Inter-rater Reliability Measures for Image Labels

Yaseen Haffejee
Computer Science
University of the Witwatersrand
Johannesburg, South Africa
1827555@students.wits.ac.za

Ziyaad Ballim
Computer Science
University of the Witwatersrand
Johannesburg, South Africa
1828251@students.wits.ac.za

Jeremy Crouch
Computer Science
University of the Witwatersrand
Johannesburg, South Africa
1598024@students.wits.ac.za

Fatima Daya
Computer Science
University of the Witwatersrand
Johannesburg, South Africa
1620146@students.wits.ac.za

***Abstract*—The report analyses the pre-processing and cleaning of various puzzle images and their masks in order to ensure the images can be utilised in a puzzle solving algorithm. Consequently any tasks performed are related to this greater task. We also investigated the inter-rater reliability of the images masks generated by different individuals utilising the Fleiss Kappa and Pearsons' Correlation Coefficient to measure the inter-rater reliability.**

***Index Terms*—Computer Vision, Image Segmentation, Image Processing**

I. Introduction

There are two primary focuses in this report: the processing and cleaning of images as well as their segmented masks, and investigating the inter-rater reliability of the available mask labels using Pearsons' correlation coefficient and the Fleiss Kappa metric. We aim to utilise the results of the investigation to curate a dataset from which we will build a puzzle solver. Consequently, careful consideration is made to relate any decisions taken, and the contribution it may make to solving the puzzle.

II. Pre-processing

The pre-processing consisted of analysing both, the RGB puzzle images and their respective masks.

A. Puzzle images

The dataset consists of 46 RGB puzzle pieces. Each image has a size of 3840 x 5120 pixels. The images are relatively large. Consequently there will be an increase in the computational expense that is coupled with per-pixel operations in order to solve the puzzle. Therefore, the images were resized to 480 x 640 pixels which is an eighth of the original size. At this size, the vital information of the image such as the edges of the puzzle, the shape of the puzzle and orientation are all preserved as seen in B. However the significant reduction in the size will lead to an increase in the computational speed of per-pixel operations.

B. Puzzle masks

A puzzle mask is a puzzle image that has been converted to grayscale and thereafter thresholded to produce a binary image. Within the binary image, the foreground (the puzzle piece) is represented by white (an intensity of 1) and the background is denoted by black (an intensity value of 0). The size of each mask is also 3840 x 5120 pixels. Consequently, we resized the masks to be 480 x 640 pixels as well for coherence with the puzzle images, and to address the concerns mentioned in Subsection II-A. It is worth noting that certain masks were inverted. This implies that within the mask, the puzzle piece was black and the background was white. The inverted pixel values would make it difficult for the puzzle solver to fit these pieces into the puzzle, since the algorithm would focus on the puzzle pixels (white pixels) whereas in these images the puzzle is represented by black pixels. These masks were identified and corrected as denoted in Appendix A.

In total the dataset has 137 masks. Each puzzle piece has a minimum of 2 masks and a maximum of 5 masks. If a puzzle piece has more than one mask, each mask was generated by a different individual. As a result, we need to measure the inter-rater reliability of these masks.

III. Inter-rater reliability

Certain puzzle pieces have multiple masks which were generated by different individuals. Hence we need to measure how reliable these labels are and to what extent are the labels the same. Inter-rater reliability enables us to find the extent to which the individuals that labelled the masks agree on which pixels are puzzle pixels and which pixels belong to the background [1]. The task was to segment the image into foreground and background pixels. Thus we have nominal data. In order to measure the reliability of the labels, we employed two metrics: the Fleiss Kappa metric and Pearsons' correlation coefficient which are appropriate when we are considering nominal data [2].

A. *Fleiss Kappa*

Fleiss' Kappa is a generalisation of the Cohen Kappa metric [2]. The Cohen Kappa metric is utilised to measure the inter-rater reliability between two raters, whereas the Fleiss Kappa metric enables us to investigate the inter-rater reliability when we have more than two raters. This is the case for most of the masks which we have.
Given m raters and n subjects to rate, the Fleiss Kappa metric is denoted by:

$$\kappa = \frac{p_a - p_\epsilon}{1 - p_\epsilon} \tag{1}$$

where:

$$p_\epsilon = \sum_{j=1}^{k} {q_j}^2 \tag{2}$$

$$q_j = \frac{1}{mn} \sum_{i=1}^{n} x_{ij} \tag{3}$$

$$p_a = \frac{1}{mn(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2 - mn \tag{4}$$

The Kappa statistic is an important evaluation since it rules out the possibility of agreement occurring between raters due to chance [4]. Consequently, mutual agreement between raters is indicative of a truth value. The Fleiss Kappa value provides a value in the range $[0, 1]$, where 1 implies perfect agreement and 0 implies no agreement [2]. The results are discussed in Section IV.

B. *Pearson's Correlation Coefficient*

Pearsons' Correlation Coefficient enables us to calculate how closely the masks suggested by different raters are correlated and agree with each other [3].
Pearsons' Correlation Coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{(}y)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

The Pearsons' Correlation Coefficient metric provides us with a value in the range $[0, 1]$. The higher the value, the greater the correlation between the masks which implies stronger agreement between the raters and vice versa. Since the calculation is a pairwise calculation, the results discussed in Section IV represent the average correlation between all the masks for a given puzzle piece.

IV. Results

In order to summarise the general agreement between raters, we utilised the averages from tables in Appendices C and D. The average agreement according the Fleiss Kappa metric is 0.9815. The average agreement according to Pearsons' Correlation Coefficient is 0.9909. Both metrics indicate a high level of agreement between the varying raters. Consequently, we can be assured that the labels provided are accurate since we are near perfect agreement between raters. The metric values for each mask can be further studied in Appendices C and D.

V. Conclusion

The pre-processing step ensured that we re-sized both the RGB puzzle images and the masks to a size of 480 x 640 pixels. Even though the reduction in size of the image will likely lead to the loss of data, none of the data that is lost is integral to the puzzle solver. All the vital information is preserved. This ensures that the images can be utilised effectively in the puzzle sorting algorithm. The reduction in size also ensures a computational speed-up since the images are smaller. The utilisation of the Fleiss Kappa and Pearsons Correlation metric indicated that the masks have a high inter-rater reliability. Consequently, we can curate the final set of masks by classifying a pixel based on the highest number of votes according to the raters.

References

[1] A. Fink. Survey research methods. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, pages 152–160. Elsevier, Oxford, third edition edition, 2010.

[2] Stephanie Glen. "fleiss' kappa" from statisticshowto.com: Elementary statistics for the rest of us! https://www.statisticshowto.com/fleiss-kappa/.

[3] Stephanie Glen. "inter-rater reliability irr: Definition, calculation" from statisticshowto.com: Elementary statistics for the rest of us! https://www.statisticshowto.com/inter-rater-reliability/.

[4] Charles Zaiontz. Fleiss' kappa.

Appendix

A. *Inverted Masks*

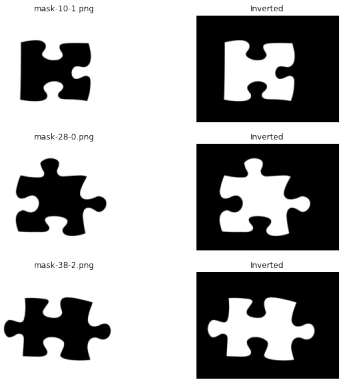

Fig. 1. Inverted masks and their corrected masks

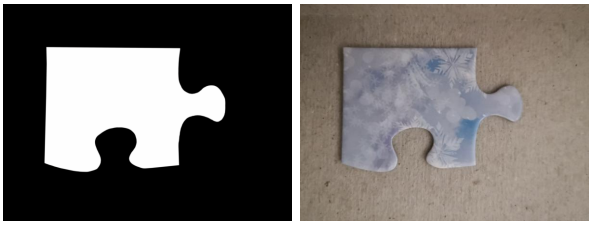

Fig. 2. Re-scaled image and mask

B. Re-sized image and mask

C. Fleiss Kappa results for each set of masks

Mask	Fleiss Kappa
mask-0	0.994332
mask-1	0.971781
mask-2	0.980728
mask-3	0.751374
mask-4	0.993802
mask-5	0.992485
mask-6	0.976068
mask-7	0.980328
mask-8	0.99028
mask-9	0.994931
mask-10	0.996075
mask-11	0.990884
mask-12	0.982699
mask-13	0.994761
mask-14	0.965092
mask-15	0.987199
mask-16	0.984237
mask-17	0.990519
mask-18	0.989317
mask-19	0.99044
mask-20	0.948888
mask-21	0.984404
mask-22	0.958586
mask-23	0.995295
mask-24	0.980063
mask-25	0.988322
mask-26	0.991019
mask-27	0.995808
mask-28	0.991998
mask-29	0.99266
mask-30	0.992094
mask-31	0.994384
mask-32	0.992898
mask-33	0.972895
mask-34	0.990826
mask-35	0.996248
mask-36	0.995971
mask-37	0.962213
mask-38	0.992853
mask-39	0.992101
mask-40	0.995963
mask-41	0.983257
mask-42	0.995582
mask-43	0.980809
mask-44	0.994319
mask-45	0.994776

D. Pearson's correlation coefficient for each set of masks

Mask	Pearson's Coefficient
mask-0	0.998754
mask-1	0.986166
mask-2	0.991246
mask-3	0.838158
mask-4	0.998309
mask-5	0.99775
mask-6	0.988713
mask-7	0.991298
mask-8	0.996304
mask-9	0.998808
mask-10	0.999475
mask-11	0.996834
mask-12	0.992962
mask-13	0.998898
mask-14	0.980615
mask-15	0.995051
mask-16	0.994932
mask-17	0.9978
mask-18	0.991088
mask-19	0.996842
mask-20	0.97196
mask-21	0.993054
mask-22	0.977932
mask-23	0.99923
mask-24	0.991086
mask-25	0.99552
mask-26	0.997107
mask-27	0.999452
mask-28	0.997692
mask-29	0.998598
mask-30	0.997374
mask-31	0.998881
mask-32	0.997682
mask-33	0.986119
mask-34	0.997087
mask-35	0.99936
mask-36	0.999263
mask-37	0.979781
mask-38	0.997992
mask-39	0.997737
mask-40	0.999258
mask-41	0.992543
mask-42	0.998896
mask-43	0.99123
mask-44	0.998204
mask-45	0.999209