Depth Estimation by Combining Binocular Stereo and Monocular Structured-Light

Yuhua Xu[1,2,*], Xiaoli Yang[2], Yushan Yu[2], Wei Jia[1], Zhaobi Chu[1], Yulan Guo[3]
[1]Hefei University of Technology, [2]Orbbec, [3]Sun Yat-sen University
xyh_nudt@163.com

Abstract

It is well known that the passive stereo system cannot adapt well to weak texture objects, e.g., white walls. However, these weak texture targets are very common in indoor environments. In this paper, we present a novel stereo system, which consists of two cameras (an RGB camera and an IR camera) and an IR speckle projector. The RGB camera is used both for depth estimation and texture acquisition. The IR camera and the speckle projector can form a monocular structured-light (MSL) subsystem, while the two cameras can form a binocular stereo subsystem. The depth map generated by the MSL subsystem can provide external guidance for the stereo matching networks, which can improve the matching accuracy significantly. In order to verify the effectiveness of the proposed system, we build a prototype and collect a test dataset in indoor scenes. The evaluation results show that the Bad 2.0 error of the proposed system is 28.2% of the passive stereo system when the network RAFT is used. The dataset and trained models are available at https://github.com/YuhuaXu/MonoStereoFusion.

1. Introduction

Depth estimation is a fundamental problem in computer vision, which has numerous applications in the fields of 3D modeling, robotics, UAVs, augmented realities (AR), and autonomous driving [1, 10, 30]. Depth estimation methods can be divided into active structured-light, binocular stereo vision, time-of-flight (TOF), and monocular depth estimation.

Since Microsoft Kinect [48] was released in 2010, consumer-grade depth sensors have been widely used. Kinect is based on the monocular structured-light method, which was also used in iPhone X released in 2017. However, it may fail to obtain depth measurements for distant objects, or outdoor scenes under strong light. The binocular stereo vision system has a larger measurement range than the structured-light system, and it can also work in outdoor environment with strong sunlight, but it is easily affected by the surface texture of the objects. In recent years, stereo matching methods based on deep learning have achieved remarkable progress. However, these methods may still fail on scenes with weak texture (e.g., white walls). And this kind of weak texture objects are very common in indoor environment. The binocular active structured-light system (e.g., Intel D435 [14]) relies on two IR cameras and an IR projector for depth estimation, which has good adaptability in both indoor and outdoor situations. To acquire texture, a third camera (i.e. RGB camera) is required. Since there is a baseline between the RGB camera and IR camera, a coordinate system conversion step is required to make the depth image aligned with the RGB image. Due to the noise of the depth map and the error of the calibration parameters, it is difficult to accurately align the RGB image and depth map. In terms of hardware, three cameras and one projector are required, which is not compact. TOF has poor adaptability to objects with low reflectivity and distant objects. In addition, TOF suffers from multipath interference [29]. The monocular depth estimation methods cannot obtain the depth maps with a certain scale [11].

In this work, we seek a compact depth sensing solution that can integrate the advantages of the monocular structured-light and binocular stereo vision.

The main contributions of this work are:

(1) We propose a novel stereo vision system, which consists of an RGB camera, an IR camera and an IR speckle projector. Especially, the IR camera is not attached with a filter. Thus the IR camera can receive IR light (invisible to human eyes) and ambient light (visible to human eyes) simultaneously. The IR camera and IR projector can form a monocular active structured-light system as Kinect, while the IR camera and the RGB camera can form a binocular stereo system. These two types of stereo systems have complementary advantages. The active structured-light system is robust to weak texture objects (e.g., white walls) which are hard to handle for the passive binocular stereo system. We can obtain a robust stereo system by fusing the initial

*Corresponding author

depth map obtained by the active structured-light system in the cost volume of stereo matching network.

(2) We build a prototype system and collect a new stereo dataset for integrating the monocular structured light and binocular stereo vision (MonoBinoStereo) to verify the effectiveness of the proposed method. The dataset will be open for further research.

(3) We find that DNN can accurately estimate the disparity map of a pair of asymmetric stereo images, where one is passive and the other is active (with speckles). To the best of our knowledge, this is the first time that DNN is used to process this kind of stereo images with asymmetric texture.

The features of the proposed stereo system are as follows:

(1) Compared with the classical binocular stereo vision, it is robust to weak texture objects and rich texture objects simultaneously in indoor environments.

(2) Compared with the existing monocular structured-light system (e.g., Kinect), it has a larger measuring distance range and better performance in outdoor environment.

(3) Compared with the existing active depth sensing system (e.g., Kinect and Intel D435), its output depth maps have better completeness. In addition, the depth map is naturally aligned with the RGB image pixel-by-pixel.

(4) For the interference of strong sunlight, it will degenerate into an ordinary passive stereo system in outdoor environments.

2. Related Work

Zbontar *et al*. [44] first use convolutional neural network (CNN) to compare two image patches (e.g., 9×9 or 11×11) and calculate their matching costs. The following steps, such as cost aggregation, disparity computation, and disparity refinement, are still traditional methods [23]. This method (i.e. MC-CNN) significantly improves the accuracy, but still struggles to produce accurate disparity results in textureless, reflective and occluded regions and is time-consuming. DispNetC [22] is the first end-to-end stereo matching network, which is more efficient, almost 1000 times faster than MC-CNN-Acrt [44]. In DispNetC, there is an explicit correlation layer. In traditional stereo matching methods, there is usually a disparity refinement module. Inspired by this, the residual refinement layers are exploited [19, 20, 24] to further improve the prediction accuracy. Besides, the segmentation information [42] and edge information [32] are incorporated into the stereo matching networks to improve the performance. Wang *et al*. [37, 38] propose a generic parallax-attention mechanism to capture stereo correspondence regardless of disparity variations. Optical flow and rectified stereo are closely related problems. RAFT [35] uses a gated recurrent unit (GRU) based operator to iteratively update the flow field using features retrieved from the correlation volume. RAFT shows good generalization performance.

GC-Net [16] first uses 3D convolutions for cost aggregation in a 4D cost volume, and utilizes the soft $argmin$ to regress the disparity. Duggal *et al*. [8] adopt the idea of PatchMatch Stereo [2], and build a thin cost volume to speed up the prediction process. The similar idea is also used in [12]. Variance-based uncertainty estimation is used to adaptively adjust disparity search space of the thin cost volume [4, 31]. Recent work [3, 8] shows that the 3D convolution can improve matching accuracy on specific datasets. However, 3D convolution is more time-consuming than 2D convolution, which makes it difficult to apply in real-time applications. In order to pursue real-time performance, StereoNet [17] performs 3D convolution at a low resolution (e.g., 1/8 resolution), and then refines the disparities hierarchically. The resulting network can run in real-time at 60 fps. However, this simplification decreases the network's accuracy.

Xu *et al*. [40] design a bilateral grid based edge-preserving cost volume upsampling module. With the upsampling module, a high quality cost volume of high resolution can be obtained from the low resolution version efficiently. The upsampling module can be embedded into many existing stereo matching networks, such as GCNet [16], PSMNet [3] and GANet [45]. The resulting networks can be accelerated by several times while maintaining comparable accuracy. HITNet [34] does not explicitly build a volume and instead relies on a fast multiresolution initialization step, differentiable 2D geometric propagation and warping mechanisms to infer disparity hypotheses. To achieve high accuracy, this method infers slanted plane hypotheses allowing to accurately perform geometric warping and upsampling operations. In order to reduce the computation burden, Yao *et al*. [43] propose a decomposition model which performs dense matching at a very low resolution (e.g., 20 × 36) and uses sparse matching at different higher resolutions to recover the disparity of lost details scale-by-scale.

ActiveStereoNet [47] is the first deep learning solution for active stereo systems. Due to the lack of ground truth, the network is designed to be fully self-supervised. Instead of formulating the depth estimation via a correspondence search problem, Riegler *et al*. [28] show that a simple convolutional architecture is sufficient for high-quality disparity estimates in a monocular structured-light system.

Our work is also related to image guided depth completion, whose task is to estimate the dense depth map from sparse depth measurement. Ma *et al*. [21] proposed to feed the concatenation of the sparse depth and the color image into an encoder-decoder deep network. Jaritz *et al*. [15] combined semantic segmentation to improve the depth completion. Cheng *et al*. [5] proposed a convolu-

tional spatial propagation network (CSPN) to post process the depth completion results with neighboring depth values. However, CSPN relies on fixed-local neighbors, which could be from irrelevant objects. Park *et al*. [25] proposed a non-local spatial propagation network for depth completion. This method can effectively avoids irrelevant local neighbors and concentrates on relevant non-local neighbors during propagation. Qiu *et al*. [27] learned surface normals as the intermediate representation. Xu *et al*. [41] modeled the geometric constraints between depth and surface normal in a diffusion module and predicted the confidence of sparse LiDAR measurements to mitigate the impact of noise. For addressing the problem of depth smearing, Imran *et al*. [13] proposed a multi-hypothesis depth representation that explicitly models both foreground and background depths in the difficult occlusion-boundary regions.

Compared with depth completion methods, our method can utilize the stereo pair and the depth guidance from the monocular structured light subsystem for disparity estimation. When the depth guidance is not available, the stereo pair can still be used to estimate the depth of the targets. The stereo images can form stronger constraints than single images.

3. System

3.1. Hardware

In this paper, we design a novel stereo camera. As illustrated in Figure 1d, the proposed stereo camera consists of an RGB camera, an IR camera and an IR projector. Its layout is similar to the monocular structured-light system (Figure 1b), e.g., Kinect. However, it is significantly different from Kinect. In Kinect, the IR camera and IR projector are used for depth estimation. To obtain the depth map aligned with RGB image, a depth-to-color step is required to convert the depth map from the IR camera coordinate system to the RGB camera coordinate system.

The proposed stereo system consists of two subsystems. First, the IR camera and the IR projector form an active monocular structured-light subsystem. Second, the IR camera and the RGB camera form a binocular stereo subsystem. The monocular structured-light subsystem is robust to weak texture objects, while the binocular subsystem has the ability to reconstruct distant objects and can work in outdoor environment. Thus the two subsystems have complementary advantages.

In the next subsections, we will show how the two subsystems are integrated.

3.2. Depth Estimation Pipeline

As mentioned before, the proposed depth camera consists of two subsystems. The input includes an RGB image, an IR image and a reference speckle image. The

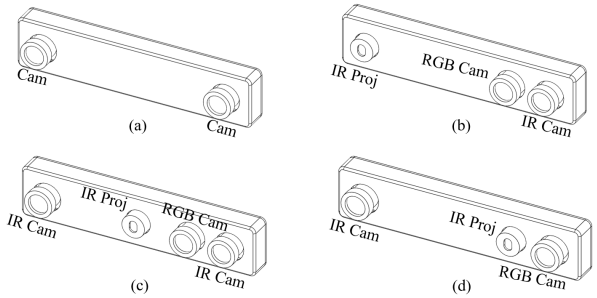

Figure 1. Layout of various depth cameras. (a) Binocular stereo camera (e.g., ZED [33]). (b) Monocular structured-light depth camera (e.g., Kinect [48]). (c) Active binocular depth camera (e.g., Intel D435 [14]). (d) Design of the proposed depth camera. Compared with (b), there should be enough baseline between the IR camera and RGB camera in the proposed depth camera since the two cameras are used for the binocular stereo subsystem. In addition, the IR camera and IR projector form a monocular structured-light (MSL) subsystem. The depth map from the MSL subsystem can provide external guidance in stereo matching networks.

reference image is pre-stored and fixed in the monocular structured-light subsystem, as shown in Figure 2. First, the current IR image of the targets and the reference speckle image are matched, and then a disparity map d_m is obtained. With the calibration parameters of the monocular structured-light subsystem, a depth map Z_m can be obtained and re-projected to the RGB camera coordinate system. We use Z'_m to denote the depth map aligned with the RGB image and d'_m to denote the corresponding disparity map. Then, the RGB image, IR image and disparity map d'_m are fed into the stereo matching network to estimate the final disparity map. The pipeline is illustrated in Figure 2.

3.3. Monocular Structured-Light

Different from the binocular stereo system, a camera is replaced by a projector in the monocular structured-light system, as shown in Figure 3. The depth estimation process is similar to Kinect [9, 36]. The current speckle image of the targets is matched to the reference image, which is a speckle image captured when the camera's optical axis is perpendicular to a planar target at a known distance Z_{ref}. In order to eliminate the influence of different brightness of the two images to be matched, we follow the method in [36] to convert these images to binary images. Then, an efficient block matching algorithm is used to calculate the corresponding relationships between the two images to obtain the disparity map d_m. The matching window size is set to 21×21. With the disparity map, we can obtain the depth map Z_m via

$$Z_m = \frac{Z_{ref}}{1 - \frac{Z_{ref} d_m}{B_m f_m}} \tag{1}$$

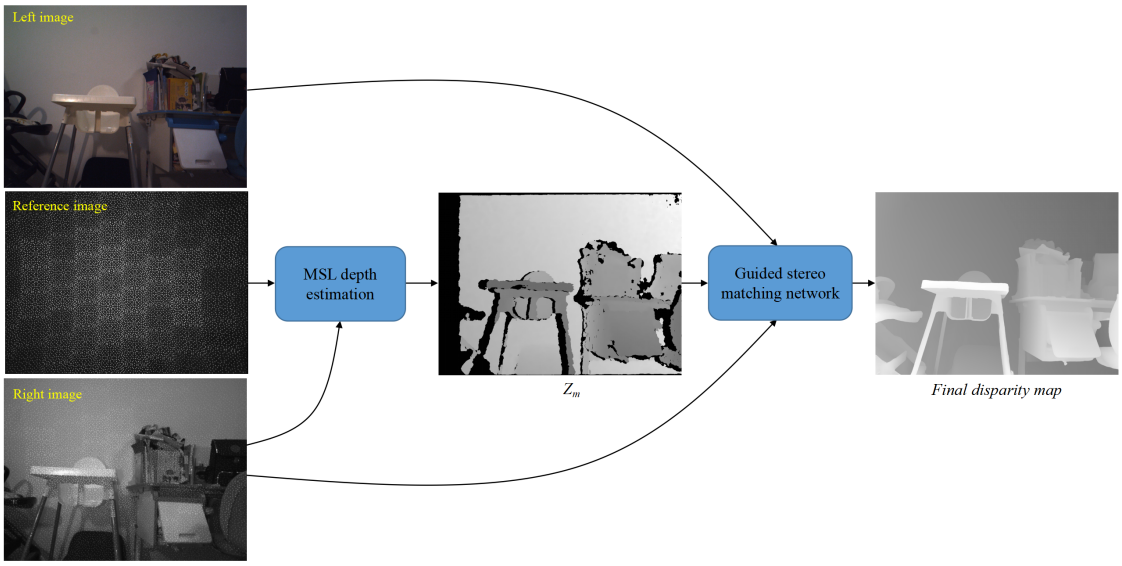

Figure 2. Pipeline of the proposed depth estimation method. First, the initial depth map is obtained with the monocular structured-light (MSL) subsystem by matching the IR image and the pre-stored reference image. Then, the IR and RGB image pairs are fed to the stereo matching network to extract features and build a cost volume. The information of the active monocular subsystem is integrated in the cost volume as done in GSM [26] to obtain high quality disparity map.

where B_m is the baseline and f_m is the focal length of the monocular structured-light system. With the calibration parameters of the cameras, we can convert the depth map Z_m onto the image plane of the RGB camera and obtain the depth map Z'_m aligned with the RGB image. Next, we can obtain the corresponding disparity map in the binocular stereo system via

$$d'_m = Bf/Z'_m \tag{2}$$

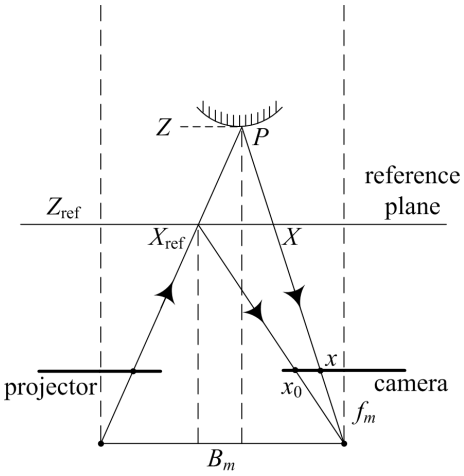

Figure 3. Principle of the monocular structured-light system. The. The change in depth will bring about the movement of the speckle spots in the horizontal direction.

where B is the baseline and f is the focal length of the binocular system.

3.4. Stereo Matching Network and Fusion Strategy

Note that, the IR camera here does not have a narrow-band filter as Kinect. So the IR camera can receive the active speckled light and ambient light. Thus the images of the two cameras are very different in appearance in indoor environments, as shown in Figure 2. It seems that it is difficult to match this kind of images. Fortunately, we find that accurate matching results can be obtained with the deep neural networks (DNN).

In order to verify the adaptability of the DNN to this kind of binocular images with asymmetric textures, we first modify the training dataset and testing dataset of Flyingthings3D. In the modified dataset, the left image remains unchanged, while tens of thousands of random speckle dots are added in the right image, as shown in Figure 4. So the stereo images in the modified dataset have asymmetric textures. The brightness of the speckles is decreased according to the distance from these points to the camera to mimicking energy attenuation of light energy. Then, we use both the original and modified training datasets to train two existing stereo matching networks, including PSMNet [3], and RAFT [35]. RAFT shows good generalization in optical flow estimation task, which requires to estimate displacement both in X and Y directions. Here, we make a small modification to estimate only the displacement in the X direction.

Table 1 indicates that these networks have good adaptability to this kind of stereo images with asymmetric textures (more details are in subsection 4.3). Figure 4 shows the qualitative results.

Although there are usually many invalid values in the depth map from the active structured-light system (Figure 2), the depth values are relatively reliable. So the valid depth values can be used as the guidance for the stereo matching network. The cost volume in stereo matching network consists of features with geometric and contextual information that allows the subsequent convolution to regress the disparity probability [3, 16, 19]. To integrate the advantage of the monocular structured-light system, we modify the cost volume according to the disparity map d'_m as done in guided stereo matching (GSM) [26], which peaks the correlation scores or the features activation related to the hypotheses suggested by the sparse hints and dampens the remaining ones.

Specifically, let g be a matrix of size $w \times h$, conveying the externally provided disparity values, and v a binary mask, specifying which elements of g are valid (i.e., if v_{xy} = 1). The cost volume is denoted as $\mathcal{C} \in \mathbb{R}^{w \times h \times D_{max} \times F}$, where D_{max} is the max disparity and F is the feature number. Given the pixel coordinate (x, y) and disparity value $g(x, y)$ from external cue g, GSM applies Gaussian function

$$f_{GSM}(x, y, d) = \lambda \cdot e^{-\frac{(d-g(x,y))^2}{2\sigma^2}} \tag{3}$$

on the features $\mathcal{C}(x, y, d)$ of the cost volume, and obtain a new cost volume $\mathcal{C}'$,

$$\mathcal{C}'(x, y, d) = (1 - v_{xy} + v_{xy} \cdot f_{GSM}(x, y, d)) \cdot \mathcal{C}(x, y, d) \tag{4}$$

where σ determines the width of the Gaussian, while λ represents its maximum magnitude and should be greater than or equal to 1.

For RAFT, the correlation values in the cost volume are normalized to $[0, 1]$ to avoid peak negative correlations via

$$\mathcal{C}(x, y, d) = \frac{< F_l(x, y), F_r(x - d, y) >}{2(||F_l(x, y)|| + \epsilon)(||F_r(x - d, y)|| + \epsilon)} + 0.5 \tag{5}$$

where, F_l and F_r are features extracted from the left and right images, d denotes the disparity, and ϵ is a small constant.

In this work, the disparity map d'_m is taken as the external guidance for the stereo matching networks.

4. Experiments

4.1. Prototype

To verify the effectiveness of the proposed system, we build a prototype system as shown in Figure 5. The system includes two synchronized CMOS cameras and an IR speckle projector. Both cameras have a focal length of 4.0 mm and a resolution of 1280×960. The maximum frame rate is 30 frames per second (fps). The baseline of the stereo subsystem is 94.14 mm and that of monocular structured-light system is 63.0 mm. The diffractive optical element (DOE) based projector can project about 11,000 speckle dots onto the scenes. This kind of projector is very cheap (less than $3). We capture a speckle image of a white wall as the reference image at a distance of 80 cm when the optical axis of the camera is perpendicular to the white wall. The RGB camera has an IR-cut filter, and the IR camera has no filter.

Method	EPE (Original)	EPE (Modified)
PSMNet-O [3]	0.895	3.922
PSMNet-M	1.212	0.955
PSMNet-OM	0.925	0.984
PSMNet-OM-G	0.666	0.686
RAFT-O [35]	0.985	1.910
RAFT-M	1.070	1.092
RAFT-OM	1.026	1.109
RAFT-OM-G	0.751	0.771

Table 1. Evaluation of networks on the original SceneFlow dataset and the modified SceneFlow dataset. We use suffixes O, M and OM to denote the models trained with the original Flyingthings3D dataset, the modified Flyingthings3D dataset and the mixture of the two datasets, respectively. The suffix G denotes the guidance is used in the network.

4.2. Dataset and Evaluation Metrics

Synthetic dataset. The synthetic SceneFlow [22] stereo dataset includes Flyingthings3D, Driving, and Monkaa. The dataset consists of 35,454 training images and 4,370 testing images of size 960×540 with accurate ground-truth disparity maps. We will use Flyingthings3D for study of the stereo matching networks. The End-Point-Error (EPE) will be used as the evaluation metric.

Real-scene dataset. To evaluate the performance of the proposed system, we collect a dataset (i.e., MonoBinoStereo) in indoor environment, which covers different indoor scenes, including offices, living rooms, and bedrooms. The stereo pairs are easy to acquire. However, it is not a easy task to acquire the corresponding ground-truth disparity maps for the stereo pairs. Here, we choose to use the space-time stereo method [7, 46] to obtain the ground truth disparities as done in [6]. 200 pairs of stereo images are captured for each scene. During the process of image capturing, thousands of moving speckles are projected. Therefore, the speckle distribution in each frame is different. The ground truth disparity maps are estimated by integrating all

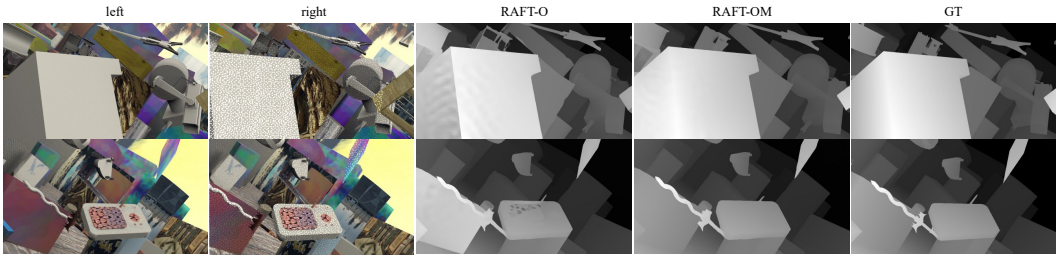

Figure 4. Evaluation on SceneFlow.

Method	projector on				projector off			
	EPE	Bad0.5 (%)	Bad1.0 (%)	Bad2.0 (%)	EPE	Bad0.5 (%)	Bad1.0 (%)	Bad2.0 (%)
PSMNet-O [3]	9.007	70.51	55.66	41.79	2.112	52.54	33.85	20.42
PSMNet-OM	2.687	57.35	39.28	24.81	**1.871**	**51.70**	33.16	19.89
PSMNet-OM-G	**0.814**	**45.63**	**15.73**	**3.81**	2.018	52.02	**32.67**	**18.33**
RAFT-O [35]	2.498	57.83	37.70	21.88	1.183	46.43	26.07	12.71
RAFT-OM	1.370	49.23	29.31	14.60	1.239	**44.21**	**23.18**	11.72
RAFT-OM-G	**0.811**	**45.13**	**16.08**	**3.59**	**1.103**	44.75	23.71	**10.51**
MSG [18]	3.092	58.85	30.32	14.25	-	-	-	-

Table 2. Quantitative evaluation on the real scene dataset. The suffix G denotes the guidance is used during training of the network models. Note that, when the projector is on, depth from MSL is used as the guidance in the models with suffix G. When the DOE projector is off (i.e., both the left and right images are passive), the guidance is not available and not used in network prediction.

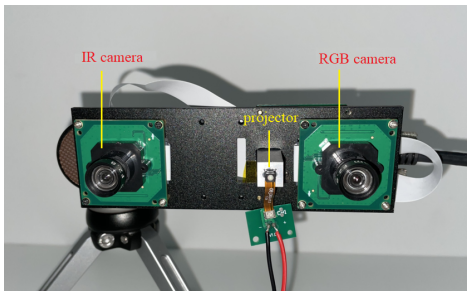

Figure 5. Prototype of the proposed depth camera.

the 200 pairs of images. A sub-pixel refinement and a left-right check (LRC) are also applied. The MonoBinoStereo dataset includes 15 scenes in total. The samples are shown in Figure 6. For each scene, we collect two stereo pairs, where the left images are always passive, while one image of the right camera is passive (with projector off) and the other is active (with projector on).

However, we lack a large training dataset in real indoor scenes. The synthetic IRS dataset [39] is considerably close to the real scenes. It contains more than 100,000 pairs of 960×540 resolution stereo images (84,946 for training and 15,079 for testing) in indoor scenes. We use the IRS dataset as the training dataset for evaluation on the MonoBinoStereo dataset. Details of the network training are presented in the supplementary material.

4.3. Quantitative Evaluation

We first evaluate the proposed method on the SceneFlow dataset. We trained PSMNet [3] and RAFT [35] with the original Flyingthings3D dataset and the modified Flyingthings3D dataset respectively. We use suffixes O, M and OM (e.g, PSMNet-O) to denote the models trained with the original Flyingthings3D dataset, the modified Flyingthings3D dataset and the mixture of the two datasets, respectively. The End-Point-Error (EPE) results are reported in Table 1. When the models are trained with the original dataset, the EPEs on the modified test dataset are large. For example, the EPE of the PSMNet-O on the modified test dataset is 3.922. When the modified training dataset is used, the EPE of the resulting model (PSMNet-M) is reduced to 0.955. However, the EPE for the original test dataset increases from 0.895 to 1.212. When both training datasets are used, the resulting model (PSMNet-OM) can balance the two test datasets. Furthermore, if the external guidance is available, we can use the strategy in GSM [26] to further improve the results. The resulting methods are denoted with a suffix G, e.g., PSMNet-OM-G. When 5% pixels of the ground truth depth map are used as the external guidance, the EPE is reduced from 0.984 to 0.686 on the modified test dataset. The results are similar for RAFT. The qualitative results are shown in Figure 4.

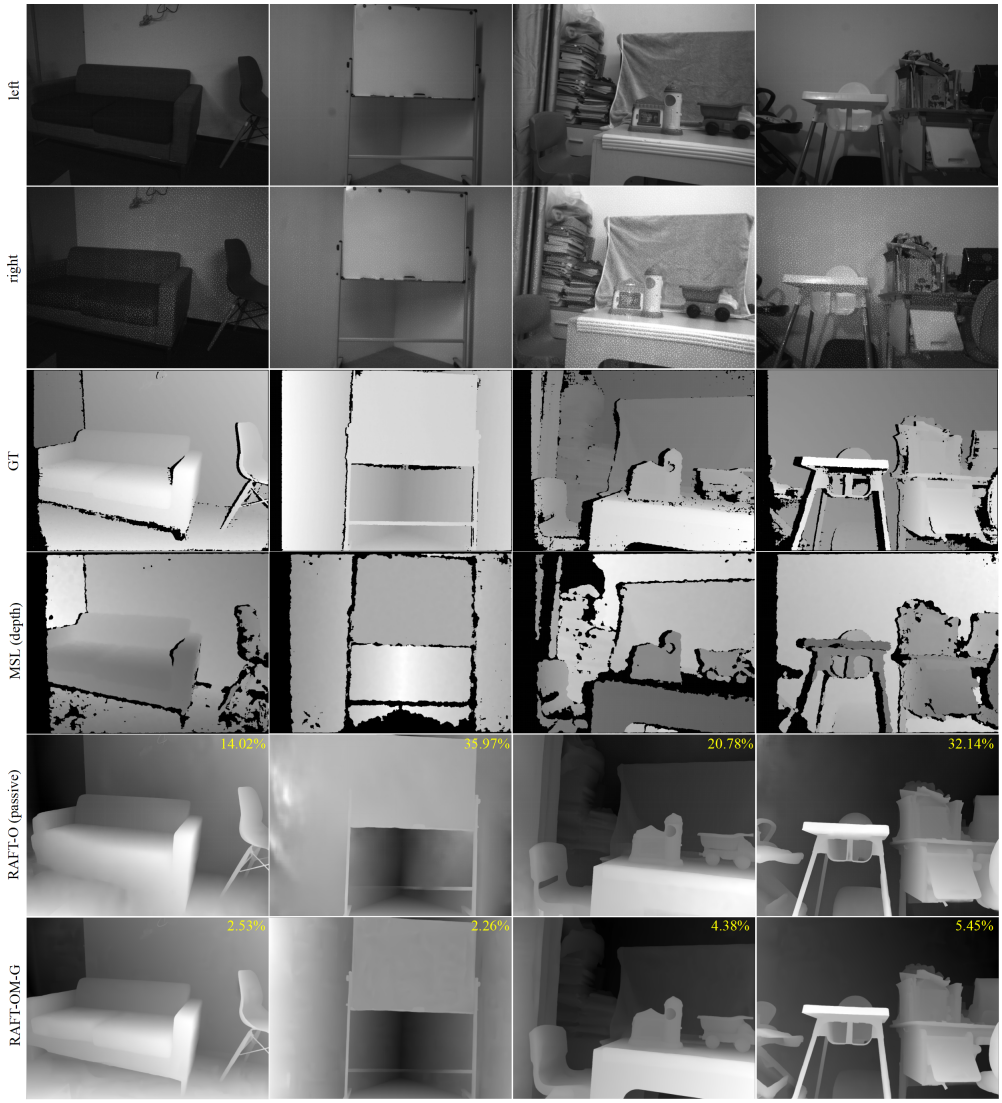

Figure 6. Comparisons on the real dataset. The first row shows the left images (The RGB images are converted to grayscale images before network prediction). The second row shows the right images with speckles (the passive right images are not shown), the third row is the ground truth disparity maps generated with the space-time stereo method [7, 46], the fourth row shows the depth images generated with the MSL subsystem, the fifth row shows the disparity maps of RAFT-O for the passive stereo images, and the last row shows the disparity maps of RAFT-OM-G, where the left image is passive and the right image is with speckles. In row 5 and row 6, Bad2.0 error is shown for each disparity map. The corresponding error maps are shown in the supplementary material.

To further verify the effectiveness of the proposed method, we evaluate the models on the collected real-scene dataset, MonoBinoStereo. The models are trained by mixing the Flyingthings3D and IRS datasets. The quantitative results are shown in Table 2. Take RAFT for example. The Bad 2.0 error of RAFT-O is up to 21.88% on the real test dataset with the DOE projector on, where only the original datasets (Flyingthings3D and IRS) are used for training. When the modified datasets are added, the Bad 2.0 error of the resulting model (RAFT-OM) is reduced to 14.60%. In our system, the depth map from the monocular structured-light subsystem can be used as the external guidance for the stereo matching networks. We use 10% of the pixels in d'_m as the guidance[1]. When this guidance is utilized, the Bad 2.0 error is reduced to 3.59% (RAFT-OM-G). In Table 2, the quantitative results of the different models on the pure passive stereo dataset (see Subsection 4.2) are also shown. Note that the guidance information of the passive mode is

[1]Since the cost volume is built at 1/8 resolution for RAFT, only 1/640 of the pixels in d'_m are used for guidance actually.

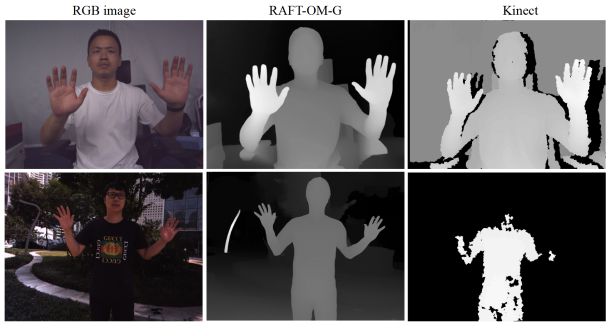

Figure 7. Qualitative comparison. The first column shows the RGB images, the second column shows the disparity maps of RAFT-OM-G, the third column shows disparity maps of Kinect. The first row is the results in indoor scenes, and the second row shows the results in outdoor scenes. It is difficult for Kinect to output stable depth map out of doors. To keep anonymous, the faces are masked.

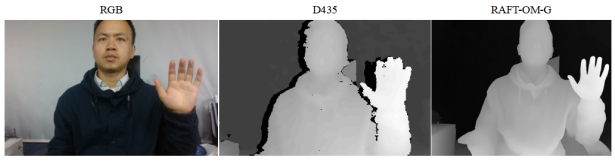

Figure 8. Qualitative comparison with Intel RealSense D435 [14]. D435 uses two cameras to obtain depth map and a third camera for texture acquiring, where occlusion is inevitable. In contrast, Our system can output depth maps naturally aligned with RGB images with only two cameras. To keep anonymous, the face is masked.

not available. We run the model RAFT-O on the passive test dataset. The Bad 2.0 error is 12.71%, which is 3.5 times of RAFT-OM-G. It indicates that the proposed method can improve stereo matching accuracy significantly. The Bad 2.0 error on the passive dataset of RAFT-OM-G is 10.51% (no guidance used), which indicates that RAFT-OM-G can be well generalized to passive scenes. The qualitative results are shown in Figure 6. Table 2 also shows that the overall performance of RAFT is better than PSMNet on MonoBinoStereo.

In addition, we compare with a depth completion method, MSG [18], on the MonoBinoStereo dataset, where 1% of the pixels in d'_m are used as the guidance. The results are shown in Table 2. The Bad 2.0 error of MSG is 18.57%, which is much larger than RAFT-OM-G.

4.4. Qualitative Evaluation

We also test the proposed system in dynamic scenes with people and outdoor scenes, where it is difficult to obtain the ground truth disparity maps. For these scenes, we present the qualitative comparison results.

In Figure 7, we compare the proposed system with Kinect V1 in indoor and outdoor scenes. Kinect can output dense depth estimation in indoor scenes. However, in outdoor scenes, there are more holes in the depth maps because the IR speckles projected are interfered by the sun light. However, for the proposed system, it will degenerate into a passive binocular stereo system, where the stereo pairs can still be used to estimate the dense depth maps of the scenes. We also compare our system with Intel RealSense D435 [14], the results are shown in Figure 8.

4.5. Limitation

In the monocular structured light system, a reference image of a planar target with known depth Z_{ref} is required. When capturing the reference image, we assume that the optical axis of the camera is perpendicular to the planar target, which is hard to guarantee in practice. Compared with the binocular stereo system, the monocular structured light system is more difficult to calibrate. The calibration error will lead to alignment error of the RGB image and the depth image Z'_m, which may cause wrong guidance in guided stereo matching network. In experiment, we find that increasing the number of guide points does not improve the accuracy (see supplementary material for details). Furthermore, if the same number of guidance points are sampled from the ground truth, the Bad0.5, Bad1.0 and Bad 2.0 errors are reduced to 12.94, 4.94, and 2.00 for RAFT-OM-G, respectively. So in the future, we will focus on the accurate calibration method of the monocular structured light system to further improve performance.

5. Conclusion

In this paper, we present a novel stereo system. This system includes a monocular structured-light subsystem and a binocular stereo subsystem. These two subsystems are combined to obtain robust depth estimation. Our system is unique in that it has only two cameras, an RGB camera and an IR camera. The RGB camera is used both for depth estimation and texture acquisition. The depth maps obtained are naturally aligned with RGB images pixel-by-pixel. We collect a real test dataset in indoor scenes. The quantitative results show that the Bad 2.0 error of the proposed system is 28.2% of the classical passive stereo system. Under strong outdoor light, the proposed system will degenerate to a passive stereo system. We hope the proposed system can provide a new solution for designing more robust depth cameras for the community.

Acknowledgements. This work was supported by Orbbec Inc. (No. W2020JSKF0547), and partly supported by the National Natural Science Foundation of China (No. U20A20185, 61972435, 62076086), Major Science and Technology Projects in Anhui Province (202103a05020001), and Key Research and Development Program in Anhui Province (202004d07020008).

References

[1] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11):1–11, 2020. 1

[2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, volume 11, pages 1–11, 2011. 2

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2, 4, 5, 6

[4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[5] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018. 2

[6] Carlo Dal Mutto, Pietro Zanuttigh, and Guido Maria Cortelazzo. Probabilistic tof and stereo data fusion based on mixed pixels measurement models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2260–2272, 2015. 5

[7] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–359. IEEE, 2003. 5, 7

[8] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4384–4393, 2019. 2

[9] Barak Freedman. Depth mapping using projected patterns. *US Application Publication, US 2010/0118123 A1*, 2010. 3

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 1

[12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2

[13] Saif Imran, Xiaoming Liu, and Daniel Morris. Depth completion with twin surface extrapolation at occlusion boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2583–2592, 2021. 3

[14] Intel. Intel realsense depthcamera d435. https://www.intelrealsense.com/. 1, 3, 8

[15] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 52–60. IEEE, 2018. 2

[16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 2, 5

[17] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for edge-aware depth prediction. In *European Conference on Computer Vision*, 2018. 2

[18] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2020. 6, 8

[19] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. 2, 5

[20] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[21] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803. IEEE, 2018. 2

[22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 2, 5

[23] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang. On building an accurate stereo matching system on graphics hardware. In *IEEE International Conference on Computer Vision Workshops*, pages 467–474. IEEE, 2011. 2

[24] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. 2

[25] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Computer Vision–ECCV 2020: 16th Euro-*

pean Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pages 120–136. Springer, 2020. 3

[26] Matteo Poggi, Davide Pallotti, Fabio Tosi, and Stefano Mattoccia. Guided stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2019. 4, 5, 6

[27] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. 3

[28] Gernot Riegler, Yiyi Liao, Simon Donne, Vladlen Koltun, and Andreas Geiger. Connecting the dots: Learning representations for active monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7624–7633, 2019. 2

[29] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer vision and image understanding*, 139:1–20, 2015. 1

[30] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 1

[31] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 2

[32] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, pages 1–21, 2020. 2

[33] Stereolabs. Zed. https://www.stereolabs.com/zed/. 3

[34] Vladimir Tankovich, Christian Häne, Sean Fanello, Yinda Zhang, Shahram Izadi, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. *arXiv preprint arXiv:2007.12140*, 2020. 2

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 2, 4, 5, 6

[36] Guijin Wang, Xuanwu Yin, Xiaokang Pei, and Chenbo Shi. Depth estimation for speckle projection system using progressive reliable points growing matching. *Applied optics*, 52(3):516–524, 2013. 3

[37] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[38] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. 2

[39] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019. 6

[40] Bin Xu, Yuhua Xu, Xiaoli Yang, Wei Jia, and Yulan Guo. Bilateral grid learning for stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12497–12506, 2021. 2

[41] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2811–2820, 2019. 3

[42] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision*, pages 636–651, 2018. 2

[43] Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, and Yuwei Wu. A decomposition model for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6091–6100, 2021. 2

[44] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1592–1599, 2015. 2

[45] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2

[46] Li Zhang, Brian Curless, and Steven M Seitz. Spacetime stereo: Shape recovery for dynamic scenes. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–367. IEEE, 2003. 5, 7

[47] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–801, 2018. 2

[48] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1, 3