Cooperative Stereo-Zoom Matching for Disparity Computation

Bo-Yang Zhuo and Huei-Yung Lin
Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan

Keywords: Stereo Matching, Zooming, Disparity, Optical Zoom.

Abstract: This paper investigates a stereo matching approach incorporated with the zooming information. Conventional stereo vision algorithms take one pair of images for correspondence matching, while our proposed method adopts two zoom lens cameras to acquire multiple stereo image pairs with zoom changes. These image sequences are able to provide more accurate results for stereo matching algorithms. The new framework makes the rectified images compliant with the zoom characteristics by the definition of the relationship between the left and right images. Our approach can be integrated with existing stereo matching algorithms under some requirements and adjustments. In the experiments, we test the proposed framework on 2014 Middlebury benchmark dataset and our own zoom image dataset. The results have demonstrated the improvement of disparity computation of the our technique.

1 INTRODUCTION

Stereo vision for machine perception is to simulate the human visual system. When an object is closer to the observer, the disparity between the left and right eyes becomes larger. With this property, the stereo vision problem can be simplified as finding the corresponding points within an image pair. This is an important topic in computer vision, and aims to provide the disparity maps from the depth computation of 3D scenes or objects. In the past few decades, stereo matching techniques have been extensively investigated by computer vision researchers and practitioners. Since the availability of Middlebury stereo datasets (Scharstein and Szeliski, 2002) and public benchmarks, many sophisticated algorithms have been proposed and evaluated for the objective of decreasing the mismatching rate. Nevertheless, most approaches take standard rectified image pairs as inputs and do not consider the image acquisition process of a real camera system.

The image pair captured by a conventional stereo vision system consists of two images taken from two different cameras. This imaging model forms the two-view geometry and the point correspondence relation between the two images is restricted by the epipolar constraint (Hartley and Zisserman, 2004). For convenience, the image pairs are commonly rectified so that the epipolar lines are parallel to the image scanlines. The stereo matching can then be carried out efficiently along the one-dimensional image scanlines at the cost of rectification computation and image warping. Since most existing stereo matching algorithms work on rectified image pairs and do not consider the image rectification step, the investigation is commonly focused on the matching cost rather than the development and evaluation of the overall stereo vision system.

In general, there are four steps adopted in stereo matching algorithms: 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, etc. Since the cost calculated by a pixel is not reliable, the cost aggregation considering a specific region is carried out to increase the robustness (Hosni et al., 2011). The disparity selection usually adopts the winner-takes-all (WTA) strategy which finds the lowest cost for the result. Alternatively, an image scaling approach with multi-block matching and sub-pixel estimation can be used to reduce the error rate (Chang and Maruyama, 2018). The refinement process aims to improve the reliability using some techniques such as the left-right consistency, matching confidence, median filter, speckle filter, and ground control points, etc.

The stereo vision research in recent years is divided into two categories, the speed-oriented and precision-oriented approaches. The speed-oriented techniques concern about the computational effi-

Zhuo, B. and Lin, H.
Cooperative Stereo-Zoom Matching for Disparity Computation.
DOI: 10.5220/0008877001570164
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 157-164
ISBN: 978-989-758-402-2; ISSN: 2184-4321

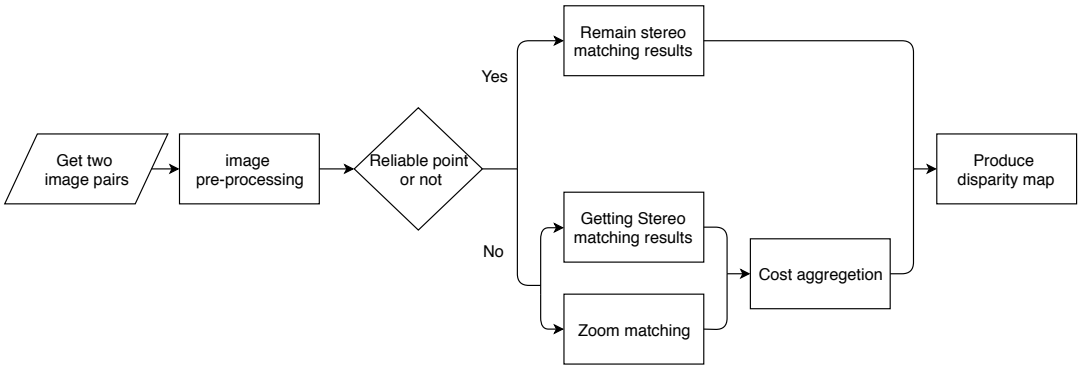


Figure 1: The system flowchart of the proposed cooperative zoom-stereo framework for disparity computation.

ciency and consider the porting to hardware dependent platforms such as FPGA, GPU, etc. (Chang and Maruyama, 2018). On the other hand, the objective of precision-oriented techniques is to increase the correctness of stereo matching results. In some recent works, possible surface structures are used to improve the accuracy (Park and Yoon, 2019), Kim and Kim use the texture and edge information as the smoothness constraints (Kim and Kim, 2016). In (Batsos et al., 2018), various input scales, masks and cost calculation methods are combined to make the algorithms more robust. Moreover, learning based techniques for stereo matching have shown significant progress in the past few years. In (Žbontar and LeCun, 2016), a method is proposed to learn a similarity measure on small image patches using the convolutional neural network, and evaluating on KITTI and Middlebury datasets. Seki *et al.* present a learning based penalties estimation method to derive the parameters of the widely used semi-global matching algorithm (Seki and Pollefeys, 2017). Cheng and Lin present a matching technique based on image bit-plane slicing and fusion (Cheng and Lin, 2015). The bit-plane slices are used to find stereo correspondences and then combined for the final disparity map. These techniques have successfully reduced the correspondence matching error, but at the cost of more sophisticated hardware requirement.

In this work, a pair of zoom lens cameras are used to capture the stereo image pairs with various focal length settings (Chen et al., 2018). The existing or newly developed stereo matching algorithms can be incorporated with the proposed zoom-stereo framework. A zoom rectification method is presented to reduce the zoom image correspondence search range. We aggregate the matching cost of stereo and zoom images to mitigate the unreliable matching. Our approach is able to improve the correspondence search results with the additional zooming constraint, and provides a robust disparity reliability check. Figure 1 shows the system flowchart of the proposed stereo with zooming technique. Two or more stereo image pairs with various zoom factors are first taken with different focal length settings. An initial disparity computation is first carried out by 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. The cost aggregation combining the matching from stereo and zoom is then performed to refine the disparity map.

2 THE APPROACH

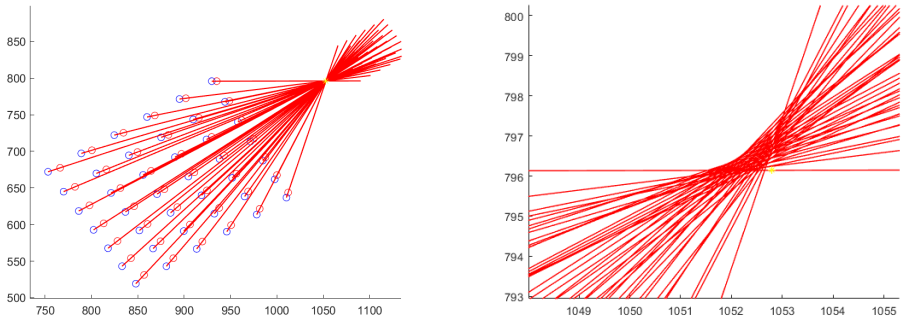
2.1 Zoom-Stereo Framework

The scale of a scene or an object appeared in the image is determined by 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 by the equation

$$\frac{|\mathbf{v}_i|}{f_i} = \frac{|\mathbf{v}_j|}{f_j} \quad \text{and} \quad \mathbf{v}_j = \lambda \mathbf{v}_i \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the zoom vectors originated from the principal point (or the image center), and λ is the focal length ratio f_j/f_i.

Normally, there are two types of zoom images, one derived from digital zoom and the other acquired with optical zoom. The former is synthesized by up- or down-sampling the original image with interpolation, and thus no additional information is generated when magnified. The optical zoom, on the other hand, involves the actual lens movement, and the zoom images are captured with independent samples. In this



(a) Ideal zoom (b) Optical zoom

Figure 2: The red lines are connected by the same correspondences appeared in different zooms. The yellow dot is the image center obtained from calibration.

work, the focus length is changed to acquire the zoom images for stereo matching. One difficulty of optical zoom is the change of principle point along with the change of focus length. As illustrated in Figure 2, 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 (see Figure 2(b)). This is different from ideal case used in digital zoom (see Figure 2(a)).

In a conventional stereo vision system, the distance z of a scene point is given by

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

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 by the left-right consistency (Jie et al., 2018). Notice that a pair of zoom lens cameras is used in our system, and several stereo image pairs are captured at a fixed location. Since 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 will change due to zooming (i.e., when the focal length is adjusted). Thus, we also need to 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 formulated as

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

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

2.2 Disparity Reliability Check

For reliable stereo matching results, we need to identify the error correspondences (also called unreliable points) in the disparity map before the calculation of matching cost aggregation. Since the error correspondences usually appear near the image edges due to the depth discontinuity (Zhang et al., 2017), Canny edge detection is first carried out on the image, followed by morphological dilation to find the unreliable points. The matching confidence also considers the pixel location difference between the smallest cost and the second small cost. In our stereo with zooming approach, Equation (2) is also used to identify the unreliable points.

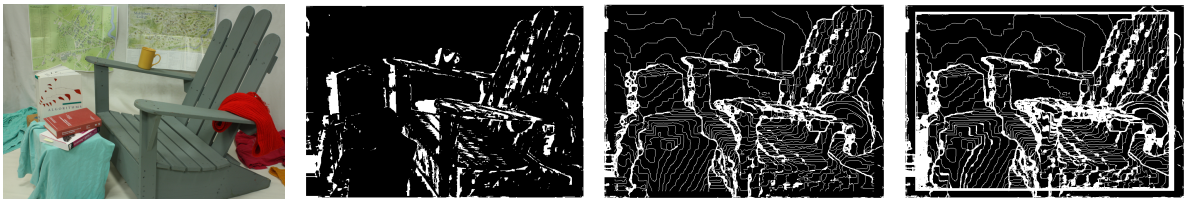
It should be noted that, Equation (2) is given in the same world coordinate system for i and j, instead of the image coordinates. Thus, the zoom correspondence matching should be performed to align the image coordinate frames. The disparity maps derived from different zooms are then subtracted to obtain the unreliable point map. Finally, we combine the matching confidence, edge discontinuity, and zooming to derive the error correspondences, as given by the equations

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

and

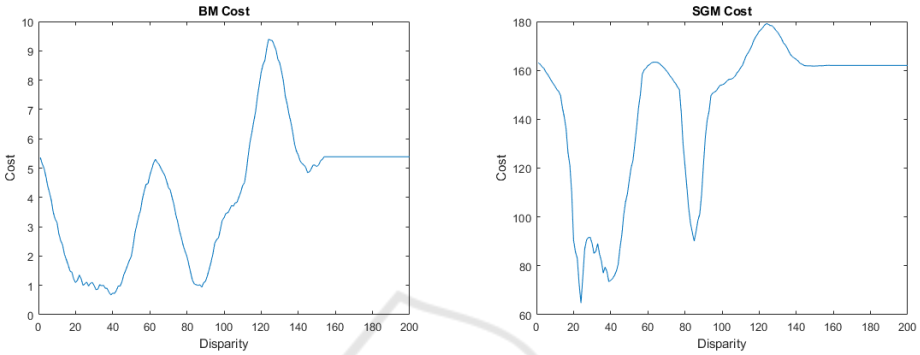
$$R[x,y] = R_{con}[x,y] \cup R_{edge}[x,y] \cup R_{zoom}[x,y] \tag{4}$$

where τ is a user parameter and set as 0.2 in the experiments. Figure 3 shows an example of the image 'Adirondack' in the Middlebury dataset processed by the proposed disparity reliability check method. We use 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.



(a) Original left image (b) Error point (c) $R_{con}(x,y) \cup R_{edge}(x,y)$ (d) Result with zoom

Figure 3: The disparity reliability check on the image 'Adirondack' (in Middlebury 2014 dataset).



(a) BM (b) SGBM

Figure 4: The disparity cost curves for block matching (BM) and semi-global matching (SGBM).

2.3 Disparity Candidates

When calculating the disparity map, we usually provide a maximum disparity (D). The parameter D is determined by the stereo vision system setup since 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 different zoom images, the time complexity will become O(D^2). Thus, it is necessary to choose only some specific disparities as candidates to reduce the computation time. In general, the candidates can be selected by the matching cost. However, it might not work well for some algorithms such as SGBM and the local minimum is used for the candidates instead. Figure 4 shows the disparity cost curves for block matching (BM) and semi-global matching (SGBM). BM provides the more stable result compared to SGBM.

(a) Zoom1 images before rectification.

(b) Zoom2 images before rectification.

Figure 5: Although two image are captured in the same place, little changes of the lens will make the image rectification results different.

2.4 Zoom Correspondence

After the main geometric construction and the matching between the stereo image pair, the next problem is to find the correspondences among different 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. As illustrated in Figure 5, two stereo image pairs are captured with different zoom positions. Due to the lens movement for the zoom change, the image rectification results are not identical even a suitable scale factor is applied.

The images captured by a camera with different zooms can thought as acquired by different cameras. It can then be considered as multiple view geometry,

and the most important 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 use homography to find the correspondences directly for planar objects since the camera translation is zero. It is a special case of the epipolar constraints, and can be formulated by

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

where $\mathbf{q}$ and $\mathbf{q}'$ are in homogeneous coordinates, R and $\mathbf{t}$ is 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 a scale factor.

To deal with the problem of the zooming property changed by image rectification, we use camera calibration to establish the relationship between different zoom images. Since this image rectification only involves a rotation transformation, it can be computed easily by homography. It should be noted that the zoom property only holds under certain circumstances as described in Figure 2. Thus, we need to add some constraints such as the position of the principal point and the fixed aspect ratio.

When the calibration is carried out with additional constraints, the calibration error indicated by the re-projection error is magnified as illustrated in Table 1. The idea zoom model assumes that two zoom images have the same principal points, i.e. the field-of-view changes along the optical axis. Our calibration result is with a little difference in rotation: $(-0.05^\circ, -1.18^\circ, -0.08^\circ)$ and translation: $(-0.7, 2.7, -1.88)$. This should be neglected in most common situations (Joshi et al., 2004). For more precise results, we adopt the SIFT features (Lowe et al., 1999) and RANSAC (Fischler and Bolles, 1981) for the correspondence matching to calculate a new zoom center.

2.5 Cost Aggregation

To combine the information of two zoom image pairs, it is necessary to study how their relationship can be used. Figure 6 shows the concept and schematic diagram of the proposed approach. The major concerns are indicated by the red arrows since there is no guarantee of correspondence matching in this part. Here we define a new cost function

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

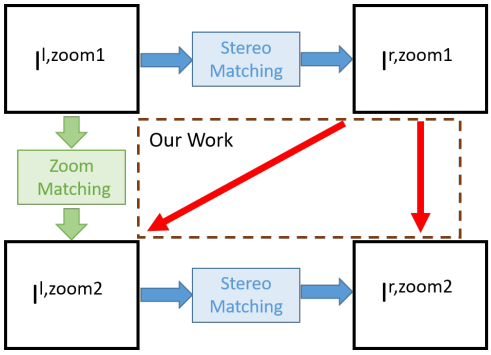


Figure 6: The relationship between two zoom image pairs. The blue arrows represent the stereo matching, and the green arrow represents the zoom correspondence.

where

$$C_n = Cost(i, j, d_n(i, j)), 1 \le n \le 3 \qquad (7)$$

and

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

We only consider the pairs which are produced by the disparity candidates, and *Cost* is a similarity metric such as SSD, CT, or NCC. 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 pair produced by zoom images are in the same place, so their distances to the camera are the same. Thus, the cost function can incorporate the disparity constraint and is written as

$$FinalDisparity = \arg\min_n \alpha(C_n^{zoom1} + C_m^{zoom2}) \\ + \beta(Disparity_n^{zoom1} - Disparity_m^{zoom2}) \\ 1 \le n \le 3, 1 \le m \le 3 \qquad (9)$$

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

3 EXPERIMENTS

The proposed technique is tested on the Middlebury stereo datasets and our own dataset. Since Middlebury datasets do not contain zoom images, we manually synthesize the zoom images by resizing. They are used to verify our method in the cost function evaluation. The stereo matching algorithms adopted in

Table 1: The calibration performed under different 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)	(0.2,0.25)

Table 2: The average minimum costs of different methods (SGBM, BM, MC-CNN).

	SGBM	BM	MC-CNN
doll	120.236	20.92	1.1056
toy brick	120.6858	21.78	1.0581
toy brick and cup	126.8454	19.55	0.9194
toy brick and lamp	113.8195	16.643	1.3634

our framework for performance comparison are BM, SGBM and 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 datasets. The parameter α is fixed as 1, and β is 20, 10 and 0.5 for SGBM, BM and MC-CNN, respectively. The cost aggregation are tabulated in Table 2, and the results from different methods are shown in Tables 3–5. We choose the 'Q' Middlebury dataset as the input and the evaluation method are BPR1.0.

In the cost function Equation (6), we only test CT and NCC because the image intensity will be shifted by the camera lens change and auto-exposure. The results from our framework are not very significant when adopted to MC-CNN. This might be due to 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 like SGBM. Figure 7 shows the results with or without the SGBM process. There are clear differences mainly because the basic algorithm cannot provide the suitable candidates.

Figure 8 shows the result comparison of several algorithms evaluated in this paper. For a fair comparison, we choose the same input size and remove some machine learning based methods. We choose SGBM+Dis as the result for comparison. The overall performance of the framework is robust in the whole dataset although some of them are not very significant mainly due to the different illumination conditions of the left and right images.

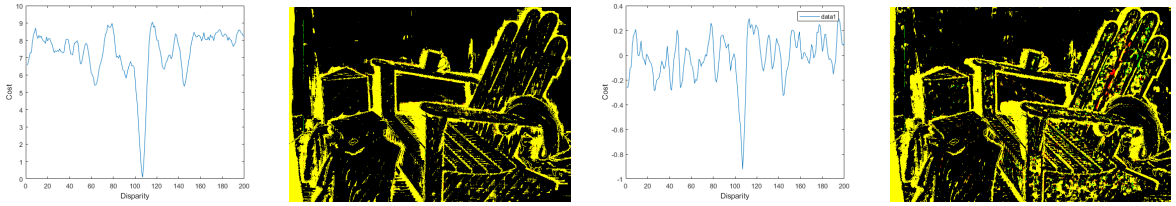
3.1 Optical Zoom Image Dataset

We use zoom lens cameras to construct a new dataset with ground truth. The evaluation method using our dataset is changed to BPR2.0 because the ground truth labeled manually is not very precise. We test the zoom correspondence methods mentioned previously, and test the effect of the candidate quantity. The toy brick+BM result appears very strange. When we calculate the BPR with zoom2 disparity map in BM, its value is near to 80%. Thus, we believe that the two initial disparity maps must have a certain level of correctness.

We then increase the number of disparity candidates to test the relationship between the result and the candidate quantity. In the disparity candidates, some algorithms such as SGBM cannot just sort the cost to choose candidates, so only MC-CNN and BM are tested. Figure 9 shows that if we increase the number of disparity candidates with no constraints, the results will be unpredictable. Thus, we add a new constraint that the correspondence pair only takes the candidates in the top three, and the results are better improved. MC-CNN encounters a problem that we cannot make sure if the top three in the cost are the same as the required candidates. The cost curve shown in Figure 7 only indicates the best disparity point. If the point is removed the whole line will look like a horizontal line with noise. With this test, it can be sure that the overall framework is stable if the number of candidate is sufficient.

4 CONCLUSION

In this work, a stereo matching framework using zoom images is proposed. With zoom image pairs, we are able to reduce the error and the uncertain region in the disparity map. Compared to the existing stereo matching algorithms, our approach can improve the disparity results with less computation. The proposed framework can adapt to the existing local and global methods for stereo matching, even the machine learning based matching algorithms. In the future work,



(a) MC-CNN+SGBM (b) Result (c) MC-CNN (d) Result

Figure 7: MC-CNN with different processes. The improvement is shown in red, and the green part indicates the incorrect change region.

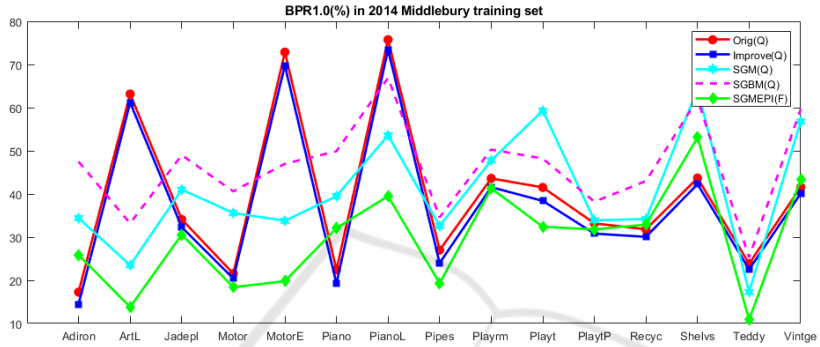


Figure 8: The result comparison of different approaches. Here we use SGBM+Dis as our result and 'Q' means that we use the quarter size image as input. SGM and SGBM are different with their basic cost functions, SGM is CT and the other is SAD. The input of SGMEPI is in full size.

Table 3: SGBM + Equation (9).

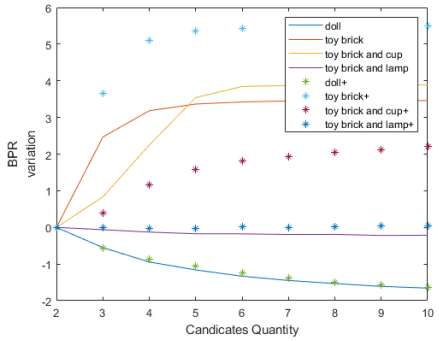
	Original	Homography	Epipole Geometry	Zoom calibration
doll	19.6503	**15.2552**	15.5919	15.801
toy brick	14.5528	13.3947	13.6271	**13.3372**
toy brick and cup	9.8251	9.2016	9.5835	**9.0291**
toy brick and lamp	19.6457	17.6351	**17.5975**	17.804

Table 4: BM + Equation (9).

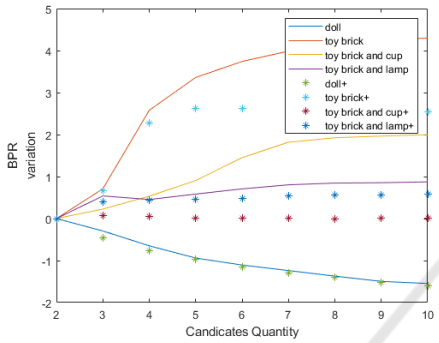
	Original	Homography	Epipole Geometry	Zoom calibration
doll	25.5569	**22.3145**	22.3699	23.0199
toy brick	**15.1138**	15.7089	15.7333	15.5487
toy brick and cup	12.797	12.1638	**11.7294**	12.1277
toy brick and lamp	26.4771	25.8682	**25.2747**	25.6811

Table 5: MC-CNN (No SGBM) + Equation (9).

	Original	Homography	Epipole Geometry	Zoom calibration
doll	22.3460	**20.2321**	21.0419	20.4651
toy brick	**18.9041**	25.8092	21.7056	24.3230
toy brick and cup	**11.8306**	12.9873	12.8016	12.9491
toy brick and lamp	23.2376	22.8388	**22.4940**	22.9996



MC-CNN (No SGBM)



BM

Figure 9: The effect of the candidate quantity, '+' means adding the constraint about the candidate order.

more investigation will be carried out to aggregate the information for machine learning methods and the cost computation.

REFERENCES

Batsos, K., Cai, C., and Mordohai, P. (2018). Cbmv: A coalesced bidirectional matching volume for disparity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2060–2069.

Chang, Q. and Maruyama, T. (2018). Real-time stereo vision system: a multi-block matching on gpu. *IEEE Access*, 6:42030–42046.

Chen, Y., Zhuo, B., and Lin, H. (2018). Stereo with zooming. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2224–2229.

Cheng, K. and Lin, H. (2015). Stereo matching with bit-plane slicing and disparity fusion. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 341–346.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.

Hosni, A., Bleyer, M., Rhemann, C., Gelautz, M., and Rother, C. (2011). Real-time local stereo matching using guided image filtering. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE.

Jie, Z., Wang, P., Ling, Y., Zhao, B., Wei, Y., Feng, J., and Liu, W. (2018). Left-right comparative recurrent model for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3838–3846.

Joshi, M. V., Chaudhuri, S., and Panuganti, R. (2004). Super-resolution imaging: use of zoom as a cue. *Image and Vision Computing*, 22(14):1185–1196.

Kim, K.-R. and Kim, C.-S. (2016). Adaptive smoothness constraints for efficient stereo matching using texture and edge information. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3429–3433. IEEE.

Lowe, D. G. et al. (1999). Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157.

Park, M.-G. and Yoon, K.-J. (2019). As-planar-as-possible depth map estimation. *Computer Vision and Image Understanding*.

Scharstein, D. and Szeliski, R. (2002). Middlebury stereo vision page. *http://vision.middlebury.edu/stereo*.

Seki, A. and Pollefeys, M. (2017). Sgm-nets: Semi-global matching with neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6640–6649.

Žbontar, J. and LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318.

Zhang, S., Xie, W., Zhang, G., Bao, H., and Kaess, M. (2017). Robust stereo matching with surface normal prediction. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2540–2547. IEEE.