DOI: 10.1049/cvi2.12141

The Institution of Engineering and Technology WILEY

GUEST EDITORIAL

Deep learning for 3D vision

With the rapid development of 3D imaging sensors, such as depth cameras and laser scanning systems, 3D data has become increasingly accessible. Meanwhile, the boost of various deep learning algorithms, such as convolutional neural networks and transformers, further increases the usability of 3D vision systems. Driven by these factors, 3D vision has become an emerging and core component for numerous applications, such as autonomous driving, augmented reality, virtual reality and robotics.

Although remarkable progress has been achieved in this area during the last few years, there are still several challenges that need to be addressed, such as the noisy, sparse, and irregular nature of point clouds, the high cost to label 3D data and the necessity to integrate geometry-based and learning-based techniques. Besides, 3D data produced by different 3D imaging sensors (e.g. structured light, stereo, LiDAR and time-of-flight) can be highly different. It is, therefore, necessary to investigate general algorithms that can mitigate the domain gap between different types of 3D data.

This special issue aims to collect and present the latest research development in learning-based 3D vision theories and their applications and to inspire future research in this area. In total, there are eight papers accepted for publication in this special issue through careful peer reviews and revisions. These accepted papers are broadly categorised into three topics, and the summary of each topic is given below.

TOPIC A—OPTICAL FLOW AND DEPTH ESTIMATION

Han et al., in their paper 'DEMVSNet: Denoising and Depth Inference for Unstructured Multi-View Stereo on Noised Images', proposed a DEMVSNet to simultaneously address the depth estimation and image denoising problems for unstructured multi-view stereo. The multi-scales feature maps for each image are wrapped to construct cost volumes containing both the depth and RGB information through differentiable homography and Gaussian probability mapping. The cost volume regularisation module is then adopted to predict the probability of depth and RGB. To avoid overfitting in multi-task learning, the gradient normalisation algorithm is utilised to dynamically fine-tune the weights between the depth prediction task and the denoising task. To evaluate the performance of proposed DEMVSNet, a noisy Technical University of Denmark dataset is generated by adding Gaussian-Poisson noise to each image, and the experimental results demonstrate the superiority of DEMVSNet on both the denoising and multi-view stereo reconstruction tasks.

Lin et al., in their paper 'EAGAN: Event-Based Attention Generative Adversarial Networks for Optical Flow and Depth Estimation', proposed an event-based attention generative adversarial network named EAGAN to simultaneously deal with optical flow and depth estimation based on monocular event camera. The generator of EAGAN is similar to U-net except that a transformer structure is introduced between the encoder and decoder. The position-coding features learnt from the transformer is added to features learnt from the encoding layer, which helps to capture the correlation between sequence information. The discriminator of EAGAN is based on a fully convolutional network and aims to distinguish whether the depth image or the optical flow image is generated by the generator. Experimental results conducted on the multi-vehicle stereo event camera dataset demonstrate the effectiveness of EAGAN on both the depth and optical flow estimation tasks.

TOPIC B—POSE ESTIMATION

Gao et al., in their paper 'Efficient 6D Object Pose Estimation based on Attentive Multi-Scale Contextual Information', proposed an end-to-end 6D pose estimation network to utilise multi-scale contextual features learnt from two heterogeneous data. First, interesting objects are detected from an RGB-D image using an existing semantic segmentation method. Then, pixel-wise geometric and colour features are learnt from 3D point clouds and 2D images respectively. Next, three pixelwise feature attention mechanism modules are utilised to exploit the inter-channel relationship of multimodal features. Finally, multi-scale features are extracted at three different scales and 6D pose is estimated through a dense regression module. Experimental results conducted on the LineMOD and YCB-Video datasets demonstrate that the proposed method achieves state-of-the-art performances in terms of average point distance and average closest point distance.

Liu et al., in their paper 'Auto Calibration of Multi-Camera System for Human Pose Estimation', proposed an iterative joint estimation of intrinsic and extrinsic parameters for a multi-camera system. Specifically, keypoints are detected with high confidence to estimate the essential matrix between two cameras, and the valid extrinsic parameters are estimated by assuming that the intrinsic parameters are known a priori. Then, the reconstructed 3D human body coordinates are projected into the pixel coordinate system, and the intrinsic parameters are estimated by minimising the projection errors. The experimental results show that the proposed method achieves better performance than commonly used calibration tools.

TOPIC C—POINT CLOUD PROCESSING AND UNDERSTANDING

Liu et al., in their paper 'Point Cloud Completion by Dynamic Transformer with Adaptive Neighbourhood Feature Fusion', utilised the adaptive neighbourhood feature extraction (ANE) module and genetic hierarchical point generation (GHG) module to accomplish the point cloud completion task. The ANE module selects k nearest points both in the spatial and feature spaces adaptively according to different target shapes. The GHG module generates finer point clouds hierarchically according to the local shape characteristics, and the shape information of current points is transferred to the next stage through a dynamic transformer structure. The experimental results conducted on the Point Completion Network and Completion3D datasets demonstrate the superiority of the proposed method.

Wang et al., in their paper 'PCCN-RE: Point Cloud Colourisation Network Based on Relevance Embedding', proposed a highly authentic point cloud colorisation network based on conditional generative adversarial (cGANs) networks. The generator network predicts the colours from the coordinates of each point, while the discriminator utilises the coordinates and the generated colours to determine the reality of input colourised point clouds. Three key components are contained in the generator. Specifically, the relevance embedding structure captures the most related local information, the weighted pooling structure aggregates the local features based on the correlation values of the covariance matrix, and the enhanced spatial transform network keeps the point clouds invariant to the geometric transformations based on weighted pooling and maximal pooling. The experimental results show that the proposed method achieves the highest Peak Signal to Noise Ratio and Structural Similarity Index on the ShapeNetCore dataset.

Fang et al., in their paper 'Sparse Point-Voxel Aggregation Network for Efficient Point Cloud Semantic Segmentation', proposed a sparse point-voxel aggregation network to overcome high computational costs in the point cloud semantic segmentation task. In the encoding layer, the local context features are learnt through a sparse convolutional network performed on the voxelised point cloud, and the individual point features are learnt through multi-layer perceptron (MLP)-based network performed on the original point cloud. In the decoding layer, these two kinds of features are aggregated at different encoding layers through simple MLP layers. The experimental results show that the proposed method achieves state-of-the-art performance on the SemanticKITTI and S3DIS datasets.

Wang et al., in their paper 'Scale Robust Point Matching-Net: End-to-End Scale Point Matching Using Lie Group', proposed an end-to-end scale point cloud matching network named SRPM-Net based on Lie Group. The extracted pointwise features are composed of point absolute coordinates, relative coordinates and point pair features of neighbouring points, and the local context features are aggregated through an attentive pooling layer. The matching matrix is computed via the exponential map of Lie group, which represents the feature similarity of points in two point clouds. The final transformation estimation problem is transferred as estimating the coefficients of the Lie algebra optimisation problem and is optimised through an iterative linear optimisation approach. The experimental results show that SRPM-Net achieves the best performance on the ModelNet40 and Stanford 3D scanning datasets.

SUMMARY/CONCLUSION

The papers published in this Special Issue show that traditional topics, such as optical flow and depth estimation, pose estimation, and point cloud processing have developed very fast in recent years. In addition, many topics have emerged in deep learning-based 3D vision, such as multi-task joint learning and multimodality intelligence. Future research in this field is expected to boost the theoretical development and potential applications of 3D vision.

Yulan Guo[1]
Hanyun Wang[2]
Ronald Clark[3]
Stefano Berretti[4]
Mohammed Bennamoun[5]

[1]*College of Electronic Science and Technology, National University of Defense Technology, Changsha, China*
[2]*School of Surveying and Mapping, Information Engineering University, Zhengzhou, China*
[3]*University of Oxford, Oxford, UK*
[4]*Media Integration and Communication Center (MICC) and at the Department of Information Engineering (DINFO), University of Florence (UNIFI), Florence, Italy*
[5]*Department of Computer Science and Software Engineering, UWA, Perth, Western Australia, Australia*

Correspondence
Yulan Guo, College of Electronic Science and Technology, National University of Defense Technology, 109 Deya Rd, Changsha, 410073, China.
Email: yulan.guo@nudt.edu.cn

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

AUTHOR BIOGRAPHIES

Yulan Guo received the BE and PhD degrees from National University of Defence Technology (NUDT) in 2008 and 2015, respectively. He has authored over 100 articles at highly referred journals and conferences. His current research interests focus on 3D vision, particularly on 3D feature learning, 3D modelling, 3D object recognition, and scene understanding. He served as an associate editor for IEEE Transactions on Image Processing, IET Computer Vision, IET Image Processing, and Computers & Graphics. He also served as an area chair for CVPR 2023/2021, ICCV 2021, and ACM Multimedia 2021. He organised several tutorials, workshops, and challenges in prestigious conferences, such as CVPR 2016, CVPR 2019, ICCV 2021, 3DV 2021, CVPR 2022, ICPR 2022, and ECCV 2022. He is a Senior Member of IEEE and ACM.

Hanyun Wang is currently an associate professor with the School of Surveying and Mapping, Information Engineering University, China. He received his PhD degree in Electronic Science and Technology from National University of Defence Technology in 2015. He was a visiting PhD student with Xiamen University from 2011 to 2014. He has authored over 30 articles in journals and conferences, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, and IEEE Transactions on Geoscience and Remote Sensing. His research interests focus on 3D computer vision, especially on point cloud registration, 3D object detection and scene understanding. He served as a reviewer for many journals, he also served as a guest editor for IET Computer Vision, and a chair of Special Session on 3D Computer Vision in ICVR 2022.

Ronald Clark received his PhD degree in Computer Science from the University of Oxford, and his BSc in Information Engineering from the University of the Witwatersrand. His work lies at the intersection of computer vision and machine learning, and mainly focuses on allowing machines to interpret and understand the 3D world around them. Before coming to the UK, he received his MSc in electrical engineering at the University of the Witwatersrand in South Africa. He has received numerous international awards including a best paper honourable mention at the Conference on Computer Vision and Pattern Recognition (CVPR). His research has been supported by a number of prestigious fellowships, including an EPSRC doctoral studentship, a Dyson Fellowship and most recently an Imperial College Early Career Fellowship.

Stefano Berretti is currently an associate professor at the Media Integration and Communication Centre (MICC) and at the Department of Information Engineering (DINFO) of the University of Florence (UNIFI), Florence, Italy. He has been also Visiting Professor at the University of Lille and the University of Alberta. His research interests are in the areas of computer vision, pattern recognition and multimedia. On these themes he published more than 200 papers in some of the most distinguished journal and conferences. He has been also actively involved in the organization of conferences and workshops, among which ICMCS 1999, ACM MM 2010 and ECCV 2012. He was the general chair of STAG 2020 and 3DOR 2022. He is an Associate Editor of ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMM) and of the IET Computer Vision journal. From 2016 to 2021 he was also the Information Director ACM TOMM.

Mohammed Bennamoun is currently a Winthrop professor with the Department of Computer Science and Software Engineering, UWA and is a researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published four books (available on Amazon), one edited book, one Encyclopaedia article, 14 book chapters, more than 120 journal papers, more than 250 conference publications, 16 invited & keynote publications. His h-index is 50 and his number of citations is more than 11,000 (Google Scholar). He was awarded more than 65 competitive research grants, from the Australian Research Council, and numerous other Government, UWA and industry Research Grants. He successfully supervised more than 26 PhD students to completion. He won the Best Supervisor of the Year Award at QUT (1998), and received award for research supervision at UWA (2008 & 2016) and Vice-Chancellor Award for mentorship (2016). He delivered conference tutorials at major conferences, including: IEEE Computer Vision and Pattern Recognition (CVPR 2016), Interspeech 2014, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) and European Conference on Computer Vision (ECCV). He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).