International Journal of Innovative Computing, Information and Control
Volume **17**, Number **2**, April **2021**

ICIC International ©2021 ISSN 1349-4198
pp. 473–482

STEREO MATCHING APPROACH USING ZOOMING IMAGES

Bo-Yang Zhuo[1,2], Huei-Yung Lin[1,2] and Chin-Chen Chang[3,*]

[1]Department of Electrical Engineering
[2]Advanced Institute of Manufacturing with High-Tech Innovations
National Chung Cheng University
No. 168, Sec. 1, University Road, Minhsiung, Chiayi 621, Taiwan
ipopo520@gmail.com; lin@ee.ccu.edu.tw

[3]Department of Computer Science and Information Engineering
National United University
No. 2, Lienda Road, Miaoli 360, Taiwan
[*]Corresponding author: chinchen.chang@gmail.com

Received November 2020; revised February 2021

Abstract. *In this paper, we present a novel stereo matching approach using zooming images. In previous approaches, a pair of stereo images (left and right images) is used for correspondence matching. In the proposed approach, we used two zoom lens cameras to acquire multiple pairs of stereo images with zoom changes. By using the acquired stereo image sequences, we can obtain accurate results for stereo matching algorithms. By defining the relationship between the left and right images of a stereo image pair, the proposed approach renders the rectified images compliant with the zoom characteristics. Moreover, the proposed approach can be integrated with existing stereo matching algorithms. The results revealed that our approach can improve disparity computation for stereo matching.*

Keywords: Stereo matching, Zooming, Disparity, Optical zoom

1. Introduction.

Stereo vision is used in many applications such as depth perception and 3D model reconstruction. Stereo vision is a vital research topic in computer vision. Developing techniques for superimposing images based on stereo vision with other information have been comprehensively studied.

Stereo vision for machine perception is critical because it enables machine vision systems to simulate the human visual system. When an object is closer to the observer, the disparity between the eyes becomes larger. Therefore, the stereo vision problem can be simplified by computing the corresponding points within a pair of stereo images. The disparity maps can be obtained using the depth computation of 3D scenes or objects. Computer vision researchers have extensively studied stereo matching techniques [4,7-9,11,16-19]. Based on the Middlebury stereo data sets [14] and public benchmarks, several algorithms have been proposed and evaluated for decreasing the mismatching rate. However, most approaches use standard rectified image pairs as inputs and do not consider the image acquisition process of a real camera system. Chen et al. [3] proposed a framework for stereo matching approach. Their approach took two or more stereo image pairs with different focal lengths. An initial disparity map is first computed by the stereo image pair with the same focal length. A process is introduced to identify the point correspondences among these stereo images. The cost aggregation is then performed to refine the disparity map.

In this paper, a stereo matching approach is presented based on zooming images. We use a pair of zoom lens cameras to capture stereo images with various focal length settings.

DOI: 10.24507/ijicic.17.02.473

In particular, a zoom rectification method is introduced to reduce the zoom image correspondence search range. We aggregate the matching cost of stereo and zoomed images to mitigate unreliable matching. The proposed approach can improve the correspondence search results with the additional zooming constraint and provide a robust disparity reliability check.

2. Related Works.

A stereo image pair captured using a conventional stereo vision system consists of two images captured from two cameras. This imaging model forms the two-view geometry and the point correspondence relationship between the two images is restricted by the epipolar constraint [6]. For convenience, the image pairs are commonly rectified and the epipolar lines are parallel to the image scanlines. Stereo matching can then be performed efficiently along the one-dimensional image scanlines at the cost of rectification computation and image warping. Because most existing stereo matching algorithms use rectified image pairs and do not consider the image rectification step, studies are commonly focused on the matching cost rather than the development and evaluation of the overall stereo vision system.

In general, stereo matching algorithms involve the following four steps: cost initialization, cost aggregation, disparity selection and refinement. The cost initialization is a process to calculate the similarity at the pixel level, such as the absolute difference or cross correlation. Because the cost calculated by a pixel is not reliable, cost aggregation considering a specific region was performed to increase the robustness. Hosni et al. [7] presented a fast algorithm for stereo matching. They used a filter to filter the cost volume. Hence, the computational cost of their approach is independent of the size of a window for stereo matching. The disparity selection typically adopts the winner-takes-all strategy, which determines the lowest cost for the result. Alternatively, Chang and Maruyama [2] proposed an image scaling approach with multi-block matching and sub-pixel estimation can be used to reduce the error rate. The refinement process aims to improve the reliability through techniques such as the left-right consistency, matching confidence, median filter, speckle filter, and ground control points. However, the accuracy and computational cost are limited by the memory size.

The stereo vision research is divided into two categories, namely the speed-oriented and precision-oriented approaches. The speed-oriented techniques concern the computational efficiency and consider the porting to hardware-dependent platforms such as FPGA and GPU [2]. Park and Yoon [13] introduced a method to compute disparity maps. Their approach divided the stereo matching into initial disparity map estimation, plane hypotheses generation, and global optimization. The results revealed that their method can deal with ambiguous regions. The precision-oriented techniques aim to increase the correctness of stereo matching results. In some studies, possible surface structures are used to improve accuracy. Kim and Kim [11] presented a stereo matching approach which used the texture and edge information as the smoothness constraints. Their approach can provide good stereo matching performance and obtain desired results. Batsos et al. [1] combined various input scales, masks, and cost calculation methods to make the algorithms more robust. Moreover, considerable progress has been achieved in learning-based techniques for stereo matching. Žbontar and LeCun [18] proposed a technique to learn a similarity measure on small image patches using the convolutional neural network and evaluate the similarity on KITTI and Middlebury data sets. Results show that their approach outperforms other approaches. Seki and Pollefeys [15] presented a learning-based penalties estimation technique to derive the parameters of the semi-global matching algorithm. They introduced a loss function to train the networks with learned penalties for semi-global matching. Cheng and Lin [4] presented a matching technique based on image bit-plane slicing and fusion.

The bit-plane slices were used to determine stereo correspondences and then combined for the final disparity map. These techniques have effectively reduced the correspondence matching error but require sophisticated hardware.

3. **Proposed Approach.** Figure 1 displays the flow of the proposed approach. We first extracted two or more stereo image pairs with various zoom factors and various focal length settings. Then, we performed an initial disparity computation using the stereo image pair acquired with the same focal length. A series of zoom images captured from the same camera is used for zoom matching, which is a process to identify the point correspondences among the zoom images. Finally, cost aggregation combining the matching from stereo and zoom is then performed to refine the disparity map.

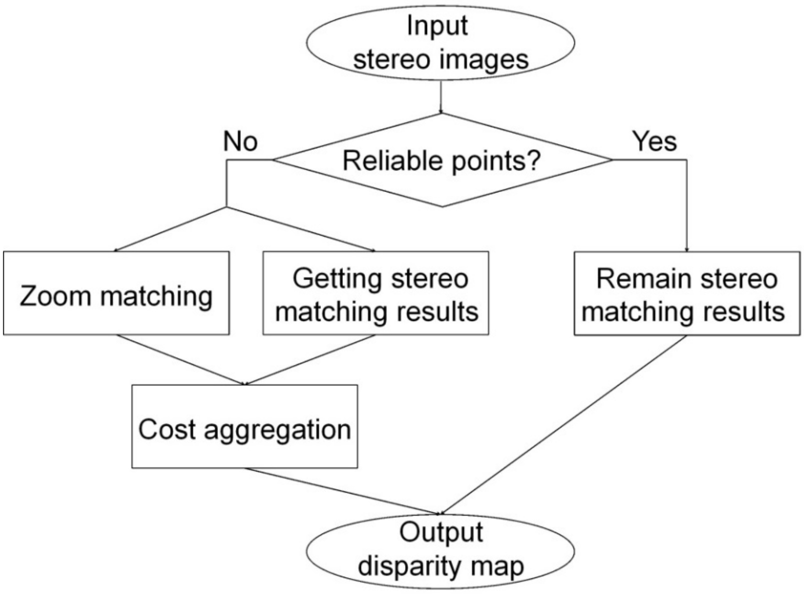

FIGURE 1. Flowchart of the proposed approach

3.1. **Zoom-stereo framework.** The scale of a scene or an object appearing in an image can be determined using the focal length of the camera or the zoom factor. A vector defined by the image center and a specific point can be used to illustrate the phenomenon when the focal length is changed. An ideal zoom model can be described as follows:

$$\frac{|v_i|}{f_i} = \frac{|v_j|}{f_j} \text{ and } v_j = \lambda \cdot v_i, \tag{1}$$

where v_i and v_j are the zoom vectors originating from the principal point, and λ is the focal length ratio f_j/f_i.

Zooming images are of two types: one derived from digital zoom and the other acquired with an optical zoom. Digital zoom is synthesized through up- or down-sampling the original image with interpolation and thus no additional information is generated when magnified. The optical zoom involves the actual lens movement, and the zooming images are captured with independent samples. In the proposed approach, the focus length is changed to acquire the zoom images for stereo matching. A difficulty of optical zoom is the change of the principle point along with the change of the focus length. The zoom vectors do not converge to a single point due to the non-ideal lens movement of a real camera system with optical zoom.

In a conventional stereo vision system, the distance z of a scene point is obtained using the following expression:

$$z = f \cdot \frac{b}{d}$$

where f is the focal length of the camera, b is the stereo baseline, and d is the stereo disparity. The stereo baseline and focal length can be obtained by camera calibration and used to check the disparity reliability using the left-right consistency [9]. A pair of zoom lens cameras was used in our system, and several stereo image pairs are captured at a fixed location. Because the distance between the camera and the scene does not change, additional geometric constraints can be constructed for the image pairs. In a real camera system, the principal point changes due to zooming. Thus, we should consider the baseline change for cooperative stereo and zoom matching. For a conventional stereo system setting, the disparity is proportional to the stereo baseline. The restriction can be expressed by

$$\frac{d_i}{d_j} = \frac{b_i}{b_j}, \tag{2}$$

where i and j represent various zoom positions for image acquisition.

3.2. **Disparity reliability.** For reliable stereo matching results, identification of the error correspondences (also called unreliable points) in the disparity map before the calculation of matching cost aggregation is critical. Because the error correspondences typically appear near the image edges due to the depth discontinuity [19], Canny edge detection is first used on the image, followed by morphological dilation to determine the unreliable points. The matching confidence also considers the pixel location difference between the smallest cost and the second smallest cost. In the proposed approach, Equation (2) is used to identify the unreliable points. Notably, Equation (2) is given in the same world coordinate system for i and j, instead of the image coordinates. Thus, the zoom correspondence matching should be performed to align the image coordinate frames. The disparity maps derived from various zooms are then subtracted to obtain the unreliable point map. Finally, we combined the matching confidence $R_{con}[x,y]$, edge discontinuity $R_{edge}[x,y]$, and zooming $R_{zoom}[x,y]$ to derive the error correspondences $R[x,y]$ as follows:

$$R_{con}[x,y] = \begin{cases} 1, & 1 - \dfrac{C^{\text{1st}}_{x,y}}{C^{\text{2nd}}_{x,y}} < \tau \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and

$$R[x,y] = R_{con}[x,y] \cup R_{edge}[x,y] \cup R_{zoom}[x,y], \tag{4}$$

where τ is a user parameter. It is not easy to adequately determine the user parameter τ. In the experiments, τ was set as 0.2 for the best performance, heuristically.

Figure 2 displays an example of the image "Adirondack" in the Middlebury data set processed by the proposed disparity reliability check method. We used the percentage of error correspondences marked as unreliable points to evaluate the results, and the values of $R_{con}[x,y]$, $R_{con}[x,y] \cup R_{edge}[x,y]$ and $R[x,y]$ are 46.87%, 59.47%, and 71.88%, respectively.

3.3. **Disparity candidates.** When calculating the disparity map, we typically provide a maximum disparity D. The parameter D is determined by the stereo vision system setup because it has to be smaller than the disparity corresponding to the closest scene distance perceivable by both cameras. If all possibilities are considered for stereo matching with various zooming images, the time complexity will become $O(D^2)$. Thus, it is necessary to

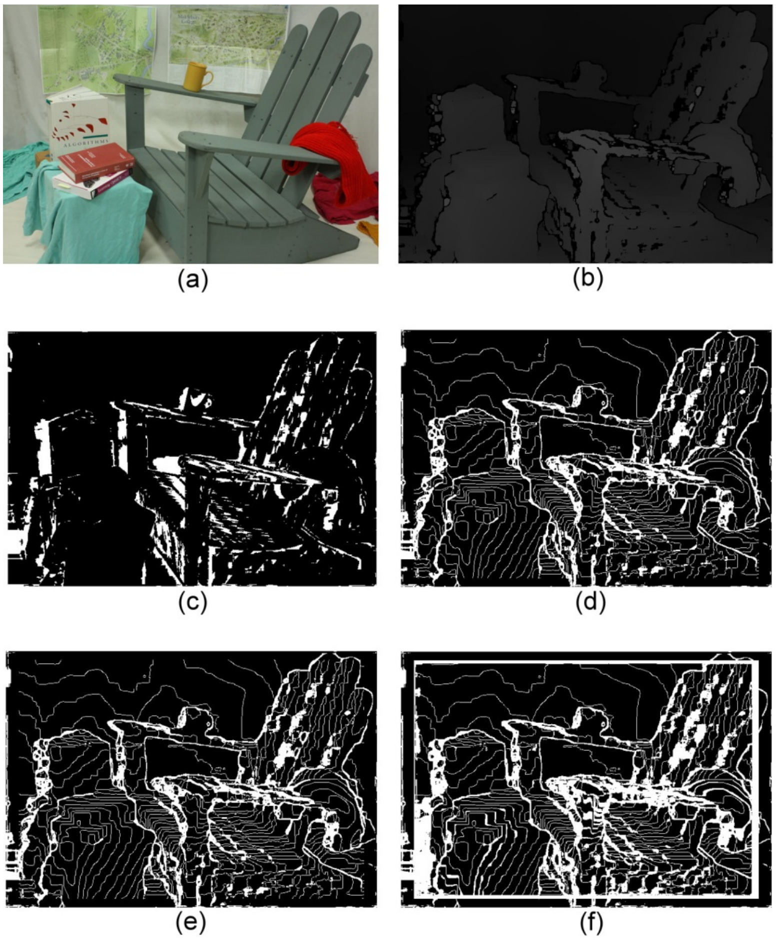

Figure 2. The disparity reliability check on the image “Adirondack”: (a) original left image, (b) disparity map, (c) error point, (d) $R_{con}[x,y] \cup R_{edge}[x,y]$, (e) resulting image with edges, and (f) resulting image with zoom

select only some specific disparities as candidates to reduce the computation time. The candidates can typically be selected by the matching cost. However, these techniques may not be appropriate for some algorithms such as semi-global block matching (SGBM) and the local minimum is used for the candidates instead. In this case, block matching (BM) provides the more stable result compared to SGBM.

3.4. **Zoom correspondence.** After the main geometric construction and the matching between the stereo image pair, the next problem involves determining the correspondences among various zoom images. For the digital zoom, the correspondence can be obtained directly by the image scale change. However, it is not a trivial task for the optical zoom.

The images captured by a camera with various zooms can be assumed as acquired by multiple cameras. Then, the multiple view geometry can be considered. The most vital relationship between the images is the epipolar constraint. Except for some special cases, the stereo matching search range can be reduced from the 2D plane to a 1D line. We used homography to determine the correspondences directly for planar objects because the camera translation is zero. It is a special case of the epipolar constraints, and can be expressed as follows:

$$s\mathbf{q}' = K'\left(R + \frac{\mathbf{t}}{d}\mathbf{n}^T\right)K\mathbf{q} \rightarrow s\mathbf{q}' = H\mathbf{q}, \tag{5}$$

where $\mathbf{q}$ and $\mathbf{q}'$ are in homogeneous coordinates, R and $\mathbf{t}$ are the rotation and translation between the cameras, $\mathbf{n}$ and d are the normal vector and distance of the object with respect to the first camera respectively, and s is the scale factor.

To address the problem of the zooming property changed by image rectification, we used camera calibration to establish the relationship between various zooming images. Because this image rectification only involved a rotation transformation, it can be computed easily through homography. However, the zoom property only holds under certain circumstances. Thus, some constraints, such as the position of the principal point and the fixed aspect ratio, are added.

When the calibration was performed with additional constraints, the calibration error indicated by the re-projection error was magnified, as indicated in Table 1. The idea zoom model assumes that two zoom images have the same principal points. The field-of-view changes along the optical axis. Our calibration result is with a little difference in rotation: $(-0.05°, -1.18°, -0.08°)$ and translation: $(-0.7, 2.7, -1.88)$. This should be neglected in most common situations [10]. For more precise results, we adopted the SIFT features [12] and RANSAC [5] for the correspondence matching to calculate a new zoom center.

Table 1. Calibration performed under various constraints

Condition	Focal X	Focal Y	Principal point	Reprojection error
Common	2209	2190	$(988, 887)$	$(0.19, 0.25)$
No distortion	2962	2884	$(1307, 414)$	$(0.33, 0.32)$
Aspect = 1	2166	2166	$(1003, 928)$	$(0.19, 0.25)$
Principal point fixed	2243	2243	$(1023.5, 767.5)$	$(0.20, 0.25)$

3.5. **Cost aggregation.** To combine the information of two zooming image pairs, it is necessary to study how their relationship can be used. Figure 3 displays the concept and schematic of the proposed approach.

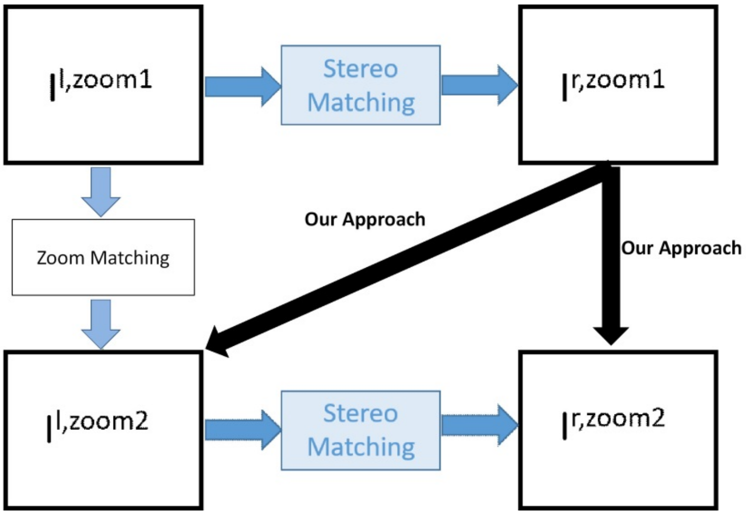

Figure 3. Relationship between two zooming image pairs

Here, we defined a new cost function:

$$Disparity = \arg\min_{n} \alpha \left(C_n^{zoom1} + C_m^{zoom2}\right) + \beta(C_{mn}), \quad 1 \le n \le 3, \ \ 1 \le m \le 3, \qquad (6)$$

where α and β are user-defined parameters; C_n is the matching cost between left and right images with the same focal length given by

$$C_n = Cost(i, j, d_n(i, j)), \quad 1 \le n \le 3, \tag{7}$$

where $d_n(i, j)$ is the disparity of the candidate pixel (i, j); C_{mn} is the matching cost between the zoom images given by

$$\begin{aligned} C_{mn} = {} & Cost(Z_{1R}(i - d_n(i, j), j), Z_{2R}(i' - d_m(i', j'), j')) \\ & + Cost(Z_{1R}(i - d_n(i, j), j), Z_{2L}(i', j')), \quad 1 \le n \le 3, \ \ 1 \le m \le 3, \end{aligned} \tag{8}$$

where Z_{1R}, Z_{2R} and Z_{2L} are the zoom1 left image, the zoom2 right image and the zoom2 left image, respectively.

We only considered the pairs produced by the disparity candidates, and *Cost* is a similarity metric such as the sum of squared differences (SSD), census transform (CT), or normalized cross correlation (NCC). Here, C_n^{zoom1} and C_m^{zoom2} can be replaced by the cost given in the stereo matching algorithms. Equation (6) is based on the texture information. In the reliability check, we use the disparity as a constraint. The correspondence pairs produced by zoom images are in the same place. Therefore, their distances to the camera are the same. Thus, the cost function can incorporate the disparity constraint and is expressed as follows:

$$\begin{aligned} & Final\ Disparity = \arg\min_n \alpha \left(C_n^{zoom1} + C_m^{zoom2} \right) + \beta \left(Disparity_n^{zoom1} - Disparity_m^{zoom2} \right), \\ & 1 \le n \le 3, \ \ 1 \le m \le 3, \end{aligned} \tag{9}$$

where α and β are user defined parameters as the previous equations.

4. **Experimental Results.** The proposed approach was evaluated on the Middlebury stereo data sets and our own data set. Because Middlebury data sets do not contain zooming images, we manually synthesize the zoom images by resizing. They are used to verify the proposed approach in the cost function evaluation. The stereo matching algorithms adopted in our system for performance comparison are BM, SGBM, and matching cost convolutional neural network (MC-CNN). BM and SGBM run with the window size of 13×13, and P_1, and P_2 are 18 and 32, respectively. MC-CNN uses the fast model trained by KITTI data sets. The parameter α is fixed as 1 and β is 20, 10 and 0.5 for SGBM, BM, and MC-CNN, respectively. We selected the "Q" Middlebury data set as the input. The bad pixel rate (BPR) representing the percentage of bad pixels in an image is employed as an objective performance measure. In the experiments, the evaluation method was BPR1.0.

The results from various methods are displayed in Table 2. From the results, the performance of zoom calibration is more prominent for most cases. However, the performance of original method is better for some cases. This is because when calculating these data sets, they do not provide correct information and may even produce misleading results. For example, the performance for the toy brick data of BM + Equation (9) is better than that of other approaches.

In the cost function Equation (6), we only tested CT and NCC because the image intensity was shifted by the camera lens change and auto-exposure. The results from our approach were not significant when adopted to the MC-CNN. This phenomenon could be because of the design of the cost function for MC-CNN training. If only the best solution is considered during training, it will not provide the best correspondence candidates for the algorithm, and the normal MC-CNN has a similar process such as SGBM. Figure 4 displays the results with or without the SGBM process. Clear differences were observed mainly because the basic algorithm cannot provide the suitable candidates. Figure 5 displays resulting images of optical zooming for doll. From the results, for disparity matching,

TABLE 2. The results from various methods

		Original	Homography	Epipole geometry	Zoom calibration
doll	SGBM + Equation (9)	19.6503	15.2552	15.5919	15.801
	BM + Equation (9)	25.5569	22.3145	22.3699	23.0199
	MC-CNN + Equation (9)	22.3460	20.2321	21.0419	20.4651
toy brick	SGBM + Equation (9)	14.5528	13.3947	13.6271	13.3372
	BM + Equation (9)	15.1138	15.7089	15.7333	15.5487
	MC-CNN + Equation (9)	18.9041	25.8092	21.7056	24.3230
toy brick and cup	SGBM + Equation (9)	9.8251	9.2016	9.5835	9.0291
	BM + Equation (9)	12.797	12.1638	11.7294	12.1277
	MC-CNN + Equation (9)	11.8306	12.9873	12.8016	12.9491
toy brick and lamp	SGBM + Equation (9)	19.6457	17.6351	17.5975	17.804
	BM + Equation (9)	26.4771	25.8682	25.2747	25.6811
	MC-CNN + Equation (9)	23.2376	22.8388	22.4940	22.9996

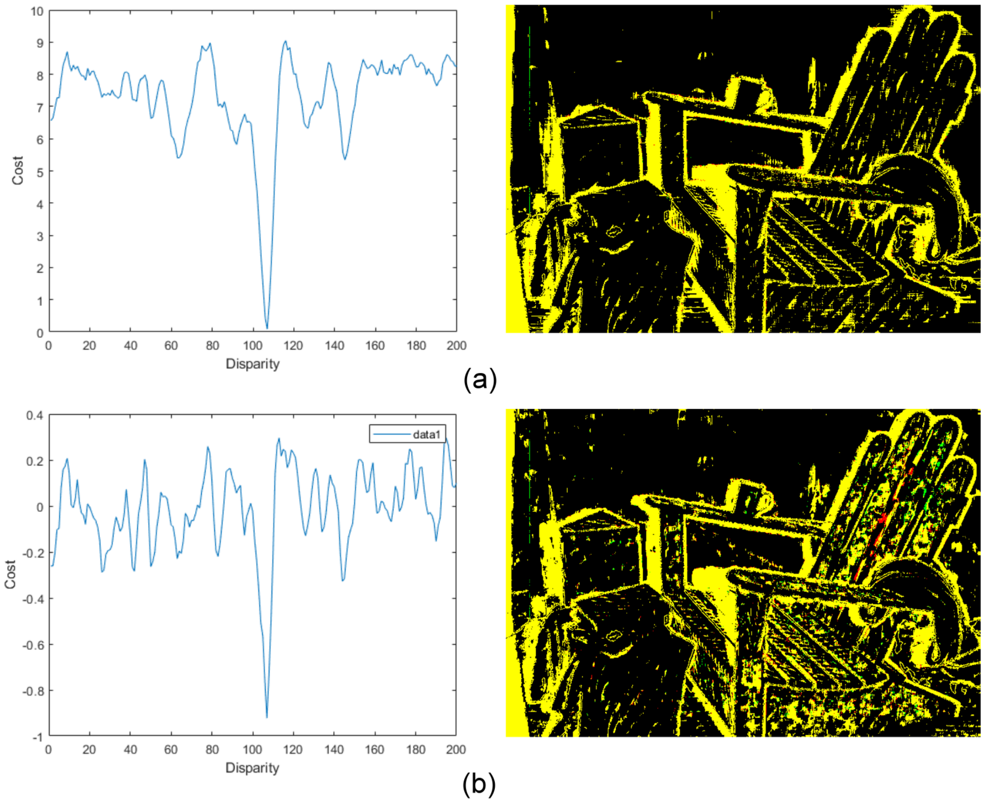

FIGURE 4. (color online) MC-CNN with various processes. The improvement is denoted in red, and the green part indicates the incorrect change region: (a) MC-CNN + SGBM and resulting image, and (b) MC-CNN and resulting image.

some parts of the disparity values were directly modified. For NCC, the original disparity map was used to do filling and minor changes.

Finally, we used zoom lens cameras to construct a new data set with the ground truth. The evaluation method using our data set was changed to BPR2.0 because the manual labeling of the ground truth is not precise. We tested the zoom correspondence methods

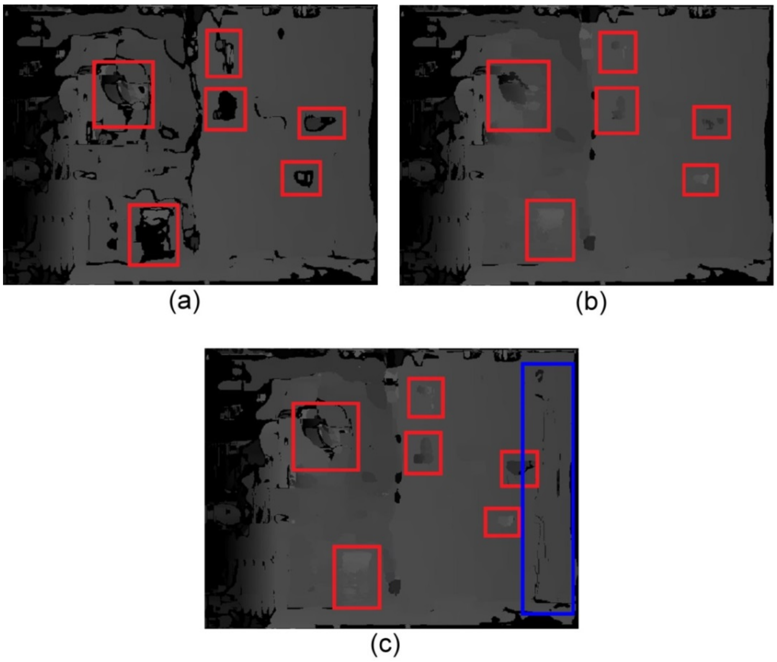

Figure 5. Resulting images of optical zooming: (a) doll + SBGM (original), (b) doll + SBGM (disparity matching), and (c) doll + SBGM (NCC)

mentioned previously, and tested the effect of the candidate quantity. The toy brick+BM result is noticeably different from other results. When we calculate the BPR with zoom2 disparity map in BM, its value was nearly 80%. Thus, the two initial disparity maps must have a certain level of correctness.

5. **Conclusions.** We have proposed a stereo matching approach using zooming images. With zoom image pairs, the proposed approach can reduce the error and the uncertain region in the disparity map. Compared with the existing stereo matching algorithms, the proposed approach can improve the disparity results with less computation. The proposed approach can be adapted to the existing local and global methods for stereo matching. In the future studies, more investigation will be performed to aggregate the information for machine learning methods and the cost computation.

Acknowledgment. This paper is a revised and expanded version of a paper entitled "Cooperative Stereo-Zoom Matching for Disparity Computation", presented at the 15th International Conference on Computer Vision Theory and Applications (VISAPP 2020), 2020.

REFERENCES

[1] K. Batsos, C. Cai and P. Mordohai, CBMV: A coalesced bidirectional matching volume for disparity estimation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2060-2069, 2018.

[2] Q. Chang and T. Maruyama, Real-time stereo vision system: A multi-block matching on GPU, *IEEE Access*, vol.6, pp.42030-42046, 2018.

[3] Y. Chen, B. Zhuo and H. Lin, Stereo with zooming, *Proc. of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp.2224-2229, 2018.

[4] K. Cheng and H. Lin, Stereo matching with bit-plane slicing and disparity fusion, *Proc. of 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp.341-346, 2015.

[5] M. A. Fischler and R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, vol.24, no.6, pp.381-395, 1981.

[6] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd Edition, Cambridge University Press, 2004.

[7] A. Hosni, M. Bleyer, C. Rhemann, M. Gelautz and C. Rother, Real-time local stereo matching using guided image filtering, *Proc. of 2011 IEEE International Conference on Multimedia and Expo*, pp.1-6, 2011.

[8] C. S. Huang, Y. H. Huang, D. Y. Chan and J. F. Yang, Shape-reserved stereo matching with segment-based cost aggregation and dual-path refinement, *EURASIP Journal on Image and Video Processing*, 2020.

[9] Z. Jie, P. Wang, Y. Ling, B. Zhao, Y. Wei, J. Feng and W. Liu, Left-right comparative re-current model for stereo matching, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3838-3846, 2018.

[10] M. V. Joshi, S. Chaudhuri and R. Panuganti, Super-resolution imaging: Use of zoom as a cue, *Image and Vision Computing*, vol.22, no.14, pp.1185-1196, 2004.

[11] K. R. Kim and C. S. Kim, Adaptive smoothness constraints for efficient stereo matching using texture and edge information, *Proc. of 2016 IEEE International Conference on Image Processing (ICIP)*, pp.3429-3433, 2016.

[12] D. G. Lowe, Object recognition from local scale-invariant features, *Proc. of the International Conference on Computer Vision (ICCV)*, vol.2, pp.1150-1157, 1999.

[13] M. G. Park and K. J. Yoon, As-planar-as-possible depth map estimation, *Computer Vision and Image Understanding*, vol.181, pp.50-59, 2019.

[14] D. Scharstein and R. Szeliski, *Middlebury Stereo Vision Page*, http://vision.middlebury.edu/stereo, 2002.

[15] A. Seki and M. Pollefeys, SGM-Nets: Semi-global matching with neural networks, *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6640-6649, 2017.

[16] Y. Shin, M. Kim, K.-W. Pak and D. Kim, Practical methods of image data preprocessing for enhancing the performance of deep learning based road crack detection, *ICIC Express Letters, Part B: Applications*, vol.11, no.4, pp.373-379, 2020.

[17] E. Song, S. Kim and M. Chang, Novel stereo-matching method utilizing surface normal data, *International Journal of Precision Engineering and Manufacturing*, vol.21, pp.1437-1445, 2020.

[18] J. Žbontar and Y. LeCun, Stereo matching by training a convolutional neural network to compare image patches, *The Journal of Machine Learning Research*, vol.17, no.1, pp.2287-2318, 2016.

[19] S. Zhang, W. Xie, G. Zhang, H. Bao and M. Kaess, Robust stereo matching with surface normal prediction, *Proc. of 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp.2540-2547, 2017.