Overview of machine learning (ML) based perception algorithms for unstructured and degraded visual environments

Narayanan, Priya, Wu, Zhenyu, Kwon, Heesung, Wang, Zhangyang , Rao, Raghuveer

Priya Narayanan, Zhenyu Wu, Heesung Kwon, Zhangyang Wang, Raghuveer Rao, "Overview of machine learning (ML) based perception algorithms for unstructured and degraded visual environments," Proc. SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, 110061T (10 May 2019); doi: 10.1117/12.2519029

Event: SPIE Defense + Commercial Sensing, 2019, Baltimore, Maryland, United States

Overview of Machine Learning (ML) based Perception Algorithms for Unstructured and Degraded Visual Environments

Priya Narayanan[a], Zhenyu Wu[b], Heesung Kwon[a], Zhangyang Wang[b], Raghuveer Rao[a]
[a]U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD, USA 20783
[b]Texas A&M University, College Station, TX, USA 77843

ABSTRACT

Machine learning based perception algorithms are increasingly being used for the development of autonomous navigation systems of self-driving vehicles. These vehicles are mainly designed to operate on structured roads or lanes and the ML algorithms are primarily used for functionalities such as object tracking, lane detection and semantic understanding. On the other hand, Autonomous/ Unmanned Ground Vehicles (UGV) being developed for military applications need to operate in unstructured, combat environment including diverse off-road terrain, inclement weather conditions, water hazards, GPS denied environment, smoke etc. Therefore, the perception algorithm requirements are different and have to be robust enough to account for several diverse terrain conditions and degradations in visual environment. In this paper, we present military-relevant requirements and challenges for scene perception that are not met by current state-of-the-art algorithms, and discuss potential strategies to address these capability gaps. We also present a survey of ML algorithms and datasets that could be employed to support maneuver of autonomous systems in complex terrains, focusing on techniques for (1) distributed scene perception using heterogeneous platforms, (2) computation in resource constrained environment (3) object detection in degraded visual imagery.

Keywords: perception, autonomous navigation, deep learning, degraded visual environment, unstructured environment

1. INTRODUCTION

Visual perception algorithms based on data driven approaches are increasingly being used for several applications including the design and development of self-driving vehicles. Sensor feeds for these algorithms mainly come from cameras, lidars, stereo sensors etc. Some of the most frequently used algorithms include object classification, object localization, lane detection and semantic scene understanding. For example, deep learning-based algorithms are typically used for detection and classification of street signs, traffic lights, and pedestrians. These detections are then used to make autonomous driving decisions. However, operationalizing many of these algorithms in real-world conditions have been challenging primarily due their reduced performance in crowded scenarios, cluttered environment with occlusion, varying lighting conditions and domain shift from training to testing conditions.

In the military domain, the technology gap is wider since autonomous systems or unmanned ground vehicles (UGVs) need to work in unstructured, combat environments. In several instances, these vehicles have to operate off-road, in rough terrain conditions, degraded visual environments, high clutter and resource constrained scenarios with limited connectivity. Several of the current autonomous technologies, including the perception algorithms have limitations in operating in such conditions preventing the vehicles from functioning reliably in these unstructured and visually degraded environments. It has to be also noted that many of these military specific autonomy challenges are not typically addressed by industrial partners or academia.

We propose using multiple heterogeneous platforms including ground and aerial vehicles distributed over a region of interest with overlapping imagery to mutually provide better situational awareness, scene understanding and holistic view of the terrain of action. The specific goal is to achieve improved object detection, tracking, recognition performance and enhanced scene understanding in unstructured environment with heavy clutter. Additionally, since these platforms frequently operate in resource constrained environments, edge computing on low SWAP (Size, Weight and Power) devices is a critical element in achieving real-time inference from these multi-domain platforms. Additionally, these platforms will operate in visually degraded conditions generated by environmental factors such as fog, mist, smoke, high altitude (aerial vehicle) and occlusions. Developing algorithms that are robust to these conditions is another important aspect that require attention.

Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications
edited by Tien Pham, Proc. of SPIE Vol. 11006, 110061T · © 2019 SPIE
CCC code: 0277-786X/19/$18 · doi: 10.1117/12.2519029

In this paper, we present a survey of algorithms and datasets that could be leveraged to support three specific areas 1) Distributed Inference using Heterogeneous Platforms 2) Computation on Resource Constrained environment 3) Degraded Visual Imagery.

2. DISTRIBUTED SCENE PERCEPTION USING HETEROGENEOUS PLATFORMS FO

Distributed inference for scene perception is one of the emerging areas of research. As mentioned in the introduction, this approach could be a potential solution for enhancing scene perception in unstructured and visually degraded environments. In this approach, multiple cameras mounted on heterogeneous platforms overlook a small area of interest and generate videos from different perspective (scale/view angle). These videos will be correlated with each other to estimate the common ground plane and objects of interest will be detected from multi-view images. Further, this approach could be leveraged to improve the performance in single/multi object trackers+, activity recognition algorithms etc. using multi-view imagery

Table 1. Summary of multi-view datasets.

Dataset	Description	Applications
PETS 2009	Data collected from 8 cameras in crowded public area with approximately forty actors	People tracking, People Counting, Event recognition
Market – 1501 Dataset	Data collected using 6 cameras infront of a supermarket. There are a total of 1501 annotated people with each person present in at least two cameras	Person Re-identification
DukeMTMC Data set	Data Collected using 8 synchronized static cameras in an outdoor setting in a university setting.	Object tracking
EPFL Pedestrian Dataset	Data collected using 3-4 Cameras in laboratory, campus, terrace, passageway, basketball game settings.	People detection and tracking
EPFL Multi-view Multi-class Detection dataset	Data collected using 6 static cameras in a university campus where there is a road with a bus stop, parking slots for cars and a pedestrian crossing.	Object Detection
ISSIA	Data collected using 6 synchronized HD cameras three for each major side of the soccer field	Sports Analysis
APIDIS	Data collected using 7 basket ball events for the entire game	Sports Analysis
IXMAS	Multiview dataset for view-invariant human action recognition. 13 daily-live motions performed each 3 times by 11 actors.	Human Action Recognition
MuHAVi	Human action video data using 8 cameras. There are 17 action classes performed by 14 actors.	Human Action Recognition

There are limited number of algorithms and datasets currently available for multi-view scene understanding. Most of the available datasets are collected from stationary multicamera network for surveillance applications. Fewer datasets are available for applications such as people/object detection and tracking, activity recognition etc. In this paper, we survey the existing datasets (Table 1) that could potentially be leveraged for multi-view scene understanding research. These include object/people detection and tracking datasets such as PETS 2009 [1], EPFL pedestrian [2], EPFL Multi-view and Multi-class Detection dataset [3]. Market-1501 [4], a dataset created specifically for person re-identification has overlapping imagery and could be treated as a mult-view dataset. Sports analysis datasets such as ISSIA [5] and APIDIS [6] and multi-view indoor dataset such as IXMAS [7] and MuHAVi [8] could be potentially used for activity and event recognition.

As is evident from Table 1, there is a severe shortage of datasets, particularly imagery simultaneously captured from moving aerial and ground platforms with a range of objects and activities. Multiview datasets captured in unstructured environment are non- existent. Hence custom data collection will be an important component for research in the field of distributed scene perception.

Just as with the case of datasets, there is severe dearth of computational methodologies and algorithms for processing and generating inferences from multi-view datasets. A survey of literature showed possibility of leveraging few algorithms from similar computer vision domains for multi-view data processing. This include few methodologies developed for comparing images taken at different viewpoints and for image correspondence. Eg. [9]. compares image patches using convolutional neural networks (CNN) using raw pixel data from the imagery. The architecture was capable of encoding 3D shapes in limited lighting conditions with very small number of multi-view images. Another architecture, SCNet [10] established semantic correspondence between different instances of same object or scene category using appearance and geometrical features learned from a CNN. Yet, another approach using proposal flow [11] established region correspondence using object proposals and their geometric relations. These approaches could be potentially used for matching local regions of imagery from multiple overlapping cameras thereby enhancing object detection, tracking, and scene perception capabilities.

There are methodologies [12][13] developed for people counting using multi-view datasets that could be directly relevant for this project. These algorithms fuse complementary information captured by multiple camera to handle heavy occlusion in crowded areas and low-quality imagery. Significant improvement in accuracy was reported when compared to results obtained from single view imagery. Few research groups have also developed detection and tracking algorithms using multi-view datasets that could be highly relevant for our study. [3] developed the Conditional Random Field (CRF) model that first estimated the ground plane locations of objects in a multiview multiclass dataset and predicted the presence of an object in the scene after taking into account contextual constraints and occlusions. These models were then tested on PETS2009 and EPFL Multi-view multi-class dataset was found to significantly outperform the CRF model. [2] used a combination of dynamic programming and generative model to successfully follow six individuals from multi-view video streams taken from different view angles. This algorithm was able to address limitations posed by occlusions and lighting changes by utilizing a multi-view dataset. Another methodology proposed by [14] provides a Bayesian approach using combined likelihood for real-time object detection and tracking of multiple human subjects using camera feeds from multiple cameras with non-overlapping field of view. The last few studies that specifically addressed object detection and tracking using multi-view imagery could be the most relevant for the proposed research in scene understanding using heterogenous platforms streaming simultaneous multi-view video feeds.

3. COMPUTATION IN RESOURCE CONSTRAINED ENVIRONMENT

Robotic platforms used for combat typically operate in constrained sensor network environment (resources, bandwidth, etc.). Conventional distributed learning methods that require transmission of raw data streams to a base station for processing is not a viable solution. Thus, local AI-ML algorithms running on edge devices within the network needs to be leveraged. In this approach, compact embedded processors on SWaP-constrained platforms generate visual analytic results. These analytic results from neighboring nodes are then combined to provide improved scene understanding of the terrain of operation. Currently, there are several parallel efforts in developing specialized low powered hardware for deep neural based algorithms. Some of the major efforts include development of Application Specific Integrated Circuits (ASIC), Neuromorphic Chips and Binarized Neural Network

3.1 Neural Network ASIC

Deep Neural Network based Application Specific Integrated Circuits uses the traditional ASIC units as an accelerator for the state-of-the-art parallel algorithms such as convolutional neural networks (CNN). This technology has made advances in leaps and bounds in the last few years which is mainly contributed by the commercial industry. Popular examples are the Tensor Processing Technology (TPU) developed by Google that achieved sixteen times performance/power improvement over GPU. However, TPU's in general are intended for server -computing system. Google has recently launched an ASIC based small edge-TPU, that can be connected to low powered devices. Similarly, AWS launched the ASIC based Inferentia ML chip and intel will soon launch its first AI ASIC chip - NNP-L1000.

3.2 Neuromorphic Chips

A neuromorphic chip is an analog circuit that mimics the biological neuron network. In recent years, there has been emphasis on designing this brain inspired chip for deep neural network-based architecture. A sub category of neuromorphic computing called Spiking Neural Network (SNN) was found be particularly appropriate for developing ultra-low powered hardware. Well known hardware using this technology include IBM TrueNorth and SpiNNaker [15].

3.3 Binarized Neural Networks

Floating point operations in neural networks are computationally very expensive. One of the approaches to address this shortcoming is to binarize the operations to make the network computationally more efficient and improve the potential for low power implementation. In this method, both the weights and activations of the network are binarized, eventually compressing the network up to 32 times. Some of the state-of the-art BNN networks include Binary Connect/ Binary Net [16], XNOR-net [17] etc.

4. DEGRADED VISUAL IMAGERY

One of the major challenges in scene understanding in unstructured environment is analyzing images captured in visually degraded environment. These degradations could be due to weather or low lighting conditions, smoke, fog, scale of image, view point and occlusions.

As with multi-view datasets, there are limited amount of publicly available degraded imagery datasets (Table 2) with the attributes such as weather conditions, view point, scale etc. In the aerial domain, there are only two known datasets currently for analysis- UAVDT [18] and Visdrone [19]. Both these datasets look at scale variation, weather conditions and viewpoint variations. While the UAVDT dataset primarily focuses on urban environment, Visdrone dataset covers a range of locations including urban, country side etc. Some of the ground based datasets with visual degradations are MOT16 [20], DETRAC [21] and KAIST [22]. These datasets are mostly collected for object detection, multi and single object tracking. Most of the deep learning-based algorithms are trained with datasets captured in ideal condition. However, the performance these algorithms plummets when tested on datasets collected in different domain such as a degraded visual environment. For example, in the case aerial imagery, an algorithm trained on imagery taken at low-altitude and daytime will not work very well on those taken at high-altitude, nighttime domain and degraded weather conditions.

One of the main approaches that has been explored to address performance drop due to domain shift is data augmentation. In this methodology the dataset is artificially expanded by adding images created by augmentation techniques using basic geometric transformations such as scaling, translation horizontal rotation etc. [23]. A more recent approach uses segmentation annotation for increasing object instances in the appropriate environment or visual context [24]. This approach could be interpreted as increasing training dataset for better performance of the algorithm or adding noise to the dataset which enables the network to learn generalized features as opposed to overfitting to the limited amount of available data. Another commonly used approach is domain adaptation or transfer learning, wherein a model trained for a task in a specific domain is used to jump start the modeling task in a different domain. Some of the most recent advances in this area include projecting source subspace to target subspace using subspace alignment for improved object detection [25]. Another approach built on the Faster R-CNN model uses image-level and instance-level adaptation to improve the performance after domain shift [26]. However, neither of these techniques can be easily generalized to unseen domains.

Table 2. Summary of datasets with attributes of degradation.

Dataset (view)	Description	Type of Degradation & Application
UAVDT (aerial)	100 video sequences with a drone platform in urban areas such as crowded squares, arterial streets, toll stations etc	S, W, V (Object Detection, Single Object Tracking, Multi-Object Tracking)
VisDrone (aerial)	263 video sequences captured by various drone-mounted cameras captured in a wide range of locations, environments with varying objects and object density. The dataset also covers various weather and lighting conditions.	S, V, W Object Detection, Single Object Tracking, Multi-Object Tracking)
MOT16(ground)	14 video sequence in crowded areas with different viewpoints, camera motions and weather conditions.	S, V, W (Multi Object Tracking)
DETRAC (ground)	100 video sequences taken at 24 different locations, representing various weather conditions, traffic types, scale and traffic types including urban highway, T-junctions etc.	S, W (Multi Object Tracking)
KAIST (ground)	95k color-thermal pedestrian dataset taken from a vehicle with regular traffic scenes at day and night time to consider changes in light conditions.	O, S, D-N (pedestrian Detection)

- S-Scale, W-Weather, V-Viewpoint, O-Occlusion, D-N – Day, Night

In order to address degradation due to object size and scale, the typical approach is to learn representations of all the objects at multiple scales. However, this approach is not superior in terms of performance gain or efficiency. [27] developed and tested a detection algorithm for low resolution imagery by coupling a supervised super-resolution algorithm with dictionary learning and E-SVM based detection algorithm. This methodology was found to produce significant gain for applications such as vehicle detections in satellite imagery. More recently, [28] developed a deep learning-based architecture that internally incorporated a feature pyramid enabling detection of objects of various different scales. This architecture was found to significantly improve the detection accuracy in COCO 2016 challenge with a reasonable framerate (6 fps) on a single GPU making it a 'good' solution for multiscale object detection.

Similarly, invariance to occlusions and deformations during object detection is primarily addressed using a data driven approach, wherein large amount of data is collected under various different conditions and scenarios including occluded conditions to train the classifier. However, this approach might not capture rare instances of occlusion and deformation. More Recently, [29] has developed adversarial network that generates rare examples of occlusions and deformations that are difficult for existing classifiers to detect. The detector and the adversary are jointly trained and there was a significant mAP boost of 2.3% and 2.6 % on VOC07 and VOC 2012 datasets respectively. Similar approaches for unsupervised domain adaptation, where the source domain has labelled data and target domain has fully unlabeled or partially labelled dataset, have become very popular. For example in [30], a deep domain adaptation algorithm is developed that combined deep feature learning algorithms and domain adaptation algorithms. This combined training process created features that are invariant to the domain shift, but still discriminative for the learning task. Similarly, in [31], imagery from aerial vehicles were classified using the features generated by a Siamese-GAN network, that are invariant to domain changes and later is feeding them into fully connected layers for classification. [23] Most recently, CyCADA [32], a Cycle-Consistent Adversarial Domain Adaptation model that adapts representations at both pixel and feature level have been applied to classification and recognition tasks and have been successfully applied for transfer from synthetic domain to real world domain. Overall, there is huge potential for applying adversarial based learning approaches for improving scene perception in degraded visual environment.

5. CONCLUSION

Development of perception systems for UGVs, particularly in the military domain, is challenging due to their sub-optimal performance in unstructured and visually degraded environment. In this paper, we propose a solution using distributed scene perception, edge computing, and algorithms for degraded visual imagery to address the problem. A comprehensive survey of literature is conducted to determine the state of the art in this area. It was concluded that there is currently severe dearth of research in this field and custom data collection and algorithm development will be necessary.

REFERENCES

[1] Ferryman, J., and A. Shahrokni. "Pets2009: Dataset and challenge." 2009*. Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*.

[2] Fleuret, F., J. Berclaz, R. Lengagne, P. Fua, Multi-Camera People Tracking with a Probabilistic Occupancy Map. 2018. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI : 10.1109/TPAMI.2007.1174.

[3] Roig Noguera,G., X. Boix Bosch, H. Ben Shitrit, P. Fua, Conditional Random Fields for Multi-Camera Object Detection. 2011. *International Conference on Computer Vision,* Barcelona, Spain, November 6-

[4] Zheng, L., L,Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. 2015.In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1116-1124)

[5] D'Orazio, T., N. Ancona, G. Cicirelli, and M. Nitti. A ball detection algorithm for real soccer image sequences.2002. In IEEE *Object recognition supported by user interaction for service robots*, vol. 1, pp. 210-213.

[6] Chen, F., C. De Vleeschouwer. Personalized production of basketball videos from multi-sensored data under limited display resolution. 2010. *Computer Vision and Image Understanding*, Vol. 114, no. 6, p. 667-680.

[7] Weinland, D. and R. Ronfard and E. Boyer. Free Viewpoint Action Recognition using Motion History Volumes. 2006. *Computer Vision and Image Understanding (CVIU), Volume 104, Number 2-3,*

[8] Singh, S., S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. *2010. 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*.

[9] Zagoruyko, S. and N. Komodakis, N. Learning to compare image patches via convolutional neural networks. 2015. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4353-4361).

[10] Han, K., R. S. Rezende, B. Ham, K. Y. K. Wong, M. Cho, C. Schmid and J. Ponce. Scnet: Learning semantic correspondence. 2017. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1831-1840).

[11] Ham, B., M. Cho, C. Schmid and J. Ponce. Proposal flow: Semantic correspondences from object proposals.2018. *IEEE transactions on pattern analysis and machine intelligence*, *40*(7), 1711-1725.

[12] Ma, H., C. Zeng and C. X. Ling. A reliable people counting system via multiple cameras. 2012. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(2), 31.

[13] Tang, N. C., Y.Y. Lin, M. F. Weng and H. Y. M. Liao. Cross-camera knowledge transfer for multiview people counting. 2015. *IEEE Transactions on image processing*, *24*(1), 80-93.

[14] Limprasert, W., A. Wallace, and G. Michaelson. Real-time people tracking in a camera network. 2013. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *3*(2), 263-271.

[15] Khan, M. M., D. R. Lester, L. A. Plana, A. Rast, X. Jin, E. Painkras and S. B. Furber, S. B. SpiNNaker: mapping neural networks onto a massively-parallel chip multiprocessor. *2008. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 2849-2856)

[16] Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., & Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. 2016. *arXiv preprint arXiv:1602.02830*.

[17] Rastegari, M., V. Ordonez, J. Redmon and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks.2016. In *European Conference on Computer Vision* (pp. 525-542). Springer, Cham.

[18] Du, D., Y. Qi., H. Yu, Y. Yang, K. Duan G. Li, G. and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. 2018. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 370-386).

[19] Zhu, P., L. Wen, D. Du, X. Bian., H. Ling, Q. Hu and X. Liu. VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results. 2018. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 0-0).

[20] Milan, A., L. Leal-Taixé, I. Reid, S. Roth and K. Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. 2016.*arXiv:1603.00831*.

[21] Wen, L., D. Du, Z. Cai, Z. Lei, M-C. Chang, H. Qi, J. Lim, M-H. Yang, and S. yu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. 2015. *arXiv preprint arXiv:1511.04136* (2015).

[22] Hwang, S., J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. 2015. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1037-1045. 2015.

[23] Data augmentation for bounding boxes: Rethinking image transforms for object detection. DataAugmentationforBoundingBoxes:Rethinkingimagetransformsforobjectdetection/, 2018.

[24] Dvornik, N., J. Mairal, and C. Schmid. On the Importance of Visual Context for Data Augmentation in Scene Understanding.2018. *arXiv preprint arXiv:1809.02492* .

[25] Raj, A., Namboodiri, V. P. and Tuytelaars, T. Subspace alignment based domain adaptation for rcnn detector. 2015. *arXiv preprint arXiv:1507.05578*.

[26] Chen, Y. Li,W., Sakaridis, C., Dai, D. , and Gool, L. V.. Domain adaptive faster r-cnn for object detection in the wild. 2018. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348.

[27] Cao, L., R. Ji, C. Wang, and J. Li. Towards domain adaptive vehicle detection in satellite image by supervised superresolution transfer. 2016. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[28] Lin, T-Y, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. 2017. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117-2125).

[29] Wang, X. , A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. 2017. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2606-2615).

[30] Ganin, Y. and V. Lempitsky. Unsupervised domain adaptation by backpropagation. 2014. *arXiv preprint arXiv:1409.7495*, 2014.

[31] Bashmal, L., Y. Bazi, H. AlHichri, M. M. AlRahhal, N. Ammour, and N. Alajlan. Siamese-gan: Learning invariant representations for aerial vehicle image categorization. 2018. *Remote Sensing*, 10(2):351.

[32] Hoffman, J., E. Tzeng, T. Park, J-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. 2017. *arXiv preprint arXiv:1711.03213*.