



# Bayesian Dense Inverse Searching Algorithm for Real-Time Stereo Matching in Minimally Invasive Surgery

Jingwei Song<sup>1,2(✉)</sup>, Qiuchen Zhu<sup>3</sup>, Jianyu Lin<sup>4</sup>, and Maani Ghaffari<sup>1</sup>

<sup>1</sup> Michigan Robotics, University of Michigan, Ann Arbor, MI 48109, USA  
{jingweso,maanigj}@umich.edu

<sup>2</sup> United Imaging Research Institute of Intelligent Imaging, Beijing 100144, China

<sup>3</sup> School of Electrical and Data Engineering, University of Technology,  
Ultimo, NSW 2007, Australia  
Qiuchen.Zhu@uts.edu.au

<sup>4</sup> Hamlyn Centre for Robotic Surgery, Imperial College London,  
London SW7 2AZ, UK

**Abstract.** This paper reports a CPU-level real-time stereo matching method for surgical images (10 Hz on  $640 \times 480$  image with a single core of i5-9400). The proposed method is built on the fast LK algorithm, which estimates the disparity of the stereo images patch-wisely and in a coarse-to-fine manner. We propose a Bayesian framework to evaluate the probability of the optimized patch disparity at different scales. Moreover, we introduce a spatial Gaussian mixed probability distribution to address the pixel-wise probability within the patch. In-vivo and synthetic experiments show that our method can handle ambiguities resulted from the textureless surfaces and the photometric inconsistency caused by the non-Lambertian reflectance. Our Bayesian method correctly balances the probability of the patch for stereo images at different scales. Experiments indicate that the estimated depth has similar accuracy and fewer outliers than the baseline methods in the surgical scenario with real-time performance. The code and data set are available at <https://github.com/JingweiSong/BDIS.git>.

**Keywords:** Stereo matching · Bayesian theory · Posterior probability inference

## 1 Introduction

Real-time 3D intra-operative tissue surface shape recovery from stereo images is important in Computer Assisted Surgery (CAS). The reconstructed depth is a crucial for dense Simultaneous Localization and Mapping (SLAM) [23, 24], AR

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-16449-1\\_32](https://doi.org/10.1007/978-3-031-16449-1_32).

system [11, 28] and diseases diagnosis [13, 18]. All stereo matching procedures follow the pinhole camera model [2] and conduct image rectification, undistortion, and disparity estimation. The stereo matching techniques are normally classified into two categories regarding disparity estimation: prior-free and learning-based. Conventional prior-free methods estimate the pixel-wise disparity using the image alignment techniques [8, 12, 14, 20, 25]. Based on the left-right image consistency assumption (photo-metric or feature-metric), they either use corner feature registration, dense direct pixel searching, or a combination. Differently, Deep Neural Network (DNN) based techniques directly learn the disparity from the training image pairs [3, 16, 26, 29, 30]. Although DNN methods are reported to be efficient, the results may be invalidated with changing parameters such as focal length and baseline or a large texture difference between the training and testing data [1, 19]. Moreover, the DNN-based methods heavily depend on the size and quality of the annotated training data, which are not accessible in many CAS scenarios.

In the category of prior-free methods, ELAS [8] is still one of the most widely used stereo matching algorithms due to its robustness and accuracy [23, 24, 32, 33]. It is also the most popular method in the industry [4, 31]. ELAS uses Sobel descriptors to match sparse corners as the supporting points and triangulate the pixel-wise disparity prior. Then, the optimal dense disparity is retrieved with its proposed maximum a-posteriori algorithm. Its two-step process requires around 0.25–1 s on a single modern CPU core. This paper aims for a faster CPU-based stereo matching method.

The Dense Inverse Searching (DIS) [14] shows the potential of dense direct matching without the time-consuming sparse supporting points alignment. By resizing the left and right images to several coarse scales, it adopts and modifies the Lucas-Kanade (LK) optical flow algorithm [17] for fast estimating the pixel-wise optimal disparity. [14] demonstrates that real-time computation is possible with its patch-based coarse-to-fine dense matching, where patch refers to an arbitrary squared image segment. However, DIS is strictly built based on the photometric consistency and surface texture abundance assumptions, which cannot always be satisfied in CAS. The two main challenges are the textureless/dark surfaces and the serious non-Lambertian reflectance. The weak/dark texture, which widely exists in CAS, leads to ambiguous photometric consistency. Meanwhile, non-Lambertian reflectance brings uneven disturbance on the surfaces, and it cannot be eliminated by just enforcing the patch normalization [21].

In this paper, to deal with photometric inconsistency and non-Lambertian reflectance in stereo matching, we propose a Bayesian Dense Inverse Searching (BDIS) to quantify the posterior probability of each optimized patch. A spatial Gaussian Mixture Model (GMM) is further adapted to quantify pixel-wise confidence within the patch. The final pixel-wise disparity is the fusion of multiple local overlapping patches, reducing the impact of those patches suffering from the textureless/dark surfaces or the non-Lambertian reflectance. In extreme cases,

it is beneficial to give up the disparity estimation of some patches identified as dubious. In particular, this work has the following contributions:

- A Bayesian approach is developed to quantify the posterior probability of the patch.
- A spatial GMM is introduced to quantify the pixels' confidence within the patch.
- To our knowledge, BDIS is the first single core CPU based stereo matching approach that achieves similar performance to the near real-time method ELAS.

## 2 Methodology

### 2.1 Multiscale DIS

Figure 1 shows the DIS (based on fast LK) algorithm for stereo matching proposed by [14]. It is a modified version of the LK algorithm. We use the fast DIS as our base framework. Note that the variational refinement module in [14] is abandoned because it has a small (less than 0.5%) contribution in promoting the accuracy. The modified fast LK based DIS is achieved by minimizing the following objective function:

$$\Delta \mathbf{u} = \operatorname{argmin}_{\Delta \mathbf{u}'} \sum_x [I_r(\mathbf{x} + \mathbf{u}) - I_l(\mathbf{x} + \Delta \mathbf{u}')]^2, \tag{1}$$

where  $\mathbf{x}$  is the processed location,  $\mathbf{u}$  is the estimated disparity in the loop,  $I_l$  and  $I_r$  are the left image patch and right image, and  $\Delta \mathbf{u}$  is the optimal update of  $\mathbf{u}$  at one loop. Different from authentic LK,  $\Delta \mathbf{u}'$  is moved from the right image to the left image patch. The improvement avoids the expensive re-evaluation of the Hessian on the right image. (1) is traversed on all patches at different scales. The disparity at the fine-scale level is initialized at the optimized coarse scale. The optimal disparity at the location  $\mathbf{x}$  is the weighted fusion with all covering patches using inverse residual:

$$\hat{\mathbf{u}}_{\mathbf{x}} = \sum_{k \in \Omega} \frac{1 / \max(\|I_l(\mathbf{x} + \mathbf{u}^{(k)}) - I_r(\mathbf{x})\|^2, 1)}{\sum_{k \in \Omega} 1 / \max(\|I_l(\mathbf{x} + \mathbf{u}^{(k)}) - I_r(\mathbf{x})\|^2, 1)} \mathbf{u}^{(k)}, \tag{2}$$

where  $\Omega$  is the set of patches covering the position  $\mathbf{x}$ ,  $\mathbf{u}^{(k)}$  is the estimated disparity of the patch  $k$  and  $\max(\cdot, \cdot)$  selects the maximum value. The pixel-wise disparity  $\hat{\mathbf{u}}_{\mathbf{x}}$  is the weighted average of the estimated disparities from all patches, wherein the weight is the inverse residual of brightness.

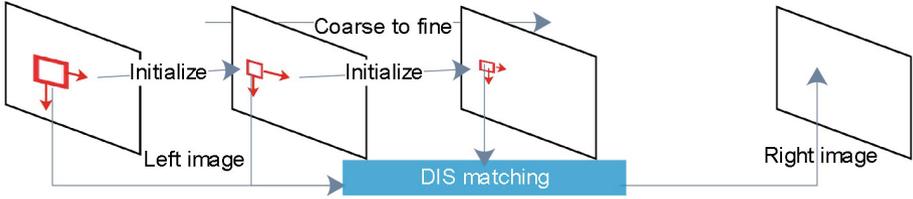


Fig. 1. The framework of the DIS algorithm [14]. It uses 3 scale levels as an example.

### 2.2 The Bayesian Patch-Wise Posterior Probability

The residual-based weighted average fusion (2) suffers from the ambiguities brought by the textureless/dark surface and non-Lambertian reflectance. The textureless/dark surface leads to ambiguous local minima of the cost function penalizing photometric inconsistency (1) and misleads the algorithm to be overconfident on the estimation. Furthermore, the photometric consistency presumption is seriously violated on the surface affected heavily by the non-Lambertian reflectance. The affine lighting changes formulation in previous large-scale SLAM studies [7] cannot fully tackle the complex and severe non-Lambertian reflectance in CAS. In both situations, the weights retrieved from the photometric residuals (2) are misleading. To overcome the difficulty in defining the confidence of the estimated disparity, we propose a Bayesian model to correctly estimate the confidence in the presence of textureless surface and non-Lambertian reflectance. Since the uncertainty distribution of both the left and right scenes is unclear, it is difficult to conduct the direct inference of the posterior probability in terms of disparity. Thus, we implicitly infer the probability with Bayesian modeling using Conditional Random Fields (CRF) [27]. The posterior probability of the patch-wise disparity  $\mathbf{u}^{(k)}$  is

$$p(\mathbf{u}^{(k)}|I_l, I_r) \propto \frac{p(I_r|I_l, \mathbf{u}^{(k)})}{p(I_r, I_l, \mathbf{u}^{(k)})} \propto \frac{p(I_r|I_l, \mathbf{u}^{(k)})}{\sum_{\mathbf{u}_i^{(k)} \in \mathcal{P}} p(I_r|I_l, \mathbf{u}_i^{(k)})} \propto \frac{p(I_r|I_l, \mathbf{u}^{(k)})}{\sum_{\mathbf{u}_i^{(k)} \in \mathcal{P}'} p(I_r|I_l, \mathbf{u}_i^{(k)})} \mathbf{r}, \tag{3}$$

where  $\mathcal{P}$  is the domain of all possible choice of  $\mathbf{u}_i^{(k)}$ . To reduce computational load,  $\mathbf{r}$  is applied as the constant compensation ratio for all patches within the window.  $\mathcal{P}$  is reduced to a small window  $\mathcal{P}'$  assuming the rest candidates are numerically trivial.

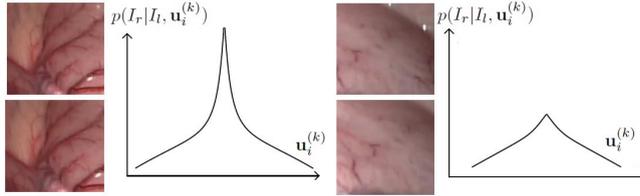
Equation (3) indicates that the posterior probability of the disparity can be obtained by traversing the probability on all possible  $\mathbf{u}_i^{(k)}$ . And the possible choice of disparity is equal to window size  $s$ . Even though the posterior probability suffers from the textureless surface and non-Lambertian reflectance, the illumination consistency probability is proportional to the residuals because the set of neighboring disparities is within one patch, and the impact of the issues is consistent. Thus, we model the illumination consistency probability  $p(I_r|I_l, \mathbf{u}_i^{(k)}, \mathbf{x})$

based on the Boltzmann distribution [15] as

$$p(I_r|I_l, \mathbf{u}_i^{(k)}) = \exp\left(-\frac{\|I_l(\mathbf{u}_i^{(k)}) - I_r(\mathbf{u}_i^{(k)})\|_F^2}{2\sigma_r^2\sigma_s^2}\right), \quad (4)$$

where  $\|\cdot\|_F$  is the Frobenius norm and  $\sigma_r$  is the hyperparameter to describe the variance of the brightness. The relative posterior probability can be obtained with (3) and (4). Generally, the absolute exponential parameter of the Boltzmann distribution denotes the entropy of the state. In our case, such entropy is defined as (4). Image with abundant texture has more entropy loss. Hence, the entropy item is highly related to the photometric inconsistency loss.

Figure 2 shows the relationship between the illumination consistency probability density function and the texture. The response is stronger on the textured surface. The residuals are always small in the textureless surface, no matter how the left and right images are aligned. (2) cannot correctly measure the weights while (4) describes the relative probability of the estimation. Moreover, it tests the local convergence to filter the Saddle point solutions.



**Fig. 2.** The probability density function of the textureless region.

### 2.3 The Prior Spatial Gaussian Probability

In addition to the patch-wise posterior probability of the disparity in the last section, a spatial GMM is adopted to estimate pixel-wise probability within the patch. Considering that medical images are natural images, a multivariate Gaussian distribution is adopted to measure the confidence of the pixel-wise probability using a Gaussian mask. In accordance with the multivariate Gaussian distribution, the center of the patch has higher confidence than the edge pixels since those central pixels preserve more information for inference. Assuming all pixels in the patch are i.i.d, we have

$$p(\mathbf{u}^{(k)}|I_l, I_r, \mathbf{x}) \propto p(\mathbf{u}^{(k)}|I_l, I_r) \exp\left(-\frac{\sum_{\xi^{(k)}(\mathbf{x})} \|\mathbf{x} - \xi^{(k)}(\mathbf{x})\|_F^2}{2\sigma_s^2}\right), \quad (5)$$

where  $\xi^{(k)}(\mathbf{x})$  is the set of all pixel positions within the patch  $k$  in image coordinate.  $\sigma_s$  is the 2D spatial variance of the probability. Note that (5) is independent

of the patch and can therefore be pre-computed before the process. Combining (3), (4) and (5), the final pixel-wise posterior probability distribution can be represented as follows,

$$p(\mathbf{u}^{(k)}|I_l, I_r, \mathbf{x}) \propto \exp\left(-\frac{\sum_{\xi^{(k)}(\mathbf{x})} \|\mathbf{x} - \xi^{(k)}(\mathbf{x})\|_{\mathbb{F}}^2}{2\sigma_s^2}\right) \frac{\exp\left(-\frac{\|I_l(\mathbf{u}^{(k)}) - I_r(\mathbf{u}^{(k)})\|_{\mathbb{F}}^2}{2\sigma_r^2\sigma_s^2}\right)}{\sum_{\mathbf{u}_i^{(k)} \in \mathcal{P}} \exp\left(-\frac{\|I_l(\mathbf{u}_i^{(k)}) - I_r(\mathbf{u}_i^{(k)})\|_{\mathbb{F}}^2}{2\sigma_r^2\sigma_s^2}\right)}.$$

Finally, it should be emphasized that (4) and (5) are not the cost functions but probability/weight for each patch or pixel. Costly optimization steps are avoided.

### 3 Results and Discussion

BDIS was compared with DIS [14], SGBM [12] and ELAS [8] on the in-vivo and the synthetic data sets<sup>1</sup>. The computations were implemented on a commercial desktop (i5-9400) in C++. DNN-based methods PSMNet [5] and GwcNet [10] were also compared for completeness and the computation was conducted on the GTX 1080ti in PyTorch. The public in-vivo stereo videos from [9] were adopted which contains 200 images with size  $640 \times 480$  and 200 images with size  $288 \times 360$ . All stereo images were rectified, undistorted, calibrated, and vertically aligned with the provided intrinsic and extrinsic parameters. We also provided a synthetic data set generated from an off-the-shelf virtual phantom of a male’s digestive system. A virtual handheld colonoscope was placed inside the colon and was manipulated to go through the colon to collect the depth and stereo images. The 3D game engine Unity3D<sup>2</sup> was used to generate the sequential stereo and depth images with a pin-hole camera in size  $640 \times 480$ . The synthetic distortion-free data has accurate intrinsic and extrinsic parameters. Both diffuse lighting (100 frames) and non-Lambertian reflectance (100 frames) were simulated.  $\gamma$  was set to 0.75 for  $640 \times 480$  and 0.25 for  $288 \times 360$  data to discard the patch without enough valid pixels.  $\sigma_r$  and  $\sigma_s$  were set to 4; the sampling within one Bayesian window was 5; the disturbance from the convergence was 0.5 and 1 pixel.

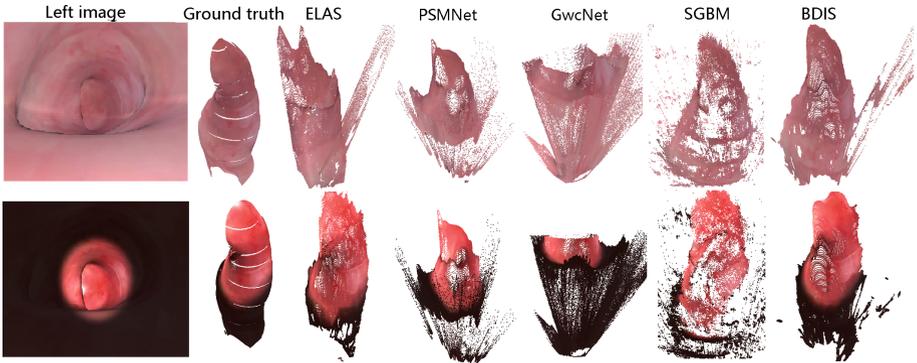
#### 3.1 Quantitative Comparisons on the Synthetic Data Set

BDIS was compared quantitatively with the baseline methods ELAS, SGBM, DIS, PSMNet, and GwcNet. The comparison between the prior-based DNN-based method and BDIS is for completeness only. The default setting of PSMNet and GwcNet were strictly followed. The pre-trained networks were adopted and finetuned with the labeled 300 (training) and 50 (validation) synthetic images for training and validation. Both were trained with Adam optimizer in 300 epochs.

Table 1 and Fig. 3 show the comparisons on the synthetic data set, which are unaffected by distortion and inaccurate camera parameters. Considering

<sup>1</sup> Readers are encouraged to watch the attached video and test the code.

<sup>2</sup> <https://unity.com/>.



**Fig. 3.** Sample reconstructions in Diffuse lighting and Lambertian reflectance scenarios.

**Table 1.** The results on the synthetic data with diffused light and Lambertian reflectance.

		ELAS	SGBM	DIS	BDIS	GwcNet	PSMnet
Diffuse light	Median error	0.178	0.512	0.251	<b>0.161</b>	0.542	0.417
	Mean error	<b>0.220</b>	1.113	0.753	0.320	0.809	0.641
	Valid pixels (1000)	166.77	103.92	288.41	208.44	100.00	301.42
Non-Lambertian reflectance	Median error	0.198	0.710	0.376	<b>0.163</b>	0.271	0.731
	Mean error	<b>0.235</b>	1.400	1.051	0.379	0.662	1.027
	Valid pixels (1000)	81.50	74.29	295.92	204.42	106.38	301.46

the median error, BDIS is the best and has 9.55% and 17.68% higher accuracy than ELAS in diffuse lighting and non-Lambertian reflectance. The results indicate that BDIS is more advantageous in the scenario of non-Lambertian reflectance over ELAS, thus more robust in surgical scenarios. Results also show that BDIS cannot handle the edges well. Figure 3 and Table 1 reveal the bad mean error comparison is attributed to the small group of far-out points on the dark regions/edges. The number of valid prediction suggest BDIS produces more predictions but suffers from inaccurate dark region predictions.

Readers may notice the bad performance of DNN, which contradicts the conclusion from [1]. The reason is that the finetuning training process does not yield satisfying model parameters. The synthetic data set for transfer learning and the data used to pre-train the DNN are significantly different in terms of textures. Studies [6, 22] indicate that the performance of the convolutional DNN is heavily dependent on the image texture, and efforts were devoted to bridging the domain gap [6, 34]. The bad training process indicates its strong dependency on the training data set, which can be avoided using prior-free methods. Further tests will be conducted on labeled in-vivo data set.

### 3.2 Qualitative Comparisons on the In-vivo Dataset

We compared ELAS, BDIS, DIS, and SGBM on the in-vivo data sets. Since no ground truth is provided, DNN-based methods cannot be implemented. We aim to show that BDIS achieves similar accuracy as ELAS since near real-time ELAS is widely used in the community. Based on the scope-to-surface distance, the samples were categorized into five groups. Results show that BDIS achieves an average 0.4–1.66 mm (median error) and 0.65–2.32 mm (mean error) deviation from ELAS’s results.

The invalid/dark/bright pixels lead to photometric inconsistency in the stereo matching process. Figure 4 shows the qualitative comparisons of ELAS, DIS, SGBM, and BDIS on the relatively well-textured images. Generally, BDIS achieves similar performance as ELAS but better matches pixels at the image edge with fewer outliers. DIS and SGBM suffer from the wrong edges. Invalid pixels inevitably exist on the edges of the rectified image after the image undistortion. Thus, in the coarse-level patch disparity estimation, patches with more invalid pixels are more likely to fail in convergence or yield local minima (abnormal depth) due to insufficient information. The dubious predictions, however, substantially influence the prediction and the initialization of the disparity at the finer-scale patch optimization (as in (2)). BDIS solves the problem by quantifying the posterior probability, discarding the patch that does not converge, and lowering the patches’ probabilities with invalid pixels. Although the discarded patch does not help yield disparity, other patches compensate for the loss. If one pixel is not covered by any patch, we follow ELAS not to optimize the pixel.

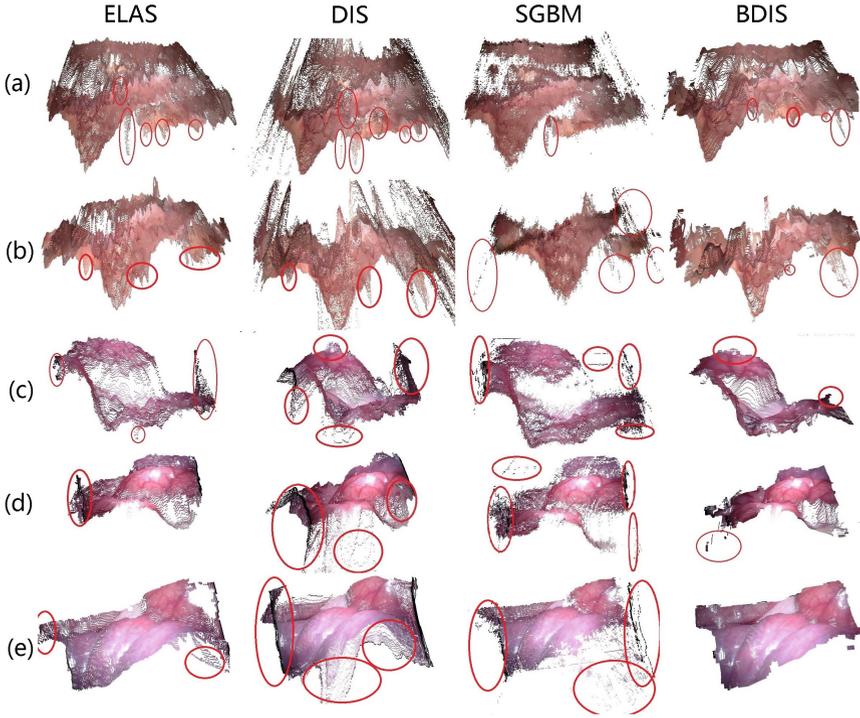
Another noticeable problem is the ambiguous local minima in the cost function, which penalizes the photometric inconsistency. Figure 4 shows BDIS has fewer local minima than DIS and SGBM and is similar to ELAS. Figure 4 (a–b) indicates that the BDIS addresses the patches’ probabilities with textured and alleviates the ambiguous disparity from the textureless surface. Figure 4 (c–e) show that the ambiguities caused by the illumination have been greatly reduced. The quantitative results also provide evidence on its side. It should be emphasized that this work does not enforce any prior smoothness constraint in the optimization process.

We additionally tested BDIS and ELAS on the surfaces with serious non-Lambertian reflectance (Fig. 5). The photometric consistency of this data deteriorates significantly. Figure 5 shows that the center of the soft tissue is exposed to intense lighting while the marginal region is dark. Figure 5 indicates that ELAS suffers from the ambiguity on the marginal dark regions while BDIS can ignore or estimate most dark pixels correctly.

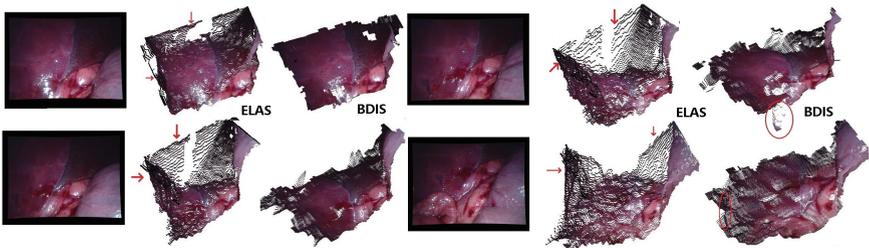
### 3.3 Processing Rate Comparison

We compared the time consumption of ELAS, DIS, and BDIS on a single core of CPU (i5-9400). BDIS runs 10 Hz on  $640 \times 480$  image and 25 Hz on  $360 \times 288$  image while ELAS achieves 4 Hz and 11 Hz. The two DNN methods GwcNet and PSMnet run 3 Hz and 5 Hz on GTX 1080ti. BDIS consumes double the

time of DIS. The majority of the extra time of BDIS is devoted to patch-wise window traversing. Since the sampling window size is 5 in the experiment, 5 more times residual estimations are needed. In general, BDIS achieves similar/better performance over ELAS but runs 2 times faster.



**Fig. 4.** Sample recovered shapes of the 5 classes. Circles mark the regions with large error.



**Fig. 5.** The qualitative comparisons on the heavy Lambertian reflectance and dark case.

## 4 Conclusion

We propose BDIS, the first CPU-level real-time stereo matching approach for CAS. BDIS inherits the fast performance of DIS while being more robust to textureless/dark surface and severe non-Lambertian reflectance. It achieves similar or better performance in accuracy as the near real-time method ELAS. A Bayesian approach and a spatial GMM are developed to describe the relative confidence of the pixel-wise disparity to achieve the performance. Experiments indicate that BDIS has fewer outliers than DIS and achieves a lower amount of outlier predictions than the near real-time ELAS.

## References

1. Allan, M., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint [arXiv:2101.01133](https://arxiv.org/abs/2101.01133) (2021)
2. Andrew, A.M.: Multiple view geometry in computer vision. Kybernetes (2001)
3. Brandao, P., Psychogyios, D., Mazomenos, E., Stoyanov, D., Janatka, M.: HAPNet: hierarchically aggregated pyramid network for real-time stereo matching. *Comput. Methods Biomech. Biomed. Eng. Imaging Visual.* 1–6 (2020)
4. Cartucho, J., Tukra, S., Li, Y.S., Elson, D., Giannarou, S.: VisionBlender: a tool to efficiently generate computer vision datasets for robotic surgery. *Comput. Methods Biomech. Biomed. Eng. Imaging Visual.* 1–8 (2020)
5. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418 (2018)
6. Chen, X., Wang, Y., Chen, X., Zeng, W.: S2R-DepthNet: learning a generalizable depth-specific structural representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3034–3043 (2021)
7. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(3), 611–625 (2017)
8. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6492, pp. 25–38. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-19315-6\\_3](https://doi.org/10.1007/978-3-642-19315-6_3)
9. Giannarou, S., Visentini-Scarzanella, M., Yang, G.Z.: Probabilistic tracking of affine-invariant anisotropic regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 130–143 (2013)
10. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3273–3282 (2019)
11. Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.O., Cotin, S.: Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 199–208. IEEE (2013)
12. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 807–814. IEEE (2005)
13. Jia, X., et al.: Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. *IEEE Trans. Autom. Sci. Eng.* **17**(3), 1570–1584 (2020)

14. Kroeger, T., Timofte, R., Dai, D., Van Gool, L.: Fast optical flow using dense inverse search. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 471–488. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46493-0\\_29](https://doi.org/10.1007/978-3-319-46493-0_29)
15. Larochelle, H., Bengio, Y.: Classification using discriminative restricted boltzmann machines, pp. 536–543 (2008)
16. Long, Y., et al.: E-DSSR: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. arXiv preprint [arXiv:2107.00229](https://arxiv.org/abs/2107.00229) (2021)
17. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia (1981)
18. Mahmood, F., Yang, Z., Chen, R., Borders, D., Xu, W., Durr, N.J.: Polyp segmentation and classification using predicted depth from monocular endoscopy. In: Medical Imaging 2019: Computer-Aided Diagnosis, vol. 10950, p. 1095011. International Society for Optics and Photonics (2019)
19. Pratt, P., Bergeles, C., Darzi, A., Yang, G.-Z.: Practical intraoperative stereo camera calibration. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 667–675. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10470-6\\_83](https://doi.org/10.1007/978-3-319-10470-6_83)
20. Rappel, J.K.: Surgical stereo vision systems and methods for microsurgery. US Patent 9,330,477, 3 May 2016
21. Shimasaki, Y., Iwahori, Y., Neog, D.R., Woodham, R.J., Bhuyan, M.: Generating Lambertian image with uniform reflectance for endoscope image. In: IWAIT 2013, pp. 1–6 (2013)
22. Song, J., Patel, M., Girgensohn, A., Kim, C.: Combining deep learning with geometric features for image-based localization in the gastrointestinal tract. *Expert Syst. Appl.* 115631 (2021)
23. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robot. Autom. Lett.* **3**(1), 155–162 (2017)
24. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: MIS-SLAM: real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing. *IEEE Robot. Autom. Lett.* **3**(4), 4068–4075 (2018)
25. Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.-Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: Jiang, T., Navab, N., Plum, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 275–282. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-15705-9\\_34](https://doi.org/10.1007/978-3-642-15705-9_34)
26. Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., Sitti, M.: Deep endovo: a recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. *Neurocomputing* **275**, 1861–1870 (2018). <https://doi.org/10.1016/j.neucom.2017.10.014>, <http://www.sciencedirect.com/science/article/pii/S092523121731665X>
27. Uzunbas, M.G., Chen, C., Metaxas, D.: An efficient conditional random field approach for automatic and interactive neuron segmentation. *Med. Image Anal.* **27**, 31–44 (2016)
28. Widya, A.R., Monno, Y., Imahori, K., Okutomi, M., Suzuki, S., Gotoda, T., Miki, K.: 3D reconstruction of whole stomach from endoscope video using structure-from-motion. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3900–3904. IEEE (2019)

29. Yang, G., Manela, J., Happold, M., Ramanan, D.: Hierarchical deep stereo matching on high-resolution images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5515–5524 (2019)
30. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint [arXiv:1705.08260](https://arxiv.org/abs/1705.08260) (2017)
31. Zampokas, G., Tsiolis, K., Peleka, G., Mariolis, I., Malasiotis, S., Tzovaras, D.: Real-time 3D reconstruction in minimally invasive surgery with quasi-dense matching. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE (2018)
32. Zhan, J., Cartucho, J., Giannarou, S.: Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 11147–11154. IEEE (2020)
33. Zhang, L., Ye, M., Giataganas, P., Hughes, M., Yang, G.Z.: Autonomous scanning for endoscopic mosaicing and 3D fusion. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3587–3593. IEEE (2017)
34. Zheng, C., Cham, T.J., Cai, J.: T2net: synthetic-to-realistic translation for solving single-image depth estimation tasks. In: Proceedings of the European Conference on Computer Vision, pp. 767–783 (2018)