User Preference and Performance using Tagging and Browsing for Image Labeling

Bruno Fruchard, Sylvain Malacria, Géry Casiez, Stéphane Huot

User Preference and Performance using Tagging and Browsing for Image Labeling

Bruno Fruchard
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL
Lille, France
bruno.fruchard@inria.fr

Sylvain Malacria
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL
Lille, France
sylvain.malacria@inria.fr

Géry Casiez*
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
Lille, France
gery.casiez@univ-lille.fr

Stéphane Huot
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL
Lille, France
stephane.huot@inria.fr

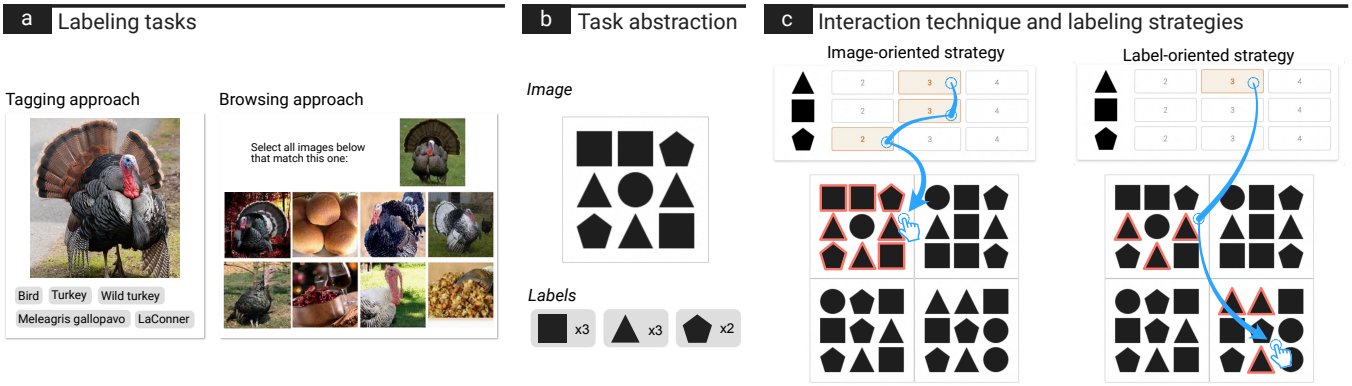

Figure 1: Image labeling tools rely on two primary approaches: a) *tagging* a single image with labels, or *browsing* all images to assign a single label. b) We characterize the labeling task and systematically study the efficiency of these approaches by measuring the performance of annotators when counting shapes in images (circles are distractors). c) Annotators can select possible shape counts (labels) using toggle buttons to tag corresponding images. Through this setting we study what strategy users adopt when they have the choice and evaluate their efficiency (outlined shapes represent the annotators' targets).

ABSTRACT

Visual content must be labeled to facilitate navigation and retrieval, or provide ground truth data for supervised machine learning approaches. The efficiency of labeling techniques is crucial to produce numerous qualitative labels, but existing techniques remain sparsely evaluated. We systematically evaluate the efficiency of tagging and browsing tasks in relation to the number of images displayed, interaction modes, and the image visual complexity. Tagging consists in focusing on a single image to assign multiple labels (image-oriented strategy), and browsing in focusing on a single label to assign to multiple images (label-oriented strategy). In a first experiment, we focus on the nudges inducing participants to adopt one of the strategies (n=18). In a second experiment, we evaluate the efficiency of the strategies (n=24). Results suggest an image-oriented strategy (tagging task) leads to shorter annotation times, especially for complex images, and participants tend to adopt it regardless of the conditions they face.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; **Empirical studies in HCI**.

KEYWORDS

Human-Computer Interaction, empirical studies, user performance, image labeling, tagging, browsing, visual complexity, open science

ACM Reference Format:
Bruno Fruchard, Sylvain Malacria, Géry Casiez, and Stéphane Huot. 2023. User Preference and Performance using Tagging and Browsing for Image Labeling. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3544548.3580926

1 INTRODUCTION

Image labeling consists in associating images with a list of labels describing its features or the concepts it represents. Such labels facilitate grouping relevant images together and navigating an image set to find specific content quickly. For instance, image libraries like

*Also with Institut Universitaire de France.

CHI '23, April 23–28, 2023, Hamburg, Germany

Flickr [23][1] or Instagram [45] enable users to tag the images they upload using free-form text or ontological keywords for organizational and personal purposes [52]. Labels can also be used as a way to provide ground truth data that is essential to train supervised machine learning algorithms using computer vision [57]. In this scenario, they usually accompany graphical elements that highlight specific areas or objects in images [22, 55] or video frames [16] to provide more descriptive power. Assistive systems can automatically assign labels to images to support annotators by reducing labeling times [70], but the task remains tedious as annotators must still verify and validate the generated labels. Manual labeling is so frequently required that previous work proposed crowdsourcing approaches to alleviate the annotator workload [13, 36, 55]. The task remains the same, and as tedious, for a single annotator uploading a picture with tags on social networks, or many annotators labeling images to build a data set.

Labeling an image is usually performed via two stereotypical tasks: *tagging* consists in assigning multiple labels to a single image, while *browsing* consists in focusing on a single label to assign to multiple images [57] (Figure 1a). Selecting pre-suggested tags when adding a photo on Flickr is an example of the former, whereas image-based CAPTCHAs are a common example of the latter [17]. While the task characteristics might greatly influence the annotator performance when labeling images, i.e., the time to produce the tags and their quality, their efficiency remains sparsely evaluated.

The literature on image labeling tools does not provide a rigorous definition of tagging and browsing tasks. The lack of a definition makes it challenging to identify their design implications, the interactions they should support, and their impact on user performance. For instance, if an annotator faces a set of images and tags each of them sequentially with all possible labels [49], are they performing a tagging or a browsing task? To systematically study these tasks and all possible variants, we rather study their underlying strategies, i.e., tagging single images sequentially (image-oriented) and tagging a single label on all images (label-oriented). Our exploration of the literature on annotation tools revealed three factors that likely impact the user performance, but were not investigated to date. The first is the number of images displayed. Displaying multiple images at once enables to rapidly scan them and identify patterns to assign labels to groups (e.g., [49]), but may also distract annotators when labeling images one-by-one and result in lower performance overall. The second is the interaction mode offered by the system. Browsing implies selecting images sequentially without changing the label to tag [68, 71], e.g., click on the label to tag then click on all relevant images, thus the label remains *persistent* until explicitly modified. The opposite, *transient* labels, require annotators to (re-)specify the label to tag for each image (e.g., [1]). While persistent labels might seem quicker [75], they may also induce mode errors and produce more errors overall [59]. The last factor is the image visual complexity. Complex images are harder to parse and require longer visual analyses to identify specific features [14, 74], which can be detrimental when browsing images.

We study these factors through an abstract task (Figure 1b) which replaces the features of a natural image (such as the group, species, and race of an animal [13] as illustrated on Figure 1) with 9 different geometric shapes that users must count (Figure 1b) using persistent or transient toggle buttons (Figure 1c). Using abstract images avoids any kind of bias linked to expert knowledge that users could leverage to quickly scan specific images. Our focus is to understand whether users tend to adopt an *image-oriented* strategy or a *label-oriented* strategy if they can choose between the two (Figure 1c), and whether one strategy is more efficient than the other.

We report the results of two experiments. The first experiment investigates what factors nudge participants into using an image-oriented or a label-oriented strategy. The second compares the efficiency of both strategies with regard to precision and time. The results suggest participants are more likely to adopt an image-oriented strategy regardless of the scenario they are facing, but they may adapt or completely switch their strategies based on the system's characteristics. They also exhibit evidence of shorter annotation times while using an image-oriented strategy, especially when analyzing images with high visual complexity. In contrast, participants appeared to perceive the label-oriented strategy as being efficient more often. These findings seem to indicate tagging tasks are more efficient than browsing tasks in general. We conclude with a list of design implications addressing the fact that systems should prioritize tagging tasks, but not disregard browsing tasks, when dealing with images with a high visual complexity.

2 RELATED WORK

In this section, we characterize image labeling tasks. We first review the type of labels used by image labeling systems and then discuss the possible impact of interactive strategies on labeling tasks. We continue with a list of metrics used for evaluating the efficiency of labeling tools, and end by exposing the potential impact of the image visual complexity on the user performance. Our literature review considers both image and video labeling tools as the latter consists essentially in annotating a set of frames [3, 16, 41, 61].

2.1 Image Labeling Tools

Image labeling is the action of tagging visual content with metadata. Sager et al.'s survey on image labeling [57] shows the corresponding data consists of free-form text [3, 55], ontological keywords [42, 67], or graphical elements to highlight parts of images and add spatial information to labels [16, 55, 60]. Hanbury [29] refers to the latter as segmentation and emphasizes segmented areas are concomitant with tags. Labeling tools such as LabelMe [55] are designed for this sole purpose. We focus our investigations on ontological labels because they can both tag entire images and be associated to image segments [28, 55, 56].

Labeling tools support *tagging* or *browsing* an image set [57]. Tools supporting tagging tasks display a single image at once [16, 22, 28, 67, 69] with sometimes means to navigate the whole set [22, 56]. Annotators might need to come up with keywords from scratch [55, 67, 69] or choose from a vocabulary [1, 42]. Tools supporting browsing tasks enable annotators to select a group of images and tag them with a given label [10, 31, 71], select a persistent label to tag on relevant images [49, 68], or drag-and-drop labels on relevant images or groups of images [62]. Image-based CAPTCHAs [17] are a

[1] the turkey image and its labels on Figure 1 was found on Flickr (https://tinyurl.com/y44fn3sz)

common example of a browsing task that sets the label to tag (Figure 1a). Assistive tagging [70] provides benefits for browsing tasks by automatically assigning labels to images that must be human-verified [68], or by grouping images together [62]. The body of work on image labeling tasks still lacks comprehensive evaluations of tagging and browsing tasks to assess their outcomes with regard to time and precision; to the best of our knowledge, only Yan et al. [75] compared time efficiency of both tasks with three users and presented comparable results. In our study, we propose a more systematic comparison of these two approaches by comparing the user performance when varying the number of images displayed, the interaction modes, and the image visual complexity.

2.2 Interaction Modes and Strategies

Interactive systems build on *modes* to either avoid performing repetitive actions or set a specific interactive context. Text editors such as Microsoft Word, for instance, enable writing bold or italic text by pressing a button in a toolbar. Similarly, vector graphics editors such as Adobe Illustrator often propose a toolbar with toggle buttons to switch between modes for drawing and moving shapes. Such modes are *persistent* and require users to explicitly switch between them. Systems supporting browsing tasks by fixing a label to tag on multiple images propose a similar mechanism [49, 68]. We refer in this case to *persistent* labels. On the other hand, these systems may enable to select relevant images before assigning them the same labels [10, 31, 71], in which case they rely on *transient* labels. Despite the advantages of modes, they are error-prone. A common caveat coined as *situational awareness* [18, 19, 72] encapsulates the fact that users might not know they triggered a mode or forgot the modes they are currently in, leading to mode errors [51, 58, 59]. We investigate the effects of persistent and transient labels on the user performance.

Tagging and browsing tasks lack rigorous definitions to clearly identify them in various scenarios. Besides, hybrid tasks such as tagging multiple labels on multiple images remain unsystematically explored (e.g., [49, 68]). To study the space of image labeling tasks, we identify and focus on the underlying labeling strategies, namely, an image-oriented strategy that consists in tagging all labels on single images and a label-oriented strategy that consists in tagging a single label on all images. Systems such as EVA [68] or ECAT [49] enable these strategies but did not directly study their efficiency.

Adopting a specific strategy can have various effects depending on the context. Mackay [44] showed that interaction techniques can be more efficient in specific interactive contexts. They compared floating palettes, marking menus, and a toolglass command selection techniques in the context of Petri-net editing tasks. Their results showed floating palettes were more efficient in a "copy" context and that users preferred toolglasses and marking menus when solving problems. Goguey et al. [26] compared variants of a tool palette and a toolglass in the context of object manipulation. They expected users to adopt an object-oriented (perform all commands on an object before moving to the next) or command-oriented (perform the same command on all relevant objects before moving to the next) strategy depending on the technique used. They proposed two metrics to assess the type of strategy adopted based on a sequence of interactions. Their results show that the toolglass technique induces an object-oriented strategy whereas tool palettes induce a command-oriented strategy. These strategies are equivalent to image-oriented and label-oriented strategies in the context of image labeling. We build on their metrics to assess the type of strategies adopted by the participants in the first experiment.

2.3 Efficiency Metrics for Labeling Tools

Task completion time is one of the primary variables used to assess the efficiency of an image labeling system. Volkmer et al. [68] compared labeling times with the mouse or keyboard and showed shorter times with the latter. Iakovidis et al. [33] compared their system to the LabelMe system [55] and showed a benefit of assistive tagging on labeling time. Yan et al. [75] proposed two models to predict the annotation time of an image set using either tagging or browsing. They evaluated their model empirically with three users and reported good accuracy. Chang et al. [8] investigated labeling techniques based on a 1D and 2D spatial categorization and reported comparable task completion times. They also compared the performance of annotators when tagging images entirely or by splitting the task hierarchically among crowd workers [9]. They reported faster completion times for the latter.

The label accuracy is also used to assess the efficiency of a system. Labeled image sets such as ImageNet [13] or LabelMe [55] do not report particular verification procedures to control the label quality, thus consider the labels as inherently accurate. To evaluate the quality of labels, one can compute the inter-user agreement between annotators to evaluate whether a consensus exists for each of them [27, 40, 68]. Once ground truth data exists, the precision of labels can be computed in comparison. For instance, in their studies, Chang et al. [7–9] compute the accuracy rate of annotators based on the number of correct labels.

Labeling tools can increase their efficiency by supporting the annotators with simpler tasks to perform, providing interactive shortcuts, or providing sequences of related images. However, the impact of user performance on the quality of the labels produced remains sparsely evaluated. Our study focuses on two main performance metrics: task completion time and accuracy rate based on ground truth data.

2.4 Visual Complexity and Task Difficulty

The label quality depends in part on the features the annotators can extract from an image. Some features *pop-out* and can be identified preemptively (e.g., colors), but others might require careful analyses and specific knowledge to appreciate details or simply identify objects. Donderi [14] reviews the concept of visual complexity and lists the main factors used throughout history to measure visual complexity, such as the characteristics of a single visual form, visual arrays, how information is picked up, or theories based on algorithmic information. Wolfe and Horowitz [74] review five forms of visual guidance and highlight two main approaches: bottom-up in which "aspects of the scene attract more attention than others", and top-down in which "attention is directed to objects with known features of desired targets". We build on these reviews to create a set of images with various visual complexities: images with the lower visual complexity rely on bottom-up guidance by providing pop-out features and images with the higher visual complexity rely

on top-down guidance building on shapes with high redundancy (see Fig.1 in [14]).

Visual search performance depends not only on an image complexity, but also on the goal of the search, the logic of the scene exposed, or the type of stimulus exposed beforehand. Wolfe et al. [73] showed evidence that searching for objects in real scenes leads to faster times than searching for items randomly. Unintentional factors can also affect performance. Kristjansson's review of the literature on implicit visual learning [37] highlighted visual searches can be faster when *primed* with features that share similarities with the target (e.g., position or shape). Their work also showed that not all visual features could produce priming effects, but some, such as color, always did [38]. A label-oriented strategy and consequently browsing tasks might benefit from the effects of priming by looking for specific features in images and ultimately provide shorter annotation times overall. We investigate these benefits in our study by allowing (experiment 1) or forcing (experiment 2) participants to focus on a single label.

3 STUDY CHARACTERISTICS

In this section, we introduce the task used in both experiments to simulate an image labeling task. We also describe the experimental setup and explain how we analyze quantitative data.

3.1 Task

Controlling the visual complexity. Building a set of natural images that provide distinct visual complexity levels and represent concepts that any annotator can relate to is a challenging task. To fully control these variables, we rather simulate an image labeling task using abstract images. These images consist of circles and polygons that represent features of a natural image, such as the type, number, and size of objects present in a scene. We vary the image *visual complexity* by manipulating the set of shapes and define three distinct levels (Figure 2). The low level differentiates shapes by colors, a known pop-out feature [74]. The medium level consists of polygons with 3 to 5 sides. The high level consists of polygons with 8 to 10 sides that lack salient features [14].

Experimental task. An image consists of 9 shapes. Each shape belongs to one out of four distinct categories. Image labels consist of shape counts for all categories of shapes except black circles that represent distractors. Shapes from a category appear between 2 and 4 times in a single image, and distractors between 1 and 2 times; there exist 3 labels per shape category for a vocabulary size of 9 (see the buttons on Figure 1c). All combinations constitute a set of 9 images. Users tag images using toggle buttons: they must select the correct shape counts and then click on the corresponding image. If all labels are tagged (Figure 1c, image-oriented example), the image vanishes and the user receives a blue or red visual feedback to indicate a right or wrong characterization. If the image set still contains untagged images that are not displayed, one of them replaces the one that just vanished. If only one or two labels are tagged (Figure 1c, label-oriented example), the image is partially labeled and labels appear on the left-hand side of the image. Users can later complete the remaining labels. We instruct participants to perform this task as quickly and precisely as possible.

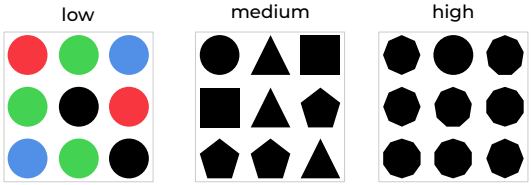

Figure 2: Levels of *visual complexity* used in the study. Black circles represent distractors (1 or 2 per image).

3.2 Experimental Setup

We used an iMac running macOS Big Sur (v11.5.2) consisting of a 27-inch display with a resolution of 5120×2880 pixels, and a 4 GHz Intel quad-core with 8 GB random-access memory. The experiment was implemented with web-technologies; we used a React [46] front-end connected to a Node.js server [24] using Socket.IO [24].

The buttons consisted of rectangles with a dimension of 150×50 pixels spaced by 10 pixels from each other, and images consisted of squares of 200×200 pixels including 9 shapes of 50×50 pixels (see Figure 1c).

3.3 Data Analysis

Analyzing results with null hypothesis significance testing (NHST) is increasingly being criticized by statisticians [4, 12, 30] and HCI researchers [5, 15, 32, 34]. Thus, our analyses rely on visual estimation techniques: we report differences between samples as plots depicting the mean of the differences and the effect sizes as 95% confidence intervals (CIs). We base our interpretations on the width of the CIs and their gap to the zero-line as recommended in the literature [12, 15]: the smaller the CIs are and the larger their distance is to the zero-line, the more evidence exists of a significant difference. While we make use of estimation techniques, a p-value approach reading of our results can be done by comparing our CIs spacing with common p-value spacing as shown by Krzywinski and Altman [39].

All plots use data aggregated by participant and depict 95% bootstrapped CIs. We compute them using the adjusted bootstrap percentile (BCa) from the *boot* [64] R package.

For both experiments, we first present a figure depicting all effects of the independent variables on the dependent variables studied, then only reference figures exhibiting interactions between factors. To support transparency [34, 47], we provide the data sets (including transcripts from the participants), and the source code of all the plots (single and combination of independent variables) at https://osf.io/dyj8p/?view_only=ca332f97a8ad4e79bf2abe8f9d86ae40.

4 EXPERIMENT 1: STRATEGY ADOPTION

This first experiment investigates the effects of the image visual complexity, the number of images displayed, and the persistence of labels on the user performance and the adoption of strategies.

The experiment follows a within-subject design involving 18 participants with three independent variables. We received a formal approval from our national ethics board to conduct this study under the identifying number 2022-14.

Independent variables. We investigate three main independent variables (IVs). The VISUAL COMPLEXITY (VC) of images varies between the three levels presented on Figure 2 (*low*, *medium*, *high*). We display images as a square matrix of size n and vary the MATRIX

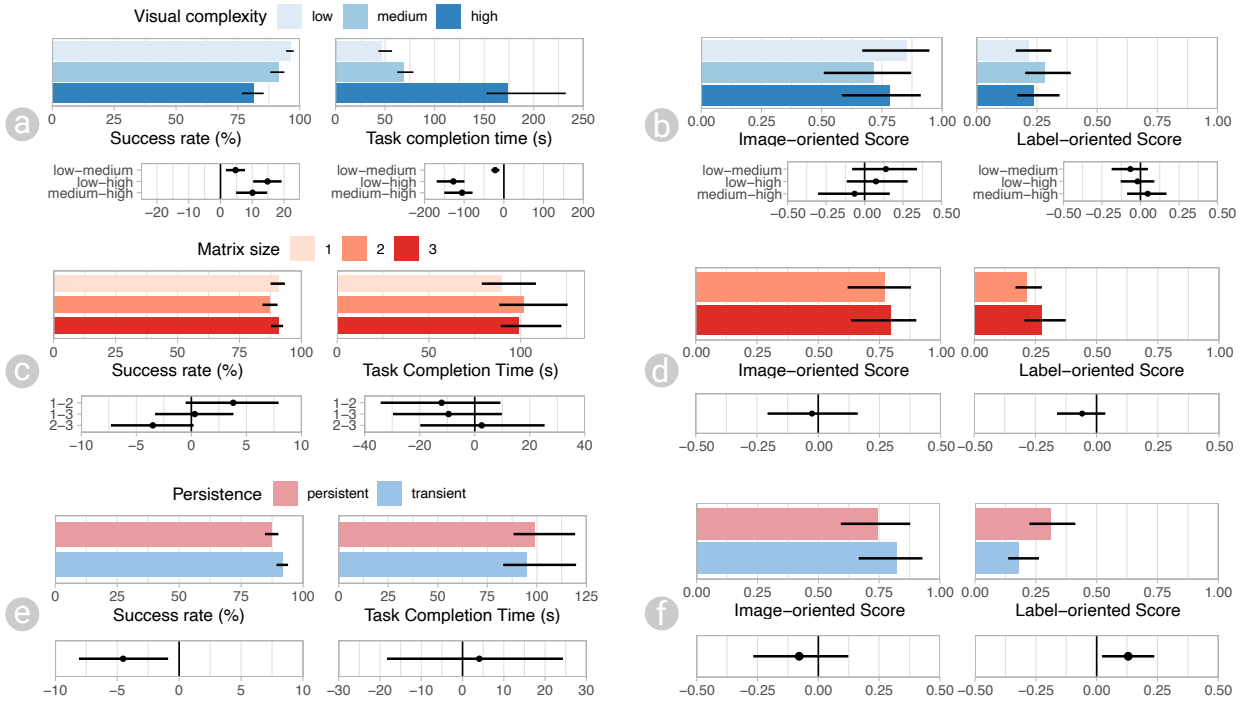

Figure 3: Effects of all IVs on the *success rate* and the *task completion time* (a,c,e), and both *strategy scores* (b,d,f). Non-bar plots represent means of the differences between samples of the plots above them. The smaller the CIs are and the larger their distance is to the zero-line, the more evidence exists of a significant difference.

SIZE (MS). We cover three unique cases by displaying a single image ($n = 1$, 1 image), a subset of the entire image set ($n = 2$, 4 images), the entire image set ($n = 3$, 9 images). We also vary the interaction mode by controlling the label PERSISTENCE (P). We switch between *persistent* and *transient* buttons: persistent buttons remain in the same state upon tagging an image, while transient buttons are toggled off.

Design. We counterbalance all independent variables and get a 3 VC × 3 MS × 2 P within-subject design. A block consists of a combination of the three independent variables and includes 9 images that participants must label. At the end of each block, we ask participants to describe the strategy they adopted. All blocks using the same VISUAL COMPLEXITY were performed in a row. The experiment starts with a 3-blocks training phase consisting of the following combinations of IVs ({VC, MS, P}): *{low, 3, persistent}, {medium, 2, persistent}, {high, 1, transient}*. The evaluation phase of the experiment consists of 18 blocks for a total of 162 images per participant.

Quantitative dependent variables. We consider four distinct dependent variables (DVs). Two of them are dedicated to measuring user performance: the *success rate* that corresponds to the percentage of correct tags produced, and the *task completion time* that corresponds to the duration of each block. We do not consider the time to characterize single images, as this factor heavily depends on the strategy adopted. Two other scores, namely *image-oriented* and *label-oriented scores*, assess the strategies used by the participants during each block of the experiment. These scores range between 0 and 1 and are based on Goguey et al.'s metrics [26]. The image-oriented score relies on image tagging actions (i.e., the user clicks on an image to tag it) and is defined by $S_{io} = 1 - \frac{P_{img}}{n-9}$ with n the number of image tagging actions and P_{img} defined as :

$$P_{img} = \sum_{i=3}^{n} \begin{cases} 1 & \text{if } Image(a_i) \neq Image(a_{i-1}) \\ & \text{and } \exists j \in [1; i-2] \text{ such as } Image(a_i) = Image(a_j) \\ 0 & \text{otherwise} \end{cases}$$

with $Image(a_i)$ representing the image tagged by the i^{th} action. If only 9 actions were performed within a block, the strategy was image-oriented, in which case $S_{io} = 1$.

The label-oriented score is based on button actions. A button action consists in a state change when the participant toggles a button on or off. We code such a change on three bits: $State(x) = B_1B_2B_3$. Each bit represents a label type (e.g., triangle, square, and pentagon for the medium complexity). A value of 1 indicates the label was changed since the last state (e.g., *010* would represent a click on the label "2 squares" on Figure 1c for both examples). We compute this score as $S_{lo} = 1 - \frac{P_{lab}}{n-3}$, with n the number of button actions and P_{lab} defined as :

$$P_{lab} = \sum_{i=3}^{n} \begin{cases} 1 & \text{if } State(a_i) \neq State(a_{i-1}) \\ & \text{and } \exists j \in [1; i-2] \text{ such as } State(a_i) = State(a_j) \\ 0 & \text{otherwise} \end{cases}$$

If only 3 actions were performed within a block, the strategy was label-oriented, in which case $S_{lo} = 1$.

We only compute the scores for a MATRIX SIZE greater than 1 as this condition forces an image-oriented strategy.

Qualitative dependent variables. We collect qualitative data from the participant explanations after each block. This data consists in

spoken comments transcribed into textual data that we cross with the strategy scores for finer analyses.

Additionally, we asked 5 open-ended questions at the end of the experiment to gather detailed feedback on the perceived effects of the IVs on user performance: *1) Did you perceive any impact of the number of images displayed on your performance? 2) ... strategies?, 3) Did you perceive any impact of the persistence on your performance? 4) ... strategies?, 5) Did you perceive any impact of the visual complexity on your strategies?*

Participants. We recruited 18 able workers from our laboratory to participate in our study (15M, 3F). We asked participants directly to confirm they did not have visual or motor impairments.

Research questions and hypotheses. We focus on the following research questions:

RQ1 Are the *success rate* and *task completion time* affected by the VISUAL COMPLEXITY, the MATRIX SIZE, and the PERSISTENCE?

RQ2 Do the *strategy scores* vary based on the VISUAL COMPLEXITY, the MATRIX SIZE, and the PERSISTENCE?

These questions translate to the following hypotheses:

H1 - The visual complexity impacts the labeling performance.

H2 - Persistent buttons facilitate a label-oriented strategy by minimizing back-and-forth movements between the buttons and the images, and leveraging priming effects.

H3 - Annotators adapt their strategies based on the interface and interactions they face to optimize their performance.

4.1 Results

RQ1 *Are the success rate and task completion time affected by the VISUAL COMPLEXITY, the MATRIX SIZE, and the PERSISTENCE?*

We depict on Figure 3 (a,c,e) the effects of all independent variables on the user performance, i.e., the *success rate* and the *task completion time*. The differences between the VISUAL COMPLEXITY levels exhibit clear deviation from the zero-line and relatively small confidence intervals (Figure 3a). These results present evidence of an effect between all visual complexity levels. As hypothesized (*H1*), the VISUAL COMPLEXITY strongly impacts the labeling performance.

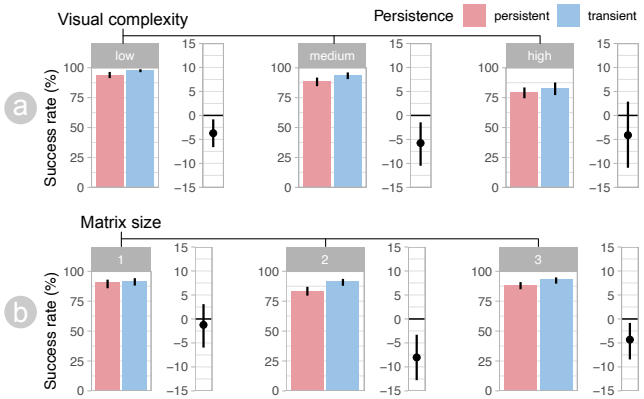

Figure 4: Interaction between the PERSISTENCE and the VISUAL COMPLEXITY (a), and the MATRIX SIZE (b), on the *success rate*.

The MATRIX SIZE did not seem to have a significant effect on the labeling performance independently of other factors (Figure 3c). We note some trends of an effect on the *success rate* that weakly suggest a MATRIX SIZE of 2 led to more errors.

We found evidence of an effect of the PERSISTENCE on the *success rate* indicating that persistent buttons seem to produce more errors (Figure 3e). On the other hand, we did not detect an effect of the PERSISTENCE on the *task completion times*.

By plotting all possible interactions between the IVs, we observed various trends and effects. First, the overall effect of the PERSISTENCE on the *success rate* seems to originate in part from labeling images with low and medium VISUAL COMPLEXITY (Figure 4a). Additionally, we observed an interaction between the PERSISTENCE and the MATRIX SIZE (Figure 4b) on the *success rate*. The results indicate a detrimental effect of persistent buttons when facing a matrix size of 2 and 3, with less evidence for the latter. One possible explanation of these effects is that complex images required more focus and a single image provided less visual information at once, so the participants were able to make fewer mistakes likely due to mode errors [59].

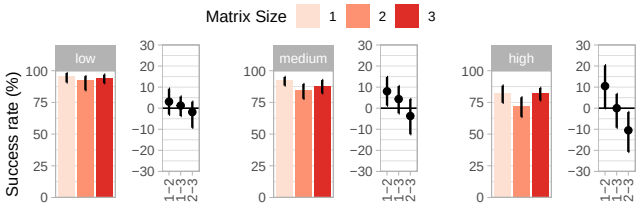

Figure 5: Interaction between the MATRIX SIZE and the VISUAL COMPLEXITY on the *success rate* when using persistent buttons.

We observed only small evidence of effects in relation to all IVs. We depict on Figure 5 the effects of the VISUAL COMPLEXITY and the MATRIX SIZE for persistent buttons. The results weakly suggest that a MATRIX SIZE of 2 produces worse labeling performance when dealing with complex images.

RQ2 *Do the strategy scores vary based on the VISUAL COMPLEXITY, the MATRIX SIZE, and the PERSISTENCE?*

We show on Figure 3 (b,d,f) the effects of all IVs on both the *image-oriented* and the *label-oriented scores*. We do not detect effects of the VISUAL COMPLEXITY or the MATRIX SIZE on either score. We found evidence of an effect of the PERSISTENCE on the *label-oriented score*, but no effect on the *image-oriented score* (Figure 3f). This seems to indicate the PERSISTENCE nudged participants into adopting a label-oriented strategy.

The results also exhibit evidence of an interaction between the PERSISTENCE and the VISUAL COMPLEXITY for the *label-oriented score* (Figure 6). This suggests a label-oriented strategy was used more often when labeling images with low and medium VISUAL COMPLEXITY with persistent buttons. Again, complex images seem to be less affected by the PERSISTENCE, hence limit the use of a label-oriented strategy.

4.2 What Triggers a Strategy Switch?

We plot the evolution of the strategy scores per participant through time on Figure 7a. Each entry on the y axis represents a block in the experiment (excluding blocks with a MATRIX SIZE of 1). Horizontal black lines denote a change of the VISUAL COMPLEXITY level.

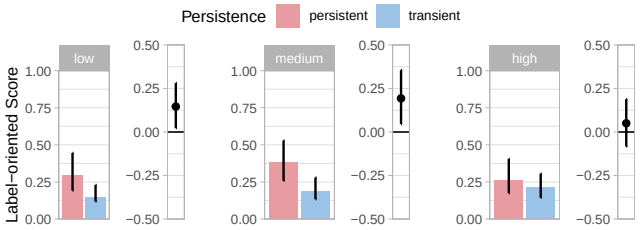

Figure 6: Interaction between the PERSISTENCE and the VISUAL COMPLEXITY on the *label-oriented score.*

We categorized the participant behaviors by crossing quantitative and qualitative results. We computed the variance of the absolute differences between consecutive blocks for both scores to understand to what extent did participant modified their strategy through time. We plot these measures in relation to each other on Figure 7b and cluster them into four groups using a k-means algorithm[2]. Using this method, the most decisive factor used for categorizing participants seems to be the image-oriented score. We note, however, that simply using the variance of the absolute differences is not perfect as it does not capture nuances from the qualitative results. P_6 and P_3, for instance, are likely outliers in their categories: P_6's strategy changed more than once without qualitative data explaining this behavior, and P_3 did not seem to change their strategy systematically compared to others in their category.

A striking result is that half of the participants exclusively used an image-oriented strategy (B1). One participant followed a fairly consistent strategy that evolved through time (B2). Three participants used an image-oriented strategy but switched transiently to the label-oriented strategy (B3). The rest of the participants adapted their strategies based on the task at hand and the opportunities the system offered (B4).

4.2.1 Single strategy (B1). P_1 commented after finishing a block with a MATRIX SIZE of 1 "*I like to have a single image, you analyze one by one [...] you don't have parasites*"$_\top$[3] and at the end of the experiment that "*I did very barely use annotations overall. I did not feel the need to do so*"$_\top$.

P_2 remarked their strategy sometimes did not match the label PERSISTENCE, which created some frustration "*I think it's more annoying to have persistent buttons. But the strategy was the same*".

P_4 made a similar comment "*the consistency of these buttons [is] really not helpful. It's rather confusing for me, I guess*", but felt otherwise later "*The buttons were persistent so, like suppose if the blue and green are same as earlier I don't need to click them. [...] It's easier. It saves time.*".

P_5 explained they adopted an image-oriented strategy based on their experience in the training phase "*if I find three triangles then I select all the images that have three triangles. [...] The other method is as follows: I just see one image and select the number of triangles, the number of squares, and the number of pentagons per image*". They later underlined benefits of using a single image "*I like the image on top of the other, [...] it's easier to compare [...] to the image that was before*", and compared it to matrix larger than one image using complex images "*This was difficult because it was too many images, it kinda make me feel like, dizzy. It's like the same image too many times, so it is more difficult than just having one image on top of the other*".

P_7 highlighted the benefits of matrices with more than one image. They remarked it enabled them to look for salient patterns in other images when looking for their next target "*When I had 4 images, I was trying to find for instance a [pattern of] 4 [shapes] to free space to have others [images]*"$_\top$ and added "*when there were several images, I tried to skim to see what was the most interesting to address*"$_\top$.

P_{10} explicitly stated using an image-oriented strategy "*Here I focused first on one of the images, and tried to finish that quick, then moved to the other ones*". They added later "*I did not look at the other squares [images] if there were some things similar, I just concentrated on one and finish one*" which indicates they did not perceive an impact of the MATRIX SIZE on their performance. Their answer to the related question at the end confirmed that "*even if the buttons were persistent, I just focused on one image, and I did not look the whole picture*".

P_{14} focused on a single image at once "*I have the impression I don't use the fact there are 4 [images] or not*"$_\top$ and mentioned facing multiple images might hinder them "*I am under the impression that for me it's easier when there is only one [image]*"$_\top$. They also remarked not leveraging the button persistence "*I think I don't really use persistence. [...] I did not try to know whether some images had similar stuff to the one I already put*"$_\top$.

Similarly to others, P_{18} remarked the MATRIX SIZE was better capped to one "*it was easier indeed when there was only one [image] because there was only one spot to look at, it changed automatically, and it implied less visual back-and-forth*"$_\top$.

We noted mixed remarks on the MATRIX SIZE: some participants felt disturbances from facing multiple images at once, while others took advantage of this to quickly scan images to find salient patterns to prioritize. We observed a similar heterogeneity for comments referring to the label PERSISTENCE. Participants described it as being useless (P_5 "*For me it was like useless*"), disturbing (P_{14} "*in the case of difficult shapes I think it rather disturbed me*"$_\top$), or helpful in some contexts (P_{17} "*sometimes I just had to click, change a 2 to a 3 and send the image*"$_\top$, P_4 "*it saves times*").

4.2.2 Evolutionary (B2). P_8 did not seem to settle on a single strategy throughout the experiment. They explained at the end of the first block "*I did by shape, which is easier for the eye [...] it's more challenging to count by [image] and to change each time*"$_\top$. When facing complex images they tended to use an image-oriented strategy "*I started with the 9 [sided shapes], because they have a bar at the bottom, so they are easier to identify. Then, [...] I counted either the 10 or 8, and I completed with the number of circles*"$_\top$.

This participant started clearly using a label-oriented strategy for the first VISUAL COMPLEXITY, which slowly faded as they seemed to use a more hybrid approach in the following blocks.

4.2.3 Strategy switch (B3). P_{13} switched their strategy when facing the highest VISUAL COMPLEXITY. They explained their strategy was to "*take the less complex shape, counted it in each image first, then do the same thing, and subtract circles*"$_\top$. It is unclear whether the

[2]we used the *kmeans* function from the *stats* R package [65]
[3]to support transparency [47], we provide all transcripts of participants with citations highlighted at https://osf.io/dyj8p/?view_only=ca332f97a8ad4e79bf2abe8f9d86ae40. Citations translated to English are marked with a $\top$

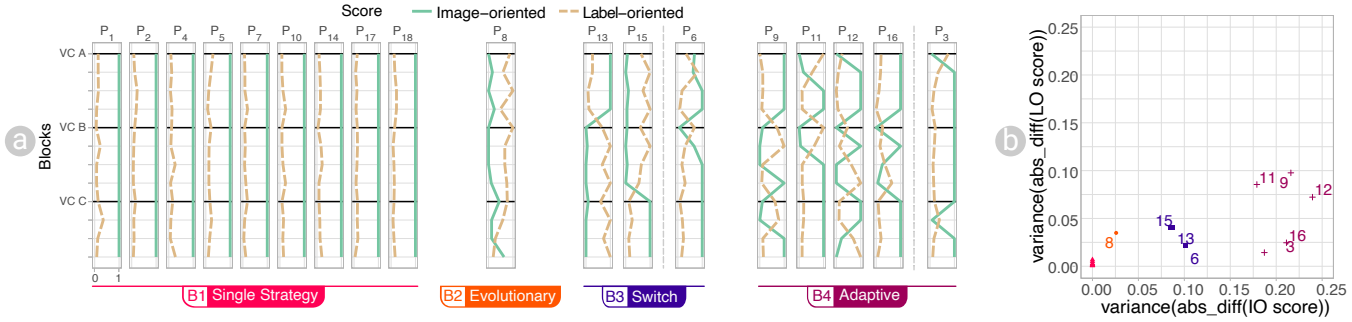

Figure 7: Evolution of the scores through time per participant (a). Each entry on the y axis represents a block. We categorized participant behaviors by crossing qualitative and quantitative results: we plot the variance of the absolute differences between consecutive blocks for both scores in relation to each other, and identify 4 clusters using a k-means algorithm (b).

participant would have switched their strategy back if they had to analyze shapes with lower complexities once more.

The trigger for P_{15} also related to the VISUAL COMPLEXITY "*I start to the first one in the grid [...] For the shapes I actively counted [...] like 'ok 3 3 and 2' [...] I think it's more obvious visually like it's not that much red compared to 1 2*". In contrast with P_{13} they switched to an image-oriented strategy when the VISUAL COMPLEXITY decreased, making it "more obvious visually" to identify patterns in each image.

P_6 did not comment on their strategy change. The scores indicate they used an hybrid strategy until they settled on an image-oriented one. We do not observe a clear *switch* in this case, thus mark this participant as an outlier in this category.

The two switches occurred in both cases on a change in the VISUAL COMPLEXITY level. The participants seemed to take that event as an opportunity to review their strategies and adopt a more efficient one. However, they did not consistently change their strategy when facing a more challenging task.

4.2.4 Adaptive strategies (B4). P_9 faced the highest VISUAL COMPLEXITY first and adopted an image-oriented strategy "*I did [image] by [image] while trying to count each element, select all buttons, and at the end reinitialize the buttons before starting the next [image]*"$_\top$. They then switched to a label-oriented strategy "*I tried to do first all triangles, then all squares*"$_\top$ and backtracked when buttons were not persistent anymore "*I did image by image because [...] with the strategy I used before [...] I had to go back on the button each time*"$_\top$. They continued with a label-oriented strategy "*I click the button, like 2 reds, and I skim, and if there are 2 reds I click [on the image]*"$_\top$ until the buttons became transient.

P_{11} started with a label-oriented strategy by relying on persistent buttons "*I tried to select all of those which had 2 triangles, 3 triangles, 4 triangles, and so on for all shapes*"$_\top$. They used an image-oriented strategy when they lost this persistence "*I did more or less the same strategy as when there is a single square [image], as it is not persistent*"$_\top$. They kept switching back and forth depending on the button persistence "*I did more or less the same strategy as with persistent 3x3 [...] in the medium difficulty*"$_\top$, and "*I did in order, not all colors on a single streak as it was not persistent*"$_\top$. They broke this cycle when facing the highest VISUAL COMPLEXITY and exclusively used an image-oriented strategy: "*I do image by image because I don't have a global perception.*"$_\top$.

P_{12} mentioned changing their strategy for the first VISUAL COMPLEXITY "*I changed in the middle*"$_\top$. They adopted a label-oriented strategy when facing the second VISUAL COMPLEXITY "*I started by [...] looking for instance all the triangles on the four images, [...] then all the squares, then all the pentagons. However, as the buttons were not persistent, I found it not efficient to navigate between the two each time*"$_\top$, but preferred an image-oriented one in the next block "*I did it differently this time. [...] The ones that looked easier, I did them in a single go, then I did the rest by looking whether a figure appeared 4 times*"$_\top$. They switched again when buttons became persistent "*as buttons were persistent, [...] I tried to [...] assess the number of triangles in the four images, and if two had the same, it allowed to click only once on the button*"$_\top$. They remarked using an image-oriented strategy for the last VISUAL COMPLEXITY "*here I did image by image again*"$_\top$, and changed with persistent buttons "*I rather did again by color because buttons are persistent*"$_\top$.

P_{16} used an image-oriented strategy for the first VISUAL COMPLEXITY, and tried a different approach when facing the highest VISUAL COMPLEXITY. They explained "*the pattern is much more difficult to analyze, I realized since I was making many errors, my strategy was likely not so good*"$_\top$ and that they tried to "*concentrate at the end only on [the shapes] that had a shape typology to do them in the whole grid*"$_\top$. They did not continue using that strategy because "*I realized buttons were not consistent, it was bothersome because I had to click them again so I will lose time again while traveling between the buttons*"$_\top$, and remarked "*I think I will try to do shape by shape [...] once the pre-selections will be saved*"$_\top$ referring to partial labels. Indeed, they switched again when buttons were persistent "*I tried to do by elimination of shape by shape as I said previously*"$_\top$. They reverted to an image-oriented strategy when facing the medium VISUAL COMPLEXITY "*I picked up again the thing where I empty first the first cell*"$_\top$.

P_3 changed back their strategy to the ones they used in the training phase on the last block "*I came back to do all... by color, all the 2, all the 3, all the 4*"$_\top$ and added "*I recalled it was a possibility and with colors it is easier to over-count. [...] It seemed less of a headache to do color by color instead of number.*"$_\top$. They did not provide any reason why they switched back to the former strategy later. This participant did not change their strategy systematically based on the experimental conditions, thus is likely an outlier in this category.

Participants mainly adapted their strategies according to the PERSISTENCE; persistent buttons induced a label-oriented strategy, as quantitative results indicate (Figure 3f). The VISUAL COMPLEXITY was also a prevailing factor in most cases, as participants either

could not apply a label-oriented strategy when the image analysis was too challenging, or they adopted this strategy to focus only on specific shapes.

4.3 Discussion

We list two important findings from this experiment. Half of the participants used an image-oriented strategy throughout the entire experiment, a behavior which did not seem to be affected by the experimental factors. This suggests that following a *tagging task* was preferred overall. Our results also hint that persistent buttons nudge participants into adopting a label-oriented strategy.

Half of the participants exclusively adopted an image-oriented strategy, regardless of the various experimental conditions, which does not support *H3*. Their remarks indicate they did not find benefits in dealing with a matrix of images and that the label persistence was mostly a hindrance. We provide two interpretations. The participants quickly found the most efficient strategy or the one they preferred, and were satisfied with their labeling performance using it. Another non-exclusive interpretation is that the system provided a "facilitate" nudge [6] that builds on the status-quo bias [54, 66]. This bias means that participants adopted a comfortable strategy that would require some effort to change, so they kept using the same throughout the experiment.

Quantitative and qualitative results both suggest the label persistence has an impact on the user performance and the adoption of a label-oriented strategy (*supporting H2*). This effect is particularly salient for the participants consistently adapting their strategies (B4); they explicitly noted that persistence played a role in switching strategies and the strategy scores bounced between extremes depending on the experimental conditions (Figure 7a). This echoes results from the literature that interaction techniques can be more efficient or more appreciated in specific scenarios [2, 26, 44], and that a system can nudge its users into specific behaviors [6, 66]. Strategy switches also happened when the visual complexity changed. This experiment is the first, to the best of our knowledge, to expose clear effects of the visual complexity on labeling tasks (*supporting H1*). Our experimental design does not allow us, however, to know whether these switches would be consistent in the case participants faced the same visual complexity multiple times.

The results also showed evidence of more errors when persistent buttons were used compared to transient buttons. When buttons are persistent, the user enters an interaction mode every time a button is activated and stays in the same mode until toggling on or off any button. Users must then stay aware of the mode they are in and can easily forget that a specific button was on [51, 58, 59]. This type of mode errors can find solutions through design [20]. They could likely be mitigated, for instance, by leveraging quasi-modes such as only toggle on labels when pressing keyboard keys.

5 EXPERIMENT 2: STRATEGY EFFICIENCY

The first experiment investigated triggers to adopt a strategy. This second experiment compares how the two strategies may impact user performance. In this regard, we fix the strategy and force participants to use one or the other. For the image-oriented strategy participants must select all types of labels to tag an image. For the label-oriented strategy, we enable a single type of label (e.g., triangle for the medium visual complexity) until all images are partially tagged with it. We then disable this type and enable another if images are not already completely labeled.

The experiment follows a within-subject design involving 24 participants with three independent variables. We used the same ethical approval as before.

Independent variables. We consider three variables: the STRATEGY (s) used, the VISUAL COMPLEXITY (vc) of images, and the PERSISTENCE (p). We fixed the matrix size to 3 to allow for a label-oriented strategy.

Structure. We counterbalance the three IVs and get a 2 s × 3 vc × 2 p within-subject design. Blocks still consist of 9 images each. Participants perform the task with one strategy, then use the other. Overall, the experiment consists of 12 blocks, thus a total of 108 images to tag per participant.

Quantitative dependent variables. Similarly to the first experiment, we measure the participant performance through their *success rate* and *task completion time*. To evaluate the perceived performance of participants, we hand them a questionnaire to fill out after finishing all the blocks using the same VISUAL COMPLEXITY, so 3 in total. The questionnaire includes five-levels continuous response scales that consist of statements the participants must agree or disagree with ("strongly disagree" ↔ "strongly agree"): (S1) *I preferred using the image-oriented strategy over the label-oriented strategy*, (S2) *The image-oriented strategy led to a high performance*, (S3) *The label-oriented strategy led to a high performance*, (S4) *When following an image-oriented strategy, using persistent buttons had a positive effect*, (S5) *When following a label-oriented strategy, using persistent buttons had a positive effect.*

Participants. We recruited 24 able workers from our laboratory who did not participate in the previous study (23M, 1F). We asked participants directly to confirm they did not have visual or motor impairments.

Research questions and hypotheses. For this second study, we focus on the following research questions:

RQ1 Does one labeling STRATEGY produce better *success rates* and *task completion times*?

RQ2 Do the VISUAL COMPLEXITY and the PERSISTENCE have an impact on the *success rate* and the *task completion time*?

RQ3 Is one STRATEGY preferred or perceived as more efficient?

These questions combined with the previous experiment results lead to the following hypotheses:

H1 - A label-oriented strategy produces shorter labeling times by leveraging priming effects [37, 38].

H2 - Persistent buttons produce mode errors and lead to worse performances.

H3 - Annotators prefer an image-oriented strategy overall.

5.1 Results

RQ1 *Does one labeling STRATEGY produce better success rates and task completion times?*

We depict on Figure 8 the effects of all IVs on the *success rate* and the *task completion time*. The results do not exhibit evidence of a difference between the two types of STRATEGY on the *success rate*, but they do indicate small evidence of a difference on the *task completion time* (Figure 8c). This suggests adopting an image-oriented strategy produces shorter labeling times overall.

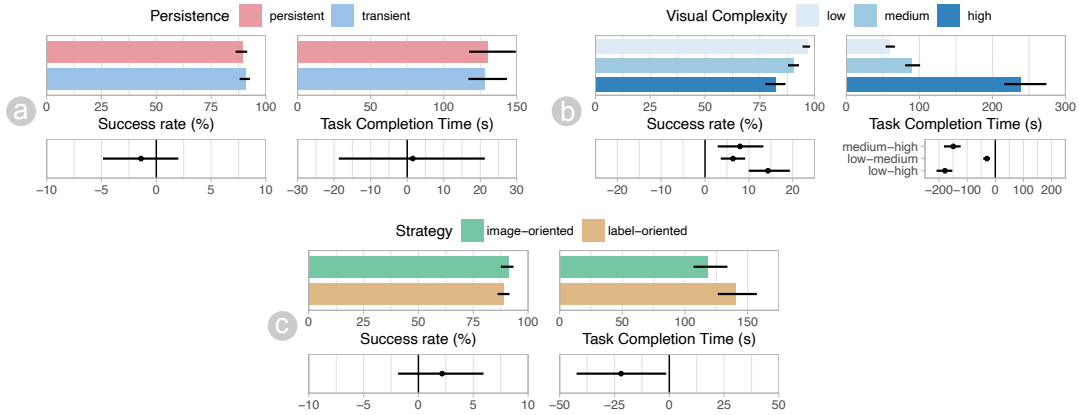

Figure 8: Effects of all IVs on the *success rate* and the *task completion time*.

We detected an interaction between the STRATEGY and the VISUAL COMPLEXITY on the *task completion time* (Figure 9). The results exhibit evidence of an effect of the STRATEGY for images with a high VISUAL COMPLEXITY. Overall, the effect of the STRATEGY on the *task completion time* seems to be particularly present when labeling images with higher VISUAL COMPLEXITY.

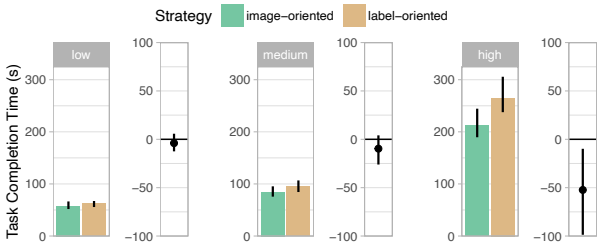

Figure 9: Interaction between the STRATEGY and the VISUAL COMPLEXITY on the *task completion time*.

RQ2 *Do the VISUAL COMPLEXITY and the PERSISTENCE have an impact on the success rate and task completion time?*

The results again provide strong evidence of an effect of the VISUAL COMPLEXITY on both the *success rate* and the *task completion time* (Figure 8b). This reinforces the fact that the VISUAL COMPLEXITY is an important factor when studying image labeling tasks and should be considered more often.

The PERSISTENCE, however, did not seem to have an effect on the user performance in this experiment. We also did not detect any interaction between this IV and the others. This result tones down the results obtained in the first experiment and tend to indicate persistent buttons are not intrinsically error-prone.

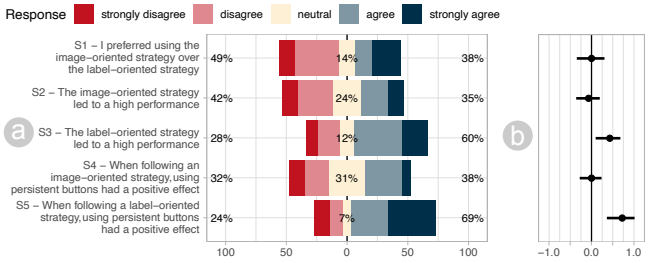

Figure 10: Aggregation of the questionnaire responses (a) and their distributions in relation to the "neutral" answer (b).

RQ3 *Is one STRATEGY preferred or perceived as more efficient?*

We aggregated the questionnaire responses for all statements on Figure 10a. We display on Figure 10b the response distributions in relation to the "neutral" response. The participants did not seem to prefer one strategy over the other consistently. To assess the effect of the VISUAL COMPLEXITY on the preference, we depict on Figure 11a the distribution of the responses for all the VISUAL COMPLEXITY levels. The results do not indicate an effect of this factor on the participant preferences.

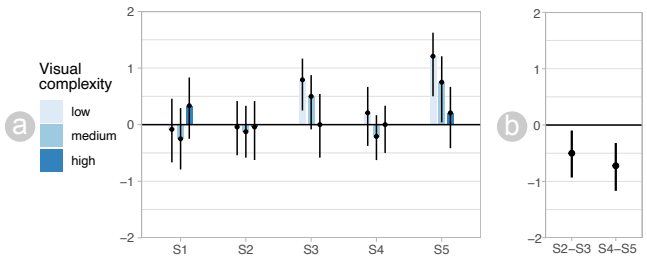

Figure 11: Distributions of questionnaire responses in relation to the "neutral" answer (a). Differences of responses between S2-S3 and S4-S5 (b).

The response distributions on Figure 10b seem to indicate the participants perceived the label-oriented strategy as being efficient, and that the PERSISTENCE provided benefits when using this strategy, but did not feel similarly about the image-oriented strategy. We compared responses from S2-S3 and S4-S5 to understand whether the participants' judgment was systematic (Figure 11b). The results exhibit strong evidence of a difference for both comparisons. This indicates the participants perceived the label-oriented strategy as being efficient more often than the image-oriented strategy, and persistent buttons being only beneficial for a label-oriented strategy.

5.2 Discussion

The major finding of this second experiment is that an image-oriented strategy seems to produce shorter labeling times overall, particularly when labeling images with a high visual complexity. Experimental results indicate that completing a block of 9 images using an image-oriented strategy took in average 21.99s [2.03, 42.97] less than using a label-oriented strategy (Figure 8c), which can have

a great impact on the long-term for large image sets. *H1* is therefore not supported. Combined with results from the first experiment, this suggests *tagging tasks* provide advantages over *browsing tasks* in specific contexts. We also found mixed results on the effects of the label persistence: results from the first experiment showed evidence of more errors with persistent buttons that results from the second did not confirm (*H2* is not supported).

We explain the advantages of the image-oriented strategy by the fact that focusing on a single image enabled the participants to rely on their working memory and scan part of the images only once. Müller and Krummenacher's review on visual search and selective attention [48] highlights the role of memory in such a scenario and discuss an effect coined as "inhibition of return" [35, 53, 63]. This effect consists in marking analyzed areas to favor not-yet-scanned locations. While we expected beneficial effects of *priming* [37, 38] (*H1*), i.e., advantages of focusing on a single shape to find it faster in a set of images (label-oriented strategy), the results did not yield such an effect. This finding implies that based on the visual complexity of the image set, an interactive system should prioritize *tagging tasks* to allow for shorter labeling times.

The results of this second experiment did not exhibit evidence of a participant preference for either strategies (*H3* is not supported). However, we found evidence that participants perceived the label-oriented strategy as leading to high performances more often than the image-oriented strategy. They also evaluated the label persistence as more beneficial when using the label-oriented strategy, an effect that quantitative results did not support. While the literature provides evidence that user preference does not always correlate to their performance [25, 50], participants seemed to feel more confident using the label-oriented strategy. This raises an interesting design challenge for labeling tools: should they maximize the labeling performance and ultimately the quantity of labels produced, or rather the positive perception annotators have of their work? The latter is likely significant to alleviate the tediousness of labeling tasks and should not be disregarded.

6 DESIGN IMPLICATIONS AND LIMITATIONS

In this section we summarize results from the study and propose a list of design recommendations for labeling tools.

The results of the first experiment suggest that an image-oriented strategy is more likely to be adopted by default, and the results of the second experiment showed evidence of shorter labeling times using this strategy compared to a label-oriented strategy. We also found evidence that the label persistence nudge annotators into using a label-oriented strategy, and that overall they tended to find this strategy as being efficient more often than the image-oriented strategy. These findings underline two pieces of information: *tagging tasks* seem to provide advantages over *browsing tasks* overall, and they highlight a trade-off in the efficiency of browsing tasks and the perception annotators have of them.

From these findings, we formulate the following design recommendations:

DR1 Prioritize tagging tasks to maximize the annotator labeling performance

The study findings hint that displaying a single image to label at a time induces consistently shorter labeling times, particularly for images with a high visual complexity. Qualitative results from the first experiment suggest that displaying multiple images can be disturbing when using an image-oriented strategy. Therefore, tools supporting tagging tasks [1, 16, 22] are likely to produce more labels faster.

DR2 Consider the image visual complexity when designing a labeling tool

Browsing tasks are not always less efficient than tagging tasks: our analysis did not detect a difference between the two tasks for images with lower visual complexities. Besides, the results from the second experiment provided evidence that participants felt more confident in their performance when using a label-oriented strategy. Therefore, we recommend considering the complexity of images when designing labeling tools to support tagging tasks only if necessary and allow annotators to choose otherwise. A labeling tool could ideally adapt its interface and interaction means by assessing the image complexity with computational methods such as [11, 21, 43].

DR3 Mind mode errors

Some labeling tools use persistent labels to support browsing tasks [49, 68]. The study results exposed mixed findings on the effects of these persistent modes: persistent labels produced more errors in the first experiment, but we did not detect a similar effect in the second. This indicates using persistent labels has a chance to lead to errors, a result supported by the literature on mode errors [51, 58, 59]. To completely avoid possible problems, we recommend leveraging transient interactions that do not seem to impact the precision while providing comparable labeling times.

6.1 Limitations

This study is the first, to the best of our knowledge, to comprehensively investigate user preference and performance using tagging or browsing in the context of image labeling. We had to make compromises that might limit its outreach. We used abstract images to precisely control a set of visual features and produce 3 levels of visual complexity. One advantage of using abstract images was to limit biases linked to user knowledge of specific visual content (e.g., dog breeds). The visual complexity levels simulate the time and precision one requires to analyze natural images, ranging from easy to difficult. Annotators, regardless of their expertise, face simple and complex images to analyze that produce comparable annotation performances than the ones observed in the experiments, for which the conclusions drawn should apply. Nevertheless, further research is required on natural images to ascertain this claim. We also only focused on a fixed vocabulary of labels and did not vary its size. Future work should investigate whether the current findings remain true based on the number of labels and their type. Additionally, our experimental design does not allow for assessing the inter-participant bias for choosing specific strategies; findings concerning the adoption of strategies might be impacted by specific user archetypes that we did not control.

7 CONCLUSION

We reported a study investigating the performance of annotators when labeling an image set. Our goal was to compare the efficiency of two conventional tasks supported by labeling tools: tagging and browsing. We characterized the labeling task and identified two underlying labeling strategies, namely image-oriented and label-oriented. In a first experiment, we investigated the triggers to adopt one or the other strategy and found evidence that the image-oriented strategy is more likely to be adopted, and that factors such as the label persistence could nudge participants into adopting a label-oriented strategy. In a second experiment, we evaluated the efficiency of both strategies and found evidence that the image-oriented strategy produces shorter labeling times, especially for images with a high visual complexity. We also observed that participants felt more confident using the label-oriented strategy. Overall, the study results indicate that tagging tasks are likely more efficient than browsing tasks, but that the latter might be better perceived by annotators despite they did not seem to provide advantages in terms of performance. Findings from this work enabled us to list three design recommendations for labeling tools.

ACKNOWLEDGMENTS

This project was funded by the ANR PPR STHP 2020 (PerfAnalytics project, ANR 20-STHP-0003). We would like to thank all participants in our study who voluntarily helped us, and Jacob Wobbrock for replying to our statistical inquiries.

REFERENCES

[1] Axel Antoine, Sylvain Malacria, Nicolai Marquardt, and Géry Casiez. 2021. Interaction Illustration Taxonomy: Classification of Styles and Techniques for Visually Representing Interaction Scenarios. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–22. https://doi.org/10.1145/3411764.3445586

[2] Caroline Appert, Michel Beaudouin-Lafon, and Wendy E Mackay. 2005. Context matters: Evaluating Interaction Techniques with the CIS Model. In *People and Computers XVIII — Design for Life*, Sally Fincher, Panos Markopoulos, David Moore, and Roy Ruddle (Eds.). Springer London, London, 279–295. https://doi.org/10.1007/1-84628-062-1_18

[3] Olivier Aubert and Yannick Prié. 2007. Advene: an open-source framework for integrating and visualising audiovisual metadata. In *Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07*. ACM Press, Augsburg, Germany, 1005. https://doi.org/10.1145/1291233.1291451

[4] Monya Baker. 2016. Statisticians issue warning over misuse of P values. *Nature* 531, 7593 (March 2016), 151–151. https://doi.org/10.1038/nature.2016.19503

[5] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290607.3310432 event-place: Glasgow, Scotland Uk.

[6] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300733

[7] Chia-Ming Chang, Yi He, Xi Yang, Haoran Xie, and Takeo Igarashi. 2022. DualLabel: Secondary Labels for Challenging Image Annotation. In *Graphics Interface 2022*. https://openreview.net/forum?id=rrBz2lFETzq

[8] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445165

[9] 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)*. IEEE, Memphis, TN, USA, 139–143. https://doi.org/10.1109/VLHCC.2019.8818828

[10] Cheng-Chieh Chiang. 2013. Interactive tool for image annotation using a semi-supervised and hierarchical approach. *Computer Standards & Interfaces* 35, 1 (2013), 9.

[11] Valeriy Chikhman, Valeriya Bondarko, Marina Danilova, Anna Goluzina, and Yuri Shelepin. 2012. Complexity of images: Experimental and computational estimates compared. *Perception* 41, 6 (2012), 631–647. Publisher: SAGE Publications Sage UK: London, England.

[12] Geoff Cumming. 2014. The New Statistics: Why and How. *Psychological Science* 25, 1 (2014), 7–29. https://doi.org/10.1177/0956797613504966 arXiv:https://doi.org/10.1177/0956797613504966 PMID: 24220629.

[13] 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*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[14] Don C. Donderi. 2006. Visual complexity: A review. *Psychological Bulletin* 132, 1 (Jan. 2006), 73–97. https://doi.org/10.1037/0033-2909.132.1.73

[15] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, 291–330. https://doi.org/10.1007/978-3-319-26633-6_13 Series Title: Human–Computer Interaction Series.

[16] 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*. ACM, Nice France, 2276–2279. https://doi.org/10.1145/3343031.3350535

[17] Jeremy Elson, John R. Douceur, Jon Howell, and Jared Saul. 2007. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In *Proceedings of the 14th ACM conference on Computer and communications security - CCS '07*. ACM Press, Alexandria, Virginia, USA, 366. https://doi.org/10.1145/1315245.1315291

[18] Mica R. Endsley. 1995. Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 65–84. https://doi.org/10.1518/001872095779049499

[19] Mica R. Endsley. 2015. Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making* 9, 1 (March 2015), 4–32. https://doi.org/10.1177/1555343415572631

[20] Mica R Endsley, Betty Bolté, and Debra G Jones. 2003. *Designing for situation awareness: An approach to user-centered design*. CRC press.

[21] Carlos Fernandez-Lozano, Adrian Carballal, Penousal Machado, Antonino Santos, and Juan Romero. 2019. Visual complexity modelling based on image features fusion of multiple kernels. *PeerJ* 7 (July 2019), e7075. https://doi.org/10.7717/peerj.7075

[22] Niklas Fiedler, Marc Bestmann, and Norman Hendrich. 2019. ImageTagger: An Open Source Online Platform for Collaborative Image Labeling. In *RoboCup 2018: Robot World Cup XXII*, Dirk Holz, Katie Genter, Maarouf Saad, and Oskar von Stryk (Eds.). Vol. 11374. Springer International Publishing, Cham, 162–169. https://doi.org/10.1007/978-3-030-27544-0_13 Series Title: Lecture Notes in Computer Science.

[23] Inc Flickr. Last visited in September 2022. Flickr. https://www.flickr.com/.

[24] OpenJS Foundation. Last visited in August 2022. Node.js. https://nodejs.org/en/.

[25] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. 2000. Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. Association for Computing Machinery, New York, NY, USA, 345–352. https://doi.org/10.1145/332040.332455 event-place: The Hague, The Netherlands.

[26] Alix Goguey, Julie Wagner, and Géry Casiez. 2015. Quantifying Object-and Command-Oriented Interaction. In *IFIP Conference on Human-Computer Interaction*. Springer, 231–239.

[27] Joey Hagedorn, Joshua Hailpern, and Karrie G. Karahalios. 2008. VCode and VData: illustrating a new framework for supporting the video annotation workflow. In *Proceedings of the working conference on Advanced visual interfaces - AVI '08*. ACM Press, Napoli, Italy, 317. https://doi.org/10.1145/1385569.1385622

[28] Christian Halaschek-Wiener, Jennifer Golbeck, Andrew Schain, Michael Grove, Bijan Parsia, and Jim Hendler. 2005. Photostuff-an image annotation tool for the semantic web. In *Proceedings of the 4th international semantic web conference*. Citeseer, 6–10.

[29] Allan Hanbury. 2008. A survey of methods for image annotation. *Journal of Visual Languages & Computing* 19, 5 (Oct. 2008), 617–627. https://doi.org/10.1016/j.jvlc.2008.01.002

[30] Lisa L Harlow, Stanley A Mulaik, and James H Steiger. 2013. *What if there were no significance tests?* Psychology Press.

[31] Alexander G. Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang, and Ming-Yu Chen. 2006. Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*. ACM Press, Santa Barbara, CA, USA, 385. https://doi.org/10.1145/1180639.1180721

[32] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besançon. 2021. Can Visualization Alleviate Dichotomous Thinking? Effects of Visual Representations on the Cliff Effect. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2021), 3397–3409. https://doi.org/10.1109/TVCG.2021.3073466

[33] D. K. Iakovidis, T. Goudas, C. Smailis, and I. Maglogiannis. 2014. Ratsnake: A Versatile Image Annotation Tool with Application to Computer-Aided Diagnosis. *The Scientific World Journal* 2014 (2014), 1–12. https://doi.org/10.1155/2014/286856

[34] Transparent Statistics in Human–Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. https://doi.org/10.5281/zenodo.1186169 (Available at https://transparentstats.github.io/guidelines).

[35] Raymond Klein. 1988. Inhibitory tagging system facilitates visual search. *Nature* 334, 6181 (Aug. 1988), 430–431. https://doi.org/10.1038/334430a0

[36] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. Crowdsourcing in Computer Vision. *Foundations and Trends® in Computer Graphics and Vision* 10, 3 (2016), 177–243. https://doi.org/10.1561/0600000071

[37] Árni Kristjánsson. 2006. Rapid learning in attention shifts: A review. *Visual Cognition* 13, 3 (Feb. 2006), 324–362. https://doi.org/10.1080/13506280544000039

[38] Árni Kristjánsson. 2006. Simultaneous priming along multiple feature dimensions in a visual search task. *Vision Research* 46, 16 (Aug. 2006), 2554–2570. https://doi.org/10.1016/j.visres.2006.01.015

[39] Martin Krzywinski and Naomi Altman. 2013. Error bars: the meaning of error bars is often misinterpreted, as is the statistical significance of their overlap. *Nature methods* 10, 10 (2013), 921–923.

[40] Walter S. Lasecki, Mitchell Gordon, Danai Koutra, Malte F. Jung, Steven P. Dow, and Jeffrey P. Bigham. 2014. Glance: rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, Honolulu Hawaii USA, 551–562. https://doi.org/10.1145/2642918.2647367

[41] Bokyung Lee, Michael Lee, Pan Zhang, Alexander Tessier, and Azam Khan. 2019. Semantic Human Activity Annotation Tool Using Skeletonized Surveillance Videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 312–315. https://doi.org/10.1145/3341162.3343807

[42] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag Ranking. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/1526709.1526757 event-place: Madrid, Spain.

[43] Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. 2015. Computerized measures of visual complexity. *Acta Psychologica* 160 (Sept. 2015), 43–57. https://doi.org/10.1016/j.actpsy.2015.06.005

[44] Wendy E. Mackay. 2002. Which Interaction Technique Works When? Floating Palettes, Marking Menus and Toolglasses Support Different Task Strategies. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '02)*. Association for Computing Machinery, New York, NY, USA, 203–208. https://doi.org/10.1145/1556262.1556294 event-place: Trento, Italy.

[45] Meta. Last visited in September 2022. Instagram. https://www.instagram.com/.

[46] Inc Meta Platforms. Last visited in August 2022. React. https://reactjs.org/.

[47] Andrew Moravcsik. 2014. Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics* 47, 1 (2014), 48–53. https://doi.org/10.1017/S1049096513001789

[48] Hermann J. Müller and Joseph Krummenacher. 2006. Visual search and selective attention. *Visual Cognition* 14, 4-8 (Aug. 2006), 389–410. https://doi.org/10.1080/13506280500527676

[49] Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, and Klaus Schoeffmann. 2019. ECAT - Endoscopic Concept Annotation Tool. In *MultiMedia Modeling*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Springer International Publishing, Cham, 571–576.

[50] Jakob Nielsen and Jonathan Levy. 1994. Measuring usability: preference vs. performance. *Commun. ACM* 37, 4 (1994), 66–75. Publisher: ACM New York, NY, USA.

[51] Donald A. Norman. 1981. Categorization of action slips. *Psychological Review* 88, 1 (1981), 1–15. https://doi.org/10.1037/0033-295X.88.1.1 Place: US Publisher: American Psychological Association.

[52] Oded Nov and Chen Ye. 2010. Why Do People Tag? Motivations for Photo Tagging. *Commun. ACM* 53, 7 (jul 2010), 128–131. https://doi.org/10.1145/1785414.1785450

[53] Michael I Posner, Yoav Cohen, and others. 1984. Components of visual orienting. *Attention and performance X: Control of language processes* 32 (1984), 531–556. Publisher: Hilldale, NJ.

[54] Ilana Ritov and Jonathan Baron. 1992. Status-quo and omission biases. *Journal of Risk and Uncertainty* 5, 1 (Feb. 1992). https://doi.org/10.1007/BF00208786

[55] 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, 1-3 (May 2008), 157–173. https://doi.org/10.1007/s11263-007-0090-8

[56] Carsten Saathoff, Krishna Chandramouli, Werner Bailer, Peter Schallauer, and Raphaël Troncy. 2011. Multimedia Annotation Tools. In *Multimedia Semantics*, Raphaël Troncy, Benoit Huet, and Simon Schenk (Eds.). John Wiley & Sons, Ltd, Chichester, UK, 223–239. https://doi.org/10.1002/9781119970231.ch13

[57] Christoph Sager, Christian Janiesch, and Patrick Zschech. 2021. A survey of image labelling for computer vision applications. *Journal of Business Analytics* 4, 2 (July 2021), 91–110. https://doi.org/10.1080/2573234X.2021.1908861

[58] Nadine B. Sarter and David D. Woods. 1992. Mode Error in Supervisory Control of Automated Systems. *Proceedings of the Human Factors Society Annual Meeting* 36, 1 (Oct. 1992), 26–29. https://doi.org/10.1177/154193129203600108

[59] Nadine B. Sarter and David D. Woods. 1995. How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 5–19. https://doi.org/10.1518/001872095779049516

[60] Alexander Sorokin and David Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Anchorage, AK, USA, 1–8. https://doi.org/10.1109/CVPRW.2008.4562953

[61] Alberto Soto, Oleguer Camerino, Xavier Iglesias, M. Teresa Anguera, and Marta Castañer. 2019. LINCE PLUS: Research Software for Behavior Video Analysis. *Apunts Educació Física i Esports* 137 (July 2019), 149–153. https://doi.org/10.5672/apunts.2014-0983.es.(2019/3).137.11

[62] Bongwon Suh and Benjamin B Bederson. 2004. Semi-Automatic Image Annotation Using Event and Torso Identification. (2004), 4.

[63] Yuji Takeda and Akihiro Yagi. 2000. Inhibitory tagging in visual search can be found if search stimuli remain visible. *Perception & Psychophysics* 62, 5 (July 2000), 927–934. https://doi.org/10.3758/BF03212078

[64] *boot* package (R programming language). Last visited in September 2022. https://www.rdocumentation.org/packages/boot/versions/1.3-28.

[65] *stats* package (R programming language). Last visited in August 2022. https://rdocumentation.org/packages/stats/versions/3.6.2.

[66] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven. OCLC: ocn181517463.

[67] Hsing-Lin Tsai, Cheng-Hsien Han, En-Hsin Wu, Chi-Lan Yang, and Hao-Chuan Wang. 2015. Tag & Link: Supporting Regional and Relational Tagging in Images with Direct Annotation. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1791–1796. https://doi.org/10.1145/2702613.2732857

[68] Timo Volkmer, John R. Smith, and Apostol (Paul) Natsev. 2005. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*. ACM Press, Hilton, Singapore, 892. https://doi.org/10.1145/1101149.1101341

[69] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*. ACM Press, Vienna, Austria, 319–326. https://doi.org/10.1145/985692.985733

[70] Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *Comput. Surveys* 44, 4 (Aug. 2012), 1–24. https://doi.org/10.1145/2333112.2333120

[71] Liu Wenyin, Susan T Dumais, Yanfeng Sun, HongJiang Zhang, Mary Czerwinski, Brent A Field, et al. 2001. Semi-Automatic Image Annotation.. In *Interact*, Vol. 1. Citeseer, 326–333.

[72] Christopher D. Wickens. 2008. Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 397–403. https://doi.org/10.1518/001872008X288420

[73] Jeremy M. Wolfe, George A. Alvarez, Ruth Rosenholtz, Yoana I. Kuzmova, and Ashley M. Sherman. 2011. Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics* 73, 6 (Aug. 2011), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3

[74] Jeremy M. Wolfe and Todd S. Horowitz. 2017. Five factors that guide attention in visual search. *Nature Human Behaviour* 1, 3 (March 2017), 0058. https://doi.org/10.1038/s41562-017-0058

[75] Rong Yan, Apostol Natsev, and Murray Campbell. 2008. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. https://doi.org/10.1109/CVPR.2008.4587380 ISSN: 1063-6919.