Effects of Image Presentation Highlighting and Accuracy on Target Category Learning

David T. Slayback, Benjamin T. Files, Brent J. Lance, and Justin Ryan Brooks

***Abstract*—This study alters various exemplar presentation parameters to determine their effects on human online category learning for a future system that combines humans and computer vision (CV). Online category learning is necessary in this system because we envision that humans will need to provide input to assist CV modules in determining category labels without reducing throughput and without necessarily having expert knowledge of each category. In our study, subjects participated in a Rapid Serial Visual Presentation paradigm in which they were asked to determine the target category from highlighted exemplar images interspersed among distractor images. In Experiment 1, the highlighting method was varied among four options and a negative (no-label) and positive (explicit, text-based) control. In Experiment 2, label accuracy was altered by incorrectly labeling some distractor and exemplar images. In both experiments, there were three levels of difficulty that varied the similarity between distractor and exemplar images. The results show that most highlighting methods resulted in equivalent accuracy to the positive control, but certain modalities were more effective at varying difficulty levels. In addition, the subject accuracy was sensitive to distractors highlighted as targets, but not to non-highlighted exemplars. Our results indicate that human online category learning can be optimized for human–system interaction.**

***Index Terms*—Attentional processes, human–automation interaction, human–systems integration, information processing.**

I. Introduction

Combining human and autonomous agents to leverage the advantages of each has been a focus of many recent studies. Human–autonomy teams already outperform either human or autonomy alone in several applications, including predictions in complex environments [1], playing strategy games [2], and planning and teleoperation [3], [4]. The present study seeks to facilitate the human's contribution in a paradigm that combines humans with computer vision (CV) for image classification by optimizing the way in which the image category of interest is communicated to the human. If the human can quickly learn the new category from a small number of labeled images, then the human could transfer that understanding by rapidly labeling more images for training the CV system.

Image classification is ripe for a combined approach because the limitations of humans and computers in this area are numerous and well documented [5], [6]. Recent research shows humans were better than CV algorithms at classifying images with small objects, color, and contrast-distorting filters, abstract representations, and strange viewpoints, whereas CV more accurately classified fine-grained distinctions (such as between species) and less common labels [7]. Using the *ImageNet* database, the best CV algorithm had an error rate of 6.8%, while a trained annotator had an error rate of 5.1% [7], [8]. A combination of human and computer, however, could ameliorate the issues of fatigue and training time for humans, while allowing their input to correct for more heavily obscured images.

Humans and computers have been teamed for image classification before, such as in object detection and classification [8]. Generally, these studies have explicitly provided a known target of interest and required the participants to label presented images via various manual and automatic mechanisms. The labels derived from the human are then combined with labels from CV to generate a system-level combined label for a given image. Within the constraints of known, communicable target categories, this paper has shown significant progress. For example, in [9], labels generated by human and CV agents were fused, resulting in a 5% increase in labeling precision even over a fusion of several CV algorithms. In [10], an ensemble of humans and CVs performed 1.6 times better than a purely automated ensemble. Human EEG output collected during rapid serial visual presentation (RSVP) tasks has been successfully used to guide satellite image classification [11], [12], and some CV algorithms, such as active learning, explicitly require human input to achieve their performance [13]. Even pure CV-based algorithms have combined humans and autonomy to create labeled datasets more efficiently, resulting in equally accurately labeled data with 40 times the number of images as could be obtained with manual labeling [14].

Despite these successes, much of the research on human-in-the-loop systems has focused on adapting vision algorithms to unreliable human inputs rather than optimizing those inputs [15], [16]. Crucially, these studies presume that explicitly communicating a target label (e.g., "fish") is both an optimal and viable strategy for teaching the human the image category of interest. This is category identification and works well in general domains that humans can be expected to know. This presumption will almost certainly fail, however, if the target category of interest is unknown to the participant. For example, a participant with no background knowledge of dog breeds is unlikely to accurately label images of Pomeranians among images of other dog breeds.

This strategy necessarily limits potential participants in such a system to experts and limits the selection of targets to those with explicit labels. An alternative approach is to add an additional step before identification, category learning, wherein the human learns through some method to recognize the target category. Previous research has been done to incorporate nonexperts using a questioning procedure for cat-

Manuscript received March 23, 2017; revised July 31, 2017, October 12, 2017, and January 9, 2018; accepted March 24, 2018. Date of publication June 6, 2018; date of current version July 13, 2018. This work was supported in part by the U.S. Office of the Secretary of Defense's Autonomy Research Pilot Initiative under MIPR DWAM31168 and in part by the U.S. Army Research Laboratory under CAST 076910227001 and under ARL-H70-HR52. This paper was recommended by Associate Editor Huiyu Zhou. *(Corresponding author: David T. Slayback.)*

D. T. Slayback, B. J. Lance, and J. R. Brooks are with the U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005 USA (e-mail: david.t.slayback.civ@mail.mil; benjamin.t.files.civ@mail.mil; justin.r.brooks.civ@mail.mil).

B. T. Files is with the U.S. Army Research Laboratory, Los Angeles, CA 90001 USA (e-mail: brent.j.lance.civ@mail.mil).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes two tables detailing the full results of the statistical analysis performed on the data from the paper study. This material is 15 KB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2830649

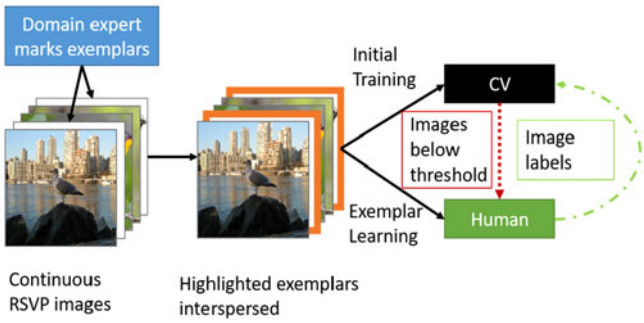

Fig. 1. Illustration of a real-time hybrid CV system.

egory learning, but for a real-time high-throughput hybrid system, this is too slow [17], [18].

Furthermore, relying on explicit category labeling requires an interruption to communicate those new instructions. In fact, many previous approaches to human-in-the-loop classification involve significant interruptions that incur additional time to process and result in reduced efficiency. Some approaches, mentioned earlier, use a guided questioning procedure [17], [18]. Others separate human and CV inputs into sequential pipelines, either using CV to highlight potential points of interest for later human labeling [19] or iterating subsets between humans and CVs to prime each for labeling [15]. Alternatively, we propose that target category learning and identification could be integrated within the presentation modality itself to prevent such disruptions of the system. To this end, we introduce a novel methodology to teach participants a category of target images, which is the focus of our experiments.

Our proposed RSVP-based system should accurately label images with a known target category, but also adapt to new target categories in near real time. In this system, each human is seated in front of a computer that displays a series of images in rapid succession, allowing the human to give faster input and thus allowing for improved efficiency over previous human–CV hybrid systems [20]. This is a similar concept to the cortically coupled computing seen in [21]. During the stream, the human labels the images both behaviorally (via button press) and via near-real-time neural classification of EEG time series data. CV modules simultaneously label the images, and the ultimate output of the system is a combined label for each image. The system adapts to new target categories by presenting exemplar images to both the CV modules and the human. It does not provide the target category to the human explicitly. There are several reasons for this. First, it allows for rapid adaptation of the overall system to new information without having to take the system offline (i.e., the time it takes to read a description of the target). Second, it allows nonexpert humans to learn what constitutes a target category while using the system; pictures are likely to convey more of the new category features than words [22] and permit nonexperts to utilize the system. To produce these pictures, an external-domain expert chooses new target categories (such as improvised explosive device), labels a small set of exemplars, and provides these to both the CVs and humans in the system. This small set of data is not enough for the CV modules to confidently label new images, so the humans determine whether images below a confidence threshold belong to the new category or not, thus providing additional training data to the CVs and decreasing their reliance on human input over time without needing expert input beyond the original exemplars (see Fig. 1).

In relation to the category learning component, there are a few models of category learning that consider different mechanisms by which the human learns. For example, the generalized context model (GCM) is an exemplar-based model positing that humans determine an image's category by examining its similarity to example images of that category (exemplars) from memory [23]. This contrasts with the prototype model in which humans draw similarity to some abstract summary representation of the category instead [23]. In both cases, each new image is considered as a point on some multidimensional feature space.

A similar category learning model, COmpetition between verbal and implicit systems [24] posits that humans learn categories implicitly or explicitly. In implicit learning, humans learn the target category by unconsciously recognizing similarities among the many exemplars they have encountered [24]. An example of this type of learning is discussed in [25], where dot patterns are presented to individuals, without explicit category knowledge, who are then asked to categorize subsequent dot patterns in a test sequence. By contrast, explicit learning is determined by conscious formation of hypothesized rules to define the category that can be verbalized or objectively measured [24]. An example would be highlighting images that are "dogs" and not those that are "cats," overtly defining categories of interest whose membership can be tested with hypotheses and explicit rules [26].

In this paper, the methodology we consider alters the common RSVP paradigm to create a real-time hybrid labeling system in which humans would be required to learn categories of real-world images "on the fly" so as to maintain high throughput and near real-time analysis, similar to previously proposed systems [12]. The value of this learning could be demonstrated in the detection of improvised explosive devices (IEDs) in images gathered from local camera feeds. Specific features of IEDs are ambiguous and difficult to express, and the pool of available experts familiar with them is small. CV agents would require a large amount of training data to correctly classify images of IEDs, but as we will demonstrate in this paper, nonexpert human labelers can learn the target category from as few as seven exemplars. Thus, instead of requiring extended offline training during a period of potentially critical risk, this system's few experts can quickly label enough images for a larger pool of nonexpert humans to learn the category and act as both a real-time labeler and force multiplier for training the CVs in real time. To investigate this paradigm, we modify the RSVP stream to present exemplar images or explicit definitions of the target category to induce category learning. Subjects are able to form hypotheses based on exemplars and test them against other exemplars and distractors, as they would in the full system. In this regard, our work more closely aligns with category learning models that utilize exemplar based, explicit learning.

Building upon our previous work, we performed two experiments in this study. The first examined different target indicator modalities (TIMs) with which the system might present images to human labelers to learn the target category [27]. TIMs are modifications to the exemplar images that contain the target of interest and are displayed to the human instead of explicitly communicating the target category. One method seen often in the previous literature is to provide bounding boxes to highlight relevant portions of the images [19]. Indeed, the dataset we used provided such annotations for a large subset. However, given that such a modality would incur significant additional cost to the domain expert in real-world systems, we compared simpler modalities (e.g., adding a border to an image) with more costly choices (e.g., bounding boxes and pauses). We also consider whether distinctiveness between categories (as a surrogate for difficulty) interacts with TIM to affect categorization accuracy. In the second experiment, we consider the case where exemplar images are mislabeled (e.g., target category is cats but images with dogs are labeled) as may happen in a real-world system. We use the TIM that performed best in a pilot experiment performed

with the Experiment 1 paradigm, vary the accuracy of exemplar and distractor labels, and examine the participant's accuracy.

In both experiments, we collected additional variables that can be used to further refine human categorization accuracy. We determine whether subjective perceptions of difficulty and response time (RT) correlate with performance measures. RT in particular may be a potential indicator to the CV as to the reliability of the human agent [28]. We investigated subjective measures because they may be a valuable indicator in this domain, and it is important to understand how they correspond to the actual performance. Past research on subjective measures has shown varying levels of correlation to performance, demonstrating unreliability in subjects' introspection that must be addressed if such measures are to be used in a hybrid system [29]. Finally, since we expect the end system to have to deal with the real-world noisy image data, we use ImageNet to provide heterogeneous images of varying display sizes, resolutions, and target saliency [30].

Results of the first experiment show most of the TIMs result in performance increases above baseline that are similar to those in the positive control. Results from the second experiment show performance decreases with increasing image similarity and higher levels of exemplar inaccuracy, but distractor inaccuracy has no significant effect, suggesting future systems should focus on accurately labeling exemplars. In addition, we see a significant correlation between accuracy and the subjective measures and RT, potentially providing another input with which the system can predict human performance. These results have implications for the design of future systems that are focused on combining humans and computer systems for accurate image labeling.

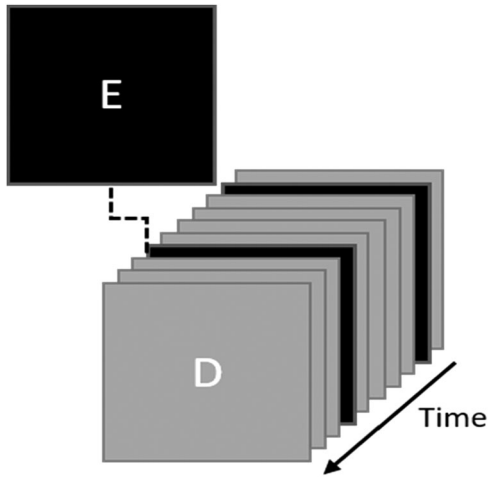

Fig. 2. Typical RSVP paradigm. D are distractors, E are exemplars of the target category.

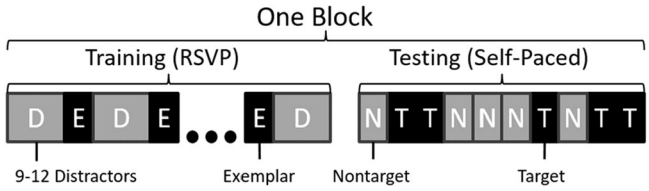

Fig. 3. Layout of an experimental block.

II. Methods and Materials

A. Participants

Total 13 subjects (one female) with ages ranging from 27 to 49 served as participants. The voluntary informed consent of the participants was obtained in accordance with U.S. Department of Defense human-use regulations observed by the Army Research Laboratory's Institutional Review Board (i.e., 32 CFR 219 and DoDI 3216.02).

B. System

Participants were seated at a desk with a standard desktop computer and keyboard in a sound-attenuated chamber. Stimuli were displayed in the center of a 23-inch monitor with 1080p resolution. The task was generated using Psychtoolbox, a commonly used MATLAB toolbox [31] for precision-timed psychophysics experiments [32]. Stimulus timing was verified using a photodiode. The images used for stimuli were drawn from ImageNet [30]. Only images that had a bounding box provided by ImageNet were used. All images with less than 200 pixels in either dimension were scaled evenly such that their smallest dimension was 200 pixels, and all images larger than 1080p were scaled to ¾ of their original size.

C. Task

1) Overall Design: Our task focused on the dynamics of target category learning and identification in an RSVP-based image labeling paradigm (see Fig. 2).

The task was divided into two experiments: The first examined the effects of different TIMs, while the second manipulated the accuracy of exemplar labeling to test the effect of labeling on the category learning. Each experiment was further subdivided into blocks, one for each unique combination of variables, each with a unique target category. Blocks were randomly assigned and sequenced for each participant in both experiments to mitigate any effect of blocks.

2) Block Design: Each block consisted of a training session and a testing session (see Fig. 3). Our primary experimental manipulations occurred in the training session, while the testing session, a category identification task, determined how effective these manipulations were to induce category learning. During the training session, participants saw a number of exemplar images from a target category interleaved with distractor images that were not members of the category. Within each training session the participant was presented an RSVP stream at 3 Hz. Each stream consisted of seven exemplar images interspersed with distractor images (see Fig. 3). Between exemplar images, there were 9 to 12 distractor images displayed. The specific number of distractors displayed between each exemplar was varied so subjects could not predict when exemplars would appear. Because each training session began and ended with distractor images, there were total eight sets of distractors used in each training session. Images containing the target category were distinguished from distractors via the TIM. In the testing session, subjects were serially presented ten images (self-paced) and asked to press the letter "q" for images from the target category and "p" for nontarget images. Each testing session contained five target images and five nontarget images in random order (see Fig. 3). In addition, after the testing session, subjects were asked to answer three survey questions pertaining to the perceived difficulty of the block: *How difficult was the last block (1 to 10)? How confident were you in what the target category was (1 to 10)? How well could you distinguish the target category from the other categories (1 to 10)?* Respectively, they will be referred to as subjective difficulty (S_{Diff}), subjective distinctiveness (S_{Dist}), and subjective confidence (S_{Conf}).

3) Experiment 1: In the first experiment, the presentation of images in the training session was manipulated according to 2 variables:

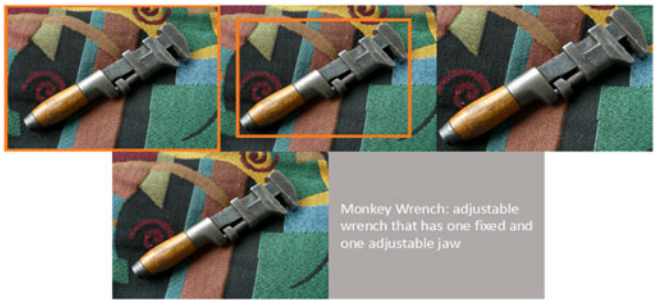

Fig. 4. BI, BT, and CO modalities, left to right in top row. Implicit and EB conditions in bottom row, left to right.

Fig. 5. Examples of task with different levels of similarity. Exemplar images are highlighted in orange. Actual number of distractor image varies between blocks.

TIM and image similarity. There were six forms of TIM and three levels of similarity, for a total of 18 blocks. Exemplar images containing the target category were modified before being displayed to the screen. The first five TIMs used were as follows:

1) bordered image (BI), where the image containing the target category was bordered with an orange (255, 102, 0 RGB) surround (see Fig. 4, left);
2) bordered target (BT), where the actual target within the image was surrounded with an orange border (see Fig. 4, center);
3) cutout (CO), where the image was cropped tightly around the target and centered (see Fig. 4, right);
4) paused, where the image was unchanged, but displayed for twice the duration of other images; and
5) implicit baseline (IB), where there were no distinguishing characteristics.

The sixth TIM was the explicit baseline (EB) condition: Instead of being presented with an RSVP stream to learn the target categories, subjects were told explicitly (via text displayed on the screen) which image category they would look for and then given the same test as in other blocks. The IB served as a negative control to ensure there was no structure in our experiment that may induce identification of a particular category, and the explicit condition served as a positive control to compare the current standard practice.

Image categories for the three levels—Easy, Medium, and Hard—of similarity were selected by a combination of depth within the ImageNet hierarchy of their common parent categories and manual selection such that subjective difficulty was the same within each level. This was necessary because the available ImageNet hierarchy has redundant links and is not completely connected. Different categories within the ImageNet hierarchy could be reached at multiple levels, meaning that for a purely depth-based selection, most categories were ambiguous. Similarity, in particular, refers to the difficulty of distinguishing the target category from the other categories. Fig. 5 shows representative exemplar and distractor images. At the low similarity level, target categories were chosen that were completely unrelated to distractor categories in the hierarchy. The Easy similarity level featured nine distinct categories, three of which were "maple," "trucks," and "domestic cats" (e.g., Fig. 5, first row, target category is "maple"). The Medium similarity level featured types of nuts, wheels (e.g., Fig. 5, second row, target category is "car wheel"), and sports balls. The highest similarity category contained species of spiders, snakes (e.g., Fig. 5, third row, target category is "hognose snake"), and old world primates. Exemplar images for a given block were sampled from images within that level of similarity (e.g., vertebrate: mammals) while distractors were selected from parallel nodes (e.g., vertebrate: amphibians, reptiles, and birds).

4) Experiment 2: In the second experiment, the presentation was manipulated according to three variables: exemplar label accuracy (percentage of labeled images that actually belonged to the target category), distractor label accuracy (percentage of nonlabeled images that actually did not belong to the target category), and image similarity. When exemplar label accuracy was not 100%, some of the images in the RSVP stream that were labeled as exemplars with a TIM were not actually members of the target category. When the distractor accuracy was not 100%, some of the images not labeled with a TIM were actually members of the target category. There were three levels of exemplar and distractor accuracy: 50%, 75%, and 100% accuracy. Image similarity retained the same three levels from Experiment 1, resulting in 27 total blocks for each participant. All blocks used the BT modality from the previous experiment, which had shown the best (though not statistically significant) results in pilot testing of the first experiment.

D. Dependent Measures and Statistical Analysis

The primary dependent variable measured was accuracy. Accuracy was determined as the number of hits plus correct rejections divided by the number of trials. RT was determined for each trial as the difference in time between presentation of the image and the button response. Subjective measures were gathered at the end of each block by having the participant click on a bar displayed on the screen. Depending on where they clicked, their response was scored on a continuous scale of 1–10. A generalized, linear mixed-effect modeling approach was used for both experiments. In each model a logistic regression was used to relate accuracy to the experimental manipulations, the responses to the subjective questions, RT, and all pairwise interactions between these variables. In the first experiment, accuracy was related to TIM, RT, target similarity, S_{Diff}, S_{Dist}, S_{Conf}, and all pairwise interactions

$$\begin{aligned}\text{Accuracy} \sim 1 + \text{similarity}^{*}\text{TIM}^{*}S_{\mathrm{Diff}}{}^{*}S_{\mathrm{Dist}}{}^{*} \\ S_{\mathrm{Conf}}{}^{*}\text{RT} + (1|\text{subject})\,. \end{aligned} \tag{1}$$

In the second experiment, the same procedure was implemented; however, the predictor variables were target similarity, S_{Diff}, S_{Dist}, S_{Conf}, distractor accuracy, exemplar accuracy, and all pairwise interactions

$$\begin{aligned}\text{Accuracy} \sim 1 + \text{similarity}^{*}\text{ExemplarAcc}^{*}\text{DistAcc}^{*} \\ S_{\mathrm{Diff}}{}^{*}S_{\mathrm{Dist}}{}^{*}S_{\mathrm{Conf}}{}^{*}\text{RT} + (1|\text{subject})\,. \end{aligned} \tag{2}$$

An iterative, backward model selection procedure was performed wherein terms in the model were sorted by p-value. The term with the largest p-value above 0.05 was then eliminated from the model. The model was refit without this term until all terms in the final model were

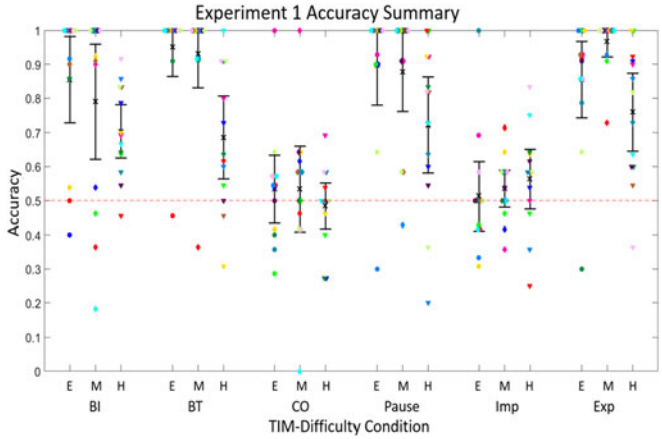

Fig. 6. Summary of Experiment 1's results with 95% confidence intervals (CIs). All CIs calculated using corrected Cousineau method [34]. Different colors are different subjects. TIMs are BI, BT, CO, Imp, and Exp. E, M, and H are easy, medium, and hard difficulties, respectively.

significant (i.e., < 0.05) [33]. For ease of interpretation, odds ratios (ORs) were calculated separately for the significant continuous and categorical predictors from the models using the formula in (3). In the continuous case, OR refers to a change in the odds of accuracy relative to a unit increase of the variable being studied. In the categorical case, OR refers to a relative increase in the odds of an accurate response from a baseline condition

$$\mathrm{OR} = e^{\mathrm{coefficient}}. \tag{3}$$

III. Results

A. Experiment 1

1) General Results: The average accuracy across all subjects and experimental conditions was (mean $\pm$ std) 0.73 ± 0.24 with an average RT of 1.26 s $\pm$ 0.73 s. As is shown in Fig. 6 there is a wide range of accuracies related to the TIM and the difficulty. The average responses to surveys were $S_{\mathrm{Diff}} = 6.03 \pm 3.24$ (5.13 Easy difficulty, 5.11 Medium, and 7.86 Hard); $S_{\mathrm{Dist}} = 5.26 \pm 3.47$ (6.13 Easy, 6.02 Medium, and 3.63 Hard); and $S_{\mathrm{Conf}} = 5.39 \pm 3.17$ (6.45 Easy, 6.45 Medium, and 3.28 Hard).

2) Modeling: As stated in Section II, a generalized logistic regression was performed that related accuracy of target categorization to the experimental manipulations (TIM and Difficulty) and the responses of the participant (RT, S_{Diff}, S_{Dist}, and S_{Conf}) and all pairwise interactions. After the backward selection procedure, the most parsimonious model contained just the TIM term; an analysis of variance (ANOVA) performed on the model yielded $F(5, 228) = 58.94, p =< 0.0001$, and $\eta^2 = 0.356$. Fig. 7 shows the OR for each of the TIMs relative to the implicit condition. Each increase of 1 in OR corresponds to a 100% increase in odds of an accurate response as compared to overall subject accuracy.

For reference, the accuracy with implicit TIM was equal to near chance 0.53 ± 0.14. The BT, Pause, Explicit (Exp), and BI modalities all show significant improvement over the Implicit (Imp) and CO conditions. BT, Exp, BI, and Pause are not significantly different from each other as indicated by overlapping 95% confidence intervals.

The coefficients and associated statistics for the full model are available as supplementary information (S1).

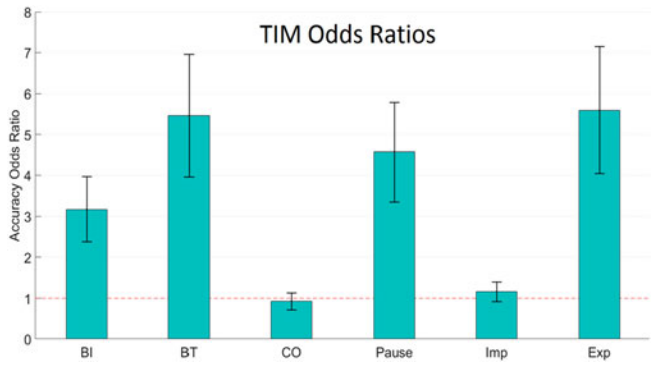

Fig. 7. OR from the reduced model. BI = 3.17, BT = 5.46, CO = 0.92, pause = 4.58, Imp = 1.17, and Exp = 5.59. The y-axis is the multiplicative improvement over chance. Red line represents baseline accuracy. Error bars represent 95% CIs.

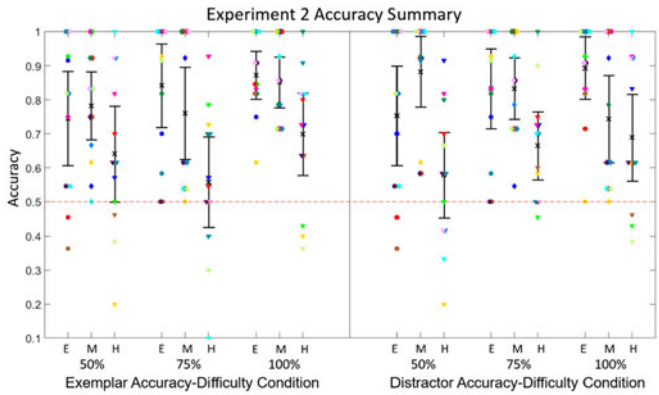

Fig. 8. Summary of Experiment 2 results with 95% CIs for easy, medium, and hard difficulties.

B. Experiment 2

1) General Results: The average accuracy across all subjects and experimental conditions was 0.77 ± 0.21 with an average RT of 1.01 s $\pm$ 0.38 s. In Fig. 8, we show the data for all participants and demonstrate the range of behavior elicited by the target and distractor accuracy conditions for varying levels of category difficulty. The average responses to surveys were $S_{\mathrm{Diff}} = 6.72 \pm 2.63$ (6.15 Easy, 5.89 Medium, and 8.12 Hard); $S_{\mathrm{Dist}} = 4.83 \pm 2.85$ (5.58 Easy, 5.56 Medium, and 3.37 Hard); $S_{\mathrm{Conf}} = 5.03 \pm 2.77$ (5.99 Easy, 6.20 Medium, and 2.89 Hard).

2) Modeling: A generalized logistic regression was performed that related accuracy of target categorization to the experimental manipulations (Exemplar Accuracy, Distractor Accuracy, and Difficulty) and the responses of the participant (RT, S_{Diff}, S_{Dist}, and S_{Conf}) and all pairwise interactions.

After the model selection procedure, the final terms were Difficulty, Exemplar Accuracy, RT, S_{Diff}, S_{Dist}, Difficulty: S_{Diff}, and RT: S_{Diff}. Since this model contained a mixture of categorical predictors (Difficulty) and continuous predictors (Exemplar Accuracy, Distractor Accuracy, RT, S_{Diff}, S_{Dist}, and S_{Conf}) the actual coefficient values are plotted separately for ease of explanation. Reporting for variables is done as ORs. In Fig. 9, the ORs from the reduced model are plotted for the categorical values. The ANOVA for the reduced model for Difficulty showed $F(2, 341) = 29.267, p < 0.001$ and for Difficulty: S_{Diff} $\mathrm{F}(2, 341) = 10.77, p < 0.001$. Medium difficulty images were not significantly different from Easy; however, Hard images were as-

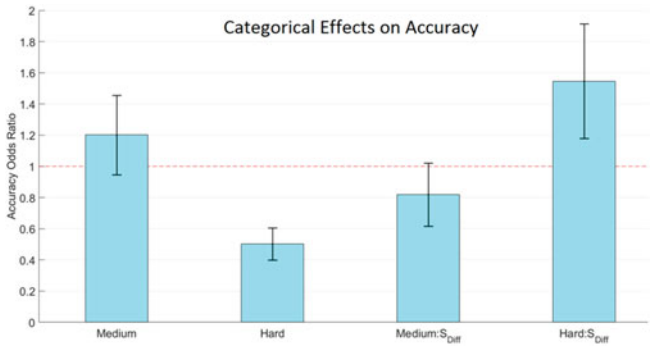

Fig. 9. Categorical OR relative to easy condition for Experiment 2 and interaction terms with 95% CIs. OR, respectively: 1.20, 0.50, 0.82, and 1.55. An OR of 1 is baseline odds of an accurate response. Red line corresponds to baseline accuracy. Nonsignificant results are shown in supplemental material.

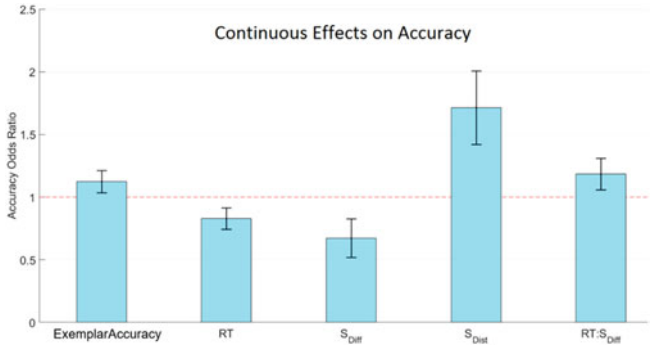

Fig. 10. Continuous OR for Experiment 2 and interaction terms with 95% CIs. OR, respectively: 1.12, 0.83, 0.67, 1.71, and 1.19. Red line corresponds to equivalent accuracy. Nonsignificant results are shown in supplemental material.

sociated with a nearly 50% reduction in performance accuracy across all other conditions. The interaction between Medium difficulty and the S_{Diff} was not significantly different from the interaction between Easy difficulty and S_{Diff}. However, interestingly, the interaction between Hard difficulty level and S_{Diff} showed a positive improvement over Easy: S_{Diff} as being associated with a near 50% increase in performance accuracy.

The significant, continuous predictors (Exemplar Accuracy, RT, S_{Diff}, S_{Dist}, and RT: S_{Diff}) are plotted in Fig. 10. The ANOVA for the reduced model for exemplar accuracy showed $F(1, 341) = 7.66, p = 0.006$; RT $F(1, 341) = 11.48, p = 0.0008$; S_{Diff} $F(1, 341) = 9.06, p = 0.003$; S_{Dist} $F(1, 341) = 32.01, p < 0.001$; and RT : S_{Diff} $F(1, 341) = 9.04, p = 0.003$. Exemplar accuracy, subjective distinctiveness, and the interaction between RT and subjective difficulty were associated with an increased likelihood of accurate categorization. RT and subjective difficulty were both related to a decreased probability of accurate categorization.

IV. Discussion

This study was motivated by the concept of a real-time human–CV hybrid system for image labeling. The idea of a hybrid vision system for tasks that cannot be performed by CV alone is not new. Tohme is a system that combines CV and crowd-sourced human input to identify pedestrian ramps at road intersections [19], and other systems have used communication between CVs and humans to distinguish fine-grained or semantic categories [17], [18], [35]. An important limitation of these methods is that they cannot be adapted to new targets in real time because their human-input paradigms are fundamentally detached from the actual labeling task or separated into batch pipelines. Our goal is to incorporate these human inputs on the same timescale as CV labeling by leveraging an RSVP paradigm. Critically, we also want our system to maintain the ability to make semantic distinctions and to use primarily nonexpert input. To this end, we needed to refine methods by which the system could teach its human agents new categories such that they could later identify them.

In the first experiment, we examined which TIM produced the best results while controlling for the difficulty of the categorization as well as which subjective measures might also predict performance. Following the GCM framework for human categorization, the purpose of our TIMs was essentially to increase the resolution of the various features the subject would use to compare new images to stored exemplars. Given that our task likely provided too few examples for implicit learning, subjects would develop explicit rules of comparison based on those features [24]. From previous research on visual image processing, it seemed likely that different modalities might increase the salience of relevant features within those exemplars, improving recall of those features in exemplar-based classification and, thus, leading to superior performance for particular TIMs [36], [37].

The two TIMs with borders (BI and BT) both demonstrated high accuracy but no significant differences from each other. The lack of differences between the two border conditions is unsurprising, as visual search literature has shown that subjects often do not take full advantage of selective highlighting methods such as BT despite the potential advantages [38]. This suggests that future systems may be able to do away with the data requirements the BT condition imposes. The lack of difference between these modalities and the EB, however, is potentially a key insight for our proposed system. Explicitly telling a participant what the target category is can be effectively approximated in our experiment by simply displaying labeled images containing the target category. The lack of correlation of accuracy with difficulty is counter-intuitive, as the GCM model would suggest that more similar nontarget images would be more difficult to distinguish [23]. We think this is primarily due to the large effect the TIMs had above the implicit and CO baselines.

Both the CO and the implicit TIMs were associated with near-chance probabilities of accuracy, demonstrating the substantial improvement observed with the other modalities was much larger than the effect of the other predictors in this experiment. The lack of difference between performance with CO and the IB could be due to the elimination of contextual information in the removed portions of the images, which has been shown to be vital to guiding attention to areas or images of interest [39], [40]. Given that the task images were of various sizes, as were the bounding boxes, it is also possible that cropping the images did not make exemplar images sufficiently distinct from distractor images.

In the second experiment, we examined how modulating the exemplar and distractor accuracy affected performance across different levels of difficulty. The results showed that participants tended to focus more on learning the target category rather than learning what was not a member. Our initial hypothesis, based on an explicit learning model, had been that mixing exemplar images with distractors would make it more difficult for humans to learn a consistent set of rules to identify the target [24]. Based on these results, it is acceptable to let some degree of target images (up to 50%) be mislabeled as distractors; however, it is not advisable to present nontarget images labeled as exemplars.

The significant effect of the Hard difficulty condition on accuracy indicates changes to TIM may have overshadowed it in Experiment 1. Previous research on image similarity among stimuli corroborates this effect [41], [42], and it fit with GCM classification as discussed

previously [23]. This supports the effectiveness of our creation of a difficulty hierarchy from the ImageNet structure that may be of benefit for other experiments. The lack of significant difference between Easy and Medium conditions, though, suggests future refinement and testing is necessary.

RT and the rating of block difficulty were both negatively related to performance. Prior research on psychomotor vigilance tasks has correlated RTs with confused or inhibited information processing [43]. However, it did appear that participants subjectively determined their accuracy with respect to block difficulty and in evaluation of their confidence.

In contrast, when participants rated a difficult block as difficult, there was an increased probability of being accurate. This indicates that participants who rated those blocks as easier may have been confident but learned the wrong target category or used an incorrect abstract distinction, and thus had worse performance. However, these results may also be related to a previously observed effect wherein participants perform better than their expectation [44].

V. Conclusion

These findings, though preliminary, have direct bearing on how information is presented to humans in a combined human–CV image-labeling system. We acknowledge our sample size is small and therefore requires replication to validate. We further suggest future research should likely focus on the effects of confidence and subjective task difficulty on learning, especially how those factors relate to more difficult tasks such as categorizing semantically categorized image databases (e.g., Places2 [45]).

In addition to the use for IED detection mentioned earlier, we envision this research work expanding crowdsourcing applications for real-time data, such as security and traffic camera monitoring, aerial image classification, and collaborative mapping with robots. Many of these applications are task-specific and/or require a certain level of expertise from their labelers; our research might allow these systems to leverage nonexpert labelers and adapt to changes in task demands. For this research work, it will also be vital to determine how to detect and reject misclassifications from these nonexpert labelers, studying how both rate and cause may change during operation, and identifying methods by which the system might adapt. This could include eliciting further feedback from domain experts in an iterative process.

It is important to be cognizant of how information is presented to the human in the future and how confident humans feel in their judgments with that information—not just which particular CV algorithm to use or what percentage of human labels to discount. The human factors in learning and categorization must be treated with the same importance as the CV parameters in training and processing. In the same way that an active-learning algorithm carefully chooses data points that will provide the most information when relabeled, we must carefully consider our inputs to the human side to enhance their understandability and, thus, maximize information they can provide.

References

[1] Y. Nagar, "Combining human and machine intelligence for making predictions," M.S. Thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2013.

[2] D. Cook, "A human–computer team experiment for 9x9 Go," in *Proc. Comput. Games Conf.*, 2010, pp. 145–155.

[3] E. de Visser and R. Parasuraman, "Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload," *J. Cognitive Eng. Decis. Making*, vol. 5, no. 2, pp. 209–231, Jun. 2011.

[4] J. Y. C. Chen and M. J. Barnes, "Human-agent teaming for multirobot control: A review of human factors issues," *IEEE Trans. Human–Mach. Syst.*, vol. 44, no. 1, pp. 13–29, Feb. 2014.

[5] J. M. Wolfe, T. S. Horowitz, M. J. Van Wert, N. M. Kenner, S. S. Place, and N. Kibbi, "Low target prevalence is a stubborn source of errors in visual search tasks," *J. Exp. Psychol., Gen.*, vol. 136, no. 4, pp. 623–628, Nov. 2007.

[6] H. A. Sholl, "Modeling of an operator's performance in a short-term visual information processing task," *IEEE Trans. Syst., Man, Cybern.*, vol. 2, no. 3, pp. 352–362, Jul. 1972.

[7] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[8] H. Lee, H. Kwon, R. M. Robinson, D. Donavanik, W. D. Nothwang, and A. R. Marathe, "Task-conversions for integrating human and machine perception in a unified task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Daejeon, South Korea, 2016, pp. 2751–2758.

[9] R. M. Robinson *et al.*, "Human-autonomy sensor fusion for rapid object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Hamburg, Germany, 2015, pp. 205–312.

[10] A. W. Bohannon, N. R. Waytowich, V. J. Lawhern, B. M. Sadler, and B. J. Lance, "Collaborative image triage with humans and computer vision," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Budapest, Hungary, 2016, pp. 4046–4051.

[11] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Aug. 2008.

[12] P. Sajda *et al.*, "In a blink of an eye and a switch of a transistor: Cortically couple computer vision," *Proc. IEEE*, vol. 98, no. 3, pp. 462–478, Mar. 2010.

[13] B. Settles, "Active learning literature survey," Univ. Wisconsin-Madison, Madison, WI, USA, Rep. 1648, 2009.

[14] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," Jun. 2015. arXiv: 1506.03365.

[15] E. Lughofer *et al.*, "Human–machine interaction issues in quality control based on online image classification," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 39, no. 5, pp. 960–971, Sep. 2009.

[16] A. R. Marathe, B. J. Lance, K. McDowell, W. D. Nothwang, and J. S. Metcalfe, "Confidence metrics improve human-autonomy integration," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, New York, NY, USA, 2014, pp. 240–241.

[17] S. Branson *et al.*, "Visual recognition with humans in the loop," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 438–451.

[18] C. Wah, S. Branson, P. Perona, and S. Belongie, "Multiclass recognition and part localization with humans in the loop," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 2524–2531.

[19] K. Hara, J. Sun, R. Moore, D. Jacobs, and J. Froehlich, "Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning," in *Proc. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA, 2014, pp. 189–204.

[20] R. Spence and M. Witkowski, "What is RSVP? And why do I need it?" in *Rapid Serial Visual Presentation*, London, U.K.: Springer, 2013, pp. 1–18.

[21] S. Saproo *et al.*, "Cortically coupled computing: A new paradigm for synergistic human-machine interaction," *Computer*, vol. 49, no. 9, pp. 60–68, Sep. 2016.

[22] C. Plaue, T, Miller, and J. Stasko, "Is a picture worth a thousand words?: An evaluation of information awareness displays," in *Proc. Graph. Interface*, 2004, pp. 117–126.

[23] R. M. Nosofsky, "The generalized context model: An exemplar model of classification," in *Formal Approaches to Categorization*, E. M. Pothos and A. J. Willis, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 18–39.

[24] C. L. Huang-Pollock, W. T. Maddox, and S. L. Karalunas, "Development of implicit and explicit category learning," *J. Exp. Child Psychol.*, vol. 109, no. 3, pp. 321–335, Jul. 2011.

[25] F. G. Ashby and W. T. Maddox, "Human category learning," *Annu. Rev. Psychol.*, vol. 56, no. 1, pp. 149–178, Feb. 2005.

[26] E. E. Smith, "The case for implicit category learning," *Cognitive, Affective, Behavioral Neurosci.*, vol. 8, no. 1, pp. 3–16, Mar. 2008.

[27] J. Brooks, D. Slayback, B. Shih, A. Marathe, V. Lawhern, and B. J. Lance, "Target class induction through image feedback manipulation in rapid serial visual presentation experiments," in *Proc. IEEE Int. Conf. Systems, Man, Cybern.*, Hong Kong, 2015, pp. 1047–1052.

[28] B. A. J. Reddi, K. N. Asrress, and R. H. S. Carpenter, "Accuracy, information, and response time in a saccadic decision task," *J. Neurophysiol.*, vol. 90, no. 5, pp. 3538–3546, Nov. 2003.

[29] K. Desender, F. V. Opstal, and E. V. den Bussche, "Subjective experience of difficulty depends on multiple cues," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 44222.

[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[31] Mathworks Inc., "MATLAB version R2014b," Mathworks Inc., Natick, MA, USA, 2014.

[32] D. H. Brainard, "The psychophysics toolbox," *Spatial Vis.*, vol. 10, no. 4, pp. 433–436, 1997.

[33] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc., Series B (Stat. Methodol.)*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[34] R. D. Morey, "Confidence intervals from normalized data: a correction to Cousineau (2005)," *Tut. Quantitative Methods Psychol.*, vol. 4, no. 2, pp. 61–64, 2008.

[35] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," presented at the *Proceeding IEEE Conf. Computer Vision. Pattern Recognition*, Seattle, WA, USA, 2016, pp. 1153–1162.

[36] M. R. Blair, M. R. Watson, R. C. Walshe, and F. Maj, "Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization," *J. Exp. Psych., Learn., Memory, Cognit.*, vol. 35, no. 5, pp. 1196–1206, Sep. 2009.

[37] B. Rehder and A. B. Hoffman, "Eyetracking and selective attention in category learning," *Cognitive Psychol.*, vol. 51, no. 1, pp. 1–41, Mar. 2005.

[38] F. P. Tamborello and M. D. Byrne, "Adaptive but non-optimal visual search behavior with highlighted displays," *Cognitive Syst. Res.*, vol. 8, no. 3, pp. 182–191, Sep. 2007.

[39] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cognitive Sci.*, vol. 11, no. 12, pp. 520–527, Dec. 2007.

[40] J. M. Cathcart, T. J. Doll, and D. E. Schmieder, "Target detection in urban clutter," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 5, pp. 1242–1250, Sep./Oct. 1989.

[41] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychol. Rev.*, vol. 96, no. 3, pp. 433–458, Jul. 1989.

[42] K. Grill-Spector and N. Kanwisher, "Visual recognition: As soon as you know it is there, you know what it is," *Psychol. Sci.*, vol. 16, no. 2, pp. 152–160, Feb. 2005.

[43] M. Basner and D. F. Dinges, "Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss," *Sleep*, vol. 34, no. 5, pp. 581–591, 2011.

[44] Y. H. Kim, C. Y. Chiu, and Z. Zou, "Know thyself: Misperceptions of actual performance undermine achievement motivation, future performance, and subjective well-being," *J. Personality Social Psychol.*, vol. 99, no. 3, pp. 395–409, Sep. 2010.

[45] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.