement which allowed for validating their approach.

Rensink [17] used four different data point distributions to test. They found that JND was a linear function of distance from $r = 1$, and the perceived relative magnitude was a logarithmic function of that amount. Then the other three cases were checked and the performance was found to be similar in all situations. The results showed that the basis of the correlation perception was not in a geometric structure, such as the shape of

a point cloud, but rather the shape of the probability distribution of points that might be inferred by a collective encoding.

Doherty et al. [18] studied experiments on four different scatter-related perceptions and found that except for the error variance, all graphical attributed the affect subjective but non-objective correlations (R) remained the same.

Hasse and Kaczmarek [19] compared the visual perception of auditory and scatterplots and found the similar correlation estimation performance in both modes. Their results showed that electrical tactile complexity of the graphics and also provided useful information to improve the future version of the tactile display.

3 Experiment

3.1 Method and Materials

In this paper, the stimuli are scatterplots that contain 100, 200 and 500 normally-distributed points (including inherent points and interference points) which are generated by Python using pseudo-random numbers. The Pseudo-random numbers which are taken from a Gaussian distribution data set that the mean and the standard deviation was set to 0.5 and 0.2 are chosen to build correlated pairs (x, y') and are then used to generate scatterplots. The x -coordinate is the first number chosen from the Gaussian distribution. Y value is then created and transformed using Eq. (2) to create y' .

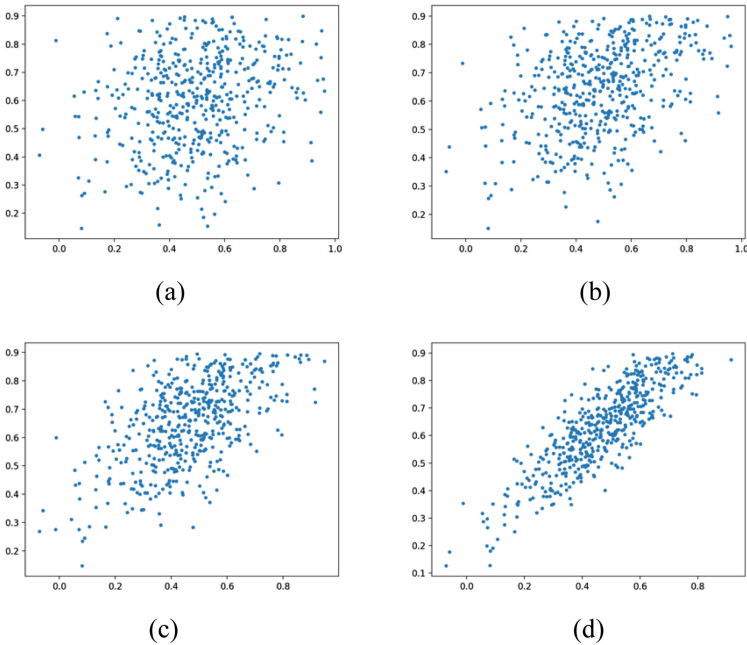








Fig. 1. Scatterplots in different distributions

$$y' = \frac{\lambda x + (1 - \lambda)y}{\sqrt{\lambda^2 + (1 - \lambda)^2}}, \text{ where } \lambda = \frac{r^2 - \sqrt{r^2 - r^4}}{2r^2 - 1} \quad (2)$$

In order to avoid points outside the range of the graph, any points which are greater than 2 standard deviations from the mean are eliminated and then generated a new point to take its place.

Figure 2 shows 200 inherent points which R value is 0.5 and interference points which R value is 0.5 in the scatterplot. The different colors of interference points are shown in the Table 1.

Table 1. The different color of interference points

Component	Value					
H	30°	60°	90°	120°	150°	180°
S	100%	100%	100%	100%	100%	100%
B	100%	100%	100%	100%	100%	100%
Sample						

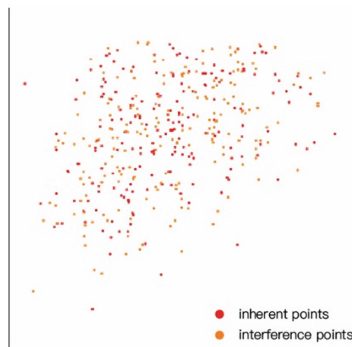


Fig. 2. Scatterplot with 200 inherent points (r = 0.5) and interference points(r = 0.5)

3.2 Experimental Design

In this experiment, there are two scatter plots presented in the screen with different R value which size are both 464 × 464 pixels. Participants need to observe the trends of these two scatterplots and decide which scatterplot has higher R value. Press ‘f’ key in keyboard to response if the left one is higher, press ‘j’ key in keyboard to response if the right one is higher.

The experiment is divided into two parts: the training session and the formal session. At the beginning of the experiment, the computer screen shows experimental

guidance. After reading the guidance, participants enter the experimental stage by pressing any key. The cross-visual guidance appears in the center of screen for 500 ms at first and then participants need to observe the trends of these two scatterplots and decided which scatterplot has higher R value. Press ‘f’ key in keyboard to response if the left one is higher, press ‘j’ key in keyboard to response if the right one is higher. In this experiment, the presentation time of each task images is set to infinite. The cross-visual guidance appears after each image. Procedure of the experiment is shown in the Fig. 3.

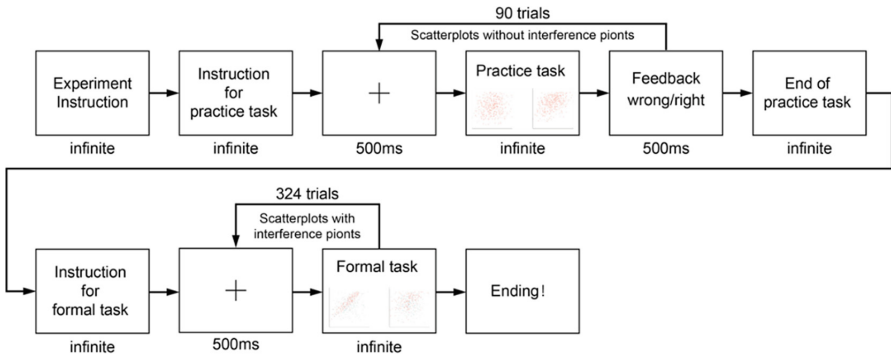


Fig. 3. The procedure of experiment

The interference points are not included in the scatterplots presented in the practice task. Training session is used to familiarize with the task for all participants. The training session consists of 90 trials. Figure 4 shows scatterplots in the training session.

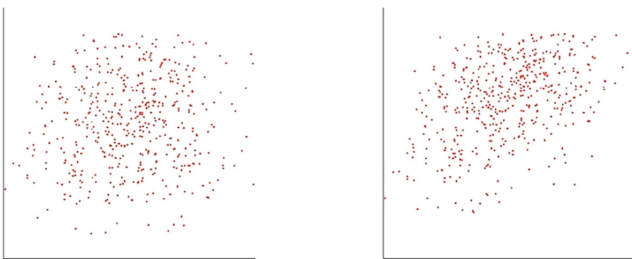


Fig. 4. Scatterplots in the training session (Color figure online)

Interference points are added in the scatterplots of the formal session. Participants are asked to make judgments based on the R value of red points (inherent points, $H = 0^\circ$, $S = 100\%$, $B = 100\%$) on which scatterplot has higher R value. Press ‘f’ key in keyboard to response if the left one is higher, press ‘j’ key in keyboard to response if the right one is higher. Figure 5 shows scatterplots in the formal session. For instances,

in Fig. 5, interference points which are blue ($H = H = 180^\circ$, $S = 100\%$, $B = 100\%$) are added in the scatterplots, participants should make decisions based on the red points. The scatterplot on the left has higher R value, so they are supposed to press ‘f’ key in the keyboard to response. In this experiment, there are 50% pictures on the left which have higher R value, 50% pictures on the right which have higher R value. The reaction time and response of each participants are recorded through E-prime.

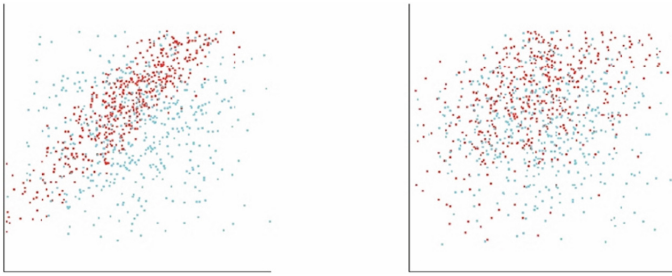


Fig. 5. Scatterplots in the formal session

25 participants attend this experiment, and each seats 64 cm from a screen with the resolution of 1366×768 -pixel laptop. All participants are tested on both training and formal sessions. Average age of participants is 24 years old who has at least some experience with scatterplots; most have made and used scatterplots on several occasions. They are given as much time as needed to complete each task and they are also mentioned that accuracy is important.

4 Result

The independent variables of this experiment are as follow: (1) The amount of interference point (Amount): 3 levels, 100, 200, 500; (2) The color of interference point (Color): 6 levels, $H = 30^\circ$, $H = 60^\circ$, $H = 90^\circ$, $H = 120^\circ$, $H = 150^\circ$, $H = 180^\circ$; (3) The correlation levels of interference point (Rvalue): 3 levels, 0.1 (low correlation), 0.5 (general correlation), 0.9 (high correlation).

The dependent variable of this experiment is the reaction time (RT) of each subject. There are 25 subjects participate in the experiment. After data analysis by using SPSS, the 12th and 24th subjects’ reaction time (RT) are eliminated because of the abnormal deviation and too many singular values in reaction time.

The analysis of variance (ANOVA) is performed on RT and ACC. The results are shown in Table 2. As shown in the Table 2, the main effect of Color ($F = 5.981$, $P = 0.000$, $P \leq 0.01$), the main effect of Rvalue ($F = 34.864$, $P = 0.000$, $P \leq 0.01$) and the main effect of Amount ($F = 32.103$, $P = 0.000$, $P \leq 0.01$) on reaction time are significant and the main effect of the second-order interaction between factors: Rvalue*Color, $F = 1.076$, $P = 0.377$, $P > 0.01$; Amount*Color, $F = 1.085$, $P = 0.370$, $P > 0.05$; Amount*Rvalue $F = 3.144$, $P = 0.014$, $P > 0.01$ are not significant. The

main effect of third-order interaction is not significant ($F = 0.897, P = 0.591, P > 0.05$). It can be seen that these three independent variables (the color of the interference point, the amount of the interference points and the correlation of the interference points) all have significant influence on the reaction time(RT) of participants. As shown in Tables 3, 4 and 5, the multiple post-test multiple comparison of the least significant difference is then performed.

Table 2. Tests of between-subjects effects

Source	Type III sum of squares	df	F	Sig.	Partial Eta squared
Rvalue	15424064.531	2	34.864	.000	.055
Color	6614853.091	5	5.981	.000	.025
Amount	14202908.242	2	32.103	.000	.051
Rvalue* color	2381025.137	10	1.076	.377	.009
Rvalue* amount	2782197.923	4	3.144	.014	.010
Color* amount	2400687.825	10	1.085	.370	.009
Rvalue* color* amount	3968684.949	20	.897	.591	.015

The results of multiple comparisons in Table 3 show that there are significant differences between the mean reaction time (RT) of $H = 30^\circ$ and the reaction time (RT) of $H = 60^\circ, H = 90^\circ, H = 120^\circ, H = 150^\circ, H = 180^\circ$ which indicates that if $H > 30^\circ$, the reaction time(RT) is not influenced by the H value changes of the interference points. As the HSB value for inherent points in scatterplots is $H = 0^\circ, B = 100\%, H = 100\%$, the ΔH between inherent and interference points is 30 which means the difference between colors is not significant. As a result, it causes a huge interference which makes test images difficult to identify for the correlation of participants' perception.

According to the result of multiple comparisons for the amount of interference point in Table 4, it can be seen that there are significant differences between the mean reaction time (RT) of 100, 200 and 500. It can be indicated that the number of interference points has a huge interference on the correlation of participants' perception.

It can be seen from multiple comparisons for the correlation of interference point (Rvalue) in Table 5 that there are significant differences between the mean reaction time (RT) of $R = 0.1$ and $R = 0.5, R = 0.9$. It can be indicated that when $R > 0.1$, the mean reaction time (RT) is not influenced by the changes of interference points' correlation. The reason is that when $R = 0.1$, the interference points are in low correlation which means the distribution for interference points is straggling and makes test images difficult to identify.

After each reaction time is compared, Fig. 6a–c are line charts which show the influence of different combinations of variables on mean reaction time.

Figure 6a exposes the relationship between color, correlation of the interference points and the mean reaction time of each subject. As it shown in the graph, when the

Table 3. Multiple comparisons (color)

(I) Color	(J) Color	Mean difference (I-J)	Std. error	Sig.
30	60	240.0427 [*]	46.23047	.000
	90	132.7625 [*]	46.23047	.004
	120	131.7375 [*]	46.23047	.004
	150	180.5395 [*]	46.23047	.000
	180	164.4614 [*]	46.23047	.000
60	30	-240.0427 [*]	46.23047	.000
	90	-107.2802	46.23047	.020
	120	-108.3052	46.23047	.019
	150	-59.5032	46.23047	.198
	180	-75.5813	46.23047	.102
90	30	-132.7625 [*]	46.23047	.004
	60	107.2802	46.23047	.020
	120	-1.0250	46.23047	.982
	150	47.7770	46.23047	.302
	180	31.6989	46.23047	.493
120	30	-131.7375 [*]	46.23047	.004
	60	108.3052	46.23047	.019
	90	1.0250	46.23047	.982
	150	48.8019	46.23047	.291
	180	32.7238	46.23047	.479
150	30	-180.5395 [*]	46.23047	.000
	60	59.5032	46.23047	.198
	90	-47.7770	46.23047	.302
	120	-48.8019	46.23047	.291
	180	-16.0781	46.23047	.728
180	30	-164.4614 [*]	46.23047	.000
	60	75.5813	46.23047	.102
	90	-31.6989	46.23047	.493
	120	-32.7238	46.23047	.479
	150	16.0781	46.23047	.728

Table 4. Multiple comparisons (amount)

(I) Amount	(J) Amount	Mean difference (I-J)	Std. error	Sig.
100	200	156.1486 [*]	32.68988	.000
	500	260.2089 [*]	32.68988	.000
200	100	-156.1486 [*]	32.68988	.000
	500	104.0604 [*]	32.68988	.001
500	100	-260.2089 [*]	32.68988	.000
	200	-104.0604 [*]	32.68988	.001

Table 5. Multiple comparisons (Rvalue)

(I) Rvalue	(J) Rvalue	Mean difference (I-J)	Std. error	Sig.
.10	.50	232.5958*	32.68988	.000
	.90	240.0262*	32.68988	.000
.50	.10	-232.5958*	32.68988	.000
	.90	7.4304	32.68988	.820
.90	.10	-240.0262*	32.68988	.000
	.50	-7.4304	32.68988	.820

interference point color is 30°(H = 30°, S = 100°, B = 100°), the mean reaction time of subjects is the longest in different correlation levels; when the interference point color is 60°(H = 60°, S = 100°, B = 100°), the mean reaction time of subjects is the shortest during these three correlation levels; when the interference point correlation level is low (R = 0.1), the mean reaction time is higher than the other two correlation levels; in all combinations of variables, the mean reaction time is the shortest when h = 60° and the correlation level is general (R = 0.5).

Figure 6b shows the relationship between the color, amount of interference points and the mean reaction time. It can be concluded that when the H component of the interference point located at 30° in the color wheel (H = 30°, S = 100°, B = 100°), the mean reaction time is the longest; when the H component of the interference point located at 60° in the color wheel (H = 60°, S = 100°, B = 100°), the mean reaction time is the shortest compared to the other five locations of H component; when the interference point number is 100, the mean reaction time is higher than the other two levels; In all the combinations of variables, the mean reaction time of subjects is the shortest when H component of the interference point located at 60° (H = 60°, S = 100°, B = 100°) in the color wheel and the amount of the interference points is 500.

Figure 6c shows the relationship between the amount, correlation of interference points and the mean reaction time of subjects. It can be seen from the figure that when the number of interference points is 100, the mean reaction time is the longest compared to 200 and 500. When the number of interference points is 500, the mean reaction time is the shortest in different correlation levels; when the correction level of the interference points is low (r = 0.1), the mean reaction time is longer than the other two levels of 0.5 and 0.9; When the correlation level is general (r = 0.5) and the amount is 500, the mean reaction time is the shortest among the all combinations of variables.

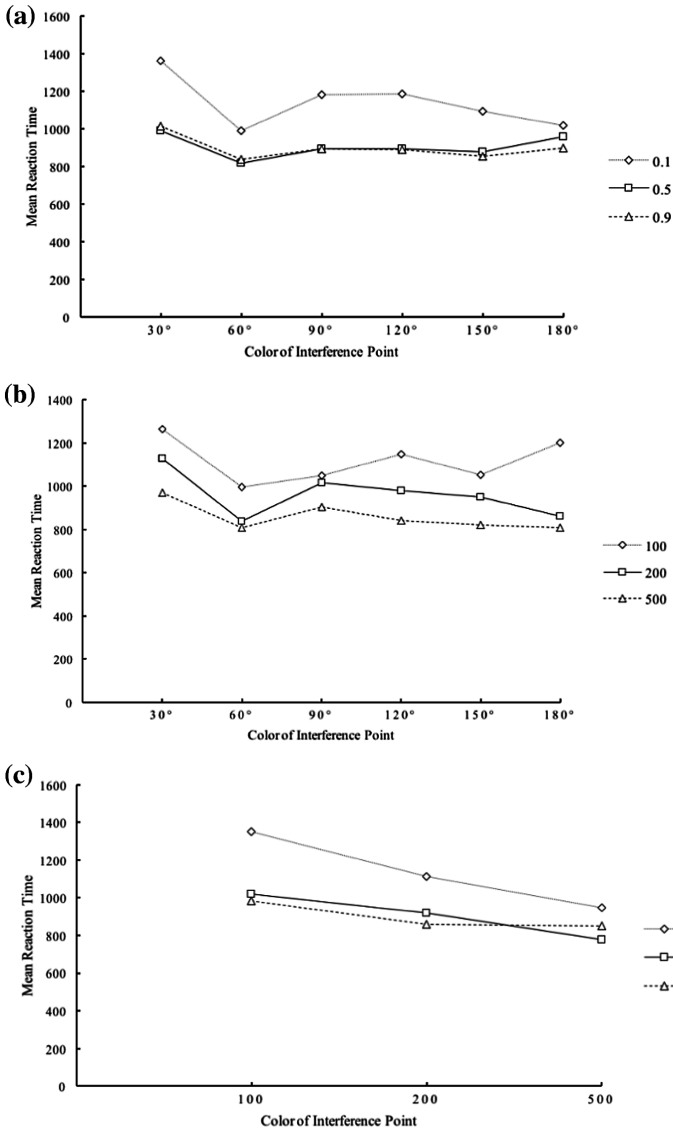


Fig. 6. a. The mean reaction time for color and correlation of the interference points, b. The mean reaction time for color and amount of the interference points, c. The mean reaction time for amount and correlation of the interference points

5 Conclusion

The effect of the color of interference points, amount of interference points and correlation of interference points on correlation recognition are analyzed in this paper according to the correlation of scatter plot correlation test. What's more, the degree of

interference is compared. The result shows that the advantages and disadvantages of the scatter plot visualization can be judged according to correlation recognition of the users. That is to say, the faster the subjects recognize the correlation, the more reasonable scatter plot visualization is.

The results of main effect indicate that the influence of the color, amount and correlation of the interference points on the reaction time is significant under the white background; the result of post-test multiple comparison analysis shows the different mean reaction time between different combinations of variables. In the end, the shortest reaction time among the combination of the three variables is concluded; the method of quantifying color based on HSB color mode in this paper can intuitively represent the color for experimental research.

References

1. Friendly, M., Denis, D.: The early origins and development of the scatterplot. *J. Hist. Behav. Sci.* **41**(2), 103–130 (2005)
2. Elmqvist, N., Dragicevic, P., Fekete, J.-D.: Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. *IEEE Trans. Visual. Comput. Graph.* **14**(6), 1141–1148 (2008)
3. Chung, D.H.S., Legg, P.A., Parry, M.L., Bown, R., Griffiths, I.W., Laramée, R.S., Chen, M.: Glyph sorting: interactive visualization for multi-dimensional data. *Inf. Visual.* **14**(1), 76–90 (2013)
4. Lewandowsky, S., Spence, I.: The perception of statistical graphs. *Sociol. Methods Res.* **18** (2–3), 200–242 (1989)
5. Kanjanabose, R., Abdul-Rahman, A., Chen, M.: A multi-task comparative study on scatter plots and parallel coordinates plots. *Comput. Graph. Forum* **34**(3), 261–270 (2015)
6. Rensink, R.A.: On the prospects for a science of visualization. In: Huang, W. (ed.) *Handbook of Human Centric Visualization*, pp. 147–175. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-7485-2_6
7. Cleveland, W.S., McGill, R.: Graphical perception: the visual decoding of quantitative information on graphical displays of data. *J. Roy. Stat. Soc.* **150**(3), 192–229 (1987)
8. Meyer, J., Taieb, M., Flascher, I.: Correlation estimates as perceptual judgments. *J. Exp. Psychol. Appl.* **3**(1), 3–20 (1997)
9. Li, J., Martens, J.B., Wijk, J.J.V.: Judging correlation from scatterplots and parallel coordinate plots. *Inf. Vis.* **9**, 13–30 (2010). Palgrave Macmillan
10. Carpenter, M.: The new statistical analysis of data. *J. Am. Stat. Assoc.* **42**(2), 205–206 (1996)
11. Sher, V., Bemis, K.G., Liccardi, I.: An empirical study on the reliability of perceiving correlation indices using scatterplots. *Comput. Graph. Forum* **36**(3), 61–72 (2017)
12. Li, J., Martens, J.B., Van Wijk, J.J.: A model of symbol size discrimination in scatterplots (2010)
13. Micallef, L., Palmas, G., Oulasvirta, A.: Towards perceptual optimization of the visual design of scatterplots. *IEEE Trans. Vis. Comput. Graph.* **23**(6), 1588–1599 (2017)
14. Gleicher, M., Correll, M., Nothelfer, C.: Perception of average value in multiclass scatterplots. *IEEE Trans. Vis. Comput. Graph.* **19**(12), 2316–2325 (2013)
15. Stenholt, R., Madsen, C.B.: Shape perception in 3-D scatterplots using constant visual angle glyphs. In: *Virtual Reality Short Papers and Posters* (2012)

16. Bertini, E., Santucci, G.: Improving 2D scatterplots effectiveness through sampling, displacement, and user perception. In: International Conference on Information Visualisation. IEEE Computer Society (2005)
17. Rensink, R.A.: The nature of correlation perception in scatterplots. *Psychon. Bull. Rev.* **24**(3), 1–22 (2017)
18. Doherty, M.E., Anderson, R.B., Angott, A.M.: The perception of scatterplots. *Percept. Psychophys.* **69**(7), 1261–1272 (2007)
19. Haase, S.J., Kaczmarek, K.A.: Electrotactile perception of scatterplots on the fingertips and abdomen. *Med. Biol. Eng. Comput.* **43**(2), 283 (2005)