Stereo Matching Based on Features of Image Patch

SHI ZHOU ^{*†}	Member,	HE LI [‡]	Non-member
MIAOMIAO ZHU [†]	Member,	ZHEN LI [†]	Member
MITSUNORI MIZUMACHI [†]	Member,	LIFENG ZHANG [†]	Member

(Received April 5, 2021, revised July 7, 2021)

Abstract: Stereo matching is a branch of 3D vision and has a wide range of applications in 3D reconstruction and autonomous driving. Recently, stereo matching methods leverage all the information of the stereo image to calculate a disparity map. However, these methods still have difficulties in texture-less areas and occlusion areas, and post-processing need to do to improve the accuracy. Therefore, there is a high computational cost in feature extraction and post-processing. This paper proposes a stereo matching method that predict the disparity of non-occlusion areas with corresponding features, instead of predicting the disparity of the full image with all features. And aggregation methods are performed to modify all kinds of mismatching pixels based on the correct disparity in the non-occlusion areas. Furthermore, we evaluated the proposed method on the Middlebury dataset. The result shows that the proposed method performs well in all areas.

Keywords: Stereo matching, Disparity prediction, Image patch, Convolutional neural networks, 3D vision,

1. Introduction

Stereo matching [1] is a fundamental problem in 3D computer vision, and its purpose is to predict the disparity map for stereo images. Usually, the stereo image is two images of the same scene, which taken from cameras at different horizontal positions but at the same vertical position. And the disparity indicates the distance in the horizontal direction. In this way, according to the parallel model of binocular vision, the 3D coordinates of objects in the scene can be calculated with the disparity map [2]. Therefore, it is widely used in object surface reconstruction, scene reconstruction, *etc.*

As a popular research topic for decades, traditional stereo matching methods comprise four steps: calculation of matching cost, aggregation of matching cost, optimization, and disparity refinement [3]. These methods usually calculate the similarity between pixels by designing an energy function. If the two pixels are very similar, the matching cost between the two pixels is low, and if the two pixels are very different, the corresponding matching cost is high. Therefore, calculating the matching cost with high accuracy is the main problem of the stereo matching method. However, the energy function is sensitive to different lighting conditions. Even for different applications, different energy functions should be designed. Therefore, traditional stereo matching methods have poor robustness to different envi-

ronments.

Recently, convolutional neural networks (CNN) have performed well in image feature extraction, and many stereo matching methods based on CNN have been proposed [4] [5]. These methods use end-to-end disparity estimation networks to integrate all steps in stereo matching. However, these methods still have difficulties in ill-posed regions, such as occlusion areas and texture-less areas. Post-processing is necessary to obtain a high accuracy disparity map. Also, the end-to-end network usually has a complex network structure, which leading to high computational cost.

In this work, we proposed a novel stereo matching method based on features extracted from image patches to achieve low computational cost disparity prediction. These features are extracted from the neural network, which can describe the image patch from different light condition well, so that the proposed method has strong robustness. According to these features, the similarity between image patches can be calculated. The matching cost is the opposite of similarity. Then, the local aggregation method will be employed to modify the scattered mismatching pixels that look like noise. And for the texture-less areas, where many mismatching pixels appeared, the semi-global method will be performed to correct these mismatching. In addition, for occlusion areas, where the object can be only observed in one of the two stereo image, the left-right consistency check method will be performed, and the disparity of background areas will be used to replace the disparity of occlusion area. After extracting features from the neural network and performing post-processing for different mismatching pixels, the output disparity map of the proposed method has a good performance on the Middlebury dataset [6].

* Corresponding: zhou.shi403@mail.kyutech.jp

[†] Department of Electrical and Electronic Engineering, Kyushu Institute of Technology

1-1 Sensui, Tobata, Kitakyushu, Fukuoka, 804-0015, Japan

[‡] School of Mechanical Engineering and Automation, Northeastern University

3-11 Wenhua, Heping, Shenyang, Liaoning, 110819, China

Figure 1: Overall architecture of proposed stereo matching method.

2. Proposed method

In this section, we detail the proposed stereo matching method. The pipeline is shown in Fig. 1, which follows the same procedure as traditional stereo matching. Firstly, according to the correlation between stereo images, the matching cost can be calculated with the features extracted from CNN. Next, the matching cost is aggregated so that mismatching pixels can be corrected. Finally, we perform post-processing to obtain a disparity map with high accuracy. When training the neural network, the input is an image patch. However, when testing, an entire image can be used as input to obtain features that describe each pixel.

2.1 Matching cost calculation The traditional method calculates the matching cost based on a self-designed energy function, which cannot adapt to different lighting conditions [7][8]. This work calculates the matching cost based on features extracted from CNN, which is robust to different lighting conditions. The overall architecture of the proposed neural network is shown in Fig. 2. Firstly, the image patch is input to the first convolutional layer. Next, these features, which describe the image patch, will be extracted after two multi-scale blocks. About the multi-scale blocks, 4 kinds of features, which are extracted from the input, will be concatenated and then input to the following 3 convolutional layers. The input image patch size is 11×11 , and the output feature size is 1×64 .

When training the network, it is considered as a binary classification problem. For each image patch from the left image, a negative image patch and a positive image patch will be extracted from the right image. In this way, a negative stereo image patch sample and a positive stereo image patch sample will be created. When training, the neural network will learn to distinguish positive and negative samples, meanwhile, the well-described features will be extracted.

The similarity score C of the sample is computed directly from the features that output from the network. And matching cost P is the opposite of similarity score.

$$C = \sum (f_l * f_r) \quad (1)$$

$$P = -C \quad (2)$$

Where f_l denotes the feature of left image patch, and f_r denotes the feature of right image patch. The loss function is defined as,

$$L = \max(0, m - P_1 + P_0) \quad (3)$$

Figure 2: The architecture of our convolutional neural network.

Where P_0 denotes the matching cost of the negative sample, and P_1 denotes the matching cost of the positive sample. The margin m between the positive and negative samples is a hyper-parameter, which is initialized to 0.2. When training the network, if the matching cost of the negative sample is m higher than that of the positive sample, we think that the network can distinguish different samples. Therefore, the network pays more attention to the distinguishing of samples with a difference between 0.2.

2.2 Matching cost aggregation For each pixel, the network extracts features only based on the information of the image patch instead of the full image. Therefore, some mismatching pixels exist, especially in texture-less areas and repeat texture areas. The matching cost calculated in Sec. 2.1 cannot be used directly to predict disparity, and further modifications are necessary. Therefore, different matching cost aggregation methods are used to modify different kinds of mismatching pixels.

Scattered mismatching pixels are very similar to the image's noise, so they can be corrected using the image

Figure 3: Matching cost aggregation.

smoothing method. Therefore, an adaptive cross window is designed to modify scattered mismatching pixels, as shown in Fig. 3(a). For pixel $\mathbf{p}$, the control condition of the adaptive cross window is,

$$\begin{cases} |I(\mathbf{p}) - I(\mathbf{q})| < \tau \\ \|\mathbf{p} - \mathbf{q}\| < \eta \end{cases} \quad (4)$$

The distance between $\mathbf{p}$ and $\mathbf{q}$ should be less than τ . The difference between $\mathbf{p}$ and $\mathbf{q}$ in the input image should be less than η . Each pixel will get an adaptive cross window W , as shown in the right part of Fig. 3(a). The matching cost of each pixel will be aggregated based on the matching cost of all pixels in its adaptive cross windows.

$$C'(\mathbf{p}, d) = \frac{1}{n} \sum_{i \in W} C_i(\mathbf{p}, d) \quad (5)$$

In the texture-less areas, there are many pixels mismatched in the texture-less areas since the pixels are very similar to each other. These mismatched pixels cannot be modified base on the nearby pixels in its adaptive cross window. Therefore, A larger region-of-interesting around the target pixel should be under consideration. Thus the semi-global method is used for modification. For adjacent pixels, if the image intensity between they are very close, their disparity will be close too. Therefore, we refine the matching cost by enforcing smoothness constraints in texture-less areas. The principle of matching cost aggregation is shown in Fig. 3(b). In direction $\mathbf{r}$, the semi-global aggregation method of matching cost is,

$$\begin{aligned} L_r(\mathbf{p}, d) = & C(\mathbf{p}, d) + \min\{L_r(\mathbf{p} - \mathbf{r}, d), \\ & L_r(\mathbf{p} - \mathbf{r}, d - 1) + P_1, \\ & L_r(\mathbf{p} - \mathbf{r}, d + 1) + P_1, \\ & \min_i [L_r(\mathbf{p} - \mathbf{r}, i) + P_2]\} - \\ & \min_k [L_r(\mathbf{p} - \mathbf{r}, k)] \end{aligned} \quad (6)$$

Where C denotes the matching cost of pixel $\mathbf{p}$ before semi-global aggregation. When the difference between adjacent pixels is 1, a penalty P_1 will be added. When the difference between adjacent pixels is larger than 1, a penalty P_2

Figure 4: Pixels classification and disparity modification.

will be added. Moreover, the parameters P_1 and P_2 are set according to the gradient of the image. Finally, we refine the matching cost by combining the four directions of up, down, left, and right.

$$C'(\mathbf{p}, d) = \frac{1}{n} \sum_{\mathbf{r}} L_r(\mathbf{p}, d) \quad (7)$$

2.3 Disparity prediction and post-processing After the matching cost is aggregated, most of the mismatching pixels are modified. The distance in the horizontal direction obtains the most negligible matching cost is considered as the predicted disparity.

The occlusion area means a set of pixels can only be seen in one image but not be seen in the other. This phenomenon happened because objects located at close shade objects located at faraway. Therefore, the occlusion area is the background. The left-right consistency check method is performed to check whether the pixel located at occlusion area, and make a correction to pixels which located at occlusion area [9].

All the pixels are divided into four classes: matching, occlusion, mismatching, and no matching, as shown in Fig. 4. For each class, different method is considered to make a better modification.

matching: no operation

occlusion: replaced by background pixel

mismatching: modified based on surrounding pixels

nomatching: replaced by near pixels

After left-right consistency checking, median filtering and disparity refinement is performed to obtain a high accuracy disparity map.

3. Experiments

3.1 Datasets We evaluate the proposed method on the Middlebury dataset. Because the dense ground truth disparity is calculated based on the data from the grating, it is difficult to control the lighting condition [10]. Therefore, this dataset is very small. To obtain a large amount of data for training, the pixels that all pixels in its corresponding image patch has ground truth will be consider as a candidate. We predict the disparity of each pixel and calculate the error rate with different error tolerance.

Table 1: Comparison of error rate and RMS error during stereo matching.

processing method	all area					nonocc area				
	>0.5	>1.0	>2.0	>4.0	RMS	>0.5	>1.0	>2.0	>4.0	RMS
CNN	24.07	17.80	15.69	13.64	14.42	15.55	9.05	7.79	7.16	9.57
1st time aggregation	21.66	15.52	13.48	11.45	13.11	12.90	6.58	5.46	4.92	7.75
2nd time aggregation	18.23	12.88	10.58	8.28	9.89	9.63	4.40	3.28	2.66	3.92
3rd time aggregation	17.90	12.28	9.99	7.70	9.78	9.43	3.94	2.82	2.27	3.72
post-processing	16.09	10.36	7.49	4.73	4.40	8.89	3.70	2.58	1.94	1.88

3.2 Implementation details We train the proposed network by using gradient descent with the momentum method. When the neural network parameters continue to update in the same direction, the speed update is faster than the stochastic gradient descent method, which will speed up the network training. When the parameters update direction changes suddenly, the process can move forward for a while due to the enormous momentum. Therefore, some local best points and saddle points will be dropped out, and

the possibility of finding the global best point will increase. In addition, we train our network with an exponential decay learning rate so that it can find the global best point. The network is build based on TensorFlow and trained on a GPU with 6GB memory.

Three times of matching cost aggregation is performed during stereo matching. The first and third times are aggregate with adaptive windows, and the second time is aggregated based on semi-global method. The aggregation

Figure 5: Test on Teddy stereo images.

Figure 6: Test on some other stereo images taken by ourselves.

based on adaptive windows before and after semi-global approaches can avoid the influence of noise on it and remove the noise it brings. The calculation cost based on adaptive windows is small, so it only takes a little time to perform it twice.

3.3 Result According to the standard evaluation metrics of the Middlebury dataset, we evaluate the proposed method, as shown in Table 1. All areas include areas with ground truth (a few area which without ground truth since the occlusion of grating). Non-occlusion (nonocc) areas denote areas with ground truth and not occluded. With different tolerance errors, the disparity map error rate and RMS error decrease with stereo matching. When the tolerance error is 4 pixels, the error rate of the disparity map is 4.73. Comparing the error rate and RMS error after CNN in non-occlusion areas with all areas, the stereo matching based on image patch features can achieve high accuracy disparity prediction in the non-occlusion areas. And almost all mismatching in non-occlusion area can be corrected after aggregation and post-processing. The disparity map output from the proposed method is shown in Fig. 5. According to the error mask of the edge area of the bear, there are only a few mismatching pixels, and the proposed method can achieve high accuracy disparity prediction. Furthermore, we test the proposed method on some other stereo images taken by ourselves, as shown in Fig. 6. The proposed method has strong robustness to different environments and is well performed in the texture-less area and repeat texture area.

4. Conclusions

This paper proposed a stereo matching method based on features of the image patch. The result on the Middlebury dataset suggests that stereo matching based on features extracted from the image patch can perform well and have a high accuracy in the non-occlusion area. Employing the correct disparity in the non-occlusion area, the occlusion area's disparity can be modified with an adaptive cross window and semi-global method. However, our method is not yet suitable for real-time applications, such as autonomous driving. Future work can focus on improving the post-processing method to achieve real-time application.

References

- [1] H. Kien, T. Pan, Z. Wang, H. Nguyen and M. Vo, "Accurate 3D shape measurement of multiple separate objects with stereo vision", *Measurement Science and Technology*, Vol.25, No.3, pp.1-7, 2014. DOI: 10.1088/0957-0233/25/3/035401
- [2] H. Yu, T. Xing and X. Jia, "The analysis of measurement accuracy of the parallel binocular stereo vision system", *8th International Symposium on Advanced Optical Manufacturing and Testing Technologies*, 2016. DOI: 10.1117/12.2243156
- [3] S. Degadwala, D. Vyas and A. Mahajan, "Review on Stereo Vision Based Depth Estimation", *International Journal for Scientific Research in Science: Engineering and Technology*, Vol.6, N.2, pp.665-671, 2019. DOI: 10.32628/IJSRSET207261
- [4] W. Luo, A. G. Schwing and R. Urtasun, "Efficient Deep Learning for Stereo Matching", *Computer Vision Foundation*, pp.5695-5703, 2016. DOI: 10.1109/CVPR.2016.614
- [5] X. Song, X. Zhao, L. Fang, H. Hu and Y. Yu, "EdgeStereo: An Effective Multi-task Learning Network for Stereo Matching and Edge Detection", *International Journal of Computer Vision*, Vol.128, No.4, pp.910-930, 2019. DOI: 10.1007/s11263-019-01287-w
- [6] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M. J. Black and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow", *International Journal of Computer Vision*, Vol.92, No.1, pp.1-31, 2011. DOI: 10.1007/s11263-010-0390-2
- [7] L. Li, S. Zhang, X. Yu and L. Zhang, "PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching", *IEEE Transactions on Circuits & Systems for Video Technology*, Vol.28, No.3, pp.679-692, 2018. DOI: 10.1109/TCSVT.2016.2628782
- [8] J. Lu, K. Zhang, G. Lafruit and F. Catthoor, "REAL-TIME STEREO MATCHING: A CROSS-BASED LOCAL APPROACH", In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.733-736, 2009. DOI: 10.1109/ICASSP.2009.4959688
- [9] Q. Yang, P. Ji, D. Li, S. Yao, and M. Zhang, "Fast stereo matching using adaptive guided filtering", *Image and Vision Computing*, Vol.32, No.3, pp.202-211, 2014. DOI: 10.1016/j.imavis.2014.01.001
- [10] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth", In *German Conference on Pattern Recognition*, pp.1-12, 2014. DOI: 10.1007/978-3-319-11752-2_3

Shi Zhou (Member) was born in Hunan, China, 1994. She received the B.S. and M.S. degrees in Mechanical Engineering and Automation from Northeastern University, China, in 2017 and 2020. She is currently a Ph.D student in Electrical and Electronic Engineering at Kyushu Institute of Technology, Japan. her research interests include 3D vision, SLAM, image processing and artificial intelligence.

He Li (Non-member) received the B.S., M.S. and Ph.D degrees in Mechanical Engineering and Automation from Northeastern University, China, in 1999, 2002 and 2005. Respectively, he is currently a professor of Northeastern University, China. His research interests include nonlinear vibration, mechanical system dynamics, and Machine Tool Dynamics. He is a member of Liaoning Association for Science and Technology, and Chinese Mechanical Engineering Society.

Miaobao Zhu (Member) received the M.S. degree in engineering from Xiaan University of Architecture and Technology, china, in 2008. Currently, she is a Ph.D. student in Electrical and Electronic Engineering at Kyushu Institute of Technology, Japan. Her research interests include image processing, image enhancement and Wireless sensor network.

Zhen Li (Member) received the B.S. degree in Communication Engineering from Yangzhou University in 2017. He is master candidate student of Engineering from Yangzhou University in China, and from Kyushu Institute of Technology in Japan since 2017. He is now a doctor candidate in Kyushu Institute of Technology. His current research interests include object detec-

tion and SLAM.

Mitsunori Mizumachi (Member) received the Bachelor degree in design from Kyushu Institute of Design and the Ph.D degree in information science from Japan Advanced Institute of Science and Technology (JAIST) in 1995 and 2000, respectively. He is currently an associate professor at Kyushu Institute of technology. His research interests include acoustic information

processing and statistical signal processing. He is a member of AES, ASA, ASJ, IEEE, IEICE, and RISP

Lifeng Zhang (Member) received the B.S. degree in Electronic Engineering from Southeast University in 1994. He received the M.S. and Ph.D. degrees in Electrical Engineering from Kyushu Institute of Technology, in 1999 and 2001. Respectively, he is currently a professor at Kyushu Institute of Technology, Japan. His current research interests include computer vision,

image processing, biometrics authentication and remote sensing system. He is a member of IIAE, IEEE, and IEICE.