Journal of King Saud University – Computer and Information Sciences 34 (2022) 1663–1673

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Stereo matching algorithm based on deep learning: A survey

Mohd Saad Hamid [a], NurulFajar Abd Manap [b], Rostam Affendi Hamzah [a,*], Ahmad Fauzan Kadmin [a]

[a] Fakulti Teknologi Kejuruteraan Elektrik & Elektronik, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia
[b] Fakulti Kejuruteraan Elektronik & Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

ARTICLE INFO

Article history:
Received 17 May 2020
Revised 7 August 2020
Accepted 23 August 2020
Available online 28 August 2020

Keywords:
Stereo matching algorithm
Deep learning
Convolutional neural network
Artificial intelligence

ABSTRACT

The development of stereo matching algorithm is still one of the challenging problems, especially in ill-posed regions. Hence, this article presents a survey on the algorithm frameworks related to the stereo matching algorithm. Based on the early survey that had been conducted, two major frameworks available in current stereo matching algorithm development, they are traditional and artificial intelligence (AI) frameworks. Most of the traditional methods are very low accuracy compared to the AI-based approach. This can be observed in the standard benchmarking dataset, such as from the KITTI and the Middlebury, where AI methods rank at the top of the accuracy list. Additionally, the trend for solving computer vision problems uses AI or machine learning tools that become more apparent in recent years. Thus, this paper is focusing on the survey between the deep learning frameworks, which is one of the machine learning tools related to the convolutional neural network (CNN). Several mixed approaches between CNN based method and traditional handcraft method, as well as the end to end CNN method also discussed in this paper.

Contents

1. Introduction

Recently the rapid growth in computer science and technology intensifies the various implementation of computing platforms in many aspects of our day-to-day activities. It changed the way we perform our daily jobs. The decision-making process in our daily

* Corresponding author.
E-mail address: rostamaffendi@utem.edu.my (R.A. Hamzah).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.08.011
1319-1578/

routines also influenced by the advancement in computing technology. For decades, the researchers in the artificial intelligence (AI) field attempt to create the ultimate intelligent machine. This perfect machine will help to make decisions based on the input given. In recent years, there are so many articles published in the area of AI. As mentioned in Chollet (2017), AI has become a concentrated topic in exaggerated publicity by the mass media. One of the exciting areas in AI these days is machine learning. The beauty of a machine learning algorithm is the capability to instruct the computer to react or making decisions on certain conditions without having to program the computer explicitly. Due to the flexibility of machine learning, the algorithm capable of learning via self-train, analysis, observation, and experience. Machine learning algorithm capable of adapting new situations through pattern and trend detection for better results. There are various machine learning applications related to our daily activities, such as computer vision, syntactic pattern recognition, natural language processing, search engines, and machine perception. Driven by the rapid development in computer hardware and software, we have unlimited possibilities to implement a machine learning algorithm. As mentioned in Forbes.com article (Bernard, 2018), the trend and direction for the manufacturing industry now and the future as disclosed in Industry 4.0, are moving towards smarter and autonomous systems. The intelligent machines will be connected and communicate with one another for critical autonomous decision-making systems through the implementation of the machine learning algorithm and fueled by multiple class of input data.

1.1. Deep learning

Deep learning existed for more than three decades ago. It is another branch of machine learning that has become the leading research focus in recent years. According to Gibson and Patterson (2016), the common definition of deep learning involves a neural network that contains more than two layers. It also inspired by how our human brain learns based on the different amounts of data. As explained by Ketkar (2017), deep learning term refers to the multiple hierarchies involve to learn from raw input data. This hierarchical learning requires neural networks with multiple layers to learn raw input data and transform it into something meaningful based on how we want to define the conclusion. Most of the implementation of a deep learning network is based on an artificial neural network that contains a hidden layer in addition to the input and output layer. The authors of Gibson and Patterson (2016) discussed four types of network architectures in their book. The types of architectures mentioned in their book were: unsupervised pre-trained networks, convolutional neural networks (CNN), recurrent neural networks (RNN), and recursive neural networks. Deep learning capable of solving problems in many areas such as computer vision, speech-audio processing, natural language processing (Goodfellow et al., 2016).

Deep learning became popular in the area of computer vision when the author of Ciresan et al. (2012) published their work on the effectiveness of CNN in the computer vision area. They showed that running CNN on the graphics processing unit (GPU) improved recognition rates in many vision benchmark databases such as MNIST, NIST SD 19, CIFAR10, and NORB. The following advancement on CNN and deep learning began to flourish the same year made by Krizhevsky et al. (2012) when they won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. The way CNN revolutionize the computer vision related to the earlier works led by LeCun. The related works (Lecun et al., 1998, LeCun et al., 1989) provide the foundation for CNN to evolve in the computer vision area. LeNet architecture presented by Lecun et al. (1998) contains seven layers illustrated in Fig. 1 (except the input layer). More layers and larger of CNNs are required to handle higher resolution images.

The constraint on the CNN technique is the computing power of the machine. However, as shown in earlier works (Ciresan et al., 2012; Krizhevsky et al., 2012), the implementation of GPU capable of catering to the limitation. The computational power required by CNN mainly in two phases (Malita et al., 2018). The first one during a training session where the learning process happens, and network weights will be set up based on the training input. The second phase is during the inference phase, where usually the real application executed and the network with the proper training data can classify the testing input. So, the advancement in computing technology such as in the high-end personal computer (PC) and the GPUs enable the researchers to reduce training time from weeks in the past into only hours for their CNNs nowadays. Flynn et al. (2016) proposed DeepStereo to solve the problem related to the image-based rendering (IBR) area. Flynn et al. (2016) also mentioned about the possible improvement on their network using GPUs to achieve real-time performance. In related work (Fangmin et al., 2017) on 3D face reconstruction also mentioned the effectiveness of the deep learning approach, which helps to build more accurate and fast method compare to the traditional way. The trend in deep learning implementations conveys a better future for deep learning to evolve in the future.

The research work on CNN helps to improve many algorithms. Smith and Smith (2018) described that deep learning through the implementation of CNNs has advanced the innovation of machine

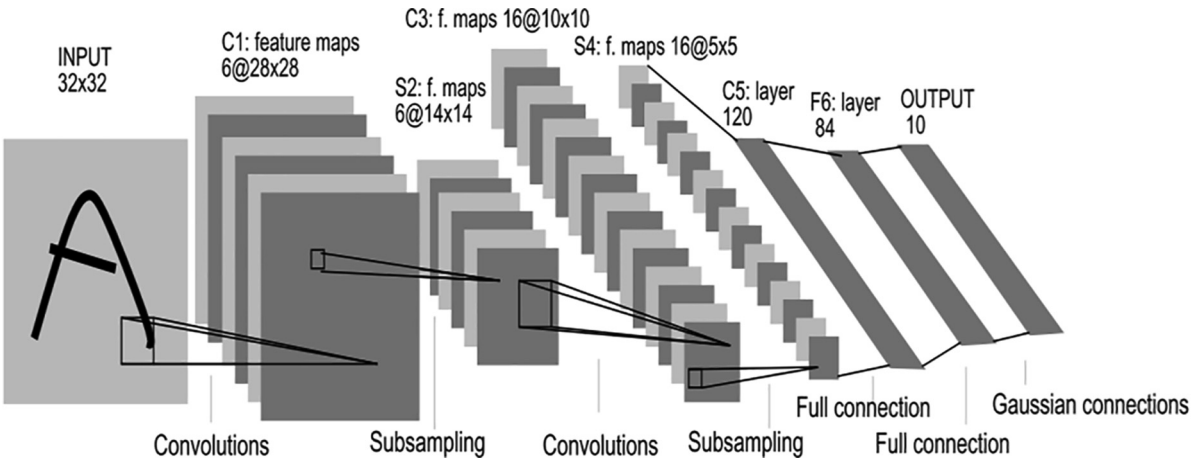

Fig. 1. Architecture of LeNet-5 for character recognition.

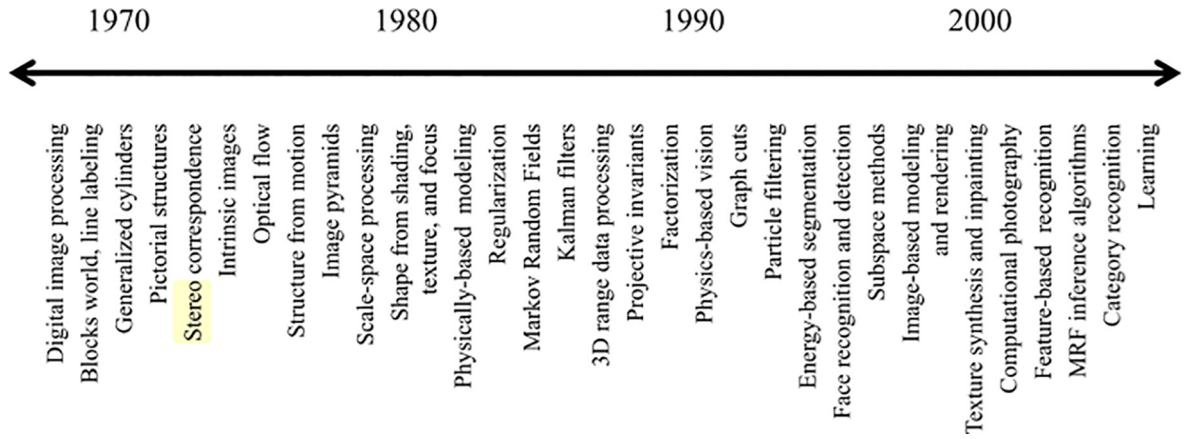

Fig. 2. Timeline Showing Active Topics in Computer Vision (Szeliski, 2011).

vision for the outdoor environment. Vu et al. (2018) successfully designed a 3D-CNN model to perform a 2D image and 3D volumetric classification by extending LeNet-5 CNN. Dense Convolutional Network (DenseNet) proposed by Huang et al. (2017), connect a layer to every other layer through the feed-forward direction. They highlighted that DenseNets could be good feature extractors and suits for many computer vision problems. Following the DenseNet, another researcher, Swami et al. (2019), implemented DenseNet in the stereo matching problem. They presented the end-to-end network for disparity estimation, DISCO (Depth Inference from Stereo using Context), to solve the stereo vision problem.

1.2. Stereo vision

3D shape and appearance reconstruction from images through mathematical techniques have been active in the computer vision research field. We are now capable of reconstructing the 3D model of the environment based on multiple images through established methods in computer vision (Flynn et al., 2016; Szeliski, 2011). As highlighted by Hodges et al. (2019), computer vision has been implemented ubiquitously in modern technologies. Computer vision also being used in multiple industrial applications such as optical character recognition (OCR), machine inspection, medical imaging, automotive safety, surveillance, and others. Based on the active topics on computer vision shown in Fig. 2, Szeliski (2011) stated that the trend to implement machine learning to solve computer vision problems becomes more prominent. Yang et al. (2019b) proposed a framework to perform a quality assessment on the stereo image using the deep belief network (DBN).

Vision-based object tracking and robot navigation are some of the examples of the research output established from research work on the computer vision field. Both are based on depth information obtained from images from the image sensor or camera. This is a part of the stereo vision research area, which provides a wide range of applications. As mentioned in (Scharstein and Szeliski, 2002), the main work in stereo vision is to get the depth information, extracted from a pair of rectified images as input. In relation to the depth information, the work on stereo vision also helps to extract 3D information from the pair image (left and right) taken from a different angle (Cambuim et al., 2017; Malekabadi et al., 2019; Salehian et al., 2018). Fig. 3 illustrated the structure of stereo vision (Hong and Kim, 2017), where both left and right camera separated by a distance of B. The view from the stereo camera also almost similar to human eye perception. The distance D between target object and cameras can be determined by using the equation (1), through camera focal length, f, and disparity, d

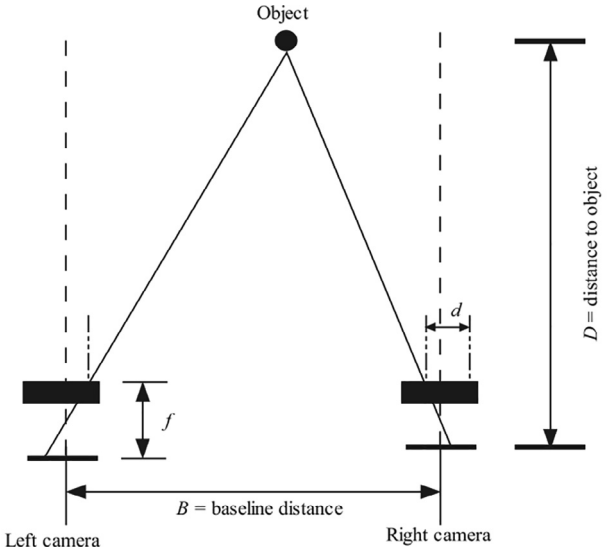

Fig. 3. Basic Structure of Stereo Vision (Hong and Kim, 2017).

where the location difference of the target object inside reference and target images yield the disparity value, d.

$$D = \frac{B.f}{d} \tag{1}$$

1.3. Stereo matching and depth/dense/disparity map

The study on the stereo vision mainly focuses on stereo matching (Damjanović et al., 2012). Stereo matching still remains a challenging area to this day (Pang et al., 2018; Xue et al., 2019; Zhu et al., 2016). In the stereo matching algorithm, the main objective is to find the disparity value calculated based on the object in the left and right image pairs. Disparity value obtained based on the differences in the pixel location of particular corresponding features recorded in the left and right images. This is similar to the images seen by the left and right eyes of the human vision system (Scharstein and Szeliski, 2002). The disparity map is the output of the stereo matching algorithm ((Chen et al., 2018). The distance between the camera and object revealed through the depth or disparity map (Fu and Liang, 2019). The importance of depth perception by stereo matching also highlighted by Smolyanskiy et al. (2018). According to the authors, the depth information enables us to use for multiple applications such as scene reconstruction,

virtual and augmented reality, obstacle avoidance, and several other applications. Another related work (Fu et al., 2019) performs stereo matching for 3D face reconstruction. They used the spatial-temporal integral image (STII) for faster matching cost computation in stereo matching for the reconstruction process. According to (Ma et al., 2019), the information from the disparity map will provide more insight into 3D projective transformation. A similar 3D depth perception study also was done based on the mantis vision system (Nityananda et al., 2018). They found that insect stereopsis becomes more efficient and robust to low-resolution image and becomes more responsive to the deviations in the pattern of luminance between the two eyes. The mentioned research works on the stereo vision will open new frontiers on the stereo matching area when we utilize the unique traits of the natural elements surrounding us.

1.4. Traditional approach for stereo matching

As mentioned by Hamzah and Ibrahim (2016), both local and global approaches are the main categories in the stereo vision algorithm. Due to the way the local approach method implemented, the disparity computation at any point in the image determined by the intensity values within a predefined support window. The local approach also called the window-based approach. Because of this behavior, the local approach capable of running faster with low computational (Popovi et al., 2018). Due to this reason, Sangeetha et al. (2018) chosen the local method for their robotic arm real-time application. The global method is another interesting topic on stereo matching. Wang et al. (2016) mentioned the global method produces disparity based on energy minimization process, which is commonly based on Markov Random Field (MRF). The global method provides better accuracy for the disparity output. However, it will incur more computational complexity (Sangeetha et al., 2018). The energy minimization in the global method focuses on data term and smoothness term (Scharstein and Szeliski, 2002). The previous study on the stereo vision algorithm by Hamzah and Ibrahim (2016), Li et al. (2016), Scharstein and Szeliski (2002) mentioned there are four main steps to produce disparity map from the stereo based algorithm. The matching step is the most important step in stereo vision (Salehian et al., 2018). The summary of the general steps in the stereo vision algorithm done by several authors (Hamzah and Ibrahim, 2016; Scharstein and Szeliski, 2002) can be illustrated as per following Fig. 4.

Scharstein and Szeliski (2002) also described that different algorithms might employ different step sequence combination. For example, based on the explanation by Szeliski and Scharstein, the aggregation step was often skipped in the global approach due to the redundant purpose of global smoothness constraint when it performs optimization step after the disparity computation step. The main steps for the stereo vision algorithm, as depicted in Fig. 4, will be explained in the following subsection.

1.4.1. Disparity computation

This step involves the calculation of the cost of assigning a special disparity to each pixel (Salehian et al., 2018). For the local approach, Sangeetha et al. (2018) implemented the sum of absolute difference algorithm (SAD) at the matching cost computation stage after performing comparisons on commonly used cost functions. They performed comparisons based on essential parameters for real-time algorithms, which are the computational time and memory requirement. The algorithm compares each block of pixels in the reference image with the matching blocks in the pair image and takes the absolute difference. These pixel differences in terms of pixel intensities then summed together to determine the dissimilarity between the two images (Sangeetha et al., 2018), as stated in Eq. (2).

$$\boldsymbol{C_{SAD}}(\boldsymbol{p},\boldsymbol{d}) = \sum_{\boldsymbol{q}\in\aleph_{\boldsymbol{p}}} |\boldsymbol{I^L}(\boldsymbol{q}) - \boldsymbol{I^R}(\boldsymbol{q}-\boldsymbol{d})| \tag{2}$$

For each position p and disparities d, the matching cost C_{SAD} will be calculated. $I^L(p)$ and $I^R(p)$ represent image intensities at position p in the left and right images. The $\aleph_p$ contains the set of locations within a fixed rectangular window centered at position p. The basic idea representing the cost is cost will be high when the two patches are centered around the image in different 3D points, and vice versa. Other methods used for cost computation are the sum of squared difference (SSD), normalized cross-correlation (NCC), Zero Mean Normalized Cross-Correlation (ZNCC), rank transform and census transform (CT) as explained in Hamzah and Ibrahim (2016) and Scharstein and Szeliski (2002).

1.4.2. Cost aggregation

The second stage in the pipeline is the cost aggregation step. As mentioned earlier, some algorithms might skip this step, especially the algorithms which implement under the global approach. Typically in the local approach, the cost aggregation step involves summing or averaging over a support region in the disparity space image (DSI) (Xu et al., 2014). The direct approach for cost aggregation, as mentioned by Zhu et al. (2016), is to equipped fixed kernel size to a low pass filter such as box filter and Gaussian filter. Another example mentioned by Scharstein and Szeliski (2002) is the binomial filter, where they used the separable Finite Impulse Response (FIR) filter. Another type of filter used for the purpose is the edge-preserving filter such as bilateral filter (BF), guided filter (GF) which preserve good edge and better results in the aggregation process (Hamzah and Ibrahim, 2016; Xu et al., 2014; Zhu et al., 2016). Zhu et al. (2016) proposed Adaptive Edge-Preserving Guided Filter (AEGF) in their work, which produced an accurate performance for the indoor and outdoor environment. Williem and Park (2018) also mentioned in their article that the guided filter performs better than the other algorithm based on their work on depth estimation.

1.4.3. Disparity computation/optimization

Winner Take All (WTA) optimization responsible for assigning the disparity map value (Cambuim et al., 2017; Hamzah and Ibrahim, 2016; Malekabadi et al., 2019; Zeglazi et al., 2018). Where in WTA, the disparity associated with the lowest cost value is chosen at each pixel (Szeliski, 2011). The equation (3) for the WTA step as follows:

$$\boldsymbol{d} = \arg\min_{\boldsymbol{d}\in\boldsymbol{d_r}}(\boldsymbol{C}(\boldsymbol{p},\boldsymbol{d})) \tag{3}$$

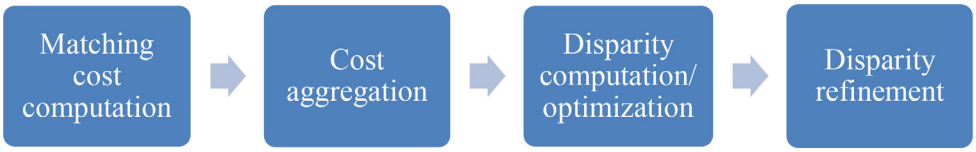

Fig. 4. Stereo Vision Algorithm Steps.

where C(p,d) is the cost volume after aggregation, dr is the range of the allowed disparity values. The maximum disparity value of dr is based on the ground truth of the disparity map. Other examples of the optimization stage are dynamic programming (DP), simulated annealing (SA), scanline optimization (SO), and graph cut (GC), as discussed by Scharstein and Szeliski (2002).

1.4.4. Disparity refinement

After the third step, the generated disparity or depth map may contain noises, errors such as invalid matches, and occlusion. Some of the methods to solve the problem is to implement methods such as slanted plane smoothing for occlusion problem. The Left-to-right-consistency (LRC) check will help to detect invalid pixels (Hamzah and Ibrahim, 2016; Zhu et al., 2016). At this step, multiple filtering techniques will be used to reduces the noises and errors. Gaussian convolution and median filter techniques are commonly used for local refinement. This last step also might introduce additional timing to the overall process due to the complexity. Huang and Zhang (2016) propose two methods, belief propagation and belief aggregation method for disparity refinement step. The author of (Damjanović et al., 2012) applied the median filter to the initial disparity maps in their works.

2. Deep learning on stereo matching development

As mentioned earlier, deep learning has become the catalyst to evolve the stereo vision area. Deep learning implementation has boosted the performance of stereo vision applications, as described by Krizhevsky et al. (2012). Furthermore, as mentioned by Ciresan et al. (2012), the traditional stereo vison cannot match the human performance for recognition tasks. But with the assimilation of deep learning into their algorithms, they can match human performance. Since then, researchers around the globe have been working to refine and implement deep learning in real-world stereo vision applications. The advancement of the implementation of machine learning also affected the study on stereo vision (Tonioni et al., 2017). For example, in the image classification area, Chauhan et al. (2019) implemented CNN for their works in transport engineering for vehicle counting and classification. They were able to classify different types of vehicles using their CNN based counting system. In the following section, we will discuss the implementation of deep learning for the stereo matching algorithm.

2.1. CNN based stereo matching algorithm

Over the years, the implementation of CNN in stereo vision has been tremendously excellent. For the area of stereo matching, the application of CNN pioneered by Zbontar and LeCun (2016), Zbontar and LeCun (2015). The authors described how they implement CNN to compute matching cost in their article. Their MC-CNN-acrt network used to produce matching costs for disparity or depth map displayed in Fig. 5. The eight layers of CNN based network fed by 9x9 gray image patches. The first layer is a convolutional layer with 32 kernels of 5x5 size. The other seven layers implemented were fully connected layers. The output vectors of the convolutional layers (Layer 1) passed through Layer 2 and 3 (fully connected layer with 200 neurons each). The vectors processed from left and right image patches concatenated together from 2 channels of the 200-dimensional vector (left and right) into a single 400-dimensional vector. This will be passed to Layer 4 of the architecture mentioned by the authors. The next layer, Layer 4, until Layer 7 are the single layer with 300 neurons each. The final Layer 8 produced a distribution of good and bad match classes.

They accompanied all the layers with Rectified Linear Units (ReLU) as the activation function except for the last layer. The MC-CNN-acrt network directly produced the matching costs for the next stereo matching algorithm step. The CNN network also paired with stereo methods to evaluate the matching costs. Following the previous work by Mei et al. (2011) and Zbontar and LeCun (2015) implemented cross based cost aggregation (CBCA) and the semi-global matching (SGM) to refine the matching cost. They performed minimization of energy function using dynamic programming, then the disparity image computed by using WTA optimization with interpolation and subpixel enhancement process. Finally, the disparity output has been applied with the 5×5 median filter followed by a bilateral filter after an enlargement process to match the original input size. They evaluated their method with the KITTI stereo dataset (Geiger et al., 2013). Their method was ranked as the best with the least error rate (2.61%) compare to other methods. They concluded that increasing the amount of the training set will help to achieve better performance.

Concerning their previous work, Zbontar and LeCun (2016) proposed two architectures: Fast (MC-CNN-fst), and Accurate (MC-CNN-acrt) architecture a. They elaborated further evaluations made on KITTI 2012, KITTI 2015, and Middlebury dataset and pro-

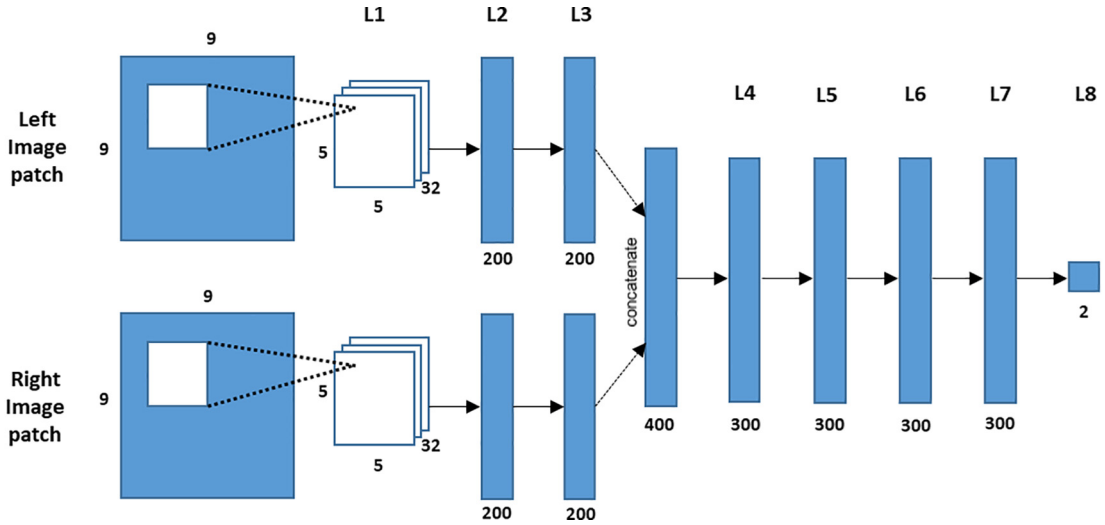

Fig. 5. MC-CNN-acrt network.

Name	Kernel	Str.	Ch I/O	InpRes	OutRes	Input
conv1	7×7	2	6/64	768×384	384×192	images
conv2	5×5	2	64/128	384×192	192×96	conv1
conv3a	5×5	2	128/256	192×96	96×48	conv2
conv3b	3×3	1	256/256	96×48	96×48	conv3a
conv4a	3×3	2	256/512	96×48	48×24	conv3b
conv4b	3×3	1	512/512	48×24	48×24	conv4a
conv5a	3×3	2	512/512	48×24	24×12	conv4b
conv5b	3×3	1	512/512	24×12	24×12	conv5a
conv6a	3×3	2	512/1024	24×12	12×6	conv5b
conv6b	3×3	1	1024/1024	12×6	12×6	conv6a
pr6+loss6	3×3	1	1024/1	12×6	12×6	conv6b
upconv5	4×4	2	1024/512	12×6	24×12	conv6b
iconv5	3×3	1	1025/512	24×12	24×12	upconv5+pr6+conv5b
pr5+loss5	3×3	1	512/1	24×12	24×12	iconv5
upconv4	4×4	2	512/256	24×12	48×24	iconv5
iconv4	3×3	1	769/256	48×24	48×24	upconv4+pr5+conv4b
pr4+loss4	3×3	1	256/1	48×24	48×24	iconv4
upconv3	4×4	2	256/128	48×24	96×48	iconv4
iconv3	3×3	1	385/128	96×48	96×48	upconv3+pr4+conv3b
pr3+loss3	3×3	1	128/1	96×48	96×48	iconv3
upconv2	4×4	2	128/64	96×48	192×96	iconv3
iconv2	3×3	1	193/64	192×96	192×96	upconv2+pr3+conv2
pr2+loss2	3×3	1	64/1	192×96	192×96	iconv2
upconv1	4×4	2	64/32	192×96	384×192	iconv2
iconv1	3×3	1	97/32	384×192	384×192	upconv1+pr2+conv1
pr1+loss1	3×3	1	32/1	384×192	384×192	iconv1

Fig. 6. DispNet Specification (Mayer et al., 2016).

ven that their MC-CNN-acrt architecture performs better than any other published method on all three datasets mentioned. In contrast, the result for their MC-CNN-fst architecture imposed some increase in error but with 90 times faster computation than MC-CNN-acrt. They also performed a comparison on their CNN method to compute matching costs with other handcraft methods such as SAD, CT, and NCC, and their architecture outperformed all the three methods (Zbontar and LeCun, 2016). The overall conclusion based on the work by Zbontar and LeCun (2016) CNN is well suited to perform stereo matching cost computation.

Chen and Yuan (2016) proposed another variation of CNN to perform matching cost computation. They mentioned due to maintaining equal weight left and right image in the convolutional layer, the relational information between the patches disappeared, which caused the result less accurate at texture-less regions. So they proposed a multi-scale CNN structure to calculate the stereo matching cost. In their comparison between the work (Zbontar and LeCun, 2016) and (Zagoruyko and Komodakis, 2015), they follow the later method (Zagoruyko and Komodakis, 2017) where they proposed to train corresponding patches in two-channel with flexible weight which lead to higher accuracy.

Mayer et al. (2016), with their proposed network, was the earlier contribution to end-to-end based networks for disparity estimation. The authors proposed 1D correlation to the approximate cost volume. They also presented a synthetic dataset with over 35,000 stereo frames. For their work on disparity estimation, they proposed DispNet. This network is based on encoder-decoder architecture. The first convolution layer on DispNet receives the image as input. As shown in Fig. 6, the main architecture contains contracting part (*conv1* to *conv6b*) and expanding part made of upconvolution (*upconvN*), convolution (*iconvN*, *prN*), and loss layers (*lossN*).

The disparity map predicted by *pr1*. They evaluated their DispNet on their synthetic dataset (FlyingThings3D, Sintel, and Moonka) and real-world datasets from KITTI 2012 (Geiger et al., 2012) and KITTI 2015 (Menze and Geiger, 2015) datasets. One of their network variant, DispNetCorr, although fall behind MC-CNN-acrt for KITTI 2015 results but it almost 1000 faster than MC-CNN-acrt. They perform better than SGM and MC-CNN for another dataset. As they fine-tuned the network based on the KITTI dataset, the network tends to produce a larger error on another evaluated dataset. The network was unable to predict image with huge object displacement. This has become one of the demerits of their network. The 2D and 3D convolutional layers in their architecture followed by batch normalization and ReLU.

In Fig. 7, GC-Net, a new method developed by Kendall et al. (2017), where they implemented Siamese convolution for their 2D convolution with shared weights. They use the unary features from both images and construct the cost volume to compute the matching cost. They explained the reason for constructing the cost volume of 4D (height, width, disparity, and feature size) enables them to maintain the geometry information of the stereo vision. The authors produced the 4D cost volume by concatenating the unary feature from both the left and right images. They claimed that the method to concatenate the features to get the cost volume has better performance than the method of subtracting features or

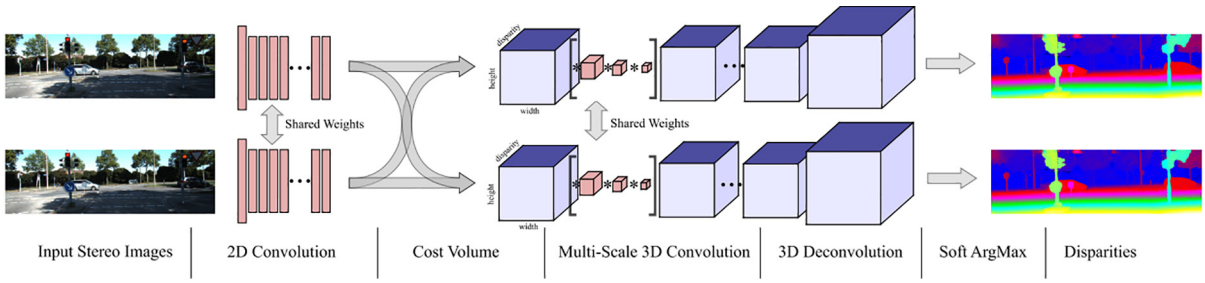

Fig. 7. Geometry and Context Network Architecture (GC-Net).

distance metric method. They implemented a 3D convolutional process to acquire the feature representations based on height, width, and disparity axis. They did mention the demerit of using 3D convolutional operation, which will incur more computational time in the overall process.

Yang et al. (2018a) introduce semantic information for their stereo matching works. They proposed their network, SegStereo. They also introduce semantic softmax loss for better accuracy on the disparity map generated. Semantic cues have been incorporated in SegStereo. Their network also integrates ResNet-50 by He et al. (2016) for the encoder part. Their decoder part performs disparity regression through several deconvolutional blocks to produce a full-size disparity map. They also embed segmentation subnetwork to extract semantic features using PSPNet-50 by Zhao et al. (2017). They achieve the state-of-the-art result on KITTI stereo benchmark dataset.

Kang et al. (2019a) extend their work by adding dilated convolution to their end-to-end network. According to Kang et al. (2019a), dilated convolution (also called as atrous convolution) enables them to utilize multi-scale context information. The authors also highlighted that the dilated convolution enhanced the receptive field to make their network more robust in low texture areas. Kang et al. (2019a) also mentioned that the dilated convolution better in terms of computational cost compared to the 3D CNN method used in Chang and Chen (2018) due to the 3D convolution computational cost. The result of the KITTI dataset demonstrates the improvement over original DispNet implementation.

Nguyen and Jeon (2019) also proposed another end-to-end network. According to them, the ill-posed region remains a challenging area for the matching algorithm to solve. Based on that, they want to overcome the disadvantages of the previous end-to-end networks, which unable to leverage the maximum capabilities of CNN. They also highlighted that the other networks also lack on exploiting the capabilities to maximize wide context information utilization. They want the network to learn a wide context using large receptive fields. They implemented Spatial Pyramid Pooling (SPP) and dilated convolution for their wide context learning network, which also used by other researchers (Chang and Chen, 2018; He et al., 2015; Kang et al., 2019b; Yang et al., 2018b). The stacked encoder and decoder network contain spatial diffusion, which responsible for refining the matching costs volume. Then the matching costs will be regressed to produce a disparity map.

Swami et al. (2019) presented their DISCO network. The author highlighted two issues related to another end-to-end based network where the loss of spatial information due to excessive downsampling. Another problem is the smaller receptive field of the network with low spatial resolution. They proposed to use dense blocks to preserve low-level spatial information. They used varying dilation rates on the dense blocks to enhance the effective receptive fields to capture contextual information. Their implementation on varying dilation on the dense block also similar to the method introduced in Yang et al. (2018b) but for the different problem which related to semantic segmentation. They also proposed a new matching cost computation module denoted as Local and Global Context Fusion (LGCF) module. Their network consists of three following subnetworks: Feature Extraction; Disparity Estimation; Disparity Refinement. Besides using a standard stereo benchmarking system such as the KIITI and Middlebury dataset, the authors also proposed their own dataset captured using a smartphone containing approximately 10,000 stereo images and ground truth images.

Song et al. (2019) also highlight the end-to-end network still suffer from the difficulties to overcome related to the ill-posed region and the accuracy of the disparity map of the near boundary region. They proposed EdgeStereo, their end-to-end network. The EdgeStereo contains a disparity network that contains the context pyramid and residual pyramid. Context information obtained using the context pyramid contains the relationship of an object and its surroundings. It is very useful in stereo correspondence. Instead of using stacking convolutional blocks like other methods, the author used the context pyramid to grab the information. The authors claimed that the residual pyramid used in the network simplified the cascaded refinement process. They also implemented an edge detection sub-network inside EdgeStereo. The edge sub-network produces an edge map that guides the disparity through their edge-aware smoothness loss for residual learning. They achieved the state-of-the-art results on KITTI stereo benchmarks.

Yang et al. (2019a) proposed another end-to-end network, HSMNet. Their implementation of the network is based on the encoder-decoder architecture and the stereo matching performed in the coarse-to-fine hierarchy. While extracting multi-scale features, the network decreases the high-resolution input into a lower resolution. The pyramid feature performs the matching in a hierarchical manner. It contains residual blocks and SPP layers with a similar purpose to increase receptive fields. They reported that their network was able to perform on-demand computation in real-time. Where their network can produce a rough estimation of large disparity objects before the pipeline complete, they achieve a better result on Middlebury and KITTI dataset as compared to other end-to-end networks (Chang and Chen, 2018; Kendall et al., 2017; Song et al., 2019).

2.2. Overall comparison

As we can see previously, the pioneer for the CNN network was developed, MC-CNN (Zbontar and LeCun, 2016) and DispNet (Mayer et al., 2016). MC-CNN-acrt which combine CNN and handcraft post-processing method performs better than DispNet in term

of accuracy in KITTI dataset. However, the demerit has been shadowed by DispNet speed, where it executes more than 1000 times faster than MC-CNN-acrt. For MC-CNN-acrt, the execution time also affected by handcraft processing time. The accuracy of the MC-CNN mostly driven by the handcraft post-processing method. For example the, in KITTI 2012 benchmark, they reported that the validation error would increase from 2.61% to 4.26% when they remove the SGM method. As discussed earlier, the demerit of DispNet is dependent on the dataset used. As they fine-tune their network towards a particular dataset, the performance for another dataset will degrade. So both methods also have their own advantages and disadvantages for further consideration. The following Table 1 illustrates the results for KITTI 2015 benchmark results for comparisons between the pioneer CNN networks and latest CNN networks.

In comparison between mixed method CNN and other end to end based networks, GC-Net and MC-CNN, GC-Net outperform the MC-CNN in terms of accuracy and speed for KITTI 2015 on all pixel and non-occluded pixels results. GC-Net also beats DispNet for accuracy, but in terms of execution speed, the DispNet is roughly 15 times faster than GC-Net on a similar benchmark. This is due to the 3D CNN computational method used in GC-Net. Same situation when comparing with DispNet. For example, PSMNet outperforms DispNet for accuracy for the *D1-all* error for KITTI 2015 benchmark. But similarly, PSMNet also lost to DispNet for speed. However, PSMNet outperforms GC-Net for both terms. Based on the architecture comparison between GC-Net and PSMNet, the spatial pyramid pooling (SPP) module and the way they implement stacked hourglass 3D CNN for cost volume aggregation enhanced the result for PSMNet as compared to GC-Net. Another end-to-end network, GA-Net, implemented stacked hourglass architecture for feature extraction in contrast to PSMNet. GA-Net aggregate the cost volume using their proposed Semi-Global Guided Aggregation (SGA) and Local Guided Aggregation (LGA) module. Whereas in PSMNet, the stacked hourglass CNN has been used for cost aggregation. In comparison for cost aggregation steps between GA-Net and PSMNet, the GA-Net perform better in term of speed and accuracy. Both GA-Net and PSMNet implement similar disparity regression applied in GC-Net. GA-Net outperformed PSMNet in terms of accuracy for the KITTI 2015 benchmark for overall performance. However, in terms of speed, GA-Net still cannot beat the execution speed shown by PSMNet.

The following Table 2 summarizes the comparison of the previous works, which mixed the CNN based approach with the traditional handcraft algorithm for the standard stereo pipeline. The work by Zbontar and LeCun (2015) inspired many other researchers on the implementation of the CNN based method for matching cost computation. Most of the mixed methods also using WTA to compute the disparity image together with several handcraft methods to refine the disparity map such as LRC to remove the rid of outlier pixels in the final disparity map. SGM method also has been used by several authors in their work for better accuracy and low computational cost. However, tuning the SGM penalty parameters quite difficult. Seki and Pollefeys (2017) try to improve the problem through their learning-based penalties estimation method, SGM-Nets. The SGM-Net results show better than hand tune SGM method (Zbontar and LeCun, 2015). This shows that the implementation of a learning-based method capable of improving the overall performance.

3. Conclusion and recommendation

In this article, several published methods have been introduced and discussed, based on the implementation of deep learning to solve the stereo matching problem. It has been highlighted most of the researchers implemented CNN in their methods to solve the problem. Basically, the work by Zbontar and LeCun (2016) inspired many other researchers to further improve the patch-based learning method for calculating matching costs. The mixture of machine learning elements such as CNN and handcraft algorithm provide variety to the stereo matching solutions. It is proven that the simple implementation of CNN could boost the performance of the algorithm. Based on the works on CNN documented in this paper, we can see that some of the researchers implemented CNN for different stages in the stereo algorithm. Some of the researchers focused on the specific subnetwork to solve certain stage in stereo algorithm steps. Some of the researchers also utilize CNN for enhancing handcraft optimization using neural networks approach such as SGM-Net. There is also another approach on CNN for the stereo algorithm pioneered by Mayer et al. (2016) and Kendall et al. (2017). The authors implemented the end-to-end network for disparity regression using CNN. Both ends to end networks, DispNet and GC-Net, also inspired several other researchers in later works on generation of the stereo based disparity map. However, this does not make the end-to-end based network superior compared to other mixed methods. As mentioned in other articles (Seki and Pollefeys, 2017; Song et al., 2019), the problem of the end-to-end network the accuracy achieved still not enough. There is still some more room for mixed and end-to-end based networks to grow. The variety of approaches on machine learning becomes more interesting with other types of networks such as Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), which could further improve the stereo matching algorithm for better disparity maps and 3D output generation. This also demonstrates the power of new machine learning tools supported by enhancement of the cutting-edge computing hardware.

Table 1
The comparison methods from the KITTI benchmarking evaluation system which these methods are based on the deep learning technique.

Models	Non Occluded			All			Runtime (sec)
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
MC-CNN (Zbontar and LeCun, 2016)	2.48	7.64	3.33	2.89	8.88	3.89	67
DispNet (Mayer et al., 2016)	4.11	3.72	4.05	4.32	4.41	4.34	0.06
GC-Net (Kendall et al., 2017)	2.02	5.58	2.61	2.21	6.16	2.87	0.90
SGM-Nets (Seki and Pollefeys, 2017)	2.23	7.44	3.09	2.66	8.64	3.66	67
PSMNet (Chang and Chen, 2018)	1.71	4.31	2.14	1.86	4.62	2.32	0.41
SegStereo (Yang et al., 2018a)	1.76	3.70	2.08	1.88	4.07	2.25	0.60
GA-Net-15 (Zhang et al., 2019)	1.40	3.37	1.73	1.55	3.82	1.93	1.80
HSMNet (Yang et al., 2019a)	1.63	3.40	1.92	1.80	3.85	2.14	0.15
EdgeStereo (Song et al., 2019)	1.69	2.94	1.89	1.87	3.61	2.16	0.70

Table 2
Summary of framework comparison on mixed CNN-based stereo matching algorithms.

Year	Author	Matching Cost Computation	Cost Aggregation	Disparity Computation/Optimization	Disparity Refinement
2015	Zbontar & LeCun	Cost computed directly from CNN network MCCNN-acrt and MC-CNN-fst Binary classification for matching cost	Cross based Cost Aggregation	SGM (Dynamic Programming) + WTA	LR cross consistency check (LRC)
2015	Chen et al.	CNN based with multiscale deep embedding model. Compute similarity in Euclidean space faster than MCCNN.	–	SGM	Left Right Check
2016	Zbontar & LeCun	Cost computed directly from CNN output	Cross based Cost Aggregation	SGM + WTA + Interpolation + Subpixel Enhancement	Median Filter and Bilateral Filter
2016	Chen & Yuan	Multi scale CNN to compute matching cost. L and R image patches downsampled and sent to Layer 1 of Convolutional Sub network and this will produce vectors result from different scale (from different conv subnetwork) Layer 8 output the initial matching cost	Multiscale Cross-Based Cost Aggregation	WTA + Scanline Optimization	LRC + subpixel enhancement
2016	Wang et al.	CNN + CRF	–	–	LRC Check + Four-direction propagation + Gradient Domain guided image filter
2016	Luo et al.	Faster CNN network for computing local matching costs as a multi-label classification of disparities using a Siamese network	Average Pooling	SGM	Slanted Plane + Interpolation
2016	Seki & Pollefeys	Based on MC-CNN-acrt and MC-CNN-fst		Similar to Zbontar & LeCun with Correspondence confidence fused with SGM	Similar to Zbontar & LeCun median filter + modified bilateral filter.
2017	Yang et al.	CNN - Matching cost directly based on L2 distance in Euclidean space + SGM (dynamic programming)	–	Multi-scale Segmentation SGM + WTA	LRC Check + Median filter
2017	Shaked & Wolf	CNN based with constant highway residual block (outer and inner-λ residual) and skip connection block.	Cross based Cost Aggregation	SGM + Global Disparity Network (CNN based)	LRC Check and Interpolation + Subpixel Enhancement + Median and Bilateral Filter
2017	Seki & Pollefeys	Author try two methods MC-CNN and ZNCC	–	SGM-Net – learning penalty parameter to predict penalty input for SGM algorithm + WTA	–
2017	Joung et al.	MC-CNN-fast + multiscale cost computation	Cross based Cost Aggregation + SGM	WTA + Interpolation + Subpixel Enhancement	Median Filter and Bilateral Filter
2017	Wen	CNN based using binary cross-entropy loss during training process	Guided filter based on ARSW	WTA	Region voting + Median Filter
2018	Yang and Lv	CNN - Matching cost directly based on L2 distance in Euclidean space + SGM (dynamic programming)	Fast cost aggregation using orthogonal integral image (OII)	Generalized SGM + WTA	LRC Check + Median filter
2018	Liang et al.	Initial Disparity Estimation Subnetwork of iResNet through correlation layer	Done in Initial Disparity Estimation Subnetwork of iResNet by concatenating left image features with the matching cost	Done in Initial Disparity Estimation Subnetwork of iResNet	Disparity Refinement Subnetwork of iResNet
2019	Song et al.	MC-CNN + DD-CNN (disparity discontinuous) classify DD region	–	SGM	–
2019	Nguyen and Jeon	MC-CNN-fst + Census Transform	Cost Volume Unary Network + Disparity Boundaries Pairwise Network + Weighted least squares (WLS) optimization framework	SGM + Interpolation + Subpixel enhancement	–
2019	Brandao et al.	Faster CNN network for computing local matching costs as a multi-label classification of disparities using a Siamese network	Average Pooling	SGM	Slanted Plane + Interpolation
2019	Xue	Same as Zbontar (Change the activation function and add batch normalization)	CBCA	SGM + WTA + Subpixel enhancement	Median Filter and Bilateral Filter
2019	Yang et al.	CNN based method (Feature Extraction + Volume Creation + Similarity Computation)	2D and 3D CNN (Aggregation Proposal + Aggregation Guidance)	Soft Argmin Function similar to Kendall et al. (2017)	–
2019	Fu et al.	Similar to Wen but with implementation of Atrous CNN to enhance receptive fields	Average Pooling + SGM	WTA	Interpolation + Median Filter
2019	Brandao et al.	Faster CNN network for computing local matching costs as a multi-label classification of disparities using a Siamese network	Average Pooling	SGM	Slanted Plane + Interpolation

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work supported by the Ministry of Higher Education (MOHE), Malaysia, and sponsored by Universiti Teknikal Malaysia Melaka with a grant number JURNAL/2018/FTK/Q00008.

References

Bernard, M., 2018. What is Industry 4.0? Here's A Super Easy Explanation For Anyone [WWW Document]. URL https://www.forbes.com/sites/bernardmarr/2018/09/02/what-is-industry-4-0-heres-a-super-easy-explanation-for-anyone/#4600feae9788 (accessed 2.11.19).

Cambuim, L.F.S., Barbosa, J.P.F., Barros, E.N.S., 2017. Hardware module for low-resource and real-time stereo vision engine using semi-global matching approach. In: Proc. 30th Symp. Integr. Circuits Syst. Des. Chip Sands – SBCCI '17, pp. 53–58. https://doi.org/10.1145/3109984.3109992.

Chang, J.-R., Chen, Y.-S., 2018. Pyramid stereo matching network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Salt Lake City Utah. https://doi.org/10.1109/CVPR.2017.730.

Chauhan, M.S., Singh, A., Khemka, M., Prateek, A., Sen, R., 2019. Embedded CNN based vehicle classification and counting in non-laned road traffic. In: Proceedings of the Tenth International Conference on Information and Communication Technologies and Development. ACM New York, NY, USA ©2019, p. Article No 5. https://doi.org/10.1145/3287098.3287118.

Chen, H., Wang, K., Yang, K., 2018. Improving RealSense by Fusing Color Stereo Vision and Infrared Stereo Vision for the Visually Impaired. In: Proceedings of the 2018 International Conference on Information Science and System. ACM New York, NY, USA ©2018, Jeju, Republic of Korea, pp. 142–146. https://doi.org/https://doi.org/10.1145/3209914.3209944.

Chen, J., Yuan, C., 2016. Convolutional neural network using multi-scale information for stereo matching cost computation. In: Proc. – Int. Conf. Image Process. ICIP 2016-Augus, 3424–3428. https://doi.org/10.1109/ICIP.2016.7532995.

Chollet, F., 2017. Deep Learning with Python. Manning Publications Co., Greenwich, CT, USA.

Ciresan, D., Meier, U., Schmidhuber, J., Cires, D., Meier, U., 2012. Multi-column Deep Neural Networks for Image Classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On. pp. 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110.

Damjanović, S., van der Heijden, F., Spreeuwers, L.J., Heijden Van Der, F., Group, S., 2012. Local stereo matching using adaptive local segmentation. ISRN Mach. Vis. 2012, 1–11. https://doi.org/10.5402/2012/163285.

Fangmin, L., Ke, C., Xinhua, L., 2017. 3D Face Reconstruction Based on Convolutional Neural Network. Proc. – 10th Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2017 2017-Octob, 71–74. https://doi.org/10.1109/ICICTA.2017.23.

Flynn, J., Neulander, I., Philbin, J., Snavely, N., 2016. DeepStereo : Learning to Predict New Views from the World's Imagery. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016). IEEE, Las Vegas, NV, USA, pp. 5515–5524. https://doi.org/10.1109/CVPR.2016.595.

Fu, J., Liang, J., 2019. Virtual view generation based on 3D-dense-attentive GAN networks. Sensors (Switzerland) 19. https://doi.org/10.3390/s19020344.

Fu, K., Xie, Y., Jing, H., Zhu, J., 2019. Fast spatial–temporal stereo matching for 3D face reconstruction under speckle pattern projection. Image Vis. Comput. 85, 36–45. https://doi.org/10.1016/j.imavis.2019.02.007.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: the KITTI dataset. Int. J. Rob. Res. 32, 1231–1237. https://doi.org/10.1177/0278364913491297.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074.

Gibson, A., Patterson, J., 2016. Deep Learning: A Practitioner's Approach. O'Reilly Media Inc..

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. https://doi.org/https://doi.org/10.1007/s10710-017-9314-z.

Hamzah, R.A., Ibrahim, H., 2016. Literature survey on stereo vision disparity map algorithms. J. Sensors 2016. https://doi.org/10.1155/2016/8742920.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.90.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2015.2389824.

Hodges, C., Bennamoun, M., Rahmani, H., 2019. Single image dehazing using deep neural networks. Pattern Recognit. Lett. 128, 70–77. https://doi.org/10.1016/j.patrec.2019.08.013.

Hong, G.S., Kim, B.G., 2017. A local stereo matching algorithm based on weighted guided image filtering for improving the generation of depth range images. Displays 49, 80–87. https://doi.org/10.1016/j.displa.2017.07.006.

Huang, G., Liu, Z., v. d. Maaten, L., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Huang, X., Zhang, Y.J., 2016. An O(1) disparity refinement method for stereo matching. Pattern Recognit. 55, 198–206. https://doi.org/10.1016/j.patcog.2016.01.025.

Kang, J., Chen, L., Deng, F., Heipke, C., 2019a. Context pyramidal network for stereo matching regularized by disparity gradients. ISPRS J. Photogramm. Remote Sens. 157, 201–215. https://doi.org/10.1016/j.isprsjprs.2019.09.012.

Kang, J., Chen, L., Deng, F., Heipke, C., 2019b. Encoder-Decoder network for local structure preserving stereo matching. In: Dreiländertagung Der DGPF, Der OVG Und Der SGPF in Wien, Österreich – Publikationen Der DGPF, Band 28, 2019. Vienna, Austria.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end Learning of Geometry and Context for Deep Stereo Regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75. https://doi.org/10.1109/ICCV.2017.17.

Ketkar, N., 2017. Deep Learning with Python, Deep Learning with Python : A Hands-on Introduction. Apress. https://doi.org/10.1007/978-1-4842-2766-4.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. pp. 1097–1105. https://doi.org/10.1145/3065386.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1, 541–551. https://doi.org/10.1162/neco.1989.1.4.541.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791.

Li, Y., Huang, J.-B., Narendra, A., Yang, M.-H., 2016. Deep Joint Image Filtering. In: European Conference on Computer Vision. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-46493-0_10.

Ma, X., Wang, S., Liu, W., Ma, F., Wang, A., Sheng, Y., Li, Y., Ming, H., 2019. Optimized stereo matching algorithm for integral imaging microscopy and its potential use in precise 3-D optical manipulation. Opt. Commun. 430, 374–379. https://doi.org/10.1016/j.optcom.2018.08.064.

Malekabadi, A.J., Khojastehpour, M., Emadi, B., 2019a. Disparity map computation of tree using stereo vision system and effects of canopy shapes and foliage density. Comput. Electron. Agric. 156, 627–644. https://doi.org/10.1016/J.COMPAG.2018.12.022.

Malekabadi, A.J., Khojastehpour, M., Emadi, B., 2019b. Comparison of block-based stereo and semi-global algorithm and effects of pre-processing and imaging parameters on tree disparity map. Sci. Hortic. (Amsterdam) 247, 264–274. https://doi.org/10.1016/j.scienta.2018.12.033.

Malita, M., Nedescu, O., Negoita, A., Stefan, G.M., 2018. Deep learning in low-power stereo vision accelerator for automotive. 2018 IEEE Int. Conf. Consum. Electron. ICCE 2018 2018-Janua. https://doi.org/10.1109/ICCE.2018.8326285.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation nikolaus. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 4040–4048. https://doi.org/10.1109/CVPR.2016.438.

Mei, X., Sun, X., Zhou, M., Jiao, S., Wang, H., Zhang, X., 2011. On building an accurate stereo matching system on graphics hardware. In: Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCVW.2011.6130280.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3061–3070. https://doi.org/10.1109/CVPR.2015.7298925.

Nguyen, T.P., Jeon, J.W., 2019. Wide context learning network for stereo matching. Signal Process. Image Commun. 78, 263–273. https://doi.org/10.1016/j.image.2019.07.008.

Nityananda, V., Tarawneh, G., Henriksen, S., Umeton, D., Simmons, A., Read, J.C.A., 2018. A novel form of stereo vision in the praying mantis. Curr. Biol. 28, 588–593.e4. https://doi.org/10.1016/j.cub.2018.01.012.

Pang, J., Sun, W., Ren, J.S.J., Yang, C., Yan, Q., 2018. Cascade Residual Learning: a Two-Stage Convolutional Neural Network for Stereo Matching. In: Proceedings – 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017. pp. 878–886. https://doi.org/10.1109/ICCVW.2017.108.

Popovi, G., Hadviger, A., Markovi, I., Petrovi, I., 2018. Computationally efficient dense moving object detection based on reduced space disparity estimation. Int. Feder. Autom. Control, 360–365. https://doi.org/10.1016/j.ifacol.2018.11.568.

Salehian, B., Fotouhi, A.M., Raie, A.A., 2018. Dynamic programming-based dense stereo matching improvement using an efficient search space reduction technique. Optik (Stuttg). 160, 1–12. https://doi.org/10.1016/j.ijleo.2018.01.021.

Sangeetha, G.R., Kumar, N., Hari, P.R., Sasikumar, S., 2018. Implementation of a Stereo vision based system for visual feedback control of Robotic Arm for space manipulations. Proc. Comput. Sci. 133, 1066–1073. https://doi.org/10.1016/j.procs.2018.07.031.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo. Int. J. Comput. Vis. 47, 7–42. https://doi.org/10.1109/SMBV.2001.988771.

Seki, A., Pollefeys, M., 2017. SGM-Nets: Semi-global matching with neural networks. In: Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 6640–6649. https://doi.org/10.1109/CVPR.2017.703.

Smith, M., Smith, L., 2018. Special issue on: Machine vision for outdoor environments. Comput. Ind. 100, 224–226. https://doi.org/10.1016/j.compind.2018.04.016.

Smolyanskiy, N., Kamenev, A., Birchfield, S., 2018. On the importance of stereo for accurate depth estimation: an efficient semi-supervised deep neural network approach. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. 2018-June, 1120–1128. https://doi.org/10.1109/CVPRW.2018.00147.

Song, X., Zhao, X., Hu, H., Fang, L., 2019. EdgeStereo: a context integrated residual pyramid network for stereo matching. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11365 LNCS, 20–35. https://doi.org/10.1007/978-3-030-20873-8_2.

Swami, K., Raghavan, K., Pelluri, N., Sarkar, R., Bajpai, P., 2019. DISCO: Depth Inference from Stereo using Context. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE. pp. 502–507. https://doi.org/10.1109/ICME.2019.00093.

Szeliski, R., 2011. Computer Vision : Algorithms and Applications, Texts in Computer Science. Springer London, London. https://doi.org/10.1007/978-1-84882-935-0.

Tonioni, A., Poggi, M., Mattoccia, S., Stefano, L. Di, 2017. Unsupervised Adaptation for Deep Stereo. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 1614–1622. https://doi.org/10.1109/ICCV.2017.178.

Vu, H., Kim, H.C., Lee, J.H., 2018. 3D convolutional neural network for feature extraction and classification of fMRI volumes. 2018 Int. Work. Pattern Recognit. Neuroimaging, PRNI 2018 1–4. https://doi.org/10.1109/PRNI.2018.8423964.

Wang, Z., Zhu, S., Li, Y., Cui, Z., 2016. Convolutional neural network based deep conditional random fields for stereo matching. J. Vis. Commun. Image Represent. 40, 739–750. https://doi.org/10.1016/j.jvcir.2016.08.022.

Williem, Park, I.K., 2018. Cost aggregation benchmark for light field depth estimation. J. Vis. Commun. Image Represent. 56, 38–51. https://doi.org/10.1016/j.jvcir.2018.08.015.

Xu, Y., Zhao, Y., Ji, M., 2014. Local stereo matching with adaptive shape support window based cost aggregation. Appl. Opt. 53, 6885. https://doi.org/10.1364/ao.53.006885.

Xue, T., Owens, A., Scharstein, D., Goesele, M., Szeliski, R., 2019. Multi-frame stereo matching with edges, planes, and superpixels. Image Vis. Comput. https://doi.org/10.1016/j.imavis.2019.05.006.

Yang, G., Manela, J., Happold, M., Ramanan, D., 2019. Hierarchical Deep Stereo Matching on High-Resolution Images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA. pp. 5510–5519. https://doi.org/10.1109/CVPR.2019.00566.

Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J., 2018. SegStereo: Exploiting Semantic Information for Disparity Estimation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), European Conference on Computer Vision(ECCV). Springer International Publishing, Cham. pp. 660–676. https://doi.org/10.1007/978-3-030-01234-2_39.

Yang, J., Zhao, Y., Zhu, Y., Xu, H., Lu, W., Meng, Q., 2019b. Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network. Inf. Sci. (Ny) 474, 1–17. https://doi.org/10.1016/j.ins.2018.08.066.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. DenseASPP for semantic segmentation in street scenes. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 3684–3692. https://doi.org/10.1109/CVPR.2018.00388.

Zagoruyko, S., Komodakis, N., 2017. Deep compare: a study on using convolutional neural networks to compare image patches. Comput. Vis. Image Underst. 164, 38–55. https://doi.org/10.1016/j.cviu.2017.10.007.

Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 4353–4361. https://doi.org/10.1109/CVPR.2015.7299064.

Zeglazi, O., Rziza, M., Amine, A., Demonceaux, C., 2018. A hierarchical stereo matching algorithm based on adaptive support region aggregation method. Pattern Recognit. Lett. 112, 205–211. https://doi.org/10.1016/j.patrec.2018.07.020.

Zbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. 17, 1–32. https://doi.org/10.1186/s13568-015-0106-7.

Zbontar, J., LeCun, Y., 2015. Computing the Stereo Matching Cost with a Convolutional Neural Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015). pp. 1592–1599. https://doi.org/10.1109/CVPR.2015.7298767.

Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.S., 2019. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.660.

Zhu, S., Wang, Z., Zhang, X., Li, Y., 2016. Edge-preserving guided filtering based cost aggregation for stereo matching. J. Vis. Commun. Image Represent. 39, 107–119. https://doi.org/10.1016/j.jvcir.2016.05.012.