Asking for Help with the Right Question by Predicting Human Visual Performance

Hong Cai and Yasamin Mostofi
Dept. of Electrical and Computer Engineering, University of California Santa Barbara
{hcai, ymostofi}@ece.ucsb.edu

***Abstract*—In this paper, we consider robotic surveillance tasks that involve visual perception. The robot has a limited access to a remote operator to ask for help. However, humans may not be able to accomplish the visual task in many scenarios, depending on the sensory input. In this paper, we propose a machine learning-based approach that allows the robot to probabilistically predict human visual performance for any visual input. Based on this prediction, we then present a methodology that allows the robot to properly optimize its field decisions in terms of when to ask for help, when to sense more, and when to rely on itself. The proposed approach enables the robot to ask the right questions, only querying the operator with the sensory inputs for which humans have a high chance of success. Furthermore, it allows it to autonomously locate the areas that need more sensing. We test the proposed predictor on a large validation set and show Normalized Mean Square Error of 0.0199, as well as a reduction of about an order of magnitude in error as compared to the state-of-the-art. We then run a number of robotic surveillance experiments on our campus as well as a larger-scale evaluation with real data/human feedback in a simulation environment. The results showcase the efficacy of our approach, indicating a considerable increase in the success rate of human queries (a few folds in several cases) and the overall performance (30%-41% increase in success rate).**

I. Introduction

Thanks to the advances in areas such as perception, navigation, and robotic manipulation, robots are becoming more capable of accomplishing complicated tasks. There, however, still exist many tasks that robots cannot autonomously perform to a satisfactory level. As such, robots can benefit tremendously from seeking human's help, as has been demonstrated in recent work [16, 17, 27, 36]. In most existing work, it is assumed that human performance is perfect. This, however, is not true in many cases. For instance, consider a task that involves visual perception. While the state-of-the-art in vision has improved tremendously due to the advances in machine learning, there can still be several sensory inputs for which the robot cannot correctly perform a visual task. The robot may have access to a remote human operator to ask for help with the task. However, human visual performance may also be far from perfect depending on the sensory input.

Fig. 1 demonstrates a real example of this for a robotic surveillance task on our campus.[1] An unmanned ground vehicle is given a visual perception task that involves finding the human at each of the 4 shown sites, based on onboard camera inputs. The robot has a limited access to a remote human operator to ask for help with this task. The figure shows the images taken by the robot after its initial sensing of each site. Equipped with a state-of-the-art vision algorithm, the robot has no problem recognizing the humans at Sites 1 and 4 itself. However, the robot's vision fails for Sites 2 and 3. On the other hand, the person at Site 3 can be easily detected by a human, while it is also hard for humans to spot the person at Site 2 based on the current sensory input. Thus, if the robot assumes that human is perfect and asks for help by sending image 2 to the human operator, it is highly likely that the operator will fail to find the human in the image.

Fig. 1: A robotic surveillance task on our campus that involves finding the person in each of the 4 sites, based on imagery inputs. The robot has a limited access to a remote operator to ask for help and needs help with Sites 2 and 3. However, human performance is not perfect and she cannot help with Site 2 if asked. If the robot could predict human visual performance for a sensory input, it could optimally ask for help with Site 3 and move to Site 2 to sense more. Readers are referred to the color pdf for details.

This example highlights three key points at the core of this paper. First, robot's vision will fail for several realistic scenarios that involve visual perception. Second, there will be many cases where humans may not be able to help the robot depending on the sensory input. Given that remote operators typically have to manage several different tasks under fatigue and time pressure, the robot should only ask for human's help when it is highly confident that humans can do the task. Third, the robot needs to move closer to those sites where both robot and human fail the task for further sensing.

In this paper, we consider robotic surveillance tasks that involve visual perception. The robot has a limited access to a remote operator to ask for help with its task. We propose

[1]Readers are referred to the color pdf to see all the images of this paper more similar to what was seen by the robot and humans during the operation. Furthermore, we note that all the images used as part of this work are at the size of 256×256, while we have to show a smaller version in the paper. We note though that the visual difficulty of the real-size images is pretty consistent with the visual difficulty of the smaller versions when viewed in color. We have included all the images of the performance evaluation section at their real size in the supplementary document (http://dx.doi.org/10.7919/F43X84K7).

a methodology that allows the robot to properly optimize its decisions in terms of which sites to ask for help, which sites to visit and further sense, and which sites to rely on itself. At the core of our approach, is a proposed probabilistic predictor for human visual performance, which allows the robot to probabilistically assess human performance for a given sensory input. This enables the robot to **ask the right questions**, only querying the operator with the sensory inputs for which humans have a high chance of success. Then, the robot will use the feedback from our predictor to optimize its human collaboration and further sensing of the field.

A. Related Work

There is a great body of work on different aspects of human-robot collaboration [18, 19, 21, 31, 33]. More related to this work are those papers that focus on robots asking human for help. For instance, [10] shows how robots can recover from difficult states or failures by asking for help. In [25, 35, 36], a robot learns from human demonstration and correction, while a robot performs object detection and recognition with human inputs in [23, 28, 32, 39]. In computer vision, a number of work have focused on designing human-machine interfaces that allow the vision algorithm to ask for human's help when it encounters difficulties [4, 9, 30, 40]. In most of these work on asking for help, however, human is assumed perfect in task accomplishment. As can be seen in Fig. 1, human visual performance is not perfect depending on the sensory input. In [34], robot asks for human help by generating unambiguous sentences for a collaborative manipulation task, a subject different from this paper.

A number of work have taken imperfect human performance into account for non-robotic vision applications. In [4], for instance, authors propose a collaborative vision task inspired by the 20-question game. [30] proposes a collaborative annotation system in which human and machine collaboratively label the objects in images. In these work, however, it is assumed that human performance is task dependent but invariant to the sensory inputs. In other words, for a specific visual task, the probability of human's task accomplishment is constant (less than 1), independent of the input image. While a good step towards considering imperfect human performance, human visual performance can largely vary for a given visual task, depending on the sensory input, as shown in Fig. 1.

A few work have attempted to estimate human visual performance based on a given input image. Johnson criteria is one of the first attempts along this line, where human visual performance is predicted based on the number of line pairs of display resolution occupied by the target on the screen [26]. In cognitive psychology, it is heavily acknowledged that human visual system is not perfect [11, 12, 13], motivating work that attempt to understand how different image features affect human performance [7]. The goal of these work, however, is not predicting human performance for a given image but understanding a certain feature's impact on human performance.

In [14, 37], authors utilize machine learning to predict the probability that a driver is able to detect a pedestrian at a glance based on hand-crafted image features, such as size and position of the pedestrian in the image. In [5, 6], imperfect human visual performance is considered in the context of a robotic field operation, emphasizing the importance of properly optimizing human's help. However, human visual performance is predicted when noise with a known variance is added to otherwise easy images. The known noise variance is then used as a hand-crafted feature to predict human performance. In general, a sensory input such as an image may have several features that can make it hard or easy for humans to perform a visual task, making identifying and hand-crafting all of them very challenging. Thus, an automated method that can predict human visual performance for any given sensory input is needed and currently lacking, which is one of the main motivations for this paper. More specifically, as compared to the existing approaches, our proposed prediction methodology is fundamentally different in that the relevant image features are selected in an automated manner during training, by properly utilizing convolutional neural networks, and then used for predicting human visual performance.

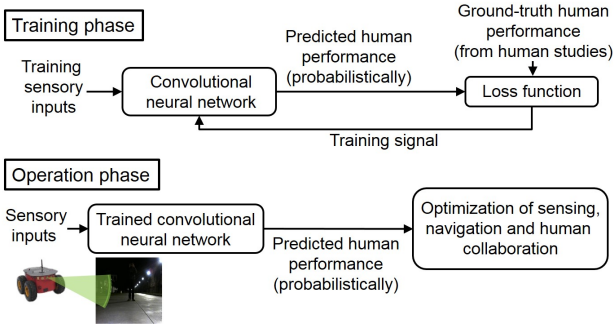

Fig. 2: High-level diagram of our approach for (top) predicting human performance using machine learning and (bottom) field operation and robot decision optimization with human performance prediction.

***Statement of contribution*:** The main contributions of this paper are then as follows: 1) we propose a machine learning-based approach that allows the robot to probabilistically predict human visual performance for a given visual input, without a need for hand-crafting any feature. Fig. 2 (top) shows a high-level diagram. We train our neural network by gathering several human data using Amazon Mechanical Turk (MTurk) [2]. We then test the proposed predictor on a large validation set and show a considerable reduction in prediction error as compared to the state-of-the-art, 2) We propose how the robot can optimize its field decisions in terms of relying on itself, asking for help, and further sensing, based on the output of the predictor as summarized in Fig. 2 (bottom), and 3) We run a number of robotic surveillance experiments on our campus, which showcase the effectiveness of our approach, indicating a considerable increase in the success rate of human queries and the overall task. We further run a larger-scale evaluation in a simulation environment, based on images taken on our campus and with MTurk users acting as human operators. To the best of our knowledge, these are the first real experiments where an unmanned vehicle optimizes its collaboration with a human operator, based on predicting human performance for each sensory input, and performs futher sensing of the field

based on this prediction.

II. Problem Formulation & Decision Optimization

In this section, we formulate the human-robot collaborative perception problem as a constrained optimization problem and discuss the optimum decisions such that robot's motion energy usage is minimized while task performance is guaranteed above a certain level. We then propose how to equip the robot with a human performance predictor in order to successfully execute the optimum decisions. In the rest of the paper, we focus on surveillance tasks for finding humans based on imagery inputs. We note that the proposed methodology is applicable to any visual perception task with any sensory input.

A. Problem Formulation

Consider a case where there is a total of N sites with a human at each site. A field robot is tasked with finding the human at each site.[2] The robot has limited access to a human operator to ask for help with the task, in the form of M maximum questions. The robot can also spend motion energy and time to move closer to a site for better sensing. The robot's goal is to successfully perform the task while minimizing its total energy usage (or equivalently operation time).

During the operation, the robot first performs initial sensing of the sites by taking a picture of each site. Based on these sensory inputs, it then estimates its own probability of task accomplishment, which is denoted by $p_{r,i}$ for the i^{th} site, for $i \in \{1,...,N\}$. If the estimated probability is high enough for a site, then the robot can rely on itself. If not, it has to decide if it should ask for help from a remote operator for this site or if it should move to the site for further sensing. In order to properly make this decision, it needs to assess the chance that the operator can perform the task successfully. Let $p_{h,i}$ denote the probability that humans can successfully perform the task for the i^{th} site, for $i \in \{1,...,N\}$. We then have the following optimization problem:

$$\begin{aligned} \min_{\gamma,\eta,\omega} \quad & \mathscr{E}^T \eta \\ \text{s.t.} \quad & \gamma \circ p_h + \omega \circ p_r + \eta \succeq p_{\text{Th}} \mathbf{1}, \\ & \mathbf{1}^T \gamma \leq M, \quad \gamma + \eta + \omega = \mathbf{1}, \quad \gamma, \eta, \omega \in \{0,1\}^N, \end{aligned} \tag{1}$$

where $\mathscr{E} = [\mathscr{E}_1,...,\mathscr{E}_N]^T$ is the motion energy cost vector to visit the sites, p_{Th} is the minimum acceptable probability of successful task accomplishment, $p_h = [p_{h,1},...,p_{h,N}]^T$, $p_r = [p_{r,1},...,p_{r,N}]^T$, $\gamma = [\gamma_1,...,\gamma_N]^T$, $\eta = [\eta_1,...,\eta_N]^T$ and $\omega = [\omega_1,...,\omega_N]^T$. Moreover, $\gamma_i = 1$ if the robot seeks human's help for site i and $\gamma_i = 0$ otherwise. $\eta_i = 1$ if the robot visits site i and $\eta_i = 0$ otherwise. Furthermore, $\omega_i = 1$ if the robot relies on itself for site i and $\omega_i = 0$ otherwise. $\mathbf{1}$ is the vector of all 1s, $\circ$ denotes the Hadamard product and $\preceq$ indicates that the inequality is component-wise.

In order to mathematically characterize the optimum decisions, we assume that if the robot moves to a site for further sensing, its probability of successful task accomplishment is 1 in Eq. 1.[3] The optimum solution of Algorithm 1 can then be easily confirmed for Eq. 1.

B. Proposed Decision Optimization

Algorithm 1 lays out the decision optimization of the robot during the operation. If the robot has a high confidence in finding the person at a site ($p_r \geq p_{\text{Th}}$), then it can simply rely on itself, eliminating the need for further sensing or asking for help. In Section IV, we discuss how robot's vision algorithm provides it with p_r so it can assess its own performance.

Algorithm 1: Decision Making Algorithm

Initialization: $\omega_i = 0,\ \gamma_i = 0,\ \eta_i = 0,\ \forall i \in \{1,...,N\}$.
Step 1: $\forall i \in \{j : p_{r,j} \geq p_{\text{Th}},\ j \in \{1,...,N\}\}$, set $\omega_i = 1$;
Step 2: $\forall i \in \{j : \omega_j = 0,\ p_{h,j} \geq p_{\text{Th}},\ j \in \{1,...,N\}\}$, set γ_i with the M largest $\mathscr{E}_i$ to 1;
Step 3: $\forall i \in \{j : \omega_j = \gamma_j = 0,\ j \in \{1,...,N\}\}$, set $\eta_i = 1$.

For sites that $p_r < p_{\text{Th}}$, the robot has to decide on which ones to visit and which ones to ask the operator. In order to make this decision, the robot needs to predict the chance that the operator can accomplish the given perception task (p_h). This motivates our proposed predictor for human visual performance, which we extensively discuss in the next section. Based on Algorithm 1, the robot then evaluates the images of these sites with its human predictor. Let N_{hard} denote the number of sites that the robot cannot rely on itself. Out of these sites, let N_{human} denote the number of sites for which the robot predicts that the human can accomplish the task ($p_h \geq p_{\text{Th}}$). If $M \geq N_{\text{human}}$, then the robot will pass all those sites to the operator. If $M < N_{\text{human}}$, then the robot will choose the M sites that cost the most energy to visit and pass them to the operator. It will then visit the remaining sites. Fig. 3 summarizes optimum decision making of the robot.

III. A Predictor for Human Visual Performance

In this part, we develop a predictor for human visual performance in a task that involves finding a human in an image. More specifically, given a visual input in the form of an image, the robot wants to predict the probability that humans can see the person in the image. In order to equip the robot with this capability, we collect several images with different levels of difficulty to train a Convolutional Neural Network (CNN), as shown in Fig. 4. We next describe the details of the training process, the machine learning algorithm, as well as the resulting performance and underlying trends.

A. Training Dataset

We have built a dataset of 3000 total images, with each image containing a human.[4] We have included images with different degrees of visual difficulty for proper training. Sample challenging cases include images in cluttered environments, images where the human is far away, and dark images. The

[2]We emphasize that the considered task in the rest of the paper is to find the human at each site, given there is a human at each site. In other words, the robot or human operators/Mturk users know there should be a human at each site and the task is to find the human at each site.

[3]In our experiments, we take the true probability of task accomplishment into account when a site is visited.

[4]Ideally each image should contain only one human to be consistent with the defined task. However, due to the difficulty of finding images online that capture challenging scenarios, a small percentage of the images contain more than one person in the image. The MTurk users are instructed to answer "yes" if they could spot a person in the image. In such cases, the task can be thought of as finding the most obvious person in the image.

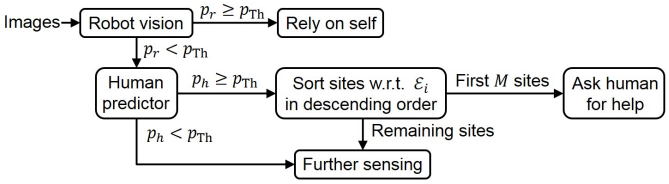

Fig. 3: Flow diagram of robot's decision making process.

images are mainly selected from the following sources: NOAA Natural Hazards Image Database [1], the SUN dataset [38], and the PASCAL VOC dataset [15]. All the images are resized to 256×256. For each acquired image, we have manually darkened it to create a new data point that is also dark.

1) Human Data Collection: For each training image, we need to evaluate human visual performance. We do so empirically by using MTurk. For each image, 50 MTurk workers are asked if they could see a human in the image. The responses are then averaged for each image to empirically assess the probability that a person can accomplish the given visual task.

2) Statistical Characteristics of the Dataset: Ideally, our dataset should be balanced in the number of difficult and easy images to avoid biasing the learning process. However, we find more easy images in online datasets than hard ones. For instance, based on the collected data from MTurk, the probability of task accomplishment is above 0.95 for about 70% of the images and below 0.9 for about 20% of the images. To avoid biasing the prediction, we utilize the commonly used technique of oversampling [8] in machine learning to create a more balanced dataset. As our extensive experimental results of Section V indicate, the number of hard images in the initial set is still sufficient as our trained predictor can well differentiate between the images that are easy and difficult for humans. We note that the design of the constrained optimization problem of Algorithm 1 is general and not affected by dataset imbalance. Building a more balanced dataset can improve the prediction performance, and the overall task success rate, when implementing this approach.

B. Using CNN to Predict Human Performance

We train a convolutional neural network to predict human performance. The high-level structure of the network is shown in Fig. 4. Our proposed network architecture is a modified version of Alexnet [22], which is among the best existing classification networks. The original Alexnet consists of 5 convolutional layers and 2 fully connected layers, followed by the output layer. Alexnet is originally designed for object classification. In this paper, however, we are interested in a different task of probabilistically predicting human performance in finding a person in an image, which is a regression task. We thus replace the output layer of Alexnet with a regression layer, which then outputs the human performance based on the output of the 7th layer of Alexnet.

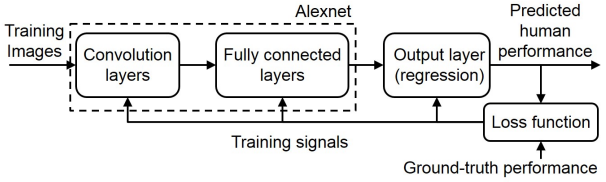

Fig. 4: Flow diagram of training a convolutional neural network to predict human performance probabilistically.

Image Categories	Normalized MSE
All	0.0199
Non-dark	0.0067
Dark	0.0330
Hard	0.0479
Hard Non-dark	0.0493
Hard Dark	0.0474

TABLE I: Performance of our human predictor – NMSEs over different subsets of images in the validation set.

By using the dataset of Section III-A as input images and the corresponding MTurk responses as ground-truth labels, we train the resulting network as follows. The parameters of the first 7 layers are initialized with the weights of the corresponding layers of an Alexnet trained for human detection (classification from a set of $\{\text{Human}, \text{No Human}\}$). The parameters of the output layer are initialized randomly. Mean Squared Error (MSE) is used as the loss function that computes the update for network weights. The underlying motivation for this design is as follows. The parameters of the first 7 layers of the Alexnet trained for human detection capture features useful for detecting human presence, which are also informative for predicting human performance in finding a person in the image. But since our task is different, all the weights need to be updated during the training based on the input from human performance. The training is performed for 50K iterations with an initial base learning rate of 10^{-5}, which is reduced by a factor of 10 after every 10K iterations. Stochastic gradient descent is used to update the network weights and the training batch size is 128. We employ the machine learning library Caffe [20] to train and test our models.

C. Evaluation of the Proposed Human Performance Predictor

We next evaluate the performance of our trained human predictor over the validation set. The validation set is randomly selected out of the original image pool and thus has a similar ratio of hard to easy images as the training pool. We first look at the Normalized Mean Squared Error (NMSE) to evaluate the prediction quality, which is calculated as follows: $\frac{1}{V}\sum_{i=1}^{V}(p_{h,i,\text{val}} - \hat{p}_{h,i,\text{val}})^2 / p_{h,i,\text{val}}^2$, where $p_{h,i,\text{val}}$ and $\hat{p}_{h,i,\text{val}}$ are the ground-truth and predicted human performance of the i^{th} validation image and V is the number of image in the validation set.[5] Table I shows a summary of NMSE over different subsets of the validation set. The NMSE over all validation images is 0.0199. The NMSE over all non-dark and dark[6] validation images are 0.0067 and 0.033 respectively. It is also of great importance to our robotic task to be able to predict images hard for humans. Define hard images such that the empirical probability of a person finding the human in the image, i.e., the ground-truth probability, is less than 0.9. The NMSEs over all such hard images, hard non-dark images and hard dark images are also summarized in Table I. As can be seen, the NMSE values are fairly small, indicating a good

[5]Although NMSE values are higher than the corresponding MSE ones since we are predicting a positive value bounded by 1, we find NMSE a more truthful metric of the performance as it is normalized by the true value.

[6]Non-dark images refer to those directly taken from the aforementioned datasets and dark images refer to those that are manually darkened.

training performance. Fig. 5 shows four sample images with the true and predicted human performance annotated (bottom-right) on each image. As can be seen, our predictor can predict the performance well for these images. Fig. 6 further shows the empirical Cumulative Distribution Function (CDF) of NMSE over the validation set. It can been seen that the NMSEs of most images (more than 90%) are upper bounded by 0.05. We note that the NMSE of the validation set and the training set is 0.0199 and 0.0112 respectively, indicating that the network did not over-fit the training data.

Fig. 5: Comparison of the true and predicted probability of successful human performance for (top-left) an easy non-dark image, (top-right) a hard non-dark image, (bottom-left) an easy dark image and (bottom-right) a hard dark image. Readers are referred to the color pdf for better viewing.

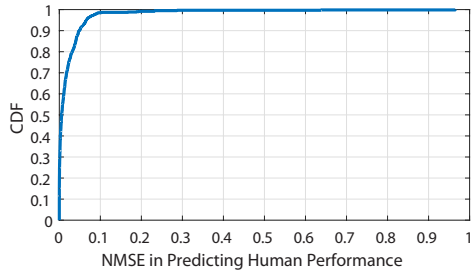

Fig. 6: The CDF of NMSEs of validation images.

To the best of authors' knowledge, there are two papers that have looked into the problem of predicting the probability with which a person is able to see a human in an image. The reported prediction Mean Absolute Error (MAE) is at least 0.22 in [37], while the prediction Mean Squared Error (MSE) is 0.04 in [14]. The comparison between our predictor and those reported in [14, 37] is summarized in Table II. As can be seen, our approach can achieve a much better performance. This is mainly because our approach is fundamentally different from the past work in that [14, 37] choose a few hand-crafted image features for prediction. In general, however, there exist many underlying features that contribute to what makes an image difficult or easy for human visual system, which are challenging to hand-craft. Thus, in our approach, the important features of an image are automatically captured by the CNN in Fig. 4, and utilized for performance prediction.

Human Predictors	MSE	MAE
Our Predictor	0.0067	0.0459
Predictor of [14]	0.04	N/A
Predictor of [37]	N/A	0.22

TABLE II: Comparison of the performance of our predictor with two existing predictors. N/A indicates that the evaluation is not available. The performance of our predictor is evaluated over all non-dark validation images, which best match the characteristics of the images used in [14, 37].

Remark 1: Table II shows what [14, 37] reported for their prediction performance. Although we evaluate the performance of our predictor over our non-dark validation images, which best resemble the images used in [14, 37], we note that our dataset is not the same as that of [14] or [37]. The performances of the three predictors are thus not fully comparable since they are evaluated over different datasets.

Remark 2: As can be seen, the NMSE values in Table I are slightly larger over the set of hard images. While we expect hard images to be more challenging for the predictor, this can also be partly due to the dataset imbalance (discussed in Section III-A2). It can, however, be seen that the NMSE values are still very small, indicating a good training performance. Building a more balanced dataset by including more hard images, as part of future work, can improve the prediction performance in more challenging settings.

D. Variability Among Human Operators and p_{Th}

In general, we should choose p_{Th} high such that the robot only asks the human about those images for which almost all the probed people have correctly done the task. This is in particular important as we consider cases where the human operator can only look at the image at a glance and does not have time to investigate the details due to work overload. Thus, the robot wants to only ask for help if it is fairly confident that the human can be of help.

Choosing a high threshold also makes the prediction more immune to the variability among operators. As expected, there will be variability among different humans in accomplishing a visual task. Furthermore, performance of the same person can vary depending on factors such as fatigue and attention overload. The variability of performance, however, is much less for easy images, as compared to hard images. This makes sense as humans can typically perform an easy task even under fatigue and stress while performing harder tasks requires more focus and energy. We thus choose $p_{\text{Th}} = 0.9$ in the next section on performance evaluation.

IV. Robot Vision

The robot uses the state-of-the-art Alexnet vision algorithm [22] for task accomplishment, which also provides it with a confidence metric to assess p_r. We train Alexnet to classify images into two classes of $\{\text{Human}, \text{Non-Human}\}$, i.e., images that contain at least one human and images that do not contain any. We fine-tune the Alexnet on our dataset (MS COCO [24] plus manually darkened images) with 230K training images

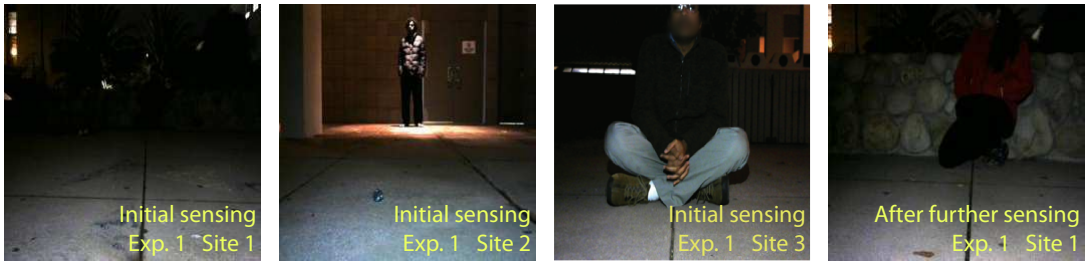

Fig. 7: Initial sensing images of 3 sites at crossroad 1 and the image taken after moving to Site 1 for further sensing based on our approach. Readers are referred to the color pdf for all the experimental results. The visual difficulty of the real-size images is pretty consistent with that of the smaller versions when viewed in color. The supplementary document contains all experimental images in their real size.

from a model pre-trained on the Imagenet data [29]. The accuracy of our model over the validation set (80K images) is 0.85. The network gives its confidence in terms of a probability for each class. We then take $p_r = \text{prob}\{\text{Human}\}$.

V. Performance Evaluation

In this section, we start by evaluating the performance of the proposed approach with 4 experiments on our campus. Additionally, we evaluate the performance at a larger scale by simulating a case with 15 sites, based on real sensing data from our campus and real human performance using MTurk. We further compare our approach with the best possible without human prediction, to which we refer as "benchmark". The benchmark optimizes the decisions without knowledge of human performance, as is summarized below:

Initialization: $\gamma_i = 0$, $\eta_i = 0$ and $\omega_i = 0$, $\forall i \in \{1,...,N\}$;

Step 1: $\forall i \in \{j : p_{r,j} \geq p_{\text{Th}},\ j \in \{1,...,N\}\}$, set $\omega_i = 1$;

Step 2: $\forall i \in \{j : \omega_j = 0,\ j \in \{1,...,N\}\}$, set γ_i with the M highest $p_{r,i}$ to 1;

Step 3: $\forall i \in \{j : \omega_j = \gamma_j = 0,\ j \in \{1,...,N\}\}$, set $\eta_i = 1$.

In summary, the robot orders the sites based on its own confidence (p_r) from highest to lowest. If the robot's confidence is high enough, it will rely on itself. It then selects the next M sites to query human operators, and visits the remaining sites. We note that in several literature on human-robot collaboration, the human is assumed perfect. We can also compare our approach to the case of perfect human. In this case, the robot should choose M most expensive sites to query the operator from the sites that are difficult for itself. However, this approach does worse than the aforementioned benchmark. We thus compare our approach with the described benchmark in this section. When comparing to any other approach, we note that the sites that the robot relies on itself will naturally be the same, as expected, since this paper is about how to optimize the decisions when the robot cannot rely on itself.

A. Robotic Experiments

We perform a number of robotic experiments at different locations on our campus to validate the proposed approach. In each experiment, the robot starts at the center of a crossroad. Each crossroad direction is a site of interest, and there is a person at each site. The robot is tasked with finding the person in each direction based on camera inputs, and is given maximum of $M = 1$ question to ask remote operators for help with finding the person in that experiment. The robot can choose to move to a site for further sensing. The goal of the robot is to find the person in each direction with a very high probability, while minimizing its total energy usage.

Site	$p_r \geq p_{\text{Th}}$	$\hat{p}_h \geq p_{\text{Th}}$	Our Method	Benchmark
1	0	0	Visit	Ask
2	1	1	Self	Self
3	1	1	Self	Self
Ave. Prob. Success			0.96	0.69
Ave. Prob. Human Success			N/A	0.07

TABLE III: Performance summary at campus crossroad 1.

The robot does an initial sensing of each direction by rotating, facing each direction, and taking a color picture. The robot then inputs the pictures to its onboard human predictor and decision making algorithm (Eq. 1). Based on the resulting optimum decisions, the robot may select a site to ask for operator's help and a number of sites for further sensing according to the strategy described in Section II-B. The robot is a Pioneer 3-AT ground vehicle [3], equipped with a webcam for sensing and a laptop for processing. Camera images are resized to 256×256 to be compatible with Alexnet. For all the images, the readers are referred to the color pdf. The supplementary document also contains the experimental images in their real size. We further note that we have blurred the faces in the images for submission anonymity.

1) Campus Robotic Experiment 1: In this experiment, the robot starts at crossroad 1 with three sites in three different directions. Fig. 7 shows the images that the robot takes during its initial sensing. Table III shows robot's performance (p_r) as well as predicted human performance ($\hat{p}_h$), as compared to the required threshold (0.9) for the three sites, with "1" indicating that the threshold is satisfied and "0" denoting otherwise.[7] As can be seen, the robot can confidently rely on itself for Sites 2 and 3. However, it has a low confidence for Site 1, which is also hard for humans, as can be seen. Without our predictor, however, the robot has no methodical way of making the right decision for this site. The table shows robot's decisions for both our approach and the benchmark. As can be seen, our human predictor accurately predicts that Site 1 is too difficult for humans and thus does not send this image to the operators. Instead, it chooses to move to this site to take the 4th image, which is now easy for itself. The benchmark, on the other hand, would inquire the operator on the first image, for which MTurkers had 0.07 chance of spotting the human. The table also shows the overall average probability

[7]The threshold $p_{\text{Th}} = 0.9$ is used throughout this section.

Fig. 8: Initial sensing images of 3 sites at crossroad 2 and the image taken after moving to Site 1 for further sensing based on our approach.

Site	$p_r \geq p_{\text{Th}}$	$\hat{p}_h \geq p_{\text{Th}}$	Our Method	Benchmark
1	0	0	Visit	Ask
2	0	1	Ask	Visit
3	1	1	Self	Self
Ave. Prob. Success			1	0.71
Ave. Prob. Human Success			1	0.27

TABLE IV: Performance summary at campus crossroad 2.

of task accomplishment, which is 0.69 for the benchmark and 0.96 for our approach (39% higher than benchmark). Average probability of success is averaged over all the sites, including the ones that the robot relies on itself.

2) Campus Robotic Experiment 2: In this experiment, the robot starts at crossroad 2 with three sites in three different directions. Fig. 8 shows the images taken by the robot during its initial sensing. Table IV shows robot's performance and predicted human performance. As can be seen, the robot can confidently rely on itself for Site 3. However, it has a low confidence for Sites 1 and 2 in this case. Site 1, however, is also hard for humans, while Site 2 is easy for humans, as can be seen.[8] The table shows robot's decisions for both our approach and the benchmark. Our human predictor accurately predicts that Site 2 is easy for humans while Site 1 is hard. The robot then chooses to move to Site 1 and send image of Site 2 to operators for help, while relying on itself for Site 3.

The benchmark instead queries the operator with Site 1 and moves to Site 2 for further sensing, which results in wasting one question, not spotting the person in Site 1, and moving to the wrong site for further sensing. This is due to the fact that p_r of Site 1 is higher than Site 2. In other words, robot's vision algorithm cannot properly predict human's performance. As a result, the average probability of human success is 0.27 in the benchmark case while it is 1 in our case (3.7 times higher). The overall average probability of success is 0.71 for the benchmark and 1 for our approach (41% higher than benchmark) in this case.

3) Campus Robotic Experiment 3: In this experiment, the robot starts at crossroad 3 with four sites in four different directions. Fig. 9 shows the initial images. As can be seen, it is hard to spot the person in Site 1, while the person in the three other sites can be easily detected. Table V shows robot's performance and predicted human performance. The predictor flags the first site as hard for humans. Then, our approach results in the robot asking about Site 4, visiting Site 1, and relying on itself for the remaining two sites. The benchmark moves to Site 4 instead and inquires operators on Site 1. The

[8]It is easy to spot the person in this image in the color pdf.

Site	$p_r \geq p_{\text{Th}}$	$\hat{p}_h \geq p_{\text{Th}}$	Our Method	Benchmark
1	0	0	Visit	Ask
2	1	1	Self	Self
3	1	1	Self	Self
4	0	1	Ask	Visit
Ave. Prob. Success			1	0.77
Ave. Prob. Human Success			1	0.12

TABLE V: Performance summary at campus crossroad 3.

average probability of human success is 0.12 in the benchmark case while it is 1 (8.3 times higher) in our case. The overall average task accomplishment probability is 1 for our approach, 30% higher than the benchmark (0.77) in this case.

4) Campus Robotic Experiment 4: Next, we show a case where the benchmark gets lucky and has the same decisions as our approach. In this experiment, the robot starts at crossroad 4 with three sites, as shown in Fig. 10. Robot's vision fails for Sites 1 and 3. As can be seen, it is hard to see the person in the first image and easy to spot the person in the other two. Our approach correctly predicts this and visits Site 1 while asking operators about Site 3. The benchmark coincides with our approach in this case. However, in general, it will be hard for the robot to ask for help without a proper human predictor.

Overall, our experimental campus results confirmed that the human predictor can properly help the robot identify which images are hard/easy for humans, allowing for the optimization of further sensing and human query. Unless all the sites are easy for the robot or for the human, the robot cannot methodically optimize its decisions without properly predicting human performance. In practice, there will be several hard cases where the robot cannot rely on itself and the sensory input is still too hard for humans, as we have shown. The proposed predictor and decision optimization approach can then be a valuable tool for the robot to achieve its best performance with minimum resources.

B. Further Evaluation over 15 Sites

In order to demonstrate the performance at a larger scale, we next show simulation results where a robot is tasked with finding humans in 15 sites with limited help from an operator. More specifically, it is given a maximum of 5 queries to operators and can also choose to visit a site after its initial sensing. All the sites have the same cost to visit. The images taken by the robot after its initial sensing are real data taken around our campus and have different levels of difficulty. The final evaluation of the performance is done by passing the chosen images to MTurk users. Table VI summarizes task difficulty of the sites for both robot and human. Without loss

Fig. 9: Initial sensing images of 4 sites at crossroad 3 and the image taken after moving to Site 1 for further sensing based on our approach.

	Sites 1-5	**Sites 6-10**	**Sites 11-15**
Robot Perf.	$p_r < p_{\text{Th}}$	$p_r < p_{\text{Th}}$	$p_r \geq p_{\text{Th}}$
Human Perf.	$p_h < p_{\text{Th}}$	$p_h \geq p_{\text{Th}}$	$p_h \geq p_{\text{Th}}$
Predicted Human Perf.	$\hat{p}_h < p_{\text{Th}}$	$\hat{p}_h \geq p_{\text{Th}}$	$\hat{p}_h \geq p_{\text{Th}}$

TABLE VI: Case of 15 sites and 5 allowed queries - The table shows robot and human performance for each site, as well as the prediction of human performance.

	Visit	**Ask**	**Self**	**Ave. Prob. Human Success**
Our Method	1-5	6-10	11-15	0.98
Benchmark	1,3,6-8	2,4,5,9,10	11-15	0.58

TABLE VII: Summary of the decisions and final performance for the case of 15 sites and 5 questions. The table shows robot's decision for each site, as well as the overall performance of the MTurk operators when asked.

of generality, we numerate the sites from hard to easy based on true human performance from 50 MTurk users (p_h). It can be seen that sites 1-5 are hard for both the robot and the human, sites 6-10 are hard for the robot but easy for human, and sites 11-15 are easy for both. The table further shows that our predictor correctly identifies the sites that are hard/easy for humans. Table VII then compares the decisions of our approach with the benchmark. As can be seen, the two methods result in very different decisions in terms of which sites to visit or query the human operator. The last column shows the average (averaged over queried sites) probability that human operators accomplish the task when asked, based on passing each image to 50 MTurk users. This probability is 0.98 for our approach and 0.58 for the benchmark.

It can be seen that by using our approach, the images selected to query the operator are those that the human operator can indeed be of help with. Without a proper prediction, however, the robot can send several hard images to the operator, instead of further sensing, thus wasting questions and incurring large performance loss.

VI. Conclusions and Future Directions

Despite great advances in vision, robot's visual perception can still fail in many cases. The robot can ask humans for help in collaborative tasks. However, human visual performance is also not perfect, depending on the sensory input.

In this paper, we proposed a machine-learning based approach that allows the robot to probabilistically predict human visual performance for any sensory input. Equipped with this tool, we then showed how the robot can optimize its field decisions, in terms of asking for help, further sensing, and relying on itself. We tested the proposed approach on our campus with a number of robotic surveillance experiments and showed a considerable improvement in the performance. Moreover, we ran a larger-scale evaluation, with real data/human feedback, in a simulation environment to further showcase the effectiveness of the approach.

While we focused on robotic surveillance tasks based on imagery inputs, the methodology can be extended to other collaborative tasks/sensory inputs. It can also be extended to a more extensive sequential optimization framework that allows for asking more questions after further sensing. The predictor can also be fine-tuned to the performance of a particular operator for a longer-term partnership, or to the time of the day, among other factors. As part of future work, building a larger training dataset that includes more hard images relevant to the task can further reduce dataset imbalance and improve the overall performance. For instance, the normalized MSE of prediction for the hard images on our campus has typically been higher than those of the validation set. This is due to the fact that the training set does not have as many hard images that capture the challenges presented by campus images. Increasing the size of the training dataset by including more hard images, as part of future work, can thus improve the overall performance in more general settings. It is also possible to use the classification time as a metric instead of human probability of correct classification.

Acknowledgments: This work was supported in part by NSF NeTS award #1321171.

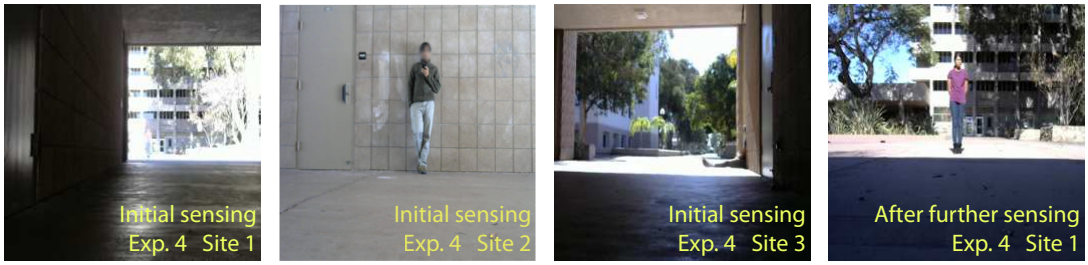

Fig. 10: Initial sensing images of 3 sites at crossroad 4 and the image taken after moving to Site 1 for further sensing based on our approach.

References

[1] NOAA Natural Hazards Image Database. http://www.ngdc.noaa.gov/hazardimages/.

[2] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome.

[3] Pioneer 3-AT Robot. http://www.mobilerobots.com/ResearchRobots/P3AT.aspx.

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *Proceedings of the European Conference on Computer Vision*, pages 438–451. 2010.

[5] H. Cai and Y. Mostofi. When human visual performance is imperfect - how to optimize the collaboration between one human operator and multiple field robots. In *Trends in Control and Decision-Making for Human-Robot Collaboration Systems*. Springer. To appear.

[6] H. Cai and Y. Mostofi. To ask or not to ask: A foundation for the optimization of human-robot collaborations. In *Proceedings of the American Control Conference*, pages 440–446, 2015.

[7] J. Capó-Aponte, L. Temme, H. Task, A. Pinkus, M. Kalich, A. Pantle, and C. Rash. Visual perception and cognitive performance. *Helmet-Mounted Displays: Sensation, Perception and Cognitive Issues*, pages 335–390, 2009.

[8] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, pages 853–867. 2005.

[9] Y. Chen, H. Shioi, C. Montesinos, L. Koh, S. Wich, and A. Krause. Active detection via adaptive submodularity. In *Proceedings of the International Conference on Machine Learning*, pages 55–63, 2014.

[10] M. B. Dias, B. Kannan, B. Browning, E. Jones, B. Argall, M. F. Dias, M. Zinck, M. Veloso, and A. Stentz. Sliding autonomy for peer-to-peer human-robot teams. In *Proceedings of the International Conference on Intelligent Autonomous Systems*, pages 332–341, 2008.

[11] M. P. Eckstein and J. S. Whiting. Visual signal detection in structured backgrounds I. Effect of number of possible spatial locations and signal contrast. *JOSA A*, 13(9): 1777–1787, 1996.

[12] M. P. Eckstein, A. J. Ahumada, and A. B. Watson. Visual signal detection in structured backgrounds II. Effects of contrast gain control, background variations, and white noise. *JOSA A*, 14(9):2406–2419, 1997.

[13] M. P. Eckstein, C. K. Abbey, and F. O. Bochud. A practical guide to model observers for visual detection in synthetic and natural noisy images. *Handbook of Medical Imaging*, 1:593–628, 2000.

[14] D. Engel and C. Curio. Pedestrian detectability: Predicting human perception performance with machine vision. In *IEEE Intelligent Vehicles Symposium*, pages 429–435, 2011.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[16] T. Fong, C. Thorpe, and C. Baur. Collaboration, dialogue, human-robot interaction. In *Robotics Research*, pages 255–266. 2003.

[17] G. Hoffman and C. Breazeal. Collaboration in human-robot teams. In *Proceedings of the AIAA Intelligent Systems Technical Conference, Chicago, IL, USA*, 2004.

[18] A. M. Howard. Role allocation in human-robot interaction schemes for mission scenario execution. In *Proceedings IEEE International Conference on Robotics and Automation*, pages 3588–3594, 2006.

[19] C. Huang, M. Cakmak, and B. Mutlu. Adaptive coordination strategies for human-robot handovers. In *Proceedings of Robotics: Science and Systems*, 2015.

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[21] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[23] R. Kurnia, M. A. Hossain, A. Nakamura, and Y. Kuno. Object recognition through human-robot interaction by speech. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pages 619–624, 2004.

[24] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. 2014.

[25] Ç. Meriçli, M. Veloso, and H. Akin. Task refinement for autonomous robots using complementary corrective human feedback. *International Journal of Advanced Robotic Systems*, 8(2):68, 2011.

[26] J. A. Ratches, R. H. Vollmerhausen, and R. G. Driggers. Target acquisition performance modeling of infrared imaging systems: past, present, and future. *IEEE Sensors Journal*, 1(1):31–40, 2001.

[27] S. Rosenthal and M. Veloso. Mobile robot planning to seek help with spatially-situated tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 4, page 1, 2012.

[28] P. Rouanet, P. Oudeyer, F. Danieau, and D. Filliat. The impact of human–robot interfaces on the learning of visual objects. *IEEE Transactions on Robotics*, 29(2): 525–541, 2013.

[29] 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, 115(3):211–252, 2015.

[30] O. Russakovsky, L. Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2121–2131, 2015.

[31] E. Sklar, S. L. Epstein, S. Parsons, A. T. Ozgelen, J. P. Munoz, and J. Gonzalez. A framework in which robots and humans help each other. In *Proceedings of the AAAI Spring Symposium: Help Me Help You: Bridging the Gaps in Human-Agent Collaboration*, 2011.

[32] A. Sorokin, D. Berenson, S. S. Srinivasa, and M. Hebert. People helping robots helping people: Crowdsourcing for grasping novel objects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2117–2122, 2010.

[33] V. Srivastava. *Stochastic search and surveillance strategies for mixed human-robot teams*. PhD thesis, University of California, Santa Barbara, 2012.

[34] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and systems*, 2014.

[35] R. Toris, H. Suay, and S. Chernova. A practical comparison of three robot learning from demonstration algorithms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 261–262, 2012.

[36] R. Toris, D. Kent, and S. Chernova. Unsupervised learning of multi-hypothesized pick-and-place task templates via crowdsourcing. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4504–4510, 2015.

[37] M. Wakayama, D. Deguchi, K. Doman, I. Ide, H. Murase, and Y. Tamatsu. Estimation of the human performance for pedestrian detectability based on visual search and motion features. In *Proceedings of the International Conference on Pattern Recognition*, pages 1940–1943, 2012.

[38] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[39] M. Yoshizaki, A. Nakamura, and Y. Kuno. Mutual assistance between speech and vision for human-robot interface. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 1308–1313, 2002.

[40] J. Zou and G. Nagy. Human-computer interaction for complex pattern recognition problems. In *Data Complexity in Pattern Recognition*, pages 271–286. 2006.