Metacognitive Reasoning of Perceptual Inconsistency for Illusion Detection

Sriram Siva and Hao Zhang
Human-Centered Robotics Laboratory
Colorado School of Mines, Golden, Colorado 80401
Email: {sivasriram, hzhang}@mines.edu

I. Introduction

When autonomous robots operate in adversarial environments, such as in tactical battlefields, they may face various misinformation attacks, such illusion and deception, by potential adversaries. Different from conventional reactive design that reacts through analyzing the effects of attacks and developing countermeasures, we propose the insight of an active design by anticipating the adversary via investigating potential attacks.

Several techniques were developed in traditional adversarial applications to defend against misinformation attacks, including data sanitization [3] and model improvement [5, 2] to protect against causative attacks, and classifier randomization [1], near-optimal evasion protection [4], and robust feature selection [6, 7] to defend against exploratory attacks. However, previous methods are not capable of addressing the challenges in robot perception in illusive and deceptive scenarios. This work aims at improving the robustness of robot perception against illusion in data-rich environments with multisensory high-dimensional observations. Specifically, in this workshop paper, we introduce a metacognitive reasoning approach for robots to reason about consistency of multisensory perception data in order to detect illusion.

II. Approach

Multisensory perception data collected from heterogeneous sensors on robots can contain misinformation faked by hostile forces by using physical illusion to directly pollute data sources (or by misinformation attacks such as hacking sensors to insert incorrect data or modify observation modalities).

For example, as illustrated in Figure 1, a sniper wearing a woodland camouflage suit can fool color cameras and LiDARs, or a human dummy model can be utilized by hostile forces to mimic a real human to physically pollute perception data. This illusion could lead to inappropriate robot decisions that can potentially interrupt human teammates and hurt ongoing operations, which emphasizes the importance of active illusion detection.

To achieve this goal, our method is to explore the redundancy and dependency of multisensory observations and automatically analyze the inconsistency between the sensing modalities. For example in Figure 1, the sniper is less likely to be recognized as a human by a robot using color cameras, but more likely to be classified as woodland. On the other hand, observations acquired by thermal sensors can indicate that the target with a higher temperature is less likely to be woodland, which is not consistent with the color sensing modalities.

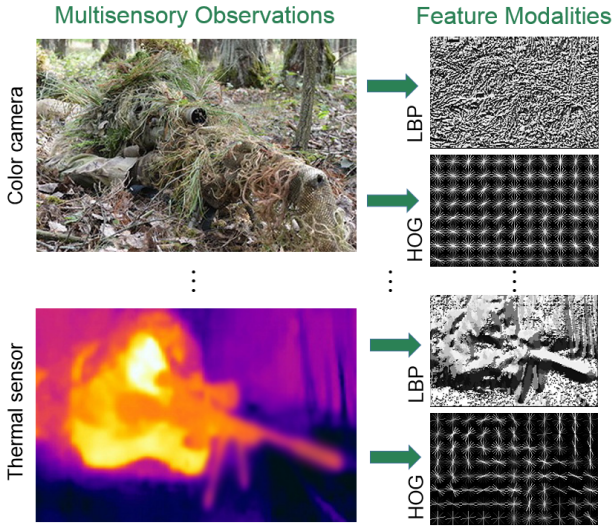

Fig. 1. Illustration of physical deception that contains inconsistency between color and thermal sensory modalities.

The proposed approach performs metacognitive reasoning about the inconsistency of multisensory observation modalities and detects sensory misinformation faked by physical illusion. Formally, given a collection of n trusted multisensory observations as training instances, the feature matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$, where $\mathbf{x}_i \in \Re^d$ represents the feature vector of the i-th observation consisting of m modalities such that $d = \sum_{j=1}^{m} d_j$. For a recognition problem of understanding a set of concepts (e.g., human recognition) represented by the category indicator matrix $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^\top \in \Re^{n \times c}$, we propose to formulate metacognitive inconsistency detection as a multi-objective optimization problem to learn a classifier pair $p(\cdot)$ and $q(\cdot)$, which simultaneously maximizes the classifier-pair consistency, the difference of modalities used by the pair of classifiers, and the classifier accuracy:

$$\min_{\mathbf{W},\mathbf{U}} \left(\mathcal{L}(\mathbf{Y}, p(\mathbf{X};\mathbf{W})) + \mathcal{R}(\mathbf{W})\right) + \left(\mathcal{L}(\mathbf{Y}, q(\mathbf{X};\mathbf{U})) + \mathcal{R}(\mathbf{U})\right) + \gamma_1 \| \mathcal{L}(\mathbf{Y}, p(\mathbf{X};\mathbf{W})) - \mathcal{L}(\mathbf{Y}, q(\mathbf{X};\mathbf{U})) \|_F + \gamma_2 \|\mathbf{W} - \mathbf{U}\|_F^{-1} \quad (1)$$

where $\mathbf{W}$ and $\mathbf{U}$ are model coefficient matrices, $\mathcal{L}(\cdot)$ is the cost function to measure the difference between the learned model

and the ground truth in the training phase, and γ_i $(i = 1, 2)$ are tradeoff hyperparameters.

The terms in the top row of Eq. (1) learn two classifiers and fuse multisensory data using sparsity-inducing norms $\mathcal{R}(\cdot)$. The first term in the bottom row of Eq. (1) models the inconsistency of the classifier pair, and the second term is a regularization term to enforce the pair of classifiers to use different modalities. Intuitively, the proposed approach constructs a pair of classifiers that produce similar prediction results when no misinformation is inserted into the multisensory observations, based on different sets of sensing modalities. Since our approach is independent of the two classifiers in general, we consider that our approach performs reasoning that combines two classifiers at the meta-cognitive level.

After solving the formulated multi-objective optimization problem in Eq. (1) and obtaining the optimal $\mathbf{W}^*$ and $\mathbf{U}^*$, given a new multisensory observation $\mathbf{x} \in \Re^d$ during online execution, we define the inconsistency score as $h(\mathbf{x}; \mathbf{W}, \mathbf{U}) = \|p(\mathbf{x}; \mathbf{W}) - q(\mathbf{x}; \mathbf{U})\|_2$ to measure the illusion level: if more misinformation by illusion is inserted into $\mathbf{x}$, our approach will output a larger value indicating the multisensory observation is more inconsistent. This approach provides the potential to improve the resilience and robustness of robot multisensory perception, and make robots less vulnerable to sensory misinformation placed by physical illusion.

REFERENCES

[1] Battista Biggio, Giorgio Fumera, and Fabio Roli. Multiple classifier systems under attack. In *Proceedings of the 9th International Conference on Multiple Classifier Systems*, MCS'10, pages 74–83, 2010.

[2] Christophe Croux, Peter Filzmoser, and M. Rosario Oliveira. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent*, 87(2), 2007.

[3] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin P. I. Rubinstein, Udam Saini, Charles Sutton, D. J. Tygar, and Kai Xia. *Misleading Learners: Co-opting Your Spam Filter*, pages 17–51. Springer US, 2009.

[4] Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing hon Lau, Steven J. Lee, Satish Rao, Anthony Tran, J. D. Tygar, and Benjamin I. Rubinstein. Near-optimal evasion of convex-inducing classifiers. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, pages 549–556, 2010.

[5] Benjamin I. P. Rubinstein, Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing hon Lau, Satish Rao, Nina Taft, and J. D. Tygar. ANTIDOTE: Understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 1–14, 2009.

[6] David Sculley, Gabriel Wachman, and Carla E. Brodley. Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers. In *Proceedings of the Fifteenth Text REtrieval Conference, TREC*, 2006.

[7] Fei Zhang, Patrick P. K. Chan, Battista Biggio, Daniel S. Yeung, and Fabio Roli. Adversarial feature selection against evasion attacks. *IEEE Trans. Cybernetics*, 46(3):766–777, 2016.