Inter-rater Reliability Measures of Labels for Image Segmentation*

Chloë Smith
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
1877342@students.wits.ac.za

***Abstract*—This report investigates the use of inter-rater reliability (IRR) measures of labels for image segmentation, comparing percentage similarity to the Kappa statistics Cohen's Kappa, and Fleiss' Kappa. We look at labels from two raters, as well as more generally from multiple raters, and find that the Kappa statistics are more conservative in their measurement of reliability. We also take a look at data cleaning and pre-processing for image segmentation, in the context of a (jigsaw) puzzle solver.**

***Index Terms*—Image Processing, Machine vision, Computer vision, Pre-processing**

I. Introduction

In this report, we investigate cleaning, pre-processing, and inter-rater reliability of mask labels for the purpose of image segmentation. We use a dataset of 46 RGB puzzle images, with 137 RGB masks, obtained by asking a group of raters to submit masks for 3 images each, not necessarily resulting in each image having at least 3 masks. Each image has at least one mask, and may have more than 3. These images are to be segmented into pixels containing puzzle pieces and pixels containing background, for the eventual purpose of solving the puzzle. We need to extract the puzzle pieces from the images so we can then fit them together at a later stage. It is important that the extraction is accurate, because if we are missing too much information around the edges, it may later be very difficult to determine whether two pieces connect and how they are orientated.

Inter-rater reliability is measured using percentage similarity, Cohen's Kappa, and Fleiss' Kappa.

II. Pre-processing

Our dataset consists of 46 RGB puzzle images of size 3840 x 5120 pixels, and labels consisting of 137 RGB masks of the same size, with the number of masks per image ranging from 2 to 5.

The images and masks are downscaled to 480 x 640 pixels, and the masks are converted to greyscale and thresholded to 0, 1, with 0 denoting background pixels and 1 denoting puzzle pixels. We note that the raw masks do not all conform to a standard of background pixels coloured black, and puzzle pixels coloured white.

There are a few masks that have inverted labels, but since these still effectively contain the information we need, and the dataset is small enough that we can easily spot-correct this, we manually correct these masks.

Fig. 1. A sample of the puzzle piece images.

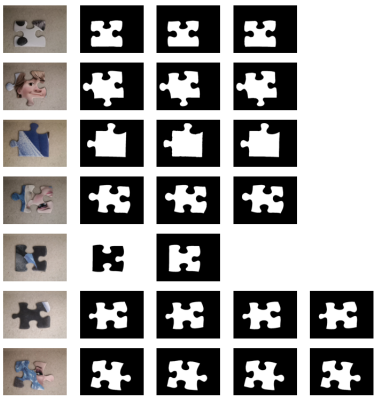

Fig. 2. A sample of the puzzle piece images and their corresponding masks before pre-processing.

III. Inter-rater Reliability

When collecting labels from multiple sources, it is important that we have some idea of the reliability of the labels [1]. This has traditionally been done using percentage agreement [2], which we note may be misleading for particular datasets, including ours.

For image segmentation masks such as the ones in our dataset, a large degree of similarity between labels can be achieved without the labels actually being correct or even

Fig. 3. Inverted masks in the thresholded labels.

Fig. 4. Corrected masks.

similar. Thus we investigate and compare alternative measures of agreement: Cohen's Kappa and Fleiss' Kappa [3], [4].

Cohen's Kappa is used to measure reliability between two raters [3], and is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement.

This is where the difference between our kappa statistics and traditional percentage similarity lies; percentage similarity does not take into account the effect of chance.

Fleiss' Kappa is a generalisation of Cohen's Kappa allowing us to measure agreement between labels from $m > 2$ raters of n subjects into k categories [5], and is calculated as:

$$\kappa = \frac{p_a - p_\epsilon}{1 - p_\epsilon}, \tag{2}$$

where $p_a = \frac{1}{mn(m-1)}[\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2 - mn]$, and
$p_\epsilon = \sum_{j=1}^{k} q_j^2$,
$q_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$.

We calculate these kappa statistics as well as percentage similarity for the labels for each image, and the results are presented in the following section.

IV. Results

We see that for both percentage similarity and our kappa statistics, we have a high degree of agreement between our labels, with an average 0.9944 percentage similarity and an average kappa similarity of 0.9851. The Kappa statistics are slightly more conservative, consistently scoring agreement lower than percentage similarity, with scores averaging 0.94 percent lower than percentage similarity.

V. Conclusion

Use of Cohen's Kappa and Fleiss' Kappa gave slightly better insight into the agreement of our labels, but considering that a high-degree of agreement can be achieved without the labels in general being accurate or correct, there is room for improvement here in the choice of statistic. Our choice in pre-preprocessing to downscale and threshold our masks may amplify noise in the labels, especially around the edges, but this may not have a significant impact on the final goal of segmenting out the puzzle pieces for the puzzle solver. We suggest that the ground-truth labels be constructed by averaging the masks for each image and then thresholding, as there is a high degree of inter-rater reliability in this dataset.

References

References

[1] Mary L. McHugh, "Interrater reliability: the kappa statistic," Biochem Med (Zagreb). 2012 Oct; 22(3): 276–282, Oct 2012.

[2] Stephanie Glen, Inter-rater Reliability IRR: From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/inter-rater-reliability/.

[3] Stephanie Glen, "Cohen's Kappa Statistic" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/cohens-kappa-statistic/.

[4] Stephanie Glen, "Fleiss' Kappa" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/fleiss-kappa/.

[5] Charles Zaiontz, "Fleiss' Kappa," J. From Real Statistics Using Excel https://www.real-statistics.com/reliability/interrater-reliability/fleiss-kappa/.