Research Statement

Introduction

My research focuses on applying novel machine learning techniques to the problem of sensor fusion [1, 2, 3]. Sensor fusion is the task of combining information from multiple sensors in an autonomous system (e.g. autonomous cars). Broadly, there are two kinds of sensor fusion, homogeneous and heterogeneous. In homogeneous sensor fusion, information is combined from sensors of the same kind. A common example of this is combining read-outs from multiple cameras to build a 3D model of an object - this process is used for motion capture systems to capture the pose of the actors [4]. In contrast, heterogeneous sensor fusion combines information from completely different kinds of sensors. An example of this would be in unmanned aerial vehicles combining GPS and inertial sensors (e.g. accelerometers, gyroscopes) to get an accurate estimate of their location [5]. Heterogeneous sensor fusion, gives more robust information since the sensors have completely different data types, but is more challenging for the same reason. I am primarily interested in researching ways of applying data-driven techniques to best utilize the information from the sensors, in both the homogeneous and heterogeneous case.

Two-view Image Matching

My earliest work was related to homogeneous sensor fusion focused on two-view image matching, with the goal of finding the relative pose between two successive images in a video sequence [1]. In such scenarios, one can use the motion field equations to model the camera motion, and find measurements using optical flow or point tracking between the images. However, such tracking is prone to noisy measurements, and some of the measurements will be outliers and not fit the motion field equations at all. We observed that outlier point measurements do not correlate well with any estimate of motion. To exploit this observation, my co-author Andrew Jaegle and I developed the Expected Residual Likelihood (ERL) method, where we regress a range of counterfactual models and weighed points based on the likelihood distribution of the points residuals. We found ERL has 45% less translation error than the standard Epipolar RANSAC method and more sophisticated lifted kernel optimization on the self-driving car dataset KITTI [6].

Multi-view Image Matching

Matching for reconstruction can be made more robust by matching points/features from three or more images. Traditional methods simply process images pairwise using techniques from the two-view matching literature. More recent work processes all images jointly [7]. Inspired by these works, my research focuses on improving performance of joint matching using data-driven methods, especially artificial neural networks.

The first work in this line of researched focused on self-supervised learning of motion representations. Training data can be very difficult to obtain when training a neural network on matching data. Hand labeling is difficult and expensive at scale. Thus my co-author and I work on developed an unsupervised method for training deep neural representations to discriminate between different kinds of global image motion [2]. We train the network by exploiting the group properties of motion, where the representations are trained to satisfy associativity, identity, and invertibility constraints. We find this simple training procedure is enough to learn relevant motion representations. We test the algorithm on synthetic 2D sequences and real-world car driving sequences. In the car sequences, we show that the representation extracts information useful for tracking, localization, and odometry.

The next work on this topic uses cycle consistency from the pairwise matches as a loss to train neural networks [3]. While pairwise matches have no guarantees, cycle consistency ensures global consistency among the matches. Previous work [7] shows that cycle consistency equivalent to a low-rank matrix factorization objective on the pairwise image matching matrix. Instead of optimizing the matrix directly, I trained a deep network to train for this and add in additional geometric consistency losses for better performance. Similar to the previous work, the unsupervised nature of this loss means the network does not require hand labeled correspondences. Due to the structure of sparse matches, we reformulate cycle consistency as a graph problem and train using graph neural networks. Our experiments show that our method is competitive with other optimization based approaches, equivalent to 25 iterations of the optimization based methods while being similar in speed.

The last chapter of my thesis works on directly solving for outlier measurement classification in standard robotic regression problems. Standard optimization in the presence of outliers typically requires using minimal solvers. However in some applications, such as pose graph optimization, minimal solvers are not available, and thus classifying the outliers is required. Recent methods have been able to perform well in such cases, even in very high percentages of outliers, but have issues with scaling [8]. Borrowing from these techniques, we proposed a deep-learning-based method to solve such problems in the presence of outliers. This could enable faster outlier classification as deep networks only require a single forward pass to compute the results. We build upon previous theoretical results and train using a primal dual solver, with guarantees on expected satisfaction of the outlier constraints [9]. We test the performance of the method on synthetic least squares problems with outliers with favorable results compared to the more standard optimizers. This work is currently in preparation for publication.

Camera-Radar Fusion

My latest research is in heterogeneous sensor fusion, namely camera-radar fusion. Cameras and radars complement each other particularly well; both sensors share information along the ground plane, but cameras have information on the semantic and elevation information of objects, while radars have velocity and depth information. Camera-radar fusion is a fairly novel area, with not many papers in the field, with limited publications but growing interest [10, 11].

To better utilize the information coming from the sensors, I research using data-driving methods to process and combine the information. In particular, due to the their flexible nature and my past research on the subject, I am researching using Graph Neural Networks (GNNs) [12], a class of machine learning algorithms with some recent impressive results. We will use the recent state-of-the-art work of Nabati and Qi [13] as the baseline.

Art History Collaborations

With the rise of computing, Digital Humanities have risen in prominence. Machine learning and deep learning in particular can introduce new possibilities for how visual analysis can be done on images. This opens new research possibilities for many new collaborations between computer vision scientists and art historians. In collaboration with the Frick Museum of Art and Art History researcher, I advised two main projects. One was using quantifying style using variational-autoencoders (VAEs). The VAEs were trained on historical portrait data, using triplet annotations from art historians. Comparing these with standard network features or embeddings trained without the triplet loss, our method performs better on zero shot classification of artists [14]. I also advised the Master's thesis of Qinyi Zhu on few-shot learning for art historical data. Training using a combination of hierarchical learning and few-shot learning from recent work [15], we classify using nearest neighbors of the network representations with respect to the cosine similarity. Combining these two we gained 60 percentage points in in top three accuracy compared to the baseline.

Future Work

This research has only scratched the surface of sensor fusion. My current work focuses on improving 3D object detection using camera-radar fusion, and using constrained learning to improve semi-supervised estimates of multi-image matching. The next immediate step would be combining these works, as radar data can be very expensive to exhaustively label. Robust semi-supervised learning can help make camera-radar fusion more efficient and robust, leveraging the advantages of data-driven methods while not extracting as high a data cost.

In the long run, I plan to generalize the lessons from camera-radar fusion to other kinds of sensors (IR cameras, Lidars, etc.). The lack of interpretability in data-driven methods concerns many in the robotics community. I also hope to improve my thesis research in order to give safety bounds in general sensor fusion results. The ultimate goal would be to create a unified framework of data driven sensor fusion, with a systematic process to combine multiple sensors in an optimal way, while still having guarantees.

References

[1] **Phillips, S.**, Jaegle, A., Daniilidis, K., "Fast, robust, continuous monocular egomotion computation." In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 773–780.

[2] **Phillips, S.**, Jaegle, A., Ippolito, D., Daniilidis, K., "Understanding image motion with group representations." In: *International Conference on Learning Representations*. 2018. URL: `https://openreview.net/forum?id=SJLlmG-AZ`.

[3] **Phillips, S.**, Daniilidis, K., "All Graphs Lead to Rome: Learning Geometric and Cycle-Consistent Representations with Graph Convolutional Networks." In: *IEEE Conference on Computer Vision and Pattern Recognition Workshop: Image Matching: Local Features and Beyond* (2019).

[4] Moeslund, T. B., Granum, E., "A survey of computer vision-based human motion capture." In: *Computer vision and image understanding* 81.3 (2001), pp. 231–268.

[5] Chang-Sun Yoo, C.-S. Y., Iee-Ki Ahn, I.-K. A., "Low cost GPS/INS sensor fusion system for UAV navigation." In: *Digital Avionics Systems Conference, 2003. DASC '03. The 22nd.* Vol. 2. 2003, 8.A.1-8.1–9 vol.2. DOI: `10.1109/DASC.2003.1245891`.

[6] Geiger, A., Lenz, P., Urtasun, R., "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[7] Wang, Q., Zhou, X., Daniilidis, K., "Multi-image semantic matching by mining consistent features." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 685–694.

[8] Yang, H., Antonante, P., Tzoumas, V., Carlone, L., "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection." In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 1127–1134.

[9] Eisen, M., Zhang, C., Chamon, L. F., Lee, D. D., Ribeiro, A., "Online Deep Learning in Wireless Communication Systems." In: *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2018, pp. 1289–1293.

[10] Maddern, W., Pascoe, G., Linegar, C., Newman, P., "1 Year, 1000km: The Oxford RobotCar Dataset." In: *The International Journal of Robotics Research (IJRR)* 36.1 (2017), pp. 3–15. DOI: `10.1177/0278364916679498`. eprint: `http://ijr.sagepub.com/content/early/2016/11/28/0278364916679498.full.pdf+html`. URL: `http://dx.doi.org/10.1177/0278364916679498`.

[11] Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I., "The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset." In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Paris, 2020. URL: `https://arxiv.org/abs/1909.01300`.

[12] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S. Y., "A comprehensive survey on graph neural networks." In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[13] Nabati, R., Qi, H., “Centerfusion: Center-based radar and camera fusion for 3d object detection.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1527–1536.

[14] Shaik, S., Bucher, B., Agrafiotis, N., **Phillips, S.**, Daniilidis, K., Schmenner, W., *Learning Portrait Style Representations*. 2020. arXiv: `2012.04153 [cs.CV]`.

[15] Chen, Y., Wang, X., Liu, Z., Xu, H., Darrell, T., *A New Meta-Baseline for Few-Shot Learning*. 2020. arXiv: `2003.04390 [cs.CV]`.