*This is a pre-copyedited, author-produced version of an article accepted for publication in Lecture Notes on Data Engineering and Communications Technologies following peer review. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-70665-4_96*

# Semantic 3D Scene Classification Based on Holoscopic 3D Camera for Autonomous Vehicles

Chuqi Cao[1], Mohammad Rafiq Swash[2], and Hongying Meng[3]

[1,2,3]Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, UK

[1] Chuqi.Cao@brunel.ac.uk
[2] Rafiq.Swash@brunel.ac.uk
[3] Hongying.Meng@brunel.ac.uk

**Abstract.** An autonomous vehicle navigates by perceiving the environment through the sensors and acting on the received data by making sense of the surroundings. In this paper, innovative holoscopic 3D scene classification based on the single aperture holoscopic 3D camera is proposed to recognise continuous 3D scene of environment. The deep learning network AlexNet is used to evaluate the proposed approach, and the outcome exhibits promising results compared to 2D images.

**Keywords:** autonomous vehicle, perception system, holoscopic 3D image, 3D image, artificial intelligence, deep learning, scene classification

## 1 Introduction

Autonomous car can guide itself to a specific target without or with minimum human conduction. It does this by using intelligent algorithms that decide on vehicle's key functions i.e. considering perception of the environment around it through the fusion of sensor observations. An autonomous car is also known as a driver-less car, robot car, self-driving car or autonomous vehicle [3].

The sensors such as cameras and laser range finders are mounted on an autonomous car in strategical points. The observations are used to monitor the movement of vehicles, pedestrians, and potential obstacles around it, much more thoroughly than an average driver[5], especially in the blind spots. As a result, the intelligent algorithms responsible for decision making and control of the vehicle, can safely react and also predict potential hazards. The high-end sensors like Light Detecting and Ranging (LiDAR) sensors, Global Positioning Systems (GPS) and Inertial Measurement Unites (IMU), are very expensive to acquire, however there are cheaper sensors like holoscopic 3D (H3D) camera, which endeavours to provide a new method of creating and representing 3D images. It is a unique method for creating a true volume spatial optical model of the object scene in the form of a planar intensity distribution through a micro-lens array [17]. It uses natural light and a single aperture camera setup with a micro-lens array in the capture process to offer a full parallax object scene as in the real-world.

With advances in deep learning and growing availability of super-fast processing capabilities such as GPU processors and inspired by work done on autonomous driving by NVIDIA [4], using the H3D camera, rich content of perception will support key points required for autonomous vehicle, like environment recognition and object detection. This will lead to a more affordable automation solution that can be realised using just the latest vision technology.

## 2 Methodology

### 2.1 Overview

In this paper, the targeted scene was captured and processed under H3D imaging system, via shooting the H3D video by H3D camera, segmenting raw video according to the scene, then frame sampling and extracting them into viewpoint images. These viewpoint images were fed into deep learning neural network, and classified as different scene labels after training. Figure 1 displays this workflow.

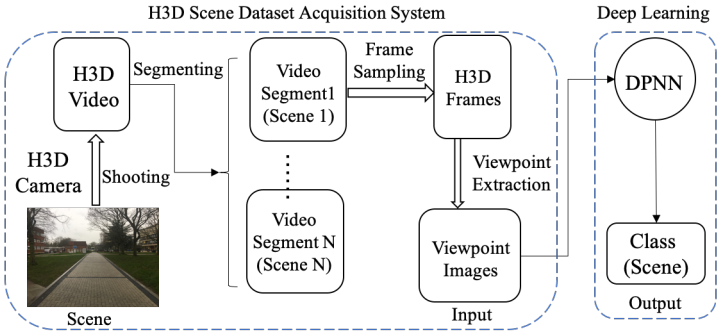


**Fig. 1.** Workflow

### 2.2 Holoscopic 3D Imaging System

H3D imaging (also referred to as integral imaging) was first introduced by Gabriel Lippmann in 1908 [11] when he proposed integral photography as a technique for recording and reproducing 3D contents. It is a true 3D imaging technology based on the "fly's eye" technique that uses coherent repetition of light to build a true spatial 3D scene [11]. This technique uses unique optical components to create and represent a true volume spatial optical model of the object scene in the form of a planar intensity distribution. H3D image is recorded by using a regularly spaced array of small lenslets closely packed together in contact with a recording device [11] [1]. Each lenslet views the scene at a slightly different angle to its neighbour so the scene is captured from many view points

and parallax information is recorded [1]. Figure 2 illustrates the design of H3D camera which has been developed in this paper according to the design specification. The main components of the H3D camera are prime lens, micro-lens array, relay lens and digital camera sensor. The micro-lens array is mounted in adjacent to the camera sensor. The prime lens image plane is formed in front of the micro-lens array, which allows the micro-lens array to capture the positions from different perspectives in the scene. A Sony alpha 7R camera sensor is used to assemble a holoscopic 3D camera that has 35mm full-frame image sensor and 4K HD to ensure ultra-good quality of videos and images captured for database. Nikon Nikkor AF 35mm f/2.0 is used as the prime lens and Rodenstock Rodagon-N APO 50mm f2.8 Enlarging Lens is used as the relay lens.

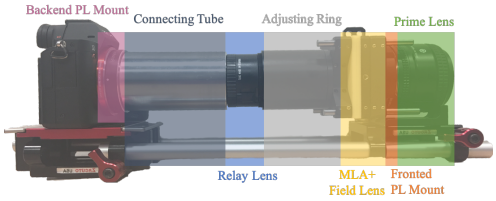


**Fig. 2.** Holoscopic 3D Camera Prototype

A holoscopic 3D image consists of multiscopic elemental images which can be used to extract multiview 2D view point images by extracting a pixel from a given index location of elemental images [2]. Figure 3 illustrates an example of H3D image and viewpoint images extracted from it.

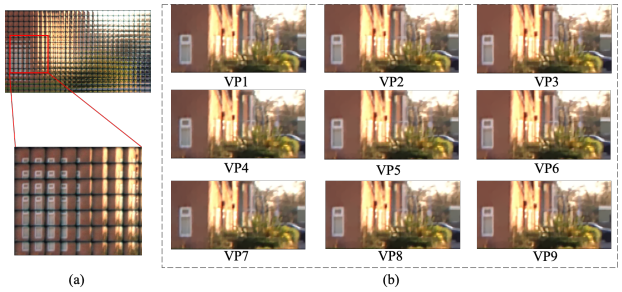


**Fig. 3.** (a) Recorded holoscopic 3D image with magnified part (b) Different viewpoint images

### 2.3 Deep Learning Model

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction, which dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains [16]. One most commonly used DNN is Convolutional Neural Networks (CNN) which is mainly used for image and video [12]. CNNs have revolutionised the computational pattern recognition process [8]. Their adoption has exploded in recent years due to two important developments. First, large, labelled data sets such as the ImageNet [10] for Large Scale Visual Recognition Challenge (ILSVRC), visual geometry group image for face descriptors [14] and CFIR dataset for different classes of objects [7] are now widely available for training and validation. Second, CNN learning algorithms are now implemented on massively parallel graphics processing units (GPUs), tremendously accelerating learning and inference ability.

AlexNet is a CNN designed by Alex Krizhevsky et al. [8], that is trained on more than a million images from the ImageNet database [10]. It won the 2012 ImageNet LSVRC-2012 competition by a large margin (15.3% VS 26.2% (second place) error rates) [9]. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of the convolutional layers are followed by max-pooling layers, and three globally-connected layers with a final 1000-way softmax. The non-saturating neurons and a very efficient GPU implementation of convolutional nets were used to make training faster [8]. The top highlight of this model is that using overlapped pooling to reduce the size of network. It reduces the top-1 and top-5 error rates by 0.4% and 0.3% respectively. Also ReLU was used instead of Tanh to add non-linearity, which accelerates the speed by 6 times at the same accuracy [6].

## 3 H3D Scene Dataset Acquisition and Preparation

The data acquisition has been conducted with the H3D camera assembled on RC car as shown in Figure 4. The route of data shooting is a sidewalk with variety of surrounding scenes, where there are coloured sidewalk bricks forming several lines on the ground which are equidistant from each other. For scene classification, these brick lines became the markers of scene segmentation automatically. The route between every two adjacent brick lines is called one segment. At video shooting stage, 10 segments contained in each video which can be determined to 10 classes, as one specific class in each video contains the same scene. In data acquisition, the H3D camera started to shoot when the RC car started to run, and stopped when the RC car reached the end of the 10th segment.

There are 28 H3D videos have been taken and time duration of each video is between 2 minutes to 3 minutes, in total there are 280 video segments with average of 12 seconds to 18 seconds each. Since the frame rate is 25 fps there are around 300 to 450 frames in every video segment. These frames are sampling in order to simplify the data processing, the sampling interval is 50 frames which means select one frame in every 2 seconds. This sampling action ensures the

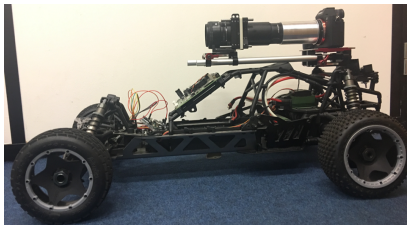

**Fig. 4.** H3D camera assembled on RC car for 3D data acquisition

sampled frames to maintain proper diversity and reduce the repetition of content and scene between adjacent frames.

While extracting H3D raw frames to viewpoint images, there are 9 viewpoints chosen from 40 to 50 in $x$ axis and from 30 to 40 in $y$ axis: $(40, 30)$, $(40, 35)$, $(40, 40)$,$(45, 30)$, $(45, 35)$, $(45, 40)$, $(50, 30)$, $(50, 35)$ and $(50, 40)$. The resolution of each viewpoint image is $3840 \times 2160$. By initial experiment of training in proposed AlexNet using small dataset, the result tends to be not as good as expected. The reason could be the bottom half of every viewpoint image is the same - the ground of the sidewalk. Therefore, the viewpoint images are cutting to half to reduce the repeated features, and the resolution of half viewpoint image is $3840 \times 1080$ px. The structure of H3D RC car video database is illustrated in Figure 5 with an example of one video segment.

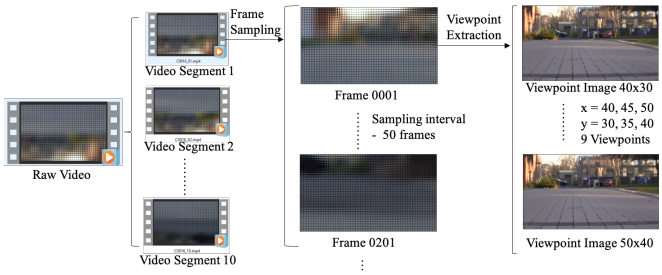


**Fig. 5.** The structure of H3D RC car video dataset

All viewpoint frames from video segment 01 of each H3D video are classified as Label 01, viewpoint frames from video segment 02 of each H3D video are classified as Label 02,...etc. There are 12276 sampled viewpoint frames altogether in 10 labels, Table 1 displays the number of frames in each label.

## 4 H3D Scene Classification using AlexNet

The standard pre-trained AlexNet implement of DeepLearning Toolbox in MATLAB is proposed to train H3D scene database for classification as it will allow

**Table 1.** Number of viewpoint frames in each label

| Class Label | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| No. of VP Frames | 1431 | 1296 | 1269 | 1071 | 1152 | 1116 | 1125 | 1143 | 1134 | 1539 |

to carry out a like-2-like comparison with 2D images and demonstrates performance of holoscopic 3D camera. The GPU used is NVIDIA Tesla K40c, which provides a high performance with its $12GB$ memory size, $745MHz$ clock speed and $288.4GB/s$ bandwidth. There are 25 layers including the input and the output layer. It contains 5 convolutional layers and 3 fully connected layers. ReLU is applied after every convolutional and fully connected layer. Dropout is applied before the second and the third fully connected layer. The input image size is $227 \times 227 \times 3$ but the images in the database have different sizes. Thus an augmented image datastore is used to automatically resize the training images. The last fully connected layer and the output layer are replaced to match the number of 10 classes of this H3D scene database.

## 5 Experimental Results and Evaluation

There are 12276 viewpoint images in the database, the train set, validation set and test set are split randomised by MATLAB at the split ratio of 70% : 15% : 15%. The initial learning rate is set to 0.0003, the validation frequency is set to 895, and the mini batch size is set to 10. The maximum epoch number is set to 50, the network takes 42950 iterations of 50 epochs in 663 min 13 sec to train.

The final validation accuracy turns to 96.47% and the test accuracy is 96.31% that is approximate to the validation accuracy. The confusion matrix of the test result is displayed in Figure 6. The accuracy of label 10 has the best accuracy of 99.6%, label 05 and 09 has the least accuracy of 92.5% and 93.5%, while all of other labels have accuracy above 95%. This final accuracy of over 96% can be determined as an outstanding performance of the network. The deep convolutional neural network of AlexNet displays its superiority on H3D scene database.

In order to compare with normal 2D scene database, an open source indoor scene recognition is used presented in here[15] [13]. This database contains 67 Indoor categories, and a total of 15620 images. Because the number of images varies across categories, ten categories with the most images are selected. There are 5844 images of 10 categories, while H3D scene database contains 12276 images, so these selected indoor scene images have been flipped to mirror images. Along with original images, there are 11688 images in this 2D indoor scene database, which is as similar size as H3D scene database.

The split ratio of 2D indoor scene database is set to 75% : 10% : 15%, thus there are 8766 images in train set and 1753 images in test set. This split ra-

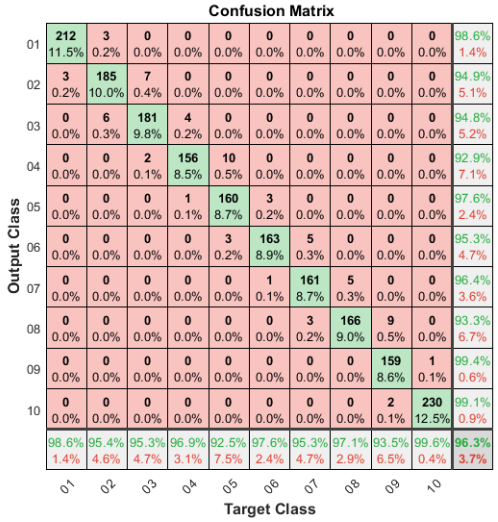


**Fig. 6.** Confusion matrix of the test result

tio maintains similar quantity of images of train set and test set as H3D scene database. Other training parameters stays same with the last training. The accuracy curve of training is almost 100% at the end of epoch 50, but the validation accuracy stays around 90%, at the same time the loss curve of training is almost 0, while the validation loss is still very high at about 0.5. The final accuracy of validation is 88.09% and accuracy of test set is 88.82%, they are not performing well compared to the proposed approach. Comparing the results of 2D scene database, AlexNet demonstrated H3D scene database delivers a remarkable performance.

## 6 Conclusion

The state of the art perception system lacks environment cognitive instead delivers only detected or recognised object information. In this paper, an innovative holoscopic 3D scene classification is proposed to provide environment cognitive by recognising continuous 3D scene where different hazard level can be incorporated. The experiment of H3D perception system was retrofitted with an RC car for holoscopic 3D capturing while the car was in driving. Holoscopic 3D database is prepared and used to train using standard deep learning AlexNet model. The experiment result demonstrated that holoscopic 3D content delivers a better accuracy of over 96% compared to traditional 2D content, which proved the concept that even compared with similar size 2D scene database, the network of AlexNet has learned the H3D database better. It is believed that by

training with different split ratio of train / validation set, and with tuning the training parameters, it could achieve a higher accuracy. In this paper, the deep convolutional neural network of AlexNet displays its superiority on H3D scene database and it also demonstrated that H3D scene database has an outstanding performance on deep learning. This innovative approach provides a more affordable automation solution that can be fairly meaningful to continuous 3D scene recognition for autonomous vehicles.

## References


1. Aggoun, A., Tsekleves, E., Swash, M.R., Zarpalas, D., Dimou, A., Daras, P., Nunes, P., Soares, L.D.: Immersive 3D holoscopic video system. IEEE MultiMedia **20**(1) (January 2013) 28–37
2. Alazawi, E.: Holoscopic 3D image depth estimation and segmentation techniques. PhD thesis, Brunel University London (2015)
3. Antsaklis, P.J., Passino, K.M., Wang, S.: An introduction to autonomous control systems. IEEE Control Systems Magazine **11**(4) (1991) 5–13
4. Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U.: Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 (2017)
5. Burke, K.: How Does a Self-Driving Car See?-Camera, radar and lidar sensors give autonomous vehicles superhuman vision.
6. Gao, H.: A Walk-through of AlexNet. (2017)
7. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (05 2012)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. (2012) 1097–1105
9. Lab, S.V.: Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). (2012)
10. Lab, S.V.: ImageNet. (2016)
11. Lippmann, G.: Épreuves réversibles donnant la sensation du relief. Journal de Physique Théorique et Appliquée **7**(1) (1908) 821–825
12. Olah, C.: Understanding Convolutions. (2014)
13. Oliva, A.: Indoor Scene Recognition - Database. (2009)
14. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Procedings of the British Machine Vision Conference 2015, British Machine Vision Association (2015)
15. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (jun 2009)
16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
17. Wang, H.x., Xu, Z.l., Li, Z.p., Wu, C.h.: 3d reconstruction from integral images based on interpolation algorithm. In: 5th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Advanced Optical Manufacturing Technologies. Volume 7655., International Society for Optics and Photonics (2010) 76552Z