ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/327980521

Computer Vision for 3D Perception A review

Conference Paper · October 2018

CITATION
1

READS
2,176

6 authors, including:

Niall O' Mahony
Institute of Technology, Tralee
34 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Sean Campbell
Institute of Technology, Tralee
19 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Lenka Krpalkova
Institute of Technology, Tralee
28 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Daniel Riordan
Institute of Technology, Tralee
66 PUBLICATIONS **113** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Machine Learning View project

Project PROPAT Integrated Process Control View project

All content following this page was uploaded by Niall O' Mahony on 01 October 2018.

The user has requested enhancement of the downloaded file.

Intelligent Systems Conference 2018
6-7 September 2018 | London, UK

Computer Vision for 3D Perception

A review

Niall O' Mahony, Sean Campbell, Lenka Krpalkova,
Daniel Riordan, Joseph Walsh
IMaR Technology Gateway
Institute of Technology Tralee
Tralee, Ireland
niall.omahony@research.ittralee.ie

Aidan Murphy, Conor Ryan
Biocomputing and Developmental Systems Research Group
University of Limerick
Limerick, Ireland

***Abstract*—This paper will review the progress which has been made in Artificial Intelligence and Computer Vision particularly in 3D computer vision. There has been a lot of activity in the development of both hardware and software in 3D imaging systems which will have a huge impact in the capabilities of robotics. This paper reviews the latest advancements in the state of the art in range imaging sensors as well as some emerging technologies. For example, Time of Flight (ToF) cameras with improved resolution and latency, low cost LiDAR, and the fusion of range imaging technologies will empower robotics with greater perception capabilities. Likewise, software approaches will be reviewed with a focus on Deep Learning approaches which are now the leading edge in data analysis and further enhancing the capabilities of intelligent robotic systems using 3D imaging. The emergence of Geometric Deep Learning for 3D computer vision in robotics will also be detailed, with a focus on object registration, object detection and semantic segmentation. Foreseeable trends which have been identified in both hardware and software aspects of 3D computer vision are also discussed.**

Keywords—3D Computer Vision; Geometric Deep Learning; Range Imaging.

I. Introduction

Artificial Intelligence (AI) and robotics are said to be undergoing a “Cambrian explosion” analogous to the dramatic increase in the diversity and capabilities of life during the Cambrian period about half a billion years ago [1]. This rapid evolution is theorised to be due to the evolution of vision and likewise the explosion in AI is largely attributed to technological developments in vision processing. Intelligence involves perception and an ability to manipulate the world. Computer Vision (CV) provides a great deal of information about the surrounding environment and so it is vital in enabling intelligent systems to move from domain-specific, brute-force heuristic approaches operating under constrained conditions to being able to perform in more diverse and unprecedented situations.

This paper will give an overview of the state of the art in 3D CV technology starting with an review of the different kinds of 3D sensors that are progressively improving in accuracy while simultaneously lowering in cost. This will be followed by a brief introduction to the traditional algorithmic approaches in 3D computer vision, methods for representing 3D data, methods for generating 3D models of objects and data enhancement and data fusion approaches. The paper will then move on to review some of the recent activities in Deep Learning (DL) for CV in terms of hardware and software, with a focus on the state-of-the-art techniques for 3D perception, namely object registration, object detection and semantic segmentation of 3D point clouds. Finally developments and possible directions of getting the performance of 3D DL to the same heights as 2D DL are discussed along with an outlook on the impact the increased use of 3D will have on CV in general.

II. Computer Vision Hardware

There has been great progress in recent years in the diversity of vision sensor technology in terms of size, cost, spectral sensitivity and depth sensing capabilities. These developments have meant more frequent deployment in a wide range of applications, e.g. self-driving cars, agriculture [2], healthcare [3], [4], Geographic Information Systems (GIS) [5] and industrial automation [6], [7]. This paper will focus on developments in the spatial domain with range imaging/ depth sensing cameras which output the 3D information of a scene.

A. Range Imaging

The ability to recognise how far away objects are in a scene is paramount in robotics applications. Due to the fact that depth information about a scene is useful in effectively any imaging application, CV is currently undergoing a transition stage from two dimensional to three dimensional techniques [8]. This section will review four primary competing technologies for depth sensing: Time of Flight (ToF), Stereo, Structured light and LiDAR as well as some up and coming technologies.

1) *Time of Flight*

ToF camera systems work by directly or indirectly measuring the time taken for light to travel from the camera to objects in a scene and back for each point/pixel in the 3D image. This measurement has been achieved in a number of different ways as described in TABLE I. . The phenomenon of vision systems have been described as disruptive in enabling autonomous robotics, lowering the cost of such systems and in doing so bringing depth acquisition to the mass market [9]. Photonic Mixer Devices are a recent development with significant advantages. The method illuminates the entire scene with modulated light and the phase delay of the continuously

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie)

modulated light measured for each pixel to generate an intelligent pixel array that provides depth measurement without the need for scanning [12] (see Fig. 1. (a)). The resolution of ToF cameras is generally quite low (e.g. 176 x 132 for the IFM o3D13) but latency is low with typical allowable frame rates of 30-60 frames per second (fps). The latest improvement in resolution and latency have been enabled by advancements in LED technology and optotronic systems.

TABLE I. TOF CAMERA COMPARISON [A]

Technology	**Photonic Mixer Devices**	**Range Gated Imagers**	**Direct ToF Imagers**
Measurement method	Phase shift of carrier in a radio frequency-modulated beam of light	The proportion of a light pulse blocked off by shutter at the receiver	Measure direct ToF required for pulse to reflect to focal plane array
Illumination	LED	Laser	Laser
Advantages	Long range (5-60m), Inexpensive,	Suitable for both long and sub-millimetre ranges, good outdoor performance (e.g. in rain/fog)	Rapid Acquisition, Complete spatial & temporal data
Disadvantages	Noisy Depth Images	Expensive	Expensive
Manufacturers	Swiss Ranger, PMD, IFM, Microsoft Kinect 1	Microsoft Kinect 2, Fraunhofer Institute	Advanced Scientific Concepts Inc

a. [7], [10], [11]

b.

2) *Stereo-vision*

Stereo systems mimic depth perception found in nature in predatory animals with front-facing eyes. Upon comparison of images from two horizontally displaced cameras, the distance to points can be computed based on their disparity (the difference in x co-ordinates) where disparity is higher for points closer to the cameras (see Fig. 1. (b)). Camera calibration is required since any lens distortion will adversely affect the depth measurements [13]. The main challenge of stereo vision is point matching in the pair of stereo images for acquiring robust depth measurements [12].

3) *Structured light*

A laser light source is projected with a known pattern and the distortion of the reflected pattern detected at the receiver is used to calculate depth based on geometry (see Fig. 1. (c)). The light pattern can be fixed or programmable to achieve better accuracy or respond to ambient light conditions or the object's optical reflection characteristics. Characteristics include superior accuracy but only at low ranges and in dark environments. The system requires several patterns to be recorded which may take a few seconds, hence it is not suitable for dynamic scenes. In [14], a method of light field imaging under structured illumination to deal with high dynamic range 3D imaging is proposed. Fringe patterns (modulated by the scene depth) are projected onto a scene and the received structured light field contains information about ray direction and phase-encoded depth which allows multidirectional depth better dynamic range and better performance on highly and lowly reflective surfaces.

4) *LiDAR*

Light Detection and Ranging (LiDAR) work by measuring the time of flight of a pulsed laser. Several laser sources are used orientated at equal spacing around 360°. The technology's main use has been in self-driving cars as it has the lowest latency and longest range compared to any other 3D imaging approach which is important in a fast-moving vehicle. Since time of flight measurements require very fast signal processing, alternative point-scanning methods also exist such as sheet-of-light/laser line triangulation measure the horizontal/vertical displacement of the imaged laser line which has been projected off-axis from the camera. These methods are very fast for point measurements, however complete imaging requires the entire unit to be rotated or translated while stitching together each single 3D profile generated one after the other [15]. Sensor manufacturers are now beginning to offer solid state lidar devices with and without moving parts and are several orders of magnitude lower in price compared to state of the art high end devices ($80,000 vs $100) with manufacturers such as Valeo, Velodyne, Quanergy and Innoviz set to have low-cost LiDAR offerings in the near future [16]. Such a price reduction is sure to have a significant impact on the market allowing the use of LiDAR in a wider range of applications which previously considered the technology to be prohibitively expensive. The SSD devices are anticipated to long and short range variants with fields of view from 170 by 60° to 50 by 20° and ranges up to 200m [16].

B. *Evaluation*

There is a diverse range of vision sensors for 3D vision in robotics applications as noted by [15] who evaluate a range of the different types and sizes of 3D scanners (in the categories listed above) on a specially designed plate to simulate varying material reflectance and structure. Other test variables investigated in comparisons of 3D cameras include; distance of the sensor to the object, environmental illumination and the surface of the object [17]. Generally, the noise in depth measurements increases with the distance for ToF Sensors and also for structured light sensors up to 3.5 m, beyond which noise increases quadratically [17]. Noise in laser line triangulation methods is also linear to mid-ranges but also increases drastically for longer ranges [18]. Some of the comparisons which have been made are summarised in TABLE II. .

In related work, the performance of an IFM o3D313 ToF camera and an Intel Realsense d435 stereo vision-camera were each evaluated in a range of scenarios. As can be seen in the scene depicted in Fig. 2., at a range of 3-4 m in an indoor environment with few features, ToF has less noise compared to stereo vision allowing more subtle features such as piping and a doorway against the wall to be distinguished. While stereo-vision has better resolution and accuracy at distances below 1m, global shutter and a wider field of view (90° compared to

60°) the bandwidth of the sensor poses a data processing challenge in embedded applications and requires down-sampling.

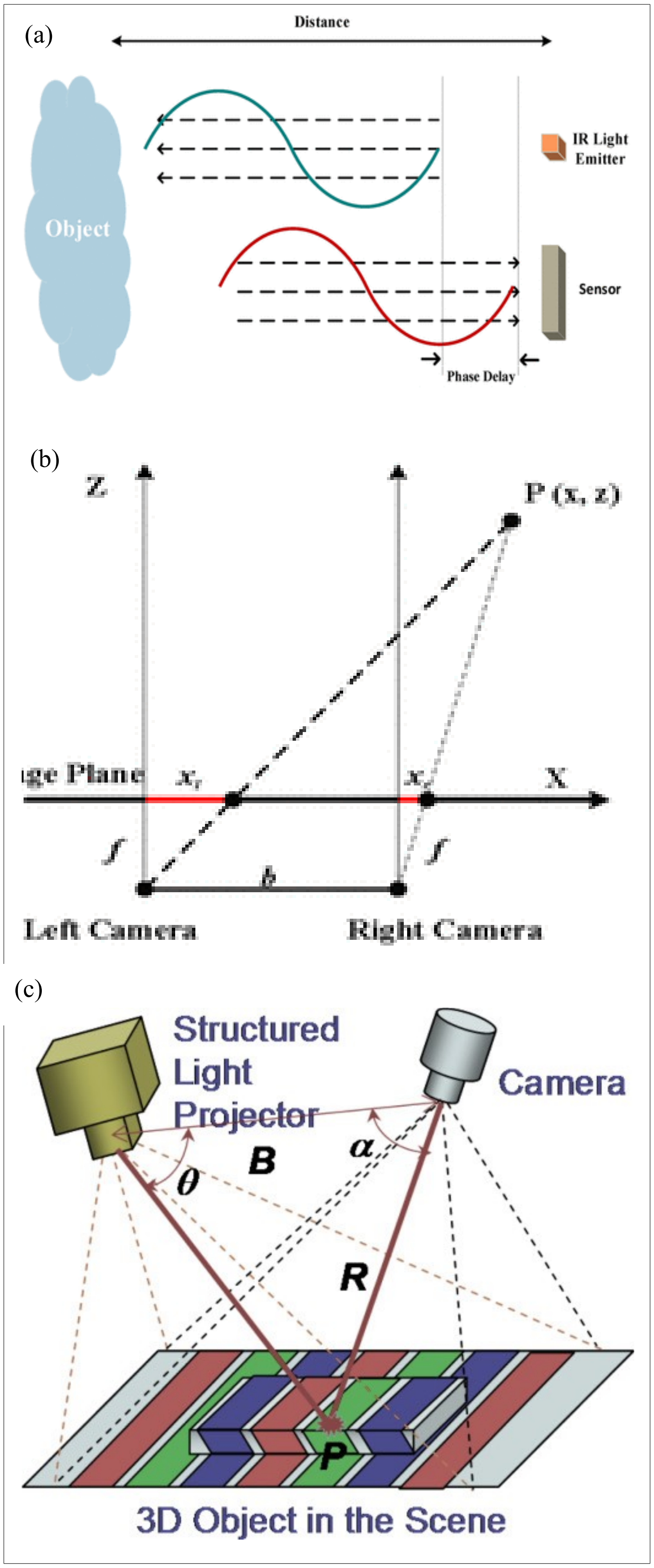

Fig. 1. Principle of operation of (a) ToF [19], (b) stereo [20] and (c) structured light [21] 3D vision systems.

TABLE II. COMPARISON OF RANGE IMAGING CAMERAS

	Time of flight/ LiDAR	Stereo vision	Structured Light
Software processing	Low	High	Medium
Latency	Low	Medium	High
Active illumination	Yes	No	Yes
Low light performance	Good	Weak	Good
Bright light performance	Medium	Good	Weak*
Power consumption	Medium/high*	Low	Medium*
Range	Short to long range (5-30m)[b] LiDAR can have longer ranges.	Mid-range (~10m) (dependant on camera spacing)	Very short to mid-range (3.5m)[b]
Resolution	Low	High (Camera Dependent)	Projected pattern dependent
Depth accuracy	High	Low	Medium

[c.] Depends on illumination power & modulation [22]
[d.]

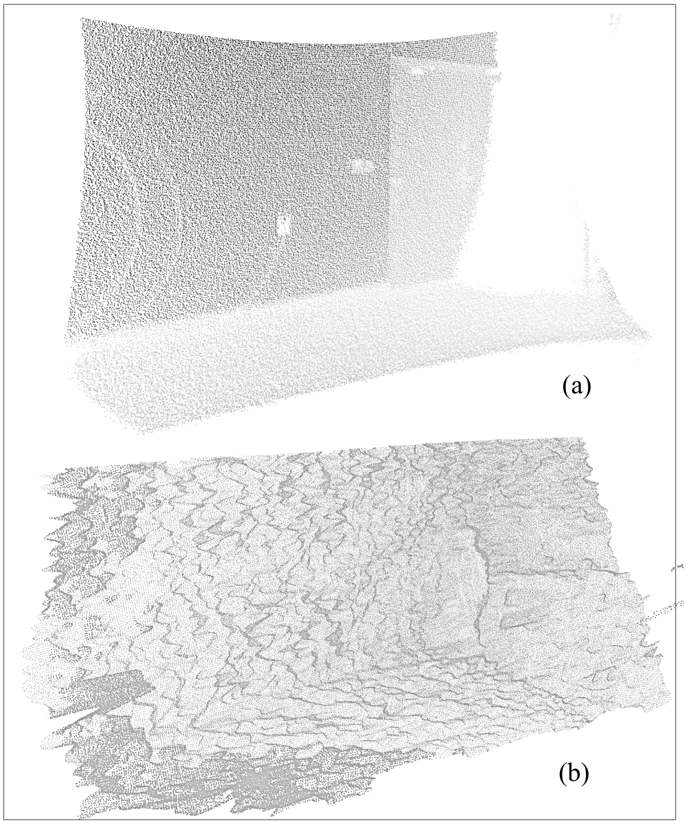

Fig. 2. Comparison of test recordings evaluating performance of (a) an IFM o3D313 ToF camera and (b) an Intel Realsense d435 stereo-vision camera that uses structured light to provide texture.

C. Trends

There are a number of opportunities for improvement for 3D vision system developers including increasing frame rate of ToF cameras, reducing cost and power consumption of long-range ToF sensors with pulsed laser or by improving LED performance, combining ToF and structural light to improve brightness tolerance and combining stereoscopic cameras with ToF to have both high depth accuracy and high resolution. One example of the combination of imaging technologies is the development of 3D hyperspectral imaging techniques by [23].

There is also active research in new passive 3D imaging approaches. For example, Active Wavefront Sampling [8] where the optical wavefront traversing a lens is sampled at two or more off-axis locations and the resulting motion of each target feature is measured and used to calculate distance. The system only requires one optical train and one sensor and hence can be smaller and lower cost compared to stereo techniques. Event-sensitive vision sensors, also called event cameras or dynamic vision systems are another emerging technology that offers low power consumption and high frame rate [20]. The event camera is a silicon retina which outputs not a sequence of video frames like a standard camera, but a stream of asynchronous spikes, each with pixel location, sign and precise timing, indicating when individual pixels record a threshold log intensity change and offers the potential to be used in dense 3D reconstruction and many other vision problems .

III. Computer Vision Processing

1) *Hand-crafted Approaches*

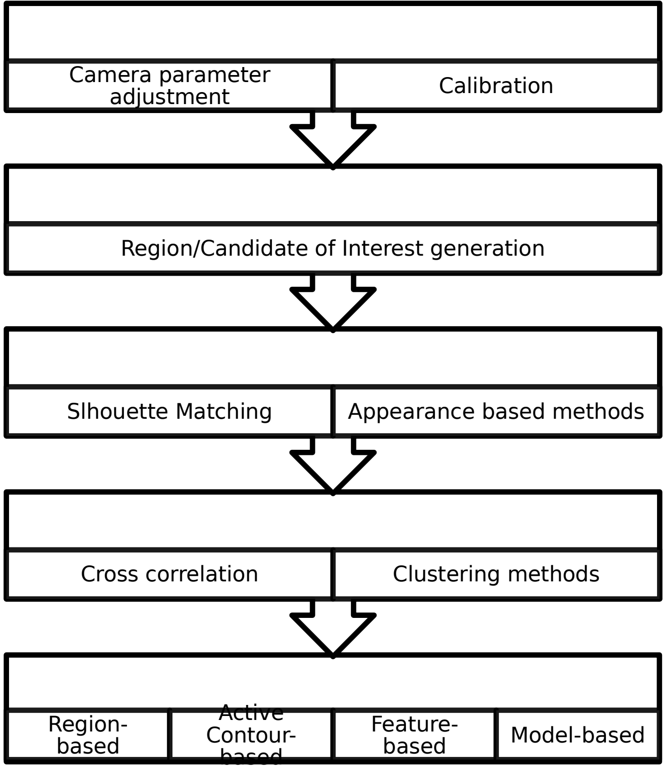

Fig. 3. Overview of traditional CV procedures for 3D vision [3]

TABLE I. gives an overview of the different categories of tasks performed in most CV applications before the advent of machine learning. The procedures involve fine tuning of parameters relating to the camera, the environment and objects of interest. Camera calibration procedures [24] and open problems such as dynamic range adjustment are issues currently being addressed in this area.

B. 3D processing

3D data representation can be split into two categories: surface representations and volume representations. The former consists of

a) depth maps, where each pixel value corresponds to the distance that point is from the camera,
b) 'surfels' (surface-pixel), which describes a local sample of the surface with its coordinates, texture, etc. [25].
c) a meshing, a group of points defined by their 3D coordinates.

[26]

Volume representations contain

a) voxels (volumetric-pixels), which contain brightness/colorimetric information of a point, but no information its 3D coordinates. Similarly, to pixels, voxel coordinates are inferred based on their position relative to other surrounding voxels.
b) the spherical harmonics allowing representation of the 3D model in spherical coordinates (which can be later converted to cartesian co-ordinates for use in robotics applications) [27].

.

This section will deal with developments in processing techniques in 3D imaging systems for producing and improving representations of 3D data.

1) *3D model construction*

3D surface registration is the process of transforming multiple 3D data sets into the same coordinate system so as get a better representation of 3D objects. For a comprehensive survey of both rigid and nonrigid (a.k.a. deformable models) registration and insights into the relationship between 3D registration and data fitting (model selection, correspondences and constraints, and optimization) refer to [28].

A novel data fitting approach for reconstructing large-scale outdoor scenes using monocular motion stereo at interactive frame rates on a modern mobile device has been demonstrated by [12]. They detect and discard unreliable depth measurements and integrate the remaining depth map into a volumetric representation of the scene using a truncated signed distance function.

Moving from mobile to localised robotic applications, robot-assisted 3D point cloud object registration has been worked on by [29] where the robot gripper and object of interest is rotated in front of a stereovision camera and its geometry is captured from different angles. Known elements (environment, robot arm and gripper) are then removed so that the object can be identified.

By learning the structure of real world 3D objects and scenes, gaps and occluded regions in 3D representations can be reconstructed as demonstrated by [30] who present a novel 3D-CNN architecture (discussed later in this paper) that learns to predict an implicit surface representation from input depth maps. Their method outperforms traditional volumetric fusion approaches in terms of noise reduction and outlier suppression.

2) *Data fusion and enhancement*

The fusion of data inputs from different sensor modalities allows the advantageous traits of each technology to be combined. Fusion of ToF and Stereo Data has been demonstrated based on:

- a probabilistic model which accounts also for depth discontinuity artefacts due to the mixed pixel effect (when a pixel represents the average of several spatial classes within its projected area, i.e. along 'jump' edges where there is a sharp transition in depth) [31].
- an extended superpixel segmentation algorithm to recover incomplete depth data from stereo cameras and low resolution ToF Cameras [32].

Data from simple sensors can also be fused with camera data, for example, [33] use monocular cameras and wheel odometry to fuse obstacle detections over time and between cameras to estimate the free and occupied space around a vehicle. The use wheel odometry to align detected obstacles in multiple depth maps eliminates the need for accurate visual inertial odometry estimation. GPS data has also been fused with ToF camera data for use in urban areas where kinematic positioning needs to rely more on vision systems [34].

Deep Neural Networks (DNNs) which will be discussed later in this paper have been used to enhance 3D data in a such a wide array of applications that it would be impossible to do a comprehensive review in a single paper. For example, DNNs have been used for depth estimation in monocular vision [35], learning shape-from-shading from monocular vision [36], stereo matching [37] and regressing disparity from a rectified pair of stereo images [38] and depth estimation fusing LiDAR and stereo-vision in self-driving cars [39].

3) *Other types of data representation*

4D data can arise from the analysis of 3D video sequences, [40] perform spatiotemporal motion analysis in a 4D representation and real-time pattern analysis based on dynamic stereo vision. The combination of spatial and spectral methods also requires for unique representation of the multi-dimensional data [41].

IV. Deep Learning

A. *An AI explosion*

There has been a big jump in ability to recognise objects in recent years. The development of Convolutional Neural Networks has had a tremendous influence in the field of machine learning and CV. This burst in progress has been enabled by an increase in computing power and in the amount of data being fed to DL models. There has been a surge in the activity in the AI research community. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [42] has fostered innovation in the field and serves as an indicator of progress in the field. Seminal papers in the field include [43]–[46].There is a multitude of tools freely available to the community for AI research including frameworks, libraries, toolkits and interfaces that have also enabled community driven progress [47]–[49].

DL methods have the ability to automatically learn complex mapping functions directly from data, eliminating the need for features to be manually defined [7] which is sometimes very difficult for high level features where poor definition of a model or changing environments can lead to a lack of robustness.

B. *Computing Hardware for Deep Learning*

DL introduces some significant challenges which are driving recent developments in computer architectures. These challenges include the memory requirements of Deep Neural Networks, the need for speed to reduce training times and reducing energy consumption for deploying DL on mobile devices [26]. There exist several different types of computer architecture including CPUs (Central Processing Unit) and GPUs (Graphics Processing Unit) which are general purpose hardware, FPGAs (Field Programable Gate Array) and ASICs (Application-Specific Integrated Circuit) which are customisable to suit DL applications and TPUs (Tensor Processing Unit) and VPUs (Vision Processing Unit) which are specially designed for AI acceleration [50], [51].

The classical CPU architecture is not very effective for training Deep Neural Networks in comparison to GPUs, parallel computing architectures which typically achieve several orders of magnitude better speed performance [19]. This is because neural networks are composed of many independent computing units or 'neurons' and lend themselves to parallel processing [19], [52]. While GPU technology continues to experience very rapid growth, TPUs and domain-specific architectures tailored for DL acceleration are considered to be the way forward for DL inference and training speedup [50]. In a test to compare the performance of TPUs, GPU's and CPU's on typical DL workloads for datacentres, Google's TPU was shown to be 15–30 times faster than its contemporary GPU or CPU, with TOPS/Watt about 30–80 times higher. Alongside these hardware developments, the operation of computing procedures can also be optimised for DL with compression, acceleration and regularisation techniques [26].

V. Deep Learning on 3D Data

2D Bounding Box (BB) localization is predominant in current CNN object detection models. This is logical as the vast majority of the images available to download for training are 2D and is also encouraged by existing benchmark data sets, such as Pascal VOC, which are also 2D. Although the developed Euclidean approaches are extremely accurate, they are limited when it comes to expressing 3D shapes or viewpoints so that we can reason about them further. 3D object representations give a greater depth of information which can aid in finer categorisation of the proposed candidates and more accurate location of points of interest for robotic applications. For example, [53] estimate the 3D geometry of objects (cars) by matching between 3D CAD models and 2D images using a DL classifier and demonstrate that more fine-grained categorization of rigid classes is achievable using the 3D object representations.

Higher-level applications which involve 3D scene understanding or 3D object tracking include video data analysis, medical imaging, robot guidance, autonomous vehicles and general categorisation tasks. These applications not only benefit from greater precision enabled by 3D geometric information but also benefit from spatial invariance built into 3D models. This invariance is due to the elimination of the translation between 2D and 3D representations which leads to distortion of the 3D data and is very beneficial in image classification tasks as several times less training samples are required [54].

Geometric Deep Learning (GDL) deals with the extension of DL techniques to 3D data. 3D data can be represented as graphs, manifolds, meshes ad point clouds depending on the application. The recent advances in GDL have been reviewed in [54]. As CNNs are the most effective DL architecture, many have adapted CNNs to take 3D data as their input and branded them 3D-CNNs. The existence of benchmark datasets such as KITTI, 3D object classes , Pascal3D+ and the RGB-D Object Dataset are useful resources for evaluating 3D-CNN models as the following works reviewed in this paper have done.

Early work in the field focused on applying GDL to pre-segmented objects in video data analysis [55], [56], feature detection from LiDAR point clouds, categorization of segmented point clouds, learning the representation of 3D shapes and integrating a volumetric occupancy grid representations [13]. More recent work has involved the use of DL to automatically segment raw point cloud data [57] and solutions for managing the large amounts of data involved.

A. Object Registration

Many recent works have looked at leveraging depth information to obtain 3D object proposals. Singh Pahwa et al. leverage the depth and multi-view scene information per frame in a RGB-D video sequence. They use efficient but robust registration to combine multiple frames of a scene in near real time and generate 3D BBs for potential 3D regions of interest which can be integrated into SLAM-based video processing for quick 3D object localization [58].

Pepik et al. designed a detector particularly tailored towards 3D geometric reasoning that generates deformable part models which include both estimates of viewpoint and 3D parts that are consistent across viewpoints [59]. The team go onto extend the successful deformable part model to include viewpoint information and part-level 3D geometry information and in turn provide richer object hypotheses than the 2D object models [60]. The generated 3D object models consist of multiple object parts represented in 3D and a continuous appearance model which provide consistently better joint object localization and viewpoint estimation than the state-of-the-art multi-view and 3D object detectors.

Discovering never-seen-before objects in 3D point clouds is an important problem to be overcome when deploying robotics in real-world scenarios. 3D-CNNs have been used by [61] to do bottom up aggregation of supervoxels obtained from 3D scenes into objects.

B. Object Detection

Recent research has been carried out into utilising 2D-CNNs and processing the inputs and output to produce 3D BBs [62], using 3D-CNNs on monocular camera data to produce 3D BBs [63] and using 3D-CNNs taking LiDAR Point Cloud Data as input and predicting 2D BBs of vehicles [53].Even more recent work has exploited stereo imagery to generate high-quality 3D object proposals in the form of 3D BBs [64]. They formulate the problem as minimizing an energy function that encodes object size priors, ground plane as well as several depth informed features that reason about free space, point cloud densities and distance to the ground. Combined with convolutional neural net (CNN) scoring, their approach outperforms all existing results on all three KITTI object classes.

C. Semantic Segmentation

Semantic segmentation is one of the key problems in the field of computer vision and is considered as one of the high-level task that paves the way towards complete scene understanding. A review of the state of the art in semantic segmentation techniques and benchmark datasets for still 2D images, video, and 3D or volumetric data is provided by [65]. One of the challenges in executing semantic segmentation of 3D point clouds is the task of accurately annotating 3D points as belonging to a specific class, especially as the data from 3D cameras can be quite noisy. One approach is to annotate the depth image in 2D using common open-source tools such as LabelImg and LabelMe [66]. This method introduces ambiguity however as it is difficult to define the boundary between pixels belonging to an object from the background accurately. This task is aided by the emergence of tools such as ScanNet [67] which allow segments to be painted over in 3D and then annotated.

Many 3D segmentation techniques use voxelization for transforming the 2D representation of 3D point clouds (e.g. x, y, z, RGB colour) into a grid of voxels of predefined size (e.g. 5 cm) resulting in a 4-channel 'image' consisting of voxel occupancy and RGB colour for example [68], [69]. This allows powerful 2D-CNN techniques to be exploited, however the extra space required for 3D data introduces a memory challenge which can require the voxel grid to be down-sampled even further (e.g. 5cm grid down to a 20 cm grid) reducing the resolution of segmentation.

One possible and common approach to refine the output of a segmentation system and boost its ability to capture fine-grained details is to apply a post-processing stage using a Conditional Random Field (CRF) which overcome the inherent invariance to spatial transformations of CNN architectures [65], [68]. SEGCloud achieved state-of-the-art performance using an end-to-end framework that allows point-level segmentation using voxelization, 3D Fully Convolutional Neural Networks (3DFCNN) and trilinear interpolation to label points according to weighted labels of surrounding voxels. The 3DFCNN was then combined with CRF which refines and enforces consistency of the labels for each point output by the final fully-connected layer.

An alternative approach for point-level semantic segmentation is PointNet and PointNet++ which avoid using voxelization by directly feeding point clouds as input to a unified architecture that outputs either class labels for the entire input or per point segment/part labels for each point of the input [70], [68]. Another approach is to avoid to representing points in Euclidean space and represent 3D data as a graph (e.g., a polygon mesh or point-based connectivity graph), convert the graph into its spectral representation, then perform convolution in the spectral domain. The main argument behind this family of GDL approaches is that they respect the permutation invariance of points in the input however they do introduce new challenges which are discussed in a comprehensive review by [54].

D. Optimising DL for 3D Data

All existing deep learning software frameworks are primarily optimized for Euclidean data where the assumption regularly structured data on 1D or 2D grid can be exploited on modern GPU hardware. Generally 3D-CNNs ensue some performance loss in comparison to their 2D counterparts, however this was experimentally verified to be minimal by [59]. Efficiency is also not as high, although it has been demonstrated that real-time operation is achievable even when deployed on embedded systems [58]. Since 3D data has a different structure there is an opportunity for computing efficiency to be optimised in this domain and possibly for the development of more suitable hardware architectures [54].

A novel framework for applying convolutional neural networks to point clouds has been developed by [71] that maps point cloud functions to volumetric functions and vice versa. In generalizing image CNNs and allowing their architectures to be readily adapted to the point cloud setting, the method is robust, efficient and able to outperform all other point cloud methods on classification, segmentation and normal estimation benchmark datasets. A network architecture has also been developed by [72] for applying convolutions to 3D data. The network uses sparse bilateral convolutional layers that maintain efficiency by using indexing structures to apply convolutions only on occupied parts of a higher dimensional lattice. The advantage of this approach over methods that feed the point cloud is that it allows flexible specification of the lattice structure enabling hierarchical and spatially-aware feature learning, as well as joint 2D-3D reasoning.

VI. Conclusion

There is significant change on the way in computer vision as 3D vision systems radically improve in terms of performance and cost. The advent of low cost LiDAR will enable a plethora of autonomous vehicle applications in self-driving cars, agriculture and factory automation. This paper has focused on improvements in 3D imaging for robotics perception which will enable greater understanding and control in applications such as autonomous vehicles, human-robot interaction and automation.

The increased deployment of range imaging sensors will facilitate the expansion of 3D datasets and enable better performance from 3D-CNNs. This paper has reviewed some recent research in DL for 3D data, an area where there are a number of unique approaches each with their own trade-offs on accuracy and efficiency. It is apparent that the number of research publications in the field of 3D computer vision is increasing and there is intense competition as the bragging rights for state-of-the-art results is constantly changing hands. Open challenges in 3D CV include deformable 3D shape correspondence, generalization of deep learning models and dealing with dynamically changing shapes in 3D video. As 3D imaging becomes more and more prolific, software algorithms and computing hardware will continue to evolve to be more suited to solving problems in 3D vision. Such advancements will enable robots to reason about 3D space more effectively.

Acknowledgment

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie)

References

[1] G. A. Pratt, "Is a Cambrian Explosion Coming for Robotics?," *J. Econ. Perspect.*, vol. 29, no. 3, pp. 51–60, Aug. 2015.

[2] M. Vázquez-Arellano, H. W. Griepentrog, D. Reiser, and D. S. Paraforos, "3-D Imaging Systems for Agricultural Applications-A Review.," *Sensors (Basel).*, vol. 16, no. 5, 2016.

[3] M. Kabir, A. Mamun, and T. Szecsi, "Development of situation recognition, environmental monitoring and patient condition monitoring service modules for hospital robots," 2012.

[4] R. Bostelman, P. Russo, J. Albus, T. Hong, and R. Madhavan, "Applications of a 3D Range Camera Towards Healthcare Mobility Aids," in *International Conference on Networking, Sensing and Control*, 2006.

[5] Information Resources Management Association., *Geographic information systems : concepts, methodologies, tools, and applications*. Information Science Reference, 2013.

[6] H. Surmann, A. Nüchter, and J. Hertzberg, "An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments," *Rob. Auton. Syst.*, vol. 45, no. 3–4, pp. 181–198, Dec. 2003.

[7] J. Molleda, R. Usamentiaga, D. F. García, F. G. Bulnes, A. Espina, and B. Dieye, "An improved 3D imaging system for dimensional quality inspection of rolled products in the metal industry," *Comput. Ind.*, vol. 64, no. 9, pp. 1186–1200, Dec. 2013.

[8] F. P. D. M. I. of T. Frigerio, "3-dimensional surface imaging using Active Wavefront Sampling," 2006.

[9] D. Costa, J. C. Cavalcanti, and D. Costa, "A Cambrian Explosion in Robotic Life," *SSRN Electron. J.*, Jan. 2011.

[10] D. Piatti and F. Rinaudo, "SR-4000 and CamCube3.0 Time of Flight (ToF) Cameras: Tests and Comparison," *Remote Sens.*, vol. 4, no. 12, pp. 1069–1089, Apr. 2012.

[11] M. Perenzoni and D. Stoppa, "Figures of Merit for Indirect Time-of-Flight 3D Cameras: Definition and Experimental Evaluation," *Remote Sens.*, vol. 3, no. 12, pp. 2461–2472, Nov. 2011.

[12] T. Schöps, T. Sattler, C. Häne, and M. Pollefeys, "Large-scale outdoor 3D reconstruction on a mobile device," *Comput. Vis. Image Underst.*, 2017.

[13] A. Wilson, "3D imaging systems target multiple applications," *Vis. Syst. Des.*, vol. 18, no. 9, 2013.

[14] Z. Cai *et al.*, "Structured light field 3D imaging," *Opt. Express*, vol. 24, no. 18, p. 20324, Sep. 2016.

[15] B. Møller, I. Balslev, and N. Krüger, "An automatic evaluation procedure for 3-D scanners in robotics applications," *IEEE Sens. J.*, 2013.

[16] J. Hecht, "Lidar for Self-Driving Cars," *Optics &Photonics News*, Jan-2018.

[17] G. Rauscher, D. Dube, and A. Zell, "A Comparison of 3D Sensors for Wheeled Mobile Robots," in *13th International Conference on Intelligent Autonomous Systems*, 2016, pp. 29–41.

[18] O. Schreer, P. Kauff, and T. Sikora, *3D videocommunication : algorithms, concepts, and real-time systems in human centred communication*. Wiley, 2005.

[19] J. Lawrence, J. Malmsten, A. Rybka, D. A. Sabol, and K. Triplin, "Comparing TensorFlow Deep Learning Performance Using CPUs, GPUs, Local PCs and Cloud," *Student-Faculty Res. Day, CSIS, Pace Univ. Pleasantville, New York*, 2017.

[20] Y. He and S. Chen, "Advances in sensing and processing methods for three-dimensional robot vision," *Int. J. Adv. Robot. Syst.*, vol. 15, no. 2, p. 172988141876062, Mar. 2018.

[21] J. Geng, “Structured-light 3D surface imaging: a tutorial,” *Adv. Opt. Photonics*, vol. 3, no. 2, p. 128, Jun. 2011.

[22] L. Li, “Time-of-Flight Camera – An Introduction,” *Texas Instruments Tech. White Pap.*, vol. SLOA190B, 2014.

[23] J. Ahlberg, I. G. Renhorn, T. R. Chevalier, J. Rydell, and D. Bergström, “Three-dimensional hyperspectral imaging technique,” 2017, vol. 10198, p. 1019805.

[24] O. Semeniuta, “Analysis of Camera Calibration with Respect to Measurement Accuracy,” in *Procedia CIRP*, 2016.

[25] J. Stückler and S. Behnke, “Multi-Resolution Surfel Maps for Efficient Dense 3D Modeling and Tracking,” *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 137–147, 2014.

[26] S. Han, “Efficient Methods and Hardware for Deep Learning,” 2017.

[27] E. U. I. Bildbehandling, E. U. I. Bildbehandling, M. Westberg, and M. Westberg, “Time of Flight Based Teat Detection.”

[28] G. K. L. Tam *et al.*, “Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 7, pp. 1199–1217, Jul. 2013.

[29] B. Jerbić, F. Šuligoj, M. Švaco, and B. Šekoranja, “Robot assisted 3D point cloud object registration,” *Procedia Eng.*, vol. 100, pp. 847–852, 2015.

[30] G. Riegler, A. Osman Ulusoy, H. Bischof, and A. Geiger, “OctNetFusion: Learning Depth Fusion from Data,” *arXiv Prepr. arXiv1704.01047v3*, 2017.

[31] C. D. Mutto, P. Zanuttigh, and G. M. Cortelazzo, “Probabilistic ToF and Stereo Data Fusion Based on Mixed Pixels Measurement Models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2260–2272, Nov. 2015.

[32] M. Van Den Bergh, D. Carton, and L. Van Gool, “Depth SEEDS: Recovering incomplete depth data using superpixels,” in *Proceedings of IEEE Workshop on Applications of Computer Vision*, 2013.

[33] C. Häne, T. Sattler, and M. Pollefeys, “Obstacle detection for self-driving cars using only monocular cameras and wheel odometry,” in *IEEE International Conference on Intelligent Robots and Systems*, 2015.

[34] S. Foix, G. Alenya, and C. Torras, “Lock-in Time-of-Flight (ToF) Cameras: A Survey,” *IEEE Sens. J.*, vol. 11, no. 9, pp. 1917–1926, Sep. 2011.

[35] C. G. Oisin, M. Aodha, and G. J. Brostow, “Unsupervised Monocular Depth Estimation with Left-Right Consistency,” *arXiv Prepr. arXiv1609.03677v3*, 2017.

[36] Jan Bednarık, Pascal Fua, and Mathieu Salzmann, “Learning Shape-from-Shading for Deformable Surfaces,” *arXiv Prepr. arXiv1803.08908v1*.

[37] Yiran Zhong, Yuchao Dai, and Hongdong Li, “Self-Supervised Learning for Stereo Matching with Self-Improving Ability,” *arXiv Prepr. arXiv1709.00930v1*, 2017.

[38] Alex Kendall *et al.*, “End-to-End Learning of Geometry and Context for Deep Stereo Regression,” *arXiv Prepr. arXiv1703.04309v1*, 2017.

[39] S. B. Nikolai Smolyanskiy, Alexey Kamenev, “On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach,” *arXiv Prepr. arXiv1803.09719v1*, 2018.

[40] B. Kohn, A. N. Belbachir, and A. Nowakowska, “Real-time gesture recognition using bio inspired 3D vision sensor,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012.

[41] H. Aasen, A. Burkart, A. Bolten, and G. Bareth, “Generating 3D hyperspectral information with lightweight UAV snapshot cameras for vegetation monitoring: From camera calibration to quality assurance,” *ISPRS J. Photogramm. Remote Sens.*, vol. 108, pp. 245–259, Oct. 2015.

[42] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., pp. 1097–1105, 2012.

[44] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *European Conference on Computer Vision*, 2014, pp. 818–833.

[45] C. Szegedy *et al.*, “Going Deeper with Convolutions,” in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[46] K. Simonyan and A. Zisserman, “Very Deep Convolutional Neural Networks for Large-Scale Image Recognition,” 2015.

[47] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, “Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers,” Jan. 2018.

[48] S. Bahrampour, N. Ramakrishnan, L. Schott, and M. Shah, “Comparative Study of Deep Learning Software Frameworks,” Nov. 2015.

[49] J. Zacharias, M. Barz, and D. Sonntag, “A Survey on Deep Learning Toolkits and Libraries for Intelligent User Interfaces,” *J. Digit. Imaging*, vol. 30, no. 4, pp. 400–405, 2017.

[50] N. P. Jouppi *et al.*, “In-Datacenter Performance Analysis of a Tensor Processing Unit TM,” in *44th International Symposium on Computer Architecture (ISCA)*, 2017, pp. 1–12.

[51] A. Cano, “A survey on graphic processing unit computing for large-scale data mining,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 1, p. e1232, Jan. 2018.

[52] H. Kim, H. Nam, W. Jung, and J. Lee, “Performance analysis of CNN frameworks for GPUs,” in *2017 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2017, pp. 55–64.

[53] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, “DepthCN: Vehicle Detection Using 3D-LIDAR and ConvNet,” 2017.

[54] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: going beyond Euclidean data,” *IEEE Sig Proc Mag*, 2017.

[55] G. Cheron, I. Laptev, and C. Schmid, “P-CNN: Pose-based CNN features for action recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[56] T. Van Hertem *et al.*, “Automatic lameness detection based on consecutive 3D-video recordings,” *Biosyst. Eng.*, 2014.

[57] J. Huang and S. You, “Point Cloud Labeling using 3D Convolutional Neural Network,” in *International Conference on Pattern Recognition (ICPR)*, 2016.

[58] R. Singh Pahwa, J. Lu, N. Jiang, and T. T. Ng, “Locating 3D Object Proposals: A Depth-Based Online Approach,” *IEEE Trans. Circuits Syst. Video Technol.*, 2017.

[59] B. Pepik, M. Stark, P. Gehler, and B. Schiele, “Teaching 3D geometry to deformable part models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[60] B. Pepik, M. Stark, P. Gehler, and B. Schiele, “Multi-View and 3D Deformable Part Models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

[61] S. Srivastava, G. Sharma, and B. Lall, “Large Scale Novel Object Discovery in 3D,” *arXiv Prepr. arXiv1701.07046v2*, Jan. 2017.

[62] B. Li, T. Zhang, and T. Xia, “Vehicle Detection from 3D Lidar Using Fully Convolutional Network,” *Robot. Sci. Syst.*, Aug. 2016.

[63] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká, “3D Bounding Box Estimation Using Deep Learning and Geometry,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[64] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, “3D Object Proposals using Stereo Imagery for Accurate Object Class Detection,” *Adv. Neural Inf. Process. Syst.*, Aug. 2016.

[65] A. Garcia-Garcia, S. Orts-Escolano, S. O. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A Review on Deep Learning Techniques Applied to Semantic Segmentation.”

[66] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A Database and Web-Based Tool for Image Annotation,” *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, May 2008.

[67] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes,” Feb. 2017.

[68] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *arXiv Prepr. arXiv1706.02413v1*, Jun. 2017.

[69] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning Deep 3D Representations at High Resolutions," *arXiv Prepr. arXiv1611.05009v4*, Nov. 2016.

[70] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation."

[71] M. Atzmon, H. Maron, and Y. Lipman, "Point Convolutional Neural Networks by Extension Operators," *arXiv Prepr. arXiv1803.10091v1*, 2018.

[72] Hang Su *et al.*, "SPLATNet: Sparse Lattice Networks for Point Cloud Processing," *arXiv Prepr. arXiv1802.08275v2*, 2018.

View publication stats