IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 22, 2016, accepted August 12, 2016, date of publication August 26, 2016, date of current version September 28, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2603220

A Depth Map Post-Processing Approach Based on Adaptive Random Walk With Restart

HOSSEIN JAVIDNIA, (Student Member, IEEE), AND PETER CORCORAN, (Fellow, IEEE)
Department of Electronic Engineering, College of Engineering, National University of Ireland, Galway SW4 794, Ireland

Corresponding author: H. Javidnia (h.javidnia1@nuigalway.ie)

This work was supported by the Strategic Partnership Program of Science Foundation Ireland (SFI) and co-funded by SFI and FotoNation Ltd., on Next Generation Imaging for Smartphone and Embedded Platforms under Project 13/SPP/I2868.

ABSTRACT Accurate depth estimation is still an important challenge after a decade, particularly from stereo images. The accuracy comes from a good depth level and preserved structure. For this purpose, a depth post-processing framework is proposed in this paper. The framework starts with the ''Adaptive Random Walk with Restart (2015)'' algorithm. To refine the depth map generated by this method, we introduced a form of median solver/filter based on the concept of the mutual structure, which refers to the structural information in both images. This filter is further enhanced by a joint filter. Next, a transformation in image domain is introduced to remove the artifacts that cause distortion in the image. The proposed post-processing method is then compared with the top eight algorithms in the Middlebury benchmark. To explore how well this method is able to compete with more widely known techniques, a comparison is performed with Google's new depth map estimation method. The experimental results demonstrate the accuracy and efficiency of the proposed post-processing method.

INDEX TERMS Stereo matching, depth map, accuracy, edge preserving.

I. INTRODUCTION

A. STEREO DEPTH MAPS

In 3D computer graphics a depth map is an image or image channel that contains information relating to the distance to the surfaces of scene objects from a viewpoint [1]. The depth information corresponds to luminance in proportion to the distance from the camera. Near surfaces are depicted as lighter while far surfaces are shown as darker. Estimating the depth can be considered an important component of understanding geometric relations within a scene. In turn, such relations help to provide a richer representation of objects and their environment, often leading to improvements in existing recognition tasks, as well as enabling further applications such as robotics. In recent years, many new economical facilities, including time-of-flight [2], [3], structured light [4], and the Kinect were introduced for depth determination from stereo images. Kinect captures pairs of synchronized depth-color images for a scene within a range of several meters. However, the depth map cannot be used directly in scene reconstruction because it has some deficiencies such as gaps due to occlusion, reflection and other optical factors.

In general stereo algorithms or stereo matching algorithms are categorized into two groups based on the taxonomy scheme of Scharstein and Szeliski [5]: i.e. local and global algorithms.

In the local algorithms, the depth value at pixel P is dependent on the intensity and color values of the window W in which P is located. The initial matching cost is pixel-wise which is often noisy with minimum information in parts of the image with smoother texture. Therefore using the cost of the neighboring regions will assign the best depth value to pixel P.

On the other hand global methods consider the overall structure of the scene and smoothen the image and then try to solve the cost optimization problem.

B. STEREO MATCHING ALGORITHMS

In the last decade stereo matching has attracted a lot of attention from researchers and many matching algorithms have been developed. Some of the most well-known and studied algorithms are LIBELAS [6], iSGM [7], DBP [8] and CostFilter [9], LIBELAS [6] has been used since 2010 in different research studies. It is inspired from the observation that despite the fact that many stereo correspondences are highly ambiguous, some of them can be robustly matched.

While the processing speed of the LIBELAS is quite fast, the accuracy of the estimated depth map is poor.

iSGM [7] is an iterative scheme of Semi-global matching (SGM) technique with refined concept of the cost integration of semi-global matching. The gathered buffer is evaluated to a prior disparity map after horizontal and vertical integration.

DBP [8] is a global matching algorithm based on energy-minimization which as all other global methods contains data and smoothness term. The main contribution in data term in this algorithm is that, it is being approximated by a color weighted correlation. Afterwards, the data term is being refined in occluded regions by employing the hierarchical loopy belief propagation algorithm.

CostFilter [9] is a framework for multiple applications such as computing the disparity maps in real-time. It is the technique which aims to be fast and edge-aware. It consists of three steps: constructing a cost volume, fast cost volume filtering and winner-take-all label selection. The estimated depth by this method suffers from blocky artifacts along the edges and corners, especially in the regions with illumination transition. This causes a broken synthetic view along the edges.

There are other methods which tried to obtain better accuracy of depth map based on the combination of Markov Random Field (MRF) and sophisticated global optimization techniques in different researches [10]–[13], but still obtaining a good accuracy in depth estimation remains a challenge, especially in images with sophisticated or very simple texture.

Another approach which has been considered to improve the accuracy of the depth map by mostly preserving the edges was using the Mutual Information (MI) and SIFT features. A multisensor synthetic aperture radar (SAR) image registration method was proposed based on MI [14] and SIFT [15]. In this application, MI was used to estimate the registration parameters which were being used later by conjugate feature selection during the SIFT matching phase to decrease the number of false matches. Following the same idea, a stereo matching method was introduced in [16], based on the combination of MI, SIFT, plane-fitting and log-chromaticity color space.

Generally finding a local matching method which performs well in terms of both speed and accuracy is not easy and straightforward. But recently employing the random walk with restart along with optimizing the matching cost proved that it is possible to have fast matching with pretty accurate estimation. ARWR is a local matching algorithm based on random walk with restart method [17] which is used as the fundamental algorithm in this paper.

At this point it is timely to introduce the field of application, which establishes requirements for a high performance stereo disparity map. This work derives from research on automotive street-scene analysis where it is important to determine small objects in order to evaluate risks in the path of a vehicle – e.g. distant pedestrians, animals, vehicles. As most automotive imaging systems employ relatively small sensors (2-4 MP) compared to consumer devices it is important to be able to run disparity mapping algorithms at full native sensor resolution – in our case 2864 * 1924 pixels.

All current methods, as outlined above, suffer from non-accurate depth around edges and corners, depth discontinuity especially in texture-less areas, depth conflict around the area with similar colors and missing depth in one depth level. By solving these challenges a depth map can present correct and accurate depth information while respecting the structure of the reference image.

C. FEATURES OF THE PROPOSED METHOD

In this paper is presented a method to refine the depth map generated by the *Adaptive Random Walk with Restart* (ARWR) algorithm in order to obtain significant improvements in accuracy. The main features of the proposed method are:

1- A guided joint filter based on the mutual information was designed by diffusing the image domain.
2- Weights are allocated dynamically to the windows as part of the joint filter. The weights are being regenerated every time the window is moving to the other patch of pixels. The pixels count in different bins of a histogram instead of storing the weights directly.
3- The important point about the proposed filter is that it is rotation invariant because of the joint mutual information. Also the filter can be applied repeatedly to remove more noise but the edges and corners will be preserved because of the mutual joint feature.
4- When using this filter, the algorithm works better on high resolution images in comparison with low resolution.
5- This filter can be used for upsampling/downsampling purposes.
6- This method has the advantage of filling the depth map in regions with missing depth values.

The rest of this paper is organized as follows:

In the next section the chosen method, ARWR is presented in detail. Section 3 provides the details of the proposed post-processing filter. The results of the evaluation as well as experimental results are presented in section 4, while conclusions are drawn in section 5. There are also 2 appendices linked to this paper presenting extended numerical and visual results.

II. INTRODUCTION TO ADAPTIVE RANDOM WALK WITH RESTART

In this section we describe the fundamental and technical details of the chosen stereo matching method, ARWR.

ARWR has an acceptable and comparable performance in terms of estimation and speed against other algorithm, but it is still far from the top stereo matching algorithm on Middlebury benchmark in terms of accuracy.

This algorithm has several important advantages which make it a suitable method for a variety of applications. It is

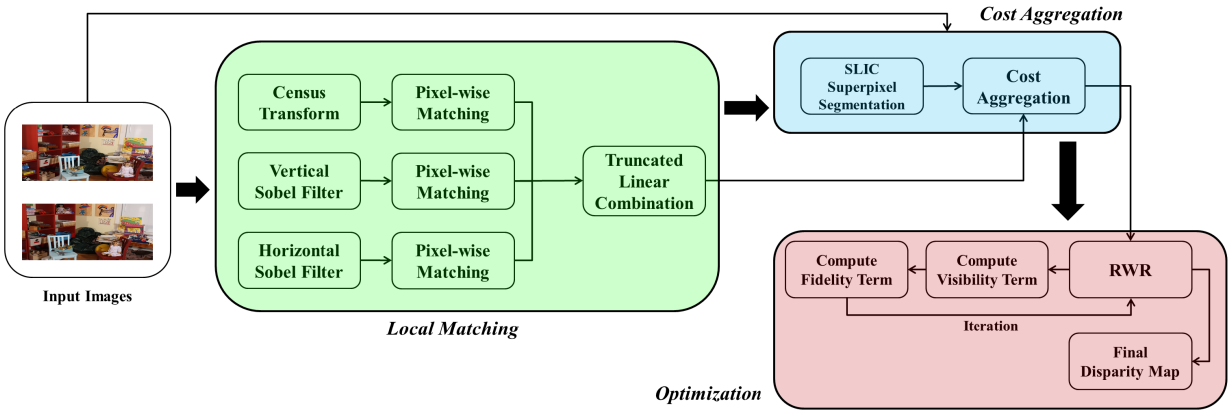

FIGURE 1. Overview of the adaptive random walk with restart.

not affected by illumination variation because of gradient and census transform, the processing time is quite fast in comparison with recently studied methods, has good performance in both outside and inside environment and gives us the option to have a estimation of the depth in low texture scenes.

One important advantage of this algorithm which convinced us to employ it as a part of our approach, is the good performance on high resolution images. A traditional way to speed up stereo computation is to use image pyramids or downsized images which also reduce the disparity range. This down-sampling in disparity computation will cause some small objects to be missed. The full disparity resolution for large distance is vital for long range object detection. The point about the chosen algorithm is that the image doesn't need to be down-sampled to speed up the method.

The comparison of this method with several others methods done in this paper showed that it has acceptable depth estimation in high resolution images, 2864 * 1924 pixels.

Acceptable depth estimation refers to the fact that the algorithm doesn't have the problem of estimating different layers of depth in one object. It respects the depth layers without conflict. This feature along with the fast processing time makes this algorithm suitable for high resolution real-time applications. Also it gives us the ability of making a more accurate filter, which is described later in the paper.

A. ALGORITHM DESIGN

The initial matching cost in ARWR is pixel-wise calculated by employing census transform and gradient image matching. Census-based matching technique or census transform was initially introduced by Zabi in 1994 [18]. It is a form of non-parametric local transform to map the intensity values of the pixels within a square window to a bit string, thereby capturing the image structure. In other words, it computes for every pixel a binary string (census signature) by comparing its grey value with the grey values in its neighborhood.

The census transform is robust to radiometric variations but the noise in the local image structure is being encoded based on the intensity of the pixels. The encoded noise brings some matching doubts especially in the area with repetitive or similar texture patterns.

To overcome this problem gradient image matching is employed as part of the local matching block in ARWR. At this stage gradient images are computed using 5×5 Sobel filters. The whole process of the ARWR is shown in Fig. 1.

The green block in Fig. 1 shows the local matching block including the transformation and matching parts.

The usual similarity criteria in stereo matching are only strictly valid for surfaces with Lambertian (diffuse) reflectance characteristics. Specular reflections are viewpoint dependent and may cause large intensity difference at corresponding image points. In the presence of specular reflection, traditional stereo methods are often unable to establish any correspondence, or the calculated disparity values tend to be inaccurate.

In this case using the gradient image matching makes the local matching method more robust on non-Lambertian surfaces.

The noise variation in the local pixel-wise matching methods can be vital in term of the performance. That is why SLIC (Simple Linear Iterative Clustering) algorithm is employed in ARWR, the blue block in Fig. 1. SLIC is one of the common super-pixeling methods [19].

The local measurements in the matching block are more robust to noise variation when the super-pixels are considered as the smallest parts of the image to be matched to the target image. Super-pixeling is considered as an alternative to pixels in pixel-wise matching which leads to a reduction in memory requirements in the whole algorithm.

At the last step of the ARWR which is shown as pink block in Fig. 1, the calculated matching cost is updated using the RWR algorithm to determine the optimum disparity with respect to occluded and discontinuity regions. The standard

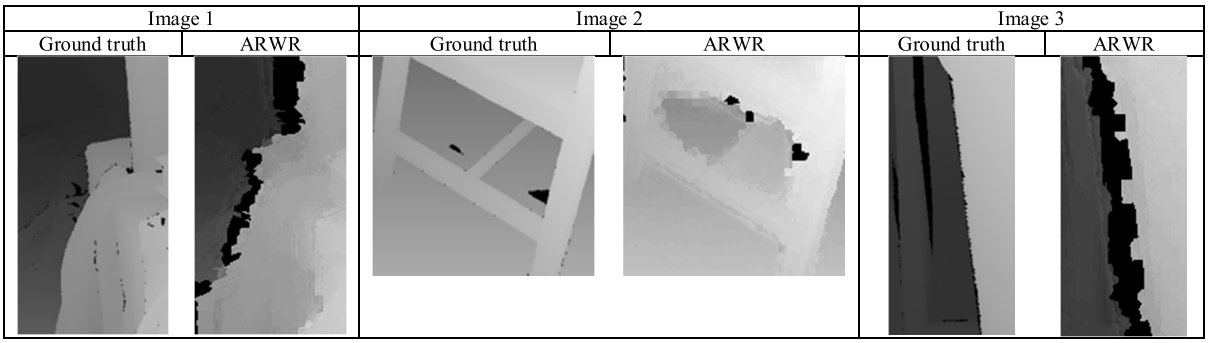

FIGURE 2. Broken edges and corners in the computed depth map by ARWR.

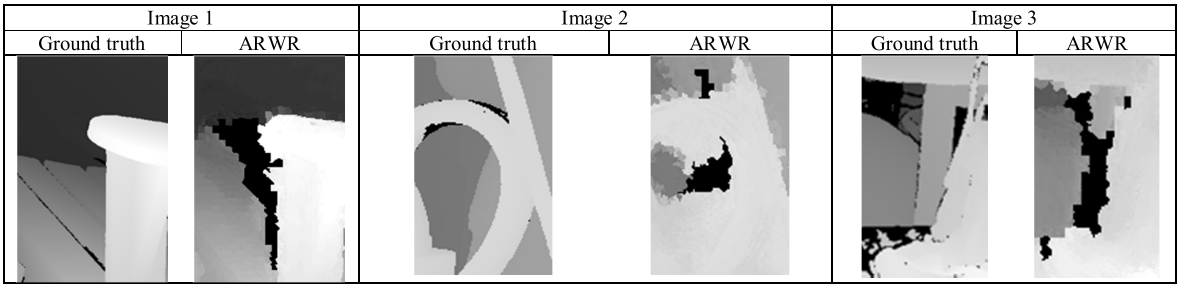

FIGURE 3. Missing patches in the computed depth map by ARWR.

cost update algorithm in RWR is modified in ARWR where the matching cost is updated adaptively by considering the position of the super-pixels in the regions of occlusion or depth discontinuity.

To recover the smoothness failure at occlusion or depth discontinuity regions in ARWR, a visibility constraint is formulated within the RWR algorithm which requires an occluded pixel to have no match on the target image, and a non-occluded pixel to have at least one match.

B. ALGORITHM TRADE OFF

There are some issues with the generated depth map based on ARWR which need to be solved to obtain a clearer and more accurate depth map.

The depth map produced by the ARWR is suffering from speckle noise and inaccurate object edges especially for objects with a detailed geometry. Basically the generated map is not preserving the edges and corners. At some parts of the computed depth map the edges are broken or they are faded into other objects which makes it unsuitable for segmentation purposes and classification. Fig. 2 shows examples of the broken edges and corners in the computed depth and the corresponding patches in the ground truth.

The other issue is the missing parts in the generated map. We demonstrate that each patch of pixels in a depth map can provide us valuable information like the scaling factor and distance to the objects. Fig. 3 represents some samples of the missing parts in the depth map and the corresponding patches in the ground truth.

The samples show that some parts of the depth map were not estimated by ARWR and it brings a false depth level which is not suitable for 3D reconstruction applications.

These issues are generally some of the most challenging problems in the current depth computation and enhancement methods. Having a map which is preserving the right edges and corners while all pixel patches are contributing in the depth level allows us to reconstruct an accurate 3D scene from the camera view point. It also provides an accurate fundamental platform for variety of applications such as classification, segmentation, distance estimation, obstacle detection and autonomous navigation.

In the next section of this paper our approach is presented and shown to provide a suitable solution to the issues mentioned above.

III. PROPOSED POST-PROCESSING FILTER

To solve the issues mentioned in the previous section, mutual information of the reference image and the depth map is used as the input of the joint weighted median filter. By employing the mutual joint filter the problem of the regions of occlusion or depth discontinuity in the initial depth map is solved. To resolve the blocky artifacts from object edges, the depth map is transferred to another domain by convolving it.

The whole process of the ARWR + proposed post-processing method is as follows:

1- Extract the initial depth by using the ARWR algorithm.
2- *A*. Apply mutual joint weighted median filter to fill the regions of occlusion or depth discontinuity in the initial depth map

B. Overwrite the structure of the RGB image on the depth map.

3- Transfer the depth map to a signal and perform normalized interpolated convolution on the domain of the signal to obtain an accurate, edges preserved depth map.

Fig. 4 presents the general overview of the whole process and Fig. 5 shows the detailed view of the ARWR + proposed post-processing method.

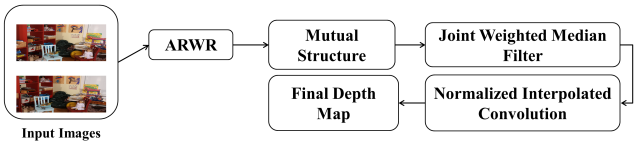

FIGURE 4. Overview of the proposed post-processing method.

A. MUTUAL-STRUCTURE

Mutual information has developed into an accurate measure for rigid and affine mono- and multimodality image registration or for two images, it is a combination of the entropy values of the images, both separately and jointly [20]. By measuring the structure similarity of two images, we can let the mutual-structure to guide the joint filtering process. Let's denote D and I as the initial depth map and the reference RGB image respectively. Also D_p and I_p are the pixel intensities in initial depth map and the reference RGB image respectively. To compute the structure similarity between two images, we consider a variety of patches in the images. One common and well-studied method to measure the structure similarity is to use normalized cross covariance (1). If we consider the images as two time series signals, then we can delay D by W samples and then calculate the cross-covariance between the pair of signals,

$$CC\,(W) = \frac{1}{M-1}\sum_{k=1}^{M}(D_{k-W} - \mu_D)(I_k - \mu_I), \tag{1}$$

Where μ_D and μ_I are the means of each time series and there are M samples in each. $CC\,(W)$ is the cross-covariance function. Normalized cross-covariance is called cross-correlation,

$$N\,(W) = \frac{CC\,(W)}{\sqrt{\sigma(D_p)\sigma(I_p)}}, \tag{2}$$

$$N\left(D_p, I_p\right) = \frac{cov(D_p, I_p)}{\sqrt{\sigma(D_p)\sigma(I_p)}}, \tag{3}$$

Where $cov(D_p, I_p)$ is the covariance of patch intensity. $\sigma\left(D_p\right)$ and $\sigma(I_p)$ denote the variances of pixel intensities in the initial depth map and RGB image respectively. The maximum value of $N\left(D_p, I_p\right)$ is 1 when two patches are with the same edges, otherwise $\left|N\left(D_p, I_p\right)\right| < 1$. Nonlinear computation makes it hard to use the normalized cross-correlation directly in the process. To solve this problem, making a connection between normalized cross-correlation and least-square regression would be helpful. If we consider $H(p)$ as a patch centered at pixel p, then the least-squared regression function would be:

$$f\left(D, I, \alpha_p^1, \alpha_p^0\right) = \sum_{q\in N(p)}\left(\alpha_p^1 D_q + \alpha_p^0 - I_q\right)^2, \tag{4}$$

Where α_p^1 and α_p^0 are the regression coefficients. This function linearly represent one patch in D corresponding with the one in I. Minimum error with the optimal α_p^1 and α_p^0 can be defined as:

$$e(D_p, I_p)^2 = \frac{min}{\alpha_p^1, \alpha_p^0}\frac{1}{|H|}f\left(D, I, \alpha_p^1, \alpha_p^0\right), \tag{5}$$

By considering the (1) and (5), we can say the mean square error is:

$$e\left(D_p, I_p\right) = \sigma\left(I_p\right)\left(1 - N\left(D_p, I_p\right)^2\right), \tag{6}$$

The relation between the mean square error and normalized cross-correlation is previously proved in [19]. When $\left|N\left(D_p, I_p\right)\right| = 1$, it means that two patches only contain mutual structure and $e\left(D_p, I_p\right) = 0$. So:

$$e(I_p, D_p)^2 = \frac{min}{b_p^1, b_p^0}\frac{1}{|H|}f\left(I, D, b_p^1, b_p^0\right), \tag{7}$$

Therefore $e\left(I_p, D_p\right) = 0$ when $\left|N\left(D_p, I_p\right)\right| = 1$. According to the above analysis, the structure similarity can be defined as:

$$S_s\,(D, I, \alpha, b) = \sum_p (f\left(D, I, \alpha_p^1, \alpha_p^0\right) + f(I, D, b_p^1, b_p^0)), \tag{8}$$

where α and b are the coefficient sets of $\left\{\alpha_p^1, \alpha_p^0\right\}$ and b_p^1, b_p^0 respectively.

Algorithm 1 computes the mutual information of D and I.

Algorithm 1 Mutual Information

Input: Image D and I
Output: Mutual Information of D and I

Initialize W, M to 0;
Initialize $\alpha = \beta(\alpha_p)$;
Initialize $b = \beta(b_p)$;
  $\mu_D \leftarrow mean(D)$;
  $\mu_I \leftarrow mean(I)$;
  $\sigma_W = M \big/ \sum (D_{M-W} - \mu_D)\,(I_M - \mu_I)$;
    foreach H in D **do**
      $\sum (\alpha_p D_{N(p)} + \alpha_p - I_{N(p)})^2$;
    end
  return $S\,(D, I, \alpha, b)$;

B. JOINT WEIGHTED MEDIAN FILTER

Median filter [21] is a nonlinear operation which runs through an image I and replaces each pixel value V by the median value of neighboring pixels within a $(2j+1)^2$ window W_p:

$$I_{median}\,(p) = median\left\{V : p_i \in W_p\right\}, \tag{9}$$

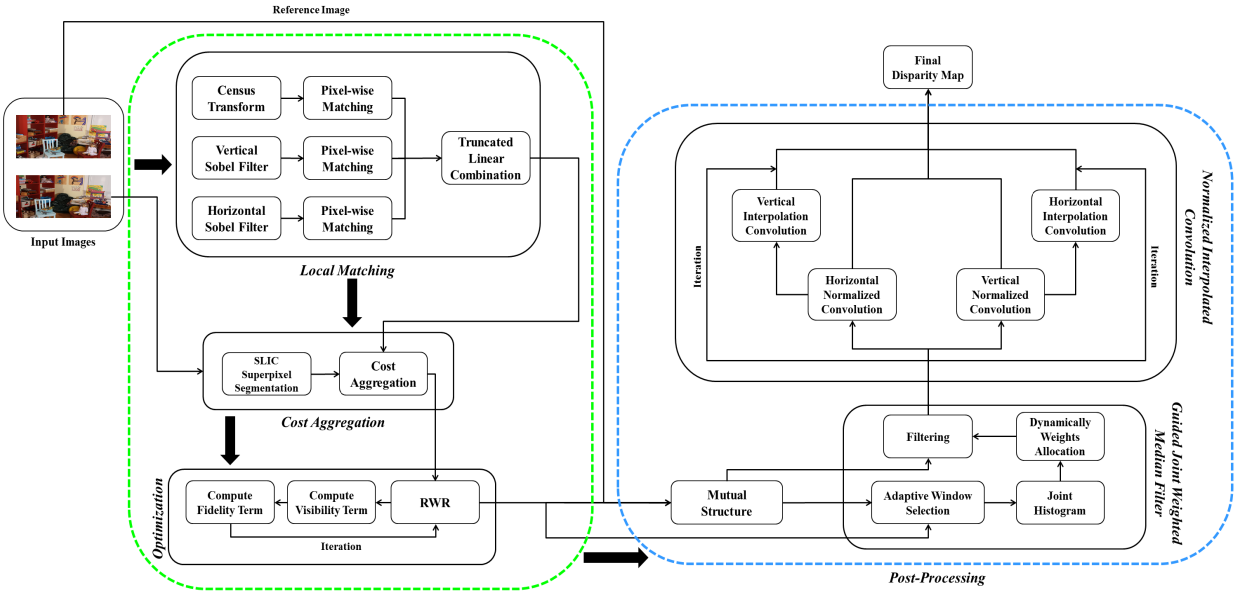

FIGURE 5. Overview of the ARWR + Proposed post-processing method.

Median filter processes all the neighbors equally and may lead to some artifacts like changing the shape of the sharp corners and make them circular or removing thin structures. Weighted median filter [22] was introduced to solve this issue. Considering $\omega(p, p^{'})$ the weight on image I, then:

$$h(p,i) = \sum_{p^{'} \in W_p} \omega(p, p^{'}) \delta(V(p^{'}) - i), \tag{10}$$

where W_p is a local windows near p, i is the discrete bin index and $\delta(.)$ is the Kronecker delta function which is 1 when the argument is 0, otherwise it is 0. $h(p, .)$ is the local histogram with the weighted pixel in it. By accumulating $h(p, i)$ the weighted median value is obtained.

Joint median filter on a depth map D with a group S of segments as masks is defined as:

$$D_{J_median}(p) = median\left\{D(p_i) : p_i \in W_p \cap S_p\right\}, \tag{11}$$

where $S_p \in S$ is the segment containing pixel p. So the new local histogram for depth map would be:

$$h_D(p,i) = \sum_{p^{'} \in W_p \cap S_p} \delta(D(p^{'}) - i), \tag{12}$$

Based on the (10) and (12), the local histogram of the joint weighted median filter on the depth map D would be:

$$h_{D_f}(p,i) = \sum_{p^{'} \in W_p \cap S_p} \omega(p, p^{'}) \delta(D(p^{'}) - i), \tag{13}$$

Using the mutual structure and joint weighted median filter gives us the capability to transfer the structural information of the reference image to the depth map, instead of transferring the whole pattern. And in addition it contributes greatly to a preservation of the edges in the depth map.

C. NORMALIZED INTERPOLATED CONVOLUTION

Joint weighted median filter based on the mutual structure provides an edge preserved and smooth depth image, but still the depth map is suffering from blocky artifact, especially on the edges. To decrease the blocky effects on the depth map, converting the image to another domain would be helpful. Let's consider a signal:

$$f(t) = [x_1; 0; 0; x_4; x_5; 0; x_7; 0], \tag{14}$$

where x_i are known samples of signals and the missing samples are replaced by 0.

A simple smoothing filter is:

$$g(t) = \left[\frac{1}{3}; \frac{1}{3}; \frac{1}{3}\right], \tag{15}$$

Filling the missing part of the $f(t)$ by applying the $g(t)$ will provide:

$$f(t) \times g(t) = \left[\frac{x_1}{3}; \frac{x_1}{3}; \frac{x_4}{3}; \frac{x_4 + x_5}{3}; \frac{x_4 + x_5}{3}; \frac{x_5 + x_7}{3}; \frac{x_7}{3}; \frac{x_7 + x_1}{3}\right], \tag{16}$$

At this level using the Normalized Convolution appends a component to each signal which expresses the confidence of a signal. This component is equal to 0 for each missed sample. If we consider the map of the component on signal $f(t)$ as $g(t)$, then:

$$c(t) = [1; 0; 0; 1; 1; 0; 1; 0], \tag{17}$$

By considering the convolution of $c(t)$, it is possible to approximate the original signal with the filled gaps. So:

$$f(t)_O = \frac{f(t) \times g(t)}{c(t) \times g(t)}, \tag{18}$$

where the $f(t)_O$ is the original signal without gaps.

This scenario previously has been studied to filter the non-uniform sampled signals [23]. If $T_\omega(ct\,(x))$ is a uniformly sampled signal in Ψ_w, then for a uniform discretization $U(\Psi)$ of the original domain Ψ, normalized convolution generates the smoothed value of a sample $q \in U(\Psi)$ as:

$$Fi\,(q) = \left(\frac{1}{J_q}\right) \sum_{l \in U(\Psi)} T\,(l)\, R\left(t\left(\hat{q}\right), t\left(\hat{l}\right)\right), \qquad (19)$$

Where $J_q = \sum_{l \in U(\Psi)} R(t\left(\hat{q}\right), t(\hat{l}))$ is a normalized factor for q and R is an arbitrary kernel. Generally interpolated surfaces in an image are smoother than the corresponding ones generated by normalized convolution. To obtain this, $Fi(q)$ can be filtered by continuous convolution as below:

$$CCF\,(q) = \int_{U_\Psi} Fi\,(x)\, R\left(t\left(\hat{q}\right), x\right) dx, \qquad (20)$$

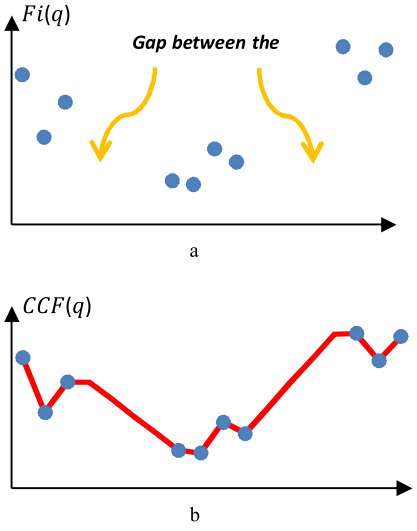

FIGURE 6. Missing samples recovery. (a) Samples of a signal with missing parts. (b) Recovered samples in domain Ψ.

Where R is a normalized kernel. Fig. 6.b shows how the missing samples of signal T are recovered in domain Ψ.

Applying the same process on a depth map generates a smooth and artifact free map by transferring it into the domain Ψ.

IV. EVALUATION

A. MIDDLEBURY BENCHMARK

The Middlebury benchmark has been widely used over the last decade to evaluate the performance of stereo matching algorithms [24]. The ARWR was applied with and without the proposed post processing on 15 standard images from the Middlebury 'dense' training dataset. Based on the average weight on metric 'bad 2.0', the first 8 algorithms from Middlebury were chosen for comparison, including GCSVR [25], INTS [26], MCCNN_Layout [25], MC-CNN+FBS [25], MC-CNN-acrt [27], MC-CNN-fst [27], MeshStereo [28], SOU4P-net [25] and the original ARWR without post-processing. As evaluation metrics we consider the ones presented in Table 1.

TABLE 1. Metrics used in this paper to evaluate the algorithms.

	Metric	Formula
1	MSE (Mean Squared Error)	$\frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}\|f(i,j)-g(i,j)\|^2$
2	RMSE (Root Mean Squared Error)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
3	PSNR (Peak Signal to Noise Ratio)	$20\log_{10}(\frac{MAX_f}{\sqrt{MSE}})$
4	SNR (Signal to Noise ratio)	$\frac{10\log_{10}(P_{signal})}{10\log_{10}(P_{noise})}$
5	MAE (Mean Absolute Error)	$\frac{1}{n}\sum_{i=1}^{n}\|f_i-y_i\| = \frac{1}{n}\sum_{i=1}^{n}\|e_i\|$
6	SSIM (Structural Similarity Index)	$\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}$
7	DSSIM (Structural Dissimilarity Index)	$\frac{1-SSIM(x,y)}{2}$

All the evaluation process in this paper is based on the high quality version of the images and all experiments were done under the same conditions.

All the images were normalized before evaluation and maximum disparity setup was defined for all algorithms. The average value of the 15 images in each metric was considered as the representing value of the corresponding algorithm. Table 2 shows the average value of metric/algorithm. To find the extended tables for each metric/image (color coded to better present relative performance of each algorithm for each evaluation metric) please refer to Appendix 1.

The best algorithm's value in each metric is emboldened. Based on the MSE, PSNR, SNR, SSIM and DSSIM metrics the proposed post-processing method has the best performance. Table 3 represents the ranking within the 10 tested algorithms of the ARWR without post-processing and with post processing applied for each of the evaluated metrics.

Fig.7 presents the results of the proposed post-processing method on three Middlebury database images.

The initial depth map is computed by ARWR. Beside the parametric evaluation, the visual comparison of the generated results and the ground truth clarify the fact that the proposed post-processing method can preserve edges and the structure. For more results of the post-processed ARWR and visual comparison with other methods please refer to Appendix 2.

While the performance of the proposed post-processing method in term of accuracy is good, the processing time is a trade-off. Fig.8 shows the processing time required by each step of the proposed post-processing method on an image with 962×1414 pixels resolution ran on Matlab R2013a. The initial disparity is estimated with a maximum disparity of 256.

TABLE 2. Average values of metric/algorithm.

	MCCNN_ Layout	MC-CNN-acrt	MC-CNN+FBS	SOU4P-net	MC-CNN-fst	MeshStereo	INTS	GCSVR	Original ARWR	Post-processed ARWR
MSE	0.0133	0.0171	0.0194	0.0133	0.0177	0.0195	0.0193	0.0235	0.0277	**0.0126**
RMSE	**0.104**	0.1199	0.1243	0.1069	0.1225	0.1357	0.1339	0.1456	0.1455	0.1041
PSNR	20.2902	19.028	19.2026	19.9836	18.842	17.6704	17.8519	17.2273	17.7517	**20.3136**
SNR	15.3891	14.1269	14.3016	15.0826	13.9409	12.7694	12.9509	12.3262	12.8506	**15.4125**
MAE	**0.0524**	0.0639	0.0915	0.0739	0.0666	0.101	0.1026	0.112	0.0867	0.0644
SSIM	0.99849	0.9981	0.9975	0.99847	0.998	0.9975	0.9976	0.9969	0.9966	**0.9985**
DSSIM	0.0008	0.001	0.0012	0.0008	0.001	0.0011	0.0012	0.0015	0.0017	**0.0007**

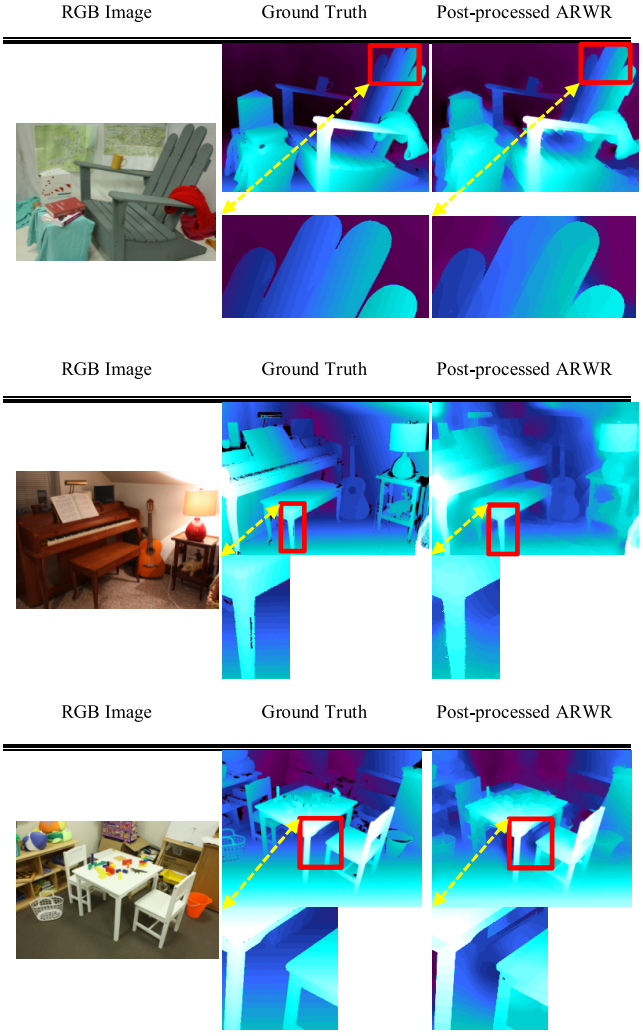

FIGURE 7. The result of the sample images from Middlebury database. Each set of figures denotes the left image, the ground truth and the proposed postprocessed depth map.

Table 4 represents the average performing time of the all algorithms applied on the same high resolution image set as per Middlebury.

The processing time of the studied method is poor, but can be readily improved as much of this work was not optimized for fast computation. The improvement of algorithm efficiency and computational speed is currently the subject of a follow-on research project to optimize for an embedded DSP or GPU implementation.

TABLE 3. Ranking of ARWR without and with post-processing out of 10 algorithms.

	ARWR without post-processing	ARWR with post-processing
MSE	9	1
RMSE	9	2
PSNR	8	1
SNR	8	1
MAE	6	3
SSIM	10	1
DSSIM	10	1

TABLE 4. The processing time of the studied algorithms on same high resolution image set.

	Algorithm	Time/Sec
1	MC-CNN-fst	1.26
2	ARWR without post-processing	21
3	MeshStereo	62
4	INTS	104
5	MC-CNN-acrt	106
6	MC-CNN+FBS	157
7	MCCNN_Layout	300
8	**ARWR with post-processing**	**440**
9	SOU4P-net	688
10	GCSVR	5891

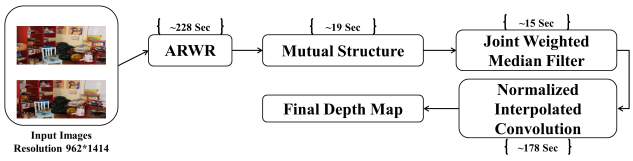

FIGURE 8. Processing time required by each step of the algorithm.

B. COMPARISON WITH GOOGLE'S DEPTH ESTIMATION TECHNIQUE

In the second part of the evaluation we referred to the recent technology which is used by the Google Camera "Lens Blur" feature in Android OS. The basic idea in this technology is to match the stereo images in the bilateral space by avoiding

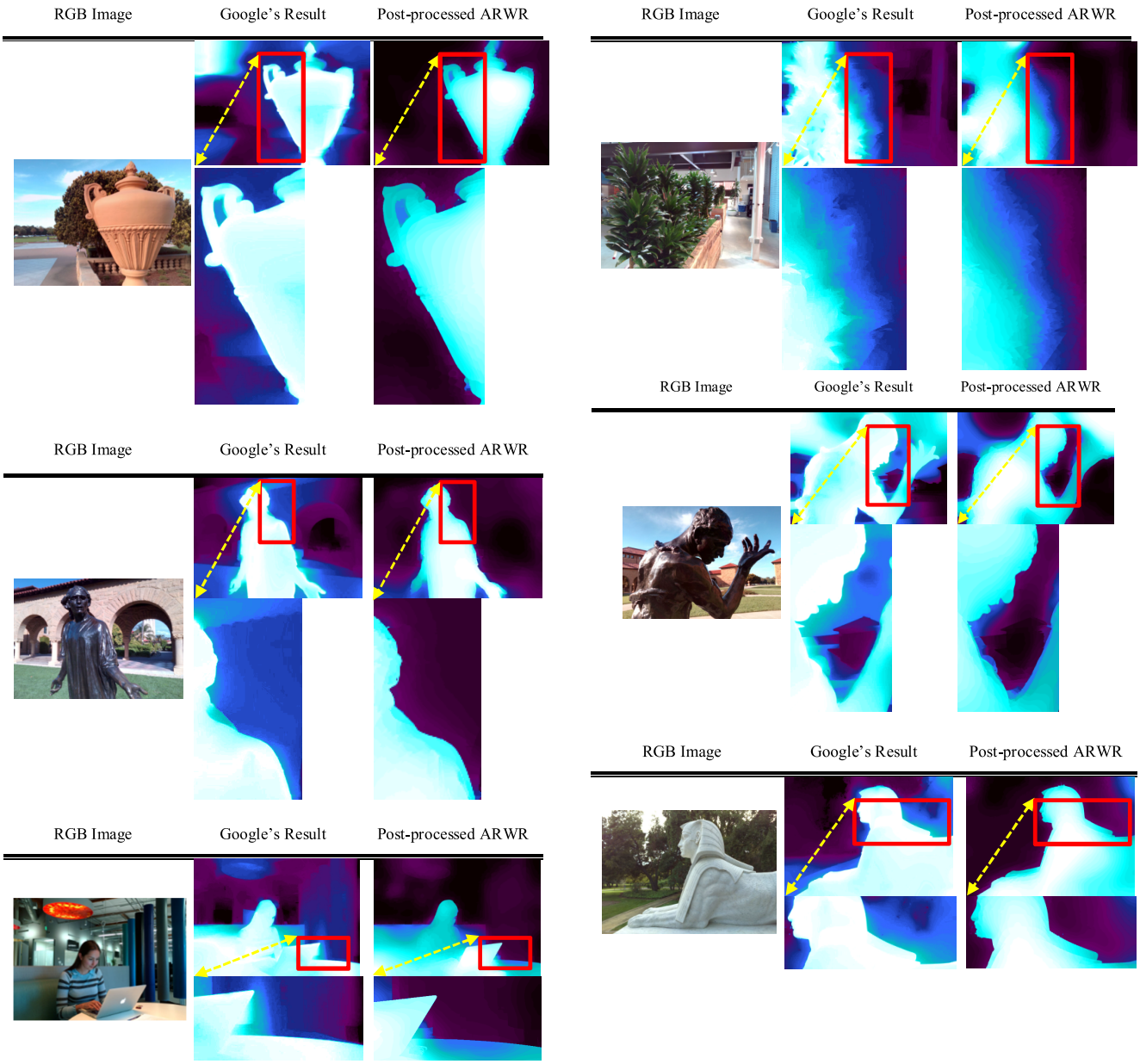

FIGURE 9. The result of the images from Google's method [29]. Each set of figures denotes the left image, Google's result and the proposed post-processed depth map.

per-pixel inference using leveraging techniques for fast bilateral filter [29]. This idea is presented in the other form to compute the depth from focus in the handheld devices by using focal stack.

A global approach is employed to generate the depth map by minimizing a cost function related to the pixel disparities. The data matching cost in their method is based on the Birchfield-Tomasi technique [30].

To satisfy the smoothness term of the cost function, the bilateral filter is used which causes a smoother image while the edges are preserved. For each pixel i of an image, one would typically consider a square (kernel) centered at i and perform a convolution.

Minimizing the cost function is extremely slow for higher resolution pictures. This problem is solved by splatting the value of each pixel into a higher dimensional bilateral space. The general idea is to; instead of applying the bilateral filter in pixel space, splat the pixels according to their location and color into a five-dimensional bilateral grid. Then blur the grid using a short range isotropic blur filter, and slice the grid in order to recover the filtered image.

According to the authors of [29], the most instinctive way to evaluate the performance of a stereo algorithm for defocus is to visually inspect the renderings produced using that algorithm. The kind of error that they cared about was related to failing to follow image edges at occlusion boundaries

TABLE 5. Structural similarity and dissimilarity of Google's method and post-processed ARWR.

	SSIM	DSSIM
Image 1	0.9939	0.0031
Image 2	0.9955	0.0022
Image 3	0.9867	0.0066
Image 4	0.9960	0.0020
Image 5	0.9907	0.0047
Image 6	0.9979	0.0011

where errors in disparity can cause rendering errors. The Middlebury error metrics are not considering this type of error. Middlebury error metrics are pixel-wise and Google's method has a poor performance on this benchmark, because their algorithm over- or under-estimate the disparity of flat texture-less regions, has disparity confusion in close shots with different level of brightness, has disparity confusion at the regions with specific pattern and sharp opposite colors.

Unfortunately there is no ground truth and benchmark based on this method. We only had access to a number of images and generated disparities which are published in [29].

To find out the structural similarity of the Google's result and the proposed post-processing method, we employed SSIM and DSSIM metrics. For two identical images the values of SSIM and DSSIM are 1 and 0 respectively. Table 5 shows how close are our results to Google's for each image and with the same disparity level setup. The visual comparison of the Google's technique and the post-processed ARWR is shown in Fig. 9. The visual comparison shows different patches of the estimated disparity by Google's and our method. This visual and numerical comparison show how close the proposed method is to Google's in terms of preserving the structure of the estimated disparity, edges and corners.

V. CONCLUSION

In this paper we proposed and evaluated a post-processing technique to increase the accuracy of the depth map computed by Adaptive Random Walk with Restart method. We demonstrated that keeping the sharp edges and corners along with main structure of the reference image in the depth map is an important factor to increase the accuracy. The proposed method uses the combination of the mutual structure of the RGB image to keep the structure and joint weighted filter to make the depth planes smooth and fill the regions of discontinuity. Transferring the depth map to another domain gave us the option to implement normalized interpolated convolution to remove the blocky artifacts of around the edges and corners. The comparison with the top 8 methods of the Middlebury benchmark and the ARWR without post-processing proved the performance quality of the proposed method. The value of the average structural similarity index which is about 0.9935 with Google's stereo matching method is another confirmation on the performance of the discussed method.

With respect to the performance of the studied method in this paper, there are still a number of open challenges such as reducing the processing time, while maintaining the same accuracy in real-time applications with low processing power. This challenge motivates our future research activity. In follow-on work it is planned to filter each image in 8-16 dimensional bilateral space instead of employing a normalized interpolation. Preliminary experiments indicate this could improve the speed of the enhanced ARWR by as much as an order of magnitude. This refinement would make the post-processed ARWR algorithm competitive in terms of computation time with the top 2-3 algorithms form Middlebury.

ACKNOWLEDGMENT

The authors would like to thank Dr. Claudia Costache for her helpful comments.

REFERENCES

[1] *Computer Arts/3D World Glossary*, Future plc, Somerset, U.K., 2011.
[2] C. Niclass, M. Soga, H. Matsubara, and S. Kato, "A 100m-range 10-frame/s 340×96-pixel time-of-flight depth sensor in 0.18 μm CMOS," in *Proc. ESSCIRC (ESSCIRC)*, Sep. 2011, pp. 107–110.
[3] C. Niclass *et al.*, "Design and characterization of a 256×64-pixel single-photon imager in CMOS for a MEMS-based laser scanning time-of-flight sensor," *Opt. Exp.*, vol. 20, no. 11, pp. 11863–11881, May 2012.
[4] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2003, pp. I-195–I-202.
[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.
[6] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. 10th Asian Conf. Comput. Vis. Comput. Vis. (ACCV)*, Queenstown, New Zealand, Nov. 2010, pp. 25–38.
[7] S. Hermann and R. Klette, "Iterative semi-global matching for robust driver assistance systems," in *Proc. 11th Asian Conf. Comput. Vis. Comput. Vis. (ACCV)*, Daejeon, South Korea, Nov. 2012, pp. 465–478.
[8] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.
[9] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3017–3024.
[10] R. Kozik, "Improving depth map quality with Markov random fields," in *Image Processing and Communications Challenges 3*, R. S. Choraś, Ed. Berlin, Germany: Springer, 2011, pp. 149–156.
[11] K.-H. Lo, K.-L. Hua, and Y.-C. F. Wang, "Depth map super-resolution via Markov random fields without texture-copying artifacts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 1414–1418.
[12] C. D. Herrera, J. Kannala, P. Sturm, and J. Heikkila, "A learned joint depth and intensity prior using Markov random fields," in *Proc. Int. Conf. 3D Vis. (3DV)*, Jun./Jul. 2013, pp. 17–24.
[13] S. Zheng, P. An, Y. Zuo, X. Zou, and J. Wang, "Depth map upsampling using segmentation and edge information," in *Proc. 8th Int. Conf. Image Graph. (ICIG)*, Tianjin, China, Aug. 2015, pp. 116–126.
[14] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2005, pp. 807–814.
[15] S. Suri, P. Schwind, P. Reinartz, and J. Uhl, "Combining mutual information and scale invariant feature transform for fast and robust multisensor SAR image registration," presented at the 75th Annu. ASPRS, Baltimore, MD, USA, 2009.

[16] Y. S. Heo, K. M. Lee, and S. U. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1094–1106, May 2013.
[17] S. Lee, J. H. Lee, J. Lim, and I. H. Suh, "Robust stereo matching using adaptive random walk with restart algorithm," *Image Vis. Comput.*, vol. 37, pp. 1–11, May 2015.
[18] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Eur. Conf. Comput. Vis. Comput. Vis. (ECCV)*, Stockholm, Sweden, May 1994, pp. 151–158.
[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[20] X. Shen, C. Zhou, L. Xu, and J. Jia, "Mutual-structure for joint filtering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3406–3414.
[21] Y. Zhu and C. Huang, "An improved median filtering algorithm for image noise reduction," *Phys. Procedia*, vol. 25, pp. 609–616, Apr. 2012.
[22] G. R. Arce, "A general weighted median filter structure admitting negative weights," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3195–3205, Dec. 1998.
[23] H. Knutsson and C.-F. Westin, "Normalized and differential convolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1993, pp. 515–523.
[24] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. 36th German Conf. Pattern Recognit. (GCPR)*, Münster, Germany, Sep. 2014, pp. 31–42.
[25] *Middlebury Stereo Evaluation—Version 3*. [Online]. Available: http://vision.middlebury.edu/stereo/eval3/
[26] X. Huang, Y. Zhang, and Z. Yue, "Image-guided non-local dense matching with three-steps optimization," in *Proc. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. III-3. 2016, pp. 67–74.
[27] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, Jan. 2016.
[28] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "MeshStereo: A global stereo model with mesh alignment regularization for view interpolation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2057–2065.
[29] J. T. Barron, A. Adams, Y. Shih, and C. Hernández, "Fast bilateral-space stereo for synthetic defocus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4466–4474.
[30] J. T. Barron, A. Adams, S. YiChang, and C. Hernandez, "Fast bilateral-space stereo for synthetic defocus—Supplemental material," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–15.

HOSSEIN JAVIDNIA received the master's degree in information technology engineering from the University of Guilan, Iran, in 2014. He is currently pursuing the Ph.D. degree in electrical engineering with the National University of Ireland, Galway. His current research interests include image processing, machine vision, and automotive navigation.

PETER CORCORAN (F'10) has co-authored over 300 technical publications and co-inventor on more than 250 granted US patents. His research interests include biometrics, cryptography, computational imaging, and consumer electronics. He is the Editor-in-Chief of the *IEEE Consumer Electronics Magazine* and a Professor with a Personal Chair at the College of Engineering & Informatics at NUI Galway. In addition to his academic career, he is also an Occasional Entrepreneur, Industry Consultant, and Compulsive Inventor.