Spatial Labeling: Leveraging Spatial Layout for Improving Label Quality in Non-Expert Image Annotation

Chia-Ming Chang
The University of Tokyo
info@chiaminghcnag.com

Chia-Hsien Lee
LeadBest Consulting Group
neil.lee@getoken.io

Takeo Igarashi
The University of Tokyo
takeo@acm.org

ABSTRACT

Non-expert annotators (who lack sufficient domain knowledge) are often recruited for manual image labeling tasks owing to the lack of expert annotators. In such a case, label quality may be relatively low. We propose leveraging the spatial layout for improving label quality in non-expert image annotation. In the proposed system, an annotator first spatially lays out the incoming images and labels them on an open space, placing related items together. This serves as a working space (spatial organization) for tentative labeling. During the process, the annotator observes and organizes the similarities and differences between the items. Finally, the annotator provides definitive labels to the images based on the results of the spatial layout. We ran a user study comparing the proposed method and a traditional non-spatial layout in an image labeling task. The results demonstrated that annotators can complete the labeling tasks more accurately using the spatial layout interface than the non-spatial layout interface.

CCS CONCEPTS

• **Human-centered computing – Interaction design;**; • **Interaction design process and methods;**; • **User interface design**;

KEYWORDS

Manual Image Labeling, Non-expert Annotator, Spatial Layout, Interface Design

ACM Reference Format:
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 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445165

1 INTRODUCTION

Image labeling is essential for building machine learning applications for images. Ideally, experts, who already have sufficient domain knowledge, should be recruited for labeling tasks as annotators. However, it is often difficult to recruit sufficient number of experts owing to the limited availability and cost [27, 28, 31]. In such cases, it becomes necessary to rely on non-expert annotators (typically, crowd workers) who lack sufficient domain knowledge for manual image labeling [7, 33]. Image labeling by non-experts can be significantly difficult and contain numerous errors [21, 23, 29]. Therefore, it is important to provide support to make the process more efficient and reduce errors. Suppose that an annotator without sufficient knowledge of dog breeds is asked to assign dog breed labels to the incoming dog images, and they assign labels to each incoming dog image by referring to the example images associated with the labels. However, the example images can be insufficient to capture the subtle differences between similar breeds. In such cases, the annotator can observe and organize the subtle differences by comparing the example images and incoming images. We believe that this observing and organizing process plays a critical role in non-expert labeling; however, it is not well supported in traditional image labeling tools. In this study, we propose to leverage the spatial layout for annotators to observe and organize the similarities and differences between images and labels before selecting a label for an image. This spatial organization process serves as tentative labeling.

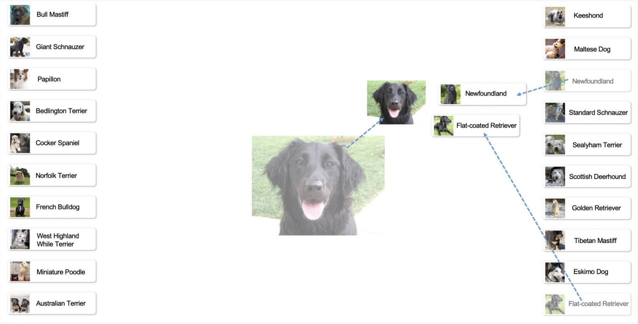


Figure 1: Screenshot of our spatial labeling prototype implementation.

Figure 1 presents a screenshot of the implementation of our prototype. An open space is provided for the annotator to spatially lay out the images and labels representing their conceptual similarity. The annotator drags the target image onto the open space and places the possible labels nearby (i.e., the image might belong to one of the labels). We expect this process to help the annotators to select a label for an image more correctly in the given domain.

We ran a user study to compare the proposed spatial layout labeling and a traditional non-spatial layout labeling in an image labeling task. The results demonstrated that the participants completed the given labeling task more accurately via the spatial layout interface (error rate: 37.63%) than the non-spatial layout interface

CHI '21, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445165

(error rate: 43.50%). Furthermore, the participants felt more confident with their selected labels in the spatial layout labeling task (confidence rate: 59.63%) than the non-spatial layout labeling task (confidence rate: 41.13%). These findings indicate that the spatial layout interface can improve the label quality in non-expert image annotation by laying out images and labels on an open space during the labeling process. In addition, the participants felt that the spatial layout interface was more helpful for manual image labeling than the non-spatial layout interface. In contrast, the participants felt that the non-spatial layout interface was more efficient than the spatial layout interface. Furthermore, most of the participants preferred the non-spatial layout (68.75%) interface than the spatial layout interface (31.25%) when performing a labeling task because, according to their experience, the non-spatial layout interface was simpler, easier, and more intuitive to use. This result indicates that we should not rely on user's subjective evaluation if we seek quality of the output in labeling. Easy-to-use is not necessarily a good criterion to estimate how well the user performs. The three main contributions of this study are as follows:

- Identifying spatial organization as important factor in non-expert annotation.
- A novel labeling interface design, spatial labeling, for tentative labeling in non-expert image annotation.
- A user study comparing a spatial layout interface to a traditional non-spatial layout interface, demonstrating the benefits of spatial layout used in non-expert image annotation.

2 RELATED WORK

2.1 Manual Image Annotation

Manual image annotation is a labor-intensive process that is significantly tedious and time-consuming. For example, ImageNet [7] is an image dataset containing more than 14 million images that were annotated manually by humans. Many studies have proposed different tools or workflows for assisting manual image annotation in a more efficient way. VIA [10] is a manual annotation tool that allows annotators to define and describe spatial regions in an image. It supports annotators to label images independently or collaboratively. LabelMe [8, 9] is a web-based annotation tool containing a "sharing" feature that allows annotators to create and share their annotation results with others instantly. Von Ahn et al. [11] introduced "ESP," which is an image annotation tool combined with a computer game. It allows annotators to provide meaningful labels for images while playing an online game. Chang et al. [1] introduced hierarchical task assignment for manual image annotation. This approach decomposes a labeling task into multiple steps and distributes it to multiple annotators to reduce their workload. Semi-automatic annotation systems assist manual image labeling using collaborative filtering and computer vision techniques [24–26]. Interactive concept learning guides users to assign labels to images that are most informative for classifiers [43–45].

The main objective of these annotation tools is to provide a more supportive, efficient, and enjoyable system to improve labor-intensive processes in manual image annotation. However, these tools generally assume that the annotators have sufficient domain knowledge; moreover, support for non-expert annotators who lack sufficient domain knowledge is still limited and must be improved. Kulesza et al. [22] introduced the notion of concept evolution in data labeling for web page classification and presented two structural labeling solutions to help annotators in defining and refining their concepts during data labeling. We share the same motivation to support annotators' concept organization in data labeling; however, they mainly focus on the changes in concepts over time in the context of web page classification, while we focus on observing items and organizing concepts in a single session in the context of image classification.

2.2 Improving Quality of Non-expert Annotation

Crowdsourcing is a popular approach used to perform labeling tasks with crowd workers, such as using Amazon's Mechanical Turk [20, 36]. Typically, crowd workers are non-expert annotators. Image labeling by non-experts is often difficult and contains many errors [21, 23, 29, 30]. Many studies have proposed solutions for improving the quality of non-expert annotation by leveraging the collaborative aspect of crowdsourcing. Revolt [19] is a collaborative crowdsourcing labeling approach for non-experts, which is based on expert annotation workflows (label-check-modify). It enables groups of workers to work together via three steps: vote (annotators select an appropriate label for an image), explain (annotators provide justifications if their selected label is different from others), and categorize (annotators review explanations from others and tag conflicting items with terms describing the newly discovered concepts). Fang et al. [32] introduced a two-round crowdsourcing framework to improve the quality of crowdsourced image labeling. In the first round, crowd workers selected a label for the target images (several labels might be assigned to each image). In the second round, crowd workers were required to select the best label for each image (referring to the results from the first round and making the best decision). Pairwise HITS [34] is a crowdsourcing workflow for quality estimation that enables evaluators to compare a pair of labeled data and select a better one. Otani et al. [35] introduced a label aggregation method for hierarchical classification tasks that can classify crowd workers in a hierarchical structure based on their response (labeled data). In addition, Liu et al. [37] proposed an interactive method that used data visualization to assist experts in verifying uncertain instance labels and unreliable workers to improve crowdsourcing annotation.

Most of these solutions aim to improve the label quality of non-expert annotation, which is based on the concept of "improving by others." This implies that these solutions support the concept of obtaining assistance from other annotators (experts or non-experts) for improving the label quality instead of relying on non-expert annotators for self-improvement. For example, worker A assigns a label to an image and worker B or C checks the labeled image and modifies it. This is a typical workflow for improving the label quality through a group of annotators working together. In this study, we aim to propose a complementary solution to improve the label quality in non-expert annotation, which is based on the concept of "improving by oneself." This means that our proposed solution will enable a non-expert annotator to work individually in a labeling task to improve the label quality by themselves without requiring any support from other annotators. In addition, these existing solutions mainly focuses on an easy problem. For example, accuracy of labeling is high (80%~90%) in an easy labeling task [19]. However, our proposed solution aims to address a new problem (difficult labeling task) where accuracy of labeling is relatively low (e.g. 56.5% in the NS labeling task and 62.37% in the S labeling task).

2.3 Spatial Layout

The concept of spatial layouts has been widely used for loosely managing information or documents. In real world, Malone [16] observed how people organize their desks using the spatial layout concept. In the digital world, a popular example is the window and icon interface on computer systems (MAC OS and Microsoft Windows), where users can manage their digital data (or files) by dragging icons on the interface. Shipman et al. [12] found that people use visual and spatial graphical layouts to express relationships between icons and visual symbols. Spatial layouts have been used in several studies, such as zoomable interfaces for information navigation [2–4]. The zoomable interface is an alternative to traditional windows and icon-based interfaces that allows people to have a zoom interaction based on a structure during information management. Mander et al. [6] observed the behavior of people in organizing information by creating piles of documents in the real world, and created a desktop interface element "pile" to support information organization. The metaphor of "pile" is also used as a tool for managing digital photo collections [13]. Bauer et al. [5] introduced a spatial tool for managing personal information collection based on the concept of zoomable interface as well as the "pile" metaphor. The spatial layout concept can also be seen in studies related to data aggregation, such as spatial aggregation. Watanabe et al. [14] introduced an interface named "Bubble Clusters," which is a technique for manipulating the spatial aggregation of graphical objects through grouping and ungrouping objects.

Many studies have indicated that the concept of spatial layout is an effective method for information management. Spatial memory (i.e., the ability to remember something in an environment) is one of the reasons that people can manage information or documents more effectively via a spatial layout interface. For example, Robertson et al. [18] introduced a "data mountain" for document management, which is based on the concept of spatial memory. In addition to information or data management, a spatial layout is used by designers to externalize thoughts and ideas [17]. We believe that the spatial layout concept can also be used in labeling interface design for annotators to observe and organize the similarities and differences between images and labels before selecting a label for an image. Spatial layout has also been used in search systems. Visualization methods map high-dimensional data items to a 2D canvas, thereby enabling users to search a data item on the canvas [46, 47]. Spatial search systems leverage user interaction on the canvas to facilitate the search process [48–50]. These systems use a spatial layout as a communication medium between the user and computer in search processes. In contrast, we use the spatial layout interface as a workspace for users to organize their thoughts during manual image annotation. The spatial layout is not created or interpreted by the system.

3 SPATIAL LABELING INTERFACE DESIGN

3.1 Overview

Our labeling interface was designed based on the concept of spatial layout, which provides an open space for annotators to spatially lay out the images and labels representing their conceptual similarity during the labeling process. Figure 2 shows the initial state of the interface. Labels (each label contains a sample image with textual name) are listed on the two sides of the interface, and a target image to be labeled is presented one by one in the middle.

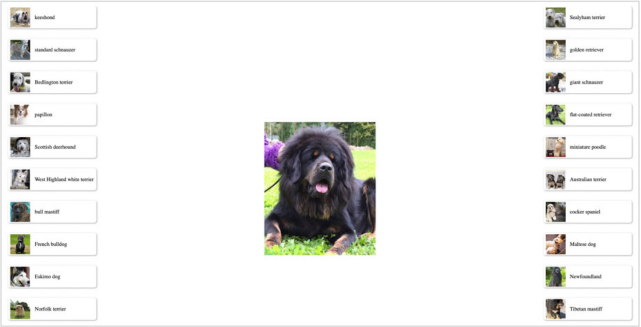


Figure 2: Initial state of the spatial labeling interface.

Figure 3 shows the working state of the interface. The annotators spatially lay out the images and labels on the open space (e.g. placing related items together) before selecting a label for an image.

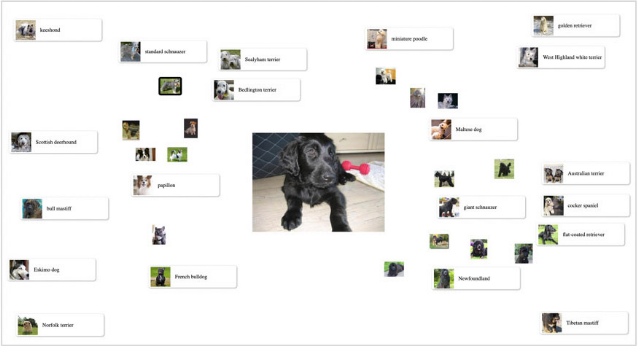


Figure 3: Working state of the spatial labeling interface.

3.2 User Interaction

The spatial layout interface contains four main functions for annotators to complete a labeling task: (a) spatially laying out images and labels, (b) assigning labels to images, (c) indicating confidence states, and (d) modifying the assigned labels (see Figure 4).

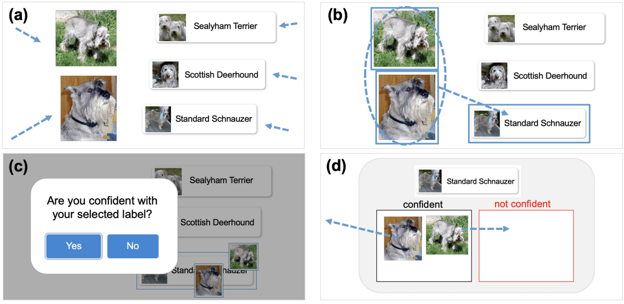


Figure 4: User interaction for the spatial layout labeling interface.

(a) Spatially laying out images and labels
In the open space, the annotator lays out the images and labels to represent their conceptual similarity. More similar images are grouped together and the possible labels are brought nearby. It serves as tentative labeling. During the process, the annotator builds domain knowledge by observing the similarities and differences between the items.

(b) Assigning labels to images
After laying out the images and labels on the open space, the annotator assigns definitive labels to the images based on the results of spatial layout. The annotator drags and drops one or more images to a label to assign the label to the images.

(c) Indicating confidence states
Immediately after assigning a label to the images (drag-and-drop operation), a pop-up window appears asking the annotator to indicate the confidence state of their label selection by answering the question: “are you confident with your selected label?”

(d) Modifying the assigned labels
The annotator can modify the labels that have been already assigned by clicking on a label to view all annotated images under this label and drag-and-drop the images to another label or open space. The annotator can also change the confidence state of the images.

3.3 Example of Usage Scenario

Figure 5 shows an example of the usage scenario. (a) The annotator first sees image A; however, it is not clear whether label X or Y is suitable for image A. Thus, the annotator brings images A and labels X and Y nearby. (b) The annotator then works on other images and places the images, which are similar to image A, labeled X and Y, nearby, thereby forming a cluster. (c) After aggregating similar images and examining them carefully, the annotator understands the key difference between labels X and Y (e.g., dog ears) and splits the cluster into two. Finally, the annotator drags and drops one of the clusters to label X and the other to label Y.

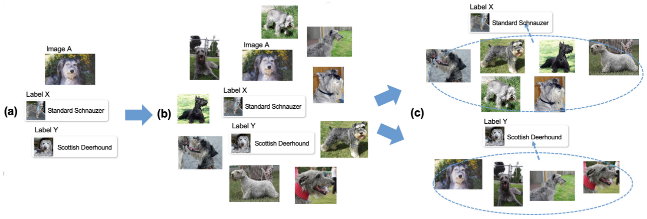


Figure 5: Example of usage scenario.

4 USER STUDY

We conducted a user study to compare our spatial layout interface with a traditional non-spatial layout interface in an image labeling task. Our hypothesis is that the spatial layout interface can improve the label quality of non-expert image annotation by providing an open space for annotators to observe and organize the similarities and differences between images and labels (i.e. a spatial organization process) during the labeling process.

4.1 Apparatus

We outsourced the execution of the user study to an outsourcing 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 $1,440 ($90 per participant). During user evaluation, the participants were asked to sit in front of a desktop computer and complete the given tasks; the labeling systems were run on Google Chrome with a screen resolution of 1440 × 900 (see Figure 6).

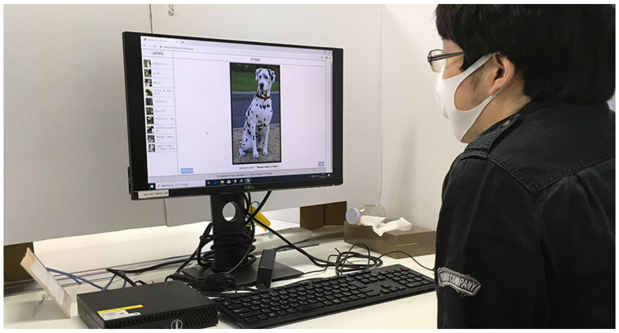


Figure 6: Photograph of user study.

4.2 Participants

Sixteen participants (eight men and eight women, in the age range of 18 to 49 years) were invited by the company to participate in the user evaluation process. None of the participants had prior experience with image labeling or specialized knowledge about image classification. Additionally, none of the participants had dogs as a pet, and none of them had specialized knowledge about dog categories. User evaluation was conducted from 27th July to 15th August 2020.

4.3 Dataset: Labels and Images

For the labeling tasks, we used the image dataset from ImageNet (ILSVRC 2012) [15]. This image dataset contains 1.3 million images in 1,000 categories. To satisfy our research requirements, we only used dog images from the ImageNet dataset, which contains 120 dog labels (breeds). This is because assigning a dog breed label to a dog image requires domain knowledge. First, we randomly selected 100 images for each dog label (breed), with a total of 12,000 images, and used it as a basic dataset. Second, we created two datasets (Dataset A and Dataset B) with comparable difficulty.

Originally, the data (labels and images) were randomly selected from the 120 dog labels (breeds) present in ImageNet for each participant. However, the results from the pilot study showed that the labels and images assigned to each participant were significantly different (some were easy to recognize and some were considerably difficult), which might affect labeling results. To ensure that the evaluation conditions were similar for each participant, we manually created two disjoint datasets (Dataset A and Dataset B). Each dataset contained 20 labels (20 dog breeds) and 50 images (50 dog images belonging to the selected 20 dog breeds). There is no overlap between the labels and images of the two datasets. The 20 labels selected for the datasets comprised three difficulty levels:

- Level 1: dog breeds are not ambiguous and easy to recognize (contains 5 labels)
- Level 2: dog breeds are ambiguous and difficult to recognize (contains 10 labels)
- Level 3: dog breeds are very ambiguous and very difficult to recognize (contains 5 labels)

Figure 7 shows the 20 dog breeds (labels) selected for Dataset A and Dataset B with the three difficulty levels.

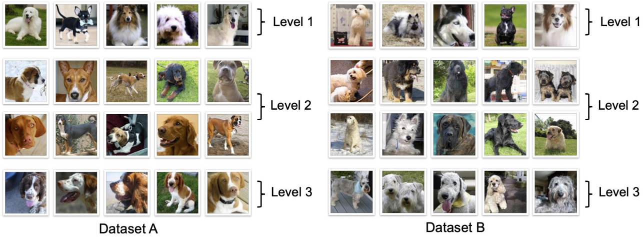
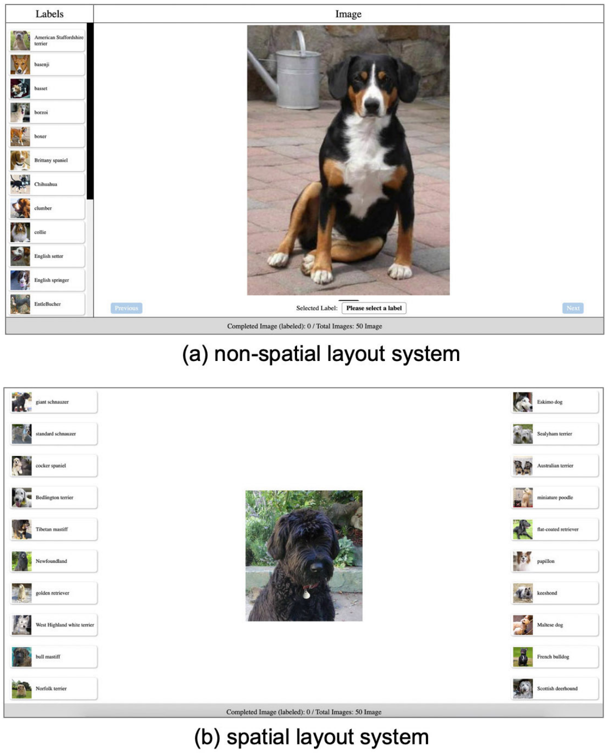


Figure 7: Twenty dog breeds (labels) used in Dataset A and Dataset B.

In addition, the distribution of the 50 images were randomly selected from the ImageNet that belongs to the 20 labels (a label contains 2∼3 images).

4.4 Task and Condition

Two online image labeling systems were developed in React.js for user evaluation: (a) non-spatial layout labeling system and (b) spatial layout labeling system.



Selected Label: Please select a label

Completed Image (labeled): 0 / Total Images: 50 Image

(a) non-spatial layout system

Completed Image (labeled): 0 / Total Images: 50 Image

(b) spatial layout system

Figure 8: Screenshots of non-spatial and spatial layout labeling systems.

Manual image labeling is significantly time consuming. To maintain the duration of evaluation reasonable, we designed a small-scale labeling task to evaluate the proposed spatial layout concept. The labeling task involved labeling 50 dog images by selecting an appropriate label from a list of 20 dog breeds (20 labels). The method of within-subjects was used, where each participant was asked to complete two labeling tasks (NS Task and S Task) using the two labeling systems.

- NS Task: Non-spatial Layout Interface and Condition
  This is the baseline condition. We mimicked the user interfaces of popular labeling systems [1]. Figure 8 (a) shows a screenshot of the non-spatial layout labeling system. The left side of the interface lists labels in both the text and sample images. The right side of the interface presents a target image to be labeled. The participant was asked to label 50 dog images by selecting an appropriate label from the list of 20 dog breeds (20 labels). This non-spatial layout interface only showed one image at one time. The participant scrolled through the list on the left to search for an appropriate label for the target image, and clicked on the label as a tentative assignment (the selected label appeared under the target image). The participant then clicked on “Next” to finalize the assignment. Subsequently, the participant was required to indicate their confidence about the selected label by answering the question: “are you confident with your selected

label?" The participant then clicked on "Next" to go to the next image. The participant was also allowed to return to previous images by clicking on "Previous" if they wanted to modify their previously selected labels.

- S Task: Spatial Layout Interface and Condition
  This is the proposed method. Figure 8 (b) shows a screenshot of the spatial layout labeling system (the spatial layout interface has been explained in Section 3). The labeling task (to label 50 dog images) included the same task that was assigned for the non-spatial layout interface. In the S task, we instructed the participants to first lay out all the images and labels representing their conceptual similarity before assigning definitive labels to the images (see Section 6.4 for more details).

Each participant worked on one non-spatial task and one spatial layout tasks in a balanced order. We prepared two image datasets and each participant worked on one of the two data sets in one of the two tasks and worked on the other in the other task. So, each participant did not see a label or image more than once in the study. The task-dataset assignment is also fully balanced.

4.5 Procedure

First, the participants were provided with an oral overview and detailed written instruction by an instructor. The evaluation itself comprised three parts (in order): instruction and trial (10–15 min), two labeling tasks (30–40 min), and questionnaire (3–5 min). The entire evaluation process was completed in approximately 1 h. After providing instructions on the labeling interfaces and given tasks, the participants were allowed to practice on a small labeling task (to label five dog images by selecting an appropriate label from a list of 10 dog breeds) via the non-spatial and spatial layout labeling interfaces. The participants were informed that there was no time restriction in the user evaluation process; they were also asked to not hurry while selecting the most appropriate label. The supplemental material shows the written instructions used in the study.

4.6 Measurement

Task Performance

Our labeling system automatically recorded and measured the time and error rate (failure in selecting an appropriate label for an image) of the labeling tasks completed by the participants. The timer started when the participants clicked on "START" and stopped when they clicked on "FINISH." The system also recorded the time spent by the participants for labeling.

Confident Label Selection

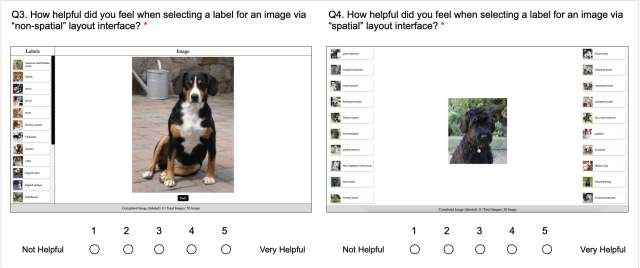
We also measured the rate of confident label selection (the labeled images that the participants were confident with).

Questionnaire

Following the evaluation of the labeling task, the participants were asked to answer a questionnaire regarding the two different labeling interfaces used in the user study. The questionnaire contained the following 10 questions.

Q1 How difficult did you feel when selecting a label for an image via the "non-spatial" layout interface?
Q2 How difficult did you feel when selecting a label for an image via the "spatial" layout interface?
Q3 How helpful did you feel when selecting a label for an image via the "non-spatial" layout interface?
Q4 How helpful did you feel when selecting a label for an image via the "spatial" layout interface?
Q5 How efficient did you feel when selecting a label for an image via the "non-spatial" layout interface?
Q6 How efficient did you feel when selecting a label for an image via the "spatial" layout interface?
Q7 What is the most difficult part when selecting a label for an image via the "non-spatial layout" interface?
Q8 What is the most difficult part when selecting a label for an image via the "spatial layout" interface?
Q9 Which labeling interface do you prefer?
Q10 Why?

We used Likert scaling (Q1–Q6) to determine the perception of participants regarding the labeling interfaces. Figure 9 displays a screenshot of the layout of the questionnaire.



Q3. How helpful did you feel when selecting a label for an image via "non-spatial" layout interface? *

Completed Image (labeled): 0 / Total Images: 50 Image

Not Helpful 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very Helpful

Q4. How helpful did you feel when selecting a label for an image via "spatial" layout interface? *

Completed Image (labeled): 0 / Total Images: 50 Image

Not Helpful 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Very Helpful

Figure 9: Screenshot of the questionnaire layout.

5 RESULTS

5.1 Task Completion Time

Figure 10 shows that the participants spent an average of 16 min 47 s and 17 min 55 s to label the 50 images using the non-spatial and spatial layout interfaces, respectively. The result of paired t-test on task completion time showed that there was no significant difference ($p > 0.05$) between the non-spatial and spatial layout interfaces.

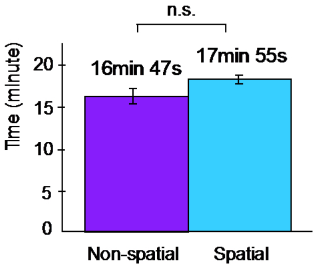


Figure 10: Labeling task completion time. N-s: mean = 16.47; SE = 1.39; S: mean = 17.55; SE = 1.63.

5.2 Error Rate

Figure 11 displays the error rates (fail to select an appropriate label for an image) for the labeling tasks completed by the participants using the non-spatial and spatial layout interfaces. The results showed that the error rate was 43.50% for the non-spatial layout interface and 37.63% in the spatial layout interface. The analysis of the error rates using paired t-test showed that there was a significant difference ($p < 0.05$) between the two labeling interfaces. This indicates that the participants could select labels more accurately via the spatial layout interface than the non-spatial layout interface.

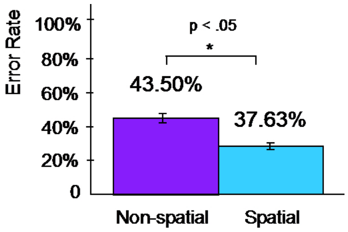


Figure 11: Error rates for the labeling tasks. N-s: mean = 43.50; SE = 2.94; S: mean = 37.63; SE = 2.41; p = 0.0416.

We analyzed the error rates in the three difficulty levels of the image datasets (see Figure 6). The results showed that there were no significant differences ($p > 0.05$) between the non-spatial and spatial layout interfaces in the level 1 and level 3, while there was a significant difference ($p < 0.05$) in the level 2 (Figure 12). This indicates that benefit of the spatial layout interface is only appeared when a labeling task contains ambiguous images. The benefit does not come when the target images are not ambiguous or too ambiguous.

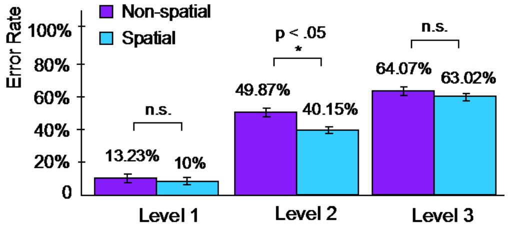


Figure 12: Error rates in difficulty levels. Level 1. N-s: mean = 13.23; SE = 2.88; S: mean = 10; SE = 2.67. Level 2. N-s: mean = 49.87; SE = 4.19; S: mean = 40.15; SE = 3.54; p = 0.0427. Level 3. N-s: mean = 64.07; SE = 5.58; S: mean = 63.02; SE = 3.48.

5.3 Relationship between Task Completion Time and Error Rate

Figures 13 and 14 show the relationship between the task completion time and error rate in the labeling task using the non-spatial and spatial layout interfaces, respectively. The results of Pearson correlation coefficient showed there was a negative correlation between the task completion time and error rate. The correlation was very weak (r = -0.288) in the non-spatial layout interface, while the correlation was moderate (r = -0.409) in the spatial layout interfaces. This indicates that when a non-expert annotator spends a shorter time for selecting a label for an image, the label quality is slightly less accurate, especially in the spatial layout interface. In addition, the correlation in the spatial layout interface was more obvious than the correlation in the non-spatial layout interface. This means that effects of the spatial layout interface in the relationship between the task completion time and error rate is stronger than the non-spatial layout interface.

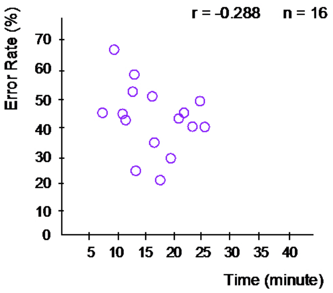


Figure 13: Error rate and time in the non-spatial layout task.

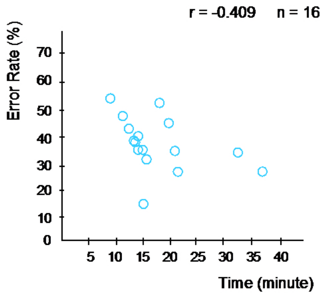


Figure 14: Error rate and time in the spatial layout task.

5.4 Confidence Rate

In the user evaluation process, the participants were asked to indicate the confidence states (confident or not confident) for their selected labels. Figure 15 shows that the participants were confident with 47.13% (24 of 50 images) of the labeled images in the non-spatial layout task, while 59.63% (30 of 50 images) in the spatial layout task. The analysis of the results with paired t-test shows that there is a significant difference ($p < 0.01$) between the non-spatial and spatial layout interfaces. This indicates that the participants felt more confident with their selected labels using the spatial layout interface than the non-spatial layout interface during the labeling process.

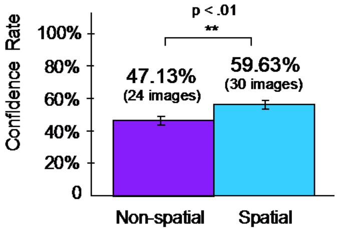


Figure 15: Confidence rate for the labeling tasks. Non-spatial: mean = 47.13; SE = 4.52; Spatial: mean = 59.63; SE = 4.12; p = 0.0008.

We analyzed the error rate in different confidence states. In the non-spatial layout task, the error rate was 19.16% for the confident cases and 63.65% for the unconfident cases (Figure 16). Similarly, in the non-spatial layout task, the error rate was 22.16% for the confident cases and 58.73% for the unconfident cases (Figure 17). The unpaired t-test showed a significant difference in both cases. This indicates that when the participants were confident with their labeled images, the label quality was significantly higher than when they were unconfident.

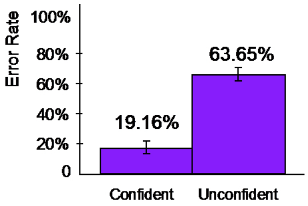


Figure 16: Error rate in the non-spatial layout task.

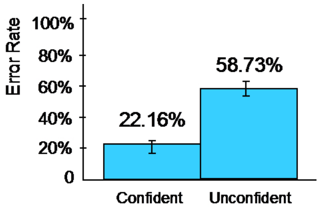


Figure 17: Error rate in the spatial layout task.

5.5 Questionnaire

Figure 18 illustrates the difficulty levels indicated by the participants regarding the labeling interfaces. The results showed that half of the participants felt that both interfaces are difficult or very difficult, while the other half felt that the interfaces were easy and very easy to use for labeling images. Interestingly, the results showed that more participants (n = 6) felt that the spatial layout interface is very difficult to use than the non-spatial layout interface (n = 3). Additionally, more participants (n = 3) felt that the spatial layout interface is easier to use than the non-spatial layout interface (n = 1). This indicates that the two interfaces have both advantages and disadvantages that contribute to the different experiences in manual image labeling.

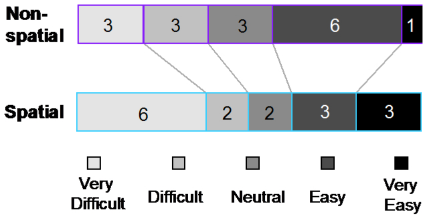


Figure 18: How difficult are the labeling interfaces?

Figure 19 shows how helpful the participants found the labeling interfaces. The results showed that more participants (n = 10) felt that the spatial layout interface was helpful or very helpful, while fewer participants (n = 3) felt that it is not helpful or not very helpful. This indicates that the spatial layout interface was perceived to be more supportive among non-expert image annotators than the non-spatial layout interface.

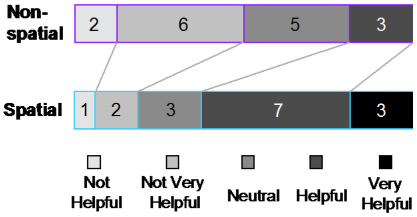


Figure 19: How helpful are the labeling interfaces?

Figure 20 displays the level of efficiency experienced by the participants regarding the labeling interfaces during the labeling process. The results showed that more participants (n = 9) felt that the non-spatial layout interface was efficient or very efficient than the spatial layout interface (n = 7), while more participants (n = 9) felt that the spatial layout was not efficient or not very efficient than the non-spatial layout interface (n = 4). This indicates that, in general, the non-spatial layout interface was perceived to be more efficient for non-expert image annotation than the spatial layout interface.

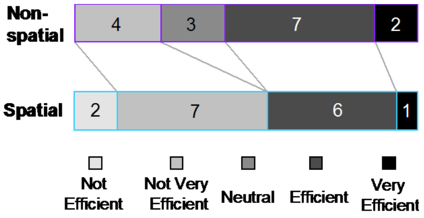


Figure 20: How efficient are the labeling interfaces?

We obtained the following answers to the question, "what is the most difficult part?" for each task. For non-spatial layout labeling, the participants answered that "they could not view all the labels and images at the same time during the labeling process" and "the sample image in each label was too small to see the detailed features of the dog." For spatial layout labeling, the participants answered that "they needed time to familiarize with the spatial layout interface and the interaction process," "it takes a long time to lay out the images and labels on the open space," and "the sample image in each label is too small to appropriately view the details."

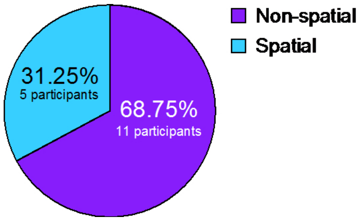


Figure 21: Preference of the labeling interfaces.

Additionally, 68.75% of the participants (n = 11) preferred the non-spatial layout interface, while 31.25% preferred the spatial layout interface (n = 5) (Figure 21). The participants who preferred the non-spatial layout interface indicated that it was simple, easy, and intuitive to use. For example, one participant indicated that "It is easy because I only need to single-click for selecting a label for an image"; one participant indicated that "The function is simple and easy in the non-spatial layout task; thus, I only need to search and select"; and another participant indicated that "In the non-spatial layout, I do not need to do many things, such as laying out the images and labels. I feel the tasks in the spatial layout are complex." In contrast, the participants who preferred the spatial layout interface indicated that it was good to see and compare all the labels and images in the spatial layout interface. For example, one participant indicated that "I can see all the labels and compare the target images"; one participant indicated that "It is easy to use because I can place images and labels anywhere I want"; and another participant indicated that "After checking all images, it becomes clear and easy to assign a label."

6 DISCUSSION

6.1 Spatial Layout Interface can Reduce Errors in Non-expert Image Annotation

The results showed that non-expert annotators can complete the given labeling tasks more accurately using the spatial layout interface than the traditional non-spatial layout interface (error rate = 37.63% and 43.50% in the spatial and non-spatial layout interfaces, respectively) without a significant increase in the task completion time (17 min 55 s for the spatial layout and 16 min 47 s for the non-spatial layout). This result implies that the spatial layout interface improves label quality in non-expert image annotation by providing an open space for annotators to lay out the images and labels (a spatial organization process) during the labeling process. More specifically, the spatial layout interface improves non-expert image annotation when a labeling task contains ambiguous images. However, it is not explicit whether the error reduction is a result of self-learning (better understanding of the problem) or the participants were simply more careful. Therefore, further investigation is required to distinguish these two possibilities. Nonetheless, the results demonstrated that our approach can reduce errors without significant additional cost, which, we believe, is a valuable insight for image annotation by non-experts in general.

6.2 Spatial Layout Interface can Increase Confidence in Non-experts for Manual Image Annotation

Confidence is a key factor that can help people to improve their learning motivation and efficiency [38–40]. We believe that confidence may be also a factor that affects non-expert image annotation. The questionnaire results showed that non-expert annotators felt more confident with their selected labels using the spatial layout interface (confidence rate = 59.63%) than the non-spatial layout interface (confidence rate = 47.13%). This indicates that the spatial layout interface can increase the subjective feeling of confidence in non-experts for manual image annotation. The evaluation results showed that the error rates for confident label selection (19.16% and 22.16% in the non-spatial and spatial layout interfaces, respectively) were significantly lower than those for unconfident label selection (63.65% and 58.57% in the non-spatial and spatial layout interfaces, respectively). This indicates that the label quality is significantly higher when the annotators are confident with their selected labels. In addition, we believe that if the labeled images contain "confident label selection" and "unconfident label selection," it would be benefit the labeling task. For example, it can be used to reduce the workload of expert annotators if they are still required in the labeling task. The expert annotators would only need to check and modify the labeled images obtained under "unconfident label selection" to improve the label quality. This is a significant aspect of indicating the confidence state in a non-expert image annotation task.

6.3 Mismatch between Perceived Usability and Annotation Quality

The results from the user evaluation process demonstrated that the spatial layout interface can help non-expert annotators to complete a labeling task more accurately. Additionally, the results of the questionnaire showed that the spatial layout interface is more helpful than the non-spatial layout interface. However, some parts of the questionnaire results showed that the participants felt that the spatial layout interface is difficult to use and it is not as efficient as the non-spatial layout interface (although the result did not show significant difference in the task completion time). The results also showed that most of the participants (non-expert annotators) preferred a traditional non-spatial layout interface (66.75%) than the spatial layout interface (31.25%). This shows that there is a discrepancy between the perceived usability and annotation quality in the two labeling interfaces. This issue has been discussed as a tension between "usability" and "functionality" in interactive systems [41, 42]. Usability mainly focuses on the interaction between a user and a system, while functionality is more focused on the functions provided by a system. According to the results, the usability of the non-spatial layout interface seems to be higher than that of the spatial layout interface. This is understandable because the non-spatial layout interface is simple, while the spatial layout interface requires additional operations. In contrast, the functionality of the spatial layout interface seems to be higher than that of the non-spatial layout interface. The quality of annotation is the most important feature in image annotation; therefore, we believe that spatial layout interface has practical value despite the usability issue. However, it should be noted that the benefit does not come without cost; reduced usability can cause discomfort and fatigue to annotators.

6.4 Importance of "Forced" Spatial Layout

Originally, we allowed the participants in the spatial layout interface to assign a label to an image immediately (without laying out the images and labels on the open space) if they were confident with the target image and label. The annotators moved the images and labels using the open space only when they wanted. However, the results from the pilot study showed that the error rate in the spatial layout interface was higher than that in the non-spatial layout interface, and the task completion time in the spatial layout interface was significantly less than that in the non-spatial layout interface. We found that they assigned a label to an image immediately without using the open space. We assumed that this might be because the participants were overconfident with their selected labels; therefore, they did not carefully lay out the images and labels on the open space before assigning a label for an image.

To exploit the benefits of the spatial layout interface and help the participants in using it efficiently, we added one more condition to the user evaluation process. The participants were "forced" to lay out all the images on the open space first (tentative) for spatial organization before assigning a label for an image (the system did not allow the participants to assign a label for an image immediately). Accordingly, the results changed, i.e., the error rate decreased. We found that the benefits of the spatial layout interface were not fully exploited by the users without the forced use of the aforementioned feature. This shows that just providing a system is not sufficient; it is crucial to specify its usage to maximally utilize the potential of the system. This is an important lesson to introduce this type of system in practice.

7 LIMITATIONS AND FUTURE WORK

A limitation of this study is the real reason of the label improvement in non-expert image annotation is not clearly demonstrated. It is unclear whether the reason of the improvement is self-learning or label carefully via the spatial layout interface during annotation. Another limitation is the size of the labeling task conducted in the user study. To ensure that the evaluation time does not exceed 1 h (participants may feel tired if it takes longer than 1 hour), we decided to use a small-scale image dataset (20 labels with 50 images) to evaluate the proposed interface. Although the results provided significant insights regarding the proposed spatial layout interface for non-expert image annotation, we believe that the spatial layout interface might perform even better in a large-scale labeling task. This is because non-expert annotators can organize conceptual similarity of images and labels during the labeling process by observing more items on the spatial layout.

Additionally, expert annotators might still be required to further improve the label quality because the label quality in the spatial layout tasks was not sufficient (the error rate of 37.63% was still not low). Our current target task includes annotation for image classification (assigning a label to an image). However, in practice, annotation for object detection in an image (assigning a label to a part of an image) is more in need. Annotation for detection is more complicated (the user not only selects a label but also selects a region); therefore, we cannot directly apply our current method for image detection. However, spatial organization would be also important in annotation for image detection, and we plan to explore methods to support this process as a next step.

In the future, we would like to investigate the real reason of the label improvement (learning or careful labeling). In addition, we would like to explore more features on the spatial layout interface to improve the label quality. One possible direction involves the use of external sources (e.g., Google) as references during annotation. Another direction could be designing a collaborative labeling interface for non-expert image annotation. We also believe that the concept of the proposed spatial layout labeling interface can be applied for developing various data annotation tools other than images.

8 CONCLUSION

In this study, we proposed a spatial layout labeling interface, called Spatial Labeling, for improving label quality in non-expert image annotation. This interface comprised an open space for annotators to lay out the images and labels (i.e. a spatial organization process) during the labeling process before selecting a label. We conducted a user study to compare the proposed spatial layout interface with a traditional non-spatial layout interface for an image labeling task. The results showed that non-expert annotators more accurately selected a label for an image using the proposed spatial layout interface than the traditional non-spatial layout interface. Moreover, it was observed that the spatial layout interface increased the confidence level of non-expert annotators during manual image labeling. Spatial labeling provides an alternative solution for improving the label quality in non-expert image annotation via spatial organization. The findings of this study have presented significant insights that could be used in the future development of annotation tools.

ACKNOWLEDGMENTS

This work was supported by JST CREST Grant Number JP-MJCR17A1, Japan.

REFERENCES

[1] 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

[2] Ken Perlin, and David Fox. 1993. Pad: An Alternative Approach to the Computer Interface. In Proceedings of the 20th annual conference on Computer graphics and interactive techniques, pp. 57-64. DOI: http://dx.doi.org/10.1145/166117.166125

[3] Benjamin B. Bederson, James D. Hollan, Ken Perlin, Jonathan Meyer, David Bacon, and George Furnas. 1996. Pad++: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics. Journal of Visual Languages & Computing 7, no. 1: 3-32. DOI: http://dx.doi.org/10.1006/jvlc.1996.0002

[4] Benjamin B. Bederson, and James D. Hollan. 1994. Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics. In Proceedings of the 7th annual ACM symposium on User interface software and technology, pp. 17-26. DOI: http://dx.doi.org/10.1145/192426.192435

[5] Daniel Bauer, Pierre Fastrez, and Jim Hollan. 2005. Spatial Tools for Managing Personal Information Collections. In Proceedings of the 33th IEEE Annual Hawaii

International Conference on System Sciences, pp. 104b-104b. DOI: http://dx.doi.org/10.1109/HICSS.2005.551

[6] Richard Mander, Gitta Salomon, and Yin Yin Wong. 1992. A "pile" Metaphor for Supporting Casual Organization of Information. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 627-634. DOI: http://dx.doi.org/10.1145/142750.143055

[7] 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

[8] 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

[9] Antonio Torralba, Bryan C. Russell, and Jenny Yuen. 2010. LabelMe: Online Image Annotation and Applications. Proceedings of the IEEE 98, no. 8: 1467-1484. DOI: http://dx.doi.org/10.1109/JPROC.2010.2050290

[10] 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

[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] Frank M. Shipman III, Catherine C. Marshall, and Thomas P. Moran. 1995. Finding and Using Implicit Structure in Human-Organized Spatial Layouts of Information. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 346-353. DOI: http://dx.doi.org/10.1145/223904.223949

[13] Daniel Bauer, Pierre Fastrez, and Jim Hollan. 2004. Computationally-Enriched 'piles' for Managing Digital Photo Collections. In 2004 IEEE Symposium on Visual Languages-Human Centric Computing, pp. 193-195. DOI: http://dx.doi.org/10.1109/VLHCC.2004.13

[14] Nayuko Watanabe, Motoi Washida, and Takeo Igarashi. 2007. Bubble Clusters: An Interface for Manipulating Spatial Aggregation of Graphical Objects. In Proceedings of the 20th annual ACM symposium on User interface software and technology, pp. 173-182. DOI: http://dx.doi.org/10.1145/1294211.1294241

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang *et al.* 2015. ImageNet Large Scale Visual Recognition Challenge. International journal of computer vision 115, no. 3: 211-252. DOI: http://dx.doi.org/10.1007/s11263-015-0816-y

[16] Thomas W Malone. 1983. How Do People Organize Their Desks? Implications for the design of office information systems. ACM Transactions on Information Systems (TOIS) 1, no. 1: 99-11. DOI: http://dx.doi.org/10.1145/357423.357430

[17] Kumiyo Nakakoji, Yasuhiro Yamamoto, Shingo Takada, and Brent N. Reeves. 2000. Two-Dimensional Spatial Positioning as a Means for Reflection in Design. In Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques, pp. 145-154. DOI: http://dx.doi.org/10.1145/347642.347697

[18] George Robertson, Mary Czerwinski, Kevin Larson, Daniel C. Robbins, David Thiel, and Maarten Van Dantzich. 1998. Data Mountain: Using Spatial Memory for Document Management. In Proceedings of the 11th annual ACM symposium on User interface software and technology, pp. 153-162. DOI: http://dx.doi.org/10.1145/288392.288596

[19] 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/10.1145/3025453.3026044

[20] 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

[21] 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

[22] 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

[23] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: Crowdsourcing Complex Work. In Proceedings of the 24th annual ACM symposium on User interface software and technology, pp. 43-52. DOI: http://dx.doi.org/10.1145/2047196.2047202

[24] Shingo Uchihashi, and Takeo Kanade. 2005.Content-Free Image Retrieval by Combinations of Keywords and User Feedbacks. In International Conference on Image and Video Retrieval, pp. 650-659. Springer, Berlin, Heidelberg. DOI: http://dx.doi.org/10.1007/11526346_68

[25] Wenyin Liu, Susan T. Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, and Brent A. Field. 2001. Semi-Automatic Image Annotation. In Interact, vol. 1, pp. 326-333. https://www.researchgate.net/profile/Liu_Wenyin/publication/2328523_Semi-Automatic_Image_Annotation/links/5650863608aeafc2aab71e41/Semi-Automatic-Image-Annotation.pdf

[26] Suh Bongwon and Benjamin B. Bederson. 2004. Semi-Automatic Image Annotation Using Event and Torso Identification. Human Computer Interaction Laboratory, University of Maryland, College Park, Maryland, USA. http://www.cs.umd.edu/hcil/trs/2004-15/2004-15.pdf

[27] 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

[28] Roland Kwitt, Sebastian Hegenbart, Nikhil Rasiwasia, Andreas Vécsei, and Andreas Uhl. 2014. Do We Need Annotation Experts? A Case 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

[29] 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

[30] Jiyi Li, Yukino Baba, and Hisashi Kashima. 2018. Incorporating Worker Similarity for Label Aggregation in Crowdsourcing. In International Conference on Artificial Neural Networks, pp. 596-606. Springer, Cham. DOI: http://dx.doi.org/10.1007/978-3-030-01421-6_57

[31] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An Overview. In Treebanks, pp. 5-22. Springer, Dordrecht. DOI: http://dx.doi.org/10.1007/978-94-010-0201-1_1

[32] 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

[33] Tao Han, Hailong Sun, Yangqiu Song, Yili Fang, and Xudong Liu. 2016. Incorporating External Knowledge into Crowd Intelligence for More Specific Knowledge Acquisition. In IJCAI, vol. 2016, pp. 1541-1547. https://dl.acm.org/doi/10.5555/3060832.3060836

[34] 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

[35] 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

[36] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In Proceedings of the ACM SIGKDD workshop on human computation. ACM, 64–67. DOI: http://dx.doi.org/10.1145/1837885.1837906

[37] Shixia Liu, Changjian Chen, Yafeng Lu, Fangxin Ouyang, and Bin Wang. 2018. An Interactive Method to Improve Crowdsourced Annotations. IEEE transactions on visualization and computer graphics 25, no. 1: 235-245. DOI: http://dx.doi.org/10.1109/TVCG.2018.2864843

[38] Mei-Mei Chang. 2005. Applying Self-Regulated Learning Strategies in a Web-Based Instruction—An Investigation of Motivation Perception. Computer Assisted Language Learning 18, no. 3: 217-230. DOI: http://dx.doi.org/10.1080/09588220500178939

[39] Robert M Klassen. 2010. Confidence to Manage Learning: The Self-Efficacy for Self-Regulated Learning of Early Adolescents with Learning Disabilities. Learning Disability Quarterly 33, no. 1: 19-30. DOI: http://dx.doi.org/10.1177/073194871003300102

[40] Juan Carlos Ortiz-Ordoñez, Friederike Stoller, and Bernd Remmele. 2015. Promoting Self-Confidence, Motivation and Sustainable Learning Skills in Basic Education. Procedia-Social and Behavioral Sciences 171: 982-986. DOI: http://dx.doi.org/10.1016/j.sbspro.2015.01.205

[41] Nancy C Goodwin. 1987. Functionality and Usability. Communications of the ACM 30, no. 3: 229-233. DOI: http://dx.doi.org/10.1145/214748.214758

[42] Fethi Calisir, A. Elvan Bayraktaroğlu, Cigdem Altin Gumussoy, Y. Ilker Topcu, and Tezcan Mutlu. 2010. The Relative Importance of Usability and Functionality Factors for Online Auction and Shopping Web Sites. Online Information Review. DOI: http://dx.doi.org/10.1108/146845210110370

[43] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In Proceedings of the sigchi conference on human factors in computing systems, pp. 29-38. DOI: http://dx.doi.org/10.1145/1357054.1357061

[44] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2009. Overview-Based Example Selection in End-User Interactive Concept Learning. In Proceedings of the 22nd annual ACM symposium on User interface software and technology, pp. 247-256. DOI: http://dx.doi.org/10.1145/1622176.1622222

[45] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Examining Multiple Potential Models in End-User Interactive Concept Learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1357-1360. DOI: http://dx.doi.org/10.1145/1753326.1753531

[46] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426. https://arxiv.org/abs/1802.03426

[47] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. journal of Machine Learning Research 9. (Nov), 2579-2605. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.457.7213

[48] Eli T. Brown, Jingjing Liu, Carla E. Brodley, and Remco Chang. 2012. Dis-Function: Learning Distance Functions Interactively. In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 83-92. IEEE. DOI: http://dx.doi.org/10.1109/VAST.2012.6400486

[49] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration. In 23rd International Conference on Intelligent User Interfaces, pp. 269-280. DOI: http://dx.doi.org/10.1145/3172944.3172950

[50] Yanir Kleiman, Joel Lanir, Dov Danon, Yasmin Felberbaum, and Daniel Cohen-Or. 2015. DynamicMaps: Similarity-based Browsing through a Massive Set of Images. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 995-1004. DOI: http://dx.doi.org/10.1145/2702123.2702224