Multi-Modal Sensor Fusion and Selection for Enhanced Situational Awareness

Brian Reily[a], Christopher Reardon[b], and Hao Zhang[a]

[a]Colorado School of Mines, Golden, CO USA
[b]University of Denver, Denver, CO USA

ABSTRACT

Collaborative multi-sensor perception enables a sensor network to provide multiple views or observations of an environment, in a way that collects multiple observations into a cohesive display. In order to do this, multiple observations must be intelligently fused. We briefly describe our existing approach for sensor fusion and selection, where a weighted combination of observations is used to recognize a target object. The optimal weights that are identified control the fusion of multiple sensors, while also selecting those which provide the most relevant or informative observations. In this paper, we propose a system which utilizes these optimal sensor fusion weights to control the display of observations to a human operator, providing enhanced situational awareness. Our proposed system displays observations based on the physical locations of the sensors, enabling a human operator to better understand where observations are located in the environment. Then, the optimal sensor fusion weights are used to scale the display of observations, highlighting those which are informative and making less relevant observations simple for a human operator to ignore.

Keywords: Multi-robot, sensor fusion, situational awareness

1. INTRODUCTION

Sensor fusion and selection enables a sensor network or multi-robot system to provide collaborative perception, fusing multiple observations in order to gain a unified understanding of an environment. Sensor networks provide observations of objects and scenes from multiple perspectives.[1] In complex environments, such as those encountered in search and rescue or military operations,[2,3] individual sensors can be obstructed or interfered with. In order to effectively fuse observations from a sensor network where some observations are not relevant, it is critical to be able to select the most informative observations and rely on these for object and scene recognition tasks.

Following the successful fusion of multiple sensor observations and the selection of the most relevant, a human operator must be made aware of this information. As sensor networks can provide a large number of observations, the amount of information available can be overwhelming and exceed the cognitive load limit of a human. In order to provide enhanced situational awareness, these observations must be displayed in an intelligent manner, highlighting the most relevant and making the less informative observations simple to ignore. In this way, situational awareness can be enabled without exceeding the cognitive processing abilities of a human operator.

In this paper, we propose a novel approach to enhanced situational awareness based on fusion of sensors in a sensor network and the selection of relevant sensor observations. We describe a formula based on regularized optimization that unifies sensor fusion for recognition with the selection of relevant observations and feature modalities. We present sparsity inducing norms in order to identify only a small number of observations and modalities. A linear combination of observations is used to approximate target objects or scenes, identifying the target with the smallest approximation error. Then, spatial positions of sensors and the identified optimal sensor fusion weights are utilized to display the observations to a human operator. By integrating sensor position

Further author information (Send correspondence to Brian Reily):
Brian Reily: breily@mines.edu
Christopher Reardon: christopher.reardon@du.edu
Hao Zhang: hzhang@mines.edu

and importance into the display, we provide enhanced situational awareness. The key contribution of this work is the proposal of utilizing a weighted selection of sensor observations to enhance situational awareness of an environment.

The remainder of this paper is structured as follows. We discuss related work in Section 2. We describe our approach for multi-modal sensor fusion and selection in Section 3. In Section 4, we describe the use of the sensor fusion weights for enhanced situational awareness. Finally, we conclude the paper in Section 5.

2. RELATED WORK

This proposed work of sensor fusion and selection builds on research in sensor coverage, active perception, and multi-view perception. We briefly review existing work in each of these areas.

Sensor coverage is the problem of placing or controlling robots in order to maximize the overall observation of an environment. While this is often done with multi-robot systems, many approaches have been developed that identify fixed sensor placements that maximally cover an environment.[4–6] The coordination of multiple robots is more common though, often as a precursor to collective sensing with observations merged together.[1] This can be done by dividing multi-robot systems to focus on specific sub-areas of an environment,[7] or dividing[8] or deploying[9,10] multi-robot systems based on the sensor capabilities they possess. In addition, approaches have considered integrating observations from sensors located on agents performing other tasks instead of coordinating them specifically to maximize sensing (such as cars driving through an environment). Research in areas such as this focus on tasks like correspondence identification, or recognizing which objects in each observation are the same object.[11] Our work takes as input observations from sensors and does not require the ability to coordinate or move them, so can operate with an active multi-robot system or from observations gained elsewhere (such as cars).

The second research area of interest is active perception, where sensors are controlled to obtain optimal views of objects or scenes, often by adjusting their positions or settings.[12] This is distinct from multi-robot sensor coverage as it can focus on tracking specific objects or planning paths to observe specific areas of an environment.[13,14] Previous work here has utilized information measures such as entropy[15] or control measures such as scheduling algorithms[16] or particle filters.[17]

Finally, the field of computer vision has focused on recognition tasks through multi-view perception. These multiple views could be obtained from a sensor network, or a single sensor over time. Recognition based on multiple views has shown promising results on human activity recognition,[18] walking gait recognition,[19] and object recognition.[20,21] While these approaches enable accurate recognition, they can be very computationally expensive as they involve more data than single view approaches. In light of this, methods have been developed to identify the most representative views[22,23] or to select two-dimensional views that can most accurately reconstruct a three-dimensional model.[24–27] Methods to integrate multiple views into unified recognition approaches have utilized neural networks[28] and have also been based on more non-traditional ideas such as semantics[29] or search algorithms.[30]

3. SENSOR FUSION AND SELECTION

In this section, we briefly describe our approach for sensor fusion and selection. * This approach provides target recognition by minimizing the approximation loss between a linear combination of input sensor observations. This linear combination, which weights each sensor observation in order to fuse them, also acts to select the most relevant or informative observations.

We begin by first defining the targets that our sensor network is used to identify. These targets can be objects, areas, faces, etc. We denote p types of targets as $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_p] \in \mathbb{R}^{d \times p}$, where $\mathbf{t}_i \in \mathbb{R}^d$ is a feature vector representing a reference observation of the i-th target. Because of this need for a reference observation, some prior knowledge of the target is necessary. This vector is of dimensionality d, with $d = \sum_{m=1}^{M} d_m$, where d_m represents the dimensionality of the m-th sensing modality provided by the sensing network, which can provide

*We note that this approach is previously described in a work under review, and is not the novel focus of this manuscript. It is described briefly to provide context for later sections.

M different modalities (e.g., RGB camera images, thermal sensor observations, etc.). The number of targets available to the sensing system, p, can be changed as the system operates - for example, as mission requirements adjust, targets can be added or removed as appropriate.

We then denote observations from n sensors as $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] \in \mathbb{R}^{n \times d}$, with $\mathbf{x}^j \in \mathbb{R}^d$ denoting the current observation of the j-th sensor. As with the reference observation of each target, the observation from each sensor is of dimensionality d, or the sum of the dimensionality of each available modality. Our approach then attempts to approximate each reference target through a linear combination of sensor observations:

$$\min_{\mathbf{w}} \|\mathbf{X}^\top \mathbf{w} - \mathbf{t}_i\|_2^2 \tag{1}$$

By finding a sensor fusion weighting $\mathbf{w}$ that minimizes this loss function for each target $\mathbf{t}_i$, we both find an optimal weighted combination of sensors ($\mathbf{w}$) but are also able to identify the target where the approximation error is the lowest ($\mathbf{t}_*$).

Our approach then introduces three regularization terms to aid in identifying an optimal $\mathbf{w}$, with the assumption that only a small number of the n available sensors and the m available modalities should be informative:

$$\min_{\mathbf{w},\mathbf{u}} \|\mathbf{X}^\top \mathbf{w} - \mathbf{t}_i\|_2^2 + \lambda_1 \|\mathbf{X}\mathbf{u} - \mathbf{w}\|_2^2 + \lambda_2 \|\mathbf{w}\|_1 + \lambda_3 \|\mathbf{u}\|_M \tag{2}$$

Here, $\lambda_k, k = \{1, 2, 3\}$ represent adjustable hyperparameters that control the importance of each regularization term. The three introduced terms have the following effects:

- $\|\mathbf{Xu} - \mathbf{w}\|_\mathbf{2}^\mathbf{2}$: This term relates the sensor fusion weighting $\mathbf{w}$ to $\mathbf{u}$, which weights the modalities available from the sensor network.
- $\|\mathbf{w}\|_\mathbf{1}$: This term applies the ℓ_1-norm on the sensor fusion weighting $\mathbf{w}$, encouraging sparsity to identify the most informative sensors.
- $\|\mathbf{u}\|_\mathbf{M}$: This term applies the ℓ_1-norm between sections within $\mathbf{u}$ that apply to each modality, encouraging the identification of discriminative modalities.

The final approach presented in Eq. (2) is minimized for each target $\mathbf{t}_i$. This identifies the target recognized by the sensor network, and also identifies an optimal sensor fusion weighting $\mathbf{w}$, which is then used to provide enhanced situational awareness, the primary focus of this work.

4. ENHANCED SITUATIONAL AWARENESS

Our main approach to provide enhanced situational awareness from a sensor network through sensor fusion and selection is described in this section. This is done through two main focuses: first, sensor observations are displayed visually based on the spatial relationships of the sensor network; and second, the visual display size of sensor observations is controlled by the optimal sensor fusion weighting $\mathbf{w}$. These two focuses mean that observations from a large sensor network can be displayed in an informative manner, by being located relevantly and by highlighting relevant observations while making less relevant observations easier to ignore. The effect of our approach is illustrated through simulation in Figure 1, with the target being an observation of the grey car.

First, we consider Figure 1. This shows an overhead view of the scene, where a multi-robot system is acting as the sensor network. The simulation environment provides both RGB images and depth camera observations of the scene, but only RGB images are displayed in later figures for simplicity. We can see that this system of five robots is providing views of the target object from a variety of angles and distances, with some obstructed by obstacles.

Figure 1(b) shows a presentation of these observations when neither spatial relationships nor relevance is considered. While this presentation provides information to the user, it is difficult to identify which sensors are providing which views, or to easily identify which views are the most informative. Naive systems that display data as such not only fail to identify view relevance, but also dismiss the valuable information provided by sensor locations.

(a) Scene Overhead

(b) Views

(c) Distributed Spatially

(d) Weighted Observations

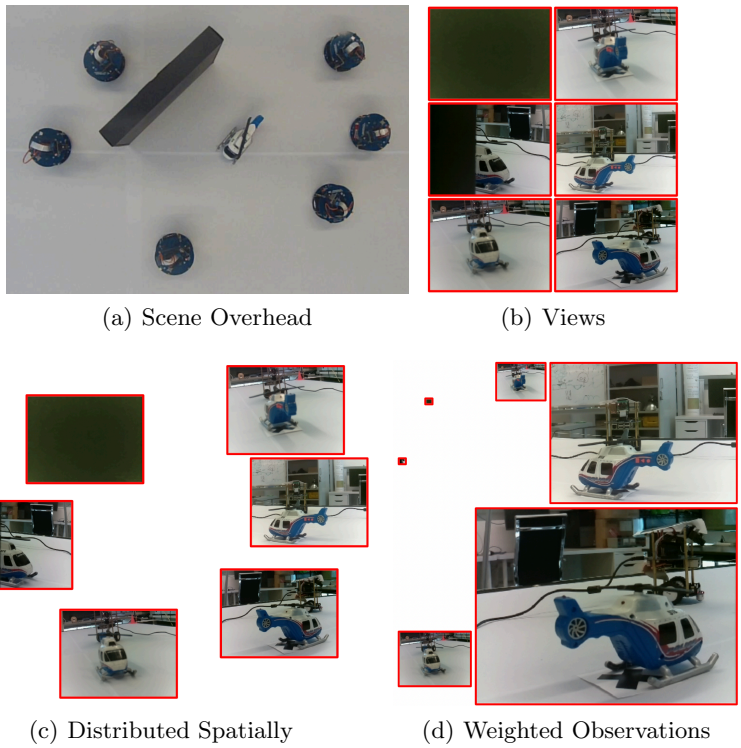
Figure 1. This series of figures shows the affect of our proposed approach. Figure 1(a) shows an overhead view of a multi-robot sensor network. Figures 1(b) - 1(d) show a progression between just presenting views normally and presenting views through our approach, which provides enhanced situational awareness.

Figure 1(c) shows the effect of the first focus of our approach. In this case, each observation is still displayed at a fixed size (possible varying from Figure 1(b) due to scaling in the manuscript). However, as opposed to a naive display, observations are now shown based on the physical locations of the sensors providing them. With fixed sensor network (such as security cameras) we would have defined measurements with their placements, and with a sensor network based on a multi-robot system such as this we could have GPS or some other localization method (e.g., SLAM) providing sensor locations. Our proposed approach takes advantage of this by displaying observations based on these known spatial locations - i.e., for each sensor observation $\mathbf{x}^j$, we have a known $(x, y)^j$ or $(latitude, longitude, altitude)^j$. By displaying observations at locations that visually relate to their real-world locations, we enable human operators to take advantage of this extra context.

Finally, our full proposed approach is based on the second focus, where we believe enhanced situational awareness is best provided by displaying the most relevant or informative observations to a human operator. To do this, we utilize the sensor fusion weight vector $\mathbf{w}$, where each element $w_k \in \mathbf{w}$ represents the weight of observation k. As we do not constrain the total sum of weights in $\mathbf{w}$, we define it such that $w_{total} = \sum_{i=1}^{n} w_i$. From this, we can determine that the display size each observation should merit is equal to its relevance:

$$\text{size}_k = w_k / w_{total} \tag{3}$$

We can the use this generate enhanced situational awareness displays such as that shown in Figure 1(d). In this display, observations are still spatially related to the sensor providing them, but their display sizes are scaled according to Eq. (3). In this figure, $\mathbf{w} \approx [0.025, 1, 0.25, 0.75, 0.125]$, corresponding to the order in Figure 2(b) and approximated for ease of explanation. We can observe that the most relevant view, provided by the second sensor, is visually the largest and a human operator would be very aware of it. The second most relevant, provided by the fourth sensor, is smaller but still visually useful. The more irrelevant observations, such as that of the first sensor, whose camera angle is completely obstructed by a guardrail, are scaled so small as to be easily ignored.



(a) Scene Overhead (b) Views

(c) Distributed Spatially (d) Weighted Observations

Figure 2. Further evaluation of our proposed approach on a physical system. Figure 2(a) shows an overhead view of a multi-robot sensor network. Figures 2(b) - 2(d) show a progression between just presenting views normally and presenting views through our approach, which provides enhanced situational awareness.

We provide further evaluation in Figure 2, on a physical multi-robot system acting as the sensor network. As with the earlier figure, Figure 2(a) shows an overhead view, with six robots observing a helicopter, with some views obstructed by a wall. Figure 2(b) shows a display of these observations with no localization or informative scaling. Non-relevant views occupy at least a third of this display, with no information about the sensor network's spatial layout provided. Figure 2(c) shows this scene displayed from the knowledge of the $(x, y, z)^j$ locations of each observation $\mathbf{x}^j$, which is provided by the overhead camera. This display provides enhanced situational awareness already, by showing that sensors on the left of the target are impeded by an obstacle. Our full approach is then shown in Figure 2(d), where the two most relevant views are highlighted, and two extremely irrelevant views (with $(w_1 + w_2) \lll 1\%$ of w_{total}, where w_1 is the top-left of Figure 2(b) and w_2 is the first-column, second-row) are almost not displayed - that is, these are the two views visible only from their red outlines.

5. CONCLUSION

Collaborative perception enables a sensor network to perceive an environment from multiple perspectives. By fusing these sensors intelligently, we can identify the sensors which have the best views, in order to obtain an optimal understanding of the environment. We provide an overview of an approach that can provide accurate recognition by fusing the observations of multiple sensors while also identifying the importance of individual sensors. Then, our novel contribution to enhance situational awareness incorporates sensors' physical locations and observation relevance, in order to display observations to a human operator based on their importance and their positions. Through evaluation in both simulation and on a physical multi-sensor system, we show that our approach can display observations in a way that provides awareness of both sensor location and importance.

REFERENCES

[1] Schmickl, T., Möslinger, C., and Crailsheim, K., "Collective perception in a robot swarm," in [*International Workshop on Swarm Robotics*], (2006).

[2] Baxter, J. L., Burke, E., Garibaldi, J. M., and Norman, M., "Multi-robot search and rescue: A potential field based approach," in [*Autonomous Robots and Agents*], 9–16 (2007).

[3] Correll, N. and Martinoli, A., "Multirobot inspection of industrial machinery," *Robotics & Automation Magazine* **16**(1), 103–112 (2009).

[4] Parker, L. E. and Emmons, B. A., "Cooperative multi-robot observation of multiple moving targets," in [*International Conference on Robotics and Automation*], (1997).

[5] Ma, J. and Burdick, J. W., "Dynamic sensor planning with stereo for model identification on a mobile platform," in [*International Conference on Robotics and Automation*], (2010).

[6] Domingo-Perez, F., Lazaro-Galilea, J. L., Wieser, A., Martin-Gorostiza, E., Salido-Monzu, D., and de la Llana, A., "Sensor placement determination for range-difference positioning using evolutionary multi-objective optimization," *Expert Systems with Applications* (2016).

[7] Inacio, F. R., Macharet, D. G., and Chaimowicz, L., "Persistent monitoring of multiple areas of interest with robotic swarms," in [*Brazilian Symposium on Robotics*], (2018).

[8] Reily, B., Reardon, C., and Zhang, H., "Representing multi-robot structure through multimodal graph embedding for the selection of robot teams," in [*International Conference on Robotics and Automation*], (2020).

[9] Singh, A., Krause, A., Guestrin, C., and Kaiser, W. J., "Efficient informative sensing using multiple robots," *Journal of Artificial Intelligence Research* **34**, 707–755 (2009).

[10] Liu, J. and Williams, R. K., "Optimal intermittent deployment and sensor selection for environmental sensing with multi-robot teams," in [*International Conference on Robotics and Automation*], (2018).

[11] Gao, P., Guo, R., Lu, H., and Zhang, H., "Regularized graph matching for correspondence identification under uncertainty in collaborative perception," (2020).

[12] Chen, S., Li, Y., and Kwok, N. M., "Active vision in robotic systems: A survey of recent developments," *International Journal of Robotics Research* **30**(11), 1343–1377 (2011).

[13] Chiu, H.-P., Zhou, X. S., Carlone, L., Dellaert, F., Samarasekera, S., and Kumar, R., "Constrained optimal selection for multi-sensor robot navigation using plug-and-play factor graphs," in [*International Conference on Robotics and Automation*], (2014).

[14] Best, G., Faigl, J., and Fitch, R., "Multi-robot path planning for budgeted active perception with self-organising maps," in [*International Conference on Intelligent Robots and Systems*], (2016).

[15] Hausman, K., Müller, J., Hariharan, A., Ayanian, N., and Sukhatme, G. S., "Cooperative multi-robot control for target tracking with onboard sensing," *International Journal of Robotics Research* **34**(13), 1660–1677 (2015).

[16] Shi, K., Chen, H., and Lin, Y., "Probabilistic coverage based sensor scheduling for target tracking sensor networks," *Information Sciences* **292**, 95–110 (2015).

[17] Dietl, M., Gutmann, J.-S., and Nebel, B., "Cooperative sensing in dynamic environments," in [*International Conference on Intelligent Robots and Systems*], (2001).

[18] Spurlock, S. and Souvenir, R., "Dynamic view selection for multi-camera action recognition," *Machine Vision and Applications* **27**(1), 53–63 (2016).

[19] Kusakunniran, W., Wu, Q., Zhang, J., and Li, H., "Support vector regression for multi-view gait recognition based on local motion feature selection," in [*Conference on Computer Vision and Pattern Recognition*], (2010).

[20] Thomas, A., Ferrar, V., Leibe, B., Tuytelaars, T., Schiel, B., and Van Gool, L., "Towards multi-view object class detection," in [*Conference on Computer Vision and Pattern Recognition*], (2006).

[21] Wu, F., Jing, X.-Y., You, X., Yue, D., Hu, R., and Yang, J.-Y., "Multi-view low-rank dictionary learning for image classification," *Pattern Recognition* **50**, 143–154 (2016).

[22] Mokhtarian, F. and Abbasi, S., "Automatic selection of optimal views in multi-view object recognition," in [*British Machine Vision Conference*], (2000).

[23] Moreira, P., Reis, L., and De Sousa, A., "Best multiple-view selection for the visualization of urban rescue simulations.," *International Journal of Simulation Modelling* **5**(4) (2006).

[24] Laga, H., "Semantics-driven approach for automatic selection of best views of 3d shapes," in [*Eurographics Conference on 3D Object Retrieval*], (2010).

[25] Mendez Maldonado, O., Hadfield, S., Pugeault, N., and Bowden, R., "Next-best stereo: extending next best view optimisation for collaborative sensors," in [*British Machine Vision Conference*], (2016).

[26] Genova, K., Savva, M., Chang, A. X., and Funkhouser, T., "Learning where to look: Data-driven viewpoint set selection for 3d scenes," *arXiv preprint arXiv:1704.02393* (2017).

[27] Wang, D., Wang, B., Zhao, S., Yao, H., et al., "View-based 3d object retrieval with discriminative views," *Neurocomputing* **252**, 58–66 (2017).

[28] Kim, S.-h., Tai, Y.-W., Lee, J.-Y., Park, J., and Kweon, I. S., "Category-specific salient view selection via deep convolutional neural networks," in [*Computer Graphics Forum*], (2017).

[29] Guérin, J., Gibaru, O., Nyiri, E., Thieryl, S., and Boots, B., "Semantically meaningful view selection," in [*International Conference on Intelligent Robots and Systems*], (2018).

[30] Wang, W. and Gao, T., "Constructing canonical regions for fast and effective view selection," in [*Conference on Computer Vision and Pattern Recognition*], (2016).