Observation and Visualization of Subjectivity-based Annotation Tasks

Rika Miura
Ochanomizu University
Tokyo, Japan
g1820536@is.ocha.ac.jp

Ami Tochigi
Ochanomizu University
Tokyo, Japan
g1620527@is.ocha.ac.jp

Takayuki itoh
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

Abstract—**Annotation is an upstream process for constructing training data for machine learning tasks. The reliability of annotation is very important for the reliability of machine learning. The annotations vary from worker to worker, and differences in these tendencies may impair the reliability of the data. This is especially relevant for tasks that depend on the subjectivity of the workers. This study aims to realize reliable annotation by observing the annotation results of workers. As a specific example, we applied the annotations of three workers who evaluated facial expressions by the Likert scale on 977 face images as a subject. We verified the reliability of the annotations from the visualization results.**

Index Terms—**subjectivity, visualization, annotation, training data**

I. Introduction

Improving the quality of training data by ensuring the reliability of the annotation process is important in order to improve the accuracy and reliability of machine learning. Here, the creation of training data is often performed manually. Therefore, variations and uncertainty may occur in the data due to human factors such as ability, knowledge, and upbringing of workers [1]. These variations and uncertainty of the data have a significant impact on the quality of training data. In particular, it is often difficult to obtain reliable training data with the annotation tasks based on the subjectivity of the worker such as age and gender estimation. This can be partly due to the lack of clear criteria and inconsistent decisions of each individual worker [1]. Such variations and uncertainty of the data vary from worker to worker. For example, Itoh [2] showed a visualization example that age estimation by eight workers with facial images varies from worker to worker. This study found that a worker gave higher age estimates for person images in the 40s to 60s while another worker gave lower age estimates. Some studies [3,4] showed that the labels and polarity of emotions annotated by multiple workers vary among individuals due to factors such as the emotional experience of the worker.

There have been many discussions on the quality of annotations based on data variability, uncertainty, and correctness, but a small number of discussions on the annotation tendency of workers. It is also important to verify the tendency of workers who create the training data in order to improve the quality of training data. Here, time factors are important while analyzing changes in the reliability of the training data during the manual annotation tasks. The time factors include the uncertainty of the annotation decision criteria due to work fatigue or the individual worker's decision criteria getting clearer as they get used to the work. Thus, we can find improvements to the problems caused by the time factor in the annotation process and obtain more efficient and reliable training data by examining the relationship between the work time and the quality of the data. Visualization of the annotation tendency and reliability of each worker is effective to understand the influence on the annotation results. Therefore, this paper aims to realize reliable annotation by analyzing and explaining the training data from workers' points of view.

This paper presents the analysis and visualization of the annotation tasks in which three workers annotated 977 face images with six facial expressions (treated as six items) using the Likert scale as the task which depends on individual subjectivity. The main contributions of this study are as follows.

1) Observation of the relationship between the time required for annotation, the elapsed time, and the quality of the data.
2) Visualization of the annotation tendency of each worker to find items that are difficult to annotate.
3) Discovery of factors in annotation errors from worker to worker.

II. Related Work

A. Reliability evaluation of training data

There have been several studies that aim to improve the reliability of training data by evaluating the reliability of annotations. Dawid [5] proposed the model which obtain reliable training data by alternately repeating estimation of the true answer and estimation of worker's ability by measuring worker error. Mitsuda [6] analyzed the reliability of the annotations by using the gaze information during the worker's tasks and improved the reliability of the training data. Komatani [7] compared two indices of Fleiss's κ [8] and Krippendorff's α [9] while assessing the reliability of annotations. The results showed that the inter-worker agreement was not extremely low when Krippendorff's α was used because the degree of disagreement was taken into account. Our study applied a similar approach to evaluate the reliability. Meanwhile, these studies did not address either the analysis of the relationship between the annotation work time and the reliability of the data or the analysis of the reliability of individual worker annotation items. This paper discussed the relationship between annotation time and data reliability by logging the time required and elapsed during the annotation process of workers. Additionally, we analyzed the tendency of annotation by evaluating the reliability of workers for each item.

B. Visualization of annotation trends

Itoh [2] developed the tool to visualize the discrepancy of annotations among workers by applying a heatmap. As a result, they showed the tendency of annotation for each worker while the task depends on subjectivity. Komatani [7] analyzed the tendency to assign scores to training data among workers by confusion matrix and regression analysis in order to understand the tendency of annotation trends. Inagaki [10] extracted minority workers with characteristic answers by using multidimensional scaling to visualize the data answered and spectral clustering to classify the data. These studies only visualized the overall trend of the training data while few studies comprehensively addressed item-by-item visualization for each worker and the relationship between data reliability and visualization results. Our study visualized the training data for each worker and analyzed them together with the reliability evaluation values to analyze the annotation tendency of each worker more in detail.

III. Processing flow of observation and visualization of annotation tasks

A. Dataset and evaluation method

In this study, we applied the FACES database as a data set [11]. The database provides 977 facial images that correspond to either of six facial expressions (happiness, disgust, anger, neutrality, sadness and fear) of 171 participants (58 young, 56 middle-aged and 57 elderly). In other words, the facial images have a label corresponding to either of the six expressions. We asked three workers (female students in their twenties) to subjectively evaluate the 977 facial images using a 5-point Likert scale for each of the six expressions without looking at the labels of the images. Table I shows the items of the impression evaluation, where we treat the six expressions as six items. We conducted a questionnaire after the annotation task was completed and analysed together with the visualisation results.

B. Reliability evaluation index

The annotation results by multiple workers do not always match [7] since this study deals with subjective annotation tasks. Therefore, we analyzed the reliability of the annotations based on the degree of agreement among workers. This paper applied Krippendorff's α as a method to evaluate the reliability of annotations. In addition, the validity of the calculated α value is verified by applying the intraclass correlation coefficient (ICC) to calculate the reliability evaluation value. Both of them are measures of inter-rater reliability. There are other typical measures of inter-rater reliability such as Kendall's coefficient of concordance and kappa statistic [8]. However, this study adopted the above two indices because we can calculate them while supposing that the scale level can be regarded as an interval scale since the training data used in this study is based on the Likert scale. The following explains these two indices.

TABLE I
Evaluation items

```
            Strongly disagree                                   Strongly agree

                 1         2         3         4         5
 happiness       |---------|---------|---------|---------|
   disgust       |---------|---------|---------|---------|
     anger       |---------|---------|---------|---------|
neutrality       |---------|---------|---------|---------|
   sadness       |---------|---------|---------|---------|
      fear       |---------|---------|---------|---------|
```

Krippendorff's α: A measure to calculate the degree of agreement between two or more workers. This measure is highly versatile because users can switch the definitions of the distance between scores based on scale levels [12]. In this study, we calculated Krippendorff's α with the interval measure by using irr package in R since this study used the data annotated with the Likert scale. As stated in a previous study [7], $\alpha > 0.8$ is generally regarded as a reliable agreement rate in sociological research. However, the annotation agreement rate of the tasks which are annotated by the subjectivity of each worker tends to be lower. Therefore, about 0.4 for α have proposed appropriate for such tasks.

Intraclass Correlation Coefficient (ICC): The intraclass correlation coefficient is a method to determine the reliability within or between examiners [14]. There are three types of intraclass correlation coefficients: Case1, Case2, and Case3. In this study, we wanted to determine the inter-test reliability of a particular examinee. Therefore, this study used Case3 in the ICC function of the psych package in R. Table II shows a criterion for ICC adopted in this study. Remark that Tsushima stated that this table had no theoretical basis because it was an application of the table of Kappa coefficients by Landis [13] to the determination of ICC [14] .

TABLE II
CRITERION FOR DETERMINING THE NUMBER OF INTRA-CLASS RELATIONSHIPS

icc	evaluate
0.0 0.2	Slight
0.21 0.40	Fair
0.41 0.60	Moderate
0.61 0.80	Substantial
0.81 1.00	Almost Perfect

C. Quality analysis for the relationship between time and data

We visualized the annotation results in order to understand how the number of annotation tasks and the time required for annotation of one face image are related to the quality of the training data. The process is as follows.

1) Preprocessed the scales to normalize them using the scikit-learn library StandardScaler in Python in order to align the scales since the unit of data for the number of elapsed times and the time required for annotation per face image are different.
2) Performed the hierarchical clustering with the normalized data by using the module Scipy in Python. We applied the Ward method to generate clusters with the Euclidean distance.
3) Determined the number of clusters. We considered the range of the number from 2 to 25 and determined the optimal number of clusters with 26 indicators using Nbclust package in Python [15]. As a result, we specified the number of clusters to 17.
4) Assigned a color to each cluster and visualized in a scatterplot.
5) Calculated the reliability evaluation value for each cluster using the two indices described in Section III-B. By logging the annotation time, the changes in the reliability of the data over time were analysed. We matched the reliability evaluation value with the visualization results and discussed the relationship between the elapsed time, the time required, and the reliability of the data.

D. *Visualization of trends of worker's annotations*

We applied the following two visualizations for multidimensional training data.

Dimension reduction: We applied principal component analysis (PCA) to visualize the multi-dimensional training data to understand the annotation tendency of each worker. This visualization allows us to observe the distribution of the annotation of all items.

Parallel coordinate plot: This study visualized the same data also using parallel coordinate plots (PCP) to finely examine the tendency of each item roughly observed by PCA. This visualization allows us to observe a specific item and/or a specific worker in detail.

IV. Results

A. *Relationship between temporal changes and reliability*

This section shows the visualizations using the shiny package of Rstudio following the procedure described in Section III-C. Fig. 1 shows the scatterplot as the result of our visualization. One point corresponds to one image that has a six-dimensional value. A specific color is assigned to each of the 17 clusters described in section III-C. The horizontal axis indicates the number of annotated face images while the vertical axis indicates the time taken to annotate one face image. Table III shows the results of the calculation of the α values for each cluster. Clusters 5, 6, and 7 were excluded in this result because the number of face images in each cluster was too small (less than three). The results showed a strong positive correlation for all items were above 0.9 between the α values and the intraclass correlation coefficient. We determined that the α values were sufficiently reliable and therefore used the scatterplot in Fig. 1 and the α values in Table III in this experiment.

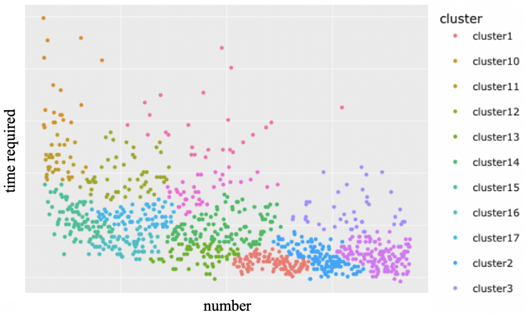


Fig. 1. Scatterplot representing the number of annotated images and the time required.

TABLE III
α VALUES FOR EACH CLUSTER

	happiness	disgust	anger	neutrality	sadness	fear
cluster1	0.968	0.485	0.403	0.793	0.818	0.729
cluster2	0.965	0.552	0.221	0.787	0.769	0.590
cluster3	0.964	0.550	0.448	0.651	0.611	0.591
cluster4	0.913	0.619	0.303	0.761	0.711	0.638
cluster8	0.903	0.338	0.028	0.724	0.638	0.340
cluster9	0.930	0.276	0.028	0.782	0.526	0.522
cluster10	0.121	0.255	0.107	0.065	-0.0	0.130
cluster11	0.539	0.253	0.535	0.142	0.283	0.057
cluster12	0.693	0.444	0.321	0.520	0.358	0.253
cluster13	0.964	0.494	0.052	0.848	0.784	0.439
cluster14	0.976	0.509	0.341	0.795	0.688	0.524
cluster15	0.920	0.426	0.397	0.369	0.104	0.429
cluster16	0.946	0.418	0.427	0.686	0.501	0.485
cluster17	0.799	0.538	0.318	0.724	0.443	0.489

Fig. 3 shows the visualization of clusters 10 and 11, which had low α values as shown in Table III. These two clusters were the image groups that were in the early stage of the annotation process and took a longer time to annotate. In contrast, Fig. 2 shows clusters 1, 2 and 4. These clusters were the image groups that were in the final stage of the annotation process and took a shorter time to annotate. Based on the comparison of Fig. 2 and Fig. 3, we found that the reliability of the data tended to be higher at the end of the annotation process than at the beginning and to be higher when the time required was shorter. This result indicates the unconfident annotation process due to the lack of confidence for the annotation in the early stages of the task. The result also implies that the task which took a long time to annotate was the task that was difficult to determine and the reliability was low. Here, we asked the workers which they felt was more accurate when comparing the early, middle and end stages of the annotation process, in the questionnaire conducted to the workers after the annotation work was completed. Two of three workers answered this question as the end of the annotation phase. As the reason, they responded that ”This is because

I have a clearer basis for my decision-making.” and ”I did not know the difference between anger and disgust at first but I began to understand the difference between the two items after I realized that making people wrinkle their noses was disgust.”. These results also show that familiarity with the annotation process makes workers’ own criteria clearer and workers’ annotations more stable. On the other hand, one of the workers who said that the accuracy of the annotation was higher at the beginning of the task said, ”I was more focused and thought more carefully at the beginning.”.

To the question ”Do you think that the accuracy of the annotations decreased due to the fatigue of the task?”, two of three workers answered ”yes.” This suggests that fatigue of the annotation task affected the quality of the data. Therefore, we suppose that improvements in the annotation environment can improve the reliability of the training data.

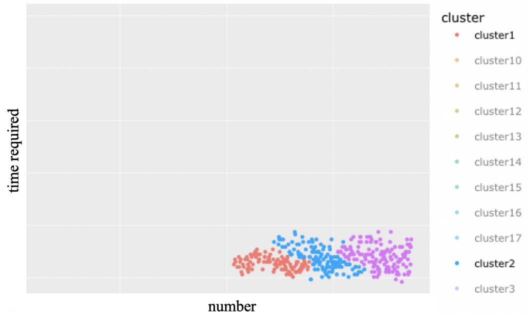


Fig. 2. cluster with higher α values.

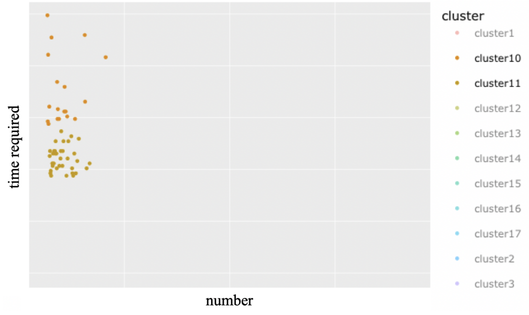


Fig. 3. cluster with lower α values.

B. Visualization by principal component analysis

This section presents the visualizations of the annotation tendency of each worker using principal component analysis (PCA).

Fig. 4 to Fig. 6 show the results of visualizing the multidimensional training data of each worker using PCA using the shiny package of Rstudio. A Specific color is assigned to each facial expression in this visualization. The arrows represent the principal components and the points are plots of the principal component values for each image in Fig. 4 to Fig. 6. Table IV shows the α values calculated for each of the six items and for each worker.

The visualization represents that ”fear” and ”sadness” had almost the same meaning for Worker A since the direction and length of the arrows of the principal components were the same in Fig. 4. This suggests that worker A had unconfident annotations of ”fear” and ”sadness.” In addition, the reliability of items similarly annotated such as ”fear” and ”sadness” or ”disgust” and ”anger” tended to be lower as shown in Table IV. Actually, the difference α values between ”fear” and ”sadness” of worker A was small. A similar result between ”anger” and ”disgust” of worker B was also observed as shown in Fig. 5 and Fig. 6. Furthermore, we found that the two items with close principal components may have close α values as well.

Then, we observed the commonalities among the three workers. The principal components of ”disgust” and ”anger” were close to each other and the face images of ”disgust” and ”anger” were distributed in similar portions. This suggests that the two items of ”disgust” and ”anger” were confusing for all workers and therefore annotations of ”disgust” and ”anger” tended to be unconfident for all workers. In fact, the two items both had lower α values and were less reliable as shown in Table IV. This result indicates it was difficult to properly distinguish between ”disgust” and ”anger” and annotate them. On the other hand, as for the two items of ”happiness” and ”neutrality,” all three workers were distinguishable from the other items. This can be also observed from the distribution in the scatterplot. From the above, we suppose two items of ”happiness” and ”neutrality” were relatively easy to annotate. Table IV also shows that α values of these two items are relatively high.

TABLE IV
α VALUES CALCULATED FOR EACH WORKER AND EACH ITEM

	A	B	C
happiness	0.9380	0.9331	0.9406
disgust	0.2805	0.4046	0.2540
anger	0.3134	0.3980	0.3469
neutrality	0.6727	0.6989	0.7368
sadness	0.5372	0.6183	0.5936
fear	0.5141	0.5364	0.4695

We asked three workers to rank the expressions from the easiness of annotations in the questionnaire conducted after the annotation task. Table V shows the results of the questionnaire. All three workers rated the happiness item as the easiest, and two of the three workers rated neutrality as the second easiest item. As shown in Table IV, ”happiness” and ”neutrality” were the items that are easy to determine also in terms of α values, while the items which the workers thought difficult to annotate tended to be less reliable. This suggests that the reliability of the data and the impressions of workers are consistent.

As mentioned above, we could infer the items which are difficult to annotate by observing the annotation tendency of each worker. Also, we could observe the relationship between the observation result and the α value. It is possible to improve the reliability of the training data by re-annotating the items which are determined to be difficult to annotate preferentially.

TABLE V
RESULTS OF RANKING THE EXPRESSIONS THAT WERE EASY TO DETERMINE

	A	B	C
first	happiness	happiness	happiness
second	neutrality	neutrality	sadness
third	sadness	sadness	disgust
fourth	anger	fear	anger
fifth	fear	disgust	fear
sixth	disgust	anger	neutrality

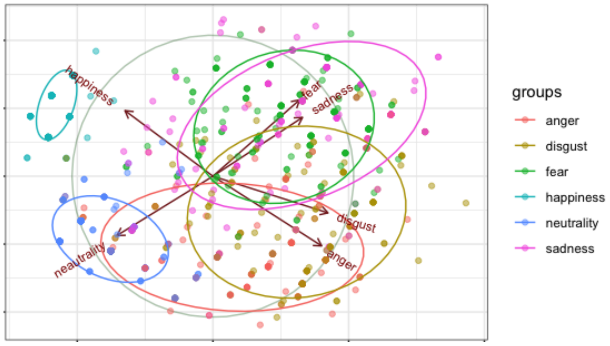


Fig. 4. Visualization of training data of worker A by PCA.

C. Visualization by parallel coordinate plots

This section presents the visualizations of the annotation tendency of each worker using a parallel coordinate plot (PCP).

We observed the tendency and variability of certain images with the visualization by PCP. We visualized the training data using HiPlot [16] published by Facebook. This section shows several examples of characteristic trends observed in the visualization results. In the visualizations shown in Fig. 7 and Fig. 9, the broken lines correspond to facial images and the seven vertical axes correspond to six items and the cluster which correspond to six different groups, one group for each expression. The annotation results for the specified range of images are displayed when a user drags the axes. Fig. 8 and Fig. 10 show the annotation results of the user-specified range of images in the PCP of Fig. 7 and Fig. 9 as a data table. Fig. 7 to 10 show the visualization of the evaluation to each item of the images those anger items are rated as ”1: totally disagree” or ”2: disagree” by worker A and worker C, even though these face images belonged to ”anger”. From Fig. 7 and Fig. 8, we can estimate that worker A tended to rate as neutrality the images which he annotated as ”disagree” even though they belonged to anger. On the other hand, Fig. 9 and Fig. 10 show that worker C tended to rate as ”sadness” the images annotated as ”disagree” even though they belonged to anger. Thus, we can also see a tendency for each worker about images that were annotated differently than originally assumed. To summarize, we can understand the tendency for each worker or for each item by visualizing the evaluation results by focusing on images that satisfy specific conditions.

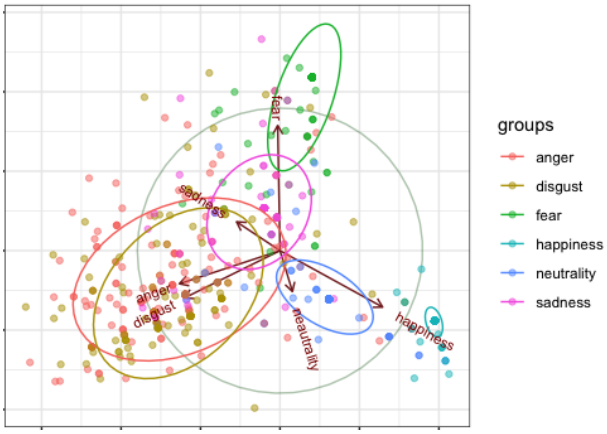


Fig. 5. Visualization of training data of worker B by PCA.

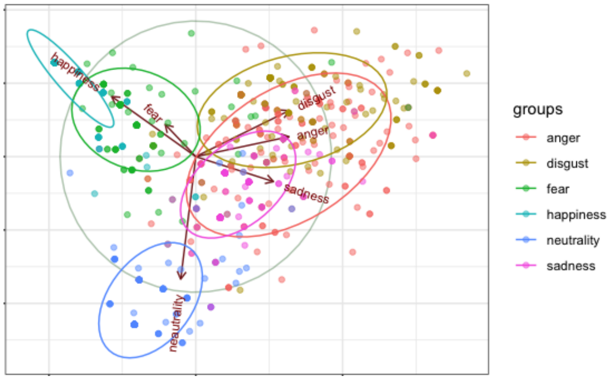


Fig. 6. Visualization of training data of worker C by PCA.

V. CONCLUSION AND FUTURE WORK

This paper proposed a technique for visualizing training data constructed by subjective annotation tasks. We discovered the relationship between the reliability of the data and the worker’s tendency and discussed the directions to achieve reliable annotation in this study. As a result, we found the following from our experiments.

1) Uncertainty and variation in the annotation of training data tends to occur at the beginning of the work or when the time required for annotation was long. This affected the reliability of training data.
2) All workers had unconfident annotations of disgust and anger and the α values for these items were lower in our experiment. This indicates that these two items were difficult to annotate.
3) Each worker had characteristics on images that were annotated differently from the original categorization.

Our future prospect is to establish a technique to improve the reliability of the training data based on the results of this study. In this study, we analyzed how the elapsed time of the annotation task and the time required to annotate a single image affected the reliability of the data. We also found groups of images that were difficult to annotate while observing the

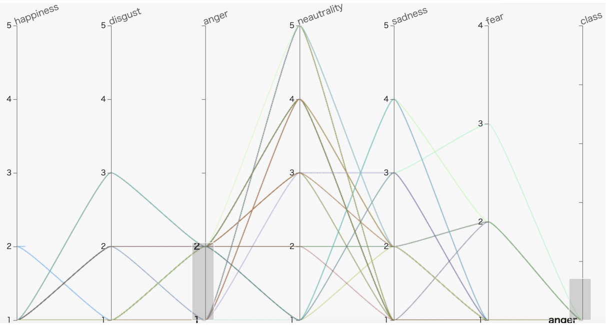
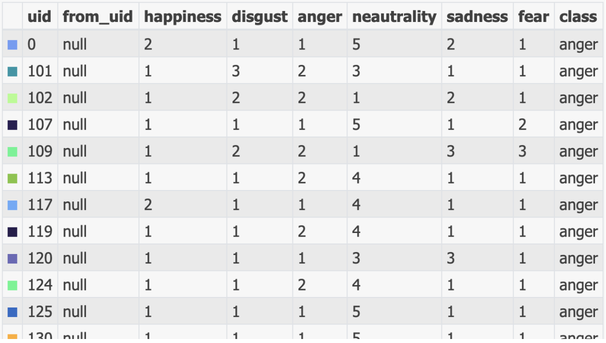


Fig. 7. Visualization result of worker A by PCP.



	uid	from_uid	happiness	disgust	anger	neautrality	sadness	fear	class
■	0	null	2	1	1	5	2	1	anger
■	101	null	1	3	2	3	1	1	anger
■	102	null	1	2	2	1	2	1	anger
■	107	null	1	1	1	5	1	2	anger
■	109	null	1	2	2	1	3	3	anger
■	113	null	1	1	2	4	1	1	anger
■	117	null	2	1	1	4	1	1	anger
■	119	null	1	1	2	4	1	1	anger
■	120	null	1	1	1	3	3	1	anger
■	124	null	1	1	2	4	1	1	anger
■	125	null	1	1	1	5	1	1	anger
■	130	null	1	1	1	5	1	1	anger

Fig. 8. Data table of the items selected in Fig. 7.

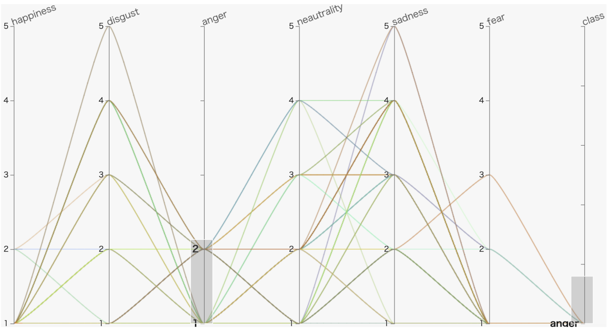
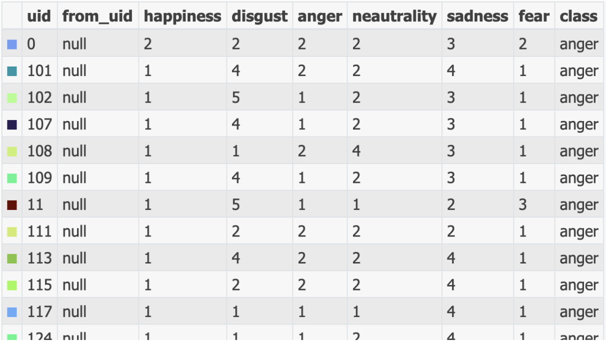


Fig. 9. Visualization result of worker C by PCP.



	uid	from_uid	happiness	disgust	anger	neautrality	sadness	fear	class
■	0	null	2	2	2	2	3	2	anger
■	101	null	1	4	2	2	4	1	anger
■	102	null	1	5	1	2	3	1	anger
■	107	null	1	4	1	2	3	1	anger
■	108	null	1	1	2	4	3	1	anger
■	109	null	1	4	1	2	3	1	anger
■	11	null	1	5	1	1	2	3	anger
■	111	null	1	2	2	2	2	1	anger
■	113	null	1	4	2	2	4	1	anger
■	115	null	1	2	2	2	4	1	anger
■	117	null	1	1	1	1	4	1	anger
■	124	null	1	1	1	2	4	1	anger

Fig. 10. Data table of the items selected in Fig. 9.

visualization results. We would like to examine how much the reliability of the training data will improve by preferentially correcting images that are calculated to have lower confidence values or analyzed to be difficult to annotate. In addition, we would like to explore visualization techniques to make it easier to examine the trend of the annotations. As a long-term issue, we will conduct further observations while increasing the number of workers.

Acknowledgment

This work was supported by Japan Society for the Promotion of Science (JSPS). We appreciate participants who contributed to the annotation tasks.

References

[1] Teramoto, T., “Mechanism and Operation for Creating Highly Accurate Teacher Data”, Speaker Deck. 2019. https://speakerdeck.com/abeja/afalsetesiyondejing-du-falsegao-ijiao-shi-detawozuo-cheng-suruwei-nibi-yao-nashi-zu-mi, (referenced 2021-05-29).

[2] Itoh, T., ”Visualization of Individual Variation of Multiple AnnotatorsWorking on Training Datasets for Machine Learning”, IEEE VIS, Posters, 2019.

[3] Barrett, L. F., Gross, J., Christensen, T. C. and Benvenuto, M., ”Knowing What You’re Feeling and Knowing What to Do About It: Mapping the Relation Between Emotion Differentiation and Emotion Regulation”, Cognition and Emotion, 15, 713-724, 2001.

[4] Niwa, A. and Matsuda, H., “A Study for Sentiment Analysis Accepting the Diverse of Emotion Sensitivities”, Proceedings of the Annual Conference of JSAI, 2021.

[5] Dawid, A. P. and Skene, A. M., ”Maximum Likelihood Estimation of Observer Error-rates using the EM algorithm”, J. Royal Statistical Society, Series C (Applied Statics), 28(1), 20-28, 1979.

[6] Mitsuda, K., Iida, R. and Tokunaga, T., ”Detecting Missing Annotation Disagreement using Eye Gaze Information”, Proceedings of the 11th Workshop on Asian Language Resources, 19-26, 2013.

[7] Komatani, K., Okada, S., Nishimoto, H., Araki, M. and Nakano, M., ”Multimodal Dialogue Data Collection and Analysis of Annotation Disagreement”, International Workshop on Spoken Dialogue Systems (IWSDS), 2019.

[8] Fleiss, J. L., ”The Measurement of Interrater Agreement”, Statistical methods for rates and proportions, 212-236, 22-23, 1981.

[9] $Krippendorff, K., \backslash Computing Krippendorff'salpha - reliability”, https://repository.upenn.edu/asc_papers/43/$, 2011.

[10] Inagaki, K., Yoshikawa, T. and Furuhashi, T., “A Study on Extraction of Minority Groups in Questionnaire Data based on Spectral Clustering”, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 988-993, 2014.

[11] Ebner, N., Riedinger, M. and Lindenberger, U., “FACES–a database of facial expression in young, middle-aged, and older women and men: development and validation”, Behavior Research Methods, 42, 1, 351–362, 2010.

[12] Kodama, T., Tanaka, R. and Kurohashi, S., “Dialogue Management by Estimating User’s Internal State Using the Movie Recommendation Dialogue”, Natural Language Processing, 28(1), 104-135, 2021.

[13] Landis, J. R. and Koch, G.G., “The Measurement of Observer Agreement for Categorical Data. “ Biometrics. 33, 159-174, 1977.

[14] Tsushima, E., ”Intra-class Correlation Coefficient as a Reliability Index”, http://www.hs.hirosaki-u.ac.jp/pteiki/research/stat/icc.pdf.

[15] Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A., NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw. 2014;61:1–36.

[16] Haziza, D., Rapin, J. and Synnaeve, G., “Hiplot, interactive high-dimensionnality plots”, https://github.com/facebookresearch/hiplot, 2020.