Baltic J. Modern Computing, Vol. 10 (2032), No. 2, pp. 251–265
https://doi.org/10.22364/bjmc.2022.10.2.11

Fusing Census and MC-CNN Cost Volumes for Stereo Matching

Sahim Giray KIVANC[1], Baha SEN[1], Ali Ozgun OK[2], Fatih NAR[1]

[1] Ankara Yıldırım Beyazıt University, Ankara, TR
[2] Hacettepe University, Ankara, TR

gkivanc@ybu.edu.tr, bsen@ybu.edu.tr,ozgunok@hacettepe.edu.tr,
fnar@ybu.edu.tr

Abstract. Stereo matching is an important and popular field of computer vision. Numerous researchers worldwide are devoted to enhancing the effectiveness of stereo matching applications. In stereo matching, determining the costs of matching is a critical step. This step generates a cost volume that quantifies the similarity of pixels, and thereafter, it is processed further to generate the final disparity map. The purpose of this study is to improve stereo matching performance by fusing two different cost-volumes, namely Census and MC-CNN. The Census transform and Hamming distance are one of the most frequently used cost functions in conventional approaches. Besides, a matching cost volume generated using a deep learning technique called MC-CNN has been shown to extract more reliable features from images than conventional approaches. Thus, both of these cost computation strategies have a number of advantages and disadvantages. By including deep learning as a cost-volume, the advantages of these two distinct cost-volumes complement one another, resulting in a better cost-volume prior to applying the smoothing operation (e.g. Semi-global matching or More-global Matching). Our findings indicate that fusing cost-volumes improves the stereo matching performance of nearly all of the benchmark stereo images we tested in the Middlebury dataset.

Keywords: Stereo Matching, SGM, MGM, Census transform, MC-CNN, Deep Learning

1 Introduction

Stereo matching is the process of determining the disparity value of a pixel that corresponds to a matching pixel in two or more rectified images. The horizontal displacement between matching points in two rectified images is defined as the *disparity map*. Stereo matching can be used in a wide range of applications to estimate depth or height maps, and it is preferred in a variety of applications such as aerial imaging (Liu et al., 2018; Kukkonen et al., 2019), autonomous driving (Yang et al., 2019; Peng et al., 2020), and robotic vision (Samadi et al., 2016; Ma et al., 2020).

Stereo matching can be broadly grouped into two approaches: local and global (Scharstein and Szeliski, 2002). Local approaches compute the disparity value by comparing the features of two pixels in the left and right images by constructing a patch around the pixel. Each pixel is subjected to a disparity calculation. Local approaches are computationally efficient and are therefore suitable for real-time applications. However, the resulting disparity maps are less accurate than those produced by global approaches. Global approaches formulate the disparity calculation as a problem of energy minimization. While this approach produces more accurate disparity maps, it is significantly slower to compute than local approaches. As a result, global approaches are unsuitable for real-time applications. Semi-global Matching (SGM) (Hirschmuller, 2007) is a technique proposed to address the challenges associated with local and global approaches. SGM is indeed an integration of the two approaches, and while SGM is slightly less accurate than global approaches, its execution time is significantly faster. Because SGM is used in this study during the aggregation step, it will be discussed in detail in the following section. An improved version of SGM called More-Global Matching (MGM) (Facciolo et al., 2015) is also preferred as a smoothing operator in this study. MGM will be discussed in detail in the following section as well.

A stereo matching algorithm consists of four parts. (i) matching costs computation, (ii) cost aggregation, (iii) disparity computation, and (iv) disparity refinement. A matching cost quantifies the similarity between pixels and is a function of (x, y and d). The image pixel coordinates are denoted by x and y and the disparity value is symbolized by d. Normalized Cross Correlation (NCC), Sum of Square Differences (SSD), Sum of Absolute Differences (SAD) (Poggi et al., 2021), the Census transform and Hamming Distance are amongst the widely used matching costs. The Census transform combined with the hamming distance is a well-known superior matching algorithm compared to other algorithms (Zabih and Woodfill, 1994). Because the census transform is independent of intensity values; as a result, it is invariant under monotonic illumination variations. It is also possible to use deep learning in the process of calculating matching costs. (Zbontar and LeCun, 2016) proposed an MC-CNN strategy for fast deep learning strategy. MC-CNN learns the similarity between image patches by training a convolutional neural network (CNN). The Siamese network (Bromley et al., 1993) was used as the base architecture. The cost volume generated by MC-CNN can be used as an input to a smoothing operation as proposed in SGM. The MC-CNN is discussed in detail in the following section.

The concept of combining multiple matching cost volumes to create a solid cost volume has been extensively studied in the past. In (Jiao et al., 2014), the authors proposed a combined cost function with a modified color census transform and truncated color and gradient absolute differences.Those cost functions were then multiplied by various weight values. By adjusting those weights, the cost function's contribution to the final result was controlled. According to the authors, their proposed algorithm ranked second in the Middlebury dataset evaluation at that time. Another study utilized a weighted sum to combine different cost volumes was presented by (Miron et al., 2014). That study combined a variant of the census transform called the cross-comparison census (Miron et al., 2012) with the mean sum of relative intensity differences. Another study (Zhan et al., 2015), proposed calculating the combined cost function using the weighted sum

of the absolute difference in image colors, the absolute difference in image gradients in both directions and the lightweight census transform of the intensity image. These functions were added using a weighted sum. Additionally each of these functions was truncated to ensure that none of them dominated the final result. In (Yang et al., 2014), the weighted sum of the truncated version of the Birchfield and Tomasi measure (Birchfield and Tomasi, 1998) and the truncated version of the gradient map's absolute difference were proposed. Another study which proposed a combined matching cost was presented in (Jeon et al., 2018). Four matching costs were primarily considered (Absolute difference summation (SAD), normalized cross correlation with zero mean, the Census transform, and Sum of Gradient Differences). In addition to the four matching costs, combinations of these matching costs were tested. These costs were combined with a single variable α. When the combined matching costs were added together, the total number of matching costs was 31. A linear equation is constructed from different cost volumes in (Shetty et al., 2020). That study adopted NCC, normalized mutual information (NMI) and SAD, and a parameter was used to merge SAD and NMI. Thereafter, the resulting matching cost from those two costs was combined with the NCC by adding the two matching costs and dividing by two. In another study, SAD and the Census transform were used to form a linear weighted matching cost with adaptive weights (Chai and Cao, 2018). The weights used in the combination process reflected the importance of the matching cost in the current support domain. Another study handled an exponential function to combine different matching costs (Stentoumis et al., 2014). The Census transform on image gradients was combined with the absolute difference in color and image gradients. The combination process was based on a robust exponential, and by utilizing the exponential combination, the study prevented the final cost from being dominated by large outlying cost penalties. Another study that favored the use of exponential function appeared in (Hamzah et al., 2017). Through the use of a weighting parameter, the Absolute difference (AD) and gradient matching (GM) were combined in a linear fashion. This output was then combined with the census transform via the exponential function. Enhanced image gradient-based cost and improved census transformation-based cost were also combined with the use of an exponential function in (Liu et al., 2020). A combination of AD, absolute gradient difference (AGD), and gradient-based census transform with the exponential function was presented in (Ma et al., 2017). The authors have achieved the second highest score in the Middlebury set. The research article (Hamid et al., 2021) was the most comparable one to our study. The cost associated with the MC-CNN was combined with the difference in directional intensity. A parameter (λ) was used to combine these two matching costs. According to the authors of that study, when tested on the Middlebury dataset their proposed method outperformed all other methods in terms of accuracy.

The census transform and the cost volume from MC-CNN are convolved in this study to generate a more accurate matching cost; and thus, improve stereo matching performance. Both of these cost volumes have their own set of advantages and disadvantages to consider. The Census transform does not depend on the actual intensity values instead it uses the ordering of the intensity values. This feature makes the census transform invariant with respect to monotonic variations of illumination. However, since it depends on the window size of a filter, it may fail to extract features in some

situations. MC-CNN can learn features better; however, it may fail on images outside of the domain of training set. Our goal is to build a new cost volume that can take advantage of the benefits of these cost volumes while also reducing their uninvited effects. In the previous studies on cost volume combination, the combination process is done through summation. In this study, cost volumes are combined using a multiplication process which adds nonlinearity to the solution. The rest of the paper is organized as follows. In Chapter 2 both of the previously used and the newly formed matching costs as well as two cost aggregation methods (SGM and MGM) will be discussed in detail. The results from the experiments will be presented in Chapter 3. Chapter 4 will discuss the results and Chapter 5 will conclude the study.

2 Proposed Fusing Method

In this section, two topics will be mentioned. Firstly SGM and MGM, cost aggregation algorithms used in this study, will be explained in detail. Secondly, the census transform and MC-CNN matching costs will be explained, then our proposed matching cost MC-Census will be explained in detail.

2.1 Cost Aggregation Methods

2.1.1 Semi-global Matching (SGM). In this study SGM is used in cost aggregation process. In order to achieve effective stereo matching performance SGM adds two penalties, P_1 and P_2 to aggregate the cost volume (Hirschmuller, 2007). The following is a mathematical expression of SGM:

$$E(D) = \sum_{p}(C(p, D_p) + \sum_{q \in N_p} P_1 T[|D_p - D_q| = 1] + \sum_{q \in N_p} P_2 T[|D_p - D_q| > 1]) \quad (1)$$

In Equation 1, the term $C(p, D_p)$ represents the sum of all pixelwise matching costs for the disparities of D. The second term represents the penalty P_1 that adds a constant penalty for all pixels q in neighborhood of N_p of p if the difference between p and q is 1. The third term represents a larger constant penalty P_2. This penalty is applied if the difference between p and q is greater than 1. SGM doesn't aggregate the cost along all dimensions to reduce the execution time. SGM constructs a star shaped 8-path or 16-path structure around the aggregated pixel and the final cost is aggregated along these paths. For example, 8-path structure can be seen in Figure 1.

2.1.2 More Global Matching (MGM). More Global Matching (MGM) was also utilized as a cost aggregation step in this research (Facciolo et al., 2015). The disadvantage of the SGM approach is that while aggregating in any direction, just the value of the previous pixel is considered. If the information is ambiguous while aggregating horizontally, the final result may be dominated by the information from the vertical aggregation. As a result, the penalty terms may lose their effectiveness. MGM has been proposed to address some of the shortcomings of the SGM approach. While SGM utilizes only the

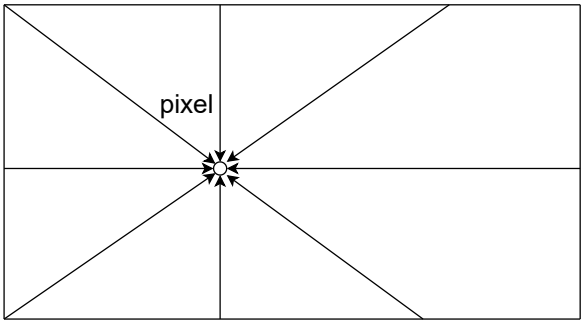

Fig. 1: Star shape structure of SGM

information from the previous pixel during the aggregation step, MGM utilizes information from both the previous and upper pixels (Figure 2b). Figure 2 depicts the outline of the MGM algorithm. MGM can be expressed in the following manner:

$$L_r(p, d_p) = C_p(d_p) + \sum_{x \in r, r'} 1/2 \min d_q \in D(L_r(p - x, d_q) + V(d_p, d_q)) \quad (2)$$

In Equation 2, x represents the neighboring pixels, r represents the left neighbor, r' represents the upper neighbor. $V(d_p, d_q)$ can be represented as in Equation 3. In addition to left and upper neighboring pixels, upper-left neighboring pixel can also be added to the summation.

$$V(d_p, d_q) = \begin{cases} 0 & if d_p = d_q \\ P1 & if |d_p - d_q| = 1 \\ P2 & if |d_p - d_q| > 1 \end{cases} \quad (3)$$

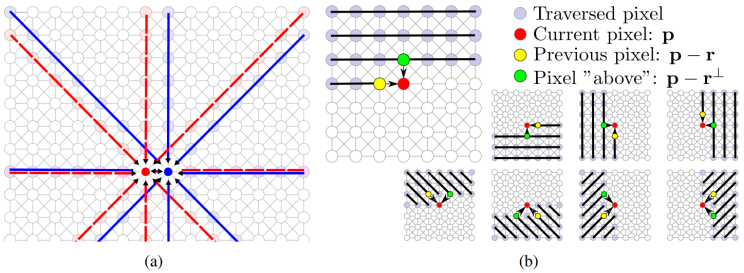

Fig. 2: a) Star shaped structure of two neighboring pixels. b) Cost aggregation of MGM along the paths r and r' (Facciolo et al., 2015)

2.2 Matching costs

2.2.1 The Census transform and Hamming Distance. There are various approaches for computing the pixel-based matching costs. The Census transform and hamming distance is one of the most popular ones. The Census transform associates each pixel with a binary string, encoding whether the pixel have smaller intensity values than each of its neighbors, one for each bit. Thereafter, the distance between two transformed images can be computed by hamming distance. The process of the census transformation and hamming distance can be seen in Figure 3. The mathematical formulation of the census transformation can be expressed as below:

$$f(p,p') = \begin{cases} 0 & if p > p' \\ 1 & if p <= p' \end{cases} \tag{4}$$

In Equation 4, p represents the center pixel of a patch and p represents the neighboring pixels. After the transform, a binary string is formed. Finally, hamming distance is used to calculate the similarity between two binary strings.

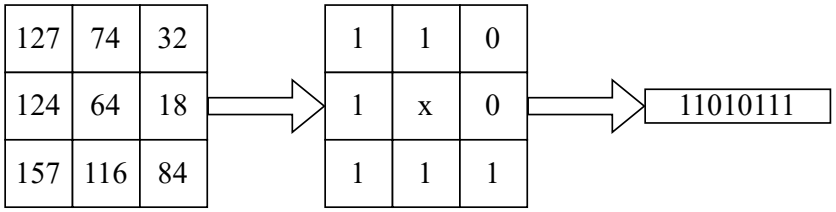

Fig. 3: An example of the census transform and hamming distance on a 3x3 patch.

2.2.2 MC-CNN. Deep learning has become an increasingly important component of the stereo matching process. A fast deep learning strategy was proposed by (Zbontar and LeCun, 2016). MC-CNN uses Siamese network architecture. Siamese networks are trained to learn the similarity between images and are mostly used in areas such as fingerprint comparison, signature fraud detection, etc. MC-CNN utilizes small image patches to train a convolutional neural network (CNN) on a similarity measure. The subnetworks output a vector containing the features of the input patch. The output vectors are then compared using the cosine similarity which gives the final output of the network. A hinge loss is used to train the network. Pairs of examples are used in the loss function. These examples are centered around the same image position where one example belongs to the positive class and one to the negative class. Positive example is generated using GT information. Negative information is generated by shifting the positive example to either the left or right by one pixel. The loss is zero when the similarity of the positive example is greater than the similarity of the negative example by at least a certain margin. The weights are shared between the networks. The CNN's generated cost volume is then forwarded to SGM for further processing. The MC-CNN algorithm

generates a cost volume with values between [-1,+1]. The lower the value, the greater the similarity between two pixel patches in the image. Figure 4 illustrates the MC-CNN fast architecture for cost computation.

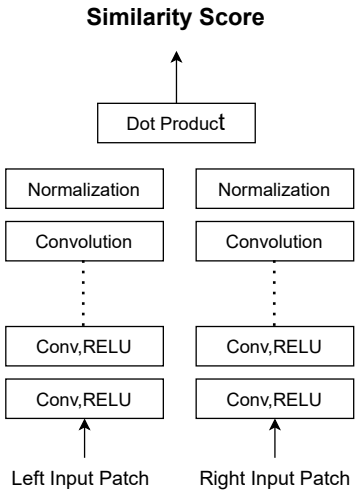

Fig. 4: The architecture of the MC-CNN model (Facciolo et al., 2015)

2.2.3 The Subset Selection. In order to select which matching cost combination to use, we carried out some experiments on a set of different matching costs. Seven matching cost functions are selected to form a set. Birchfield and Tomasi dissimilarity, the census transform and hamming distance, gradient direction, gradient magnitude, intensity features difference, MC-CNN, and the norm matching cost. These cost functions are selected based on the usage of each of them in the literature. On this set, we applied subset selection to form subsets containing different combinations of these matching costs. Since there are seven matching costs the total of number of 127 combinations are formed. These combinations can be calculated as $2 * 2 * 2 * 2 * 2 * 2 * 2$ - 1. In each of these combined cost volumes, the costs are combined using the technique proposed in this study. 127 combined cost volumes applied on all of the test images. The results are then given to SGM. On all of the test images, 4 specific combinations of the cost functions outperformed the other combinations. These combinations are, the census (alone), the census and MC-CNN, the census and Birchfield and Tomasi, the census and MC-CNN and Birchfield and Tomasi. Among these four combinations, the census combined with MC-CNN outperformed the other combinations, as a good trade-off between accuracy and computational complexity.

2.2.4 Proposed Method Both of the previously mentioned cost volumes have advantages and disadvantages. The census transform is advantageous for mitigating the

effects of monotonic illumination variations. Image features can be extracted in a more efficient manner by using MC-CNN, and this may lead to a higher degree of similarity score. The Census transform cost volume values range between zero and the maximum length of the binary string. Zero denotes the closest distance (identical), while the maximum length denotes the greatest distance (dissimilar), and MC-CNN outputs a cost volume with values between [-1, +1]. The goal of this study is to convolve cost volumes into a single volume that can be used in cost aggregation. Equation 5 describes the fusion process.

$$G = F \circ H \tag{5}$$

where G is obtained cost volume, F is MC-CNN cost volume, H is the Census cost volume, and $\circ$ is element-wise product operation, Hadamard Product, between cost volume F and cost volume H. Before the fusion, MC-CNN cost volume is min-max normalized between 0 and 1 so that the obtained cost volumes have the positive values. The negative values in obtained cost volume may cause errors in penalty application process in SGM. The disparity signals for the three cost volumes from three different images can be seen in Figure 6. It's easier to grasp the goal of our method after looking at Figure 6. In Figure 6, Adirondack column the Census matching cost has two minimum points. Compared to that MC-CNN have only one. In this scenario, the census transform may output the incorrect disparity value because of multiple minimum points. The same scenario may apply to MC-CNN as well. These two matching costs may produce correct matching scores in different regions. Our proposed matching cost combines the advantages of the census transform and MC-CNN matching costs. The advantage of our proposed method can be explained better with disparity plots in Figure 5. If one of the cost volumes has a minimum point, then the point will remain in the fused cost volume. If both of the cost volumes have the correct minimum point, then the fused minimum point will be even smaller. Multiple minimum points may exist in one of the cost volumes, if either one of them has a minimum point, the fused cost volume will have reduced number of minimum points. If two different cost volumes have different minimum points, then fused cost volume will have a minimum point closer to GT. The fused cost volume will have a smoother and smaller cost volume.

3 Results and Discussion

In this section, the numerical and visual results using three cost volumes, the census transform, MC-CNN and our proposed method MC-Census are given and discussed in detail. BAD 2.0 metric is preferred when comparing the numerical results because it is one of the conventional metrics used in Middlebury data set evaluation. BAD 2.0 metric compares the generated disparity map with the ground truth (GT) in areas specified by a mask. If the difference between disparity values in GT and generated disparity map is greater than 2 pixels, then the pixel is marked for error. Experiments were carried out on the images from the Middlebury set (Scharstein and Szeliski, 2003; Hirschmuller and Scharstein, 2007; Scharstein et al., 2014). Visual results comparing the three cost volumes with the left image from the image pair and the ground truth image can be seen

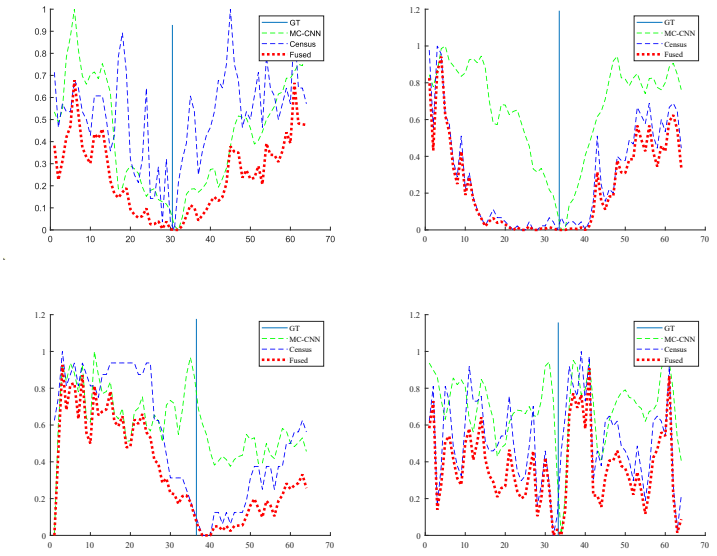

Fig. 5: Disparity signals for the three cost volumes for Teddy image.

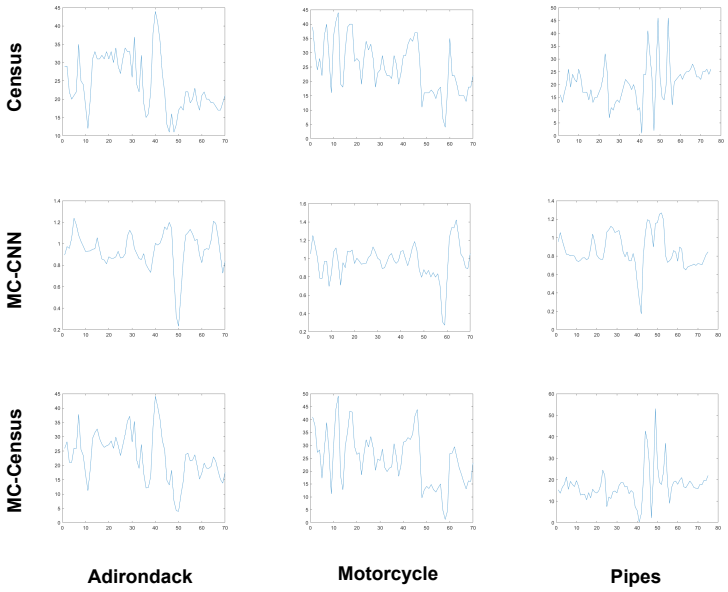

Fig. 6: Disparity signals for the three cost volumes for three different images. Pixel coordinates (x,y) are (371,219) for Adirondack, (196,385) for Motorcycle, (372,361) for Pipes.

Table 1: RESULTS FROM MGM AND SGM ACCORDING TO BAD METRIC. FOR EACH IMAGE, BEST VALUES ARE GIVEN IN BOLD.

	SGM			**MGM**		
Image Name	Census	MC	MC-Census	Census	MC	MC-Census
Adirondack	0.0624	0.0431	**0.0395**	0.0590	0.0455	**0.0388**
ArtL	**0.1172**	0.1190	0.1188	**0.1147**	0.1187	0.1172
Jadeplant	0.2317	0.1817	**0.1731**	0.2441	0.1879	**0.1806**
Motorcycle	0.0649	0.0510	**0.0479**	0.0634	0.0523	**0.0467**
Piano	0.1304	0.1061	**0.1026**	0.1264	0.1103	**0.1004**
Pipes	0.1038	0.0956	**0.0880**	0.1007	0.0946	**0.0872**
Playroom	0.1528	0.1246	**0.1066**	0.1368	0.1342	**0.1066**
Playtable	0.2160	0.1081	**0.0995**	0.1977	0.1122	**0.0999**
Recycle	0.0848	0.0696	**0.0632**	0.0821	0.0753	**0.0645**
Shelves	0.2849	**0.2155**	0.2234	0.2729	0.2158	**0.2092**
Teddy	**0.0553**	0.0657	0.0570	**0.0543**	0.0684	0.0578
Vintage	0.2602	**0.1982**	0.2081	0.2509	0.1953	**0.1943**

in Figure 7. In Figure 8 the generated error maps can be seen. Also, numerical results generated by SGM and MGM using the three cost volumes can be seen in Table 1.

We collect images from different Middlebury sets to conduct our experiments. Our collected set has 12 image pairs. Our method has been tested on all of these image pairs. Visual results are given in Figure 7. When the visual results are examined, it can be seen that our proposed method generates more accurate disparity maps. The error maps generated from the experiments are also given in Figure 8. The error maps are in binary scale. White areas represent the falsely classified pixels. Black areas represent the correctly classified pixels. The comparison of different matching costs can be evaluated better on error maps. The first row of Figure 8 is an image called ”Adirondack”. When the error map of the census transform is examined, it can be seen that the errors are concentrated around the chairs arm and coffee mug. The census cost seems to perform poorly in these regions. However, MC-CNN seems to have better accuracy around the coffee mug. Nevertheless, MC-CNN seems to perform poorly on the edges of the book and the upper part of the chair. Most of the errors are grouped in those areas. However, the Census transform performs better than MC-CNN around the edges of the book and the upper part of the chair. Another example can be seen in the 9th row of the Figure 8. This image is called ”Recycle”. When the error map from the census transform is examined, on the one hand, the census transform performs poorly on the textureless regions on the recycling bin but achieves good accuracy on the edges of curtains. On the other hand, MC-CNN performs better on the textureless regions on the recycling bin but achieves lower accuracy than the census transform on the edges of the curtains. As a result, after proposed fusing strategy, the advantages of these two cost volumes are combined. Around the aforementioned regions, the proposed method outperforms both the results achieved by the census transform and MC-CNN. In addition to visual results, the numerical results from our experiments are presented in Table 1. The numerical results corroborate our visual observations. We observed that proposed cost volume,

Fig. 7: Disparity maps generated with MGM using cost volumes Census, MC-CNN, MC-Census starting from second column and ends in fourth column left to right. Fifth column is the GT values.

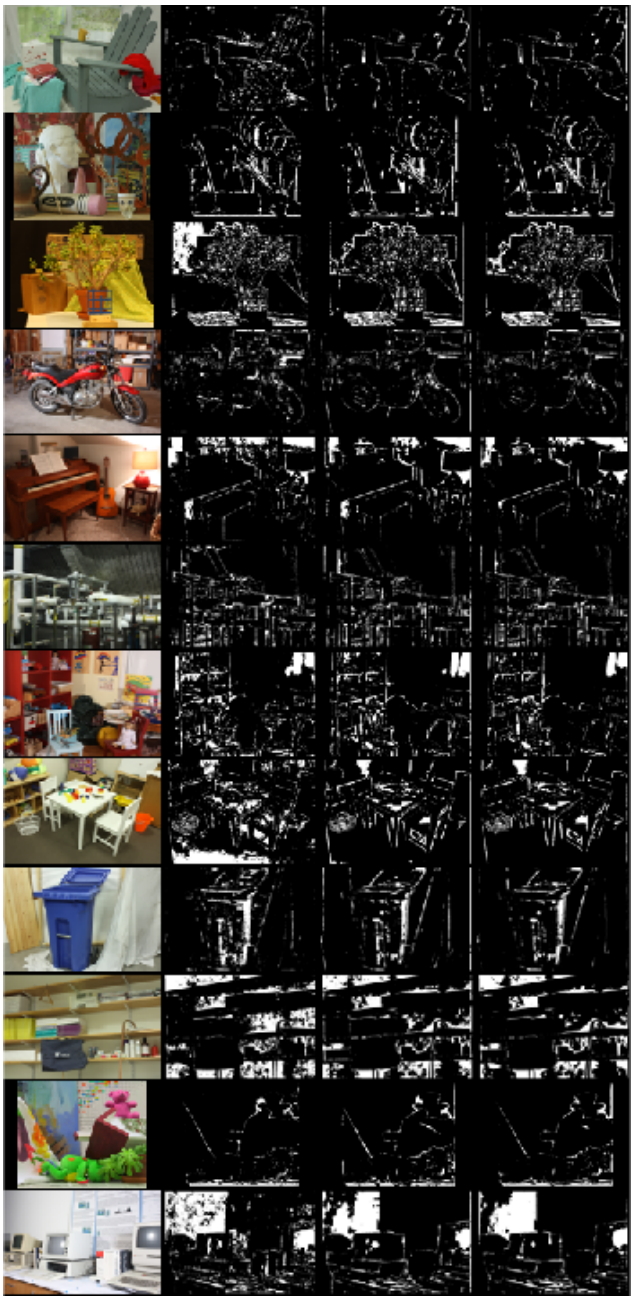

Fig. 8: Error maps generated with MGM using cost volumes Census, MC-CNN, MC-Census starting from second column and ends in fourth column left to right (White pixels represent falsely classified pixels, black pixels represent correctly classified pixels).

MC-Census, performs better than both the census transform and MC-CNN. When the numerical results are examined, it can be seen that MC-Census achieved better accuracy than the other two matching costs on most of the subsets. While using MGM as a smoothing process, only on ArtL and Teddy images that the census transform achieved better accuracy but only by a small margin. However, when other image subsets are examined, it can be seen that MC-Census achieved better scores than the other two. On the small portion of the subset (Jadeplant, Vintage), the accuracy gain is small, but on the large portion of the subsets accuracy gain is quite high. We can also deduce from Table 1 that aggregation method is also important in stereo matching. It can be seen that MGM improves the overall stereo matching performance. As a result, our proposed matching cost MC-Census performs better while using MGM as a smoothing operator.

The studies described in the introduction section use several types of aggregating methods and conduct a variety of post-processing processes. Some of the research is also done on the earlier version of the Middlebury collection. Because the goal of our research is to generate a new cost volume, alternative strategies can be employed as an aggregation algorithm. This may boost or decrease the algorithm's accuracy. As a result, a direct accuracy comparison between those studies seems to be impractical. However, when the overall accuracy results from our trials are compared to the findings from the research described in the introduction section, it is clear that our method obtains comparable accuracy results to the mentioned methods.

4 Conclusion

In this study, a better cost volume is formed by fusing the census transform and MC-CNN cost volumes. The results show that, our cost volume performs better on almost all of the images. On some images, MC-Census appears to reduce the error slightly, while on others, the error appears to be significantly reduced. On image sets like Adirondack, Motorcycle, Pipes, Playroom, Playtable, and Recycle, it can be seen that the error is reduced by 0.1 or 0.2. Post-processing techniques can be used to improve the overall accuracy of all matching costs. However, no post-processing technique is used in this study in order to maintain the clarity of the final results. Due to the fact that the proposed strategy has to calculate two matching cost volumes instead of one, the execution time eventually increases. Nevertheless, fusing the two cost volumes does not require significant additional time.

The number of combined matching costs can be increased in order to get a better combination in future studies. This may lead to an increase in accuracy since increasing the cost volumes will increase the chance of achieving a better cost volume. However, increasing the number of matching costs will also increase the execution time of the algorithm since the number of combinations between cost volumes will increase, so the trade off between accuracy gain and execution time must be handled well. Fusion techniques other than multiplication can be employed to examine the impact of different methods on the final result.

References

Birchfield, S., Tomasi, C. (1998). A pixel dissimilarity measure that is insensitive to image sampling, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(4), 401–406.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R. (1993). Signature verification using a" siamese" time delay neural network, *Advances in neural information processing systems* **6**.

Chai, Y., Cao, X. (2018). Stereo matching algorithm based on joint matching cost and adaptive window, *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, IEEE, pp. 442–446.

Facciolo, G., De Franchis, C., Meinhardt, E. (2015). Mgm: A significantly more global matching for stereovision, *BMVC 2015*.

Hamid, M. S., Manap, N., Hamzah, R. A., Kadmin, A. F. (2021). Stereo matching algorithm based on hybrid convolutional neural network and directional intensity difference, *Artificial intelligence (AI)* **14**, 16.

Hamzah, R. A., Ibrahim, H., Hassan, A. H. A. (2017). Stereo matching algorithm based on per pixel difference adjustment, iterative guided filter and graph segmentation, *Journal of Visual Communication and Image Representation* **42**, 145–160.

Hirschmuller, H. (2007). Stereo processing by semiglobal matching and mutual information, *IEEE Transactions on pattern analysis and machine intelligence* **30**(2), 328–341.

Hirschmuller, H., Scharstein, D. (2007). Evaluation of cost functions for stereo matching, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1–8.

Jeon, H.-G., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.-W., Kweon, I. S. (2018). Depth from a light field image with learning-based matching costs, *IEEE transactions on pattern analysis and machine intelligence* **41**(2), 297–310.

Jiao, J., Wang, R., Wang, W., Dong, S., Wang, Z., Gao, W. (2014). Local stereo matching with improved matching cost and disparity refinement, *IEEE MultiMedia* **21**(4), 16–27.

Kukkonen, M., Maltamo, M., Korhonen, L., Packalen, P. (2019). Comparison of multispectral airborne laser scanning and stereo matching of aerial images as a single sensor solution to forest inventories by tree species, *Remote Sensing of Environment* **231**, 111208.

Liu, H., Wang, R., Xia, Y., Zhang, X. (2020). Improved cost computation and adaptive shape guided filter for local stereo matching of low texture stereo images, *Applied Sciences* **10**(5), 1869.

Liu, J., Ji, S., Zhang, C., Qin, Z. (2018). Evaluation of deep learning based stereo matching methods: From ground to aerial images., *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* **42**(2).

Ma, N., Men, Y., Men, C., Li, X. (2017). Segmentation-based stereo matching using combinatorial similarity measurement and adaptive support region, *Optik* **137**, 124–134.

Ma, W.-P., Li, W.-X., Cao, P.-X. (2020). Binocular vision object positioning method for robots based on coarse-fine stereo matching, *International Journal of Automation and Computing* **17**(4), 562–571.

Miron, A., Ainouz, S., Rogozan, A., Bensrhair, A. (2012). Cross-comparison census for colour stereo matching applied to intelligent vehicle, *Electronics letters* **48**(24), 1530–1532.

Miron, A., Ainouz, S., Rogozan, A., Bensrhair, A. (2014). A robust cost function for stereo matching of road scenes, *Pattern Recognition Letters* **38**, 70–77.

Peng, L., Deng, D., Cai, D. (2020). Geometry-based occlusion-aware unsupervised stereo matching for autonomous driving, *arXiv preprint arXiv:2010.10700* .

Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K., Mattoccia, S. (2021). On the confidence of stereo matching in a deep-learning era: a quantitative evaluation, *arXiv preprint arXiv:2101.00431* .

Samadi, M., Othman, M. F., Talib, M. F. (2016). Fast and robust stereo matching algorithm for obstacle detection in robotic vision systems, *Jurnal Teknologi* **78**(6-13).

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth, *German conference on pattern recognition*, Springer, pp. 31–42.

Scharstein, D., Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *International journal of computer vision* **47**(1), 7–42.

Scharstein, D., Szeliski, R. (2003). High-accuracy stereo depth maps using structured light, *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 1, IEEE, pp. I–I.

Shetty, A. A., George, V., Nayak, C. G., Shetty, R. (2020). Multiple data cost-based stereo matching method to generate dense disparity maps from images under radiometric variations, *International Journal of Intelligent Systems Technologies and Applications* **19**(4), 393–404.

Stentoumis, C., Grammatikopoulos, L., Kalisperakis, I., Karras, G. (2014). On accurate dense stereo-matching using a local adaptive multi-cost approach, *ISPRS Journal of Photogrammetry and Remote Sensing* **91**, 29–49.

Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B. (2019). Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 899–908.

Yang, Q., Ji, P., Li, D., Yao, S., Zhang, M. (2014). Fast stereo matching using adaptive guided filtering, *Image and Vision Computing* **32**(3), 202–211.

Zabih, R., Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence, *European conference on computer vision*, Springer, pp. 151–158.

Zbontar, J., LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches, *Journal of Machine Learning Research* **17**, 1–32.

Zhan, Y., Gu, Y., Huang, K., Zhang, C., Hu, K. (2015). Accurate image-guided stereo matching with efficient matching cost and disparity refinement, *IEEE Transactions on Circuits and Systems for Video Technology* **26**(9), 1632–1645.

Received May 27, 2022 , accepted June 17, 2022