SAND2017-2940C

Rapid abstract perception to enable tactical unmanned system operations

Stephen P. Buerger*[a], Anup Parikh[a], Steven J. Spencer[a], Mark W. Koch[b]
[a]Robotics R&D; [b]Sensor Exploitation Applications, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM USA 87185

ABSTRACT

As unmanned systems (UMS) proliferate for security and defense applications, autonomous control system capabilities that enable them to perform tactical operations are of increasing interest. These operations, in which unmanned systems must match or exceed the performance and speed of people or manned assets, even in the presence of dynamic mission objectives and unpredictable adversary behavior, are well beyond the capability of even the most advanced control systems demonstrated to date. In this paper we deconstruct the tactical autonomy problem, identify the key technical challenges, and place them into context with the autonomy taxonomy produced by the US Department of Defense's Autonomy Community of Interest. We argue that two key capabilities beyond the state of the art are required to enable an initial fieldable capability: rapid abstract perception in appropriate environments, and tactical reasoning. We summarize our work to date in tactical reasoning, and present initial results from a new research program focused on abstract perception in tactical environments. This approach seeks to apply semantic labels to a broad set of objects via three core thrusts. First, we use physics-based multi-sensor fusion to enable generalization from imperfect and limited training data. Second, we pursue methods to optimize sensor perspective to improve object segmentation, mapping and, ultimately, classification. Finally, we assess the potential impact of using sensors that have not traditionally been used by UMS to perceive their environment, for example hyperspectral imagers, on the ability to identify objects. Our technical approach and initial results are presented.

Keywords: Perception, autonomy, semantic labeling, collaborative control, unmanned systems

1. INTRODUCTION

Unmanned systems (UMS) continue to gain increasing attention for defense and physical security applications, both as assets and potential threats. If outfitted with autonomous control systems that enable them to perform tactical operations, UMS could allow humans to be removed from some of the most dangerous and unpredictable engagements. For example, the US Department of Defense's (DoD's) Third Offset Strategy [1] seeks to employ robotics and unmanned systems as a primary means of maintaining tactical superiority. As part of this vision, the Deputy Secretary of Defense has identified robotic breaching as a goal to be operational by 2025 [2]. Similar concepts are prevalent across the services: The Army's Robotics Collaborative Technology Alliance seeks capabilities that enable robots to operate as peers within a squad, e.g. being able to "watch the back of that building and report suspicious activity" on command [3]. DoD's UMS Integrated Roadmap envisions capabilities to include: "autonomous 'wingman' capable of human-like tactical behaviors, in-stride support of Marine Corps rifle squads, including tactical decision making while in enemy contact; advanced perception of individual humans in urban environments" by 2022 [4]. These tactical operations will require UMS to match or exceed the performance and speed of people or manned assets, even in the presence of dynamic mission objectives and unpredictable adversary behavior. This represents an extraordinary challenge for perception and control. In this paper we describe the challenges of tactical autonomy, and identify the key next steps to achieving this vision. We summarize our prior work and introduce a new approach to abstract perception that will serve as a key enabler for tactical operations with UMS.

*sbuerge@sandia.gov; phone 1 505 284-3381; fax 1 505 844-8323; sandia.gov

[+]This research was funded by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

1.1 Summary of prior work toward tactical uses of UMS

To date, deployments of robots and UMS in defense and security environments have overwhelmingly used teleoperation, in which human operators control relatively low-level actions, e.g. with joysticks [5]. Even when GPS waypoint control is available, operators must direct essentially every change in action. This approach extends poorly to tactical operations because operator performance is limited by situational awareness and by communication latencies. Immersive operator interfaces may improve situational awareness, but require large communications bandwidth to provide sufficient data. Such bandwidth may not be available in many tactical environments [6]. Achieving remote situational awareness is even more challenging when multiple agents must collaborate toward common mission goals in a shared operational space. We argue that these challenges make direct, low-level operator control fundamentally impractical for tactical UMS missions. Instead, some functions must be executed autonomously onboard one or more UMS. To do this, UMS will need to gain their own situational awareness and choose actions based on observations.

Substantial progress has been made over the past decade in mapping and autonomously navigating through environments with obstacles. Simultaneous localization and mapping (SLAM) enables UMS to build maps of their surroundings using 3D sensing (e.g. LIDAR, stereo vision, or structured light sensors) and to concurrently determine their own locations. Maps and locations are updated via Bayesian methods as the UMS moves and obtains additional sensor data [7,8]. Mapping can be paired with online path planning to enable efficient exploration or traversal through spaces toward a goal [9]. While SLAM and path planning dramatically improve the ability of UMS to autonomously navigate through unknown environments, even without GPS, they do not themselves enable behaviors beyond moving from point A to point B. This is because the maps constructed are exclusively based on geometry, effectively constructing a 3D occupancy grid, rather than providing any abstract characterization of obstacles or objects.

To close "observe, orient, decide, act" loops onboard UMS requires abstract perception. Objects must be segmented from each other and *semantically labeled* as belonging to particular classes. This enables mission behaviors to be referenced to these objects. Much of the prior work in semantic labeling has focused on "personal robotics" applications in home, office, retail or medical environments. Several techniques have been demonstrated including machine learning [10], hierarchical feature-based approaches [11], and combinations of the two [12]. These methods are largely derived from image-processing methods, with depth data used as an enhancement [13]. Results are often tested against a common database of RGB+D (red, green, blue and depth) images [14]. Demonstrations of actual robot operations using these approaches are often painstakingly slow, as appropriate for human-assistive environments. A major drawback of image-based approaches is that they can be readily fooled by visual decoys.

The emerging field of self-driving cars is also advancing real-time perception. Multi-sensor fusion (RGB+LIDAR) was a key enabler to the winning approach in the second DARPA Grand Challenge [15]. For cars driving in traffic, object identification is critical. Today's most prominent developers use radar, LIDAR and optical cameras with deep learning enabled by vast data describing the background environment. Background model changes are propagated through the fleet in real-time [16]. Caution must be used in trying to extend such data-heavy approaches to tactical environments, in which training data and background models may be limited, or out of context.

Limited prior work has explored the use of more exotic sensors, to improve perception (e.g. by identifying material composition) and assist UMS operations. For example, a simple two-band multispectral imager was used to detect the presence of chlorophyll, enabling an autonomous off-road vehicle to distinguish boulders from bushes [17]. Physics-based approaches to fusion have also been used to successfully label scenes. Thermal material properties were estimated from active IR data and the application of physical laws, and fused with imagery [18].

The challenge of autonomously collaborating teams has been addressed via the development of consensus-based swarm control methods [19,20]. While numerous impressive cooperative behaviors have been demonstrated, limitations emerge upon closer scrutiny. Typical swarm control methods limit the cooperating team to a single objective. Agents are usually homogeneous, and each only use data from a single sensor in executing behaviors. The behaviors are frequently limited to moving the agents in a coordinated fashion, e.g. to travel in formation, optimize wireless network topology, or spread around a perimeter. Thus while swarm control methods provide important behavior primitives, they are insufficient to enable diverse, unpredictable tactical operations using teams of heterogeneous multi-functional UMS.

1.2 Paper structure

Section 2 describes the spectrum of capabilities and technologies required to achieve functional tactical autonomy. Section 3 summarizes an approach to achieving tactical behaviors by empowering human operators to direct operations

conducted by multiple heterogeneous UMS. Section 4 describes our overall approach to rapid abstract perception for tactical applications. Section 5 describes a proof-of-concept demonstration of this approach. Section 6 provides discussion, conclusions and future work.

2. TECHNICAL CHALLENGES FOR TACTICAL AUTONOMY

2.1 Spectrum of capabilities for tactical autonomy

Figure 1 shows our model for the mission capabilities and technologies required to achieve tactical UMS autonomy. To increase the level of autonomy, increasingly difficult technical challenges must be met. At the bottom of Figure 1, low-level actions must be directly controlled by an operator. GPS and path planning allow the user to instruct the UMS via waypoints in known, modeled outdoor environments. Known obstacles may be avoided by planning paths around them, but the operator must actively circumvent any unmodeled obstacles. SLAM enables much more sophisticated navigation of complex, unknown, and GPS-denied environments. With this capability, operators can, for example, send one or more UMS into an unknown facility to explore, and the UMS can return with a map and sensor data. SLAM research is sufficiently mature to enable such missions in the near future. However, SLAM does not provide the ability to take further mission action in response to sensor data – for instance to retrieve particular objects or maneuver to physically isolate individuals from key pieces of equipment. These capabilities require two significant additional technologies that are extremely immature today: *rapid abstraction*, and *tactical reasoning*.

Rapid abstraction is the ability to segment and classify all mission-relevant objects in an operating space by human-interpretable names in near real-time, and to link those to a 3D map. Mission-relevant objects include anything of which knowledge is required to facilitate mission execution, including navigation aids (e.g. doors, stairs, control panels), mission targets (equipment, supplies, etc.) or many other things. Rapid abstraction would enable human operators to provide object-based mission instructions, for example "look for the door inside the facility with the control panel, use the key card to open the door, and retrieve all of the radios from the room behind the door."

Tactical reasoning is the ability of an UMS to observe situations and activity, and take appropriate action to achieve higher-level objectives. For instance, one or more UMS could observe multiple threats approaching a sensitive asset, and maneuver to place themselves between the threats and the asset. There is a very broad spectrum of tactical reasoning capability, and increasing capability could enable increasingly high level objectives to be entrusted to autonomous UMS.

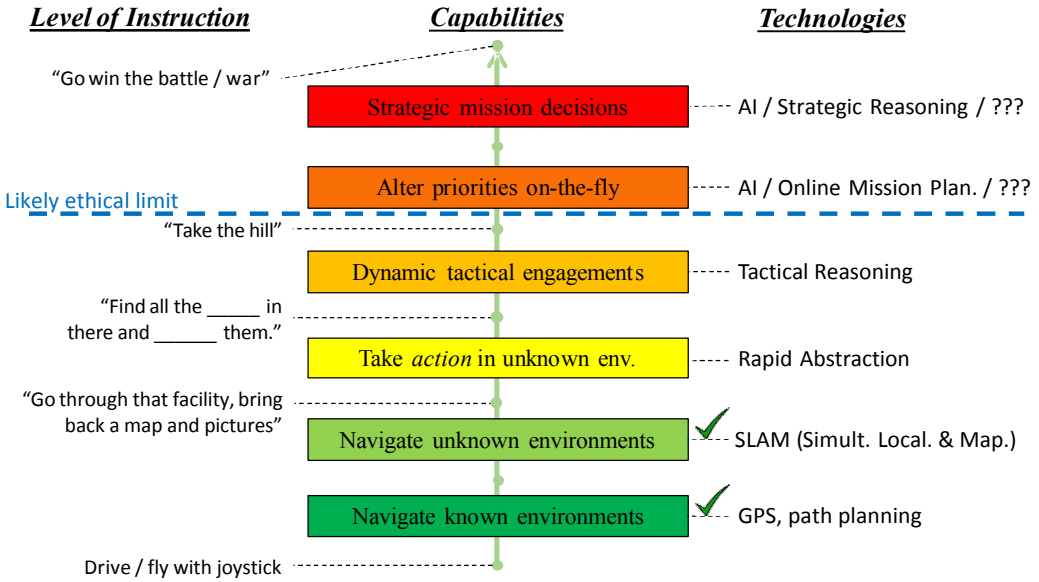

Figure 1. Progression of tactical autonomy from least (bottom) to most autonomous (top). Examples of instructions enabled by each capability are given (left). Core technologies required to achieve each capability are listed (right).

The deployment of these future capabilities could enable UMS to operate tactically within limited short-term mission constraints, approaching the effectiveness of one or more humans or manned vehicles. For instance, tactical UMS could be Army squad support members (e.g. carrying loads and maneuvering intelligently to protect humans from attack), or security sentries (e.g. investigating initial alarms, evaluating threats and requesting appropriate support).

Beyond that, additional capabilities could give UMS even greater authority. This would almost certainly exceed desired ethical limits, indicated by the dashed blue line. They are also well beyond the capability of any current technology.

2.2 Relationship to DoD autonomy community of interest taxonomy

This section places our model for tactical autonomy (Figure 1) into context with the more general autonomy taxonomy developed by the US Department of Defense (DoD) Autonomy Community of Interest (COI). The COI provides a framework for DoD to coordinate within and across agencies and evaluate technology progress toward the use of autonomous systems for its missions. The work discussed in this paper supports this goal, with a particular focus on tactical operations using UMS that will become more autonomous as technology enables progression from the bottom toward the middle of Figure 1. The COI has established a taxonomy for autonomy technology with the following four areas composing its top tier [21]:

- Machine perception, reasoning, and intelligence (MPRI)
- Human/autonomous system interaction and collaboration (HASIC)
- Scalable teaming of autonomous systems (STAS)
- Test, evaluation, validation and verification (TEVV)

The key elements of tactical autonomy map to the first three of these technical areas. Both rapid abstraction and tactical reasoning are key elements of the COI's MPRI thrust. The left column of Figure 1 addresses the level of instruction that humans can expect with tactical autonomous systems at different capability levels. This column lists just a few examples of instruction, generally described in words, but a rich suite of HASIC technologies could enable interaction and collaboration in many different forms. The next section describes one model that allows humans to direct collaborating unmanned teams by efficiently providing both abstract perception and tactical reasoning. Finally, increasing the level of autonomy of individual and collaborating agents enables operations to be scaled up by requiring less human supervision. In particular, the architecture described in the next section enables operations to become increasingly complex (in terms of mission objectives as well as number of agents) as UMS become capable of more advanced autonomous behaviors.

3. OPERATOR-DIRECTED TACTICAL BEHAVIORS

Since 2010 we have worked to develop the Sandia Architecture for Heterogeneous UMS Control (SAHUC), which enables a single operator to control a team of multiple heterogeneous UMS in dynamic, tactical operations [22-24]. SAHUC establishes a framework for operator-directed autonomy with the goal of exploiting, rather than replacing, native autonomous capabilities of various UMS. This approach, which is based on executing objectives specified by the human operator in real-time, can enable effective operations using UMS with a variety of autonomy capabilities. Implementations to date have focused mostly on enabling the human operator to provide essentially all high-level perception and tactical reasoning for the team, while executing all lower-level behaviors autonomously.

3.1 Sandia architecture for heterogeneous UMS control (SAHUC)

An annotated schematic view of SAHUC is shown in Figure 2. The framework consists of control elements that may be flexibly located throughout the system, and messages that are passed between them. SAHUC enables control of multiple agents using several automatic control layers between the operator and the low-level behavior of individual agents. At the lowest level, individual asset control migrates each UMS from the current state to a desired new state in real-time. This layer may or may not include advanced navigation such as obstacle avoidance and advanced path planning. If these capabilities are not available in the native control, they may be provided by the next highest control layer.

At the highest level, the human operator interacts with the system primarily by observing live sensor feeds and a 3D common operating picture, and specifying and prioritizing desired tactical outcomes in real-time. These goals, captured in the *Objective* message (and prioritized in the *WeightList* message), articulate desired outcomes rather than UMS

actions – e.g. "show me continuous imagery of that red truck" rather than "UAV#2, fly to the northwest 500 m and loiter." This allows the system to autonomously optimize mid-level behaviors to achieve mission outcomes. Desired outcomes are translated into lower-level actions by the "high level optimizer" (HLO) and "mid level behavior controllers and estimators" (MLBCs), two SAHUC-specific control layers. In short, the HLO continuously determines which agents in the system should address which *Objectives*. Assignments may include multiple agents per objective (e.g. swarming) or multiple simultaneous objectives per agent (multitasking), and are revisited continuously in real-time. We have demonstrated automatic real-time task handoffs between agents as warranted by evolving conditions [24]. Assignments are made to minimize overall cost as defined by a multi-objective cost function. The operator indirectly manipulates the overall cost function in real-time by adding, removing, and prioritizing objectives. The operator can also change behavior through other messages (not shown) that trade overall system priorities such as power consumption, speed of execution and redundancy / robustness against each other. Costs of each UMS to execute each task are computed by MLBCs, which also execute the assigned behaviors by directing lower-level agent actions in real-time. MLBCs can be devoted to individual agents or collaborating teams (e.g. swarms). The HLO and MLBCs may be implemented centrally or distributed throughout the system, and are modular to enable rapid algorithm updates. The operator can also access lower-level messages to force lower-level actions if the autonomous controllers are not producing the desired outcomes.

Through these tools, the operator is able to sculpt tactical missions in real-time, and respond to observed events. Because high-level perception and tactical reasoning are substantially more challenging to automate than the lower-level vehicle navigation functions, this framework attempts to divide those functions along those lines, allowing the operator to concentrate mostly on the former and the system to automate the latter.

3.2 Demonstrations of tactical autonomy principles

The ability to execute dynamic, tactical missions using SAHUC has been demonstrated using teams of up to four unmanned ground and aerial vehicles, controlled by a single operator using multiple levels of control. Demonstrations included physical security scenarios with active adversaries, and remote inspection scenarios with simulated hazardous materials [23-25]. In each case, nothing was scripted or pre-planned; the operator constructed the mission entirely on the fly and in response to real-time observations. In these demonstrations, the vehicles were capable of autonomously navigating to desired states, but the operator provided all high-level perception and tactical reasoning. The lack of autonomous perception significantly limited the complexity of the system behaviors and the available *Objective* types. Mid-level behaviors were generally limited to simple behaviors such as shadowing intruders, patrolling perimeters, or searching areas. If individual agents had the ability to semantically perceive objects in their environment, substantially more complex behaviors based on these perceptions could be specified. For instance, agents could map points of entry in a structure or find all radio communication equipment. Figure 3 shows several images from a past demonstration.

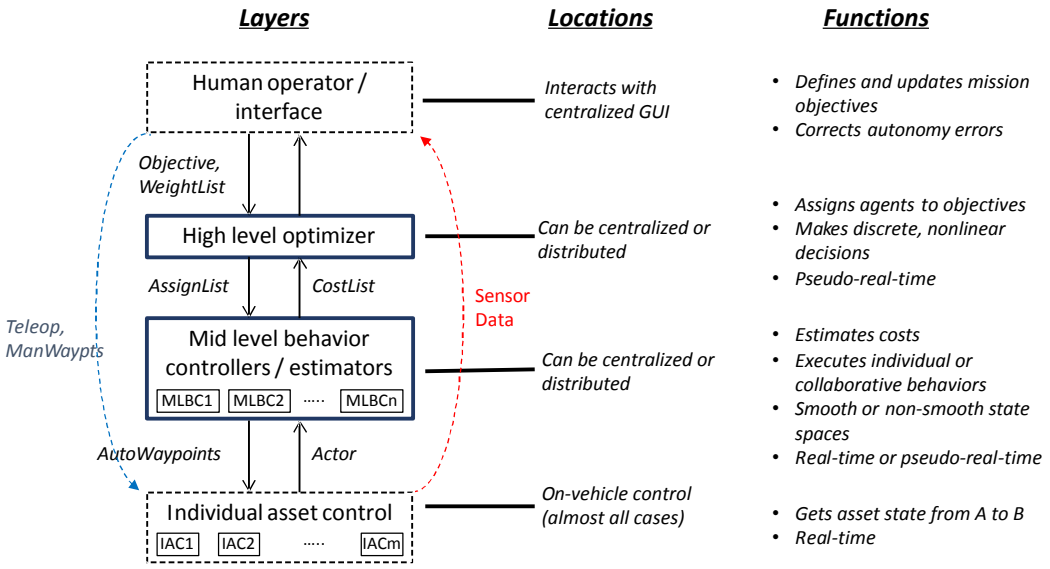

Figure 2. Schematic representation of SAHUC. Control layers and a subset of critical messages are indicated at left. The locations of algorithms that implement the layers (center) and the functions performed by each layer (right) are listed.

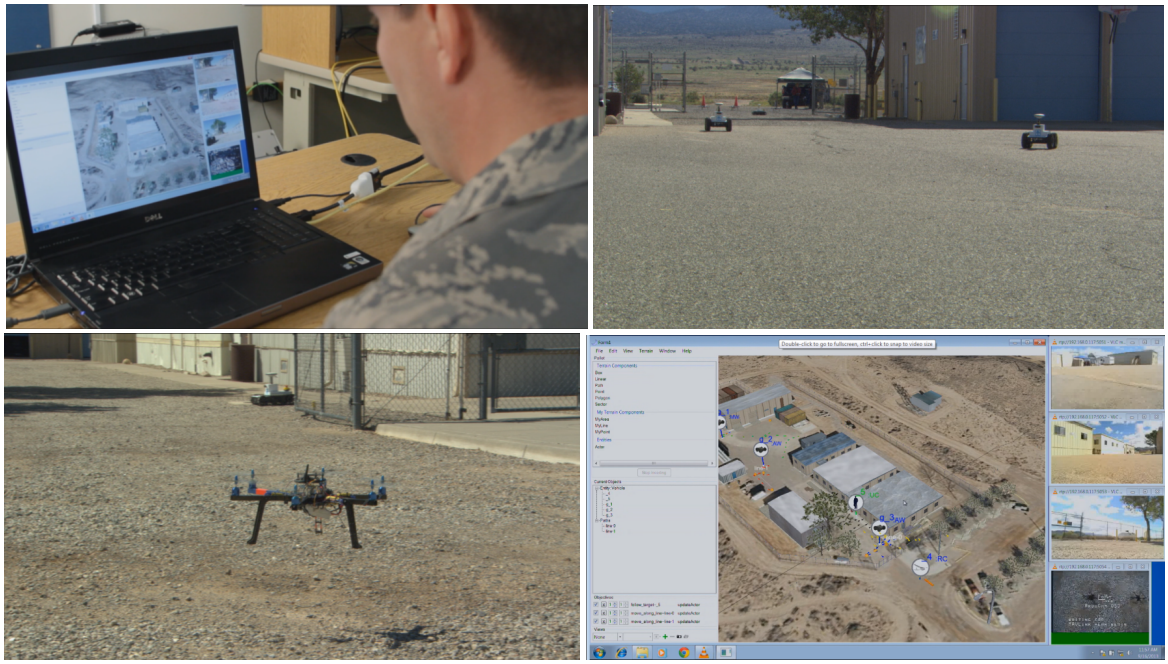

Figure 3. Screenshots from a demonstration of a single operator (top left) directing a heterogeneous team of four unmanned ground (top right) and air (bottom left) vehicles in a dynamic, tactical operation. The user interface (bottom right) includes live sensor feeds (right panels) a 3D common operating picture (middle), and a series of graphical tools that enable the operator to build missions in real-time in response to unpredictable observations and adversary behavior.

4. EXPANDING TACTICAL BEHAVIORS VIA RAPID ABSTRACTION

In our model for tactical autonomy (Figure 1), rapid abstract perception is needed to increase the complexity of behaviors that can be executed without direct operator intervention. In the context of SAHUC, the ability to quickly identify all relevant objects could enable much more advanced mid-level behaviors, allowing an operator to control more agents and coordinate more elaborate missions by focusing more on tactical reasoning and less on perception. Abstract perception would enable tactical behaviors to be defined in terms of objects and actions in the environment – just as humans describe behaviors to each other.

4.1 Goals for rapid abstraction in complex environments

The long-term goal of our perception work is to develop a set of algorithms that integrate mapping and object labeling in unpredictable environments in real-time. The intent is to label all objects of mission or operational significance in a scene in order to enable behaviors that are referenced to those objects. For defense applications, this must be done in a variety of highly dynamic environments and ambient conditions, with incomplete intelligence and without comprehensive training data. Our technical approach seeks to explicitly manage these constraints.

4.2 Technical approach

Our approach comprises three thrusts: 1) Physics-based multi-sensor fusion is used to classify / identify objects based on multiple measured properties; 2) Sensor dynamics, including the location / perspective of moving sensors, are controlled to minimize classification uncertainty; and 3) A rich sensor suite is employed to provide *a priori* information that is unavailable to more commonly-used RGB+D sensors. These thrusts are summarized below.

Physics-based multi-sensor fusion: Much of the recent literature on object identification has focused on powerful learning methods, particularly deep neural networks, that learn to classify objects by training with extensive labeled data. For many tactical applications, the lack of complete, relevant training data in the actual environment renders these "black box" methods insufficient. Furthermore, tactical operations may require classification decisions to be traceable to specific data features, which is not generally possible when learning networks are applied at a high level.

In contrast, our approach uses two stages to classify and label objects. First, physics properties and features of objects are measured and quantified. This first stage may use heuristics or learning methods, as the networks are trained to measure specific *properties* in a variety of contexts, as opposed to identifying specific target *objects*. Second, object classes / labels are determined based on several properties / features and the relationships between them.

These relationships are often depicted using graphs [26]; specific graphical descriptions may be defined a priori or learned. Figure 4 shows an example of a graph describing a face. A graph contains nodes and arcs. Here, the nodes are represented by the rectangular pictures and the arc are represented by the black lines. The nodes show parts of a face with black text indicating the label for node. This graph shows a face is made up of a head that contains objects such as a nose, mouth, and eye pair. The eye pair is further broken down to left eye and right eye. The arcs show the relationships between the different parts. These relationships are labeled by the italicized red text. For example, the nose is in the middle of the head and the mouth is in the lower half of the head. Each node is further describe by geometrical and physical attributes. These attributes can be encapsulated by statistical feature based or machine learning algorithms.

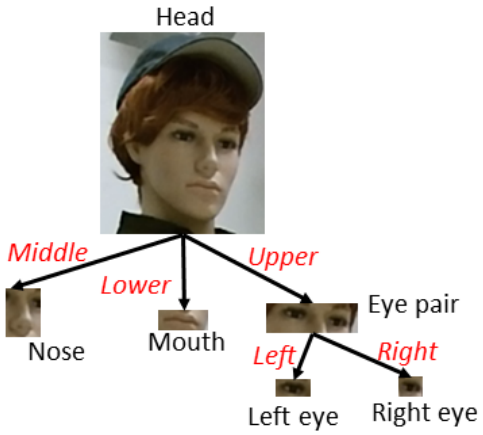

Figure 4. Example of a graph that defines a face.

Figure 5 shows an abstract graph-matching example. The red graph depicts a model graph *M* of an object that the system has loaded in its database. The green graph shows a graph representation of a scene *S* that the system wants to break into different objects. The nodes of *S* are created by running various object detectors over the scene. The arcs determine the relationship between the objects. The objective of graph matching is to determine the function *f* that matches nodes and arcs in *M* to nodes and arcs in *S*. Graph-theoretic matching algorithms that can handle missing objects and spurious objects can be found in [26]. Unfortunately, these algorithms have exponential time complexity and are impractical except for the simplest of problems. Ransac type algorithms [27] can also be devised to match partial graphs. Here, one selects a match of a model node to a scene node with a high similarity and tries to add other nodes that support the match. This will not necessarily give an optimal solution, but in general will be close to optimal.

Condition random fields (CRF) [28] can also be used to represent the graph matching probabilistically and infer the most likely scene labeling by maximizing the posterior probability of the random field. Graph cuts with alpha expansion are used for this multilabel optimization problem [28]. CRF's are a special case of a Markov random filed (MRF). An MRF describes the $\Pr(A,S)$ using a graphical model. Here, $S=\{s_1,...,s_M\}$ represents the M scene nodes. The variable $A=\{a_1,...,a_M\}$ represents the set of scene assignments to model nodes. Here a_k is the model node label for scene node s_k. Thus each a_k is a number between 0 and N, where 0 represents the unknown label and N is the number of model nodes. The MRF can be used as generative or inferential model. The problem is that s_i and features derived from s_i can be highly correlated and/or redundant, and an MRF is poor at describing these correlations and requires a densely connected graph.

The CRF is an MRF conditioned on the data: $\Pr(A|S)$ This removes the correlations and redundancies in S, since S is now assumed given. The model is no longer generative, but that is not a problem, since we want to do inference. Typically the CRF objective function is represented as $-\log(\Pr(A|S))$ and is called the energy function. Thus, instead

of maximizing the posterior probability we want to find the labeling that minimizes the energy function. The second order energy function for a CRF is given as:

$$E(A) = \sum_{i \in V} \psi_i(a_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(a_i, a_j)$$

Here V corresponds to the set of scene node indices and the neighbor set N_i corresponds to the set of scene node indices that are adjacent (connected by a scene arc) to scene node i. The term $\psi_i(a_i)$ describes the energy of assigning scene node i the label a_k and is call the *unary* term. The pairwise term $\psi_{ij}(a_i, a_j)$ represents the energy of the compatibility between nodes i and j. A low energy for $a_i = a_j$ would favor the two scene nodes having the arc as define in the scene graph.

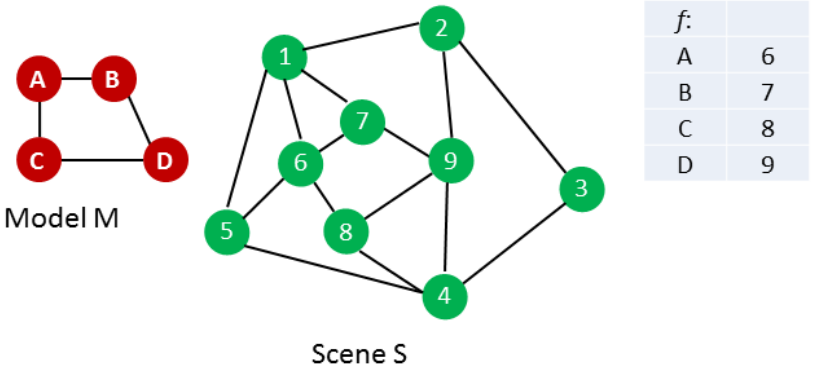

Figure 5. Abstract graph matching example.

Dynamic sensor control: Because abstract perception is such a critical capability for the intended applications, we anticipate that the perception system will have at least partial authority over controllable dynamic properties of the sensors. At a minimum this includes moving sensors to change perspective relative to target objects, but could also include controlling sensor gains, wavelength sensitivities / filters, focus and other properties in response to measured data. We are developing real-time control algorithms to minimize the classification uncertainty of key objects by controlling sensor dynamics in response to measurements.

Changing the sensor perspective can improve the performance of our physics-based fusion algorithms in several different ways. Different perspectives can aid in the object segmentation problem: separating objects from "background." Sensor movement may be used to reduce the uncertainty in the estimate of a particular physics property (e.g. by getting closer to improve effective resolution), or to produce a view of an additional property or feature that is obscured (e.g. by moving over an object to determine its depth). Finally, a new perspective may help to clarify the relationships between different properties and features as described in graphical representations of objects. Specifically controlling perspective mimics the behavior of animals and humans, who intuitively move to obtain additional information when perceiving their environment.

Our initial work in this area is focused on quantifying the impact of open-loop perspective changes on classification uncertainty. Ultimately perspective will be controlled with optimal closed-loop techniques, e.g. by extending existing techniques that identify "next best views" for mapping [29,30].

Sensor suite: Much of the current literature in semantic labeling focuses primarily on using purely optical (RGB) or optical plus depth (RGB+D) imagers [10-13]. This enables the use of commodity sensors as well as existing data sets. These sensors can fundamentally only measure the color, shape and size of objects. While they are data-rich, powerful and versatile, these sensors may be particularly limited in less structured and adversarial environments. They can be fooled by decoys, made less effective by environmental conditions (e.g. lighting, obscurants, weather, etc.), or may simply be incapable of differentiating certain critical objects in cluttered, unfamiliar environments. It is highly desirable to augment these with additional measurements that can generate distinct properties and features of objects, notably the composition and function. The ability to identify different material classes (e.g. metals, organics, conductors vs. insulators, etc.) could be particularly valuable and complementary to the RGB+D suite. We are testing several options

for identifying material types including hyperspectral imaging, thermal imaging (e.g. measuring thermal capacitance or conductance [18]), electromagnetic sensing, and optical surface characterization. To characterize function, we are considering additional sensing modalities including passive acoustic, passive RF and chemical detection.

The next section describes a proof-of-concept implementation of this integrated approach that demonstrates the benefits of each of these thrusts and the power of their combination.

5. RAPID ABSTRACTION PROOF-OF-CONCEPT DEMONSTRATION

An initial proof-of-concept demonstration was conducted to evaluate the effectiveness of each of these technical thrusts on the perception problem. While our ultimate goal is to identify all of the relevant objects in an operational area, the initial demonstration focused on the narrower, representative problem of distinguishing humans from decoys (including mannequins and cardboard cutouts) in a laboratory environment. This scenario illustrates the performance improvements provided by a diverse sensor suite, multi-sensor physics-based fusion, and changes in perspective.

5.1 Detection of multiple features

Experiments used an RGB+D sensor (a Microsoft Kinect 2.0) and a hyperspectral imager (HSI: BaySpec OCI-U-2000), sensitive to wavelengths from 600-1000 nm. Three separate features were identified using different elements of the sensor suite. These features were detected in a scene that included two full-size cardboard cutouts, two 3D mannequins, one mannequin head, and one human. The detection elements included:

RGB head detection and tracking:

Head detection is performed using Matlab's *cascade object detector* with the RGB optical data. The algorithm is based on the work of Viola and Jones [31]. The algorithm is highly efficient in terms of feature computation and applying more computation to regions most likely contain a head. The features are based on Haar wavelets [32], which are computed using an integral transform in constant time for any scale. This makes the detector scale and location invariant. The classifier is a cascade of subclassifiers that are trained to detect a very high percentage of target objects (heads) while rejecting a certain fraction of nonobject patterns (nonheads). By having a large number of subclassifiers in this cascade, the false alarm rate can be driven down to a very small fraction. As one moves down the cascade, computation is applied only to the regions that survive the previous subclassifier. In general, the detector is most effective on frontal images of heads and can be somewhat sensitive to lighting conditions.

To associate frames from one image to the next we use a standard detect-predict tracker. Any detections that do not correspond to a previous track are used to start new tracks. Currently, we are using a naïve predictor. We assume camera movement and object movements are small enough that a good prediction for the location of the head in the current frame is position in the previous frame. If necessary, a continuously adapted mean-shift (Camshift) [33] algorithm could be used to provide more accurate prediction estimates. To assign detections to tracks we create a matrix where we compare distances of all the current detections to the predict locations of the head for each track. The optimal assignment is found using James Munkres' variant of the Hungarian assignment algorithm [34]. Tracks are then updated with the assignments. A track is allowed to coast if there is no corresponding detection found, but is deleted if it has been invisible for too long. Here, we assume there was a spurious detection without enough follow-on detections, the track has moved so it is no longer detected, or the track has moved outside the frame. The result is a track identifier assigned to all the detections in the current frame that can be associated to detections in previous frames. Figure 6 shows an example of head tracks that have been sustained for several frames. Human, mannequin, and cardboard cutout heads are tracked, as are several more obvious false positives.

RGB+D skeletal geometry:

In addition to raw color and depth, the Kinect sensor provides estimates of the skeletal pose of humans in the field of view. Specifically, the 3D positions of skeletal joints relative to the sensor are estimated. Figure 7 shows estimated skeleton poses for the same frame as Figure 6. The skeleton tracking is heavily reliant on the presence of a face (using its inbuilt face tracker), especially during detection; for example, removing the mask from the wooden mannequin (second from left) severely degrades tracking performance. Furthermore, the Kinect does not strongly discriminate based on typical human figures beyond overall size and aspect ratios; the cardboard cutouts in Figure 7 (red and yellow) are easily tracked when faced head-on despite the planarity of the figure and the lack of a gap between the legs, or arms and torso.

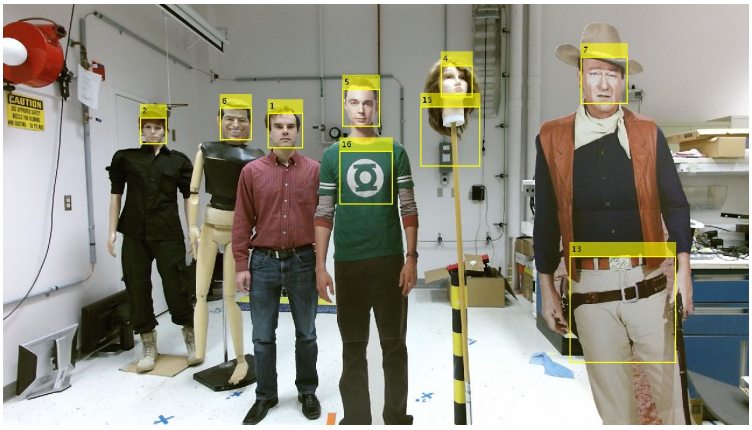

Figure 6. Example of RGB head detection and tracking.

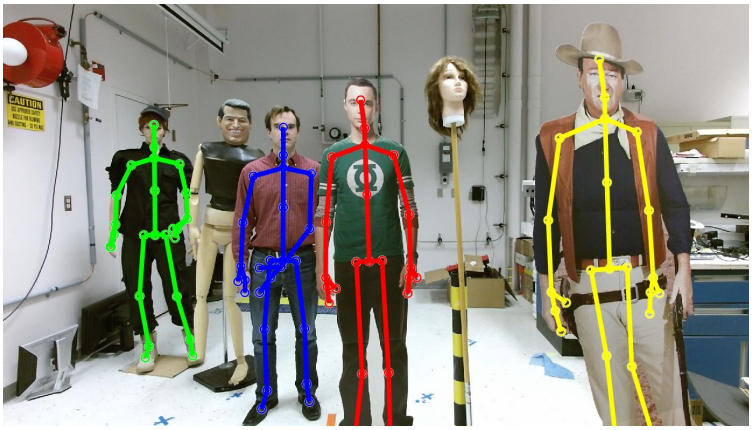

Figure 7. Example of skeleton tracking using RGB-D sensor.

Figure 8. Example of successful skin detection using hyperspectral imagery. Skin on the human hands and face are clearly distinguished from faces not having skin.

HSI skin detection:

Snapshot HSI images were collected and formed into hyperspectral cubes with 25 bands between 600-1000nm. Many researchers have previously demonstrated the value of HSI for skin detection [35-39]. However, many of these approaches take advantage of features of skin spectra above 1000nm, because the longer wavelengths exhibit less variation with skin type. For this reason, we used a different approach. To detect skin, we first trained a linear discriminant classifier using data sampled from one image in the dataset. Sampled data consisted of 30 pixels from each of the following objects in the scene: human forehead, human cheek, human hand, cutout 1 forehead (later called

"Cooper"), cutout 2 cheek ("Wayne"), plastic head, mannequin head, mask ("Nixon"), floor, wooden leg, and tool chest. For training the classifier, samples were labeled as skin or not-skin. Results of the classifier on an image from the test set are shown in Figure 8. The classifier successfully detected the human face and hands, and, in spite of some false positives, it did not detect skin on the other (false) faces. Fewer false positives could likely be obtained using imagers capable of imaging above 1000nm and using more advanced skin detection methods.

5.2 Multi-sensor fusion

Figures 6-8 show that while each of the three feature detection schemes can identify the desired human features, each suffers from significant false positives induced by the decoys, and none is a successful "human detector" in isolation. In the spirit of our technical approach as described in section 4, each detector uses independent methodology and targets a particular feature of people (head, skeleton, and skin); it remains to combine their outputs to reliably detect humans.

By securely mounting the hyperspectral imager to the Kinect and jointly (intrinsically and extrinsically) calibrating both devices, the RGB, depth, and hyperspectral data can be mapped to a common reference frame. The color and depth imagery from the Kinect are used to produce an RGB point cloud. Points in the cloud that map to within the head bounding boxes are identified as head points. The same is true for points that map to the skin regions in the HIS data. For the skeleton, a 3D bounding box is generated based on the 3D joint positions, expanded to account for typical head sizes and body depths, and used to identify points that correspond to a body structure. Because we are not yet building a graph of all features and objects in a scene, the graph methods described in section 4.2 were not used. Instead, more simply, points in the point cloud belonging to all three feature classes are identified as belonging to a human for that frame. Figure 9 shows each cluster of points in the cloud, highlighted to indicate each feature and the composite (lower right) for one example frame. For this frame, only the true human head satisfies all three feature constraints.

To counteract false positive and negative human classifications in individual frames, a temporal Bayes update is performed across multiple frames, using the data association provided by the head tracking technique described in section 5.1. Points in the point cloud that are classified as human and map into the bounding box for a particular head track are marked as a detection event for that track. The probability that a particular face track refers to a human is updated using Bayes rule with probabilities P(Detection | Human) = 0.8 and P(Detection | ~Human) = 0.5, with initial probability P(Human) = 0.1 for all face tracks.

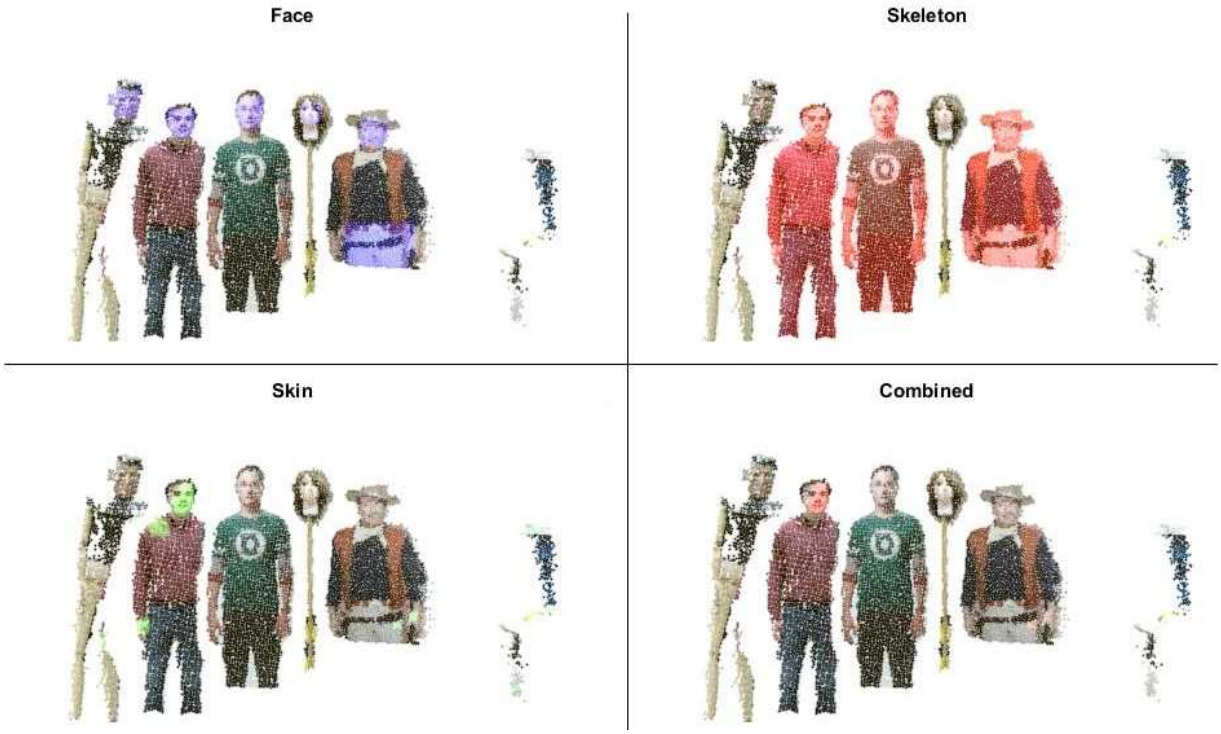

Figure 9. Points in the Kinect point cloud can be identified as belonging to a head (purple in top left), skeleton (red in top right) or skin (green in bottom left). The intersection of those sets is classified as human (red in bottom right).

5.3 Experimental results

Decoys were used to evaluate the ability to distinguish humans from similar objects, including a wooden mannequin with a rubber mask (denoted "Nixon"), a plastic mannequin (denoted "Mannequin"), two human-sized cardboard cutouts (denoted "Cooper" and "Wayne"), and a plastic head mounted on a pole (denoted "Head"). A human and the 5 decoys were arranged in the scene before each test. Two data sets were collected with varying motion trajectories of the imagers: (1) pure rotation from a fixed spatial position, and (2) changing perspective by combined translation / rotation. The results of these tests demonstrate that (a) varying imager perspective can improve performance, and (b) the multi-sensor approach is effective in identifying humans.

Benefits of perspective change:

Changing perspective, versus just rotating, proves beneficial in determining the overall geometry of 3D objects in a scene. For the task considered in this proof-of-concept demonstration, changing perspective helps distinguish the planar cardboard cutouts from humans and mannequins. Specifically, the Kinect does not place skeleton trackers on cardboard cutouts when they are viewed from the side, as exemplified in Figure 10. This is overall impact of perspective change is demonstrated by Figures 11 and 12, which plot the percentage of frames in which each feature is detected for the two trials. Focusing in particular on the center plots, skeletons for the cardboard cutouts (Cooper and Wayne) are detected significantly less often in the perspective-varying test (Figure 12) than the rotation only test (Figure 11). This is because the perspective-varying test exposes angles in which the lack of true skeletal geometry is more obvious to the Kinect sensor. This demonstrates the straightforward yet important result that even simple perspective variation can significantly aid classification.

Figure 10. Skeleton tracking for the cardboard cutouts (e.g., Cooper) degrades as they are viewed from the side.

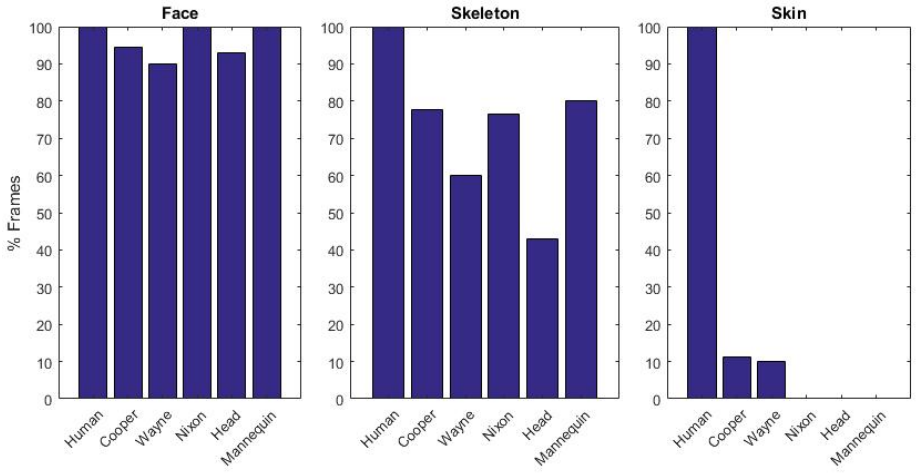

Figure 11. Percentage of frames (out of total number of frames the body is in view) where the various bodies were classified with a face, skeleton, or skin, for the dataset with purely rotational motion.

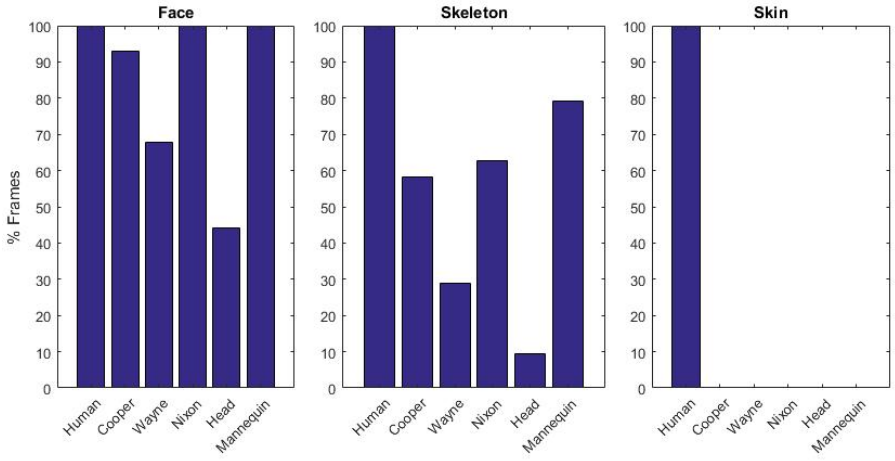

Figure 12. Percentage of frames (out of total number of frames the body is in view) where the various bodies were classified with a face, skeleton, or skin, for the dataset with changing perspective.

Composite human detection results:

Figures 11 and 12 elucidate the need for multiple, independent sensors. Using either head detection from an RGB imager or skeletal tracking from an RGB+D sensor exclusively would result in numerous false positives. Skin detection with the HSI appears to be especially accurate, but it is important to note that Figures 11 and 12 only show skin detections *within the face bounding box for each of the bodies*. As shown in Figure 8, many regions in the background or outside the face are incorrectly classified as skin, and would yield incorrect human classifications if not used in conjunction with the other sensors. The final results of the experiments are shown in Figure 13, which plots the evolving probabilities that each head track is human. By fusing the output of multiple sensors and detection algorithms, and accumulating information with simple temporal filtering as data are collected, the head track corresponding to the human quickly reaches a high probability while the others are exposed as false.

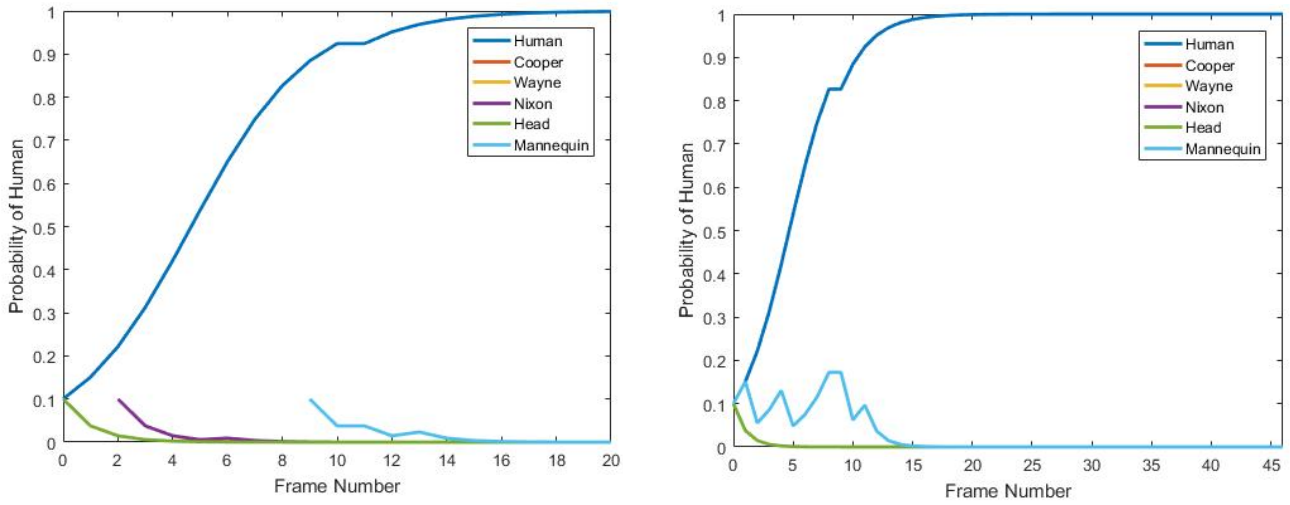

Figure 13. The evolution of probabilities of being human for the head tracks associated with the human and decoys. (Left) Pure rotation dataset. (Right) Changing perspective dataset.

6. DISCUSSION, CONCLUSIONS AND FUTURE WORK

The experiment described in section 5 demonstrates several key elements of our technical approach for achieving rapid abstraction, and provides a simple validation of several of these concepts. First, three detectors were implemented to

detect independent known features of people. These detectors each used very different constructs, including combinations of heuristics and training. Second, their output was then combined using an *a priori* model of the relationships between the features to robustly detect humans. Results show very successful rejection of the significant false positives and false negatives that plague the individual detectors. Third, the detector results were mapped onto 3D map data that was generated by the same sensor suite in real-time with no prior map or localization knowledge. Fourth, the results for skeleton tracking show that varying perspective, even over a relatively small distance and relative angle, can significantly improve the detection and quantification of features.

Substantial work remains to achieve our overall perception goal of identifying all relevant objects in a scene and integrating labeling with mapping. Feature graphs and graph searching methods need to be implemented; this requires a much larger library of object templates than used in the work to date. Substantial work also remains to integrate the development of scene graphs with online SLAM mapping and object segmentation. Ultimately geometric segmentation from SLAM and semantic object segmentation from scene graphs should be integrated and should each boot-strap each other to improve performance. Closed-loop control of perspective variation to optimize object classification has not yet been implemented. Finally, we will continue to explore additional sensors to complement our current suite.

Once this rapid abstraction capability is established, it will be integrated within SAHUC to enable a significant improvement in the execution of tactical mission complexity.

REFERENCES

[1] Work, B., "The third U.S. offset strategy and its implications for partners and allies," As-delivered remarks at Willard Hotel, Washington DC, 28 January 2015. http://www.defense.gov/News/Speeches/Speech-View/Article/606641/the-third-us-offset-strategy-and-its-implications-for-partners-and-allies (9 March 2017).

[2] Pellerin, C., "Work: Human-machine teaming represents defense technology future," *DoD News*, 8 November 2015. http://www.defense.gov/News-Article-View/Article/628154/work-human-machine-teaming-represents-defense-technology-future (9 March 2017).

[3] Bornstein, J. and Mitchell, B., "Robotics collaborative technology alliance – Foundations of autonomy for ground robotics," Army Research Laboratory, Adelphi, MD, 26 April 2012. https://www.arl.army.mil/www/pages/392/OverviewBriefingRoboticsCTA.pdf (9 March 2017).

[4] "Unmanned Systems Integrated Roadmap. FY2013-2038." Department of Defense, UnderSecretary of Defense Acquisition, Technology & Logistics, Washington, DC, 71 (2014).

[5] Fong, T. and Thorpe, C., "Vehicle teleoperation interfaces," Autonomous robots 11(1), 9-18 (2001).

[6] Sorelle, R., "How to equip the U.S. military for future electronic warfare," NDIA Industry Insights, January 2013. http://www.nationaldefensemagazine.org/archive/2013/January/Pages/HowtoEquiptheUSMilitaryForFutureElectronicWarfare.aspx (9 March 2017).

[7] Thrun, S. and Leonard, J., "Simultaneous localization and mapping," in Springer handbook of robotics, Springer Berlin Heidelberg, 871-889 (2008).

[8] Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I. and Leonard, J.J., "Past, present and future of simultaneous localization and mapping: Toward the robust-perception age," IEEE Transactions on Robotics 32(6), 1309-1332 (2016).

[9] Meyer, J.-A. and Filliat, D., "Map-based navigation in mobile robots: II. A review of map-learning and path-planning strategies," Cognitive Systems Research 4(4), 283-317 (2003).

[10] Koppula, H.S., Anand, A., Joachims, T. and Saxena, A., "Labeling 3d scenes for personal assistant robots," arXiv preprint arXiv: 1106.5551 (2011).

[11] Wu, C., Lenz, I. and Saxena, A., "Hierarchical semantic labeling for task-relevant RGB-D perception," Robotics: Science and Systems (RSS) (2014)

[12] Rusu, R.B., Marton, Z.C., Blodow, N., Holzbach, A. and Beetz, M., "Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments," IEEE/RSJ Int. Conf. on Intell. Rob. & Sys., 3601-3608 (2009).

[13] Husain, F., Schulz, H., Dellen, B., Torras, C. and Behnke, S., "Combining semantic and geometric features for object class segmentation of indoor scenes," IEEE Rob. and Autom. Letters 2(1), 49-55 (2017).

[14] Silberman, N., Kohli, P., Hoiem, D. and Fergus, R., "NYU Depth Dataset V2." http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html (9 March 2017).

[15] Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G. and Lau, K., "Stanley: The robot that won the DARPA grand challenge," J. Field Rob. 23(9), 661-692 (2006).
[16] Faust, A., "Self-driving cars, quadrotors and computing, different faces of motion planning," University of New Mexico Computer Science Department Colloquium Series, 6 April 2016, Albuquerque, NM.
[17] Bradley, D.M., Thayer, S., Stentz, A. and Rander, P., "Vegetation detection for mobile robot navigation," *CMU-RI-TR-04-12*, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2004).
[18] Nandhakumar, N. and Aggarwal, J.K., "Physics-based integration of multiple sensing modalities for scene interpretation," *Proceedings of IEEE* 85(1), 147-163 (1997).
[19] Vincent, R. et al, "Distributed Multirobot Exploration, Mapping, and Task Allocation," *Ann. Math. Artif. Intell.* 52(2), 229-255 (2009).
[20] Scardovi, L. et al., "Stabilization of Three-Dimensional Collective Motion," *Comm. Inf. Sys.* 8(4), 473-500 (2008).
[21] Bornstein, J., DoD Autonomy Roadmap Autonomy Community of Interest, NDIA 16th Annual Science & Engineering Technology Conference/Defense Tech Exposition, 24-26 March 2015.
[22] Buerger, S.P., Neely, J., Little, C., Amai, W., Joyce, R. and Love, J.A., "A layered control architecture for single-operator control of heterogeneous unmanned system teams," Proc. SPIE 8387, 838702 (2012).
[23] Love, J., Amai, W., Blada, T., Little, C., Neely, J., Buerger, S., "Enhanced physical security through a command-intent driven multi-agent sensor network," in Schmorrow, D., Fidopiastis, C. (eds) [Foundations of Augmented Cognition], Springer International Publishing, 784-795 (2012).
[24] Love, J., Amai, W., Blada, T., Little, C., Neely, J. and Buerger, S., "The Sandia architecture for heterogeneous unmanned system control (SAHUC)," Proc. SPIE 9464, 94640E (2015).
[25] Buerger, S.P., Love, J.A., Little, C.Q., Amai, W.A., Neely, J.C., Blada, T.J., Kuehl, M.A. and Wilder, D.J., "Command intent on the future battlefield: One operator controlling many unmanned systems," Sandia National Laboratories Report SAND2013-10713 (2013).
[26] Ballard and Brown, [Computer vision], Prentice Hall, Englewood Cliffs, New Jersey (1982).
[27] Fischler, M.A. and Bolles, R.C., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM 24(6), 381-395 (1981).
[28] Boykov, Y., Veksler, O. and Zabin, R., "Fast approximate energy minimization via graph cuts," IEEE Trans. Pattern Anal. Mach. Intell., 23(11), 1222-1239 (2001).
[29] Potthast, C. and Sukhatme, G.S., "A probabilistic framework for next best view estimation in a cluttered environment," J. Vis. Commun. Image R, 25, 148-164 (2014).
[30] Krainin, M., Curless, B. and Fox, D., "Autonomous generation of complete 3D object models using next best view manipulation planning," IEEE Int. Conf. on Robotics and Automation (ICRA), 5031-5037 (2011).
[31] Viola, Paul and Michael J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, 511–518 (2001).
[32] Haar, Alfréd, "Zur Theorie der orthogonalen Funktionensysteme", Mathematische Annalen, **69** (3): 331–371 (1910).
[33] Bradski, G. and A. Kaehler, *"*Learning OpenCV :Computer Vision with the OpenCV Library*,"* O'Reilly Media Inc.: Sebastopol, CA (2008).
[34] Munkres, J., "Algorithms for assignment and transportation problems," Journal of the Society for Industrial and Applied Mathematics 5(1), 32-38 (1957).
[35] Bebis, G., Gyaourova, A., Singh, S., and Pavlidis, I., "Face recognition by fusing thermal infrared and visible imagery." Image and Vision Computing, 24(7), 727-742 (2006).
[36] Pan, Z., Healey, G., Prasad, M., and Tromberg, B., "Face recognition in hyperspectral images." IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12), 1552-1560 (2003).
[37] Nunez, A.S. and Mendenhall, M. J., "Detection of human skin in near infrared hyperspectral imagery," In Geoscience and Remote Sensing Symposium, 2008. IEEE International 2, II-621 (2008).
[38] Steiner, H., Sporrer, S., Kolb, A. and Jung, N., "Design of an active multispectral SWIR camera system for skin detection and face verification," Journal of Sensors (2016).
[39] Mendenhall, M.J., Nunez, A. S. and Martin, R. K., "Human skin detection in the visible and near infrared," Applied optics, 54(35), 10559-10570 (2015).