Spring 2022

Visual Homing for Robot Teams: Do you see what I see?

Damian Lyons

Noah Petzinger

Visual Homing for Robot Teams: Do you see what I see?

Damian M. Lyons and Noah Petzinger

Robotics & Computer Vision Lab., Fordham University, New York, USA

ABSTRACT

Visual homing is a lightweight approach to visual navigation which does not require GPS. It is very attractive for robot platforms with a low computational capacity. However, a limitation is that the stored home location must be initially within the field of view of the robot. Motivated by the increasing ubiquity of camera information we propose to address this line-of-sight limitation by leveraging camera information from other robots and fixed cameras. To home to a location that is not initially within view, a robot must be able to identify a common visual landmark with another robot that can be used as an 'intermediate' home location. We call this intermediate location identification step the "Do you see what I see" (DYSWIS) task. We evaluate three approaches to this problem: SIFT based, CNN appearance based, and a semantic approach.

Keywords: Mobile robots, GPS-denied, navigation, visual homing

1. INTRODUCTION

Autonomous and semi-autonomous ground and air vehicles typically use GPS coordinates for wide-area (i.e., beyond immediate sensor range) outdoor navigation. In some outdoor applications, such as precision farming, accurate GPS navigation is in fact essential (DeBaerdemaeker 2013).[1] Other sensors, such as vision and laser ranging, are employed primarily for obstacle avoidance or other local navigation between GPS waypoints. However, for some outdoor applications, a vehicle may need to operate reliably in situations where GPS can be interrupted, is unavailable or is denied. Much of the prior work in this area addresses GPS-denied localization, especially in the context of UAVs (Bachrach 2010).[2] In contrast, we address the issue of *wide-area navigation* in the absense of GPS: traversal to a target destination beyond sensor range known only by its appearance. We propose an approach that leverages a light-weight visual navigation algorithm in conjunction with exchange of information about visual landmarks seen in common among a team of vehicles.

Visual homing, e.g., (Liu 2012),[3] is a lightweight approach to visual navigation: using the stored visual information of a home location, a robot can navigate back to this location from any other location at which this location is visible by comparing the home image to the current image. It does not require a stored map of the environment and can be combined with obstacle avoidance functionality (Fu & Lyons 2018)[4] for generality. This makes visual homing very attractive for robot vehicles with a low computational capacity objective. Examples include small UAV drones and ground robots. A robot might store (or be given) multiple different home location images related to tasks or activities it needs to perform and home between them as necessary.

A limitation of visual homing is that the stored home location must be within the field of view of the robot to start homing. It may be possible to break a path into 'line of sight' segments and home from one to another to reach an intially out of view location (Steltzer et al 2018).[5] However, this additional map storage departs from the lightweight visual homing approach, reducing its attractiveness for some applications. Our approach addresses the issue of the home location not being in the field of view by seeing the robot as a member of a team of robots (and perhaps fixed camera assets) operating in the wide-area over which navigation must occur, and then leveraging communiating between the robots in the team.

If robot A needs to travel to a specific home location, but that location is not in view, then robot A attempts to find another robot that can see the home location. If robot B can see the home location, then A and B must next identify a common visual landmark that can be used as an 'intermediate' home location that A can use to travel to the vicinity of B and from there to its home location. While this example uses only two robots and

Further author information: dlyons@fordham.edu one shared visual landmark, the approach generalizes to n robots and $n-1$ landmarks. This algorithm and the assumptions necessary for its success will be discussed in more detail in section 3; however, this paper will focus on the key supporting problem of robot A and robot B agreeing on a common landmark that can be used as an intermediate waypoint.

We call this common landmark problem the "Do you see what I see" or *DYSWIS* task. This is a non-trivial visual problem to address: apart from the usual issues of lighting, scale and orientation, the components of a visual scene (objects) may appear in different spatial locations when seen from two distant vantage points. The principal result reported in this paper is an experimental study of several approaches to the DYSWIS problem for two robots. The input visual information is panoramic imagery from each robot: in our case, two sets of six images from a 60 degree FOV camera. The output from the problem is a set of paired image regions from the two sequences, where each paired image region represents a potential shared landmark.

We evaluate three approaches to this problem:

1. A SIFT feature-based approach, in which SIFT features are evaluated for each frame in both sequences, and each frame for the first robot is then matched against each for the second robot.

2. A CNN image-based approach in which object recognitions from pairs of frames, one from each robot, are compared.

3. A semantic approach that combines CNN and SIFT-based approaches.

The methods will be compared using a ROS/Gazebo simulation of a wide area suburban landscape. Candidate common landmarks will be evaluated using metrics such as landmark triangulation and comparison with the Gazebo 3D model locations. We will demonstrate that the CNN-based approach improves dramatically on the SIFT based approach, and the semantic approach improves on both

The next section will present an overview of the relevant existing work and motivate our contribution. Section 3 presents our novel approach to wide area visual navigation and describes the common landmark (DYSWIS) problem. Section 4 presents the framework and evaluation for the feature-based and the CNN-based approaches. Section 5 introduces the semantic approach and compares it to the previous two. The final section reviews our results and discusses future work.

2. PRIOR LITERATURE

Prior work on GPS denied navigation has focused on the problem of localization. This problem is crucial for UAVs that need accurate estimates of velocity for stability. Indoor UAVs often use an instrumented flight area for this reason; however, outdoor UAVs don't have that choice. Instead IMU data is very typically used to stabilize the platform (Bachrach 2014).[2] Scaramuzza et al[6] use onboard visual sensing for localization. They also integrate visual sensing data from a team of UAVs to form a wide-area SLAM map, though this is done off-line. Druen (Druen 2020)[7] addresses GPS-denied localization for a mobile ground platform, employing a zero velocity update algorithm to generate accurate estimates of distance travelled. While our focus is also on navigation without GPS, we address the issue of wide-area navigation rather than stabilization or localization. As such, the most direct comparison therefore is with Scaramuzza et al.'s generation of a wide-area map for use in what they refer to as global navigation. But our objective is to carry out wide-area (global) navigation without the need for such a map, dispensing with the cost and off-line generation latency/time and the need to keep such a map up to date with the motion of objects and with changing appearances.

Visual homing is a lightweight approach to visual navigation, initially inspired by study of insect navigation, but applied extensively in robotics (we do not list all prior homing work for space reasons). There are a number of algorithms for visual homing of which the most straight-forward is the Average Landmark Vector (ALV) algorithm (Lambrinos et al. 2000).[8] Correspondance methods (often using SIFT) improve on ALV, removing the need for global compass information (Zhu 2012).[9] Nirmal (Nirmal & Lyons)[10] further enhance the correspondance approach using stereovision-derived depth information.

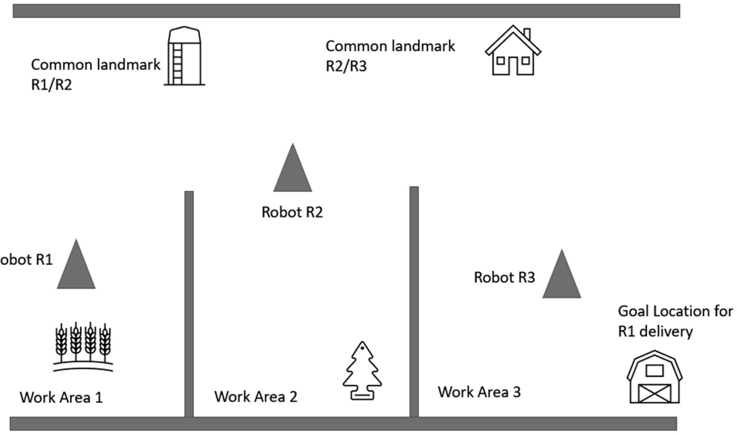

Figure 1. Wide Area Navigation Example

While visual homing avoids the memory requirement and sensor uncertainty issues associated with maintaining a metric map, it only works if the target location is initially within view. Steltzer et al.[5] develop a topological map approach, trail maps, in which intermediate visual waypoints are calculated and stored to bridge travel between distant locations. Although this is a lighter data structure than that of Scaramuzza, it still has the disadvantages that it requries exploration and time to generate, and it is sensitive to appearance change in long-term deployments. Our approach is to leverage the instantaneous geographic spread of a team of autonomous robots (and possibly fixed camera assets), exchanging visual information to identify the intermediate landmarks. The approach relies on an assumption that the team is continually distributed across the wide area to be navigated in such a manner that they can see some areas in common. In fact, this assumption is not unusual; robot teams that do metric mapping (e.g., Reid and Braunl,[11] Scaramuzza et al[6]) also have coverage and overlap requirements.

3. GPS-DENIED WIDE AREA VISUAL NAVIGATION

Our application scenario focuses on team of robots operating for long durations in an outdoor area or an area not well explored. Examples include a team of autonomous reconnaissance vehicles moving through a novel, outdoor area, or a team of autonomous agriculture vehicles working in a large outdoor area through all times, weathers and seasons. GPS navigation is the principal choice for navigation for these applications. However, in both example cases there are situations where GPS may not be available. GPS may be purposefully denied by adversaries or other bad actors looking to interfere with reconnaissance operations or with automated agricultural infrastucture. Our main motivation is not so much GPS denial as GPS reliance. Precision GPS is a mainstay of automated agricultural[1] and its cost and setup can be a barrier to small farmers leveraging automated agrculture to improve their competitive position with respect to large agribusiness.

3.1 Wide Area Visual Navigation

We address wide-area navigation for a team of robots using visual homing, assuming that GPS will be unavailable and that an exhaustive topological map will be unfeasible. Fig. 1 shows an example of wide-area navigation in an agricultural application. Robot R1 needs to deliver its load of produce to a distant temporary produce storage area. R1 knows what the area looks like (or has been instructed by communication) and searches in its immediate visual panorama for a match but cannot find one. It concludes that the goal location is not in view, and it must utilize information from other robots (or fixed camera assets) distributed across the work area.

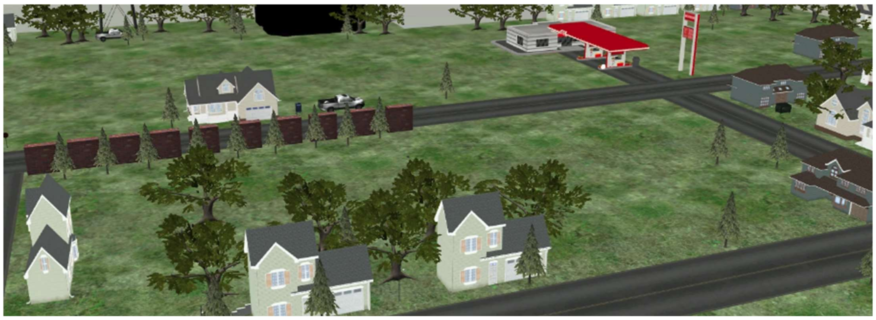

Figure 2. Section of the ROS/Gazebo suburban simulation used for testing.

R1 broadcasts its visual target information and requests every robot in the team to check for the goal location in their visual panoramas. In this example, robot R3 replies that the location is in view. R1 and R3 now attempt to identify a visual landmark that they both can see, a common landmark. They fail in this example, initiating a search for any robot that has a common landmark with R3; that will just be R2 in our example. R1 then successfully attempts to find a common landmark with R2. R1 now has a sequence of visual waypoints to traverse to bring it to the vicinity of the goal location.

When this approach is generalized for a team of n robots, the common landmark search can be carried out as a breadth-first tree search (in the absence of any way to order all the robots that see a common landmark). Assuming that at most b robots see any common landmark, the worst case search complexity for common landmarks is $O(b^d)$ for maximum path length d, where $d \approx log_b(n+1)$. For example, for a team of $n = 50$ robots where at most $b = 3$ can see a common landmark, the longest path will have $d = 4$ intermediate waypoints.

3.2 Common Landmarks

We argue that the key novel problem in this approach is the robust identification of common landmarks, the "do you see what I see" (DYSWIS) problem, and this is the main topic addressed in the remainder of the paper. Evaluating approaches to this problem is difficult because it requires labelling whether a candidate common landmark identified in the visual panorma for each robot does indeed refer to the same object. The process is simplified if the common landmark approaches are evaluated first using 3D simulation: Because the simulation model file contains all the objects that can give rise to visual images and their locations and orientations, it's possible to triangulate candidate common landmarks and determine whether there is a model at the triangulated location.

Fig. 2 shows an example scene from our ROS/Gazebo suburban simulation. The simulation models a $130 \times 180m^2$ flat, suburban area filled with grass, trees, buildings, vehicles and other readily available models, over 180 models in total. To evaluate a proposed approach to identifying common landmarks, robots are placed at two locations (between 2 and $20m$ apart), and the visual panoramas stored (e.g., Fig. 3). Common landmarks are identified using one of the methods to be evaluated and triangulation used to validate the landmark. The benefit of simulation is that this process can be repeated for many methods and for many positions with no manual intervention.

4. IDENTIFYING COMMON LANDMARKS: THE DYSWIS PROBLEM

The input to the DYSWIS problem is two sets of panoramic imagery. In our case, this will be two sets of six images from a 60^o FOV camera, one from each robot:

$$P_r = \{I_\alpha : \alpha \in \Theta\}, r \in \{A, B\}, \Theta = \{-120, -60, 0, 60, 120, 180\} \qquad (1)$$

The images are taken at angular intervals of 60^o starting at -120^o as shown in Fig. 3(a,b). In our case, the robot is rotated to each angle in turn to generate the panorama. It would also be possible to simply have six

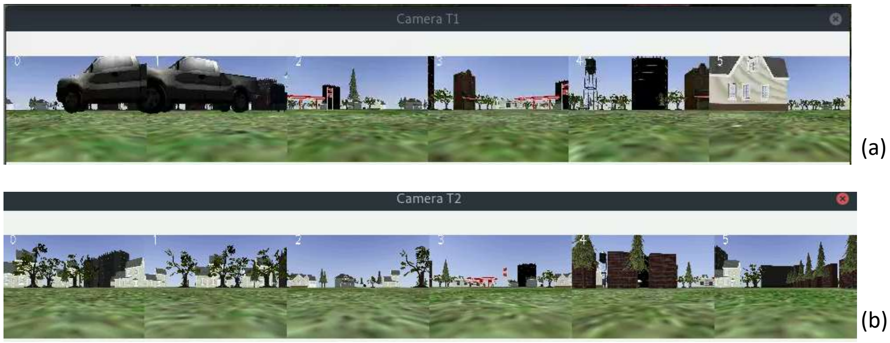

Figure 3. Example visual Panorama, robot A (a) and B (a).

separate cameras. Either approach provides better resolution than a panoramic camera with COTS equipment. Prior work has shown that panoramic stiching of the imagery is not crucial (Nirmal & Lyons 2015).[10]

4.1 Features-based approach

The first DYSWIS approach proposed is to extract SIFT features from each image and then to match each image from robot A against each from B. Any part of A's panorama that can also be seen from B should result in a common landmark. The first step is to extract the SIFT features:

$$s_{r,\alpha} = sift(I_{r,\alpha}),\ where\ I_{r,\alpha}\ is\ I_\alpha \in P_r \tag{2}$$

The match matrix $M_{A,B}$ is constructed by matching features from each image of A with those from each image of B, using RANSAC to eliminate outliers, and quantifying the goodness of match, written $m_{\alpha,\beta}$, as the average distance between matched features for the two images $I_{A,\alpha}$ and $I_{B,\beta}$.

$$M_{r,q} = [m_{\alpha,\beta} = m(s_{r,\alpha}, s_{q,\beta})]\ where\ \alpha, \beta \in \Theta \tag{3}$$

where $m(s, s')$ matches SIFT feature lists s and s' returning the average goodness of match. The set of common landmarks $CL_{A,B}$ is the set of thresholded, best matches seen by robot A:

$$CL_{r,q} = \{(\alpha, k) : k = argmax_\beta\ m_{\alpha,\beta}\ \&\ k > \tau\}\ for\ goodness\ threshold\ \tau \tag{4}$$

Fig. 4(a) shows an example common landmark identified by this approach. The image from robot A is on the left and B on the right. The colored lines between images join the image locations of matching SIFT features. The approach also generates a number of matches that are not common landmarks, e.g., Fig. 4(b).

4.2 CNN-based approach

A key characteristic for a part of a scene corresponding to a common landmark is that it is the 2D image of a 3D object in the scene such as a tree, a building or a car. Therefore a reasonable second DYSWIS approach would be to carry out object recognition on each image $I_{r,\alpha}$. Yolo (Redmon & Farhadi 2018)[12] differs from other CNN-based approaches to image recognition because it only requires a single pass to generate multiple object class match probabilities and their associated image regions. It employs a CNN architecture with 24 convolutional layers followed by 2 fully connected layers. The image is divided into a fixed-size grid, and when an object is recognized by a grid cell, that grid is responsible for predicting the object class probability and bounding box. Yolo offers a convenient and fast method to propose what regions of an image correspond to 3D objects in the scene. If SIFT matching is restricted to these regions then effects such as can be seen in Fig. 4(b) may be eliminated.

We use a CUDA-enabled YOLOv3* running on an NVIDIA GeForce RTX 3080 trained on the open images dataset[13] at 30 fps. This dataset includes many object classes with corresponding graphical models in our

*https://pjreddie.com/darknet/yolo/

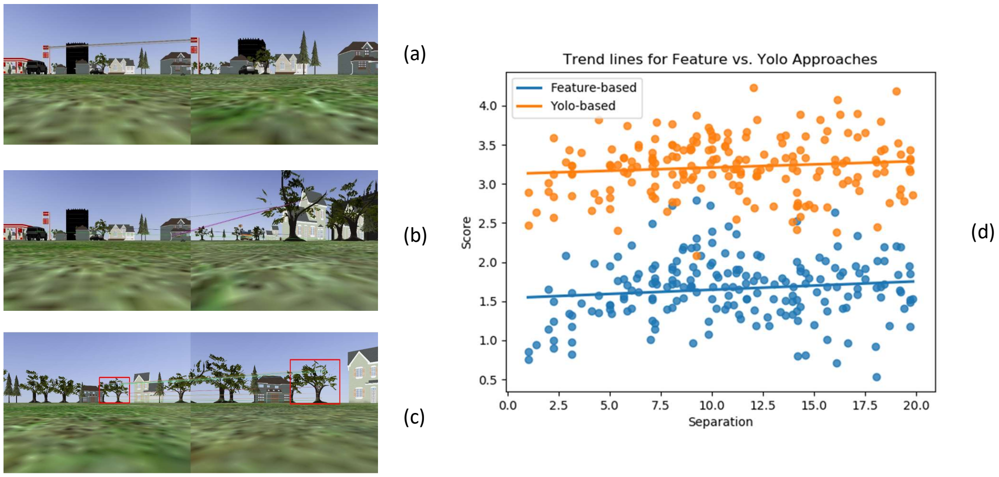

Figure 4. Example common landmark images, SIFT (a,b), Yolo(c); Comparison of SIFT and Yolo (d).

Gazebo subsurban scene. Although Yolo was not trained on graphical imagery from our simulation, it has no difficulty recognising the houses, vehicles and others parts of the scene. We integrate Yolo functionality into common landmark detection by modifying eq. (2) to include a function composition as follows

$$s_{\alpha,r} = sift \circ Yolo_1(I_{r,\alpha}) \tag{5}$$

where $Yolo_1(I)$ returns bounding box region of the highest probability object detection in image I. Common landmarks can be identified as before using eq.(4). Fig. 4(c) shows an example common landmark recognition using this method. The red bounding box outline can be seen in each corresponding image; SIFT matching has now been limited to features within these regions.

4.3 Evaluating approaches

To objectively rate which of these approaches is better, we randomly generate pairs of positions (n=200) in our ROS/Gazebo suburban scene, place the robots A and B at these postions, and collect all common landmarks using both methods. The positions were generated with separations between $2m$ and $20m$.

The locations and names of all simulation models were automatically extracted from the Gazebo world file to a list $Models$. Each common landmark extracted by the feature-based or Yolo-based DYSWIS method was scored as follows:

1. For the feature-based approach: are the matches widely distributed across the image (e.g., Fig. 4(b)) or more spatially concentrated (e.g., Fig. 4(a,c))?
2. For the Yolo approach, do both robots agree on the object class and is the triangulated Gazebo model of the same class?
3. For both: triangulating a 3D model location using the image location of the landmark in each image, does the Gazebo simulation have an object model close to the location?

Triangulation of landmark location is done as follows: First, the centroid of the landmark is calculated for each image. For the feature based approach, this is the average location of matched features; for Yolo, this is the center of the bounding box. The horizontal coordinate of the centroid is converted to a horizontal angle θ:

$$\theta = (\bar{x} - 0.5w)\frac{FOV}{w} + \gamma, \gamma \in \Theta \tag{6}$$

where γ is the robot pose at which the image was captured, w is the image width, and FOV is the camera field of view. The ray from the position of robot A with angle θ_A is tested for intersection with that from B at angle θ_B to get the predicated landmark location $lm = ray(A, \theta_A) \cap ray(B, \theta_B)$, and if lm exists, $Models$ is searched for a model within radius $\epsilon = 10m$ of lm.

4.4 Experimental method and results

The results of comparing all feature-based and all Yolo-based common landmarks for n=200 randomly generated pairs of positions are presented in this section. The feature based approach located 1232 common landmarks, and the Yolo-based approach identified 1135 common landmarks in the 200 trials.

Figure 4(d) shows a scatter graph of non-zero scores for both approaches plotted against the separation between the robots. The orange points show the Yolo scores and the blue shows the feature-based scores. The Yolo method can be seen to dominate across all separations. The graph includes trend lines for both scores, and both trend up slightly with separation. Feature based scores have a mean score of 1.63 with a standard deviation of 0.44, while Yolo scores have a mean score of 3.12 with a standard deviation of 0.66. Yolo scores are significantly better than the feature-based scores (t-test $p = 1.24 \times 10^{-18}$).

5. SEMANTIC APPROACH FOR THE DYSWIS PROBLEM

Although theYolo-based approach (Yolo$_1$) clearly outperforms the feature-based approach, it occasionally suffers from false matching of candidate regions from each robot. An example is shown in Fig. 5(a): Yolo$_1$ for each robot has selected a pine tree model, but each selected a different tree from the group of trees. Although we would predict that this aliasing problem is less severe in natural scences, we still expect it to be significant. For example, two different red sedans in a scene might still easily be mismatched. Our final proposed approach addresses this issue: Rather than looking for common landmarks consisting of a single object, which could be easily aliased, we will instead add the constraint that a landmark consist of a cluster of objects. We call this the semantic approach because it leverages higher-level object groupings rather than single objects or (just) image features.

Yolo$_1$ will return the bounding box of the highest probability object detection from Yolo. We add to this with Yolo$_2$, returning the top two probability detections, and Yolo$_3$, the top 3 detections. The framework of eq.(5) will be modified to return a set $S_{r,\alpha}$ rather than a single value:

$$S_{r,\alpha} = sift \ \circ \ \{Yolo_n(I_{r,\alpha i}) : i \in \{1, \ldots, n\}\}, n = 2, 3 \tag{7}$$

The generation of the match matrix M needs to be modified, since now the input to matching is two sets of regions and not two regions. This is a classical data association problem: Associate each region from $S_{r,\alpha}$ with one from $S_{q,\beta}$ so that the sum of the n region matching scores is a maximum. The Kuhn-Munkres algorithm is a typical solution for this:

$$M_{r,q} = [m_{\alpha,\beta} = km(S_{r,\alpha}, S_{q,\beta})] \ where \ \alpha, \beta \in \Theta \tag{8}$$

where $km()$ applies the Kuhn-Munkres algorithm and returns the maximum score sum. Eq.(4) can still be used, though now with M being a maximum score sum matrix, to pick common landmarks.

Fig. 5(b-d) presents an example match for $Yolo_3$: Each robot generates the top 3 object recognitions for its image, shown here as red, green and blue bounding boxes (in order of probability of object recognition for that robot). Kuhn-Munkres is used to select an optimal data association, pairing regions from each robot image based on the SIFT matching goodness scores. For each image in its panorama, robot A selects the data association with robot B's panorama images that maximizes the sum of match scores. Fig. 5(b) through (c) shows the highest value data association pairing for the leftmost image from robot A.

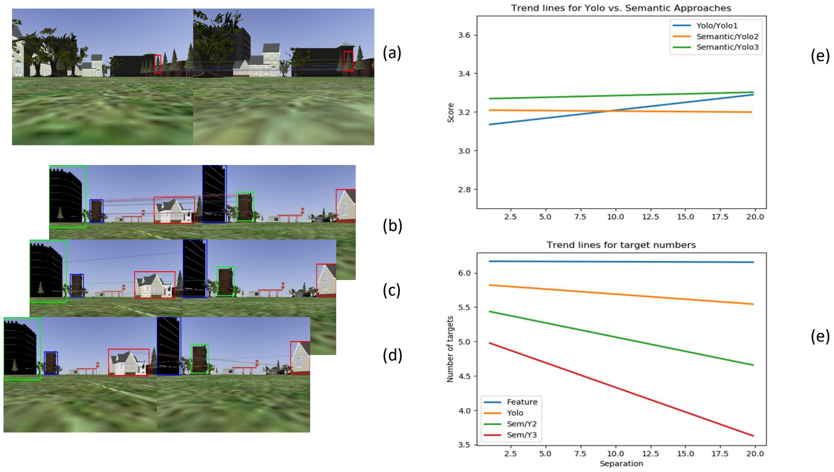

Figure 5. Example of poor detection with Yolo1 (a); Example of semantic approach, 3 cluster (b-d); Performance of semantic approaches (e,f).

5.1 Experimental method and results

The evaluation procedure employed in section 4.4 was also employed here. The same n=200 paired locations were used to compare landmarks generated by the single Yolo landmark approach and the semantic approaches consisting of clusters of Yolo-identified landmarks. The clusters were scored similarly to the single Yolo landmarks (section 4.3), with the score for cluster being the maximum score for any of the cluster's pairs of regions. The rational for this choice was that selecting a sum of scores would complicate the comparison to the single Yolo landmark approach, and selecting average generated unfairly low scores in the event that only two regions could be matched, even if they matched very well.

Fig. 5(e) shows a trend-line graph for evaluation score versus robot separation for three methods: Single Yolo landmarks, and the two semantic approaches: clusters of 2 landmarks and clusters of 3 landmarks. Clusters of 2 landmarks perform better than single Yolo landmarks for small separations, but after separations of approx. 9m, they become worse. Clusters of 3 landmarks perform better than the single Yolo landmarks for all separations tested, but the benefit drops with separation. While it is not possible to significantly distinguish the single Yolo and 2 cluster approaches, the 3 cluster approach is significantly better than the single Yolo approach ($p = 0.02$)

Fig. 5(f) shows a graph of average number of landmarks detected versus separation for the Feature-based approach, the single Yolo approach and the two semantic approaches. The Feature-based approach shows a constant number of landmarks, just over 6 per pair of robot positions, across all separations. We have seen that this approach fares the least well when the landmarks are evaluated. The Yolo-based approach generates a relatively consistent number of targets at all separations, but the number does drop slightly at larger separations. The trend is much more dramatic for the semantic approaches and for the cluster of 3 approach, an average of just greater than 3.5 cluster landmarks are seen per pair of robot positions when the pair separation is 20m. As separation between the pair increases, it becomes increasingly difficult to use objects identified by Yolo as common landmarks. Despite this, the score of the fewer such landmarks identified does increase, leading to fewer but better landmarks.

6. CONCLUSIONS

In this paper we have addressed the issue of *wide-area navigation* in the absense of GPS: traversal to a target destination beyond the immediate sensor range of a robot, and known only by its appearance. Our application scenario is a team of robots operating for long durations in an outdoor area or an area not well explored. Examples include a team of autonomous reconnaissance vehicles moving through a novel, outdoor area, or a team of autonomous agriculture vehicles working in a large outdoor area through all times, weathers and seasons. We proposed an approach that leverages a light-weight visual navigation algorithm, visual homing, in conjunction with exchange of information about visual landmarks seen in common among a team of vehicles. We argued that the key novel problem in this approach is the robust identification of common landmarks, the "do you see what I see" (DYSWIS) problem. We evaluated three approaches to the problem: a purely feature-based approach, looking for SIFT matches between the visual imagery of the two robots; a CNN-based approach, using Yolo to identify objects and then carry out SIFT matching; and a semantic approach, looking for groups of objects. We evaluted the approaches using a ROS/Gazebo simulation of a $130 \times 180m^2$ flat, suburban area filled with grass, trees, buildings, vehicles and other readily available models. Pairs of positions were randomly selected (n=200) with separations varying between 2 and 20 meters and all three methods used to select common landmarks, which were then evaluated by determining whether they were the images of object models in the simulation.

Our results show that the CNN-based approach was dramatically better than the feature based approach ($p = 1.24 \times 10^{-18}$). However, it suffered from an aliasing problem, likely exacerbated by the use of multiple instances of simulation models, but a problem we can also expect to see in non-simulation testing. Two semantic approaches were tested: clusters of 2 Yolo identified objects and clusters of 3. The clusters of 3 method did improve on the Yolo approach ($p = 0.02$) but seriously reduced the number of landmarks identified, yielding fewer but better landmarks.

Our next step is to validate these results in lab testing using a physical landscape of potential landmarks. We also believe that the performance of the semantic approach can be improved by looking at the *diversity* of objects types seen in clusters: we predict a higher diversity will reduce aliasing even further than we see now.

REFERENCES

[1] DeBaerdemaeker, J., "Precision agriculture technology and robotics for good agricultural practices," *IFAC Proceedings* **46**(4) (2013).

[2] Bachrach, A. et al., "Range - robust autonomous navigation in gps-denied environments," *IEEE/RSJ Int. Conf. on Int. Rob. & Sys. (IROS)* (2010).

[3] Liu, M., Pradalier, C., Pomerleau, F., and Siegwart, R., "The role of homing in visual topological navigation," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)* (2012).

[4] Fu, F. and Lyons, D., "An approach to robust homing with stereovision," *SPIE Defense & Security 2017 Conference on Unmanned Systems Technology XX* (April 2018).

[5] Stelzer, A., Vayugundla, M., Mair, E., Suppa, M., and Burgard, W., "Towards efficient and scalable visual homing," *Int. J. Robotics Research* **37**(2-3) (2018).

[6] Scaramuzza, D. et al., "Vision controlled micro flying robots," *IEEE/ Rob. & Aut. Magazine* (Sept. 2014).

[7] Druen, S., "Robotic navigation in gps-denied environments using the strapdown navigation algorithm with zero-velocity updates," *M.S. Thesis, Naval Postgraduate School* (2020).

[8] Lambrinos, D. et al., "Mobile robot employing insect strategies for navigation," *Robotics and Autonomous Systems* **30** (2000).

[9] Zhu, Q., Liu, C., and Cai, C., "A novel robot visual homing method based on sift features," *Sensors* **15** (2015).

[10] Nirmal, P. and Lyons, D., "Homing with stereovision," *Robotica* **34**(12) (2015).

[11] Reid, R. and Braunl, T., "Large-scale multi-robot mapping in magic 2010," *IEEE 5th International Conference on Robotics, Automation and Mechatronics (RAM)* (2011).

[12] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," *arXiv* (2018).

[13] Krasin, I. et al., "Openimages: A public dataset for large-scale multi-label and multi-class image classification.," *Dataset available from https://github.com/openimages* (2016).