An Empirical Study on the Effect of Quick and Careful Labeling Styles in Image Annotation

Author 1*
Organization or School

Author 1*
Organization or School

Author 1*
Organization or School

Abstract

Giving a label to difficult data takes a lengthy amount of time, particularly when non-expert annotators try to select the best possible label. However, there are no detailed studies exploring a label selection style during annotation. It can be very important and may affect the annotation efficiency and quality. In this study, we explore the effects of labeling styles in data annotation and machine learning. We ran an empirical study comparing "quick labeling" and "careful labeling" styles in image labeling tasks with three levels of difficulty. We also ran a machine learning experiment using the labeled images from the two labeling styles. The results showed that quick and careful labeling styles have both advantages and disadvantages in annotation efficiency, label quality, and machine learning performance. Specifically, it shows that careful labeling improves label accuracy only when the task is moderately difficult whereas it is a waste of time when the task is easy or extremely difficult.

Keywords: Cognitive Psychology, Labeling Style, Non-Expert Data Annotation, Data Collection, Machine Learning.

Index Terms: • Computing methodologies~Artificial intelligence ~Philosophical/theoretical foundations of artificial intelligence ~Cognitive science

1 Introduction

A large data size with high quality is always necessary to obtain better machine learning results. However, it is too costly if we recruit a large number of expert annotators (who have sufficient domain knowledge) to work on it. Recruiting non-expert annotators (typically, crowd workers) is cheaper and easier, so it is often the only viable option in practice [20, 21]. However, the label quality is a critical issue in non-expert data annotation (i.e. crowdsourcing tasks) [22, 23, 35]. Various annotation methods and tools have been introduced for addressing this issue [13, 14, 29, 32, 40]. However, there are no detailed studies looking into this issue from a human perspective (i.e. cognitive psychology) such as to investigate the effect of a label selection style during annotation (i.e. how a user makes a label decision). This can be important and may affect annotation efficiency and quality. We expect that different labeling styles used by annotators affect annotation efficiency as well as annotation quality.

This paper presents an empirical study comparing two labeling styles (quick labeling and careful labeling) for a manual image labeling task, with three datasets under different levels of data difficulty, easy (MNIST), moderately difficult (Fashion-MNIST), and extremely difficult (Kuzushiji-MNIST). We then conducted a machine learning experiment by using the labeled images using quick labeling and careful labeling styles and compared the machine learning results (classification accuracy), see Figure 1.

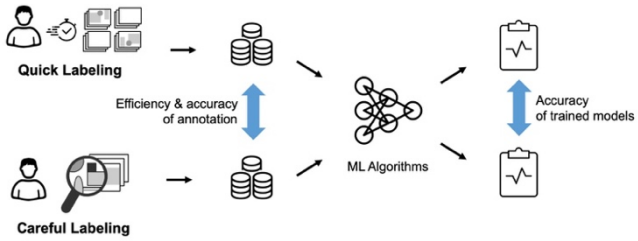

Figure 1: Research Overview.

These results demonstrated that a labeling style affects annotation efficiency (task completion time) and label accuracy. The careful labeling style showed significantly higher accuracy than the quick labeling style when the data set is moderately difficult. Moreover, the results of the machine learning experiment showed that the labeled images (training data) collected via the careful labeling style can also achieve better machine learning performance (higher classification accuracy) than the labeled image collected via the quick labeling style. However, the results showed that the careful labeling style did not bring benefits when the images are easy (i.e. label accuracy is already high via the quick labeling style) or extremely difficult (i.e. the careful labeling style cannot significantly improve label accuracy). We discussed the effects of the two labeling styles according to three different levels of data difficulty (easy, moderately difficult, and extremely difficult), and discussed three factors that need to be carefully considered when selecting an appropriate labeling style for an annotation task. This study makes the following contributions:

- Identifying labeling styles as a variable in non-expert image annotation and machine learning.
- An empirical study comparing quick labeling and careful labeling styles, demonstrating the benefits of the careful labeling style in an image annotation task.
- A machine learning experiment using labeled images via different labeling styles, demonstrating the effects of labeling styles in image classification accuracy.

* email address

LEAVE 0.5 INCH SPACE AT BOTTOM OF LEFT COLUMN ON FIRST PAGE FOR COPYRIGHT BLOCK

2 RELATED WORK

2.1 Manual Data Annotation and Challenge

A manual data annotation is a basic practice for machine learning. Large size is often necessary to improve the machine learning results. Some popular datasets including ImageNet [28], AudioSet [30], and YouTube-8M [31], all of which were manually labeled by human annotators, were applied. The manual data annotation process is extremely tedious and time-consuming, and many studies have proposed annotation tools for assisting the tedious process of manual data annotation. For example, LabelMe [10] is a web-based image annotation tool that allows multiple annotators to label an image and share their labeling results instantly. ESP [11] is an image annotation tool combined with a computer game that provides an enjoyable labeling process for annotators, whereas TagATune [27] is an audio annotation tool that shares the same idea. VIA [12] is an annotation tool that allows annotators to define and describe spatial regions in images, audio segments, and video frames, and iVAT [32] is a video annotation tool that supports manual, semi-automatic, and automatic video annotation.

Most of these tools provide supportive, efficient, and enjoyable systems for improving the tedious processes of manual data annotation. These tools generally assume that the annotators have sufficient domain knowledge in the labeling tasks. However, data annotation tasks often rely on non-expert annotators who lack sufficient domain knowledge because access to a sufficient number of expert annotators is limited and costly [18, 19]. In such cases, labeling tasks can be significantly difficult for non-expert annotators, and the labeled data may contain numerous errors [23] [33] [36]. These annotation tools may not be able to address this issue when the annotators are non-experts.

2.2 Annotation Workflows for Improving Label Quality

Many data annotation workflows have been proposed for improving the label quality, particularly for annotation tasks conducted through crowdsourcing. Revolt [13] is a collaborative crowdsourcing labeling workflow that applies concepts from expert annotation workflows (label-check-modify). This specific workflow can produce a higher label quality than a traditional labeling workflow. Pairwise HITS [14] is a labeling workflow for a quality estimation that allows annotators to compare a pair of labeled data and select a better one. Fang et al. [32] introduced a two-round workflow to improve the quality of the crowdsourced image labeling. In the first round, annotators select a label for the target images (several labels might be assigned to each image). In the second round, annotators are required to select the best label for each image (referring to the results from other annotators). In addition, Baba [29] introduced two types of labeling workflows (parallel and interactive) that allow multiple annotators to be involved in an annotation task in different ways to improve the label quality. In addition, several studies used the concept of hierarchical classification in a data annotation to increase the labeling efficiency and label quality [15] [16].

The main concept of these annotation workflows is the gathering of knowledge from multiple people. This is a typical workflow for improving the label quality by involving a group of annotators (non-exerts or experts) to collaborate together for a labeling task. In this study, we aim to explore the "labeling styles" rather than the "labeling workflows" during a data annotation task. In addition to machine learning, data annotation has been used as an approach in various research areas. For example, social scientists annotate data to discover interesting phenomena and establish theories [7] [9], and data annotation is often used in the process of analyzing qualitative data such as a thematic analysis approach [8]. Data annotation is not only a labeling process but also a cognitive process by which annotators view data, organize concepts, and make labeling decisions. Concept organization plays an important role in data annotation, and Kulesza et al. [17] indicated that annotators often organize their conceptual similarity by observing more items during data annotation. Chang et al. [40] shared the same concept and proposed a spatial layout labeling interface for a concept organization during the annotation process. We believe that cognitive processes (i.e., how to make a labeling decision) in a manual data annotation task is important and might affect the label quality and cost.

2.3 Intuitive and Systematic Decision-Making

A cognitive style (or thinking style) is a term used in cognitive psychology to describe how individuals organize and process information, and finally make decisions [1] [2] [3]. Intuitive and systematic decision-making are two types of cognitive styles. Intuitive decision-making is known as a type of associative thinking that relies on intuition, and systematic decision-making is known as a type of rule-based thinking that relies on a logical evaluation [4]. Both cognitive styles have advantages and disadvantages. For example, intuitive decision-making requires a shorter time than systematic decision-making. However, systematic decision-making contains a deeper consideration process than intuitive decision-making.

These two cognitive styles have been used and discussed in various fields. Sagiv et al. [3] conducted a study analyzing different intuitive and systematic cognitive styles used by art, accounting, and mathematics students. Their study found that different students prefer different cognitive styles in their class, and cognitive style is consistent with an individual's personal attributes. Ma-Kellams and Lerner [5] compared intuitive and systematic cognitive styles in the understanding of other people's feelings and found that a systematic cognitive style can produce better empathic accuracy than an intuitive cognitive style. In addition, Hwang and Lee [6] explored the impacts of intuitive and systematic cognitive styles of consumers on their visual attention patterns in online shopping environments. The results showed that consumers pay different visual attention to web pages when they use different cognitive styles to make purchasing decisions. Such studies showed the effects of the cognitive styles used in various activities. In this study, we share a similar concept to explore the effects of the cognitive styles (labeling styles) in a manual data annotation, where annotators complete a labeling task (i.e. select an appropriate label for an image) through intuitive (quick labeling) and systematic decision-making (careful labeling).

3 LABELING STYLE AND USER INTERFACE

We define two labeling styles of manual image annotation based on the theory of cognitive psychology: quick labeling and careful labeling.

Quick labeling. The style of quick labeling refers to the concept of intuitive decision-making. In this labeling style, annotators are asked to select a label for an image as quickly as possible even when they lack confidence in the target image and label. They are strongly encouraged to select a label within a short time.

Careful labeling. The style of careful labeling refers to the concept of systematic decision-making. In this labeling style, annotators are asked to select a label for an image as carefully as possible particularly when they are not confident with the target image and label. They are strongly encouraged to spend sufficient time before making a label decision.

A labeling system was developed to evaluate the quick labeling and careful labeling styles in a manual image annotation task. Figure 2 shows a screenshot of the image labeling interface. The left side of the interface lists the labels (10 labels/categories in Fashion-MNIST dataset), and the right side of the interface

presents the target image. During the labeling task, annotators were asked to select an appropriate label from the label list and apply it to the target image. After selecting a label for an image (by clicking a label), the system automatically moves to the next image. The annotators are not allowed to return to the previous images after selecting a label for an image.

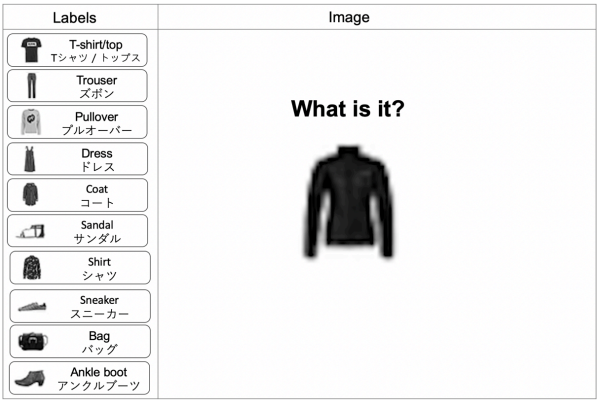

Figure 2: Screenshot of the labeling interface.

4 User Study

A user study was conducted to compare the quick labeling and careful labeling styles applied to an image labeling task. Our goal is to observe the effects of the two labeling styles in the image labeling tasks, especially in the three different levels of data difficulty (easy, moderately difficult, and extremely difficult). We compared the quick labeling and careful labeling styles in terms of label accuracy and labeling time of the given image labeling tasks.

4.1 Apparatus

In order to control user study quality, especially the use of quick labeling and careful labeling styles during the image labeling tasks. We outsourced the execution of the user study to a professional company, and the company asked their employees to participate in the user evaluation process as a part of their job. The total cost was approximately $1,600, i.e., $640 for the quick labeling task ($53 per participant) and $960 for the careful labeling task ($80 per participant). During the user study, the participants were asked to sit in front of a desktop computer and complete the given image labeling tasks (Figure 3).

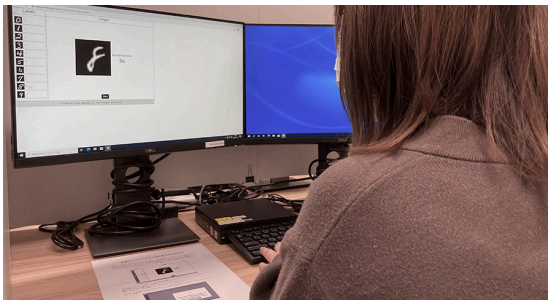

Figure 3: Photography of user study.

4.2 Participants

A total of 24 participants (12 men and 12 women, 18 to 49 years in age) were invited by the company to participate in the user study. All the participants are Japanese (i.e. understand Japanese Hiragana letters). Most of the participants (n = 19) had no prior experience with data annotation, four had less than half a year of experience, and one had between half to a full year of experience.

4.3 Dataset

Three datasets, MNIST [24], Fashion-MNIST [25], and Kuzushiji-MNIST [26], were used for the labeling tasks in the user study. Each dataset contains 60,000 training images and 10,000 testing images in ten categories (labels). Figure 4 shows the ten categories for each dataset.

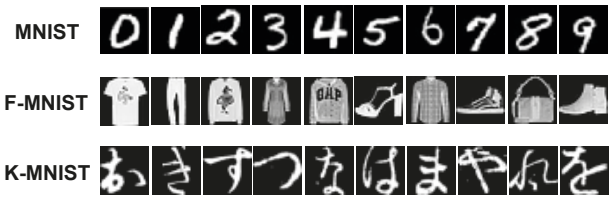

Figure 4: Ten categories in MNIST, F-MNIST, and K-MNIST datasets.

These three datasets have varying levels of difficulty. MNIST is an "easy" dataset. These handwritten number digits are not difficult for human users to recognize even they are in difficult cases, see Figure 5 (a). Fashion-MNIST is a "moderately difficult" dataset, which is more difficult than the MNIST dataset It is because it contains some difficult (confusing) items (e.g. "Pullover" and "T-shirt/Top"). Figure 5 (b) shows examples of easy and difficult (confusing) cases. Kuzushiji-MNIST is an "extremely difficult" dataset, which is much more difficult than the MNIST and Fashion-MNIST datasets. These handwritten Japanese Hiragana letters are very difficult to be recognized (even for Japanese users), especially when the letters are in difficult cases. Figure 5 (c) shows examples of easy and difficult cases.

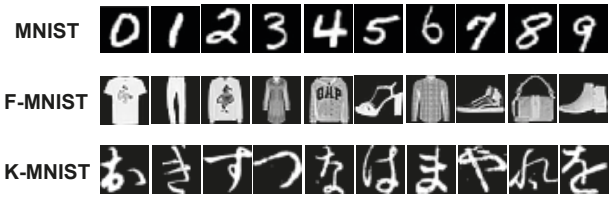

Figure 5: Examples of easy (upper part) and difficult (lower part) cases from MINIST, F-MNIST and K-MNIST.

We randomly selected 100 images (10 images for each label) from each dataset. We then created twelve 100-image datasets (e.g., Datasets 1–12) in which none of the images overlapped. These twelve 100-image datasets were used for the 12 participants in the quick labeling task and 12 participants in the careful labeling task (the same datasets were used for both labeling styles).

4.4 Task and Condition

The image labeling tasks involved labeling 300 images (100 images for each dataset) by each participant. During the image labeling task, the participants were requested to select an appropriate label from a 10-category list (10 labels) for each image. A between-subject method was used, by which 12 of the participants were asked to complete the labeling task using the quick labeling style and the other 12 participants were asked to complete the labeling task using the careful labeling style.

Quick labeling condition. The participants of the quick labeling task were provided the following instructions for the labeling task:

"Please select a label for an image as quickly as possible. In this case, if you are not sure about the target images and labels,

please just select the most appropriate one (i.e., make a guess based on your intuition). Please do not spend too much time considering before making a label selection. We STRONGLY ENCOURAGE you to select a label for an image within 5 SECONDS. A 5-second timer is provided on the labeling interface. After selecting a label for an image, the system will automatically show the next image. This means that you are not allowed to change your selected label."

Our labeling system for the label-quick task contains a 5 s timer for labeling each image. An alert message "Time's up! Please select a label now" is shown when the timer ends. See Figure 6.

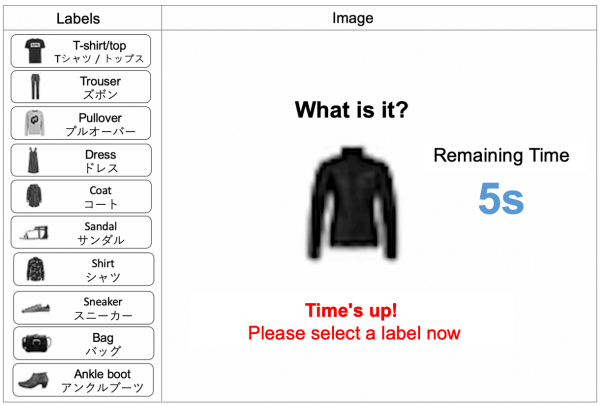

Figure 6: Timer and alert message in the labeling system.

The system does not automatically move to the next image until the participants select a label for the target image, even the timer ends. The timer was designed as a reminder for the participants during annotation. In addition, the participants were not allowed to return to the previous image after selecting a label for the image.

Careful labeling condition. The participants in the careful labeling task were provided the following instruction for the task:

"Please select a label for an image as carefully as possible. There is no time limitation for labeling each image. You have sufficient time to think carefully before making a label decision, particularly when the target images are difficult or when you are not confident with the images and labels. We STRONGLY ENCOURAGE you to spend sufficient time (there is no time limitation) before making a label decision. After selecting a label for an image, the system will automatically show the next image. This means that you are not allowed to change your selected label."

The labeling system for the careful labeling task is the same as the labeling system used for the quick labeling task, the only difference being that there is no timer shown on the labeling interface (see Figure 2). The participants were allowed to spend as much time as necessary for making a label selection. In addition, the participants were not allowed to return to the previous image after selecting a label for an image. The participants were informed that they can receive 1.5 times higher rewards for the careful labeling task than a normal labeling task (i.e. quick labeling task).

4.5 Procedure

An oral overview and detailed written instructions were provided by an instructor to the participants. The evaluation itself is composed of three parts (in order): instruction and trial (5–10 min), labeling tasks (20–45 min), and questionnaire (3–5 min). The entire evaluation process was completed between 40 and 60 min (depending on the labeling styles). After providing instructions on the labeling interfaces and the given tasks, the participants were allowed to practice on a small labeling task (to label three images for each dataset that differed from the images used in the formal tasks) before starting the given image labeling tasks.

4.6 Measurement

Our labeling system automatically recorded and measured the time and accuracy of the image labeling tasks completed by the participants. The timer started when the participants clicked on "START" and stopped when they clicked on "FINISH." In addition, the system recorded the time spent by the participants for each image labeling during the image labeling process. Following the image labeling tasks, the participants were asked to answer a questionnaire regarding the labeling process. The questionnaire contained 3 Likert scale questions for each labeling style (see section 6.4).

5 Machine Learning Experiment

In the user study, a total of 7,200 labeled images were collected (2,400 for each dataset and 1,200 for each labeling style in a dataset). The training dataset contains errors made by the participants (i.e., the accuracy rate of the training data is not 100%). We used them as training data to run a machine learning experiment to evaluate the effects of the labeling styles (data collected via the quick labeling and careful labeling styles) in machine learning accuracy (image classification). Three common machine learning algorithms (logistic regression, K-nearest neighbors, and support vector machine) were selected for the case study in the machine learning experiment. We did not use more advanced techniques (e.g. deep learning) because the training dataset is too small and our goal is not to pursue high machine learning accuracy but compare the difference between the two labeling styles. The testing data used in the machine learning experiment were 30,000 images (10,000 for each dataset), which are different from the training dataset (7,200 labeled images).

6 Results

6.1 Task Completion Time

Figure 7 shows the task completion time for different labeling styles and datasets. The results from the quick labeling task showed that the participants spent an average of 4 min and 23 s, 5 min and 6 s, and 5 min and 37 s to label the 100 images with MNIST, Fashion-MNIST, and Kuzushiji-MNIST datasets, respectively. The results showed that the participants spent an average of 4 min and 42 s, 6 min and 58 s, and 10 min and 13 s to label the 100 images in the careful labeling task. The results of an unpaired t-test on the task completion time showed that the difference was insignificant ($p > 0.05$) between the quick labeling and careful labeling styles in the MNIST dataset, whereas there were significant differences ($p < 0.01$) in the Fashion-MNIST and Kuzushiji-MNIST datasets. This indicates that the careful labeling style requires a longer time to complete the labeling tasks than the quick labeling style when the images are moderately difficult and extremely difficult. However, the task completion time is comparable between the quick and careful labeling styles when the images are easy.

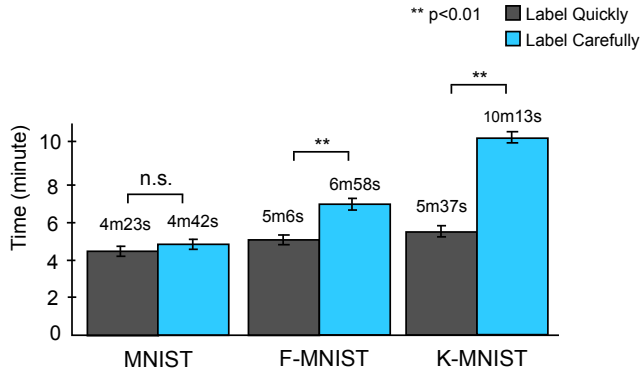

Figure 7: Task completion time. MNIST. LQ: mean = 4.23; SD = 0.73; LC: mean = 4.42; SD = 0.61. F-MNIST. LQ: mean = 5.06; SD = 0.76; LC: mean = 6.58; SD = 1.13. K-MNIST. LQ: mean = 5.37; SD = 0.90; LC: mean = 10.13; SD = 3.55.

6.2 Annotation Accuracy

Figure 8 shows the accuracy of the labels given by the participants in the quick labeling and careful labeling styles. The results from the quick labeling task showed that the accuracy was 97.58%, 72.08%, and 58.08% with the three datasets, respectively, and the results from the careful labeling tasks showed accuracy of 97.58%, 76.83%, and 60.08%, respectively. The analysis of the accuracy using an unpaired t-test showed that the difference was insignificant ($p > 0.05$) in the MNIST and Kuzushiji-MNIST dataset, whereas there was a significant difference ($p < 0.05$) between the two labeling styles in the Fashion-MNIST dataset. This indicates that the careful labeling style can help non-expert annotators to select labels more correctly when the images are moderately difficult (Fashion-MNIST), whereas no clear benefit was observed when the images were easy (MNIST) and extremely difficult (Kuzushiji-MNIST). This indicated that it is a waste of money for conducting a careful labeling task when the task is easy or too difficult.

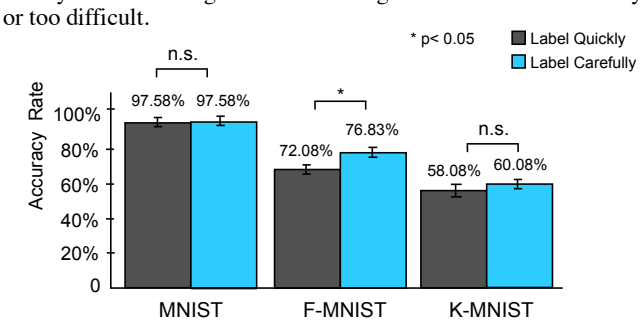

Figure 8: Accuracy of the labeling tasks. MNIST. LQ: mean = 97.58; SD = 1.44; LC: mean = 97.58; SD = 1.56. F-MNIST. LQ: mean = 72.08; SD = 4.17; LC: mean = 76.83; SD = 6.64. K-MNIST. LQ: mean = 58.08; SD = 6.53; LC: mean = 60.08; SD = 7.82.

6.3 Temporal Effect

Task Completion Time

Figure 9 shows the average time for the labeling process for the first half (1-50 images) and second half (51-100 images) in the three datasets, using the quick labeling style. The results showed that the participants spent an average of 2 min 6 s and 2 min 17 s to complete the first half and second half in the MNIST dataset, 2 min 36 s and 2 min 30 s in the Fashion-MNIST dataset and 2 min 29 s and 3 min 8 s in the Kuzushiji-MNIST dataset. The results of paired t-test showed that the difference was not significant ($p > 0.05$) between the first half and second half of the labeling process in the MNIST and Fashion-MNIST datasets, but the difference was significant ($p < 0.05$) in the Kuzushiji-MNIST dataset. This indicates that there is a temporal effect in all levels of data difficulty when using the quick labeling style. However, the results interestingly showed that the participants even spent a longer time completing the second half of the image labeling task.

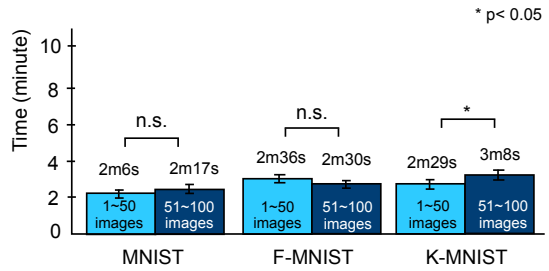

Figure 9: Average time of first and second 50 images in the quick labeling style. MNIST (1-50 images): mean = 2.06; SD = 0.38; MNIST (51-100 images): mean = 2.17; SD = 0.35. F-MNIST (1-50 images): mean = 2.36; SD = 0.28; F-MNIST (51-100 images): mean = 2.30; SD = 0.48. K-MNIST (1-50 images): mean = 2.29; SD = 0.21; K-MNIST (51-100 images): mean = 3.08; SD = 0.27.

Figure 10 shows the average time for the labeling process for the first half (1-50 images) and the second half (51-100 images) using the careful labeling style in different datasets. The results showed that the participants spent an average of 2 min 24 s and 2 min 18 s to complete the first half and second half in the MNIST dataset, 4 min 5 s and 2 min 53 s in the Fashion-MNIST dataset, and 4 min 21 s and 5 min 52 s in the Kuzushiji-MNIST dataset. The results of paired t-test showed that the difference was not significant ($p > 0.05$) between the first half and second half of the labeling process in the MNIST datasets, but the difference was significant ($p < 0.05$) in the Fashion-MNIST and Kuzushiji-MNIST dataset. This indicates that there is no temporal effect in the use of the careful labeling style when the images are easy. However, there was a temporal effect when the images are moderately difficult. The participants significantly increased their labeling speed in the second half of the image labeling task. In addition, the participant spent more time for the second half when the images are extremely difficult, which is the same as the quick labeling style.

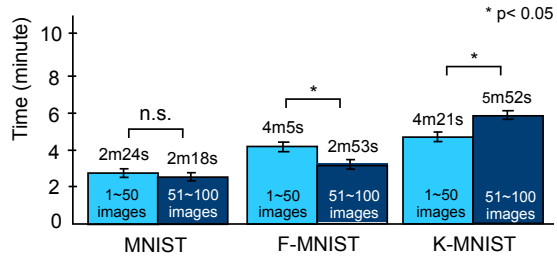

Figure 10: Average time of first and second 50 images in the careful labeling style. MNIST (1-50 images): mean = 2.24; SD = 0.33; MNIST (51-100 images): mean = 2.18; SD = 0.29. F-MNIST (1-50 images): mean = 4.05; SD = 0.57; F-MNIST (51-100 images): mean = 2.53; SD = 0.38. K-MNIST (1-50 images): mean = 4.21; SD = 1.02; K-MNIST (51-100 images): mean = 5.52; SD = 0.73.

Annotation Accuracy

Figure 11 shows the accuracy of the labeling process in the first half (1-50 images) and the second half (51-100 images) via the

quick labeling style. The results showed that the accuracy rate was 97.5% and 97.67% in the first half and second half in the MNIST dataset, 73.17% and 71% in the Fashion-MNIST dataset, and 56.67% and 59.5% in the Kuzushiji-MNIST dataset. The results of paired t-test showed that the difference was not significant ($p > 0.05$) between the first half and second half of the labeling process in all datasets. This indicates that there is no temporal effect in all levels of data difficulty when using the quick labeling style. The label accuracy was not significantly affected (improved) by different labeling styles.

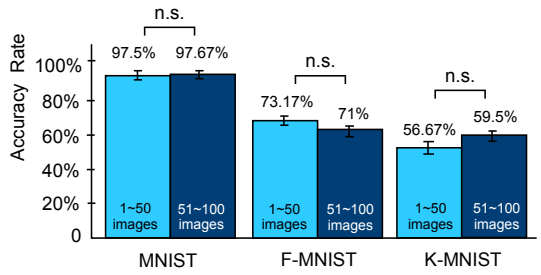

Figure 11: Average time of first and second 50 images in the quick labeling style. MNIST (1-50 images): mean = 97.50; SD = 1.27; MNIST (51-100 images): mean = 97.67; SD = 1.53. F-MNIST (1-50 images): mean = 73.17; SD = 3.93; F-MNIST (51-100 images): mean = 71; SD = 4.31. K-MNIST (1-50 images): mean = 56.67; SD = 6.95; K-MNIST (51-100 images): mean = 59.50; SD = 5.47.

Figure 12 shows the accuracy of the labeling process in the first half (1-50 images) and the second half (51-100 images) via the careful labeling style. The results showed that the accuracy rate was 97.17% and 98% in the first half and second half in the MNIST dataset, 73.67% and 80% in the Fashion-MNIST dataset, and 62.83% and 57.33% in the Kuzushiji-MNIST dataset. The results of paired t-test showed that the difference was not significant ($p > 0.05$) between the first half and second half of the labeling process in the MNIST and Kuzushiji-MNIST datasets. However, the difference was significant ($p < 0.05$) in the Fashion-MNIST dataset. This indicates that there is no temporal effect when the images are easy and extremely difficult, with the use of the careful labeling style. However, there was a temporal effect when the images are moderately difficult. The participants can significantly improve label accuracy in the second half of the image labeling task.

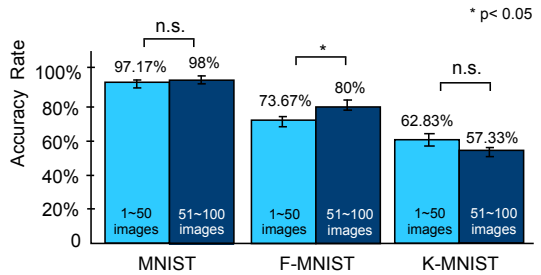

Figure 12: Average time of first and second 50 images in the careful labeling style. MNIST (1-50 images): mean = 97.17; SD = 1.83; MNIST (51-100 images): mean = 9; SD = 1.35. F-MNIST (1-50 images): mean = 73.67; SD = 4.95; F-MNIST (51-100 images): mean = 80; SD = 6.58. K-MNIST (1-50 images): mean = 62.83; SD = 8.13; K-MNIST (51-100 images): mean = 57.22; SD = 7.22.

To sum up, the careful labeling style brings temporal effect during the labeling process in the task completion time and accuracy rate (i.e. reduce time and increase accuracy) ONLY when the images are moderately difficult. If the images are too easy there is no temporal effect. Interestingly, if the images are too difficult task completion time is longer in the second half.

6.4 Questionnaire

Figure 13 displays how confident the participants felt in the given image labeling tasks. In the MNIST dataset, the results showed that most of the participants felt very confident and or confident when selecting a label for an image using either the quick labeling ($n = 10$) or careful labeling ($n = 9$) styles, whereas none of the participants felt not very confident or not confident. For the Fashion-MNIST dataset, the results showed that only one participant felt very confident when selecting a label for an image through the quick labeling style and only two participants felt confident when selecting a label for an image through the careful labeling style. In the Kuzushiji-MNIST dataset, the results showed that no participants felt confident or very confident when selecting a label for an image through either the quick labeling or careful labeling styles. More of the participants of the careful labeling style felt not confident ($n = 4$) or not very confident ($n = 7$) in comparison to the participants of the quick labeling style (not confident, $n = 3$; not very confident, $n = 6$). This indicates that the labeling styles do not affect the subjective impression of the participants' confidence during annotation. However, the ambiguities in the data do affect the confidence of the participants in selecting a label for an image during annotation.

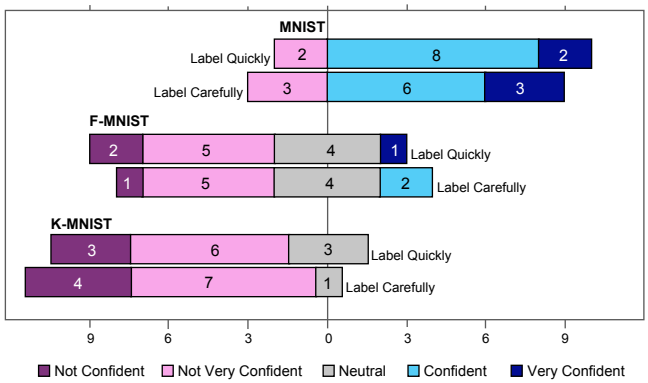

Figure 13: How confident the participants felt when selecting a label with MNIST, F-MNIST, and K-MNIST datasets?

6.5 Results of Machine Learning Experiment

Figure 14 shows the machine learning results (i.e. image classification accuracy) for the three datasets with the quick and careful labeling styles.

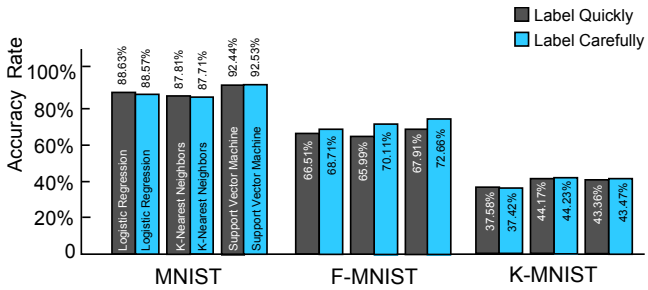

Figure 14: Accuracy of machine learning models in the MNIST, F-MNIST, and K-MNIST datasets.

In the data analysis, we were not computing accuracy for each participant because the data set is too small. We combined all the annotations and trained a model and measured the performance of the model.

MNIST Dataset. The accuracy of the training data was 97.58% for both the quick labeling and careful labeling styles (Figure 12). Based on the accuracy of the training data, the machine learning performance (accuracy) also showed that there were almost no differences between the two labeling styles (logistic regression, LQ = 88.63%, LC = 88.57%; K-nearest neighbors, LQ = 87.81%, LC = 87.71%; support vector machine, LQ = 92.44%, LC = 92.53%). These results were not surprising as the label accuracy of training data was totally the same.

Fashion-MNIST Dataset. The accuracy of the training data was 72.08% for the quick labeling style and 76.83% for the careful labeling style (Figure 12), which was a significant difference between the two labeling styles. Based on the accuracy, the machine learning performance (accuracy) also showed that there were differences between the two labeling styles (logistic regression, LQ = 66.51%, LC = 68.71%; K-nearest neighbors, LQ = 65.99%, LC = 70.11%; support vector machine: LQ = 67.91%, LC = 72.66%). The differences were between 2.2% and 4.12%. Although machine learning algorithms often work well even if the labels given to difficult data contain errors. Our results indicate that the improvement of label accuracy via the careful labeling style also can improve machine learning accuracy.

Kuzushiji-MNIST Dataset. The accuracy of the training data was 58.08% for the quick labeling style and 60.08% for the careful labeling style (Figure 12). It was a 2% small difference between the two labeling styles, but it was not significant according to a paired t-test analysis. Based on the accuracy of the training data, the machine learning performance (accuracy) showed that the differences were very small (almost the same) between the two labeling styles (logistic regression, LQ = 37.58%, LC = 37.42%; K-nearest neighbors, LQ = 44.17%, LC = 44.23%; support vector machine, LQ = 43.36%, LC = 43.47%). This indicates that the small difference of the label accuracy in the training data almost cannot affect machine learning performance.

7 Discussion

7.1 Effects of Labeling Styles in Annotation Efficiency and Label Quality

In psychology, decision-making is a cognitive process in which the cognitive styles of the individuals affect the decision-making process and the decision outcomes and quality [37] [38] [39]. In manual data annotation, a labeling style can be known as a decision-making process (i.e. select an appropriate label for an image) that may affect label outcomes and quality. In general, a quick labeling style requires less time to complete an annotation task, whereas the careful labeling style requires more time; and we normally expect that a careful labeling style bring higher quality data than a quick labeling style. However, it depends on data difficulty and annotation tasks. Our results demonstrate that there was no significant difference between the quick and carful labeling styles in the task completion time and label quality, when the data is easy (i.e. MNIST). However, it shows differences between the two labeling style when data is getting difficult. For example, the careful labeling style requires more time to complete a labeling task which contains moderately difficult images than the quick labeling style. At the same time, it significantly improves label quality. However, if a labeling task contains extremely difficult images the careful labeling style cannot improve label quality but requires much longer time to complete the labeling task. These results imply that a labeling style affects annotation efficiency (task completion time) and label quality (accuracy rate) in non-expert data annotation. However, the effects are dependent on the data difficulty of an annotation task. In addition, the questionnaire results showed that the subjective impression of annotator's confidence during annotation was not affected by the labeling styles in all the labeling tasks with different data ambiguities. However, annotator's confidence was affected by data ambiguities (higher confidence in less difficult data, lower confidence in more difficult data).

7.2 Temporal Effects in the Quick and Careful Labeling Tasks

Temporal effect has been used to analyze task performance during an image labeling task [16]. It describes how people change their behavior over time, which is a method to analyze the efficiency of an activity or a study [42, 43, 44]. Our results showed that there was a significant temporal effect ($p < 0.05$) during the labeling process via the careful labeling style when the images are moderately difficult images (Fashion-MNIST). The participants can reduce the task completion time in the second half of the image labeling task by using the careful labeling style. This indicates that the careful labeling style not only can improve label quality but also bring temporal effect during annotation in a labeling task containing moderately difficult images. Furthermore, there were significant temporal effects ($p < 0.05$) during annotation via the careful labeling style when the images are extremely difficult images (Kuzushiji-MNIST). But, the participants spent longer time to complete the second half of the image labeling task than the first half. In addition, there was no significant temporal effect ($p > 0.05$) during annotation via both the quick and careful labeling styles when the images are easy (MNIST). However, the reason of the temporal effect was not clearly demonstrated.

7.3 Effects of Labeling Styles in Machine Learning Performance

Data quality plays a critical role in machine learning. Our user study showed that the careful labeling style can significantly improve the label quality of the image labeling task which contains moderately difficult images, and slightly improves the label quality when the task contains extremely difficult images. Our machine learning experiment also showed similar results. The labeled data collected via the careful labeling style can bring better machine learning performance (higher accuracy) than the labeled data collected via the quick labeling style when the images are moderately difficult (Fashion-MNIST). However, the machine learning performance showed almost no difference between the labeled data collected via the two labeling styles when the images are easy (MNIST) and very small difference when the data is extremely difficult (Kuzushiji-MNIST). Machine learning algorithms often work well even if the labels given to difficult data contains errors. This means that different label quality may bring no difference in machine learning accuracy. In such cases, a labeling style may be not a variable in machine leaning (only in manual data annotation). However, our results showed that the improvement of label quality via the careful labeling style also can increase machine learning accuracy when the data is moderately ambiguities. This finding has shown that the careful labeling style can bring benefits for the both data annotation and machine learning. However, it depends on data ambiguities. Our machine learning experiment used basic algorithms that only can

shows labeling style is a variable in machine leaning performance. A machine learning experiment with advanced techniques (e.g. deep learning with a large-scale dataset) is still necessary to be implemented in the future.

7.4 Three Factors for Selecting an Appropriate Labeling Style for An Annotation Task

Our user study and machine learning experiment have shown that different labeling styles have its advantages and disadvantages in different annotation tasks. For example, conducting a careful labeling task is costlier (requires longer time to complete a task) than a quick labeling task, but it cannot guarantee the improvement of label quality in all levels of data difficulty. So, it is important for deciding a reasonable labeling style for an annotation task, otherwise it may be just a waste of time and money if the improvement is not clear. Here we discuss three factors that should be carefully considered when selecting a labeling style for an annotation task.

(1) Data Difficulty

Our results showed that data difficulty is a key factor that affects annotation results when using different labeling styles. For example, the careful labeling style only shows benefit (i.e. improve label quality) when the data is easy and extremely difficult, while the quick labeling study only shows benefit (i.e. require less time) when the data is easy. According to the results, we suggest that the quick labeling style is a reasonable choose when conduct an easy annotation task. However, when an annotation task contains difficult data the careful labeling style can be a worth choose for consideration.

(2) Annotator Type and Task Condition

Although this study only focused on non-expert annotators, we believe that annotator's experience (i.e. domain knowledge in the given task) is another important factor that may highly affect annotation results of using different labeling styles. For example, the data difficulty also depends on individual experience and subjective impression. We suggest that a qualification is important and necessary for recruiting annotators when conducting an annotation task with different labeling styles. In addition, task condition (e.g. crowd tasks and in-person task) is also a factor when deciding a labeling style for an annotation task. Crowdsourcing is a very popular approach for conducting annotation tasks such as Amazon Mechanical Turk [41]. However, the quality of crowd tasks is a critical issue what have been discussed for many years [22, 35]. We believe that this issue also happens when using different labeling styles in a crowd task. This is the main reason we recruited our participants (annotators) for the user study via a professional company. This helps us to explore the effects of labeling styles more precisely (in order to prove the research concept). However, crowdsourcing is still an indispensable approach for conducting an annotation task. We suggest that an online workflow should be carefully designed for controlling the annotation quality even with different labeling styles.

(3) Instruction for Implementing a Labeling Style

After deciding a labeling style, it is important to make sure that annotators can fellow and implement the labeling style precisely. In this study, we provided a textual and oral instruction of each labeling style by an instructor before start a formal task (also contain a trial). This is only for the user study. We believe that an instruction for using a labeling style is not enough in a realistic annotation task. It is because some annotators may inherently careful to fellow an assigned labeling style, but and some may inherently sloppy. In order to avoid this kind of bias, we suggest that a specific labeling workflow (or labeling interface) should be designed and provided to afford different labeling styles, or "FORCE" annotators to fellow specific steps. For example, to design a workflow that requires annotators to double-check or spend certain time for consideration before making a label decision when an annotation task conducted via a careful labeling style.

8 Limitation and Future Work

One limitation of this study is that the size of the training data (1,200 labeled images collected via each labeling style, for each dataset) was small in the machine learning experiment. This is the main reason for the much lower machine learning accuracy of our study in comparison with the benchmarks [24] [25] [26]. Another limitation is that we only used basic machine leaning algorithms for training and testing with our collected data. However, the main purpose of this study is not to pursue high accuracy of machine learning results, and instead to focus on the effects of the labeling styles. Our results show that the careful labeling style can improve the label accuracy in a manual data annotation as well as increase the machine learning accuracy. We believe that the labeling styles might have an even greater effect in large-scale labeling tasks and advanced machine learning techniques (e.g. deep learning).

In the future, we plan to run a large-scale user study via crowdsourcing and test more machine learning algorithms. In addition, we would like to explore more details regarding the careful labeling style. For example, the reason of the temporal effect and the effect of compensation were not clearly demonstrated in current study that we plan to investigate it. Another interesting possibility is the dynamic control of labeling styles during an annotation. If a system can judge the difficulty of each data item before an annotation, it might be possible to ask an annotator to use an appropriate labeling style (use a careful labeling style for data with moderate difficulty only).

9 Conclusion

In this study, we investigated the effects of labeling styles in non-expert data annotation and machine learning. We ran a user study to compare a quick labeling style and a careful labeling style for a manual image annotation task, and used the labeled data (as training data) to run machine learning experiment. Our results showed that labeling style is a variable in a data annotation process and machine learning performance. The careful labeling style improves label accuracy only when the task is moderately difficult whereas it is only increases the cost without improving accuracy when the task is easy or extremely difficult. These findings provide insights for helping annotators to select an appropriate labeling style for an annotation task. This can be an alternative solution for improving non-expert annotation.

References

This work was supported by ******** (removed for the review).

References

[1] Kozhevnikov, Maria. 2007. Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. Psychological bulletin 133, no. 3: 464. DIO: http://dx.doi.org/10.1037/0033-2909.133.3.464

[2] Lynna J. Ausburn, and Floyd B. Ausburn. 1978. Cognitive styles: Some information and implications for instructional design. Ectj 26, no. 4: 337-354. https://www.jstor.org/stable/30219783

[3] Lilach Sagiv, Adi Amit, Danit Ein-Gar, and Sharon Arieli. 2014. Not all great minds think alike: Systematic and intuitive cognitive styles. Journal of Personality 82, no. 5: 402-417. DIO: http://dx.doi.org/10.1111/jopy.12071

[4] Eliot R. Smith, and Jamie DeCoster. 2000. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and social psychology review 4, no. 2: 108-131. DIO: http://dx.doi.org/10.1207/S15327957PSPR0402_01

[5] Christine Ma-Kellams, and Jennifer Lerner. 2016. Trust your gut or think carefully? Examining whether an intuitive, versus a systematic, mode of thought produces greater empathic accuracy. Journal of personality and social psychology 111, no. 5: 674. DIO: http://dx.doi.org/10.1037/pspi0000063

[6] Yoon Min Hwang, and Kun Chang Lee. 2015. Exploring the impacts of consumers' systematic and intuitive cognitive styles on their visual attention patterns in online shopping environments: emphasis on eye-tracking method. International Journal of Multimedia and Ubiquitous Engineering 10: 175-182. DIO: http://dx.doi.org/10.14257/ijmue.2015.10.12.18

[7] Juliet Corbin, and Anselm Strauss. 2014. Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications.

[8] Virginia Braun, and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, no. 2: 77-101. DIO: https://doi.org/10.1191/1478088706qp063oa

[9] Kathleen M. MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. Cam Journal 10, no. 2: 31-36. DIO: https://doi.org/10.1177/1525822X980100020301

[10] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. International journal of computer vision 77, no. 1-3: 157-173. DOI: http://dx.doi.org/10.1007/s11263-007-0090-8

[11] Luis Von Ahn, and Laura Dabbish. 2004. Labeling Images with a Computer Game. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 319-326. DOI: http://dx.doi.org/10.1145/985692.985733

[12] Abhishek Dutta, and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the 27th ACM International Conference on Multimedia, pp. 2276-2279. DOI: http://dx.doi.org/10.1145/3343031.3350535

[13] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 2334-2346. DOI: http://dx.doi.org/10.1145/3025453.302604

[14] Takeru Sunahase, Yukino Baba, and Hisashi Kashima. 2017. Pairwise HITS: Quality Estimation from Pairwise Comparisons in Creator- Evaluator Crowdsourcing Process. In Thirty-First AAAI Conference on Artificial Intelligence. https://dl.acm.org/doi/abs/10.5555/3298239.3298383

[15] Naoki Otani, Yukino Baba, and Hisashi Kashima. 2015. Quality Control for Crowdsourced Hierarchical Classification. In 2015 IEEE International Conference on Data Mining, pp. 937-942. DOI: http://dx.doi.org/10.1109/ICDM.2015.83

[16] Chia-Ming Chang, Siddharth Deepak Mishra, and Takeo Igarashi. 2019. A Hierarchical Task Assignment for Manual Image Labeling. In 2019 IEEE Symposium on Visual Languages and Human -Centric Computing (VL/HCC), pp. 139 -143. DOI: http://dx.doi.org/10.1109/VLHCC.2019.8818828

[17] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 3075 – 3084. DOI: http://dx.doi.org/10.1145/2556288.2557238

[18] Stefanie Nowak and Stefan Rüger. 2010. How Reliable are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation. In Proceedings of the international conference on Multimedia information retrieval, pp. 557-566. DOI: http://dx.doi.org/10.1145/1743384.174347

[19] Roland Kwitt, Sebastian Hegenbart, Nikhil Rasiwasia, Andreas Vécsei, and Andreas Uhl. 2014. Do We Need Annotation Experts? A Case 17 Study in Celiac Disease Classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 454-461. Springer, Cham. DOI: http://dx.doi.org/10.1007/978-3-319-10470-6_57

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[21] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139-147. https://dl.acm.org/doi/10.5555/1866696.1866717

[22] Jiyin He, Jacco van Ossenbruggen, and Arjen P. de Vries. 2013. Do You Need Experts in the Crowd? A Case Study in Image Annotation for Marine Biology. In Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 57-60. https://dl.acm.org/doi/10.5555/2491748.2491763

[23] Donghui Feng, Sveva Besana, and Remi Zajac. 2009. Acquiring High Quality Non-Expert Knowledge from On-demand Workforce. In Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web), pp. 51-56. https://dl.acm.org/doi/10.5555/1699765.1699773

[24] Yann LeCun. 1998. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/

[25] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 https://arxiv.org/abs/1708.07747

[26] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. 2018. Deep learning for classical Japanese literature. arXiv preprint arXiv:1812.01718. DOI: https://dx.doi.org/10.20676/00000341

[27] Law, Edith LM, Luis Von Ahn, Roger B. Dannenberg, and Mike Crawford. 2007. TagATune: A Game for Music and Sound Annotation. In ISMIR, vol. 3, p. 2. DOI: https://dx.doi.org/10.5281/zenodo.1415568

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[29] Yukino Baba. 2018. Statistical Quality Control for Human Computation and Crowdsourcing. In IJCAI, pp. 5667-5671. DOI: https://doi.org/10.24963/ijcai.2018/806

[30] Jort F. Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776-780. IEEE. DOI: http://dx.doi.org/10.1109/ICASSP.2017.7952261

[31] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675. https://arxiv.org/abs/1609.08675

[32] Bianco Simone, Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2015. An interactive tool for manual, semi-automatic and automatic video annotation. Computer Vision and Image Understanding 131: 88-99. DOI: https://doi.org/10.1016/j.cviu.2014.06.015

[33] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing, pp. 27-35. https://www.aclweb.org/anthology/W09-1904/

[34] Yi-Li Fang, Hai-Long Sun, Peng-Peng Chen, and Ting Deng. 2017. Improving the Quality of Crowdsourced Image Labeling via Label Similarity. Journal of Computer Science and Technology 32, no. 5: 877-889. DOI: http://dx.doi.org/10.1007/s11390-017-1770-7

[35] Wei Wang, and Zhi-Hua Zhou. 2015. Crowdsourcing label quality: a theoretical analysis. *Science China Information Sciences* 58, no. 11: 1-12. DOI: https://doi.org/10.1007/s11432-015-5391-x

[36] Ofer Dekel, and Ohad Shamir. 2009. Vox Populi: Collecting High-Quality Labels from a Crowd. In COLT. https://www.cs.mcgill.ca/~colt2009/papers/037.pdf

[37] Raymond G. Hunt, Frank J. Krzystofiak, James R. Meindl, and Abdalla M. Yousry. 1989. Cognitive style and decision making. Organizational behavior and human decision processes 44, no. 3: 436-453. DOI: https://doi.org/10.1016/0749-5978(89)90018-6

[38] Matteo Cristofaro. 2016. Cognitive styles in dynamic decision making: a laboratory experiment. International Journal of Management and Decision Making 15, no. 1: 53-82. DOI: https://doi.org/10.1504/IJMDM.2016.076840

[39] Jill R. Hough, and D. T. Ogilvie. 2005. An empirical test of cognitive style and strategic decision outcomes. Journal of Management Studies 42, no. 2: 417-448. DOI: https://doi.org/10.1111/j.1467-6486.2005.00502.x

[40] Chia-Ming Chang, Chia-Hsien Lee, and Takeo Igarashi. 2021. Spatial Labeling: Leveraging Spatial Layout for Improving Label Quality in Non-Expert Image Annotation. In CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3411764.3445165

[41] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6, 1, 3–5. DOI: http://dx.doi.org/10.1177/1745691610393980

[42] Gur Mosheiov. 2001. Parallel machine scheduling with a learning effect. Journal of the Operational Research Society 52, no. 10: 1165-1169. DOI: https://doi.org/10.1057/palgrave.jors.2601215

[43] Koun-tem Sun, Yuan-cheng Lin, and Chia-jui Yu. 2008. A study on learning effect among different learning styles in a Web-based lab of science for elementary school students. Computers & Education 50, no. 4: 1411-1422. DOI: https://doi.org/10.1016/j.compedu.2007.01.003

[44] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H. Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 625-634. DOI: https://doi.org/10.1145/1518701.1518797