Title: Mobile - Previous Work on Scene Understanding

Author: Henry Lao

Overview:

Scene understanding is an important research topic in the field of computer vision which has applications in a wide field, such as image and video retrieval, robotics, the medical field, and video surveillance. Scene understanding is an extremely challenging problem due to its dependence on techniques to perform semantic segmentation and the method in which extracted features and associated labels are used to model perception of the scene. In this proposal I will discuss the continuation of development of a traffic scene perception (TSP) with a focus on the classes of objects: traffic signs, cars, and cyclists. The continuation of the prior work carried out by Hu et al. involves recreation of a single framework that consolidates the classes for faster feature detection for alerts.

Table of Contents:

Introduction and Problem Definition

The overarching problem of scene understanding involves 3 classes of objects thus there are 3 subproblems to be solved. Solving the problem requires using both still images and video for extraction of semantics from a scene. The images and video will be from a user's camera. TSP involves 3 stages: object detection, object recognition and tracking of objects of interest for generating signals for alerting the user of safety conditions.

Sensor Possibilities

Past attempts in scene understanding involved a variety of sensors such as radar, lidar, ultrasonic, camera and far-infrared

Findings and Readings

Prior work focused on specific detectors for the 3 classes in contrast to basis of this proposal which focuses on a single learning framework to handle for all 3 classes. One of the techniques with traffic sign perception utilized images where signs were extracted using color and shape information. More recent approaches employ techniques for detecting texture and gradient features which are invariant to image distortion but are still unable to handle cases of extreme deformation.

Car detection methods in the past involved usage of a sliding window; however, often failed to detect cars. More recently, usage of a deformable parts model (DPM) to be adapted for car detection has been used. In addition to DPM, visual subcategorization techniques have been applied to improve generalization of the car detection model.

Cyclist detection methods are sparse, and many cyclist detection methods have spun out of pedestrian detection approaches. Some these approaches are holistic detection, part-based detection and motion-based detection.

Success and Failures

Car detection utilizing the sliding window have demonstrated promising results in human detection; however, they often fail when working with cars due to variation in viewpoints. More successful attempts have resulted from the adaptation of latent SVMs into DPM detectors.

Sign detection approaches that have failed include segmentation by thresholding images and shape-based approaches involving adaptation of a genetic algorithm due to computational limitations. However, there have been successes in texture-based approaches employing histogram of gradients (HOG) features with support vector machines (SVM) as well as convolutional neural networks (CNN).

Cyclist detection has seen success when employing lasers along with a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method on subareas. SVMs, Decision Tree (DT) classifiers, Mutliple Hypothesis Tracking (MHT) algorithm and Kalman filter have been tested and validated in real road environments showing promising results.

The Challenges

The challenge of solving scene understanding involves consolidation of discussed subproblem solutions and refinement of subproblem solutions. Some techniques employed have shown promising results; however, they remain as specific solutions that still need to be integrated into the overarching solution for TSP.

The Future and Conclusions

Devising a TSP system that can perceive and extract meaning from a traffic scene is challenging and requires refinement of smaller solutions for composing the overarching solution to the problem.

References:

[1] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2641-2649, doi: 10.1109/ICCV.2015.303. https://ieeexplore.ieee.org/abstract/document/7410660

[2] Chapel, Marie-Neige, and Thierry Bouwmans. "Moving Objects Detection with a Moving Camera: A Comprehensive Review." ArXiv.org. January 15, 2020. Accessed September 23, 2021. https://arxiv.org/abs/2001.05238.

[3] Divvala, Santosh K., Alexei A. Efros, and Martial Hebert. "How Important Are Deformable Parts in the Deformable Parts Model?" ArXiv.org. June 16, 2012. Accessed September 23, 2021. https://arxiv.org/abs/1206.3714.

[4] Girshick, Ross B., Jeff Donahue, Trevor Darrell and J. Malik. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014): 580-587. https://arxiv.org/abs/1311.2524

[5] Mohammed, A.S.; Amamou, A.; Ayevide, F.K.; Kelouwani, S.; Agbossou, K.; Zioui, N. The Perception System of Intelligent Ground Vehicles in All Weather Conditions: A Systematic Literature Review. Sensors 2020, 20, 6532. https://doi.org/10.3390/s20226532

[6] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 2005, pp. 886–893. https://ieeexplore.ieee.org/document/1467360

[7] Papineni, Kishore, S. Roukos, T. Ward and Wei-Jing Zhu. “Bleu: a Method for Automatic Evaluation of Machine Translation.” ACL (2002). https://dl.acm.org/doi/10.3115/1073083.1073135

[8] P. Dhar, M. Z. Abedin, T. Biswas and A. Datta, "Traffic sign detection — A new approach and recognition using convolution neural network," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, pp. 416-419, doi: 10.1109/R10-HTC.2017.8288988. https://ieeexplore.ieee.org/abstract/document/8288988

[9] Redmon, Joseph & Divvala, Santosh & Girshick, Ross & Farhadi, Ali. (2016). You Only Look Once: Unified, Real-Time Object Detection. 779-788. 10.1109/CVPR.2016.91. https://arxiv.org/abs/1506.02640

[10] Ren, Shaoqing, Kaiming He, Ross B. Girshick and J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2015): 1137-1149. https://arxiv.org/abs/1506.01497

[11] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 142-158, 1 Jan. 2016, doi: 10.1109/TPAMI.2015.2437384. https://ieeexplore.ieee.org/document/7112511

[12] Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

[13] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." ArXiv.org. April 10, 2015. Accessed September 23, 2021. https://arxiv.org/abs/1409.1556.

[14] Siogkas, George & Skodras, Evangelos. (2012). Traffic Lights Detection in Adverse Conditions Using Color, Symmetry and Spatiotemporal Information. VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications. 1. https://www.researchgate.net/publication/230601628_Traffic_Lights_Detection_in_Adverse_Conditions_Using_Color_Symmetry_and_Spatiotemporal_Information

[15] Staudemeyer, Ralf & Morris, Eric. (2019). Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. https://arxiv.org/abs/1909.09586

[16] Vinyals, Oriol & Toshev, Alexander & Bengio, Samy & Erhan, Dumitru. (2016). Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39. 1-1. 10.1109/TPAMI.2016.2587640. https://arxiv.org/abs/1609.06647

[17] W. Li, Z. Qu, H. Song, P. Wang and B. Xue, "The Traffic Scene Understanding and Prediction Based on Image Captioning," in IEEE Access, vol. 9, pp. 1420-1427, 2021, doi: 10.1109/ACCESS.2020.3047091. https://ieeexplore.ieee.org/document/9306804

[18] Xu, Ke, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, R. Salakhutdinov, R. Zemel and Yoshua Bengio. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” ICML (2015). https://arxiv.org/abs/1502.03044