IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Edge preserving suppression for depth estimation via comparative variation

ISSN 1751-9659
Received on 21st May 2017
Revised 31st October 2017
Accepted on 3rd December 2017
E-First on 12th January 2018
doi: 10.1049/iet-ipr.2017.0506
www.ietdl.org

Eu-Tteum Baek[1], *Yo-Sung Ho*[1] ✉

[1]*Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, South Korea*
✉ *E-mail: hoyo@gist.ac.kr*

Abstract: Most applications in computer vision manage to suppress textures and noise while maintaining meaningful structure based on colour intensity variation, but it is intractable due to texture patterns or error. This study presents an edge-preserving suppression method for depth estimation. The authors formulate a functional energy function based on the relative total intensity and space variation, and they minimise the energy function via iteratively reweighted least squares. Assuming that textural edges most likely correspond to depth discontinuities, they exploit the comparative variations of the colour image to produce a more accurate depth map. The experimental results demonstrate the usefulness of the proposed approach, and show that texture patterns are suppressed while meaningful edges are preserved. According to the results of the depth acquisition methods, the proposed depth estimation methods generate the accurate and robust results.

1 Introduction

Computer vision has been widely studied during the past decades and focused on many tasks such as recognition, motion analysis, scene reconstruction, and image restoration. It has relevance to the theory behind artificial systems that extract information from images. Typically, extracting major information from a computer is tough. As shown in Fig. 1, the human visual system is fully capable of understanding images which involve complex textures, colour contrast, or noise, but computer vision system is more demanding to complex textural images such as brick walls, railroad boxcars, and subways; carpets, sweaters, and other fine crafts contain various geometric patterns. Ambient noise means the noise caused by external influences such as illumination, temperature, or transmission. The pattern of noise typically is irregular. Texture usually refers to surface patterns that are similar in appearance and local statistics. The texture could be regular, near-regular, or irregular.

To remove noise and to obtain meaningful information from a complex texture image, simple linear filters with explicit kernels, such as the mean, Gaussian, and Laplacian filters [1], have been actively implemented in image restoration, blurring/sharpening. Several non-linear filters are also proposed, for example, median filtering [2], weighted median filter [3], bilateral filter [4], and guided image filter [5].

Rudin and co-workers [6] first introduced total variation (TV) regularisation in an image processing context for noise suppression. After that, many approaches were proposed by exploiting TV. As these approaches simply use a weight to enforce structural similarity between the input and output, the TV regulariser has limited ability to distinguish between significant structural and textural edges [7, 8]. Farbman *et al.* optimised an objective function via weighted least squares, and Xu *et al.* introduced L0 gradient minimisation. However, they also did not obtain an optimal solution [9, 10]. Li Xu *et al.* [11] used relative TV method to overcome the weaknesses of previous methods. Even though this approach distinguishes between structure and texture well, the regions of major contours are blurred as shown in Fig. 2.

Fig. 1 *Complex textural images*

Recently, a variety of cost aggregation approaches for stereo matching have been proposed. Yoon and Kweon [12] first presented to filter the cost volume using a joint bilateral filter. The idea is that pixels having a colour similar to the centre pixel are likely to lie on the same object, therefore have similar depth, and effectively preserves depth boundaries. He *et al.* [5] proposed a guided image filter, which has linear runtime in the number of image pixels. This filter shows leading speed and accuracy performance [13]. Yang [14] proposed a non-local cost aggregation method, which enlarges the window size to the whole image. The non-local cost aggregation can be performed very fast by computing a minimum spanning tree derived from the graph. Recently, Zheng *et al.* [15] presented a cross-scale cost aggregation, which is one of the methods to estimate accurate disparity values in homogeneous regions. This method constructs a hierarchical structure to aggregate matching costs.

The acquisition of depth maps is a significant requirement in many three-dimensional (3D) applications. Time-of-flight (ToF) sensors are widely used to generate the depth map in real time, but low resolution is one of the crucial drawbacks. Therefore, many approaches for generating high-resolution depth methods are presented. Kopf *et al.* [16] presented filter-based depth upsampling algorithms exploiting joint bilateral upsampling. This approach generates a high-quality depth map, but a texture copy problem occurs. Thus, Chan *et al.* [17] proposed a noise-aware filter for depth upsampling, which can overcome the texture copy problem. However, these methods often induce the over-blurred problem near the depth discontinuities. Some approaches based on the Markov random field have been proposed for solving its problem [18, 19], but error propagation problem occurs during the optimisation process.

Section 2 explains relative total intensity and space variation. Section 3 introduces a novel cost aggregation for stereo matching

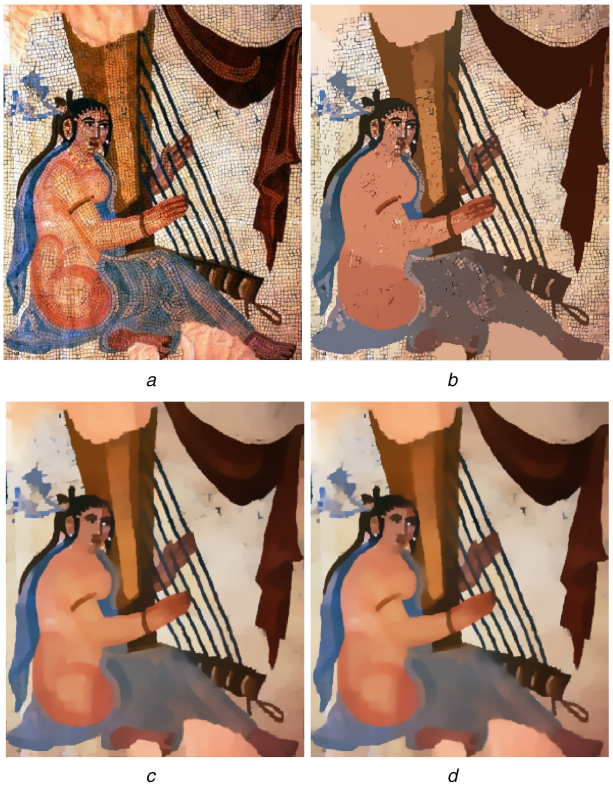

Fig. 2 *Results and comparison*
(a) Original image, ***(b)*** L0 gradient minimisation, ***(c)*** RTV, ***(d)*** Proposed

using relative total intensity and space variation in detail. Section 4 presents a depth upsampling method exploiting relative total intensity and space variation. In Section 5, the proposed technique is validated using experiments. Finally, Section 6 concludes the paper.

2 Relative total intensity and space variation

Variational optimisation deals directly with finding the optimal functions themselves. This means that the solution space is composed of functions as elements. Typically, the variational energy function is composed of two terms which involve a data-driven energy term E_{d} and a smoothness-based energy term E_{s}. A data-driven energy term denotes a reasonable optimisation criterion which a reconstructed image should be close to an input image. A smoothness-driven energy term forces adjacent pixels to have the same or relatively close intensity with comparatively small colour changes inside each region. The energy function incorporating the two terms is formulated by

$$E = E_{\mathrm{d}}(\hat{R}_p, I_p) + \lambda \cdot E_{\mathrm{s}}(\hat{R}_p) \text{where} \quad E_{\mathrm{d}} = \sum_p (\hat{R}_p - I_p)^2, \quad E_{\mathrm{s}} = \sum_p \| \nabla \hat{R}_p \|_2^2 \tag{1}$$

where R is an original unobserved image, $\hat{R}$ is a reconstructed image, and I denotes a given image. $\sum_p \| \nabla \hat{R}_p \|$ is a TV regulariser, expressed as

$$\| \nabla \hat{R}_p \|_2^2 = (\partial \hat{R}/\partial x)^2 + (\partial \hat{R}/\partial y)^2 \tag{2}$$

The TV of a signal measures how much the signal changes between signal values. TV is exploited as a regulariser to take advantage of its edge-preserving nature and is used for applications such as image denoising, inpainting, and deconvolution.

Relative TV (RTV) is a straightforward and efficient method to remove texture while conserving meaningful edges [11]. The inherent characteristic difference between textural and structural edges provides a cue to distinguish those edges as shown in Fig. 3. The textural edges fluctuate considerably, but the structural edges are increased or decreased steadily. RTV exploits the ratio between the sum of absolute variation and the absolute sum of variation [11]. RTV is defined as

$$\mathrm{RTV}(p) = \frac{D_x(p)}{L_x(p) + \varepsilon} + \frac{D_y(p)}{L_y(p) + \varepsilon}$$
$$\text{where} D_x(p) = \sum_{q \in R(p)} g_{p,q} \left| (\partial_x \hat{R})_q \right|, D_y(p) = \sum_{q \in R(p)} g_{p,q} \left| (\partial_y \hat{R})_q \right|$$
$$L_x(p) = \left| \sum_{q \in R(p)} g_{p,q} (\partial_x \hat{R})_q \right|, L_y(p) = \left| \sum_{q \in R(p)} g_{p,q} (\partial_y \hat{R})_q \right| \tag{3}$$

where $g_{p,q}$ is Gaussian convolution, and ε is a small positive number to avoid division by zero. q is a pixel in a window $R(p)$ centred at pixel p. The proposed method employs the RTV method to formulate an objective function. In order to obtain an initial near-optimal solution, the proposed method concerns the weights for intensity and space. Relative total intensity and space variation is expressed as

$$Z_x(p) = \sum_{q \in R(p)} b_{p,q} \left| (\partial_x \hat{R})_q \right|, Z_y(p) = \sum_{q \in R(p)} b_{p,q} \left| (\partial_y \hat{R})_q \right|,$$
$$N_x(p) = \left| \sum_{q \in R(p)} b_{p,q} (\partial_x \hat{R})_q \right|, N_y(p) = \left| \sum_{q \in R(p)} b_{p,q} (\partial_y \hat{R})_q \right| \tag{4}$$
$$\text{where} \quad bp,q = \frac{1}{W_p} \sum_{q \in R(p)} g_{\sigma 1}(\| p - q \|) g_{\sigma 2}(|i_p - i_q|) i_q$$

where $Z(p)$ is windowed TVs, and $N(p)$ is windowed inherent variation. $b_{p,q}$ is a spatial and intensity weight. $g_{\sigma 1}(\cdot)$ denotes spatial weight, and $g_{\sigma 2}(\cdot)$ is a weight for colour difference. W_p is a normalisation term. The relative total intensity and space variation in (3) are used to formulate an objective function. The objective function is written as

$$\arg\min_{\hat{R}} \sum_p (\hat{R}_p - I_p)^2 + \alpha \left(\frac{Z_x(p)}{N_x(p) + \varepsilon} + \frac{Z_y(p)}{N_y(p) + \varepsilon} \right) \tag{5}$$

where $(\hat{R}_p - I_p)^2$ is a data-driven energy term which makes an input and a result not wildly deviate. The second term is a regularisation (smoothness) term that preserves major edges and suppresses textures. We transform it into a matrix form to solve the objective function [20]. The objective function can be written as

$$(\boldsymbol{u} - \boldsymbol{g})^{\mathrm{T}} (\boldsymbol{u} - \boldsymbol{g}) + \alpha (\boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}_x \boldsymbol{A}_x \boldsymbol{W}_x \boldsymbol{D}_x \boldsymbol{u} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{D}_y \boldsymbol{A}_y \boldsymbol{W}_y \boldsymbol{D}_y \boldsymbol{u}) \tag{6}$$

where $\boldsymbol{u}$ and $\boldsymbol{g}$ are the vector representation of $\hat{R}$ and I, respectively, and the matrices $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ are discrete differentiation operators. $\boldsymbol{A}_x$, $\boldsymbol{A}_y$, $\boldsymbol{W}_x$, and $\boldsymbol{W}_y$ are diagonal matrices. The minimisation solution of the linear system in an iteration is represented as

$$\left(1 + D_x^T A_x^t W_x^t D_x + D_y^T A_y^t W_y^t D_y\right) \cdot u^{t+1} = g \tag{7}$$

where $\left(\mathbf{1} + D_x^T A_x^t W_x^t D_x + D_y^T A_y^t W_y^t D_y\right)$ is $\boldsymbol{W}_p^{-1}$ and $\mathbf{1}$ is an identity matrix. The sparse matrix $\boldsymbol{W}_p^{-1}$ is represented as (see (8))

The optimisation technique is similar to an iteratively reweighted least squares (IRLS) [21]. However, the weight matrix is very huge, but most of the elements are zero. Thus, operations using standard dense-matrix are extremely slow and inefficient. In addition, memory is much wasted. This paper exploits a sparse matrix strategy to handle the large matrix. In numerical analysis, a sparse matrix is a matrix in which most of the elements are zero, and the matrix is typically stored as a 2D array. Each entry in the array represents an element $a_{i,j}$ of the matrix and is accessed by the

$$\boldsymbol{W}_p^{-1} = \begin{bmatrix} y_{1,1} & z_{1,2} & 0 & \cdots & 0 & \lambda_1 & 0 & 0 & \cdots & 0 \\ x_{2,2} & y_{2,2} & z_{2,3} & 0 & \cdots & 0 & \lambda_2 & 0 & 0 & \cdots \\ 0 & x_{3,2} & y_{3,3} & z_{3,4} & 0 & \cdots & 0 & \lambda_3 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & x_{w-1,w-2} & y_{w-1,w-1} & 0 & 0 & 0 & 0 & \ddots \\ \lambda_1 & 0 & 0 & 0 & x_{w,w-1} & y_{w,w} & z_{w,w+1} & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & x_{w+1,w} & y_{w+1,w+1} & y_{w+1,w+2} & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 & x_{w+2,w+1} & y_{w+2,w+2} & z_{w+2,w+3} & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & 0 & \ddots & 0 & \cdots & 0 & \ddots & \ddots \end{bmatrix} \tag{8}$$

two indices i and j. A sparse LU factorisation method is used to solve the sparse matrix.

3 Cost aggregation for stereo matching using relative total space and variation

Stereo matching aims to identify the corresponding points and estimate their displacement to generate a depth map. It has been one of the most important tasks in the computer vision field. Many depth estimation methods are presented, and it can be classified into global and local approaches according to the strategies [22, 23]. In this section, a cost aggregation method for stereo matching is presented by using relative total intensity and space variation. Fig. 4 illustrates overall framework of the stereo matching method [24]. Our algorithm performs the following steps: (i) constructing cost volume, (ii) cost aggregation, (iii) disparity selection, and (iv) disparity refinement.

3.1 Cost volume

A raw cost volume is constructed by calculating matching costs for each pixel p at all possible disparity levels between the left the image and the right image. A truncated absolute intensity differences and truncated absolute difference of gradients in the x-direction are chosen. The absolute difference of intensity is represented as

$$D(d) = \sum_{i,j \in W} |I_{\mathrm{r}}(x+i, y+j) - I_{\mathrm{t}}(x+i+d, y+j)| \tag{9}$$

where I_{r} and I_{t} are reference and target images, respectively, the absolute difference of gradients is computed as

$$G(p,d) = |\nabla_x(I_{\mathrm{r}}(p)) - \nabla_x(I_{\mathrm{t}}(p-d))| \tag{10}$$

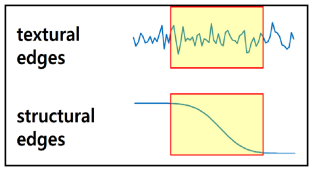

Fig. 3 *Textual edges and structural edges*

where $\nabla_x(I(p))$ denotes the gradient in x-direction computed at pixel p. The overall cost function is expressed as

$$C(p,d) = \lambda \cdot \min(T_{\mathrm{c}}, D(d)) + (1-\lambda) \cdot \min(T_{\mathrm{g}}, G(p,d)) \tag{11}$$

where λ balances the per-pixel matching cost and gradient terms and T_{c}, T_{g} are the census and gradient truncation values.

3.2 Cost aggregation using relative total intensity and space variation

Cost aggregation has a significant impact on stereo matching methods because it enforces piecewise constancy of disparity, over the support region of each pixel. Therefore, we use the proposed relative total intensity and space variation to aggregate each level of the cost volume. Exploiting a reference image I_{r}, the proposed relative total intensity and space variation is calculated to compute the aggregated cost as follows:

$$C^R(p,d) = \boldsymbol{W}_p C(p,d) \tag{12}$$

where $C^R(p, d)$ denotes the aggregated cost using the proposed relative total intensity and space variation, $\boldsymbol{W}_p$ is the inverse weight matrix computed based on the structural vector $\boldsymbol{u}$. The inverse weight matrix $\boldsymbol{W}_p$ depends on image I_{r}, which is the reference image. $\boldsymbol{W}_p$ is defined as follows:

$$\boldsymbol{W}_p = (1 + D_x^T A_x^t W_x^t \boldsymbol{D}_x + D_y^T A_y^t W_y^t \boldsymbol{D}_y)^{-1} \tag{13}$$

Here, the matrices $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ are discrete differentiation operators. $\boldsymbol{A}_x$, $\boldsymbol{A}_y$, $\boldsymbol{W}_x$, and $\boldsymbol{W}_y$ are diagonal matrices from (5).

3.3 Disparity selection and post-processing

Once the cost volume is aggregated, the winner-takes-all strategy is applied to choose the best disparity value for each pixel p. Disparity selection method is defined as

$$d_p = \arg\min_{d \in D} C^R(p,d) \tag{14}$$

where D denotes the set of all allowed disparities. Streak-like artefacts in the result are produced. Therefore, this method uses a

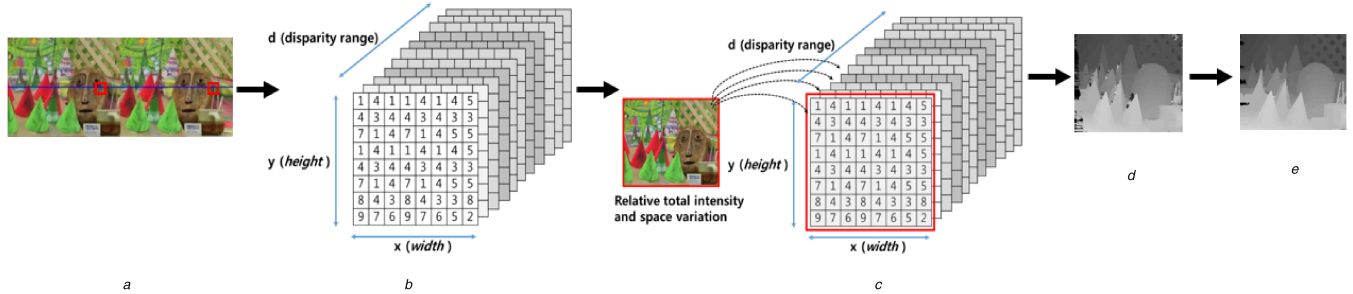

Fig. 4 *Whole process of the proposed stereo matching method*

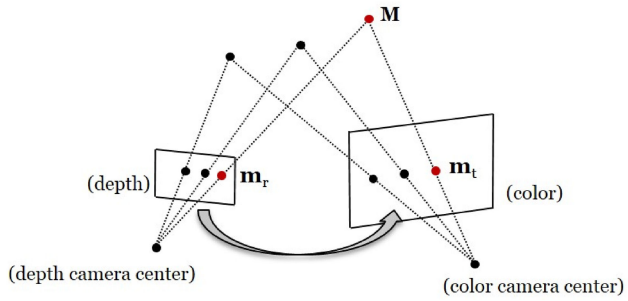

Fig. 5 *Projection from a low-resolution depth map to a high-resolution sparse depth map in the colour camera coordinate system*

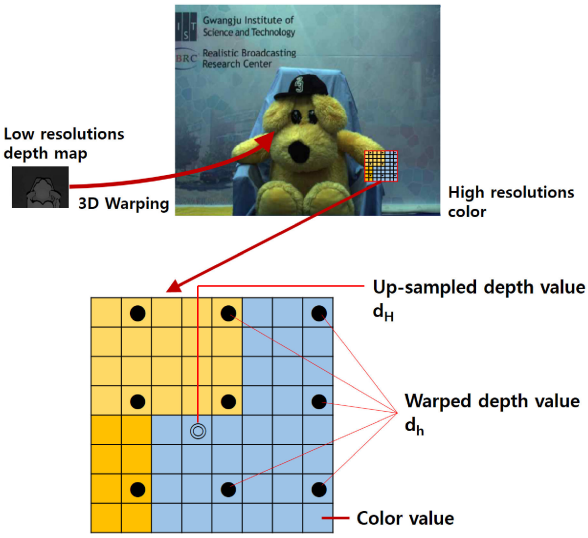

Fig. 6 *Structure after the 3D warping using low-resolution depth map*

weighted median filter to smooth the filled regions for reducing these artefacts [3].

4 Depth up-sampling

The limitation of low resolution captured by active depth cameras such as ToF camera and structured light depth camera is critical in computer vision applications. Therefore, this section presents a new interpolation method for the low-resolution depth map to effectively enhance the resolution. The solution is also globally optimised with IRLS.

4.1 3D warping

The hybrid camera system combining an active depth camera and multiple colour cameras are captured at different positions. To obtain the high-resolution depth map on the same position of the corresponding colour image, the 3D warping technique is implemented [25]. As shown in Fig. 5, the 3D image warping consists of two steps: backward projection with depth data and forward projection. Let m_l and $_{mt}$ be the corresponding pixels in the viewpoint image of the depth camera and the viewpoint image of the colour camera, respectively. Based on the pin-hole camera model, the pixel point m_l can be defined by the camera parameters as

$$\boldsymbol{m}_r = \boldsymbol{K}_r \cdot \boldsymbol{R}_r \cdot \mathbf{M} + \boldsymbol{K}_r \cdot \boldsymbol{t}_r \tag{15}$$

where $\boldsymbol{M}$ is the point in the world coordinates, and $\boldsymbol{m}_r$ is the point in the depth camera point. $\boldsymbol{K}$ is an intrinsic parameter, $\boldsymbol{R}$ denotes a rotation matrix, and $\boldsymbol{t}$ represents a translation vector. The next step is finding the corresponding pixel position m_t in the view point of the colour camera. The point in the world coordinates $\boldsymbol{M}$ is projected onto the viewpoint of the colour camera using its camera parameters as

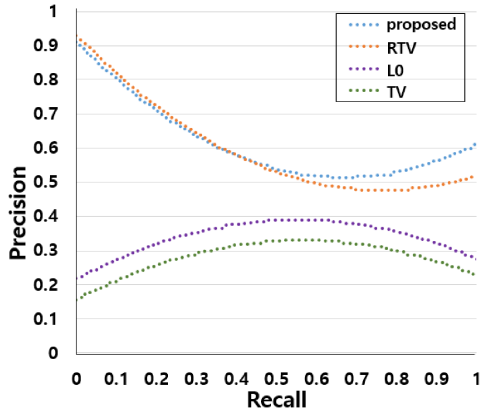

Fig. 7 *Precision–recall curve*

$$\begin{aligned}\boldsymbol{m}_c &= \boldsymbol{A}_t \cdot \boldsymbol{R}_t \cdot \mathbf{M} + \boldsymbol{A}_t \cdot \boldsymbol{t}_t \\ &= \boldsymbol{A}_t \cdot \boldsymbol{R}_t \cdot \boldsymbol{R}_r^{-1} \cdot \boldsymbol{A}_r^{-1} \cdot \boldsymbol{m}_r - \boldsymbol{A}_t \cdot \boldsymbol{R}_t \cdot \boldsymbol{R}_r^{-1} \cdot \boldsymbol{t}_r + \boldsymbol{A}_t \cdot \boldsymbol{t}_r\end{aligned} \tag{16}$$

4.2 Depth up-sampling using relative total intensity and space variation

After warping depth data, a high-resolution sparse depth map in the colour camera coordinate is generated. Fig. 6 represents the resolution difference between the colour image and the depth map of the hybrid camera system, and it shows the structure after the 3D warping process. This algorithm exploits a nearest-neighbour interpolation method to interpolate the high-resolution sparse depth map. As in the case of (12), exploiting a reference image I_r, the proposed relative total intensity and space variation can be used to produce a more accurate up-sampling result as follows:

$$d_{\mathrm{H}} = \boldsymbol{W}_p d_h^i \tag{17}$$

where d_h^i is the interpolated depth map, and d_{H} is the high-resolution dense depth map, and $\boldsymbol{W}_p$ is the inverse weight matrix computed based on the structural vector $\boldsymbol{u}$.

5 Experimental results

This section evaluates the proposed method quantitatively and quantitatively on both the synthetic data and real data. The comparison is performed with other conventional methods. First, the proposed algorithm was tested using a dataset which contains 200 'structure + texture' images, provided by [11]. A precision–recall method is chosen to evaluate the results quantitatively. Typically, precision–recall curves are used in binary classification to study the output of a classifier. After applying the conventional smoothing methods and the proposed method, this approach extracts structures from the results. Moreover, it compares the structure images and the ground truth label. The precision–recall curves for the four algorithms are plotted in Fig. 7. The results of precision–recall curves show that the proposed methods outperform conventional method.

Fig. 2*a* shows a 'Bishapur zan' image. Many marbles form the image. Thus, making an extraction is very challenging. The results from conventional methods are presented from Fig. 2*b* and *c*. L_0 gradient minimisation preserves and enhances sharp edges, but it does not deal with textures. RTV removes textural edges while maintaining meaningful edges. The proposed method suppresses textural edges, preserves major edges, and especially sharpens edges, compared with the conventional methods. Close-ups in Figs. 8*a* and *b* depict the results of proposed method have clear edges comparing with RTV. Fig. 9 represents the 1D examples of clear edge enhancement. The proposed method performs as an image structure extraction like RTV but has better behaviour near the edges.

Fig. 10 represents the results of the proposed method. Textural edges are well suppressed while preserving primary structures. The

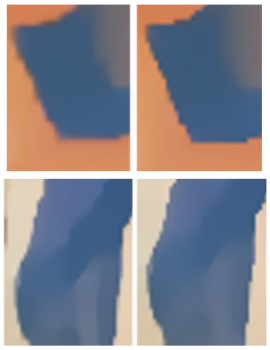

Fig. 8 *Results of close-ups and comparison*
(a) Results of RTV, ***(b)*** Results of proposed

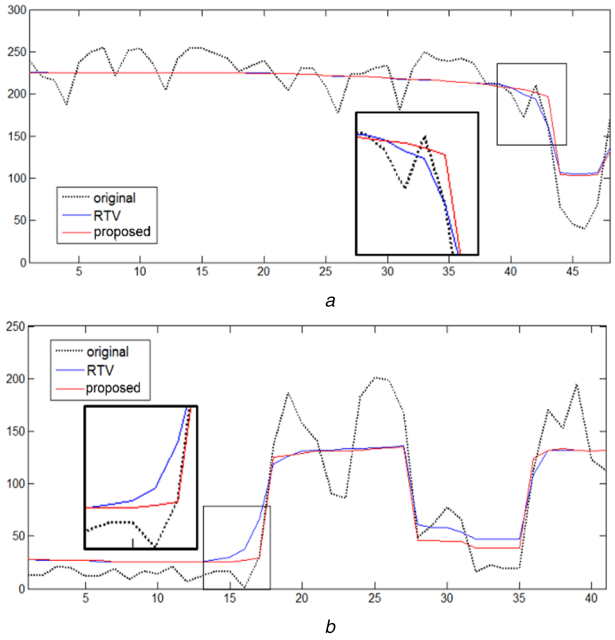

Fig. 9 *Original and denoised signal with RTV and proposed methods*

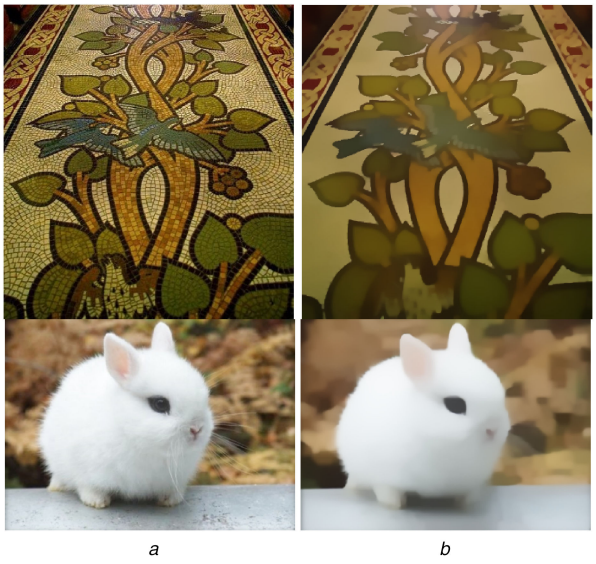

Fig. 10 *Results of proposed method. The first column images are the original images and the second-column images are deblurred images*
(a) Original image, ***(b)*** Proposed

datasets computed using the conventional filter-based cost aggregation methods, and the proposed cost aggregation method are presented in Fig. 11. Fig. 11 shows that the proposed cost aggregation method provides more accurate results, and errors are

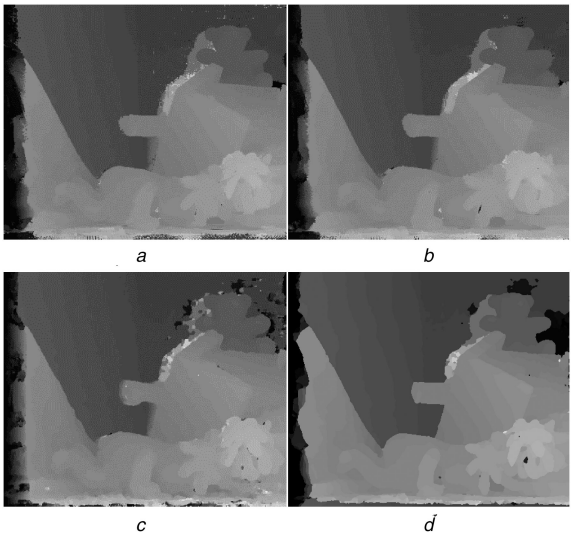

Fig. 11 *Visual comparison with conventional cost aggregation methods and proposed cost aggregation method*
(a) Bilateral filter, ***(b)*** Guided image filter, ***(c)*** Box filter, ***(d)*** Proposed

Fig. 12 *Results of the proposed stereo matching method. First column images are original images, second column images are ground truth images, and last column images are the results of the proposed method*

well removed from the Middleburry dataset [26]. Fig. 12 shows the results of the proposed stereo matching method and its ground truth.

To evaluate the performance of proposed stereo matching method objectively, we exploit the percentages of mismatching pixels (BPR) with known ground truth disparity. Table 1 summarises the percentage of the bad matching pixels between the results of the proposed method and ground truths. This measure is computed non-occluded denoted as 'nonocc'. The results exhibit robust performance compared to conventional methods. Fig. 13 represents the up-sampling results of art, book, cone, laundry, and teddy from the Middlebury dataset.

For comparison, additional experiments of the stereo matching were carried out with Middlebury dataset [26]. Fig. 14 shows the results of the proposed stereo matching method. The proposed method generates the more accurate depth maps in the texture region than the conventional method. Table 2 summarises the quantitative comparison results on the Middlebury dataset [26]. Root mean squared error (RMSE) measures objective depth map quality. RMSE is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it makes a general purpose error metric for numerical predictions. It can be observed from Table 2 that the proposed method performs well compared to other approaches. This approach also exploited real-world examples to test the proposed depth up-sampling method.

Table 1 Performance comparison with conventional methods and proposed method

Algorithm		CSBP [27]	Box	Bilateral [12]	Non-local [28]	Segment-tree [29]	Guided [5]	Proposed
Tsukuba	nonocc	14.54	2.00	6.08	5.46	6.31	5.77	5.01
Venus	nonocc	16.26	1.48	2.03	2.58	3.18	2.03	1.59
Teddy	nonocc	11.2	11.10	7.14	7.19	8.22	4.38	7.02
Cones	nonocc	15.69	5.98	9.37	8.21	9.43	7.61	4.55

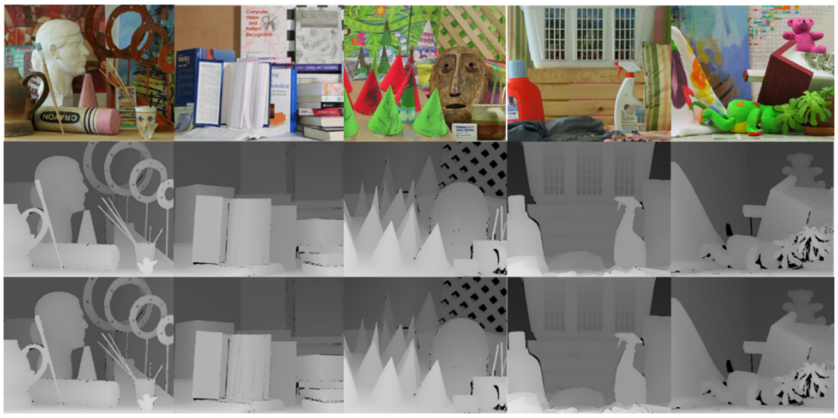

Fig. 13 *Upsampling results of art, book, cone, laundary, and teddy from the Middleburry dataset. First row images are original images, second-row images are 2× upsamping results, and last row images are 4× upsamping results*

Fig. 14 *Results of the proposed stereo matching method. Top row images are colour images, second-row images are ground truth images, third-row images are results of bilateral filter [12], fourth-row images are the result of guided filter [5], and bottom row images are the results of the proposed method*

Fig. 15 display the results of the proposed depth up-sampling method using the images captured by a ToF camera and a colour camera. Fig. 16 represents cafe and newspaper sequences obtained from GIST [34]. Since a depth map captured by ToF camera has low resolution and has fundamental problems, such as the ambiguity of depth information in the shiny and dark surface, these real-world examples are challenging with complicated boundaries and thin objects.

6 Conclusions

Ambient noises and textures are factors that affect the quality of results. Therefore, it is very important to remove textures and noises in the data before some image processing such as edge detection, object recognition, and image segmentation. In this paper, an edge-preserving smoothing method for depth estimation via comparative variation is proposed. After formulating the energy function, reweighted least squares methods are used to minimise the energy function. From the experimental results, it is confirmed that the proposed method preserves major structure while texture edges are removed, and has better behaviour near the edges, compared with the conventional algorithms, and it generates the accurate results of the stereo matching and the depth up-sampling methods.

7 Acknowledgment

This work was supported by 'The Cross-Ministry Giga KOREA Project' grant funded by the Korea government (MSIT) (GK17C0100, Development of Interactive and Realistic Massive Giga- Content Technology).

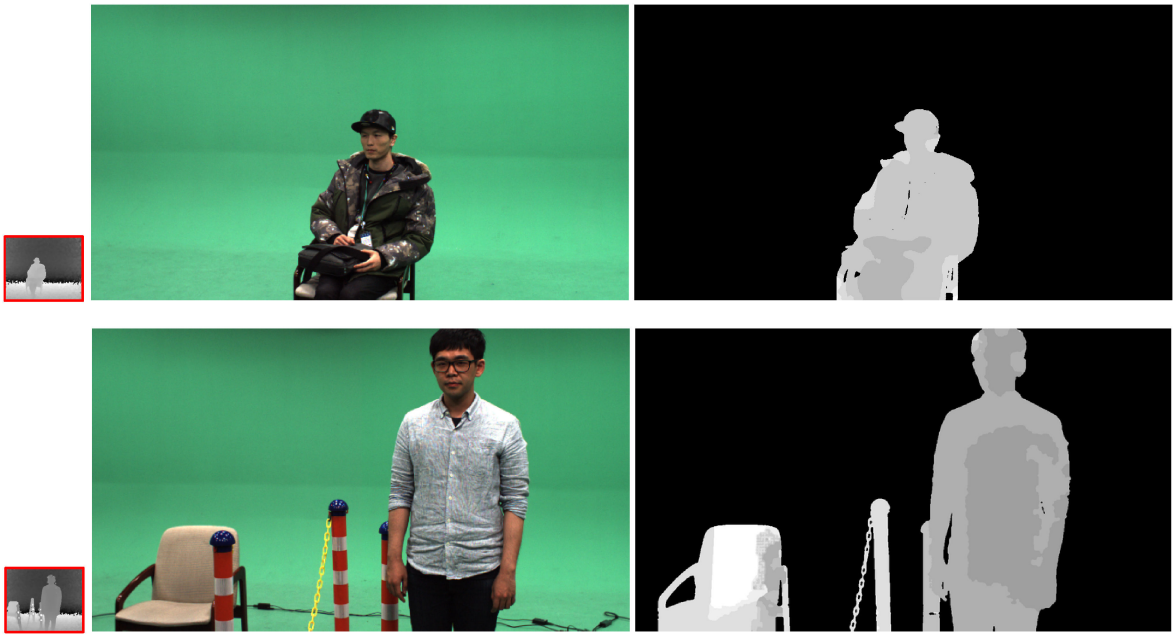

Fig. 15 *Upsampling results of depth maps captured by the ToF camera which resolution is 176 × 144, and the colour camera which resolution is 1280 × 720*

Fig. 16 *Upsampling using real datasets (newspaper and cafe) [34]*

Table 2 Quantitative comparison of the synthetic data from the middlebury dataset

	Art		Laundry		Doll		Book	
	2×	4×	2×	4×	2×	4×	2×	4×
CLMF [30]	1.19	1.77	0.96	1.56	0.87	1.44	0.9	1.48
JGF [31]	2.36	2.74	2.18	2.4	2.09	2.24	2.12	2.25
NLM-MRF [32]	1.66	2.47	1.34	1.73	1.19	1.56	1.19	1.47
MRF [33]	1.24	1.69	0.78	1.12	0.75	1.04	0.74	1.04
proposed	0.97	1.76	0.96	1.65	1.24	1.96	1.07	1.87

8 References

[1] Gonzalez, R., Woods, R.: '*Digital image processing*' (Prentice Hall, Upper Saddle River, NJ, USA, 2002, 2nd edn.)

[2] Huang, T., Yang, G., Tang, G.: 'A fast two dimensional median filtering algorithm', *IEEE Trans. Acoust. Speech Signal Process.*, 1979, **27**, (1), pp. 13–18

[3] Yin, L., Yang, L., Gabbouj, M., *et al.*: 'Weighted median filters: a tutorial', *IEEE Trans. Circuits and Syst. II, Analog Digit. Signal Process.*, 1996, **43**, (3), pp. 157–192

[4] Tomasi, C., Manduchi, R.: 'Bilateral filtering for gray and color images'. Proc. IEEE (ICCV), 1998, pp. 839–846

[5] He, K., Sun, J., Tang, X.: 'Guided image filtering', Proc. European Conf. on Computer Vision, Heraklion, Crete, September 2010, pp. 1–14

[6] Osher, S., Rudin, L., Fatemi, E.: 'Nonlinear total variation based noise removal algorithms', *Physica D*, 1992, **60**, pp. 259–268

[7] Yin, W., Goldfarb, D., Osher, S.: 'Image cartoon-texture decomposition and feature selection using the total variation regularized l1 functional'. Int. Workshop Variational, Geometric, and Level Set Methods in Computer Vision (VLSM), 2005, pp. 73–84

[8] Aujol, J., Gilboa, G., Chan, T., *et al.*: 'Structure-texture image decomposition —modeling, algorithms, and parameter selection', *Int. Comput. Vis.*, 2006, **67**, (1), pp. 111–136

[9] Kass, M., Solomon, J.: 'Smoothed local histogram filters', *ACM Trans. Graph.*, 2010, **29**, (4), pp. 100:1–100:10

[10] Xu, L., Lu, C., Xu, Y., *et al.*: 'Image smoothing via l0 gradient minimization', *ACM Siggraph Asia*, 2011, **30**, (6), pp. 174:1–174:12

[11] Xu, L., Yan, Q., Xia, Y., *et al.*: 'Structure extraction from texture via relative total variation', *ACM Trans. Graph.*, 2012, **31**, (6), pp. 139:1–139:10

[12] Yoon, K., Kweon, I.: 'Adaptive support-weight approach for correspondence search', *PAMI*, 2006, **28**, (4), pp. 650–656

[13] Hosni, A., Bleyer, M., Rhemann, C., *et al.*: 'Real-time local stereo matching using guided image filtering'. ICME, 2011, pp. 1–6

[14] Yang, Q.: 'A non-local cost aggregation method for stereo matching'. IEEE Conf. Computer Vision and Pattern Recognition, 2012, pp. 1402–1409

[15] Zheng, K., Fang, Y., Min, D., *et al.*: 'Cross-scale cost aggregation for stereo matching'. IEEE Conf. Computer Vision and Pattern Recognition, 2014, pp. 1590–1597

[16] Kopf, J., Cohen, M., Lischinski, D., *et al.*: 'Joint bilateral upsampling', *ACM Trans. Graph.*, 2007, **26**, (3), pp. 1–5

[17] Chan, D., Buisman, H., Theobalt, C., *et al.*: 'A noise-aware filter for real-time depth upsampling'. ECCV Workshop on Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications, 2008, pp. 1–12

[18] Diebel, J., Thrun, S.: 'An application of markov random fields to range sensing', *Adv. Neural Inf. Process. Syst.*, 2006, **18**, pp. 291–298

[19] Kim, D., Yoon, K.: 'High-quality depth map up-sampling robust to edge noise of range sensors'. Int. Conf. Image Processing, 2012, pp. 553–556

[20] Lischinski, D., Farbman, Z., Uyttendaele, M., *et al.*: 'Interactive local adjustment of tonal values', *ACM Trans. Graph.*, 2006, **25**, (3), pp. 646–653

[21] Farbman, Z., Fattal, R., Lischinski, D., *et al.*: 'Edge-preserving decompositions for multi-scale tone and detail manipulation', *ACM Trans. Graph.*, 2008, **27**, (3), pp. 67:1–67:10

[22] Kolmogorov, V., Zabih, R.: 'Computing visual correspondence with occlusions using graph cuts'. IEEE Int. Conf. Computer Vision, 2001, pp. 508–515

[23] Ben-Ari, R., Sochen, N.: 'Stereo matching with Mumford-shah regularization and occlusion handling', *IEEE Trans. Pattern Anal. Mach. Intell*, 2010, **32**, pp. 2071–2084

[24] Baek, E., Ho, Y.: 'Occlusion and error detection for stereo matching and hole-filling using dynamic programming', *Electron. Imaging*, 2016, p. 5o: 1–6

[25] Fehn, C.: 'Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV', *SPIE Stereoscopic Disp. Virtual Real. Syst. XI*, 2004, **5291**, pp. 93–104

[26] Scharstein, D., Pal, C.: 'Learning conditional random fields for stereo'. IEEE Conf. Computer Vision and Pattern Recognition, 2007, pp. 1–8

[27] Yang, Q., Wang, L., Ahuja, N.: 'A constant-space belief propagation algorithm for stereo matching'. IEEE Conf. Computer Vision and Pattern Recognition, 2010, pp. 1458–1465

[28] Liu, T., Zhang, P., Luo, L.: 'Dense stereo correspondence with contrast context histogram, segmentation-based two-pass aggregation and occlusion handling' in Wada, T., Huang, F., Lin, S. (Eds.): '*Advances in image and video technology*' (Springer, Berlin, Heidelberg, 2009), pp. 449–461

[29] He, K., Mei, X., Sun, X., *et al.*: 'Segment-tree based cost aggregation for stereo matching'. IEEE Conf. Computer Vision and Pattern Recognition, 2013, pp. 313–320

[30] Lu, J., Shi, K., Min, D., *et al.*: 'Cross-based local multipoint filtering'. IEEE Conf. Computer Vision and Pattern Recognition, 2012, pp. 430–437

[31] Liu, M., Tuzel, O., Taguchi, Y.: 'Joint geodesic upsampling of depth images'. IEEE Conf. Computer Vision and Pattern Recognition, 2013, pp. 169–176

[32] Park, J., Kim, H., Tai, Y., *et al.*: 'High quality depth map upsampling for 3d-tof cameras'. Proc. Int. Conf. Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 1623–1630

[33] Diebel, J., Thrun, S.: 'An application of markov random fields to range sensing'. Conf. Neural Information Processing Systems (NIPS), 2005, pp. 291–298

[34] Ho, Y., Lee, E., Lee, C.: 'Multiview video test sequence and camera parameters'. ISO/IEC JTC1/SC29/WG11 MPEG2008/M15419, 2008