

Collaborative Image Triage with Humans and Computer Vision

Addison W. Bohannon^{*†}, Nicholas R. Waytowich^{†‡}, Vernon J. Lawhern[†], Brian M. Sadler[§] and Brent J. Lance[‡]

^{*}Applied Mathematics, Statistics, and Scientific Computing

University of Maryland, College Park, MD 20742

Email: addisonb@math.umd.edu

[†]Department of Biomedical Engineering

Columbia University, New York, NY 10027

[‡]US Army Research Laboratory

Aberdeen, MD 21005

[§]US Army Research Laboratory

Adelphi, MD 20783

Abstract—As the technology for acquiring and storing images becomes more prevalent, we are faced with a growing need to sort and label these images. At this time, computer vision algorithms cannot parse abstract concepts from images like a human. As a result, there may be performance gains possible from the integration of human analysts with computer vision agents. We present an image triage system which facilitates the collaboration of heterogeneous agents through a novel unsupervised meta-learning technique. The system iteratively allocates images for binary classification among heterogeneous agents according to the Generalized Assignment Problem (GAP) and combines the classification results using the Spectral Meta-Learner (SML). In simulation, we demonstrate that the proposed system achieves significant speed-up over a naive parallel assignment strategy without sacrificing accuracy.

I. INTRODUCTION

While the Large Scale Visual Recognition Challenge has shown that computer vision algorithms can label large databases of images of objects as well as a human [1], mastery of complex concepts such as scene recognition [2], the identification of "danger," or the interpretation of visualized data eludes current computer vision algorithms. In contrast, visual analytics [3] and human computation applications [4] leverage the human visual system to identify patterns in data and solve problems that are difficult or intractable to current artificial intelligence technology. It is possible that with larger data sets and more training time, computer vision systems will achieve human-level performance in these tasks [5], but it will require a human to provide or verify labels. This points to the potential benefit of integrating human analysts with computer vision agents in ways that leverage the unique capabilities of each. In this paper, we introduce an image triage system which synchronizes the effort of human analysts manually labeling images, augmented human analysts that use a brain-computer interface (BCI) to more rapidly label images, and deep-learned computer vision to efficiently and accurately classify a two-class database of images, and we test this system in simulation.

U.S. Government work not protected by U.S. copyright

Recent work has shown how BCI technologies, among many other applications [6], can be used to improve the speed of collecting labels for images from human analysts [7], [8]. In one particular paradigm, Rapid-Serial Visual Presentation (RSVP), brain-signals are recorded from a person while passively viewing images at high rates of speed (2-10Hz) and images are classified according to the neural signature they elicit. Thus, RSVP augmented human analysts can achieve higher rates of throughput than traditional human analysts, but this speed-up comes at the cost of a reduction in labeling accuracy. Sajda, et al. address this trade-off in [9] through a collaborative system of a computer vision algorithm and a human performing image triage via RSVP. The authors explore a serial implementation in which either the computer vision algorithm or RSVP agent first screen the database before the other agent classifies the images.

This human-system approach is not unlike active learning approaches which team a human oracle with a computer vision system [10]–[12]. In [12], Joshi, et al. present human-machine system for multi-class image labeling in which a computer vision system iteratively clusters images and prompts a human for binary decisions.

The field of optimal crowd-sourcing takes a different approach to address the solution of problems such as image triage that can be decomposed into a large set of simple problems. Crowd-sourcing applications facilitate the large-scale collaboration of non-expert humans by distributing simple tasks to numerous low-cost human agents (i.e. Amazon Mechanical Turk) [13]–[19]. Necessary to these approaches are decisions about task allocation and meta-learning. For example, Karger, et al. use a random assignment of tasks with the assumption of agent and task homogeneity while implementing a belief propagation technique for meta-learning [15]. In [14], Ho, et al. implement an exploration phase to infer the difficulty of tasks and reliability of agents before using the Generalized Assignment Problem (GAP) to determine an optimal allocation of tasks in the subsequent exploitation phase.

As these systems can increase the pool of potential agents,

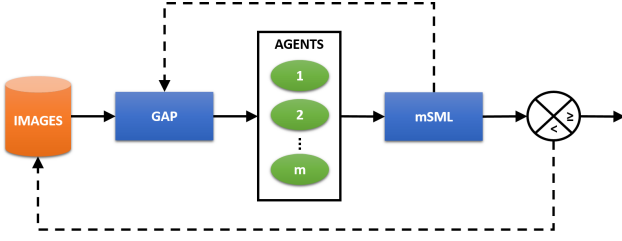


Fig. 1. The image triage system. An assignment node distributes images to agents in parallel. The agents label images, and the results are consolidated at a fusion node. At this point, the confidence, s_i , in the image classification label is used to threshold images for completion or routing back to the image database for re-assignment. Two forms of feedback occur in the system: the return of images to the database for further assignment and the inference of agent reliability, r_j and image confidence, s_i , provided to the assignment node for further assignment. Here, agents can be human analysts, computer vision, or RSVP analysts.

the cost of recruitment must be balanced with the expected performance of an agent [13]. This prioritizes the ability to infer agent performance without labeled data. Numerous approaches have addressed combining labels of noisy agents and inferring the performance of individual agents from the aggregated responses [15]–[17], of which, the work of Parisi, et al. provides an elegant computational approach to achieve both through the Spectral Meta-Learner (SML) [19].

In the proposed image triage system, we present three primary scientific contributions: a heterogeneous multi-agent system for image labeling, an extension to SML, and an unsupervised meta-learning approach. Previous approaches to image labeling have included human-machine systems (human-computer, BCI-computer) or homogenous ensembles of agents, but here we present a heterogeneous multi-agent system comprising computer vision, human, and BCI agents. Additionally, we apply SML to a full-rank sub-matrix of the incomplete data matrix required for SML. This extension of SML allows the meta-data gained from SML to inform subsequent image assignments through the Generalized Assignment Problem (GAP). By iterating image allocation and SML, the system improves both its estimate of agent capabilities and confidence in image labels over time, resulting in a novel unsupervised meta-learning approach (see Figure 1). Finally, we test our proposed image triage system in simulation and achieve human-level accuracy while finding a balance in the trade-off between time-cost and accuracy.

II. APPROACH

A. Image Assignment

Consider a system with n heterogeneous images and m heterogeneous agents. Denote $I = \{1, \dots, n\}$ as the index of images and $J = \{1, \dots, m\}$ as the index of agents. Any agent j may perform a binary decision, $\{-1, 1\}$, for any assigned image i . Each image has a confidence parameter, $s_i \geq 0$, which reflects the current confidence in the meta-label of that image, and each agent has a reliability, $0 \leq r_j \leq 1$, which reflects the system's assessment of the agent's accuracy. Additionally, each agent has a budget, b_j , which reflects the time available

to an agent on a given iteration. The assignment of an image i to an agent j will have an associated cost, $c_{ji} \geq 0$, which accounts for the time required for labeling an image, and value, v_{ji} . This image assignment problem, where we are looking for the optimal assignment policy over all images and agents, $\{x_{ji}\}_{i \in I, j \in J}$, can be mapped onto the Generalized Assignment Problem (GAP) as in [14]:

$$\mathbf{x} = \arg \max_{\mathbf{x}} \sum_{i \in I} \sum_{j \in J} v_{ji} x_{ji} \quad (1)$$

- 1) $\sum_{i \in I} c_{ji} x_{ji} \leq b_j, j \in J$
- 2) $\sum_{j \in J} x_{ji} = 1, i \in I$
- 3) $x_{ji} \in \{0, 1\}$
- 4) $v_{ji} = r_j - s_i + \max_{i^* \in I} s_{i^*}$

Besides having known solutions, GAP captures both the decision problem and the inherent trade-off between assignment value and time [20].

B. Joint Classification

Upon the receipt of image labels from the agents, we can consider the binary decision of the agents as conditionally independent discrete random variables, $A_j : \{-1, 1\} \rightarrow \mathbb{R}$, and the set of decisions from m agents for a single image, $(A_1, \dots, A_m)^t$, as a joint random variable, $\mathbf{A}^i : \{-1, 1\}^m \rightarrow \mathbb{R}$. If the true label of an image is a discrete random variable, $Y : \{-1, 1\} \rightarrow \mathbb{R}$, then we seek the decision rule, d , which maximizes $\mathbb{P}(d(\mathbf{A}^i) = y_i)$.

The obvious choice for this decision rule would be the decision which maximizes the log-likelihood [21] of the predictions of the individual classifiers:

$$d(\mathbf{a}^i) = \arg \max_{y_i \in \{-1, 1\}} \sum_{j \in J} \log \mathbb{P}_{A_j|Y}(a_j|y_i) \quad (2)$$

If we define the balanced accuracy, π_j , of an agent j as

$$\pi_j = \frac{1}{2}(\psi_j + \eta_j), \quad (3)$$

where ψ_j is sensitivity, $\mathbb{P}(a_j = 1|y_i = 1)$, and η_j is specificity, $\mathbb{P}(a_j = -1|y_i = -1)$, then, as shown in [19], the decision rule can be written in terms of the sensitivity and specificity of each agent,

$$d(\mathbf{a}^i) = \text{sign} \sum_{j=1}^m a_j (\log \alpha_j + \log \beta_j), \quad (4)$$

where $\alpha_j = \frac{\psi_j \eta_j}{(1-\psi_j)(1-\eta_j)}$ and $\beta_j = \frac{\psi_j(1-\psi_j)}{\eta_j(1-\eta_j)}$. This form of the maximum likelihood estimate invites an expectation-maximization (EM) approach to improve the decision rule [19], [21]. We implement a variant of the SML, further referred to as mSML, which accepts non-fully-populated results and uses the first order approximation of the SML for a full-rank sub-matrix to label images not in the set of images classified by all agents (Algorithm 1).

This procedure provides more than simply a decision rule for joint classification. The absolute value of the maximum

Algorithm 1 Modified Spectral Meta-Learner (mSML)

```

procedure MSML( $\{A^i\}_{i \in I}$ )
   $I^* = \{i \in I \mid |A^i| = m\}$ 
   $\mathbf{Q} = \frac{1}{|I^*|-1} \sum_{i \in I^*} (\mathbf{a}^i - \bar{\mathbf{a}})(\mathbf{a}^i - \bar{\mathbf{a}})^T$ 
   $\mathbf{v} = \{\mathbf{v}_\ell \mid \ell = \arg \max_\ell \lambda_\ell\} \text{ s.t. } \mathbf{Q}\mathbf{v}_\ell = \lambda_\ell \mathbf{v}_\ell$ 
   $k = 0$ 
  for  $i \in I^*$  do
     $s_i^k = \mathbf{v}^T \mathbf{a}^i$ 
     $d^k(\mathbf{a}^i) = \text{sign}(s_i^k)$ 
  end for
  while not converged do
    for  $j \in J$  do ▷ E-step:
       $\psi_j^{k+1} = P(a_j = 1 \mid d^k(\mathbf{a}^i) = 1)$ 
       $\eta_j^{k+1} = P(a_j = -1 \mid d^k(\mathbf{a}^i) = -1)$ 
       $r_j^k = \frac{1}{2}(\psi_j + \eta_j)$ 
    end for
    for  $i \in I^*$  do ▷ M-step:
       $s_i^{k+1} = \sum_{j=1}^m a_j (\log \alpha_j^{k+1} + \log \beta_j^{k+1})$ 
       $d^{k+1}(\mathbf{a}^i) = \text{sign}(s_i^{k+1})$ 
    end for
     $k = k + 1$ 
  end while
  for  $i \in I \setminus I^*$  do
     $s_i^k = \sum_{j \in J} r_j^k a_j$ 
     $d^k(\mathbf{a}^i) = \text{sign}(s_i^k)$ 
  end for
end procedure

```

likelihood estimate provides a measure of confidence in the image label, s_i , and the updated estimate of the agent balanced accuracy provides a measure of reliability, r_j . These values, in turn, update the assignment value, v_{ji} , in (1) for the next iteration.

C. System Implementation

Beyond providing a conceptual framework, we implemented in software a flexible and scalable system depicted in Figure 2 which achieves task parallelism among heterogeneous agents performing image labeling at distributed workstations. We envisioned a system which works with as few as three agents and scales to dozens of agents while also being agnostic to the type of agents connected.

At each iteration, the image assignment problem incorporates parameters estimated from the mSML. This feedback necessitates consideration of the stability and convergence of the system. We first define convergence as all images labeled with sufficient confidence. Ideally, this is where the system would stop, but many things could happen to prevent this; images could be continually assigned to the same reliable agent who has already seen the image, or some images may be equally hard for all agents and never achieve sufficient confidence. In practice, there is no guarantee that all images

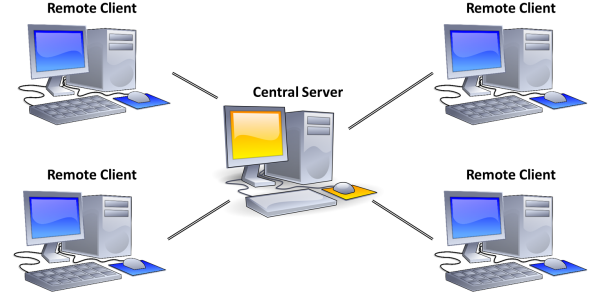


Fig. 2. Image triage system. The remote clients reside on networked workstations with read-only access to a shared image database on the central server. The central server generates assignments and communicates with the remote clients through a custom message passing interface.

will achieve a given confidence threshold even if all agents labeled the image, and the value of this threshold will likely impact the number of agents who must see an image for it to be classified. In order to address duplicate assignments, the value of an assignment which has previously been made is set to zero for all subsequent iterations. This is not a hard barrier, such as $v_{ji} = -\infty$, but it discourages duplicate assignments unless necessary to satisfy the constraints of (1). As the system evolves, and images achieve confidence threshold, there will be the same number of agents available to label fewer images; however, more of those agents will have seen some or all of the images, and the active constraints for the system will be the inequality constraints of the low-throughput agents. If we relax these constraints and increase the budget of all agents, we can facilitate system convergence. This comes at a cost of increasing the length of iteration intervals, so we introduce a maximum interval length, L_{max} , to bound the time required to achieve system convergence. The dynamic budget and maximum interval length result in problems which are pseudo-infeasible, which we define as being unable to satisfy constraint [2] of (1) without zero-value assignments. Thus, we amend our definition of system convergence to include the alternative stopping condition that the assignment problem becomes pseudo-infeasible.

In order to initiate the system, prior to any inferred knowledge of the reliability of the agents, we use a batch assignment to all m agents of $m(m+1)$ images in order to adequately estimate the sample covariance matrix, \mathbf{Q} , of Algorithm 1.

III. SIMULATIONS

All software was developed in MATLAB R2015a and later releases. Simulations ran on a Unix-based desktop computer with two Intel Xeon 2.67 GHz processors, for a total of 8 cores to support independent processes. Agent performance was provided by simulation modules, which randomly generated both image labels and pauses for their interface with the RemoteAgent objects. This simulation framework allowed the experiment to take place on a single multi-core workstation in which each agent and the central server ran on a separate dedicated instance of MATLAB.

TABLE I
PROPERTIES OF SIMULATED AGENTS

Type	Accuracy (p_j)	Cost (c_{ji})	Service Time (μ_j)
CV	0.75	1	0.01s
RSVP	0.85	1	0.1s
Human	0.95	1	1.0s

The agents generated labels randomly according to a Bernoulli distribution, $f_{A_j|Y}(a_j|y) = \text{bern}(p_j)$. The agents generated service times for each image according to an exponential distribution, $T \sim \text{exp}(\mu)$. The budget was a function of the speed of the agent, $b_j = \frac{L_k}{\mu_j}$, where L_k is the desired interval length for iteration k .

We wanted to simulate a scenario in which very little training data is available for the computer vision algorithms such that both the human and RSVP analysts possess an accuracy advantage. We encoded this expected performance into the simulation parameters, with the human agent and RSVP agent having a higher relative accuracy than the computer vision, while maintaining an order of magnitude difference in speed between each agent type. See Table I for a summary of the agent parameters. For all simulations, we measured balanced accuracy (3), wall time, the objective elapsed time while the system runs to convergence, and the number of overall assignments to reach convergence.

Assigning all images to all agents in parallel already represents a speed-up over the serial labeling of all images by all agents, but we want to specifically decrease the wall time of such a system; therefore, we compared three assignment conditions of the proposed framework against this naive parallel implementation:

- Naive - all images assigned to all agents in parallel in a single batch.
- GAP-2 - images assigned in parallel according to (1); images classified if confidence meets or exceeds two, $s_i \geq 2$.
- GAP-3 - same as GAP-2, but with $s_i \geq 3$.
- GAP-4 - same as GAP-2, but with $s_i \geq 4$.

All methods use Algorithm 1 to combine image labels. Additionally, we considered three distinct ensembles of agents for the proposed system:

- $CV \times 6$
- $\{CV \times 2, RSVP \times 2, H \times 2\}$ (Mixed)
- $H \times 6$

A. Analytical Results

We can determine the expected performance of the naive assignment condition analytically, which provides a true performance ceiling to which the GAP assignment conditions can be measured.

The accuracy from the joint classification using the first-order approximation of the SML is bounded from below by the accuracy of the best individual agent in the ensemble to

TABLE II
ANALYTICAL RESULTS OF NAIVE ASSIGNMENT CONDITION

Agent Ensemble	Accuracy (π_j)	Wall Time ($\mu_T \pm \sigma_T$)
$CV \times 6$	0.75	$2.2 \pm 0.1s$
Mixed	0.95	$208.0 \pm 12.0s$
$H \times 6$	0.95	$218.3 \pm 9.7s$

within an additive constant (Lemma S2 (iii) of [19]). This result depends on the strict conditional independence of all classifiers and that all classifiers are better than random, which is easily satisfied by the simulated agents in this experiment.

For the naive assignment condition, the wall time for a single agent to complete classification of n images will be an Erlang random variable, $T_j \sim \text{Erlang}(n, \mu_j)$, and the system wall time will be the maximum of m independent, non-identical Erlang distributions, $T = \max_{j \in J} T_j$. We can numerically evaluate the probability density function,

$$f_T(t) = F_T(t) \sum_{j \in J} \frac{f_{T_j}(t)}{F_{T_j}(t)}, \quad (5)$$

to calculate the mean and standard deviation of the system wall time, $\mu_T = \mathbb{E}(T)$ and $\sigma_T = \sqrt{\mathbb{E}((\mu_T - T)^2)}$.

B. Simulation 1

We compared the performance of a mixed ensemble over all four assignment conditions and collected results from 30 trials of six remote agents classifying 200 images for each assignment condition.

C. Simulation 2

We compared the performance of three distinct agent ensembles using the GAP-2 assignment condition: $CV \times 6$, Mixed, and $H \times 6$. These results provide context to any speed-up in the results of Simulation 1 by comparing the mixed agent ensemble against a fully automated implementation and a fully human ensemble. Again, 30 trials were collected for six remote agents classifying 200 images.

IV. RESULTS

A. Analytical Results

The analytical results for all ensemble conditions under the naive assignment condition are reported in Table II. The mixed agent ensemble matches the lower bound of balanced accuracy of the human ensemble while benefiting from a lower expected wall time. With the automated ensemble, a $95\times$ and $99\times$ fold speed-up can be expected versus the mixed agent and human ensembles respectively, but this speed-up incurs a 21% decrease in the lower bound of balanced accuracy.

B. Simulation 1

All four methods exceeded the analytical lower bound of the mixed agent ensemble. The results are summarized in Table III. We first performed a one-way analysis of variance (ANOVA) which reveals significant differences among the

TABLE III
SIMULATION 1 RESULTS (MEAN AND STANDARD DEVIATION)

Condition	Balanced Accuracy	Wall Time	Images Assigned
Naive	0.988 ± 0.011	204.1 ± 7.9	1200
GAP-2	0.974 ± 0.014	124.1 ± 19.3	879.9 ± 16.3
GAP-3	0.975 ± 0.011	147.9 ± 21.8	983.1 ± 15.1
GAP-4	0.978 ± 0.011	204.4 ± 12.3	1047.6 ± 6.4

assignment strategies ($F(3, 116) = 8.8, p = 2.6 \times 10^{-5}$). For the three GAP assignment conditions, there is not a significant difference in the means of the data at the $p < 0.05$ level, but there is a significant difference between the GAP conditions and the control in a multiple comparisons test (GAP-2: $p = 7.7 \times 10^{-5}$, GAP-3: $p = 2.4 \times 10^{-4}$, GAP-4: $p = 3.8 \times 10^{-3}$).

In the case of wall time, much more noticeable differences arise. An ANOVA reveals significantly different mean wall times between the assignment conditions ($F(3, 116) = 186.5, p < 1.0 \times 10^{-9}$). GAP-2 and GAP-3 achieve significantly lower run-times than the naive assignment condition under a multiple comparisons test (GAP-2: $p = 3.8 \times 10^{-9}$, GAP-3: $p = 3.8 \times 10^{-9}$). The mean of the GAP-2 condition achieves a $1.6\times$ speed-up over the mean of the naive condition, while the GAP-3 achieves a $1.4\times$ speed-up.

Again, an ANOVA reveals significant differences between the number of assignments ($F(2, 87) = 1205.8, p < 1.0 \times 10^{-9}$). Under a multiple comparisons test, the number of overall assignments for all assignment conditions is significantly different than that for all other assignment conditions at the $p < 1.0 \times 10^{-9}$ level.

The mixed agent system clearly achieved superior wall time under the GAP-2 and GAP-3 assignment conditions versus the naive assignment condition, but the results of the accuracy were less clear. Although the naive condition achieved significantly better balanced accuracy than any of the GAP conditions, we are more specifically testing the hypothesis that the GAP conditions achieved or exceeded the analytical lower bound of the naive assignment condition. In this light, the results are more favorable. Under a one sample t-test, it is almost certain that the accuracy of the mixed ensemble GAP assignment conditions achieved or exceeded the analytical lower bound of the mixed ensemble naive assignment (p-values of 8.5×10^{-11} , 7.7×10^{-14} , and 1.3×10^{-14} for the GAP-2, GAP-3, and GAP-4 conditions respectively). Under this more relaxed hypothesis, the GAP-2 and GAP-3 conditions achieved the accuracy performance of the naive assignment condition while significantly decreasing the wall time required to do so.

C. Simulation 2

Not surprisingly, the ensemble with computer vision agents achieved the minimum wall time and also the minimum accuracy, and the ensemble of human agents achieved the maximum accuracy and the maximum wall time. The effect

TABLE IV
SIMULATION 2 RESULTS (MEAN AND STANDARD DEVIATION)

Ensemble	Balanced Accuracy	Wall Time	Images Assigned
$CV \times 6$	0.898 ± 0.030	$6.3 \pm 0.3s$	913.8 ± 13.8
Mixed	0.974 ± 0.014	$124.1 \pm 19.3s$	879.9 ± 16.3
$H \times 6$	0.999 ± 0.003	$294.2 \pm 18.3s$	770.1 ± 7.2

of varying the agent ensemble while using the GAP-2 framework mirrored the analytical results of the naive assignment condition. Results are summarized in Table IV.

Significant differences in balanced accuracy are confirmed under a one-way ANOVA ($F(2, 87) = 255.47, p < 1.0 \times 10^{-9}$). Specific significant differences between the mixed ensemble and the automated ($p < 1.0 \times 10^{-9}$) and human (2.2×10^{-6}) ensembles are confirmed under a multiple comparisons test. Significant differences also arose in wall time ($F(2, 87) = 2667.44, p < 1.0 \times 10^{-9}$), and under a multiple comparisons test, the automated and human ensembles were significantly different than the mixed ensemble, $p < 1.0 \times 10^{-9}$ and $p < 1.0 \times 10^{-9}$ respectively. Similar results were found in the overall number of assignments to reach system convergence between the three ensembles ($F(2, 87) = 1000.9, p < 1.0 \times 10^{-9}$). Under a multiple comparisons test, the mixed ensemble requires a significantly different number of assignments than either the fully-automated ensemble ($p < 1.0 \times 10^{-9}$) or the human ensemble ($p < 1.0 \times 10^{-9}$).

V. DISCUSSION

We presented an image triage system which leverages the collaboration of heterogeneous agents through a novel unsupervised meta-learning technique, and we demonstrated the benefit of such a system through a simulated comparison a naive parallel implementation or a similar homogeneous ensemble of agents. Three types of agents were simulated, representing varying levels of accuracy and throughput. The system dynamically inferred the performance of these agents using the mSML and incorporated that information into subsequent assignments using the GAP.

Even in the naive parallel implementation, the performance of the mixed agent ensemble provides a superior lower bound in balanced accuracy to that of the automated ensemble. Additionally, the mixed ensemble provides a decrease in the expected wall time of the system. This substantial increase in accuracy requires a sacrifice of a similar scale in wall time. This trade-off underlies the challenge of optimizing such a system, but the proposed image triage system attempts to mitigate the time cost of this trade-off through an intelligent assignment framework.

In Simulation 1, we observed the same mixed ensemble achieve similar accuracy in significantly less time under the GAP assignment conditions. For the case of the GAP-2 assignment condition, the system exceeded the guaranteed accuracy of the mixed ensemble naive assignment condition while also affording a $1.6\times$ speed-up over the mixed ensemble

naive assignment condition. It apparently accomplished this time savings by making 27% fewer overall image assignments.

The proposed framework minimizes the number of assignments in the system to achieve a desired confidence. This goal is similar in nature to the CrowdSynth system proposed in [13], where the control logic attempts to calculate the marginal value of recruiting another agent for labeling the image. Here, that decision is encoded in the GAP problem and made simultaneously over all images. Interestingly, we observed a speed-up from this approach in the mixed ensemble, but the automated ensemble and human ensemble required more time in the GAP assignment conditions than that expected from the naive assignment condition. The time cost of this additional decision process is apparently detrimental in the case of a homogeneous ensemble, which corroborates the work of Karger, et al. in [15].

It is likely that in a real-world implementation, the results of incorporating a human into the system will be even more significant. As a result of the simulation, all agents here provided truly conditionally independent labels; however, this quality could be hard to realize in practice. Even different models trained on the same data will introduce conditional dependence to the system. In application, heterogeneous agents may provide the optimal means of introducing independence into the ensemble.

VI. CONCLUSION

The proposed heterogeneous multi-agent image triage system achieved human-level accuracy while minimizing the time required to do so. These results introduce a framework for the collaboration of heterogeneous agents in a multi-agent system through a novel approach to meta-learning in crowdsourcing applications. More immediately, we showed that the introduction of a human to an automated image triage system can instantly increase the expected accuracy of the system to the performance ceiling, and an intelligent assignment policy can minimize the time cost incurred. Future work will confirm these findings in an image triage task with actual human, computer vision, and BCI agents.

ACKNOWLEDGMENT

This project was supported by the Office of the Secretary of Defense Autonomy Research Pilot Initiative program MIPR DWAM31168 and by the U.S. Army Research Laboratory, under Cooperative Agreement Number W911NF-10-2-0022. This research was also supported in part by an appointment to the Student Research Participation Program at the U.S. Army Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and USARL. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [2] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, *Places2*, 2016 (accessed 8 July 2016). [Online]. Available: <http://places2.csail.mit.edu/>
- [3] M. Streit and O. Bimber, "Visual analytics: Seeking the unknown," *Computer*, vol. 46, no. 7, pp. 20–21, 2013.
- [4] P. Michelucci, *Handbook of human computation*. Springer, 2013.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [6] B. J. Lance, S. E. Kerick, A. J. Ries, K. S. Oie, and K. McDowell, "Brain-computer interface technologies in the coming decades," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1585–1599, 2012.
- [7] P. Sajda, A. Gerson, and L. Parra, "High-throughput image search via single-trial event detection in a rapid serial visual presentation task," in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*. IEEE, 2003, pp. 7–10.
- [8] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 16, no. 5, pp. 432–441, 2008.
- [9] P. Sajda, E. Pohlmeier, J. Wang, L. Parra, C. Christoforou, J. Dmochowski, B. Hanna, C. Bahlmann, M. Singh, and S.-F. Chang, "In a Blink of an Eye and a Switch of a Transistor: Cortically Coupled Computer Vision," *Proceedings of the IEEE*, vol. 98, no. 3, pp. 462–478, Mar. 2010.
- [10] B. Collins, J. Deng, K. Li, and L. Fei-Fei, "Towards scalable dataset construction: An active learning approach," in *European Conference on Computer Vision*. Springer, 2008, pp. 86–98.
- [11] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 762–769.
- [12] A. J. Joshi, F. Porikli, and N. P. Papanikolopoulos, "Scalable active learning for multiclass image classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2259–2273, 2012.
- [13] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 2012, pp. 467–474.
- [14] C.-J. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 534–542.
- [15] D. R. Karger, S. Oh, and D. Shah, "Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems," *Operations Research*, vol. 62, no. 1, pp. 1–24, Feb. 2014.
- [16] P. G. Ipeirotis, F. Provost, V. S. Sheng, and J. Wang, "Repeated labeling using multiple noisy labelers," *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 402–441, Mar. 2013.
- [17] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine Learning*, vol. 95, no. 3, pp. 291–327, Oct. 2013.
- [18] Q. Hu, Q. He, H. Huang, K. Chiew, and Z. Liu, "A formalized framework for incorporating expert labels in crowdsourcing environment," *Journal of Intelligent Information Systems*, pp. 1–23, Jul. 2015.
- [19] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, "Ranking and combining multiple predictors without labeled data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1253–1258, Jan. 2014.
- [20] O. E. Kundakcioglu and S. Alizamir, "Generalized assignment problem Generalized Assignment Problem," in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Springer US, 2008, pp. 1153–1162.
- [21] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics*, pp. 20–28, 1979.