CONFIDENTIAL. Limited circulation. For review only.

Unsupervised Learning of Assistive Camera Views by an Aerial Co-robot in Augmented Reality Multitasking Environments

William Bentz, Sahib Dhanjal, and Dimitra Panagou

***Abstract*— This paper presents a novel method by which an assistive aerial robot can learn the relevant camera views within a task domain through tracking the head motions of a human collaborator. The human's visual field is modeled as an anisotropic spherical sensor, which decays in acuity towards the periphery, and is integrated in time throughout the domain. This data is resampled and fed into an expectation maximization solver in order to estimate the environment's visual interest as a mixture of Gaussians. A dynamic coverage control law directs the robot to capture camera views of the peaks of these Gaussians which is broadcast to an augmented reality display worn by the human operator. An experimental study is presented that assesses the influence of the assitive robot on reflex time, head motion, and task completion time.**

I. Introduction

On the eighth day of the STS-120 mission, the deployment of a newly installed solar panel array aboard the International Space Station (ISS) was abruptly halted when astronauts observed a tear between two panels via the Mobile Servicing System (MSS), a robotic arm outfitted with video cameras. With the station now underpowered and unfit to withstand external loads, e.g., docking and undocking, crew and ground controllers worked tirelessly for over 72 hours to plan and execute a spacewalk to repair the array. The improvised procedure was greatly dependent on MSS to transport an astronaut to the damage site, and then downlink external video of the repair to those assisting on the ground [1]. This case study illustrates the importance of effective human-robot collaboration, in part through assistive camera views, in sustaining situational awareness during spaceflight.

As NASA begins to integrate augmented reality (AR) into spacesuit helmets [2], [3] robotic cameras will be able to directly inform spacewalking astronauts without the need for real-time ground communication–a necessity for deep space missions beyond the Earth-Moon system. These missions will substantially benefit from some level of autonomy in the generation of assistive camera views as the aforementioned contemporary camera systems depend upon teleoperation–a task that is less than trivial when considering the difficulty in establishing a common frame of reference during space walks. Astronauts routinely change their frame of reference when communicating with ground controllers–often multiple times in a single sentence [4], [5]. A simple request such as "pan the camera up" can become incredibly difficult in an environment where directions are somewhat obscure.

The authors are with the Department of Aerospace Engineering, University of Michigan, Ann Arbor; wbentz@umich.edu, sdhanjal@umich.edu, and dpanagou@umich.edu.

The authors would like to acknowledge the support of an Early Career Faculty grant from NASAs Space Technology Research Grants Program

The novel contribution of this work is an unsupervised learning algorithm by which an aerial co-robot can stream assistive camera views, which are unknown *a priori*, to a human whose attention is split between an arbitrary number of complex tasks. This is accomplished by tracking the human's head motions during multitasking and then fitting this data to a *visual interest function*, modeled as a mixture of Gaussians, via online expectation maximization. This function informs a dynamic coverage controller [6], [7] which directs the robot to patrol the regions that are most visually interesting to the human. To the best of our knowledge, no prior work has considered this specific problem framework. Though motivated by its space applications, this algorithm also has the potential to compensate for age-related difficulties in multiple task monitoring, e.g., the simultaneous preparation of all elements of a meal [8], [9].

This paper is organized as follows: Section II overviews related works in the fields of augmented reality, UAV's, and online learning from human visual fields. Section III presents our visual acuity model as well as our robot kinematic and sensing models, Section IV describes how particle resampling and expectation maximization are employed to estimate the visual interest function, and Section V presents our control law derivation. Section VI presents our experimental results and conclusions and future directions are discussed in Section VII.

II. Related Work

A number of authors have already begun studying AR in the context of human-robot collaboration. In [10], the authors consider how visually displaying an autonomous robot's intent, i.e., waypoints and trajectories, can inform a human teammate in servicing spatially distributed tasks. This group also studied AR in the context of providing UAV pilots engaged in aerial inspection tasks with a visual display of the robot's current sensing field [11]. In these works, experimental results are presented through statistical analysis of both objective task performance metrics as well as survey-based subjective measurements of user satisfaction.

AR for supplementing a human's field of view via UAV cameras has been explored in [12]–[14]. In [12], aerial images of a construction site are captured from a static location and fused with virtual elements to aid in site management. Semi-autonomous structural inspection is considered in [13] where human head motions are used to perturb the UAV's automated reference trajectory online. The authors of [14] consider exocentric control of a UAV within an occluded environment. The vehicle effectively provides the operator

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

with "x-ray vision", i.e., they can see through walls in a virtual manner. A common thread in these works is that the human's actions are largely unproductive, i.e., they act primarily to direct the vehicle rather than engage in their own tasks.

Online learning from gaze and head tracking data as well as egocentric video streams are studied in [15]–[17]. In [15], the authors present an assistive co-robot which can aid quadriplegic individuals by retrieving desired objects. Gaze information directs the robot towards a search area where object recognition is achieved via a vocabulary tree (VT) based recognition algorithm [18]. In [16] and [17], the authors record egocentric video streams from humans participating in daily activities, e.g., preparing coffee, during a learning stage. Their algorithm detects task relevant objects and will display video clips of proper usage, obtained during the learning stage, to the headsets of future users interacting with the same objects. Instructional videos displayed via AR are also considered in [19]. For further context, please see the literature reviews on egocentric activity recognition [20] as well as human-robot collaboration facilitated through AR [21] and online learning [22].

III. Problem Formulation

Consider a convex domain $\mathcal{D} \subset \mathbb{R}^3$ which contains a set $\mathcal{I}$ of N visual regions of interest denoted $\mathcal{I}_i, \forall i \in \{1, ..., N\}$. These regions may contain individual tasks in the multitasking environment as well as any relevant information sources. A human H, occupying $\mathcal{D}$ and wearing an augmented reality headset, must split their attention between all the elements of $\mathcal{I}$ in order to achieve abstract multitasking objectives. A free-flying robot R, also occupying $\mathcal{D}$, is tasked with streaming live camera views of $\mathcal{I}_i$ to the headset in order to increase the efficiency by which H achieves their abstract objectives. Note that while H is required to provide R with a value for N, the precise locations of $\mathcal{I}_i$ may be unknown *a priori*. Our system geometry, in the context of our experimental setup, is illustrated in Fig. 1.

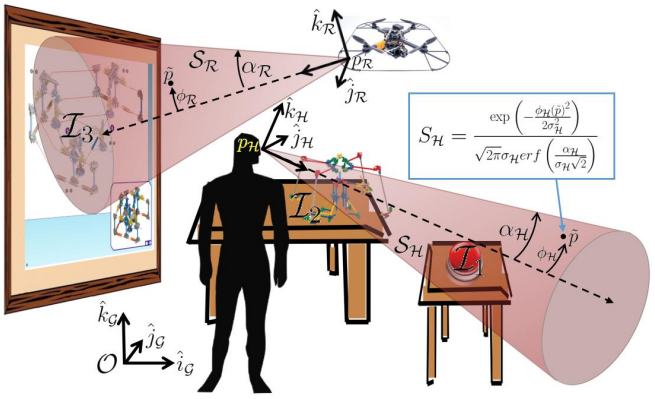

Fig. 1. In this sample scenario, the human is engaged in an assembly task that contains three visual regions of interest $\mathcal{I}$. Our model for human visual acuity at a given point $\tilde{p}$ is indicated.

A. Human Visual Acuity Model

We model the visual field of the human as a spherical sector $\mathcal{S}_\mathcal{H}$. The vertex of this sector is located at $p_\mathcal{H} = [x_\mathcal{H} \;\; y_\mathcal{H} \;\; z_\mathcal{H}]^T$ relative to the origin $\mathcal{O}$ of a global Cartesian coordinate frame $\mathcal{G}$. Note that all position vectors in the sequel are similarly taken relative to $\mathcal{O}$ and resolved in $\mathcal{G}$. We define the visual field frame $\mathcal{H}$ as having origin $p_\mathcal{H}$ and $\hat{i}_\mathcal{H}$ axis extending through the centerline of the sector. As $\mathcal{S}_\mathcal{H}$ is radially symmetric about $\hat{i}_\mathcal{H}$, $\hat{j}_\mathcal{H}$ may be chosen as any unit vector orthogonal to $\hat{i}_\mathcal{H}$ and $\hat{k}_\mathcal{H}$ completes the right-handed frame. The state vector of $\mathcal{S}_\mathcal{H}$ may thus be defined as $\bar{q}_\mathcal{H} = [p_\mathcal{H} \;\; \Phi_\mathcal{H} \;\; \Theta_\mathcal{H} \;\; \Psi_\mathcal{H}]^T$ where the latter 3 states are the 3-2-1 Euler angles of $\mathcal{H}$ relative to $\mathcal{G}$.

Human visual acuity is anisotropic in nature, e.g., it tends to degrade in quality towards one's periphery. Therefore, assuming that visual regions lying along $\hat{i}_\mathcal{H}$ are of the greatest interest, we define our model for visual acuity as:

$$S_\mathcal{H}\left(\bar{q}_\mathcal{H}, \tilde{p}\right) = \frac{\exp\left(-\frac{\phi_\mathcal{H}(\tilde{p})^2}{2\sigma_\mathcal{H}^2}\right)}{\sqrt{2\pi}\sigma_\mathcal{H}\,\mathrm{erf}\left(\frac{\alpha_\mathcal{H}}{\sigma_\mathcal{H}\sqrt{2}}\right)}, \tag{1}$$

where $\tilde{p}_i = [\tilde{x} \;\; \tilde{y} \;\; \tilde{z}]^T$ is the position of a point within $\mathcal{S}_\mathcal{H}$ with respect to $\mathcal{G}$ and $\phi_\mathcal{H}(\tilde{p})$ is angle between $\tilde{p}_i - p_\mathcal{H}$ and $\hat{i}_\mathcal{H}$. We define (1) using the probability density function associated with a zero-mean truncated normal distribution taking nonzero values between $\pm\alpha_\mathcal{H}$ and whose underlying Gaussian has standard deviation $\sigma_\mathcal{H}$. We chose this truncated Gaussian as $\alpha_\mathcal{H} = 60°$ and $\sigma_\mathcal{H} = 8°$ provide a relatively good fit for the clinically-inspired plots of visual acuity with respect to angle from fovea that are presented in sensory physiology textbooks [23], [24]. Furthermore, it enforces that visual acuity beyond the periphery bound, i.e. $\alpha_\mathcal{H}$, is zero. Note that although visual acuity reduces with respect to distance as well, we assume that the human can comfortably focus upon all objects in $\mathcal{D}$ and thus this dependence is omitted.

As H shifts their attention between the elements of $\mathcal{I}$, the volume of space occupying $\mathcal{S}_H$ varies. Thus, the accumulated visual acuity is defined as:

$$Q\left(t, \tilde{p}\right) = \int_0^t S_\mathcal{H}\left(\bar{q}_\mathcal{H}\left(\tau\right), \tilde{p}\right) d\tau, \tag{2}$$

where the time dependence of $q_\mathcal{H}$, nominally omitted to reduce clutter, is notated here for clarity.

B. Robot Kinematic and Sensor Model

We have adopted the robot kinematic and sensor model of our prior work [7]. Let us assume that R is equipped with a camera whose sensing footprint is a spherical sector $\mathcal{S}_\mathcal{R}$ similar in shape, though not in quality of sensing, to $\mathcal{S}_\mathcal{H}$. $\mathcal{S}_\mathcal{R}$ is subject to 3-D rigid body kinematics [25]:

$$\begin{bmatrix} \dot{x}_\mathcal{R} \\ \dot{y}_\mathcal{R} \\ \dot{z}_\mathcal{R} \end{bmatrix} = \begin{bmatrix} \cos\Theta_\mathcal{R}\cos\Psi_\mathcal{R} & \sin\Phi_\mathcal{R}\sin\Theta_\mathcal{R}\cos\Psi_\mathcal{R} - \cos\Phi_\mathcal{R}\sin\Psi_\mathcal{R} & \cos\Phi_\mathcal{R}\sin\Theta_\mathcal{R}\cos\Psi_\mathcal{R} + \sin\Phi_\mathcal{R}\sin\Psi_\mathcal{R} \\ \cos\Theta_\mathcal{R}\sin\Psi_\mathcal{R} & \sin\Phi_\mathcal{R}\sin\Theta_\mathcal{R}\sin\Psi_\mathcal{R} + \cos\Phi_\mathcal{R}\cos\Psi_\mathcal{R} & \cos\Phi_\mathcal{R}\sin\Theta_\mathcal{R}\sin\Psi_\mathcal{R} - \sin\Phi_\mathcal{R}\cos\Psi_\mathcal{R} \\ -\sin\Theta_\mathcal{R} & \sin\Phi_\mathcal{R}\cos\Theta_\mathcal{R} & \cos\Phi_\mathcal{R}\cos\Theta_\mathcal{R} \end{bmatrix} \begin{bmatrix} u_\mathcal{R} \\ v_\mathcal{R} \\ w_\mathcal{R} \end{bmatrix}, \tag{3}$$

$$\begin{bmatrix}\dot{\Phi}_{\mathcal{R}}\\ \dot{\Theta}_{\mathcal{R}}\\ \dot{\Psi}_{\mathcal{R}}\end{bmatrix}=\begin{bmatrix}1 & \sin\Phi_{\mathcal{R}}\tan\Theta_{\mathcal{R}} & \cos\Phi_{\mathcal{R}}\tan\Theta_{\mathcal{R}}\\ 0 & \cos\Phi_{\mathcal{R}} & -\sin\Phi_{\mathcal{R}}\\ 0 & \sin\Phi_{\mathcal{R}}\sec\Theta_{\mathcal{R}} & \cos\Phi_{\mathcal{R}}\sec\Theta_{\mathcal{R}}\end{bmatrix}\begin{bmatrix}q_{\mathcal{R}}\\ r_{\mathcal{R}}\\ s_{\mathcal{R}}\end{bmatrix}, \quad (4)$$

where $p_{\mathcal{R}} = [x_{\mathcal{R}}\ y_{\mathcal{R}}\ z_{\mathcal{R}}]^T$ is the position of the vertex of $\mathcal{S}_{\mathcal{R}}$. The state vector of $\mathcal{S}_{\mathcal{R}}$ is thus $\bar{q}_{\mathcal{R}} = [p_{\mathcal{R}}\ \Phi_{\mathcal{R}}\ \Theta_{\mathcal{R}}\ \Psi_{\mathcal{R}}]^T$ and the latter 3 states are the 3-2-1 Euler angles of the sector's body-fixed frame $\mathcal{R}$ relative to $\mathcal{G}$. $\mathcal{R}$ has origin $p_{\mathcal{R}}$ and $\hat{\imath}_{\mathcal{R}}$ axis extending through the centerline of $\mathcal{S}_{\mathcal{R}}$. $[u_{\mathcal{R}}\ v_{\mathcal{R}}\ w_{\mathcal{R}}]^T$ and $[q_{\mathcal{R}}\ r_{\mathcal{R}}\ s_{\mathcal{R}}]^T$ are respectively the linear and angular velocities of $\mathcal{S}_{\mathcal{R}}$ resolved in $\mathcal{R}$. As with $\mathcal{S}_{\mathcal{H}}$, $\mathcal{S}_{\mathcal{R}}$ is radially symmetric and thus $\hat{\jmath}_{\mathcal{R}}$ may be chosen as any unit vector orthogonal to $\hat{\imath}_{\mathcal{R}}$ with $\hat{k}_{\mathcal{R}}$ completing the right-handed frame.

The quality of sensing associated with $\mathcal{S}_{\mathcal{R}}$ is also anisotropic in nature; however, its angular degradation is not nearly as rapid as (1). As in [7], we consider the following sensing constraint functions:

$$c_1 = \beta R_{\mathcal{R}}^2 - (\tilde{x}-x_{\mathcal{R}})^2 - (\tilde{y}-y_{\mathcal{R}})^2 - (\tilde{z}-z_{\mathcal{R}})^2, \quad (5a)$$

$$c_2 = \alpha_{\mathcal{R}} - \phi_{\mathcal{R}}(\tilde{p}), \quad (5b)$$

for $\beta = \min\{1, \mu\left((\tilde{x}-x_{\mathcal{R}})^2 + (\tilde{y}-y_{\mathcal{R}})^2 - (\tilde{z}-z_{\mathcal{R}})^2\right)\}$ with real constant $\mu >> 1$. $R_{\mathcal{R}}$ is the sensing range, and $\alpha_{\mathcal{R}}$ and $\phi_{\mathcal{R}}(\tilde{p})$ are defined in the same manner as their counterparts in $\mathcal{S}_{\mathcal{H}}$. Let us denote $\max\{0, c_k\} = C_k$. One can define the quality of information available at each point over $\mathcal{S}_{\mathcal{R}}$ as:

$$S_{\mathcal{R}}(\bar{q}_{\mathcal{R}}, \tilde{p}) = \begin{cases} \frac{C_1 C_2}{C_1 + C_2}, & \text{if } card\left(\bar{C}\right) < 2 \wedge \|\tilde{p} - p_{\mathcal{R}}\| > 0; \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\bar{C}$ is the set of zero elements in C_k. $S_{\mathcal{R}}(\tilde{q}_{\mathcal{R}}, \tilde{p})$ takes a value of zero outside of $\mathcal{S}_{\mathcal{R}}$. Note that $S_{\mathcal{R}}(\tilde{q}_{\mathcal{R}}, \tilde{p})$ is defined over all of $\mathcal{D}$ and thus has static bounds. $S_{\mathcal{R}}(\tilde{q}_{\mathcal{R}}, \tilde{p})$ is continuous in $\tilde{p}$ while taking a value of zero along the sector's boundary $\partial\mathcal{S}_{\mathcal{R}}$. In verifying this continuity, it is important to note that $S_{\mathcal{R}}(\tilde{q}_{\mathcal{R}}, \tilde{p})$ approaches zero from within $\mathcal{S}_{\mathcal{R}}$ in the limit that either $card\left(\bar{C}\right) = 2$ or $\|\tilde{p} - p_{\mathcal{R}}\| = 0$ are satisfied. The former condition may be verified by taking a limit of the first piecewise definition of (6) as C_1 and C_2 tend to zero. The latter condition results from our definition of β.

Remark 1: The purpose of this work is to derive a control strategy to facilitate persistent monitoring of the regions of space occupied by $\mathcal{I}_i$, which are unknown *a priori*. We aim to evaluate whether or not the increased situational awareness of H can improve performance, e.g., through reductions in task completion time and physical effort. This must be accomplished in a safe manner that guarantees collision avoidance between R and H.

IV. Online Learning of Relevant Views

One cannot draw immediate conclusions with respect to the locations of $\mathcal{I}_i$ directly from integration of raw visual acuity data in (2). Instead, some form of online clustering is necessary in order to guide the motion of R. As our control strategy shall be gradient-based, the natural choice for clustering is expectation-maximization of a Gaussian mixture model (EM-GMM). This approach allows for the raw data, upon resampling, to be converted into a mixture of Gaussians whose gradient is smooth and defined over all of $\mathcal{D}$.

When implemented in software, let us assume that values for $Q(t, \tilde{p})$ are computed at a finite number of discrete points $\tilde{P} \subset \mathcal{D}$ on a three dimensional grid. We consider the values of Q for each element of $\tilde{P}$, i.e., each sample of $\tilde{P}$, to be an *importance weight* associated with that *sample*. This terminology is consistent with descriptions of the importance resampling step of the particle filter as presented in [26]. Using importance resampling, we select elements of $\tilde{P}$ at random where the likelihood of selection is proportional to the weight of the sample. The selected elements, which can include multiple copies of individual points, form the input for our EM-GMM implementation.

We employ the *select with replacement* algorithm as it is among the most common resampling methods used in particle filters [27]. The algorithm generates $M = card\left(\tilde{P}\right)$ samples from a uniform distribution in $(0, 1)$. This set, T, is sorted into ascending order and then augmented with an additional element equal to 1. A second array, Q_c, is then defined as the cumulative sum of the normalized elements of $Q\left(t_r, \tilde{P}\right)$ where t_r is the time of resampling. Starting with $Q_c[1]$, each ascending element of T is compared against $Q_c[1]$ with $\tilde{P}[1]$ selected for output repeatedly until an element of T is found that exceeds $Q_c[1]$. The process resumes with $Q_c[2]$ until M points have been chosen. The effect is that elements of $\tilde{P}$ with small values for Q are removed. This is described in full detail in Algorithm 1 where the array of filtered points is denoted $\tilde{P}_f$. The EM-GMM implementation requires N and $\tilde{P}_f$ as inputs in order

Algorithm 1 Select with Replacement

Inputs: $\tilde{P}, Q\left(t, \tilde{P}\right)$
Initialize: $i \leftarrow 1, j \leftarrow 1,\ M \leftarrow card\left(\tilde{P}\right)$
while $i < M + 1$ **do**
  if $i < 2$ **then**
    $Q_c[i] \leftarrow \frac{Q\left(t_r, \tilde{P}[i]\right)}{\sum_{i=1}^{M} Q\left(t_r, \tilde{P}[i]\right)}$, $T[i] \leftarrow rand(0, 1)$
  else
    $Q_c[i] \leftarrow Q_c[i-1] + \frac{Q\left(t_r, \tilde{P}[i]\right)}{\sum_{i=1}^{M} Q\left(t_r, \tilde{P}[i]\right)}$
    $T[i] \leftarrow rand(0, 1)$
  end if
end while
$T \leftarrow sort(T)$, $T[M+1] \leftarrow 1.0$, $i \leftarrow 1$
while $i < M + 1$ **do**
  if $T[i] < Q_c[j]$ **then**
    $\tilde{P}_f[i] \leftarrow \tilde{P}[j]$, $i \leftarrow i + 1$
  else
    $j \leftarrow j + 1$
  end if
end while
return $\tilde{P}_f$

to estimate our human visual interest function:

$$\bar{\psi}_{\mathcal{H}} = \sum_{i=1}^{N} \pi_i \mathcal{N}\left(\mu_i, \Sigma_i\right), \tag{7}$$

where π_i and μ_i are the mixing coefficients and means outputted from the EM-GMM solver. Note that Σ_i is not necessarily the covariance outputted from EM-GMM. We found in practice that values for Σ_i tended to be very small and thus allowing for these to be increased as tuning parameters had the effect of amplifying the gradient to which R follows towards μ_i. A description of EM-GMM is available for reference in [28]. An example $\bar{\psi}$ associated with the task scenario in Fig. 1 is presented in Fig. 2.

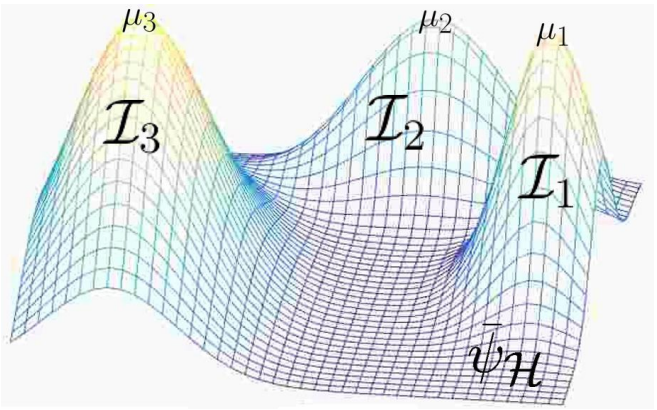

Fig. 2. This mixture of Gaussians corresponds to a cross section of the task scenario presented in Fig. 1 with mixing coefficients $\pi_1 \approx 0.35$, $\pi_2 \approx 0.30$, and $\pi_3 \approx 0.35$.

As it would be redundant for R to observe the element of $\mathcal{I}$ that presently holds the attention of H, it is necessary to correct $\bar{\psi}_{\mathcal{H}}$ such that it may drop to zero inside of the current region of space occupied by $\mathcal{S}_{\mathcal{H}}$. One simple solution would be to scale $\bar{\psi}_{\mathcal{H}}$ by a Heaviside step function with crossover point at $\phi_{\mathcal{H}} = \alpha_{\mathcal{H}}$; however, this would introduce an undesirable discontinuity into the derivation of our control laws. Instead, we adopt the logistic function as a smooth alternative to Heaviside step in our corrected visual interest function:

$$\psi_{\mathcal{H}} = \frac{\bar{\psi}_{\mathcal{H}}}{1 + \exp\left(-k\left(\phi_{\mathcal{H}} - \alpha_{\mathcal{H}}\right)\right)}, \tag{8}$$

where choosing $k >> 1$ increases the barrier steepness.

V. Controller Design

Driving $\mathcal{S}_{\mathcal{R}}$ to encompass those regions of $\mathcal{D}$ for which $\psi_{\mathcal{H}}$ has a higher value is akin to reducing the following cost function:

$$J\left(t\right) = \int_{\mathcal{D}} \max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)^2 d\tilde{p}, \tag{9}$$

where the integrand is continuously differentiable. Employing the Leibniz integral rule, one can differentiate (9) with respect to time to get $\dot{J}\left(t\right) = \int_{\mathcal{D}} \frac{\partial}{\partial t} \max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)^2 d\tilde{p} + \int_{\partial\mathcal{D}} \max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)^2 \mathbf{v} \cdot \mathbf{n} dA$ where the latter term is an area integral along the boundary of the domain $\partial\mathcal{D}$. $\mathbf{n}$ is the unit vector normal to $\partial\mathcal{D}$ while $\mathbf{v}$ is the velocity of $\partial\mathcal{D}$. As $\mathcal{D}$ is stationary, $\mathbf{v} = 0$ for all points along $\partial\mathcal{D}$ and thus we have $\dot{J}\left(t\right) = \int_{\mathcal{D}} \frac{\partial}{\partial t} \max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)^2 d\tilde{p}$ which expands to:

$$\dot{J}\left(t\right) = \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\left(\frac{\partial S_{\mathcal{R}}}{\partial t} - \frac{\partial \psi_{\mathcal{H}}}{\partial t}\right) d\tilde{p}. \tag{10}$$

Employing the chain rule, we have $\frac{\partial S_{\mathcal{R}}}{\partial t} = \frac{\partial S_{\mathcal{R}}}{\partial x_{\mathcal{R}}}\dot{x}_{\mathcal{R}} + \frac{\partial S_{\mathcal{R}}}{\partial y_{\mathcal{R}}}\dot{y}_{\mathcal{R}} + \frac{\partial S_{\mathcal{R}}}{\partial z_{\mathcal{R}}}\dot{z}_{\mathcal{R}} + \frac{\partial S_{\mathcal{R}}}{\partial \Phi_{\mathcal{R}}}\dot{\Phi}_{\mathcal{R}} + \frac{\partial S_{\mathcal{R}}}{\partial \Theta_{\mathcal{R}}}\dot{\Theta}_{\mathcal{R}} + \frac{\partial S_{\mathcal{R}}}{\partial \Psi_{\mathcal{R}}}\dot{\Psi}_{\mathcal{R}}$. This expression may be rewritten in terms of the body-fixed linear and angular velocities by substituting the expressions in (3) and (4) for the elements of $\dot{\tilde{q}}_{\mathcal{R}}$:

$$\begin{aligned}\frac{\partial S_{\mathcal{R}}}{\partial t} = &\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial x_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial y_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial z_{\mathcal{R}}}\end{bmatrix} M^{\ell}\begin{bmatrix}u_{\mathcal{R}} & v_{\mathcal{R}} & w_{\mathcal{R}}\end{bmatrix}^T \\ &+ \begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial \Phi_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Theta_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Psi_{\mathcal{R}}}\end{bmatrix} M^{a}\begin{bmatrix}q_{\mathcal{R}} & r_{\mathcal{R}} & s_{\mathcal{R}}\end{bmatrix}^T,\end{aligned} \tag{11}$$

where we denote the 3×3 matrices in (3) and (4) as M^{ℓ} and M^{a} respectively. Substituting (11) into (10) we have:

$$\begin{aligned}\dot{J}\left(t\right) = -a_0\left(t\right) + &\begin{bmatrix}u_{\mathcal{R}}\left(t\right) & v_{\mathcal{R}}\left(t\right) & w_{\mathcal{R}}\left(t\right) & q_{\mathcal{R}}\left(t\right) & r_{\mathcal{R}}\left(t\right) & s_{\mathcal{R}}\left(t\right)\end{bmatrix} \\ &\begin{bmatrix}a_1\left(t\right) & a_2\left(t\right) & a_3\left(t\right) & a_4\left(t\right) & a_5\left(t\right) & a_6\left(t\right)\end{bmatrix}^T,\end{aligned} \tag{12}$$

where:

$$\begin{aligned}
a_0\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\frac{\partial \psi_{\mathcal{H}}}{\partial t} d\tilde{p} \\
a_1\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial x_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial y_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial z_{\mathcal{R}}}\end{bmatrix} M^{\ell}\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T d\tilde{p}, \\
a_2\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial x_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial y_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial z_{\mathcal{R}}}\end{bmatrix} M^{\ell}\begin{bmatrix}0 & 1 & 0\end{bmatrix}^T d\tilde{p}, \\
a_3\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial x_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial y_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial z_{\mathcal{R}}}\end{bmatrix} M^{\ell}\begin{bmatrix}0 & 0 & 1\end{bmatrix}^T d\tilde{p}, \\
a_4\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial \Phi_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Theta_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Psi_{\mathcal{R}}}\end{bmatrix} M^{a}\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T d\tilde{p}, \\
a_5\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial \Phi_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Theta_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Psi_{\mathcal{R}}}\end{bmatrix} M^{a}\begin{bmatrix}0 & 1 & 0\end{bmatrix}^T d\tilde{p}, \\
a_6\left(t\right) &= \int_{\mathcal{D}} 2\max\left(0, S_{\mathcal{R}} - \psi_{\mathcal{H}}\right)\begin{bmatrix}\frac{\partial S_{\mathcal{R}}}{\partial \Phi_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Theta_{\mathcal{R}}} & \frac{\partial S_{\mathcal{R}}}{\partial \Psi_{\mathcal{R}}}\end{bmatrix} M^{a}\begin{bmatrix}0 & 0 & 1\end{bmatrix}^T d\tilde{p}.
\end{aligned}$$

If the only objective were to guarantee $\dot{J}\left(t\right) \leq 0$, one could achieve this with:

$$\bar{u}_{\mathcal{R}} = -K_u\left(a_1\left(t\right) - a_0\left(t\right)a_1\left(t\right)^{-1}K^{-1}\right), \tag{13a}$$

$$\bar{v}_{\mathcal{R}} = -K_v\left(a_2\left(t\right) - a_0\left(t\right)a_2\left(t\right)^{-1}K^{-1}\right), \tag{13b}$$

$$\bar{w}_{\mathcal{R}} = -K_w\left(a_3\left(t\right) - a_0\left(t\right)a_3\left(t\right)^{-1}K^{-1}\right), \tag{13c}$$

$$q_{\mathcal{R}} = -K_q\left(a_4\left(t\right) - a_0\left(t\right)a_4\left(t\right)^{-1}K^{-1}\right), \tag{13d}$$

$$r_{\mathcal{R}} = -K_r\left(a_5\left(t\right) - a_0\left(t\right)a_5\left(t\right)^{-1}K^{-1}\right), \tag{13e}$$

$$s_{\mathcal{R}} = -K_s\left(a_6\left(t\right) - a_0\left(t\right)a_6\left(t\right)^{-1}K^{-1}\right), \tag{13f}$$

where $K_{()} > 0$ are tunable gains and K is the sum of all $K_{()}$ gains. The terms subtracted within the parentheses in (13) have the effect of cancelling out $a_0\left(t\right)$ in (12). This may be physically interpreted as allowing for the agents to respond to not only the current shape of $\psi_{\mathcal{H}}$ but also its rate of change as well, e.g., a sudden rotation of $\mathcal{S}_{\mathcal{H}}$ (whose boundary encodes a smooth zero barrier in $\psi_{\mathcal{H}}$) towards R would drive R away from $\mathcal{S}_{\mathcal{H}}$ with a portion of its input proportional to $\frac{\partial \psi_{\mathcal{H}}}{\partial t}$.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

Note that ensuring $\dot{J}(t) \leq 0$ is not sufficient as we must also encode collision avoidance with respect to $p_{\mathcal{H}}$ and the most likely positions of $\mathcal{I}_i$. We accomplish this through the addition of inverse barriers in our translational control laws:

$$u_{\mathcal{R}} = \bar{u}_{\mathcal{R}} + \hat{i}_{\mathcal{R}} \cdot \left(\rho_{\mathcal{H}} + \sum_{i=1}^{N} \rho_{\mathcal{I}_i} \right), \tag{14a}$$

$$v_{\mathcal{R}} = \bar{v}_{\mathcal{R}} + \hat{j}_{\mathcal{R}} \cdot \left(\rho_{\mathcal{H}} + \sum_{i=1}^{N} \rho_{\mathcal{I}_i} \right), \tag{14b}$$

$$w_{\mathcal{R}} = \bar{w}_{\mathcal{R}} + \hat{k}_{\mathcal{R}} \cdot \left(\rho_{\mathcal{H}} + \sum_{i=1}^{N} \rho_{\mathcal{I}_i} \right), \tag{14c}$$

with:

$$\rho_{\ell} = \left(\kappa \left(\|p_{\mathcal{R}} - p_{\mathcal{H}}\|^2 - d_{\mathcal{H}} \|p_{\mathcal{R}} - p_{\mathcal{H}}\| \right) \right)^{-1} \left(M^{\ell} \right)^{-1} (p_{\mathcal{R}} - p_{\mathcal{H}}),$$

$$\rho_{\mathcal{I}_i} = \left(\kappa \left(\|p_{\mathcal{R}} - \mu_i\|^2 - d_i \|p_{\mathcal{R}} - \mu_i\| \right) \right)^{-1} \left(M^{\ell} \right)^{-1} (p_{\mathcal{R}} - \mu_i).$$

where κ is a tuning parameter for the steepness of the barrier while $d_{\mathcal{H}}$ and d_i are minimum safe distances of R from $p_{\mathcal{H}}$ and mu_i respectively. In the limit that R approaches either H or any of the elements of $\mathcal{I}$, the magnitude (14) approaches ∞ in the direction of the vector pointing from the obstacle to R.

VI. Experimental Results

We conducted an experimental trial in which ten volunteers, all students between the ages of 18-26, were asked to assemble K'NEX toy building kits on a lab bench with the instructions posted on a board to the left side of the task environment. The subjects were also asked to monitor a secondary task placed on a bench behind them and to their right. This secondary task focused upon a button which would illuminate at random times as generated by a Poisson counting process. They were expected to continually monitor the button in order to press it as quickly as possible upon illumination while simultaneously attempting to complete their primary task. The set up is illustrated in Fig. 1.

In all trials, the subjects wore a bicycle helmet with motion capture markers, from which our VICON motion capture system estimated head pose data, as well as a Vufine wearable display. This inexpensive AR set allows for a transparent video feed to be placed over the subject's visual field. An outfitted subject is illustrated in Fig. 3. Each subject completed two K'NEX kits (sailboat and rain cloud) both with and without the aid of the UAV video feed thus resulting in a total of four trials per subject. The video feed was provided by an Ascending Technologies Hummingbird which was controlled via the kinematic commands derived in Section V. During each experimental trial, R was launched 30 seconds after H began working. The evolution of $\psi_{\mathcal{H}}(t)$, which we recompute every 0.5 seconds in parallel via CUDA [29], is presented in Fig. 4.

Subjects will undoubtedly improve in assembly time on their second attempt at any given K'NEX model. To account for this, half of the subjects completed task sequence A {sailboat, rain cloud, sailboat, rain cloud}, and half of the subjects completed task sequence B {rain cloud, sailboat, rain cloud, sailboat} while the trial type sequence was always administered as {experimental, control, control, experimental}. In this manner, we can compute the average time until completion for both the sailboat and raincloud with and without the aid of augmented reality. Given a statistically significant number of subjects, the improvement due to task experience alone should average out. Statistics on assembly time and secondary task reaction time are presented in Table I. The mean time required to complete the assembly task was increased in the range of 10-16 seconds. However, this expense seemed to be at the benefit of the secondary task's reaction time which was reduced on average by 0.47 seconds per subject.

Fig. 3. Subjects wore a Vufine wearable display, a bicycle helmet with motion capture markers, and interacted with an Asctec Hummingbird.

TABLE I
Assembly and Reaction Time Statistics

	Mean Assembly Time (sec)	Std. Dev. (sec)	Mean Reaction Time (sec)	Std. Dev. (sec)
Boat Exp.	161.1	38.6	5.7	4.5
Boat Ctrl.	151.0	49.3	6.0	4.6
Cloud Exp.	190.6	46.1	5.4	4.0
Cloud Ctrl.	174.0	45.7	5.9	4.1

We also aim to assess how the robotic assistant effects physical exertion by the human. During the first 100 seconds of multitasking, subjects showed an average reduction in integrated head motion of 2.23 radians. Trends for three subjects are presented in Fig. 5. An additional note of interest is that all but one of the subjects had a background in engineering and computer science. The outlier, an artist referred to as Subject 2 in Fig. 5, found the augmented reality the least helpful. This implies that the performance of our algorithm may be highly dependent upon the technical background of the human subject.

VII. Conclusions

In this paper, we presented a novel method by which an assistive aerial robot can learn the relevant camera views within a domain in order to assist a human collaborator in completing multitasking objectives. We integrated the

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

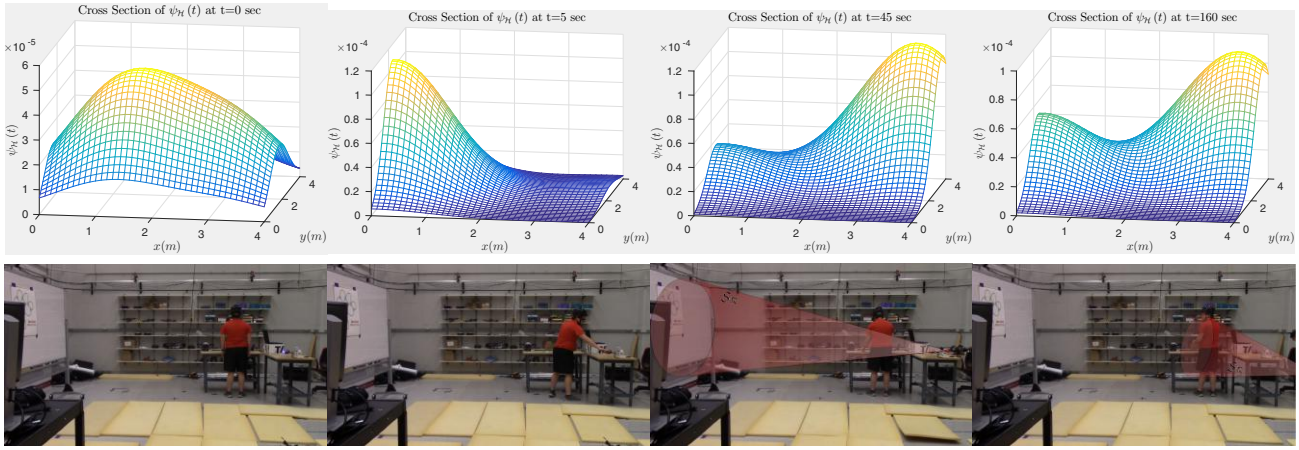

Fig. 4. For the experimental trial picture above, we have plotted the evolution of a 1.3 meter height cross section of $\psi_{\mathcal{H}}(t)$ taken at four distinct time instances. Upon initialization of the algorithm, the human's head motions indicate that substantial weight should be placed upon the set of task instructions on the left side of the environment. Through the progression of the trial, the weight becomes more balanced with the secondary buzzer task located on the right side of the environment.

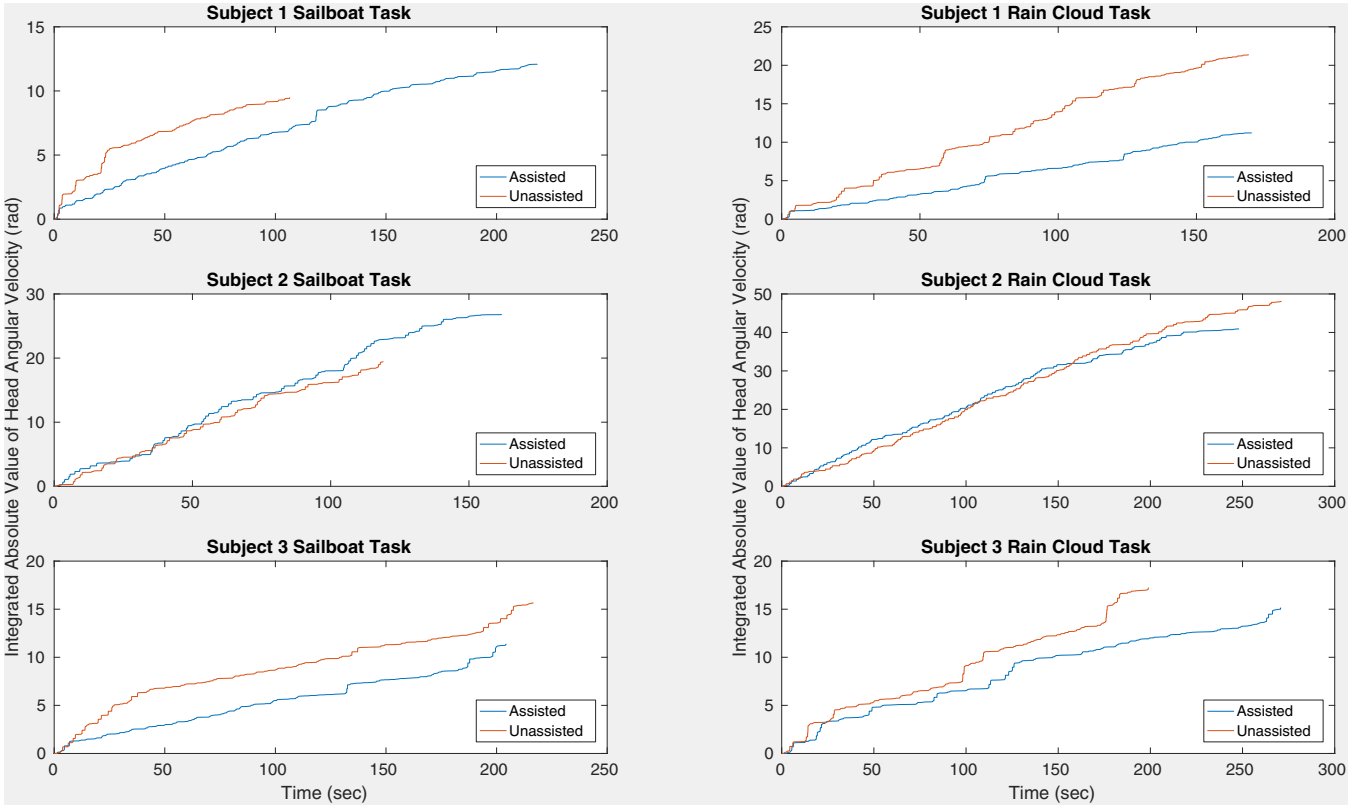

Fig. 5. Subjects 1 and 3 both showed substantial reductions in head motion during their assembly tasks. It is interesting to note that these two subjects study engineering and computer science while Subject 2 is an artist.

human's visual field and used EM-GMM to compute the environment's visual interest function. We derived a gradient-based dynamic coverage controller which directed the aerial robot to observe those regions of the domain which were most interesting to the human and broadcast these views to the human's augmented reality display. This resulted in reduced head motions on the part of the human as well as improved reaction time. However, it also increased the time required to complete the human's primary objective. In our future work, we will aim depth cameras at the environment and eye trackers at the human to more precisely characterize the locations of visual interest. We will also gather survey data to assess potential correlations with respect to academic background or familiarity with video games.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

REFERENCES

[1] S. Aziz, "Lessons learned from the STS-120/ISS 10A robotics operations," *Acta Astronautica*, vol. 66, no. 1-2, pp. 157–165, 2010.

[2] C. E. Carr, S. J. Schwartz, and I. Rosenberg, "A wearable computer for support of astronaut extravehicular activity," in *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on*. IEEE, 2002, pp. 23–30.

[3] O. Doule, D. Miranda, and J. Hochstadt, "Integrated display and environmental awareness system-system architecture definition," in *AIAA SPACE and Astronautics Forum and Exposition*, 2017, p. 5269.

[4] C. K. L. F. J. S. R. A. R. B. R. S. L. H. A. S. J. G. T. M. B. Terrence Fong, Illah Nourbakhsh and J. Scholtz, "The peer-to-peer human-robot interaction project," in *SPACE 2005*, Long Beach, CA, 2005.

[5] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, July 2005.

[6] W. Bentz and D. Panagou, "Persistent coverage of a two-dimensional manifold subject to time-varying disturbances," in *Proc. of the 56th IEEE Conference on Decision and Control*, Melbourne, Australia, Dec. 2017. [Online]. Available: http://www-personal.umich.edu/~dpanagou/assets/documents/WBentz_CDC17.pdf

[7] W. Bentz, T. Hoang, E. Bayasgalan, and D. Panagou, "Complete 3-D dynamic coverage in energy-constrained multi-UAV sensor networks," *Autonomous Robots*, pp. 1–27, 2017.

[8] I. Todorov, F. Del Missier, and T. Mntyl, "Age-related differences in multiple task monitoring," *PLOS ONE*, vol. 9, no. 9, pp. 1–7, 09 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0107619

[9] F. I. M. Craik and E. Bialystok, "Planning and task management in older adults: Cooking breakfast," *Memory & Cognition*, vol. 34, no. 6, pp. 1236–1249, Sep 2006. [Online]. Available: https://doi.org/10.3758/BF03193268

[10] M. Walker, H. Hedayati, J. Lee, and D. Szafir, "Communicating robot motion intent with augmented reality," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 316–324.

[11] H. Hedayati, M. Walker, and D. Szafir, "Improving collocated robot teleoperation with augmented reality," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 78–86.

[12] M.-C. Wen and S.-C. Kang, "Augmented reality and unmanned aerial vehicle assist in construction management," in *2014 International Conference on Computing in Civil and Building Engineering*. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/9780784413616.195

[13] C. Papachristos and K. Alexis, "Augmented reality-enhanced structural inspection using aerial robots," in *2016 IEEE International Symposium on Intelligent Control (ISIC)*, Buenos Aires, Argentina, Sept 2016.

[14] O. Erat, W. A. Isop, D. Kalkofen, and D. Schmalstieg, "Drone-augmented human vision: Exocentric control for drones exploring hidden areas," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, pp. 1437–1446, April 2018.

[15] J. Zhang, L. Zhuang, Y. Wang, Y. Zhou, Y. Meng, and G. Hua, "An egocentric vision based assistive co-robot," in *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–7.

[16] D. Damen, O. Haines, T. Leelasawassuk, A. Calway, and W. Mayol-Cuevas, "Multi-user egocentric online system for unsupervised assistance on object usage," in *European Conference on Computer Vision*. Springer, 2014, pp. 481–492.

[17] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas, "You-Do, I-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance," *Computer Vision and Image Understanding*, vol. 149, pp. 98–112, 2016.

[18] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. Ieee, 2006, pp. 2161–2168.

[19] M. Goto, Y. Uematsu, H. Saito, S. Senda, and A. Iketani, "Task support system by displaying instructional video onto AR workspace," in *2010 IEEE International Symposium on Mixed and Augmented Reality*, Oct 2010, pp. 83–90.

[20] T.-H.-C. Nguyen, J.-C. Nebel, F. Florez-Revuelta *et al.*, "Recognition of activities of daily living with egocentric vision: A review," *Sensors*, vol. 16, no. 1, 2016.

[21] S. A. Green, M. Billinghurst, X. Chen, and J. G. Chase, "Human-robot collaboration: A literature review and augmented reality approach in design," *International journal of advanced robotic systems*, vol. 5, no. 1, pp. 1–18, 2008.

[22] M. A. Z. Hernandez, E. C. Marin, J. Garcia-Rodriguez, J. Azorin-Lopez, and M. Cazorla, "Automatic learning improves human-robot interaction in productive environments: A review," *International Journal of Computer Vision and Image Processing (IJCVIP)*, vol. 7, no. 3, pp. 65–75, 2017.

[23] R. F. Schmidt, *Fundamentals of Sensory Physiology*, 3rd ed. Springer-Verlag Berlin Heidelberg, 1986, p. 159.

[24] H. Hunziker, *Im Auge Des Lesers [The eye of the reader: foveal and peripheral perception - from letter recognition to the joy of reading] (in German)*. Zürich: Transmedia Stubli Verlag, 2006.

[25] R. W. Beard, "Quadrotor dynamics and control," 2008, lecture notes. [Online]. Available: http://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2324&context=facpub

[26] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005, pp. 80–82.

[27] A. W. Hoover, "Lecture notes: Particle filter," 2016, lecture notes. [Online]. Available: http://cecas.clemson.edu/~ahoover/ece854/lecture-notes/lecture-pf.pdf

[28] C. M. Bishop, *Patern Recognition and Machine Learning*. Springer, 2006, pp. 435–439.

[29] E. Battenberg, "ggmm," http://ebattenberg.github.io/ggmm/, 2014.