Exploratory Data Analysis for DSFB Project

Group 6: Awen, Mahmoud, Joshua, Zhan

A. Task Overview

In this delivery, we annotated 120 image samples for DSFB Project. Each image is supposed to contain a string of 11 characters. This report is a exploratory data analysis on the statistics and visualisation based on these 120 samples. We pay attention on position of the characters, the character itself and the order.

B. Character Distribution

A total of 13 unique characters appear in our sample, with the number of occurrences of each letter showing a **long-tailed** distribution as in Figure 1. Of these, 5, 0, 8, and 6 occur far more often, probably because our code have a fixed prefix that contains these four characters. 5 occurs 279 times (the most frequent), while X occurs only 18 times (the rarest).

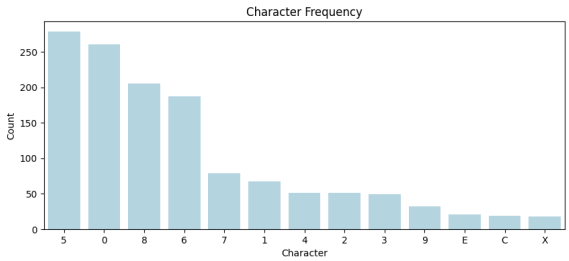

Fig. 1. Character Distribution

C. Character Size

Area refers to the area of the rectangle covered by the character from the top left corner to the bottom right corner of the character. The distribution of the area occupied by each character is shown as Figure 2 and their mean value is **1903.79** pixels. Most of them have an aspect ratio between 1:1.5 and 1:1.8. The distribution of aspect ratio is reported in Appendix Figure 5.

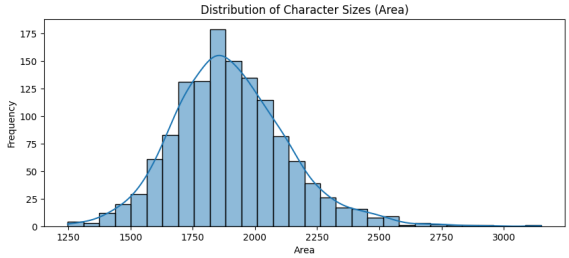

Fig. 2. Size Frequency

D. Character Location

From the heat map as Figure 3 of the position of different characters, we observe that the code is divided into two segments, *prefix* + *suffix*. The x coordinate position of the code is between 800 and 880 and the y coordinate position is between 400 and 800.

Fig. 3. Heatmap of Character Locations

E. Character Order

The occurrence of each characters at different order is reflected in Figure 4 and also the contingency table I. These codes have a common prefix so characters 1-5 are certain. The 6th character is one of {6,7,8,9}, with 8 being the most likely. The 7th character is uniformly distributed and takes values in {1,2,3,4,5,6}. The 8th character has a very high probability of being 0. Characters 9-11 are evenly distributed across the character set.

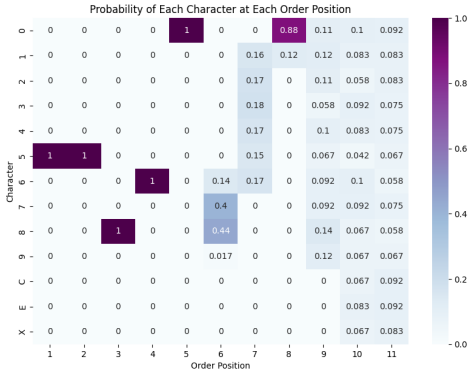

Fig. 4. Heatmap of Character Order

APPENDIX

TABLE I

CONTINGENCY TABLE FOR ORDER OF EACH CHARACTER

character \ Order	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0	120	0	0	105	13	12	11
1	0	0	0	0	0	0	19	15	14	10	10
2	0	0	0	0	0	0	21	0	13	7	10
3	0	0	0	0	0	0	22	0	7	11	9
4	0	0	0	0	0	0	20	0	12	10	9
5	120	120	0	0	0	0	18	0	8	5	8
6	0	0	0	120	0	17	20	0	11	12	7
7	0	0	0	0	0	48	0	0	11	11	9
8	0	0	120	0	0	53	0	0	17	8	7
9	0	0	0	0	0	2	0	0	14	8	8
C	0	0	0	0	0	0	0	0	0	8	11
E	0	0	0	0	0	0	0	0	0	10	11
X	0	0	0	0	0	0	0	0	0	8	10

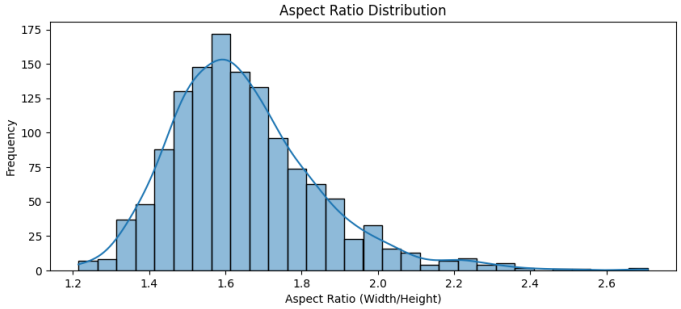

Fig. 5. Aspect Ratio frequency