International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.12 (2015), pp.237-246
http://dx.doi.org/10.14257/ijsip.2015.8.12.23

Kinect Sensor based Object Feature Estimation in Depth Images

Kajal Sharma*

Piorville Apartment, 24 Namyang-Dong, Seongsangu, Changwon, Korea
kajal175@gmail.com

Abstract

Kinect is a motion-sensing device which was originally developed for the Xbox 360 gaming console. This recently developed low-cost sensor detects the body position, motion, and voice; it consists of a microphone, a RGB camera, and a depth sensor. Kinect is PC-centric sensor which allows developers to develop real-life applications with human gestures and body motions. This paper presents an approach to interpret the indoor room objects in order to match the objects features in depth images captured from an RGBD video database. The dataset consists of color and depth image pairs gathered in real-time indoor home environment. The objects features are matched in depth image pairs with the feature association method to detect stable features at different time instances.

Keywords: *kinect sensor, image acquisition, video processing, object tracking*

1. Introduction

Recently, there has been great interest in 3-D imaging applications with advancement of low-cost Kinect sensor. Kinect has brought about a revolution with hands-free gaming and allow developers to build diverse applications in the field of robotics, imaging, education, security, and so on. Kinect offers a particularly attractive set of capabilities, and can simultaneously captures 3D depth images and 2D color images. Figure 1 shows two scenes typical of the captured dataset: first row shows the images extracted from color video captured with RGB camera. Second row shows the depth images captured with Kinect depth camera.

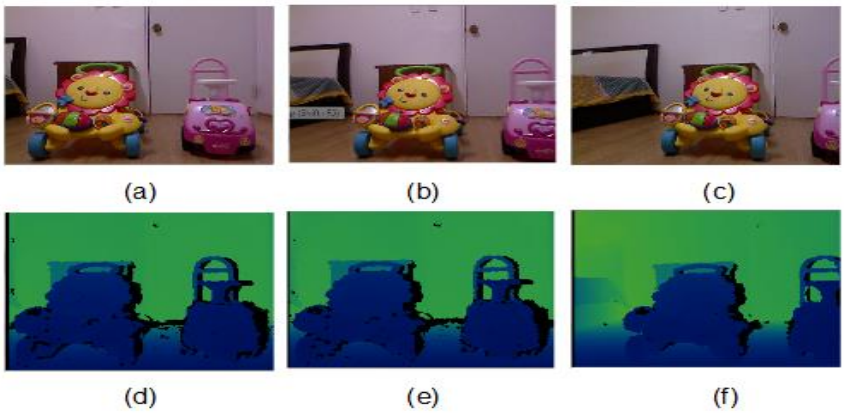

Figure 1. Two Scenes Typical of the Captured Dataset with Kinect (a) Color Frame #18 (b) Color Frame #81 (c) Color Frame #321 with One Occluded Object (d) Depth Frame #18 (e) Depth Frame #81 (f) Depth Frame #321 with One Occluded Object

* Corresponding Author

ISSN: 2005-4254 IJSIP

RGB-D depth sensing device mainly consists of RGB information along with depth information. Originally, Kinect is developed to recognize human gesture but its low cost and depth features made it available to develop 3D object reconstruction and other depth data based applications. Microsoft (MS) released the Kinect software developer's kit (SDK) to allow developer to develop various real time applications by using depth data, voice information and gesture data. The SDK supports color stream, depth stream, and skeleton stream to recognize voice and to track object and human motion. In addition, 3D objects can be tracked with the Kinect SDK which has the special depth sensing feature to recognize depth of object and can estimate the distance of depth pixels. Kinect has outstanding capability to quickly access RGB-D data in real-time, thus can be very advantageous for 3D object reconstruction and tracking. A wide range of 3D reconstruction using depth image and tracking algorithms has been proposed recently but less research is done on tracking with depth sensing capability of Kinect.

This paper presented a novel approach to detect the object features in depth images with the improved feature matching method. The color and depth video datasets are captured with the color and depth cameras of Kinect. The different depth image pairs of the depth video are passed to the keypoint calculation step to obtain the keypoints with invariant viewpoints which is further passed to the neural network. The matching in different depth images are performed with feature reduction using neural network.

This paper is organized as follows: Section 2 introduces some of the research related to Kinect and Section 3 given an overview of Kinect components. Section 4 describes the proposed method for matching the features in depth images. Section 5 presents the implementation of the proposed technique and comparison with recent research. Finally, conclusions are drawn in Section 6.

2. Review of Related Work

Numerous previous efforts have been detailed in the literature in collecting the object information and 3D reconstruction of indoor environments based on vision. The 3D model of an object can be reconstructed using depth image registration. The two main types of 3D reconstruction with the depth images consist of patch-based and voxel-based 3D reconstruction of an object. The patch based reconstruction method is based on the distance metric using depth data while the image based approaches uses object visual features. The iterative closest point (ICP) algorithm detailed in [1-2] is the most common technique to register the depth images. Other variants of the ICP techniques are proposed in the literature with the objective of speeding the convergence rate. In [2], the authors proposed an approach in which distance metric is based on correspondences between surface points and nearby tangent planes on the other surface. The surface normal measurements are available which minimizes the point-to-tangent collection directly and does not require point-to-point matches [3-1].

The point correspondences require high computation to obtain the closest point in iterative closest point (ICP) algorithm. Thus, Blais [4] proposed a speed up projective algorithm for the data association. In the field of augmented reality and robotics, a wide range of algorithms has been proposed on simultaneous localization and mapping (SLAM). A patch based 3D reconstruction algorithm has been proposed in [5] for the RGB-D SLAM to map the large indoor environment. This algorithm reconstructs the 3D model using visual and shape information gathered with RGB-D camera. The scale invariant feature transform (SIFT) features are used as the initial point pairs for ICP algorithm. The aim of RGB-D is to build 3D models of objects with shape and appearance information; in addition also perform the task of alignment and registration.

In [6], the researchers used the Kinect depth information in indoor environments to detect the people by using depth data. This method generates the output region with the detected people and it uses 2-D chamfer distance matching to scan the overall image data.

The resulted output region is then passed to the region growing algorithm to obtain the full contour of the detected human body. Rougier *et. al.*, [7] and Zhang *et. al.*, [8] proposed the algorithm for person segmentation and localization with the depth images. These methods use the background subtraction algorithm where depth background image is obtained from a number of training of background images.

A large-scale RGB-D object datasets and their annotation software have been made publicly available to the research community by the authors of [9-10]. In [9], the dataset consists of multiple views of a set of objects and the objects are organized into a hierarchical category structure. The dataset in [10] consists of registered RGBD images, detailed object labels, and annotated physical relations between objects in the scene which can be used to design indoor scene analysis applications. The authors proposed an integrated approach in [11] with the RGB and depth information of an object at the category and instance levels.

In [5], a sparse feature matching approach is presented for both appearance and shape matching via an ICP algorithm. A graph pose optimization is incorporated by using RGB feature correspondences. Other algorithms have been presented in [12-13] to improve the performance of this algorithm. The scale invariant feature transform (SIFT) [14] step for feature extraction and description in [5] is implemented with FAST feature descriptor [15] and SURF descriptor [16] to reduce the computation complexity of the algorithms.

3. Components of Kinect Sensor

The Kinect sensor bar contains depth sensor, color camera, a special infrared light source, and four microphones [14-15]. The major components of the Kinect sensor are shown in Figure 2. A tilt motor working as the base enables the device to be tilted in upward and downward direction. The list of Kinect sensor components are given below:

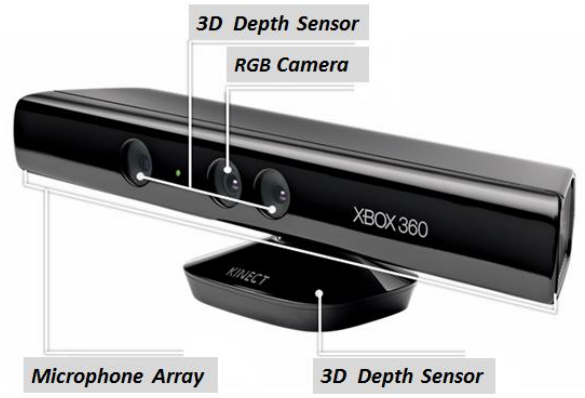

Figure 2. Components of Kinect sensor

3.1. Color Camera

The color camera has the ability to capture and stream the color video data. The Kinect camera can capture color stream at frame rate of 30 frames per second (FPS) and can detect the red, blue, and green colors. The video stream consists of various image frames and has a resolution of 640 x 480 pixels. The field of view (FOV) for the color camera ranges from 43 degrees vertical by 57 degrees horizontal.

3.2. Infrared (IR) Emitter and IR Depth Sensor

Kinect has the ability to provide the 3D information of a scene or an object. The depth map of the environment in front of the camera can be obtained directly with the Kinect device. The processing of the depth signals is done entirely inside the sensor device and the generated depth map is later transmitted similar to the color image. The only difference is that the pixel of each depth image contains the distance information; the sensor transmits the distance values for each depth pixel. Figure 3 shows the depth sensing process to obtain the distance information of any scene or an object.

The Kinect device contains two depth sensors: IR emitter and IR depth sensor. The IR emitter depth sensor is mounted as a camera on Kinect but in actual it is an IR projector which emits the infrared light on the objects in a "random dot pattern". The infrared light is projected on the objects in the dot pattern which is captured by IR depth sensor. IR depth sensor capture depth information from the dotted light reflected off different objects. This invisible dot information is used to calculate the distance between the sensor and the object from where the IR dot was read and is transformed into depth data.

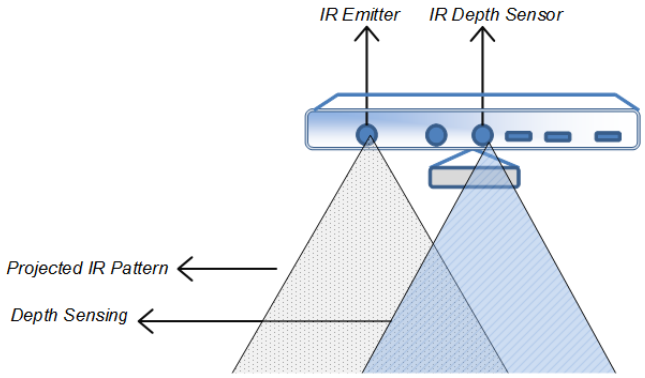

Figure 3. Kinect Depth Sensing Process to Obtain the Distance Information with Infrared (IR) Emitter and IR Depth Sensor

3.2.1. Depth Data Processing: The depth stream contains a number of depth frames where the pixels in each frame contain the distance information in millimeters. The three resolutions supported by the depth stream are 640 x 480 pixels, 320 x 240 pixels, and 80 x 60 pixels, and the depth data contains the distance information to the nearest object from the camera plane at a particular (x, y) coordinate. The depth sensor's field of view range remains the same as the field of view of color camera. The Kinect sensor uses an IR emitter and an IR depth sensor that is a monochrome CMOS (Complimentary Metal-Oxide- Semiconductor) sensor to capture the 3D information of an object. The steps of depth data processing are detailed in Figure 4.

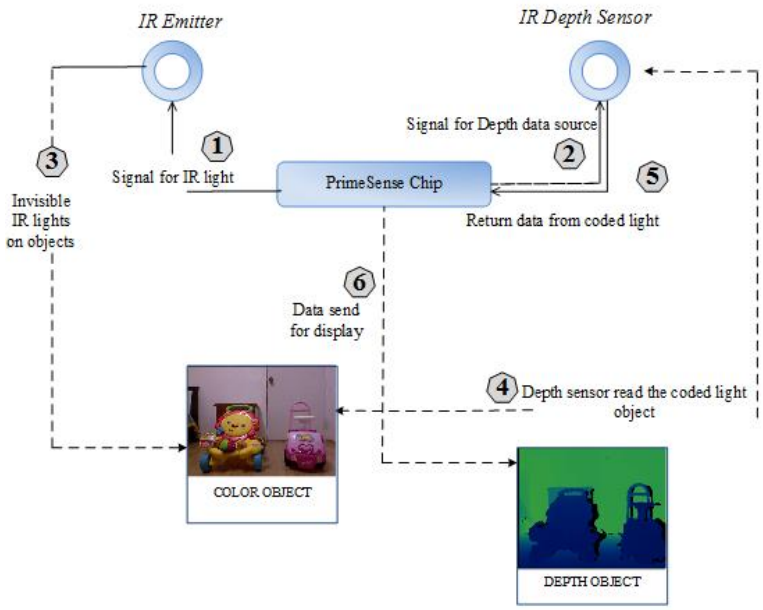

Figure 4. Steps of the Kinect Depth Data Processing with IR Emitter and IR Depth Sensor

The flow diagram steps are explained as follows: (1) The PrimeSense chip sends a signal to the IR emitter to turn on the infrared light to capture the depth data. (2) In addition, the chip also sends a signal to the IR depth sensor to initialize the depth sensor. (3) The IR emitter starts emitting an electromagnetic radiation to the objects in front of the camera. The sensor's IR lights are invisible because the wavelengths of the radiations are longer than the wavelength of the visible light (4) The IR depth sensor capture depth information and obtain the distance between the sensor and the object from where the IR dot was read. (5) The depth sensor returns the coded depth light to the PrimeSense chip. (6) The PrimeSense chip process the depth stream and form a frame by frame depth stream to create the output display data and form a depth image ready for the display.

3.3. Tilt Motor

The tilt motor connects the base and body of the sensor with a small motor which has a vertical field of view that ranges from -27° to +27°. The Kinect sensor can be shifted upwards or downwards by 27 degrees, thus increasing the range of view to capture the color and depth data. The motor can be controlled to adjust the elevation angle of the sensor in order to get the best view of the scene or an object.

3.4. Microphone Array and LED

The Kinect uses the four microphones in the sensor bar which are arranged in a linear fashion to locate sound. It has the ability to detect the audio sound and can displays the angle from the sensor to any sound source. The Kinect bidirectional microphone has the advantage of capturing and recognizing the audio beam effectively with enhanced noise suppression, echo cancellation, and beam-forming technology. An LED in the Kinect device is used to indicate the status that the Kinect device drivers have loaded properly. It shows green color when the Kinect is connected to the computer and tells that device is ready for use to create applications. It is placed between the projector and the camera.

4. Depth Image Matching with the Proposed Method

This paper presents a new depth image feature matching method in which the stable matched features at different time instances are optimized with the neural network. The kinect captured video is passed to the matching process to obtain the stable depth feature which is helpful in finding path for autonomous vision applications. The depth video consists of multiple depth images. This proposed method presents unsupervised feature selection and category classification of depth features of the toy objects appear in front of the camera with change in time. The proposed method extracts the feature points of the toy vehicles in the different depth images and calculates the descriptors using scale invariant feature transform (SIFT). The SIFT is a high dimensional feature vector; thus to detect the features in multiple images, it requires very high computations. This paper introduces a method to reduce the depth features with neural network and obtains stable features at different time in the depth images.

The initial step in this proposed algorithm is to pass each depth image with the difference-of-Gaussian function to obtain the interest points that are invariant to scale and orientation. For each depth image, the SIFT replaces the images by a set of scale and orientation-invariant feature descriptors using gradient orientation histograms. The SIFT method divides the depth image into 4 × 4 sub-regions, and sums the gradient strength in each sub-region. SIFT uses eight directions in each sub-region to generate an eight-dimensional vector. The local image gradients are transformed into 128 dimensional representations resulting into a keypoint descriptor. A SIFT descriptor is a 3-D spatial histogram of the image gradients and each pixel gradient is formed by the pixel location and the gradient orientation. These descriptor vectors are passed to proposed neuro image matching step to calculate the similar keypoints in different depth images. The keypoint vector in the depth images passed to the feature reduction step. The optimization and reduction of depth features is done with the winning pixels estimation in the depth images with the unsupervised self-organizing map (SOM). For each depth image pair, the SIFT descriptors and histograms of selected SIFT descriptors are reduced and matched using SOMs. Thus, the proposed method enables an unsupervised feature matching where there is no requirement for parameter setting for the number of category classification. The features in depth image pairs are clustered with the SOM where the distance between the input depth feature vector x(t) and the weight vector w(t) is computed by:

$$d_k(t) = \left\| x(t) - w_k(t) \right\| \text{ where } k = 1 \ldots n \quad (1)$$

where d denotes the distance vector usually the Euclidean distance. The best matched neuron (bmn) is obtained by the calculation of minimum distance with the input feature vector and is calculated as in (2):

$$d_{bmn}(t) = \min_k (d_k(t)) \quad (2)$$

The output reduced depth feature vector values are mapped onto low dimension hexagonal gridmap. The feature matching is performed between different image pairs and the depth pixels in the images are represented in terms of the winning neurons in the SOM network. The winner neuron is obtained in the images and invariant depth pixels are associated in the various image pairs. The same process is performed for all the image pairs in the video frames of the depth video captured with the Kinect. The optimization is done in the step 1 in Eq. (1) in terms of computational time. If the computation in Eq. (1) is done only on the nonzero values; the overall computations are highly reduced. The 128 dimension descriptor vector is passed to the Eq. (1) in order to pass for optimization and only the winner depth pixels are passed to the matching step. The resulted descriptor set consists of low dimension winning feature vector in the dataset of different depth image pairs. The Eq. 1 can be recomputed by following notation:

$$d_k(t) = \sum_{x_i != 0} x_i(t)(x_i(t) - 2w_{ki}(t)) + \sum_{i=1}^{n} w_{ki}(t)^2 \quad (3)$$

The computations are highly reduced by taking into consideration only the nonzero depth feature values. The complexity is reduced from O (N^2 * n) to O (N^2 * fnon zero * n) and computation is done over the nonzero depth image feature vectors. The running time is reduced to nonzero values and only 2 * xi(t) * $w_{ki}(t)$ computation is required in each iteration; thus the overall complexity is reduced to O (N^2 * fnon zero * n).

In addition, the computational overhead is further reduced in the neighborhood function calculation step, by considering only the first three neurons having the smallest distance from the input depth feature vector. The contribution of the third neuron neighborhood feature element is less than 1/9, which reduces the overall computation and thus the update neighborhood phase is decreased to O(N^2) with a constant of six since three weight vector columns and its squared components need to be updated.

5. Results & Discussion

A number of experiments have been conducted to investigate the performance of the proposed system. In the proposed system, the stable features are detected in the depth images and 3D model is reconstructed with the stable object features. Figure 5 shows an example frame observed with the RGB-D camera and depth image along with the 3D model information. Table 1 show the different experimental parameters used in the proposed system. The results of the feature matching in the different frames at different time intervals are shown in Figure 6.

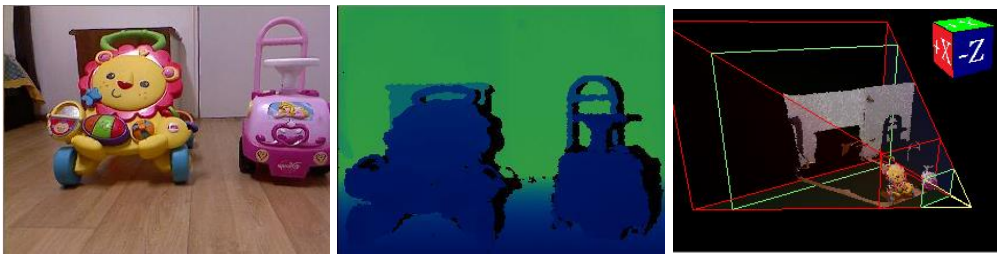

Figure 5. (left) RGB Image, (Middle) Depth Information Captured by an RGB-D Camera and (Right) 3D Model of an Object

Table 1. Experimental Parameters Set for the Proposed System

Parameters	Values					
Data Stream	Color			Depth		
Resolution	Color Image: 640 x 480 pixels			Depth Image: 640 x 480 pixels		
components	Color camera	Infrared (IR) emitter	IR depth sensor	Tilt motor	Microphone array	LED
LED	If LED turns GREEN, the Kinect device drivers are installed properly					
Tilt motor	The motor can be tilted vertically upwards or downwards by 27 degrees					

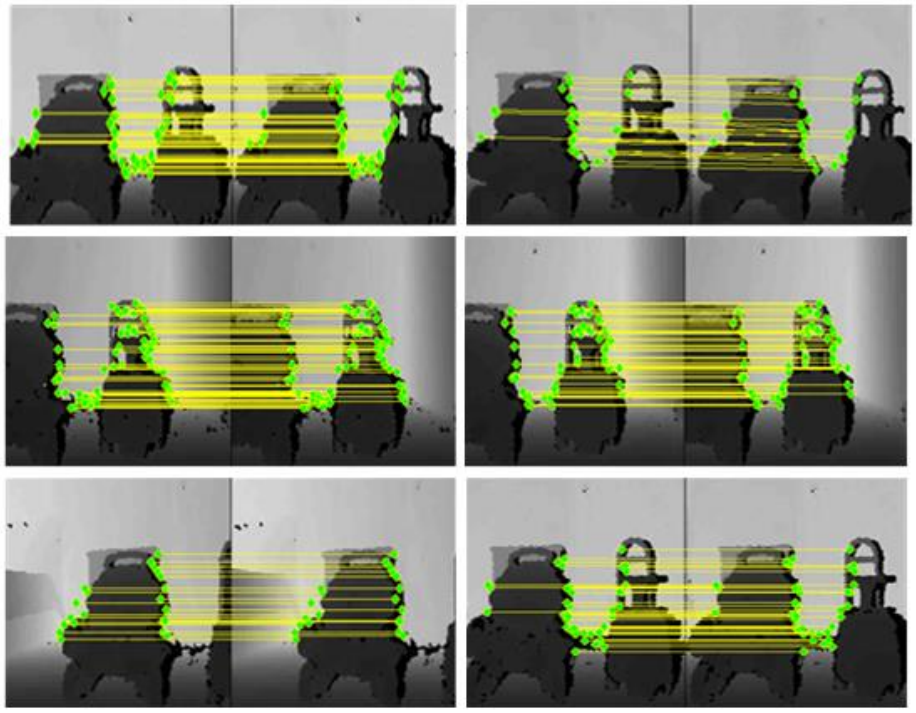

Figure 6. The Results of the Feature Matching in the Different Depth Frames at Different Time Intervals are shown. The Results showing that the Proposed Method can detect the Features in the Depth Image under Different Viewpoint Changing Conditions. The Points and Lines in the Depth Images Show the Features Points that are matched in the Depth Image

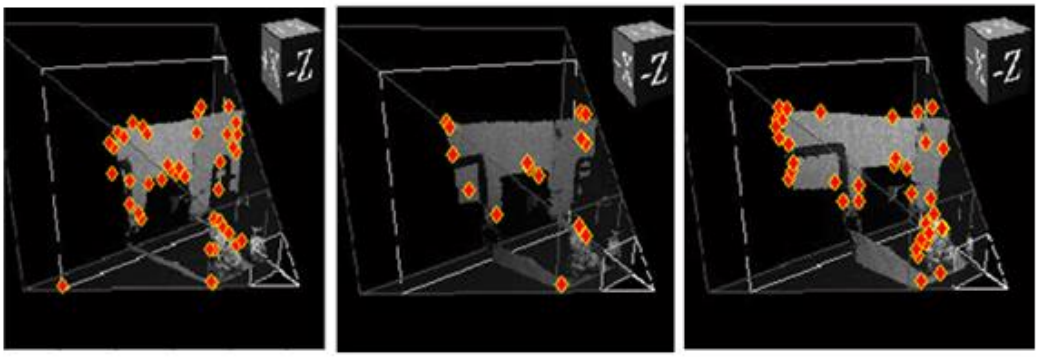

Figure 7. Demonstration of the 3D Model of the Objects for the Different Time Instances. The Colored Points in the Middle are Representatives of the Stable Depth Feature in Different Time Interval

During experiments, the camera was moved to capture the depth image dataset at different viewpoint changes such as rotation, scaling, occlusion, *etc*. The performance of the experiments is evaluated by comparing the results with the recent techniques. As shown in Figure 6, the systems resulted into efficient matched and stable depth features with very less computations as compared to the matching with SIFT method. The computations on the depth images are reduced as compared to the computations on depth images with the SIFT method. The 3D model of the object can be reconstructed with the proposed method using the stable depth features. The 3D models of the objects for the different time instances with stable features are shown in Figure 7 with the use to colored

points. The effectiveness of the system is proved with the feature matching time and the number of feature detection. The results of the experiments are given in Table 2.

Table 2. Results of Pixel Matching in Depth Image Pairs at Different Time Instances with the SIFT Method and the Proposed Method

Depth Image Pairs	Computational time			
	SIFT		Proposed Method	
	T	F	T	F
Depth pair (1, 2)	0.9150	134	0.0177	131
Depth pair (8, 9)	1.2980	95	0.0147	129
Depth pair (19,20)	1.1606	134	0.0148	131
Depth pair (34,35)	1.1098	134	0.0160	131
Depth pair (85,86)	1.1670	134	0.0156	131
Depth pair (132,133)	1.1721	134	0.0178	131
Depth pair (146,147)	1.1773	134	0.0149	131
Depth pair (180,181)	1.0646	0	0.0067	44
Depth pair (207,208)	0.9994	0	0.0064	38
Depth pair (227,228)	1.0325	82	0.0073	51
Depth pair (247,248)	1.1592	118	0.0096	52
Depth pair (269,270)	1.2133	118	0.0070	52
Depth pair (293,294)	1.0957	118	0.0074	52
Depth pair (360, 361)	1.2129	111	0.0144	125
Depth pair (369,370)	1.2036	112	0.0121	105
Depth pair (383,384)	1.2272	103	0.0131	110
Depth pair (397,398)	1.2141	118	0.0075	52
Depth pair (416,417)	1.2427	130	0.0075	36
Depth pair (427,428)	1.2801	10	0.0064	49
Depth pair (449,450)	1.3660	133	0.0076	59

6. Conclusion

In this paper, a feature matching technique in Kinect depth image pairs is proposed using Kinect captured depth videos. At first, an improved matching with feature reduction is proposed with self-organizing map. Next the features are optimized and compared with the recent matching techniques to obtain the stable depth features in the depth videos. The experimental results show out its good performance with higher stable matched features matched in fewer computations. The proposed method generates optimized matched features in the depth images. This method will be helpful for the design of various real-time applications such as autonomous path finding, traffic surveillance, and simultaneous localization and mapping (SLAM). In the future work, the tracking of objects by using depth features will be implemented to develop autonomous vision based path finding system.

References

[1] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images", Image and Vision Computing, vol. 10, no. 3, **(1992)**, pp. 145–155.
[2] P. Besland N. McKay, "A Method for Registration of 3-D Shapes", IEEE Trans. Pattern Anly. Mach. Intelli., vol. 14, no. 2, **(1992)**, pp. 239–256.
[3] P. J. Neugebauer, "Hochgenaue Objektlokalisation in Tiefenbildern", Visualisierung — Rolle von Interaktivit¨at und Echtzeit, Sankt Augustin, **(1992)**, Germany.

[4] G. Blais and M. Levine, "Registering Multiview Range Data to Create 3D Computer Objects", IEEE Trans. Pattern Anly. Mach. Intelli., vol. 17, no. 8, **(1995)**, pp. 820–824.
[5] P. Henry, M. Krainin, E. Herbst, X. Ren and D.Fox, "RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments", in Proc. of the Int. Symposium on Experimental Robotics (ISER) **(2010)**.
[6] L. Xia, C. Chen and J. Aggarwal, "Human detection using depth information by Kinect", in Proc. Int. Workshop HAU3D, **(2011)**, pp. 15–22.
[7] C. Rougier, E. Auvinet, J. Rousseau, M. Mignotte and J. Meunier, "Fall detection from depth map video sequences", in Proc. ICOST, **(2011)** pp. 121–128.
[8] Z. Zhang, W. Liu, V. Metsis and V. Athitsos, "A viewpoint-independent statistical method for fall detection", in Proc. ICPR, **(2012)**, pp. 3626–3630.
[9] K. Lai, L. Bo, X. Ren and D. Fox, "A large-scale hierarchical multiview RGB-D object dataset", in Proc. IEEE Int. Conf. Robot. Autom., **(2011)**, pp. 1817–1824.
[10] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor segmentation and support inference from RGBD images", in Proc. Eur. Conf. Comput. Vision, **(2012)**, pp. 746–760.
[11] K. Lai, L. Bo, X. Ren and D. Fox, "Sparse distance learning for object recognition combining RGB and depth information", in Proc. IEEE Int. Conf. Robot. Autom., **(2011)**, pp. 4007–4013.
[12] P. Henry, M. Krainin, E. Herbst, X. Ren and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3-D modeling of indoor environments", Int. J. Robot. Res., vol. 31, no. 5, **(2012)**, pp. 647–663.
[13] N. Engelhard, F. Endres, J. Hess, J. Sturm and W. Burgard, "Real-time 3-D visual SLAM with A hand-held camera", in Proc. RGB-D Workshop 3-D Perception Robot. Eur. Robot. Forum, **(2011)**.
[14] D. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision, vol. 60, no. 2, **(2004)**, pp. 91–110.
[15] E. Rosten, R. Porter and T. Drummond, "Faster and Better: A machine learning approach to corner detection", IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, **(2010)**, pp. 105–119.
[16] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "SURF: Speeded up robust features", Comput. Vision Image Understanding, vol. 110, no. 3, **(2008)**, pp. 346–359.
[17] Kinect camera. http://www.xbox.com/en-US/ kinect/default.htm
[18] A. Jana, " Kinect for Windows SDK Programming Guide," Packt Publishing, **(2012)**.

Author

Kajal Sharma received a B.E. degree in computer engineering from University of Rajasthan, India in 2005, and M.Tech. and Ph.D. degrees in computer science from Banasthali University, Rajasthan, India in 2007 and 2010. From October 2010 to September 2011, she worked as a postdoctoral researcher at Kongju National University, Korea. Since October 2011 to April 2013, she worked as a postdoctoral researcher at the School of Computer Engineering, Chosun University, Gwangju, Korea. Presently she is working as an independent researcher in Korea. Her research interests include image and video processing, neural networks, computer vision, and robotics. She has published many research papers in various national and international journals and conferences.

