2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Dependable Dense Stereo Matching by Both Two-layer Recurrent Process and Chaining Search

Sehyung Lee, Youngbin Park, and Il Hong Suh, *SeniorMember,IEEE*

Abstract— **Disparity computation in occluded or texture-less regions is considered to be a fundamental issue in dense stereo matching, but there is another practical issue that must be resolved before it can be used effectively in various robotics applications. This issue is the problem of intensity difference between corresponding pixels of an image pair. To tackle such problems, we present a dependable stereo matching algorithm using two-layer recurrent process and chaining search. Two-layer process integrates pixel and region-levels information through recurrent interaction. To estimate the precise disparities in occluded regions, reliable disparities in non-occluded region are propagated to occluded regions by the proposed chaining search. To test our algorithm, it was compared with two outstanding algorithms in Middlebury benchmark using Gaussian noisy images. The results validated the effectiveness of our approach.**

I. INTRODUCTION

In robotics applications, depth information is important to understand space and objects and to execute tasks efficiently. The recently launched Microsoft Kinect device has become a widely used means of acquiring depth data from the environment due to its competitive advantages, namely, price, resolution, frame rate, accuracy, etc. However, Kinect does not function well in outdoor environments, especially environments exposed to direct sunlight, and its minimum-maximum working range is inadequate for some robotics applications. These limitations mean that Kinect and devices like it are an imperfect solution to the requirements of robotics applications. Thus, stereo vision is still considered a promising area of research for providing robots with the ability to estimate depth.

In this study, we investigate the use of the dense stereo matching algorithm in various robotics applications. In the field of computer vision, the most well-known challenge in dense stereo matching is the calculation of disparity in all image regions, including occluded and textureless regions. However, a practical problem arises when the target application for stereo matching lies in robotics. In most computer vision studies, stereo depth estimation is performed based on the assumption that corresponding pixels in the two images have the same intensity. For instance, the Middlebury benchmark data set is composed of such ideal pairs of images. However, pairs of stereo images captured in real life environments, such as those encountered in robotics applications, often have pixels with different intensities at corresponding locations due to non-uniform illumination [1]. Therefore, the two problems that need to be addressed in order to be able to utilize stereo depth maps in robotics applications are the classic disparity problem and the practical non-uniform illumination problem.

Scharstein & Szeliski [2] described most existing dense stereo algorithms using the following four basic building blocks:

- Computation of a matching cost function for every pixel in both input images.
- Aggregation of matching costs computed inside support regions for every pixel in each image.
- Finding the optimum disparity value for every pixel of one image.
- Refining the resultant disparity map.

This study focuses on the third and fourth procedures. Dense matching algorithms used for disparity computation and refinement are classified as local or global. Local approaches [3], [4] compute the disparity at a given point based on intensity values within a finite support region. However, the disparity of each pixel is calculated independently, without considering the disparities of neighboring pixels. Global methods aim to minimize a global cost function, which combines data and smoothness terms and takes into account the whole image.

Global methods are further classified into three categories according to the type of elements that are assigned depth: pixel-based, region-based, and pixel- and region-based. Pixel-based methods calculate depth for each pixel [5], [6]. Region-based algorithms estimate depth for each region, improving robustness against local uncertainty but reducing the level of detail of the disparity map at pixel-level. In this approach, it is assumed that depth discontinuities occur on the boundaries of segmented regions, where a segment is considered to be a region [7]. In pixel/region-based methods, pixel-layer and region-layer depth information is combined to obtain a more reliable and detailed results [8], [9], [10].

In order to address the two problems mentioned earlier in this paper, the proposed model is based on a two-layer recurrent process and chaining search (TRC). Most global algorithms are computationally expensive but very accurate, while area-based local algorithms are less accurate but computationally inexpensive. The proposed model can be said to lie in the middle of these two types of

Sehyung Lee is with the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, shl@incorl.hanyang.ac.kr

Youngbin Park is with the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, pa9301@incorl.hanyang.ac.kr

Il Hong Suh is with the Division of Computer Science and Engineering, College of Engineering, Hanyang University, Korea. All correspondences should be addressed to Il Hong Suh, ihsuh@hanyang.ac.kr

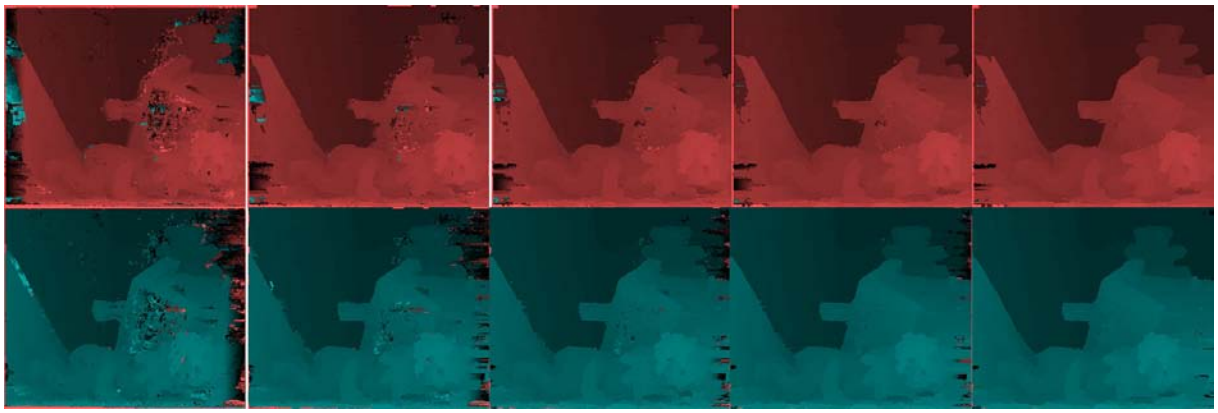

Fig. 1. Changes in the disparity map over several iterations. The top and bottom rows show the left and right disparity maps, respectively. The disparity is represented using saturation and hue.

approaches; while disparity computation is not performed by global optimization over the entire image, the disparities of neighboring pixels are factored into the depth estimation of each pixel. In addition to employing this semi-global approach, the proposed model utilizes a pixel/region-based approach.

We develop a two-layer structure, consisting of the pixel and region layers, and integrate pixel- and region-level information through recurrent interaction. Bayerl & Neumann [11] mentioned that the human visual system shows similar properties. At the pixel layer, in order to precisely estimate depth in the occluded regions, we develop a chaining search algorithm to propagate reliable disparity from non-occluded regions to occluded regions.

II. STEREO MATCHING

In the proposed model, over-segmentation is employed for preprocessing of stereo matching. A two-layer recurrent process and chaining search is then executed to achieve reliable dense stereo matching. This section describes our model in more detail.

A. Over-Segmentation

Many stereo matching algorithms [7]–[10] employ a segmentation method during preprocessing; these algorithms assume that pixels belonging to the segment have similar disparities. This assumption often fails when segments straddle disparity boundaries. Over-segmentation is a common technique used to resolve this issue, and is often implemented using the Mean-shift algorithm [12]. However, as dividing an image into small segments takes a significant amount of time, we developed a simpler and faster over-segmentation method.

The proposed approach assumes that a segment is a small extension of a homogeneous region, and thus, they usually contain similar disparities. Our over-segmentation procedure involves two steps. In the first step, homogeneous regions are detected. In the second step, homogeneous regions are used as seed points for the region-growing technique, which is used to construct the initial over-segmentation map. The decision to include a pixel in the homogeneous region is based on the variance and mean of surrounding pixels. A pixel (p_x, p_y) is included in a homogeneous region,

$$\begin{aligned} if\ \sigma^2(p_x,p_y,n_b) - \sigma^2(p_x,p_y,n_s) \le Th_{\sigma^2}\ and \\ \mu(p_x,p_y,n_b) - \mu(p_x,p_y,n_s) \le Th_{\mu}, \end{aligned} \quad (1)$$

where p_x, p_y are the x- and y-coordinates of a pixel, n denotes the window size, and Th_{σ^2}, Th_{μ} are the threshold values of the variance and average, respectively. A pixel is included in the homogeneous region if the differences between the average and variance of the large window n_b and small window n_s are less than the respective threshold values Th_{σ^2}, Th_{μ}. Homogeneous regions are composed of a set of adjacent pixels satisfying Eq. (1). Next, the threshold values Th_{σ^2} and Th_{μ} are increased to detect new homogeneous regions. If the detected pixels are contiguous to an existing homogeneous region, they are added to that homogeneous region. Otherwise, new homogeneous regions are created. Finally, the over-segmented image is constructed by the region growing method, using the homogeneous regions as seeds. A segment derived in the over-segmentation step is considered as a region in subsequent steps.

B. Two-layer Recurrent Process and Chaining Search

The disparity of a pixel can be computed using local methods only, but the results may not be reliable. Higher information is also needed to obtain a reliable estimate of disparity. Our model is composed of two layers, the pixel layer and the region layer. The region layer is comprised of a set of segments obtained in the over-segmentation stage. They interact mutually through a feed-forward process from the pixel layer to the region layer, and a feedback process from the region layer to the pixel layer and chaining search. In the feed-forward process, the disparity at the region level is estimated, while the disparity at the pixel level is updated in the feedback process.

In the feed-forward process, the disparity for each region s^j is estimated as follows:

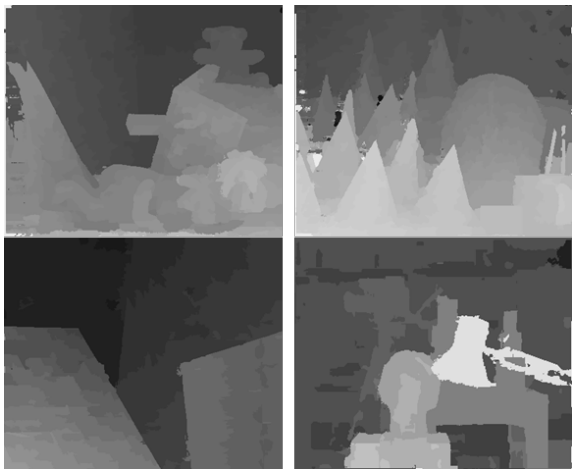

Fig. 2. Disparity maps constructed using Middlebury data set.

Images	Tsukuba	Venus	Teddy	Cones	AVBP
Bad pixels	2.34%	0.28%	6.17%	4.05%	6.35%

TABLE I
PERCENTAGE OF BAD PIXELS

$$\begin{aligned} P(s_{t+1}^j = d) &= \sum_{x^{1:N}} P(s_{t+1}^j = d, x_t^{1:N} | I) \\ &\approx \sum_{x^{1:N}} P(s_{t+1}^j = d | x_t^{1:N}) P(x_t^{1:N} | I) \\ &\approx P(s_{t+1}^j = d | x_t^{1:N} = d) P(x_t^{1:N} = d | I). \end{aligned} \quad (2)$$

where s^j is the region containing pixels $x^{1:N}$ belong. N is the number of pixels included in the j-th region, and I is stereo pair image. The second row of Eq. (2) is rewritten as the third row because conditional probability is estimated only when the disparity of a pixel and a region is identical.

The initial probability of a disparity in the pixel-layer is defined as follows:

$$P(x_0^i = d | s_0^j, I) = P(x_0^i = d | I) = \alpha \frac{1}{SAD(i,d) + \varepsilon}, \quad (3)$$

where x^i represent a pixel, d is its disparity, α is a normalization constant, and $\varepsilon = 0.001$. The matching score is calculated using the sum of absolute differences (SAD) method, and depends on the similarity of the color of the pixels. SAD is calculated for each reference pixel by comparing against pixels within a particular search range of the matching image. We add a small constant to ensure that the matching score is not zero. The disparity probability is computed by taking the inverse of the matching score and normalizing it.

The initial disparity maps of the left and right images are estimated using Eq. (3). As mentioned earlier, if only local matching is processed, the disparity probability is sensitive to the intensity of the neighboring pixels and the SAD window size. After disparities of all pixels have been calculated, the results are validated through a mutual consistency check (MCC) process. The MCC is defined as

$$MCC^p(x_t^i) = \begin{cases} \lambda, \text{ if } D_r^p(x_t^i) - D_m^p(x_t^i - D_r^p(x_t^i)) \leq 1 \\ 1 - \lambda, \text{ otherwise} \end{cases} \quad (4)$$

where the superscript p denotes the pixel-layer, and subscripts r and m refer to the reference and matching images, respectively. $D_r^p(x_t^i)$ is the disparity of x_t^i, $D_m^p(x_t^i - D_r^p(x_t^i))$ is the disparity of the pixel corresponding to x_t^i, and λ is a constant greater than 0.5.

After the MCC process, all pixels are classified as either stable or unstable pixels. A stable pixel is one that has the value λ; the disparity of a stable pixel is quite reliable. On the other hand, unstable pixels have a value of $1-\lambda$ and their disparities are ambiguous. Unstable pixels are usually located in low-texture and occluded areas.

The estimated pixel disparities are propagated to their respective regions. The probability of a region disparity is calculated as the weighted sum of the pixel disparities as follows:

$$P(s_{t+1}^j = d | x_t^{1:N} = d) = \alpha \sum_{x^i \in s^j} P(x_t^i = d | I) MCC^p(x_t^i), \quad (5)$$

Stable pixels have a greater influence on the disparity of a region than unstable pixels, which means that the MCC determines the weighting. In addition, pixel disparities that have a high matching score and few well-matched competitors have greater influence on the disparity of the region. After the feed-forward process, new disparity maps are constructed at the region level. These disparity maps are validated through a region-level MCC process, which is defined as

$$MCC^r(s_t^j) = \begin{cases} \lambda, \text{ if } \sum_k \left| D_r^r(s_t^j) - D_m^r(\prime s_t^k) \right| \leq 1 \\ 1 - \lambda, \text{ otherwise} \end{cases} \quad (6)$$

where $\prime s_t^k$ is the k-th region which overlaps with s_t^j in the matching image when the region s_t^j is projected according to its calculated disparity. D_r^r is a region level disparity map of the reference image, and D_m^r is a region level disparity map of the matching image. Next, regions are labeled as either stable or unstable. Unstable regions usually have a large number of pixels located in an occluded area. Chaining search is performed for all pixels belonging to an unstable region as follow:

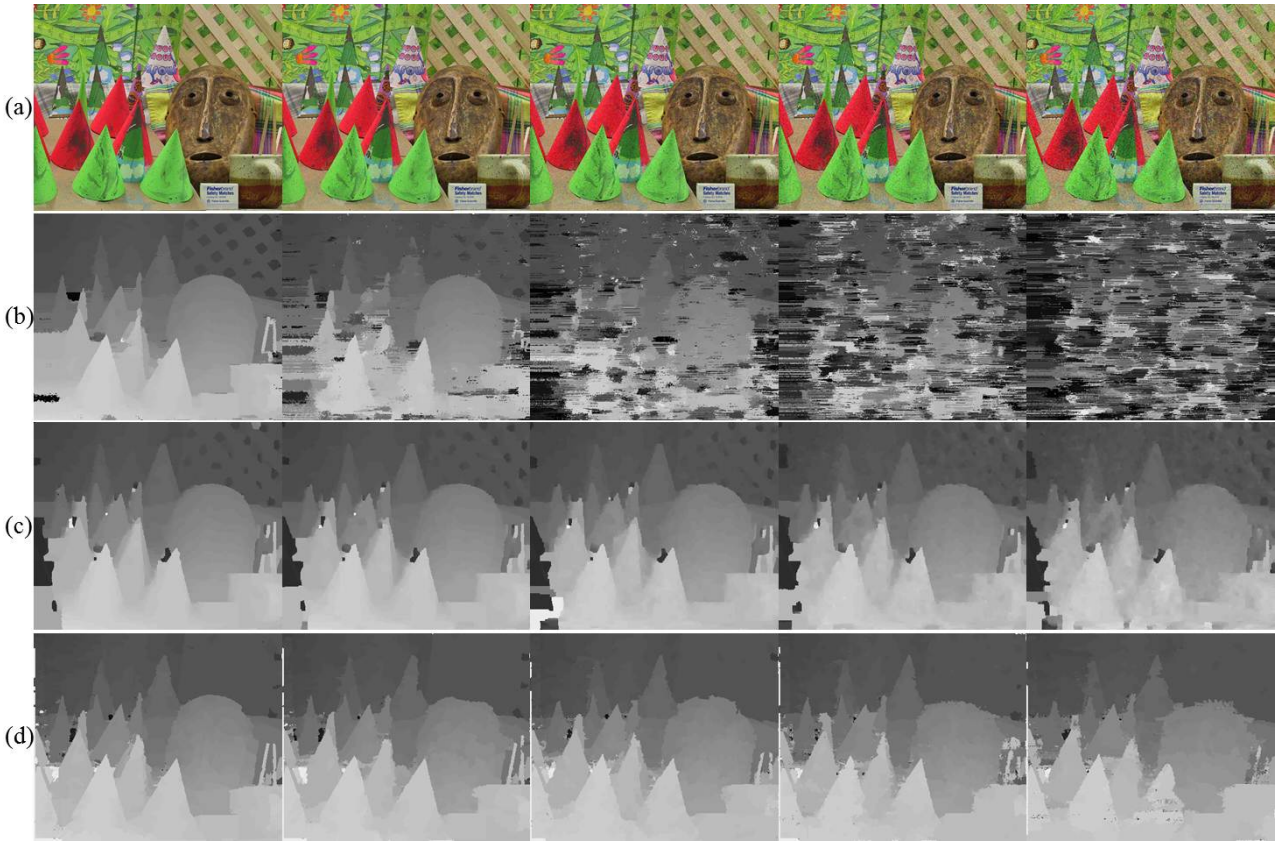

Fig. 3. The first row shows the left images of stereo pairs with varying PSNR. The second, third, and fourth rows show the results of the CVF method, the DBP method, and the proposed method

Step 1.

$$'x_t^i = x_t^i - D_r^p(x_t^i)$$

Step 2.

If $MCC^p('x_t^i) = \lambda$

$$P('x_t^i = d|I) = \alpha P(x_t^i = d|I)P('x_t^i = d|I)$$

Else if $MCC^p('x_t^i) = 1-\lambda$

$$'x_t^i = {}'x_t^i - D_r^p('x_t^i)$$

Go to Step 2. (7)

where $'x_t^i$ is the pixel corresponding to x_t^i, and $P('x_t^i = d)$ is the updated disparity probability. If the corresponding pixel is a stable pixel, the disparity probability is updated by multiplying the probability of the reference pixel by the probability of its corresponding pixel, followed by the termination of the chaining search. If the corresponding pixel is an unstable pixel, the coordinates of the corresponding pixel are updated, and step.2 is repeated. By this iterative process, an unstable pixel must be met a stable pixel. The disparity probabilities of pixels belonging to unstable regions are updated through this process. After the chaining search is complete, the updated pixel disparities are propagated to the unstable region to which the pixel belongs:

$$P(s_{t+1}^j = d|'x_t^{1:N} = d) = \alpha \sum_{'x^i \in s^j} P('x_t^i = d|I)MCC^p('x_t^i) \quad (8)$$

In the feedback process, a disparity for each pixel x^i is estimated as follows:

$$\begin{aligned} P(x_{t+1}^i = d|s_{t+1}^j, I) &= \alpha P(s_{t+1}^j, I|x_{t+1}^i = d)P(x_{t+1}^i = d) \\ &\approx P(s_{t+1}^j|I, x_{t+1}^i = d)P(I|x_{t+1}^i = d) \\ &\approx P(s_{t+1}^j|x_{t+1}^i = d)P(x_{t+1}^i = d|I) \\ &\approx P(x_{t+1}^i = d|s_{t+1}^j)P(s_{t+1}^j)P(x_{t+1}^i = d|I) \\ &\approx P(x_{t+1}^i = d|s_{t+1}^j = d)P(s_{t+1}^j = d)P(x_{t+1}^i = d|I) \\ &\approx P(s_{t+1}^j = d)P(x_t^i = d|I) \end{aligned} \quad (9)$$

The probability of $P(x_{t+1}^i = d)$ in the first row can be omitted safely since we assume that the probability is uniform. I in the first term of the second row is eliminated due to the Markov blanket, and the second term is reformulated by applying Bayes' rule and the assumption of uniformity. The fourth row is rewritten as the fifth row because we estimate likelihood only when the disparities of a pixel and a region are identical. Thus, the probability of $P(x_{t+1}^i = d|s_{t+1}^j = d)$ in the fifth row is 1. $P(s_{t+1}^j = d)$ is calculated by the feed-forward process.

The proposed algorithm can be summarized as follows:

- The disparity of a pixel is estimated with high confidence by a pixel-level MCC.
- In the feed-forward process, the disparities of pixels are propagated to the region layer in order to estimate the disparities of regions.
- The disparities of regions are verified by a region-level MCC.
- Disparities of pixels located in unstable regions are estimated using a chaining search. The resulting updated pixel disparities have a higher confidence level than the initial estimates.
- In the second feed-forward process, only the disparities updated after the chaining search are propagated to unstable regions. Consequently, the disparities of unstable regions are estimated with higher confidence than the initial estimates.
- In the feedback process, pixel disparities are refined using high confidence region disparities.

Fig. 1 shows the improvement in the calculated disparity that occurs with an increasing number of iterations. The images in the first row are disparity maps of left images, while images in the second row are disparity maps of right images. We represent the estimated disparity using saturation and hue. The saturation level represents the magnitude of disparity and the hue represents the direction of disparity. After the first iteration, many pixels do not have precise disparities. With increasing iterations, unreliable disparities are gradually improved. In particular, most pixels located in occluded areas have incorrect hue or saturation values after the first iteration. However, more accurate disparity values are assigned to these pixels by the chaining search.

III. EXPERIMENTS

We tested the proposed algorithm using three types of images. First, the four images of the Middlebury benchmark data sets (Tsukuba, Venus, Teddy, Cones) were used to evaluate performance in standard scenarios. The second experiment was conducted using contaminated images, where there was an intensity difference between corresponding pixels of stereo images, similar to images captured in a real life environment. Finally, our algorithm was applied to practical situations. In all experiments, the parameters of our algorithm were set as follows: SAD size is 5x5, and $\lambda = 0.9$.

A. Middlebury Dataset

Figure. 2 shows the disparity maps of four Middlebury images computed using our algorithm. Quantitative results are shown in Table I, where the percentage of bad pixels in the disparity map compared to ground truth image is calculated. Our method was ranked 62nd out of over 110 methods in the Middlebury evaluation.

B. Gaussian noisy images and realistic images

As mentioned earlier, intensities of corresponding pixels can vary among images captured in real life. In order to quantitatively evaluate the robustness of our algorithm to this issue, Gaussian noise is added to four Middlebury images. We changed the PSNR (Peak Signal-to-Noise Ratio) of the images (Tsukuba, Venus, Teddy, Cones) to five different levels. Our algorithm was compared against two high-performing algorithms [3], [8] using these noisy images. Fig. 3 shows disparity maps with increasing PSNR, where the first row shows the left images of stereo pairs. The first column shows the original image, and the following columns show 38dB, 33dB, 29dB, 27dB, and 25dB images from left to right. The results of CVF (Cost-Volume Filtering [3]), DBP(Double Belief Propagation [8]), and the proposed algorithm (TRC) are shown in the second, third, and fourth

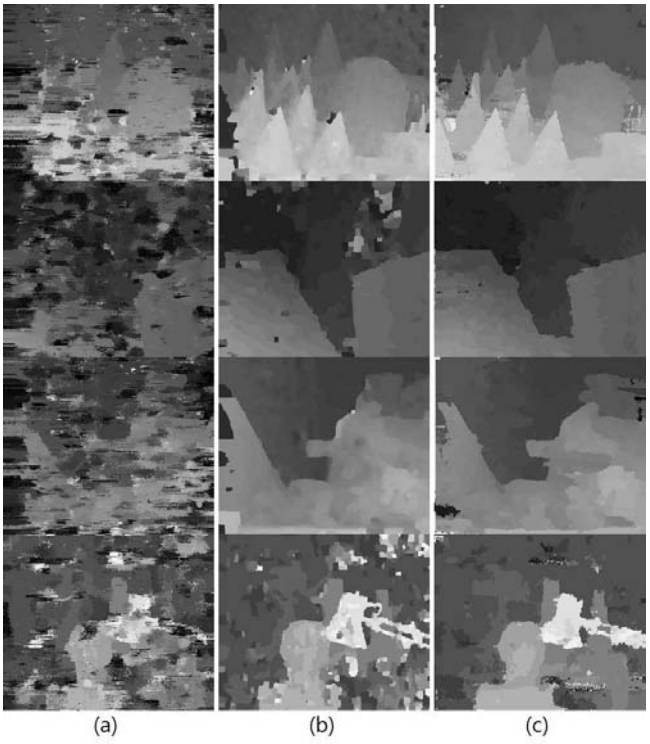

Fig. 4. Disparity maps of 29dB images. The first column shows the results of CVF, the second column shows the results of DBP, and the third column shows the results of the proposed algorithm.

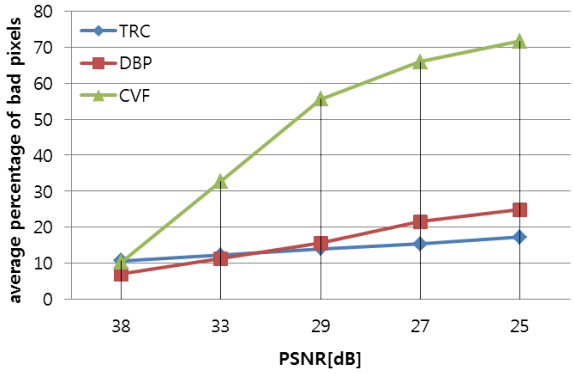

Fig. 5. Average percentage of bad pixels according to PSNR.

Fig. 6. Disparity maps of real-life image.

rows, respectively. Each disparity map is computed using the original or noisy input image from the same column. CVF was observed to be very sensitive to noise because it is a local method. DBP performs better than CVF, but the number of incorrect disparities is significant. Out of the three algorithms tested, our method obtained the most reliable results. This is explained by the fact that our algorithm refines unreliable pixels and regions by an iterative process using a two stage feed-forward, feedback and chaining search approach, as described in Section 2.B.

Fig. 4 shows disparity maps for four stereo pairs where PSNR is 29dB. The disparity maps from the first to fourth rows are calculated using the Cones, Venus, Teddy, and Tsukuba images, in that order. The results of the CVF, DBP, and TRC methods are shown in the first, second, and third columns, respectively.

Fig. 5 shows the numerical results of the comparative experiment for different PSNR values. The ratios of bad pixels for the four images were averaged. We found that all approaches offered similar performance with high PSNR images. As PSNR is reduced, however, the comparative performance of CVF decreases rapidly. DBP performs better than CVF at high PSNR, but its performance declines gradually and becomes lower than TRC beyond 29dB.

In the last experiment, we applied our algorithm to real-life scene. In Fig. 6, the four columns, from left to right, show input image and the disparity maps of CVF, DBP and TRC.

IV. CONCLUSIONS

This study proposes a recurrent two-layer process and chaining search for dense stereo matching. In this approach, the disparity map is calculated through the iterative integration of pixel and region layers to improve the reliability and accuracy of results; disparities in occluded regions are calculated by the chaining search algorithm. Performance in a practical scenario was evaluated quantitatively by comparing the proposed approach with two high-performing algorithms using a contaminated Middlebury benchmark data set. The results show that the proposed method outperforms the other two algorithms.

V. ACKNOWLEDGEMENTS

This research was supported by the Global Frontier R&D Program on ”Human-centered Interaction for Coexistence” funded by the National Research Foundation of Korea grant funded by the Korean Government (MEST) (NRF-M1AXA003- 2011-0028353). All correspondences should be addressed to I. H. Suh.

REFERENCES

[1] A. Malik, T. Choi, and H. Nisar, “Depth map and 3d imaging applications: Algorithms and technologies,” pp. 397–417, 2012.

[2] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[3] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, “Fast cost-volume filtering for visual correspondence and beyond,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3017–3024.

[4] K. Yoon and I. Kweon, “Adaptive support-weight approach for correspondence search,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 650–656, 2006.

[5] 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*, vol. 2. IEEE, 2001, pp. 508–515.

[6] J. Sun, N. Zheng, and H. Shum, “Stereo matching using belief propagation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 787–800, 2003.

[7] C. Zitnick and S. Kang, “Stereo for image-based rendering using image over-segmentation,” *International Journal of Computer Vision*, vol. 75, no. 1, pp. 49–65, 2007.

[8] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, “Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 492–504, 2009.

[9] A. Klaus, M. Sormann, and K. Karner, “Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 15–18.

[10] M. Bleyer and M. Gelautz, “Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions,” *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 127–143, 2007.

[11] P. Bayerl and H. Neumann, “A fast biologically inspired algorithm for recurrent motion estimation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, pp. 246–260, 2007.

[12] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[13] I. Gallo, E. Binaghi, and M. Raspanti, “Neural disparity computation for dense two-frame stereo correspondence,” *Pattern Recognition Letters*, vol. 29, no. 5, pp. 673–687, 2008.

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

[15] Y. Heo, K. Lee, and S. Lee, “Robust stereo matching using adaptive normalized cross correlation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 99, pp. 1–1, 2011.