ACCURATE DENSE STEREO MATCHING FOR ROAD SCENES

Oussama Zeglazi, Mohammed Rziza, Aouatif Amine, Cédric Demonceaux

ACCURATE DENSE STEREO MATCHING FOR ROAD SCENES

Oussama Zeglazi [(a)]*, Mohammed Rziza* [(a)]*, Aouatif Amine* [(b)]*, Cédric Demonceaux*[(c)]

[(a)] LRIT URAC 29, Faculty of Sciences, Mohammed V University, B.P. 1014, Rabat, Morocco
[(b)] LGS, National School of Applied Sciences, Ibn Tofail University, B.P. 241, Kenitra, Morocco
[(c)] Le2i, FRE CNRS 2005, Arts et Métiers, Univ. Bourgogne Franche-Comté, France

ABSTRACT

Stereo matching task is the core of applications linked to the intelligent vehicles. In this paper, we present a new variant function of the Census Transform (CT) which is more robust against radiometric changes in real road scenes. We demonstrate that the proposed cost function outperforms the conventional cost functions using the KITTI benchmark[1]. The cost aggregation method is also updated for taking into account the edge information. This enables to improve significantly the aggregated costs especially within homogenous regions. The Winner-Takes-All (WTA) strategy is used to compute disparity values. To further eliminate the remainder matching ambiguities, a post-processing step is performed. Experiments were conducted on the new Middlebury[2] dataset, as well as on the real road traffic scenes of the KITTI database. Obtained disparity results have demonstrated that the proposed method is promising.

Index Terms— Stereo vision, Census Transform, Cross Comparison Census, Cross based aggregation.

1. INTRODUCTION

Stereo matching is one of the most widely studied problems in computer vision. It has found various applications in the image processing domain e.g. $3D$ reconstruction, object detection and intelligent vehicles. Stereo matching algorithms described here typically operate on rectified images. They can be roughly classified into two categories : global and local approaches. Global approaches consider whole pixels in the image to produce depth values. Thus, these approaches usually produce quite accurate depth results. However, this kind of algorithms has a high complexity problem. Local approaches consider specific pixels in the image to estimate depth values. Thus, they are computationally cheap. However, they usually produce less accurate depth values. A wide range of stereo matching algorithms have been proposed and their performance has been examined in various surveys [1, 2]. A stereo matching algorithm can be performed in four major steps : cost computation, cost aggregation, disparity computation and disparity refinement [3]. In the first step, a matching cost is computed for each pixel at all possible disparity levels. Cost functions are either based on absolute intensity differences, squared intensity differences or normalized cross correlation. Yet, cost functions based on pixel intensities are sensitive to radiometric changes. Thus, other cost functions based on image transformations were developed in [4] such as the non-parametric rank and the census costs. These costs might be replaced or even merged with other ones in order to produce robust variants against radiometric changes, textureless areas or to regions with proximity to occlusion borders [5, 6, 7, 2]. Recently, Authors in [8] have proposed an adaptive fusion strategy of multiple cost matching functions. In the cost aggregation stage, costs are merged either by summing up or averaging over a predefined support region for each pixel. The commonly used shape of such support regions is rectangular window or its variations [9]. Fixed support region with varying weights according to many considerations as color similarity and distance to the center pixel [10]. Adaptive support regions with adjusting the size and shape of the window for each pixel[11]. In the case of the disparity optimization, an optimal disparity is computed for each pixel with local or global optimization approaches. For local-based methods, the WTA strategy is used. While, an optimized energy function, defined over all image pixels with some constraints, is performed in the case of the global approaches. A wide range of approaches have been developed based on the dynamic programming[12], belief propagation [13] and graph-cuts[14]. For disparity refinement many approaches have been studied. These approaches included scanline optimization, median filtering, subpixel estimation, mismatched area detection as well as interpolation [15, 11, 6].
In this paper, we propose a new cost function which deals better with radiometric changes. It is based on the CT cost function which has been recognized to be robust against such issues. One of most widely used aggregation technique [11] relies only on spacial proximity and color similarly to define the adaptive region which is turn out to be insufficient. Thus, in this paper, in order to enhance the quality of disparities, the aggregation cost is modified to further incorporate informa-

[1] http://www.cvlibs.net/datasets/kitti
[2] http://vision.middlebury.edu/stereo/data/

tion as edge information. Once these steps achieved, a post-processing stage is performed to remove any noise left.

2. THE PROPOSED METHOD

2.1. Cost Computation

Census Transform [4] has been extensively used in stereo matching algorithms regarding its robustness against radiometric differences. Recently, authors in [16] have developed a C_{CCC} cost function as variant of C_{CT} one. The CCC bit-string is obtained by comparing each pixel in the support window with those in the immediate neighborhood in a clockwise direction.

The C_{CCC} is computed through the Hamming distance (D_H) between a pixel $p = (x, y)$ in the reference image (I_1) and its hypothetical corresponding pixel $q = (x, y - d)$ in the target image (I_2) at a disparity d as follows :

$$C_{CCC}(x, y, d) = D_H(CCC_{I_1}(x, y), CCC_{I_2}(x, y - d)), \tag{1}$$

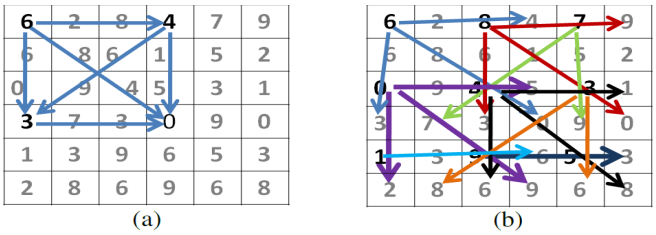
where the bit-string, CCC, is defined by:

$$CCC_I(u, v) = \otimes_{\substack{i=1:step:n \\ j=1:step:m}} (\xi(I(u+i, v+j), I(u+i^{'}, v+j^{'})), \tag{2}$$

where $n \times m$ represents the support window, $\otimes$ denotes the concatenation operation, and ξ function is defined as follows :

$$\xi(p, q) = \begin{cases} 1 & p \leq q \\ 0 & otherwise \end{cases} \tag{3}$$

In Eq.(2), $(i^{'}, j^{'}) = \{(i, j + step); (i + step, j + step); (i + step, j); (i + step, j - step)\}$, which represents the immediate neighborhood in clockwise direction of a pixel with (u, v) coordinates. Note that $(j + step) < m$, $(i + step) < n$ and $(j - step) \geq 0$. *step* is an empirical value, chosen to jump some pixels in the support window. Although the CCC cost enables to take into account the neighboring pixel information, it happens that in some cases the most significant information is located other than the pixels provided by the chosen step. In addition, the step value should be taken in order to not exceed the half size of the support window, otherwise the resulted bit-string might fail at providing a consistent bit-string for describing the considered patch. Figure 1(a) presents a typical case where the *step* value is set to 3 for a support window (5×5). Therefore, in the current paper, we propose to take two different step values. The first one $(step_1)$ is used to take one pixel and to skip it's neighbor while the second one $(step_2)$ is used for the comparison procedure. This latter is chosen to be more flexible in way that it looks for the pixel conveying the most significant information in wider manner. To more illustrate the proposed cost, figure 1(b) presents the proposed bit-string with $step_1$ is set to 2, and $step_2$ is set to 3. Moreover, since image gradients are less sensitive to radiometric changes and repetitive patterns [7], the proposed cost function is implemented on principal gradient $\delta I/\delta x$ and $\delta I/\delta y$ directions for the input image. Thus, the proposed bit-string is defined as follows :



(a) (b)

Fig. 1. Bit-string construction for original CCC: "000000" (a) and the proposed one: "000100111110111110" (b)

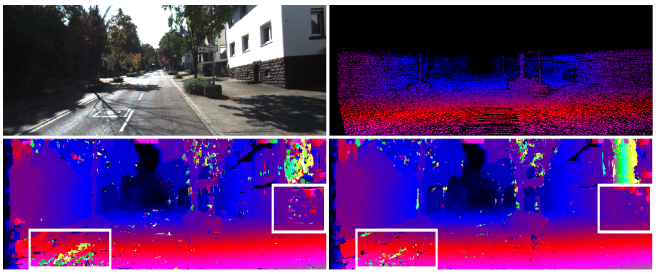


Fig. 2. From top to bottom and left to right : the left image, ground truth disparity map, disparity map computed based on C_{CCC} and disparity map using proposed cost

$$GCCC_I(u, v) = \otimes_{\{\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y}\}} \otimes_{\substack{i=1:step_1:n \\ j=1:step_1:m}} (\xi(\delta I(u + i, v + j), \delta I(u + i^{'}, v + j^{'})) \tag{4}$$

where $(i^{'}, j^{'}) = \{(i, j + step_2); (i + step_2, j + step_2); (i + step_2, j); (i + step_2, j - step_2)\}$. Similar to The C_{CCC} cost matching function, the proposed cost function C_{GCCC} is computed through the Hamming distance as follows :

$$C_{GCCC}(x, y, d) = D_H(GCCC_{I1}(x, y), GCCC_{I2}(x, y - d)) \tag{5}$$

The performance of the proposed cost function applied on #0 stereo pair from KITTI training dataset is depicted in figure 2 which denotes the improvement of disparity results, indicated areas, compared to the original C_{CCC}.

2.2. Cost Aggregation

Our attention is paid to the cross-based aggregation method as described by [11]. Thus, we propose to add the edge information as a supplementary criteria in the region construction strategy. According to study presented in [17], the Canny's edge detection method is the top performer technique. Therefore, this detector algorithm is retained in our study. Performing this edge detector for a given image provides a binary

image I_b. Then, for each pixel $p = (x, y)$, we determine its four arm lengths (i.e. the left, right, top and bottom) represented by $\{hp^+, hp^-, vp^+, vp^+\}$, separately. The algorithm proceeds in two steps, in the first step, for a pixel **p** its left arm respectively (right, up, bottom) stops when it finds an end point p_l that exceeds one of the following rules :

(i) $D_c(p_l, p) < \tau$, where $D_c(p_l, p)$ represents the color difference between p_l and p, and τ is a preset threshold value. $D_c(p_l, p) = max_{i=R,G,B}|I_i(p_l) - I_i(p)|$

ii) $D_s(p_l, p) < L$, where $D_s(p_l, p)$ represents the spatial distance between p_l and p, and L is a preset maximum length. $D_s(p_l, p) = |p_l - p|$.

iii) $I_b(p_l) = I_b(p)$, which represents the edge information between p and p_l.

The first two rules describe above set limitations on both color similarity and arm length with two thresholds τ and L. In the case of homogenous regions, where the color similarity is higher, the two above rules are insufficient to describe the local region structure. For this purpose, the edge information criteria is employed to define the adaptive region. In the second step, based on the arms length $\{hp^+, hp^-, vp^+, vp^+\}$ detected for a pixel p, two orthogonal cross segments, the horizontal segment $H(p)$ and the vertical one $V(p)$ are considered. Therefore, the adaptive shape support region $U(p)$ for the pixel p is given by:

$$U(p) = \bigcup_{q \in V(p)} H(q), \tag{6}$$

where q is the neighborhood pixel located on the vertical segment $V(p)$. The same process is performed to determine the support region $U^{'}(q)$ of the hypothetical corresponding pixel $q = (x, y - d)$ in the target. Then, the combined window involves the intersection of the two support regions as follows :

$$C(p, d) = \frac{1}{\|U_d(p)\|} \sum_{q \in U_d(p)} C(q, d), \tag{7}$$

where $U_d(p) = \{(x, y) \mid (x, y) \in U(p), (x, y - d) \in U^{'}(q)\}$, $\|U_d(p)\|$ is a the number of pixels in $U_d(p)$ used to normalize the aggregated cost $C(p, d)$.

2.3. Post-processing

The obtained disparity results provided using the previous steps contain outliers in occluded regions and along depth discontinuities. To detect these outliers, the Left right consistency check method is performed between the left and right disparity maps. Then, to fill the detected outliers, the hole filling method [18] is used. It consists to fill them with the lowest disparity values of the closest reliable disparities located on the same scanline (pixel row). The latter one can generates streak-like artifacts in the disparity map. To solve this problem, we propose to apply a median smoothing via cross-based region [7]. Finally, a median filter, with the size of (5×5) is performed in order to further smooth the final disparity map by eliminating the remainder noises.

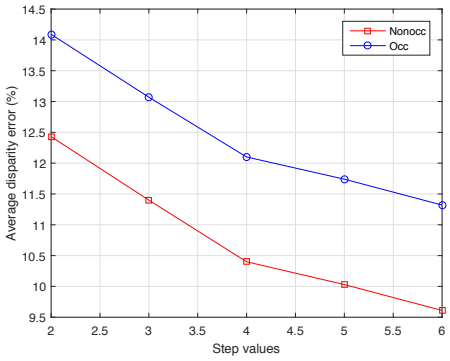


Fig. 3. Average disparity errors with respect to $step_2$ values for first 10 images from KITTI training sets

Table 1. Percentage of erroneous disparities in occluded and non-occluded regions for the KITTI training set

	3-px threshold	
	Non-occ	Occ
$C_{DIFFCensus}$ [2]	12.97	14.91
$C_{DIFFCCC}$ [2]	13.53	15.47
Proposed cost	**11.93**	**13.89**

3. EXPERIMENTAL RESULTS

3.1. Evaluation of the proposed cost function

In order to illustrate the efficiency of the proposed cost function, we performed an evaluation of this one over the top cost functions defined in the survey [2], $C_{DIFFCensus}$ and $C_{DIFFCCC}$ for the KITTI benchmark, where real radiometric changes exist. No refinement process was performed in this comparison. The cost adaptive aggregation method was used. The local WTA strategy was adopted in order to generate the disparity maps. Optimal parameter values defined in [2] were considered. In order to find the optimal step values, $step_1$ was set to 2 and we performed several tests on the first 10 KITTI images. Figure 3 presents the disparity errors, with respect to the different values of $step_2$. The lower error rate obtained was the one with $step_2 = 6$, which was therefore selected for the coming tests.

Table 1 presents the average percentage of erroneous pixels with both occluded and non occluded regions, computed at 3 default pixels error threshold for the KITTI training set. The obtained results using our cost are the lowest ones in both re-

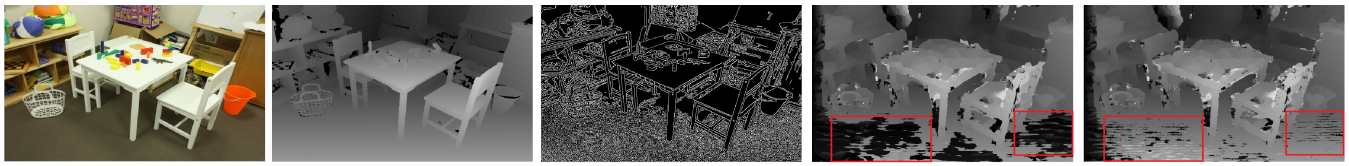


Fig. 4. From left to right : left image, ground truth disparity map, edge image, disparity map computed based on cross based adaptive region, and disparity map obtained via the modified one

Table 2. Percentage of erroneous disparities in non-occluded regions for the Middlebury training set

Aggregation Method	1 px threshold Nocc	2 px threshold Nocc	3 px threshold Nocc
Original method	19.69	14.93	13.41
Proposed one	**18.56**	**14.11**	**12.61**

gions. Indeed, the improvement is of the order 1.04, 1.60 for non-occluded region and of 1.02, 1.58 for other zones, with respect to $C_{DIFFCensus}$ and $C_{DIFFCCC}$ costs, respectively. This evaluation demonstrates that proposed cost is more robust against the real outdoor radiometric changes than the top performer in radiometric changes.

3.2. Evaluation of the modified cross based aggregation method

In this section, we performed tests to assess the effect of adding the edge information in the cross based aggregation method compared to the conventional one. Figure 4 depicts an example of the performance of the proposed aggregation method applied for the pair Playtable from the Middlebury training set which presents the improvements of disparity results when incorporating the edge information, especially, in the homogenous regions. Moreover, we carried out experiments on the whole new Middlebury training set. No refinement process was used as well in this evaluation. Table 2 presents the average of erroneous error pixels in non-occluded region for the quarter resolution. The errors are computed at 3 different pixels error thresholds. The obtained results clearly demonstrated an improvements of disparity results using the modified cost aggregation.

3.3. Evaluation of the proposed algorithm on the KITTI platform

Finally, we achieved our experiments by evaluating the performance of presented algorithm with respect to the current state of-the-art on KITTI testing dataset [19]. Parameters sets of the proposed algorithm are: the size of the support windows was set to $(m \times n) = (9 \times 9)$. The spatial and the color similarity thresholds were fixed at $L = 9$ and $\tau = 20$, respectively. In addition, a mild Gaussian filter size $((3 \times 3), r = 0)$ was performed on the gray scale images before calculating partial derivatives, in the cost computation step, in order to reduce noise and smooth around image edges. Our experiments were implemented on a Desktop equipped with a 3.0GHz Intel core i5 CPU and a 4GB of memory. A CPU implementation of the proposed algorithm has leads to an average computational time of 125s for KITTI dabaset. Table 3 presents the result of evaluation from the KITTI platform. Our algorithm is ranked 64th amongst more than 84 stereo matching algorithms and is one of the best local methods. Indeed, it overcomes may known local algorithms, such as Cross-Census[11], GF(Census)[18], LAMC-DSI[7], S+GF(Cen)[20].

Table 3. Results from the KITTI evaluation platform for the default 3 pixel threshold. Columns from left to right: method; percentage of erroneous pixels in non-occluded regions and in total; average disparity error in non-occluded areas; average disparity error in total. Date of evaluation: January 17, 2017

Method	Out-Noc	Out-All	Avg-Noc	Avg-All
Our Method	**8.71 %**	**10.05 %**	**2.1 px**	**2.7 px**
Deep-Raw	8.93 %	11.07 %	3.9 px	4.9 px
S+GF (Cen)	9.03 %	11.21 %	2.1 px	3.4 px
CrossCensus	9.46 %	10.86 %	2.3 px	2.7 px
SymST-GP	9.79 %	11.66 %	2.5 px	3.3 px
SM_GPTM	9.79 %	11.38 %	2.1 px	2.6 px
LAMC-DSI	9.82 %	11.49 %	2.1 px	2.7 px
IIW	10.78 %	12.62 %	3.3 px	4.3 px
SDM	10.95 %	12.14 %	2.0 px	2.3 px
HLSC_mesh	11.22 %	12.82 %	2.3 px	2.9 px
GF (Census)	11.65 %	13.76 %	4.5 px	5.6 px

4. CONCLUSION

In this paper, we presented a new stereo matching algorithm based on a new variant of the Census cost function for the cost computation stage. Experimental results, using real road scenes of the KITTI dataset, have demonstrated that the proposed variant leads to the lowest disparity mean errors compared to the top performer in this dataset. Moreover, a local method based on cross aggregation is updated to incorporate the edge information. The modified aggregated costs have leaded to an improvement of disparity results. A post-processing step is performed to remove any noise left. The obtained disparity results are considered promising.

5. REFERENCES

[1] Heiko Hirschmuller and Daniel Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sept. 2009.

[2] Alina Miron, Samia Ainouz, Alexandrina Rogozan, and Abdelaziz Bensrhair, "A robust cost function for stereo matching of road scenes.," *Pattern Recognition Letters*, vol. 38, pp. 70–77, 2014.

[3] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

[4] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proceedings of the Third European Conference on Computer Vision (Vol. II)*, Secaucus, NJ, USA, 1994, ECCV '94, pp. 151–158, Springer-Verlag New York, Inc.

[5] Andreas Klaus, Mario Sormann, and Konrad Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, Washington, DC, USA, 2006, ICPR '06, pp. 15–18, IEEE Computer Society.

[6] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and Xiaopeng Zhang, "On building an accurate stereo matching system on graphics hardware," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov 2011, pp. 467–474.

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

[8] Gorkem Saygili, Laurens van der Maaten, and Emile A. Hendriks, "Adaptive stereo similarity fusion using confidence measures," *Computer Vision and Image Understanding*, vol. 135, pp. 95 – 108, 2015.

[9] C. Lawrence Zitnick and Takeo Kanade Sing Bing Kang and, Jon A. Webb and, "A multibaseline stereo system with active illumination and real-time image acquisition," in *ICCV*, 1995, pp. 88–93.

[10] Kuk-Jin Yoon and In So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 650–656, April 2006.

[11] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit, "Cross-based local stereo matching using orthogonal integral images.," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 19, no. 7, pp. 1073–1079, 2009.

[12] O. Veksler, "Stereo correspondence by dynamic programming on a tree," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 2, pp. 384–390 vol. 2.

[13] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787–800, July 2003.

[14] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2001, vol. 2, pp. 508–515 vol.2.

[15] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, Feb 2008.

[16] A. Miron, S. Ainouz, A. Rogozan, and A. Bensrhair, "Cross-comparison census for colour stereo matching applied to intelligent vehicle," *Electronics Letters*, vol. 48, no. 24, pp. 1530–1532, November 2012.

[17] R. Maini Raman and H. Aggarwal, "Study and comparison of various image edge detection techniques," *International Journal of Image-processing*, vol. 3, no. 1, pp. 1–12, 2009.

[18] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *CVPR 2011*, June 2011, pp. 3017–3024.

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[20] Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, Shuicheng Yan, and Qi Tian, "Cross-scale cost aggregation for stereo matching," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 1590–1597.