

RGB-D datasets using microsoft kinect or similar sensors: a survey

Ziyun Cai¹ · Jungong Han² · Li Liu² · Ling Shao²

Received: 1 December 2015 / Revised: 2 February 2016 / Accepted: 15 February 2016 /
Published online: 19 March 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract RGB-D data has turned out to be a very useful representation of an indoor scene for solving fundamental computer vision problems. It takes the advantages of the color image that provides appearance information of an object and also the depth image that is immune to the variations in color, illumination, rotation angle and scale. With the invention of the low-cost Microsoft Kinect sensor, which was initially used for gaming and later became a popular device for computer vision, high quality RGB-D data can be acquired easily. In recent years, more and more RGB-D image/video datasets dedicated to various applications have become available, which are of great importance to benchmark the state-of-the-art. In this paper, we systematically survey popular RGB-D datasets for different applications including object recognition, scene classification, hand gesture recognition, 3D-simultaneous localization and mapping, and pose estimation. We provide the insights into the characteristics of each important dataset, and compare the popularity and the difficulty of those datasets. Overall, the main goal of this survey is to give a comprehensive description about the available RGB-D datasets and thus to guide researchers in the selection of suitable datasets for evaluating their algorithms.

✉ Ling Shao
ling.shao@northumbria.ac.uk

Ziyun Cai
cziyun1@sheffield.ac.uk

Jungong Han
jungong.han@northumbria.ac.uk

Li Liu
li2.liu@northumbria.ac.uk

¹ Department of Electronic and Electrical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK

² Department of Computer Science and Digital Technologies, Northumbria University, Newcastle Upon Tyne NE1 8ST, UK

Keywords Microsoft Kinect sensor or similar devices · RGB-D dataset · Computer vision · Survey · Database

1 Introduction

In the past decades, there has been abundant computer vision research based on RGB images [3, 18, 90]. However, RGB images usually only provide the appearance information of the objects in the scene. With this limited information provided by RGB images, it is extremely difficult, if not impossible, to solve certain problems such as the partition of the foreground and background having similar colors and textures. Additionally, the object appearance described by RGB images is not robust against common variations, such as illuminance change, which significantly impedes the usage of RGB based vision algorithms in realistic situations. While most researchers are struggling to design more sophisticated algorithms, another stream of the research turns to find a new type of representation that can better perceive the scene. RGB-D image/video is an emerging data representation that is able to help solve fundamental problems due to its complementary nature of the depth information and the visual (RGB) information. Meanwhile, it has been proved that combining RGB and depth information in high-level tasks (i.e., image/video classification) can dramatically improve the classification accuracy [94, 95].

The core of the RGB-D image/video is the depth image, which is usually generated by a range sensor. Compared to a 2D intensity image, a range image is robust to the variations in color, illumination, rotation angle and scale [17]. Early range sensors (such as Konica Minolta Vivid 910, Faro Lidar scanner, Leica C10 and Optech ILRIS-LR) are expensive and difficult to use for researchers in a human environment. Therefore, there is not much follow-up research at that time. However, with the release of the **low-cost** 3D Microsoft Kinect sensor¹ on 4th November 2010, acquisition of RGB-D data becomes cheaper and easier. Not surprisingly, the investigation of computer vision algorithms based on RGB-D data has attracted a lot of attention in the last few years.

RGB-D images/videos can facilitate a wide range of application areas, such as computer vision, robotics, construction and medical imaging [33]. Since a lot of algorithms are proposed to solve the technological problems in these areas, an increasing number of RGB-D datasets have been created so as to verify the algorithms. The usage of publicly available RGB-D datasets is not only able to save time and resources for researchers, but also enables fair comparison of different algorithms. However, it may not be practical and also not efficient to test a designed algorithm on all available datasets. At certain situation, one has to make a sound choice depending on the target of the designed algorithm. Therefore, the selection of the RGB-D datasets becomes important for evaluating different algorithms. Unfortunately, we fail to find any detailed surveys about RGB-D datasets and their collection, classification and analysis. To the best of our knowledge, there is only one short overview paper devoted to the description of available RGB-D datasets [6]. Compared to that paper, this survey is much more comprehensive and provides individual characteristics and comparisons about different RGB-D datasets. More specifically, our survey elaborates 20 popular RGB-D datasets covering most of RGB-D based computer vision applications. Basically, each dataset is described in a systematic way, involving dataset name, ownership,

¹<http://www.xbox.com/en-US/xbox-360/accessories/kinect/>, Kinect for Xbox 360.

context information, the explanation of ground truth, and example images or video frames. Apart from these 20 widely used datasets, we also briefly introduce another 26 datasets that are less popular in terms of their citations. In order to save the space, we only add them into the summary tables. But we believe that the readers can understand the characteristics of those datasets even though we only provide compact descriptions. Furthermore, this survey proposes five categories to classify existing RGB-D datasets and corrects some careless mistakes on the Internet about certain datasets. The motivation of this survey is to provide a comprehensive and systematic description of popular RGB-D datasets for the convenience of other researchers in this field.

The rest of this paper is organized as follows. In Section 2, we briefly review the background, hardware and software information about Microsoft Kinect. In Section 3, we describe 20 popular publicly available RGB-D benchmark datasets according to their application areas in detail. In total, 46 RGB-D datasets are characterized in three summary tables. Meanwhile, discussions and analysis of the datasets are given. Finally, we draw the conclusion in Section 4.

2 A brief review of kinect

In the past years, as a new type of scene representation, RGB-D data acquired by the consumer-level Kinect sensor has shown the potential to solve challenging problems for computer vision. The hardware sensor as well as the software package are released by Microsoft in November 2010 and have a vast of sales until now. At the beginning, Kinect acts as an Xbox accessory, enabling players to interact with the Xbox 360 through body language or voice instead of the usage of an intermediary device, such as a controller. Later on, due to its capability of providing accurate depth information with relatively low cost, the usage of Kinect goes beyond gaming, and is extended to the computer vision field. This device equipped with intelligent algorithms is contributing to various applications, such as 3D-simultaneous localization and mapping (SLAM) [39, 54], people tracking [69], object recognition [11] and human activity analysis [13, 57], etc. In this section, we introduce Kinect from two perspectives: hardware configuration and software tools.

2.1 Kinect hardware configuration

Generally, the basic version of Microsoft Kinect consists of a RGB camera, an infrared camera, an IR projector, a multi-array microphone [49] and a motorized tilt. Figure 1 shows the components of Kinect and two example images captured by RGB and depth sensors, respectively. The distance between objects and the camera is ranging from 1.2 meters to 3.5 meters. Here, RGB camera is able to provide the image with the resolution of 640×480 pixels at 30 Hz. This RGB camera also has option to produce higher resolution images (1280×1024 pixels), running at 10 Hz. The angular field of view is 62 degrees horizontally and 48.6° vertically. Kinect's 3D depth sensor (infrared camera and IR projector) can provide depth images with the resolution of 640×480 pixels at 30 Hz. The angular field of this sensor is slightly different with that of the RGB camera, which is 58.5 degrees horizontally and 46.6° degrees vertically. In the application such as NUI (Natural User Interface), the multi-array microphone can be available for a live communication through acoustic source localization of Xbox 360. This microphone array actually consists of four microphones, and the channels of which can process up to 16-bit audio signals at a sample rate of 16 kHz. Following



Fig. 1 Illustration of the structure and internal components of the Kinect sensor. Two example images from RGB and depth sensors are also displayed to show their differences

Microsoft, Asus launched Xtion Pro Live,² which has more or less the same features with Kinect. In July 2014, Microsoft released the second generation Kinect: Kinect for windows v2.³ The difference between Kinect v1 and Kinect v2 can be seen in Table 1. It is worth noting that this survey mainly considers the datasets generated by Kinect v1 sensor, but only lists a few datasets created by using other range sensors, such as Xtion Pro Live and Kinect v2 sensor. The reason is that the majority of RGB-D datasets being used are generated with the aid of Kinect v1 sensor.

In general, the technology used for generating the depth map is based on analyzing the speckle patterns of infrared laser light. The method is patented by PrimeSense [27]. For more detailed introductions, we refer to [30].

2.2 Kinect software tools

When Kinect is initially released for Xbox360, Microsoft actually did not deliver any SDKs. However, some other companies forecast an explosion in using Kinect and thus provide unofficial free libraries and SDKs. The representatives include CL NUI Platform,⁴ OpenKinect/Libfreenect,⁵ OpenNI⁶ and PCL.⁷ Although most of libraries provide basic algorithmic comments, such as camera calibration, automatic body calibration, skeletal tracking, facial tracking, 3-D scanning and so on, each library has its own characteristics. For example, CL NUI Platform developed by NUI researchers can obtain the data from RGB camera, depth sensor and *accelerometer*. Open Kinect focuses on providing free and open source libraries, enabling researchers to use Kinect over Linux, Mac and Windows. OpenNI is an industry-led open source library which can program RGB-D devices for NUI

²<http://www.asus.com>, Asus Corporation, Xtion Pro Live

³<http://www.xbox.com/en-GB/xbox-one/accessories>, Microsoft Corporation, Kinect v2 for Xbox 360.

⁴<http://codelaboratories.com/kb/nui>, CL NUI Platform [Online].

⁵<https://github.com/OpenKinect/libfreenect/>, OpenKinect [Online].

⁶<http://www.openni.org/>, OpenNI [Online].

⁷<http://www.pointclouds.org/>, PCL [Online].

Table 1 Comparison between Kinect v1 and Kinect v2

		Kinect for windows v1	Kinect for windows v2
Color	Resolution	640×480	1920×1080
	fps	30fps	30fps
Depth	Resolution	640×480	512×424
	fps	30fps	30fps
Sensor		Structured light	Time of flight
Range		1.2 ~ 3.5m	0.5 ~ 4.5m
Joint		20 joint / people	25 joint / people
Hand state		Open / closed	Open / closed / Lasso
Number of Apps		Single	Multiple
Body Tracking		2 people	6 people
Body Index		6 people	6 people
Angle of View	Horizontal	62 degree	70 degree
	Vertical	48.6 degree	60 degree
Tilt Motor		Yes	No
Aspect Ratio		4:3	6:5
Supported OS		Win 7, Win 8	Win 8
USB Standard		2.0	3.0

applications. It is not specifically built for Kinect, and it can support multiple PrimeSense 3D sensors. Normally, users need to install SensorKinect, NITE, and OpenNI to control the Kinect sensor, where SensorKinect is the driver of Kinect and NITE is the middleware provided by PrimeSense. The latest version of OpenNI is the version 2.2.0.33 until June 2015. The Point Cloud Library (PCL) is a standalone open source library which provides SLAM-related tools such as surface reconstruction, sample consensus, feature extraction, and visualization for RGB-D SLAM. It is licensed by Berkeley Software Distribution (BSD). More details and publications about PCL can be found in [74].

The official version of Kinect for Windows SDK⁸ was released in July 2011, which provides a straightforward access to Kinect data: depth, color and disparity. The newest version is the SDK 2.0. It can be applied for Windows 7, Windows 8, Windows 8.1 and Windows Embedded 8 with C++, C# or VB.NET. The development environment uses Visual Studio 2010 or higher versions. Regarding the software tool, it mainly contains skeletal tracking, higher depth fidelity, audio processing and so on.

The comparison of Kinect Windows SDK and unofficial SDK, e.g., OpenNI, can be summarized below. The detailed same and difference between the Kinect Windows SDK and unofficial SDK can be seen in Table 2.

Kinect Windows SDK:

- 1) It supports audio signal processing and allows to adjust the motor angle.
- 2) It provides a full-body tracker including head, feet, hands and clavicles. Meanwhile, some details such as occluded joints are processed meticulously.

⁸<http://www.microsoft.com/en-us/kinectforwindows/>, Microsoft Kinect SDK [Online].

Table 2 Comparison between the Kinect Windows SDK and unofficial SDK

	Kinect windows SDK	Unofficial SDK
Supported OS	Windows 7×86/×64 Windows 8, Windows 8.1 and Windows Embedded 8	Windows XP/Vista/7×86/×64 Windows 8, Windows 8.1 and Windows Embedded 8 LinuxUbuntu×86/×64 Mac OS Android
Development language	C++, C#	C, C++, C#, Java
Commercial use	No	Yes
Supports for audio and motor/tilt	Yes	No
Supports multiple sensors	Yes	No
Consumption of CPU power	More	Less
Full body tracking	Includes head, hands, feet, clavicles Calculates positions for the joints, but not rotations Only tracks the full body, no hands only mode	No head, hands, feet, clavicles Calculates both positions and rotations for the joints Supports for hands only mode
Supports for Unity3D game engine	No	Yes
Supports for record/playback to disk	No	Yes
Supports to stream the raw InfraRed video data	No	Yes

3) Multiple Kinect sensors can be supported.

OpenNI/NITE library:

- 1) Commercial use of OpenNI is allowed.
- 2) Frameworks for hand tracking and hand-gesture recognition are included in OpenNI. Moreover, it automatically aligns the depth image and the color image.
- 3) It consumes less CPU power than that of Kinect Windows SDK.
- 4) It supports Windows, Linux and Mac OSX. In addition, streaming the raw Infrared video data becomes possible.

In conclusion, the most attractive advantage of OpenNI is the feasibility for multiple operational platforms. Besides it, using OpenNI is more convenient and can obtain better results for the research of colored point clouds. However, in terms of collection quality of the original image and the technology for pre-processing, Kinect for Windows SDK seems to be more stable. Moreover, Kinect for Windows SDK is more advantageous when requiring skeletal tracking and audio processing.

3 RGB-D benchmark datasets

We will describe publicly available RGB-D datasets for different computer vision applications in this section. Since the Kinect sensor was just released a few years ago, most RGB-D datasets are created in a time range from 2011 to 2014. To have a clear structure, this paper divides the RGB-D datasets into 5 categories depending on the facilitated computer vision applications. More specifically, the reviewed datasets fall into object detection and tracking, human activity analysis, object and scene recognition, SLAM (Simultaneous Localization and Mapping) and hand gesture analysis. However, each dataset may not be limited to one specific application only. For example, object RGB-D can be used in detection as well. Figure 2 illustrates a tree-structured taxonomy that our review intends to follow.

In the following sections each dataset is described in a systematic way, attending to a collection of name, general information of dataset, example video images, context, ground truth, applications, creation procedure, creation environment and the published papers that used this dataset. In each category, the datasets will be presented in a chronological order. If several datasets are created in the same year, the dataset with more references will be introduced ahead of the others. General information of dataset includes the creator as well as the creation time. The context contains the information about the scenes, the number of objects and the number of RGB-D sensors. The ground truth reveals information concerning what type of knowledge in each dataset is available, such as bounding boxes, 3D geometries, camera trajectories, 6DOF poses and dense multi-class labels. Moreover, the complexity of the background, change of illumination and occlusion conditions are also discussed. At last, a list of publications using the dataset is also mentioned. In order to have a direct comparison of all the datasets, the complete information is compiled in three tables. It is worth noting that we describe the following representatives in more details. The characteristics of other datasets which are not popular are only summarized in the comparison tables due to the limited space. Moreover, the link cites of datasets, data size and citation are added into these tables as well.

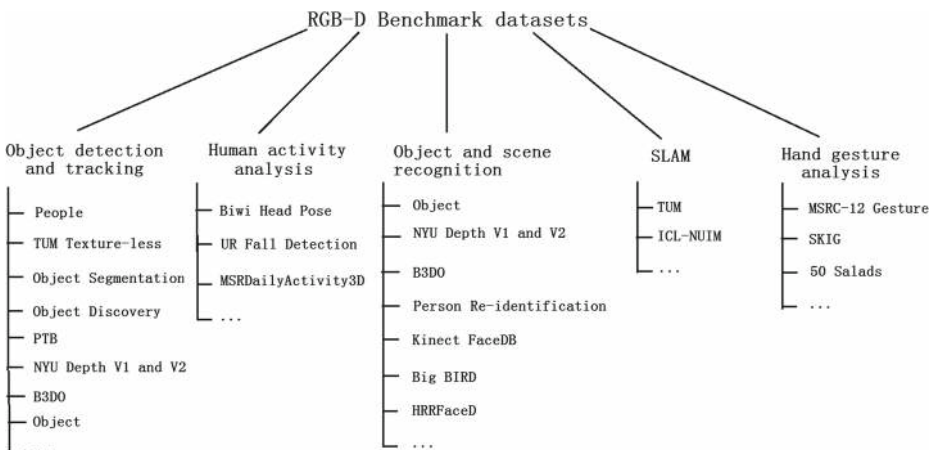


Fig. 2 Tree-structured taxonomy of RGB-D datasets reviewed in this paper

3.1 RGB-D Datasets for object detection and tracking

Object detection and tracking is one of the fundamental research topics in computer vision. It is an essential building-block of many intelligent systems. As we mentioned before, the depth information of an object is immune to changes of the object appearance or/and environmental illumination, and subtle movements of the background. With the availability of the low-cost Kinect depth camera, researchers immediately noticed that the feature descriptor based on depth information can help significantly detect and track the object in the real world where all kinds of variations occur. Therefore, RGB-D based object detection and tracking have attracted great attention in recent a few years. As a result, many datasets are created for evaluating proposed algorithms.

3.1.1 RGB-D people dataset

RGB-D People dataset [59, 83] was founded in 2011 by social Robotics Lab (SRL) of University of Freiburg with the purpose of evaluating people detection and tracking algorithms for robotics, interactive systems and intelligent vehicles. The data information is collected in an indoor environment (lobby of a large University canteen) with unscripted behavior of people during the lunch time. The video sequences are recorded through a setup of three vertically combined Kinect sensors (the field of view is $130^\circ \times 50^\circ$) at 30 Hz. The distance between this capturing device and the ground is about 1.5m. This guarantees that the three images can be acquired synchronously and simultaneously, and meanwhile, it is also able to reduce the IR projector cross-talk among these sensors. Moreover, in order to avoid detector bias, some background samples are recorded from another building within the University campus.

RGB-D people dataset collects more than 3000 frames of multiple persons walking and standing in the University hall from different views. To make the data more realistic, occlusions among persons appear in most sequences. Regarding to the ground truth, all frames are annotated manually to contain bounding box in the 2D depth image space and the visibility status of subjects.

To facilitate the evaluation of human detection algorithms, in total 1088 frames including 1648 instances of people have been labeled. Three sampled color and depth images from this dataset can be found in Fig. 3.

3.1.2 TUM texture-less 3D objects dataset

TUM Texture-Less 3D Objects dataset [36] was constructed by Technical University of Munich in 2012. It can be widely used for object segmentation, automatic modeling and



Fig. 3 Three sampled color (*left*) and depth (*right*) images from RGB-D People Dataset [59]

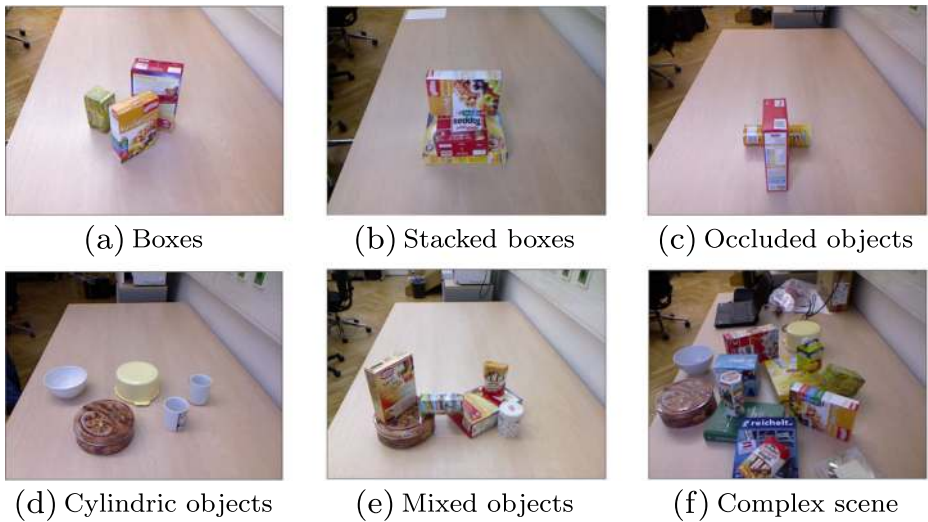


Fig. 4 Examples from Object Segmentation dataset.⁹ From left to right in order, (a) Boxes, (b) Stacked boxes, (c) Ocluded objects, (d) Cylindric objects, (e) Mixed objects, (f) Complex scene

3D object tracking. This dataset consists of 18,000 images describing 15 different texture-less 3D objects (ape, bench vise, can, bowl, cat, cup, duck, glue, hole puncher, drill, iron, phone, lamp, egg box and cam) accompanied with their ground truth poses. In the collection process, each object with its markers that can provide the corresponding ground truth poses was stuck to a planar board for model and image acquisition. Afterwards, through a simple voxel based approach, every object was reconstructed based on several images and the corresponding poses. At last, close and far range 2D and 3D clutters were added into the scene. Each sequence comprises more than 1,100 real images from different views ($0^\circ \sim 360^\circ$ around the object, $0^\circ \sim 90^\circ$ tilt rotation, $65\text{cm} \sim 115\text{cm}$ scaling and $\pm 45^\circ$ in-plane rotation). With respect to the ground truth, 6DOF pose was labeled for each object in each image. More details about this dataset can be found in [36, 37].

3.1.3 Object segmentation dataset (OSD)

Vision for robotics group in Vienna University of Technology created Object Segmentation dataset in 2012 for the evaluation of segmenting unknown objects from generic scenes [72]. This dataset is composed of 111 RGB-D images representing stacked and occluded objects on a table in six categories (boxes, stacked boxes, occluded objects, cylindric objects, mixed objects and complex scene 11). The labels of segmented objects for all RGB-D images are provided as the ground truth. Examples from this dataset are shown in Fig. 4.

3.1.4 Object disappearance for object discovery datasets

Department of Computer Science in Duke University created Object Disappearance for Object Discovery datasets [60] in 2012 for evaluating object detection, object recognition,

⁹<http://www.acin.tuwien.ac.at/?id=289>.

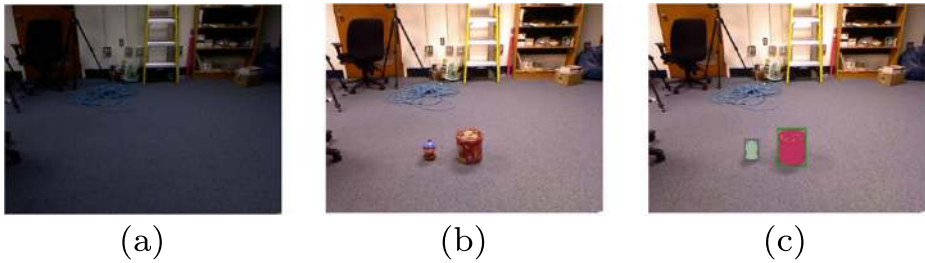


Fig. 5 Sampled images from the small dataset [60]. From left to right are: (a) image from the objects vanishing, (b) image when the objects appear and (c) image with the segmentations, respectively

localization and mapping algorithms. In this dataset, there are three sub datasets with gradually increased size and complexity. All images are recorded through a single Kinect sensor which is mounted on the top of a Willow Garage PR2 (robot). The RGB and depth images in these datasets are with the resolution of 1280×960 and the resolution of 640×480 , respectively. As the main objective is to facilitate the object detection, the image capturing rate is rather low, which is only 5 Hz. In order to minimize the range errors, Kinect is placed at a distance of 2 meters to the object. For the sake of clarity, we call these three datasets as small dataset, medium dataset and large dataset, respectively. Example images can be found in Figs. 5, 6 and 7.

Let's now elaborate each sub dataset. The small dataset consists of totally 350 images, in which 101 images are captured from a static scene without any objects, 135 images describe the same scene but with two objects, and 114 images which remove these two objects which is equivalent to 110 images before the objects appear. There are two unique objects, and 270 segmented objects found by hand.

In the medium dataset, there are totally 484 frames. The step-by-step video capturing procedure can be explained as follows. The robot firstly observes a table with objects, i.e., a dozen of eggs and a toy. It then looks away while the objects are removed. After a short while, it observes the table again. Lately, the robot travels approximately 18 meters and repeats this procedure with a counter. To make the dataset more challenging, the lighting during the recording keeps changing. The hand-segmentation results in 394 segmented objects from a total of four unique objects.

The large dataset contains the whole cover of several rooms of a $40\text{m} \times 40\text{m}$ office environment, resulting in totally 397 frames. In the first process, the robot shoots some

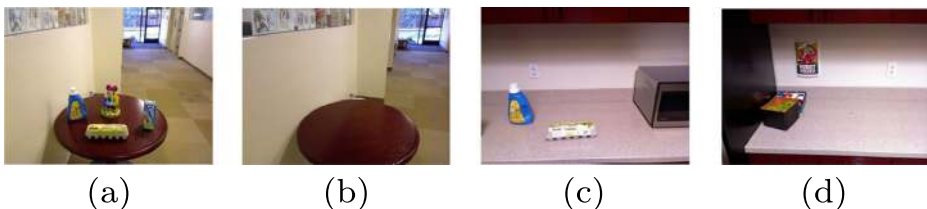


Fig. 6 Sample images from the medium dataset [60]. From first to end: (a) image with objects, (b) image without objects, (c) image of the counter with objects and (d) image of the counter without objects



Fig. 7 Sample images from the large dataset [60]. From left to right there are four different places involved in the video sequence recording. The objects in the environment are unique

objects (i.e., toy and different shapes of boxes) in each room. In the second process, the robot observes the rooms after all the objects have been removed. There are seven unique objects, and 419 segmented objects found by hand.

3.1.5 Princeton tracking benchmark dataset

Princeton Tracking Benchmark (PTB) [81] was created in 2013 with 100 videos, covering many realistic cases, such as deformable objects, moving camera, different occlusion conditions and a variety of clutter backgrounds. The object types in PTB are rich and varied in the sense that it includes both deformable objects and relatively rigid objects that only have rotating and translating motions. The deformable objects are mainly animals (including rabbits, dogs and turtles) and humans, while the rigid objects contain human heads, balls, cars and toys. The movements of animals are made up of out-of-plane rotations and deformations. The scene types consist of a few different kinds of background, e.g., a living room that has a changeless and stationary background and a restaurant that has complex backgrounds with many people walking around. Several occluding cases are also involved in the videos. In the paper [81], authors provide the statistics of movement, object category and scene type about PTB dataset, which can be summarized in Table 3.

The ground truth generation of this dataset is purely based on manual annotations. That is, they draw a bounding-box around an object on each frame. To obtain a high consistency, all frames are manually annotated by the same person. The drawing rule is applied such that the target is covered by an initialized minimum bounding box on the first frame. The bounding-box will be adjusted while the target is moving or its shape is changing over time. Concerning the occlusion cases, they have several rules. For instance, if the target is occluded, the bounding box will only cover the visible part of the target. Otherwise, no bounding box will be provided in case the target is completely occluded. Figure 8 shows some example frames obtained from PTB dataset.

In view of detailed descriptions for the above five datasets, we come to the conclusion that RGB-D People dataset is more challenging than the others. The difficulty of this dataset is that the majority of people are dressed similarly and the brightness suddenly changes among frames. Due to its realistic, most related papers prefer to test their algorithms on this dataset. The recent tracking evaluation on this dataset shows that the best algorithm achieves 78 % MOTA (avg. number of times of a correct tracking output with respect to the ground truth), 16.8 % false negative (FN), 4.5 % false positives (FP) and 32 mismatches (ID) [59]. Besides it, some algorithm-comparison reports based on other datasets, e.g., TUM Textureless, Object Segmentation, Object Discovery and PTB, can be found in [7, 61, 73] and [29], respectively.

Table 3 Statistics about Princeton tracking benchmark dataset

movement	Kinect	stationary	moving
	movement	85 %	15 %
movement	occlusion	no occlusion	target occluded in some frames
		44 %	56 %
	object speed (<i>m/s</i>)	<0.15	0.35~0.45
		13 %	13 %
type	passive	active	
	30 %	70 %	
object category	size	large	small
		37 %	63 %
scene type categories	deformability	deformable	non-deformable
		67 %	33 %
scene type categories	type	human	rigid object
		42 %	37 %
	office	room	concourse
		33 %	4 %
office	library	sports field	
	19 %	5 %	
office	shop		
	24 %		
office	animal		
	21 %		
office	shop		
	24 %		



Fig. 8 Samples from the Princeton Tracking Benchmark dataset include deformable objects, various occlusion conditions, moving camera, and different scenes [81]

3.2 Human activity analysis

Apart from offering a low-cost camera sensor that outputs both RGB and depth information, another contribution of Kinect is a fast human-skeletal tracking algorithm. This tracking algorithm is able to provide the exact location of each joint of a human body over time, which makes the interpretation of complex human activities easier. Therefore, a lot of works are devoting to deducing human activities from depth images or the combination of depth and RGB images. Not surprisingly, many RGB-D datasets that can be used to verify human activity analysis algorithms arose in recent a couple of years.

3.2.1 Biwi kinect head pose dataset

Biwi Kinect Head Pose dataset [25] was generated by computer vision laboratory of ETH Zurich in 2013 for estimating the location and orientation of a person's head from the depth data. This dataset is recorded when some people are facing to a Kinect (about one meter away) and turning their heads around randomly. The turning angle covers a range of ± 75

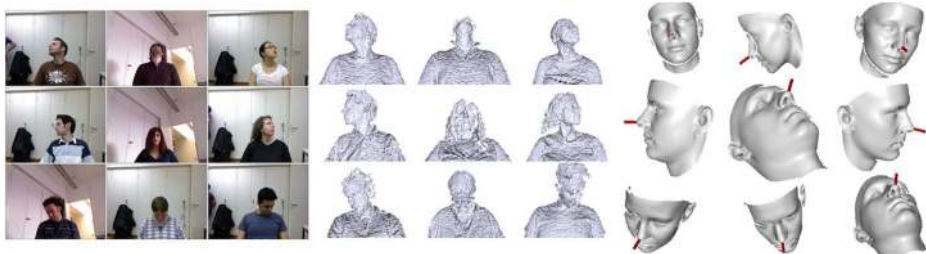


Fig. 9 From left to right: RGB images, depth images and depth images with annotations from Biwi Kinect Head Pose Dataset [25]

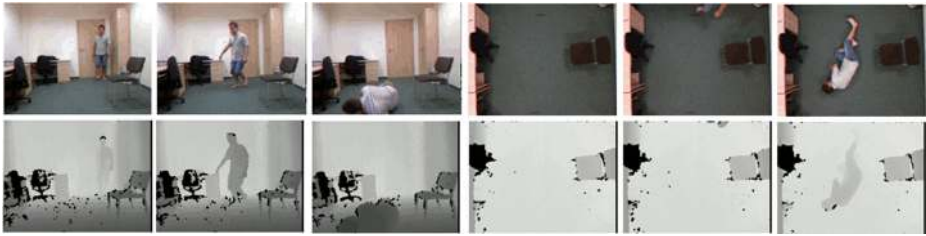


Fig. 10 Corresponding sample RGB images (*top*) and depth images (*bottom*) from UR Fall Detection dataset¹⁰

degrees for yaw, ± 60 degrees for pitch and ± 50 degrees for roll. Biwi Kinect Head Pose dataset consists of over 15K images from 24 sequences (6 women, 14 men and 4 of them are with glasses). It provides pairs of depth image and RGB image (640×480 pixels at 30Hz) as well as the annotated ground truth (see Fig. 9). The annotation is done by using the software “face shift” in the form of the 3D location of the head (3D coordinates of the nose tip) and the head rotation angles (represented as Euler angles). It can be seen from the sample images (Fig. 9) that a red cylinder going through the nose indicates the nose’s position and the head’s turning direction. Some algorithms tested on this dataset can be found in [4, 56, 71].

3.2.2 UR fall detection dataset

University of Rzeszow created UR Fall Detection dataset in 2014 [51], which devotes to detecting and recognizing human falls. In this dataset, the video sequences are recorded by two Kinect cameras. One is mounted at the height of approximate $2.5m$ such that it is able to cover the whole room ($5.5m^2$). The other one is supposed to be parallel to the fall with a distance about $1m$ from the ground.

In this dataset, there are totally 60 sequences that record 66 falls when conducting common daily activities, such as walking, taking or putting an object from floor, bending right or left to lift an object, sitting, tying laces, crouching down and lying. Meanwhile, corresponding accelerometer data are also collected using an elastic belt attached to the volunteer. Figure 10 shows some images sampled from this dataset.

3.2.3 MSRDailyActivity3D dataset

MSRDailyActivity3D dataset [97] was created by Microsoft Research in 2012 for evaluating the action recognition approaches. This dataset is designed to discover human’s daily activities in the living room, which contains 16 daily activities (drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up and sit down). There are 10 subjects performing each activity twice with the postures of standing and sitting respectively. In general, this dataset is fairly challenging, because the 3D joint positions extracted by the skeleton tracker become unambiguous when the performer is close to the sofa, which is a common situation in a living room. Meanwhile, most of the activities contain the humans-object interactions. Examples of RGB images, raw depth images in this dataset are illustrated in Fig. 11.

¹⁰<http://fenix.univ.rzeszow.pl/mkepski/ds/uf.html>.

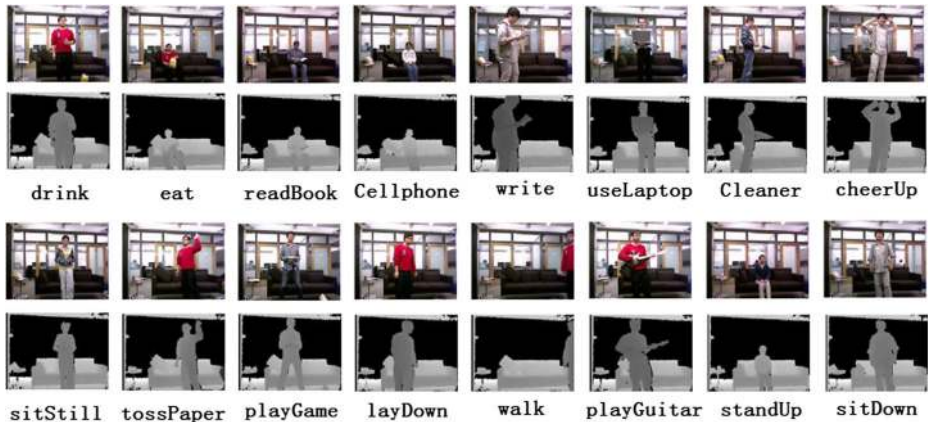


Fig. 11 Selected RGB (*top*) and raw depth images (*bottom*) from MSRDailyActivity3D dataset

Biwi Kinect Head Pose dataset was created in 2013 but it already has nearly 100 citations. The best result found in the literature shows that the detected yaw error, pitch error and roll error are $3.5^\circ \pm 5.8^\circ$, $3.8^\circ \pm 6.5^\circ$ [25] and $4.7^\circ \pm 4.6^\circ$ [20] respectively. Seen from the result, it is clear that more research efforts are needed in order to achieve a better result. UR Fall Detection dataset is a relatively new RGB-D dataset so that we only find a few algorithms tested on this dataset. According to [51], the best baseline results are achieved by Threshold UFT method [8], which are 95.00 % accuracy, 90.91 % precision, 100 % sensitivity and 90.00 % specificity. MSRDailyActivity3D dataset is a very challenging dataset, and it has the largest number of citations which is over 300 to the date. The best result of the action recognition accuracy achieved on this dataset can only reach 85.75 % [98].

3.3 Object and scene recognition

Object recognition aims to answer the question whether the image contains the pre-defined object, given an input image. Scene recognition is a sort of extension of object recognition, densely labeling everything in a scene. Usually, an object recognition algorithm relies on the feature descriptor, which is able to distinguish the different objects, and meanwhile, tolerate various distortions of the object due to the environmental variations, such as change of illumination, different levels of occlusions, and reflections, etc. Usually, the conventional RGB-based feature descriptors are sufficiently descriptive, but they may suffer from the distortions of an object. RGB information, by nature, is less capable of handling those environmental variations. Fortunately, the combination of RGB and depth information may potentially enhance the robustness of the feature descriptor. Consequently, many object/scene descriptors assembling RGB and depth information are proposed in recent a few years. In accordance with the research growth, several datasets are generated for the public usage.

3.3.1 RGB-D object dataset

University of Washington and Intel Labs Settle released this large-scale RGB-D object dataset on June 20, 2011 [52]. It contains 300 common household objects (i.e., apple, banana, keyboard, potato, mushroom, bowl, coffee mug) which are classified into 51



Fig. 12 Sample objects from the RGB-D object dataset (*left*), examples of RGB image and depth image of an object (*right top*) and RGB-D scene images (*right bot*) [52]

categories. Each object in this dataset was recorded from multiple view angles with resolution of 640×480 at 30 **Hz**, thus resulting in 153 video sequences (3 video sequences for each object) and nearly 250,000 RGB-D images. Figure 12 illustrates some selected objects from this dataset as well as the examples of RGB-D images. Through WordNet hyponym/hypernym relations, the objects are arranged in a hierarchical structure, which helps many possible algorithms. Ground truth **pose** information and per-frame bounding boxes about all these 300 objects are offered in the dataset. On April 5, 2014, the RGB-D scenes dataset was upgraded to v.2, adding 14 new scenes with the tabletop and furniture objects. This new dataset further boosts the research on applications such as category recognition, instance recognition, 3D scene labeling and object pose estimation [9, 10, 53].

To help researchers use this dataset, RGB-D object dataset provides code snippets and software for RGB-D kernel descriptors, reading point clouds (MATLAB) and spinning images (MATLAB) on their website. The performance comparison of different methods tested on this dataset is also reported on the web.

3.3.2 NYU Depth V1 and V2

Vision Learning Graphics (VLG) lab in New York University created the NYU Depth V1 for indoor-scene object segmentation in 2011. Compared to most works in which the scenes are in a very limited domain [55], this dataset is collected from much wider domains (the background is changing from one to another), facilitating multiple applications. It records video sequences of a great diversity of indoor scenes [79], including a subset of densely labeled video data, raw RGB images, depth images and accelerometer information. On the website, users can find a toolbox for processing data, and suggested training/test splits. Examples of RGB images, raw depth images and labeled images in the dataset are illustrated in Fig. 13. Besides the raw depth images, this dataset also provides some pre-processed images on which the black areas with missed depth values have been filled (see Fig. 14 for an example). The sampling rate of the Kinect camera is varying from 20 frames per second to 30 frames per second. As a result, there are 108,617 RGB-D images captured from 64 different indoor scenes, such as bedroom, bathroom and kitchen. Every 2 to 3 seconds, frames extracted from the obtained video are processed with dense multi-class labeling. This special subset contains 2347 unique labeled frames.

NYU Dataset V2 [65] is an extension of NYU Dataset V1 and was founded in 2012. This new dataset includes approximately 408,000 RGB images and 1449 aligned RGB-D images

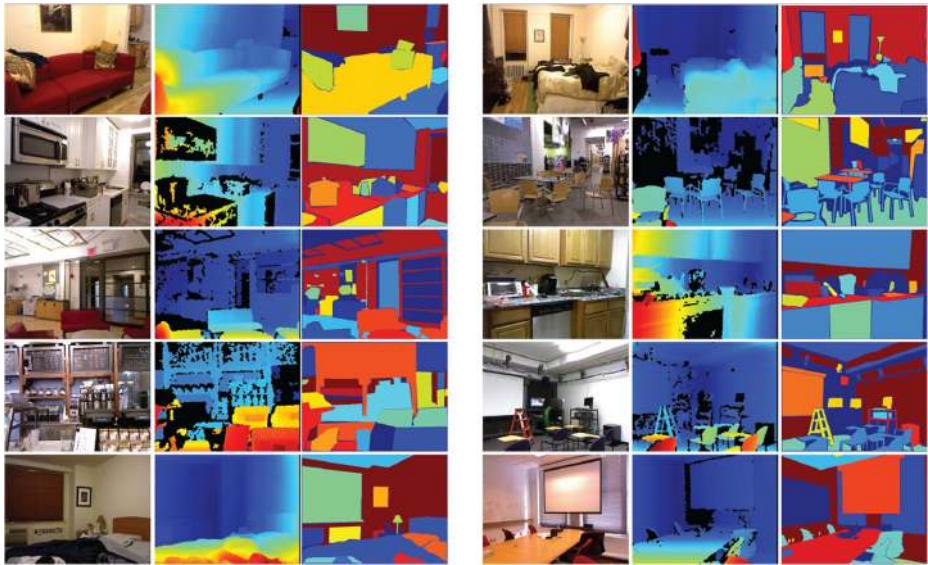


Fig. 13 Selected examples of RGB images, raw depth images and class labeled images in NYU dataset¹¹

with detailed annotations from 464 indoor scenes across 26 scene classes. Obviously, the scale of this dataset is even larger and it is more diversified than NYU dataset V1. The RGB-D images are collected from numerous buildings in three US cities. Meanwhile, this dataset includes 894 different classes about 35,064 objects. Particularly, to identify multiple instances of an object class in one scene, each instance in this scene is given a unique label. The representative research work using these two datasets as the benchmark for indoor segmentation and classification can be found in [77, 94, 95].

3.3.3 B3DO: berkeley 3-D object dataset

B3DO (Berkeley 3-D object dataset) was publicized in 2011 by University of California-Berkeley to accelerate progress in the field of evaluating approaches of indoor scene object recognition and localization [41]. The organization for this dataset is different in the sense that the data collection effort is continuously crowdsourced by many members in the research community and AMT (Amazon Mechanical Turk), instead of collecting all the data by one single host. By doing so, the dataset will have a variety of appearances over scenes and objects.

The first version of this dataset annotates 849 images from 75 scenes of more than 50 classes (i.e., table, cup, keyboard, trash can, plate, towel), which have been processed for alignment and inpainting in both real office and domestic environments. Compared to other Kinect object datasets, the images of B3DO are taken in “the wild” [35, 91] places by an automatic turntable setting. During the capturing, camera viewpoint and the lighting condition are changed. The ground truth is represented by the bounding box labeling at a class level on both RGB images and depth images.

¹¹http://cs.nyu.edu/~silberman/datasets/nyu_depth.v1.html.

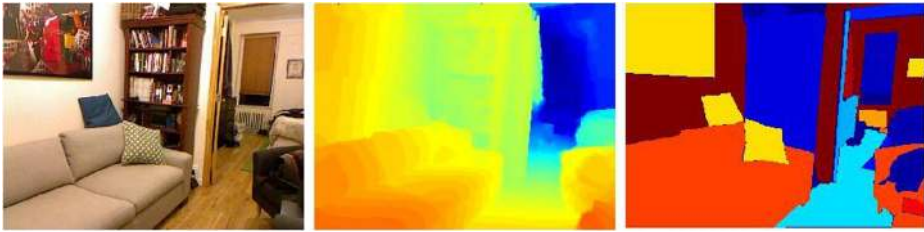


Fig. 14 Output of the RGB camera (*left*), pre-processed depth image (mid) and class labeled image (*right*) from NYU Depth V1 and V2 dataset¹²

3.3.4 RGB-D person re-identification dataset

RGB-D Person Re-identification dataset [5] was jointly created in 2012 by Italian Institute of Technology, University of Verona and University of Burgundy, aiming to promote the research of the RGB-D person re-identification. This dataset dedicates to simulating the difficult situations, e.g., changing the participant’s clothing during the observation. Compared to the existed datasets for appearance-based person re-identification, this dataset has a wider range of applications and consists of four different parts of data. The challenging level goes up gradually from the first part to the fourth part. The first part (“collaborative”) of data has been gained through recording 79 people with four kinds of conditions: walking slowly, a frontal view, with stretched arms and without occlusions. All of these are shot while passersby are more than two meters away from the Kinect sensor. The second and third groups (“walking 1” and “walking 2”) are made when the same 79 people are walking into the lab with different poses. The last part (“backwards”) actually records the departure view of the people. For increasing the challenge, all the sequences in this dataset are recorded during multiple days, which means that the clothing and accessories on the passersby may be varying.

Furthermore, four synchronized labeling information are annotated for each person: the foreground masks, the skeletons, 3D meshes and an estimate of the floor. Additionally, using the method “Greedy Projection”, a mesh can be generated from the person’s point cloud in this dataset. Figure 15 describes the computed meshes from the four kinds of groups. One published work based on this dataset is [76].

3.3.5 Eurecom kinect face dataset (Kinect FaceDB)

Kinect FaceDB [64] is a collection of different facial expressions in varying lighting conditions and occlusion levels based on a Kinect sensor, which was jointly developed by University of North Carolina and the department of Multimedia Communications in EURECOM in 2014. This dataset provides different forms of data including 936 processed and well-aligned 2D RGB images, 2.5D depth images, shots of 3D point cloud face data and 104 RGB-D video sequences. During the data collection procedure, totally 52 people (38 males and 14 females) are invited in the project. Aiming to gain multiple facial variations, these participants are selected from different age groups (27 to 40) with different nationalities and six kinds of ethnicity (21 Caucasian, 11 Middle East, 10 East Asian, 4 Indian, 3 African-American and 3 Hispanic). The data is obtained from two sessions with a time interval

¹²http://cs.nyu.edu/~silberman/datasets/nyu_depth.v1.html.

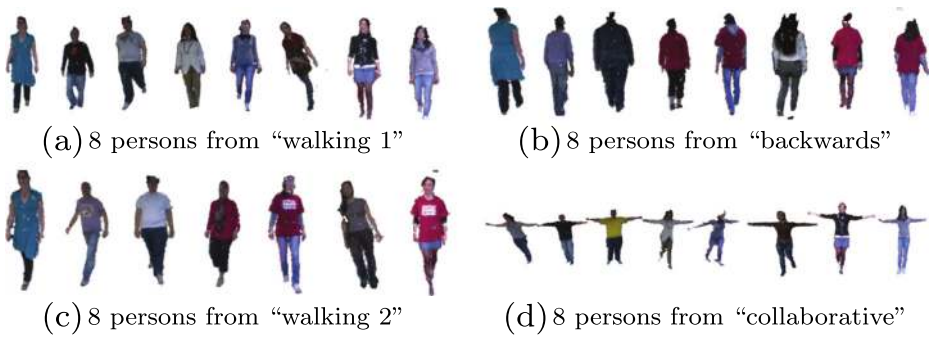


Fig. 15 Illustration of the different groups about the recorded data from RGB-D Person Re-identification dataset¹³

about half a month. Each person is asked to perform 9 kinds of different facial expressions in various lighting and occlusion situations. The facial expressions contain neutral, smiling, open mouth, left profile, right profile, occluding eyes, occluding mouth, occluded by paper and strong illumination. All the image are acquired at a distance about $1m$ away from the sensor in the lab at EURECOM Institute. All the participants follow the protocol that turns hand around slowly: the horizontal direction $0^\circ \rightarrow +90^\circ \rightarrow -90^\circ \rightarrow 0^\circ$ and the vertical direction $0^\circ \rightarrow +45^\circ \rightarrow -45^\circ \rightarrow 0^\circ$.

During the recording process, the Kinect sensor is fixed on top of a laptop. For the purpose of providing a simple background, a white board is placed on the opposite side of the Kinect sensor at the distance of $1.25m$. Furthermore, this dataset is manually annotated, providing 6 facial anchor points: left eye center, right eye center, left mouth corner, right mouth corner, nose-tip and the chin. Meanwhile, information about gender, birth, glasses-wearing and shooting time are also associated. Sample images highlighting the facial variations of this dataset are shown in Fig. 16.

3.3.6 Big BIRD (Berkeley instance recognition dataset)

Big BIRD dataset was designed in 2014 by Department of Electrical Engineering and Computer Science of University of California Berkeley, aiming to accelerate the developments in graphics, computer vision and robotic perception, particularly 3D mesh reconstruction and object recognition areas. It was first quoted in [80]. Compared to the previous 3D vision datasets, it tries to overcome the shortcomings, such as few objects, low-quality objects, low-resolution RGB data. Moreover, it also provides calibration and the pose information, enabling better alignments of multi-view objects and scenes.

Big BIRD dataset consists of 125 objects (keep growing), in which 600 12 megapixel images and 600 RGB-D point clouds spinning all views are provided for each object. Meanwhile, accurate calibration parameters and pose information are also available for each image. In the data collection system, the object is placed in the center of a controllable turntable based platform on which multiple Kinects and high-resolution DSLR cameras from 5 polar angles and 120 azimuthal angles are mounted. The collection procedure for one object takes roughly 5 mins. In this procedure, four adjustable lights are put in

¹³<http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html>.

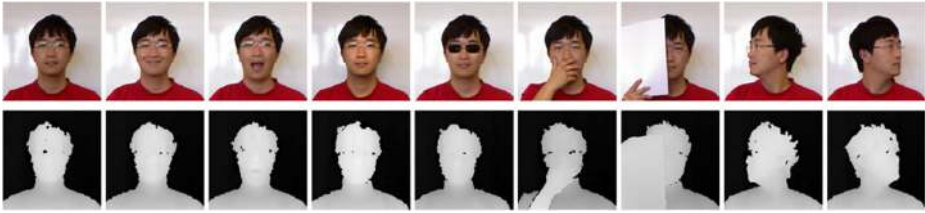


Fig. 16 Sampled RGB (*top row*) and depth images (*bottom row*) from Kinect FaceDB [64]. *Left to right*: neutral face with normal illumination, smiling, mouth open, strong illumination, occlusion by sunglasses, occlusion by hand, occlusion by paper, turn face right and turn face left

different places, illuminating the recording environment. Furthermore, in order to acquire calibrated data, one chessboard is placed on the turntable, while the system ensures that at least one camera can shoot the whole vision. The data collection equipment as well as the environment are shown in Fig. 17. More details about Big BIRD can be found in [80].

3.3.7 High resolution range based face dataset (HRRFaceD)

The image processing group in Polytechnic University of Madrid created high resolution range based face dataset (HRRFaceD) [62] in 2014, intending to evaluate the recognition of different faces from a wide range of poses. This dataset was recorded by the second generation of Microsoft Kinect sensor. It consists of 22 sequences from 18 different subjects (15 males, 3 females and 4 people from them are with glasses) with various poses (frontal, lateral, etc.). During the collection procedure, each person is sitting about 50 cm away from Kinect, while the head is at the same height as the sensor. In order to obtain more information from the nose, eyes, mouth and ears, all persons continuously turn their heads. Depth images (512×424 pixels) are saved with 16 bits format. One recent published paper about HRRFaceD can be found in [62]. Sample images from this dataset are shown in Fig. 18.

Among these object and scene recognition datasets mentioned above, RGB-D Object dataset and NYU Depth V1 and V2 have the largest number of references (> 100). The challenge of RGB-D Object dataset is that it contains both textured objects and texture-less objects. Meanwhile, the lighting conditions have large variations over the data frames. The category recognition accuracy (%) can reach 90.3 (RGB), 85.9 (Depth) and 92.3 (RGB-D) in [43].

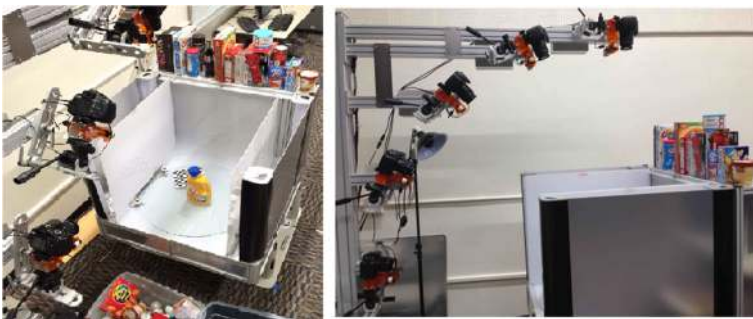


Fig. 17 The *left* is the data-collection system of Big BIRD dataset. The chessboard aside the object is used for merging clouds when turntable rotates. The *right* is the side view of this system [80]

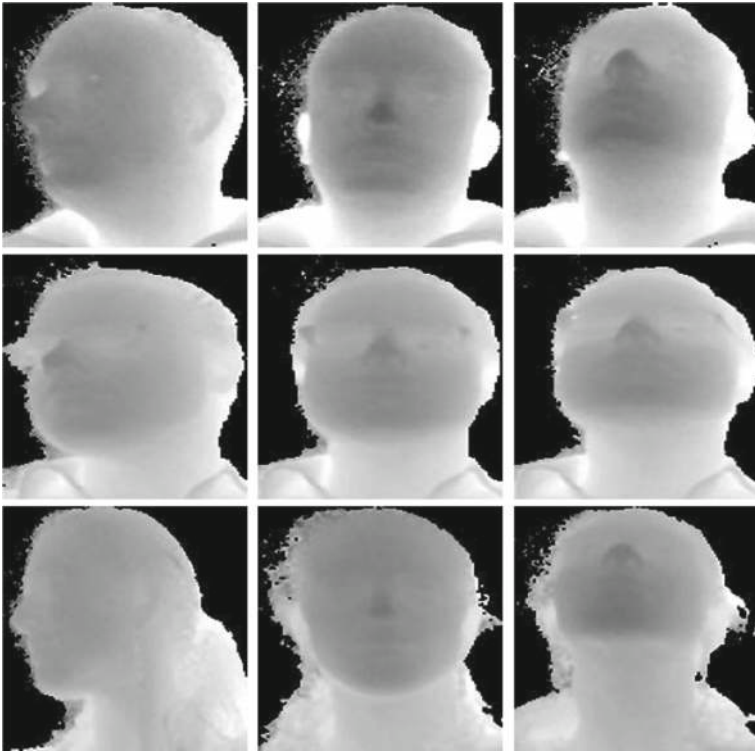


Fig. 18 Sample images from HRRFaceD [62]. There are three images in a row displaying a person with different poses. *Top row*: a man without glasses. *Middle row*: a man with glasses. *Bottom row*: a woman without glasses

The instance recognition accuracy (%) can reach 92.43 (RGB), 55.69 (Depth) and 93.23 (RGB-D) in [42]. In general, NYU Depth V1 and V2 dataset is very difficult for scene classification since it contains various objects in one category. Therefore, the scene recognition rates are relatively low, which are 75.9 ± 2.9 (RGB), 65.8 ± 2.7 (Depth) and 76.2 ± 3.2 (RGB-D) reported in [42]. The latest algorithm performance comparisons based on B3DO, Person Re-identification, Kinect FaceDB, Big BIRD and HRRFaceD can be found in [32, 40, 64, 66] and [62] respectively.

3.4 Simultaneous localization and mapping (SLAM)

The emergence of new RGB-D camera, like Kinect, boosts the research for SLAM due to its capability of providing depth information directly. Over the last a few years, many excellent works have been published. In order to test and compare those algorithms, several datasets and benchmarks have been created.

3.4.1 TUM benchmark dataset

TUM Benchmark dataset [85] was founded by University of Technology Munich in July 2011. The intention is to build a novel benchmark for evaluating visual odometry and visual

SLAM (Simultaneous Localization and Mapping) systems. It is noted that this is the first RGB-D dataset for visual SLAM benchmarking. It provides RGB and depth images (640×480 at $30Hz$) along with the time-synchronized ground truth trajectory of camera poses generated by a motion-capture system. TUM Benchmark dataset consists of 39 sequences which are captured in two different indoor environments: an office scene ($6 \times 6m^2$) and an industrial hall ($10 \times 12m^2$). Meanwhile, the IMU accelerometer data is provided from the Kinect.

This dataset is recorded by moving the handheld Kinect sensor with unconstrained 6-DOF motions along different trajectories in the environments. It contains totally 50 GB Kinect data and 9 sequences. For having more variations in the dataset, the angular velocities (fast/slow), conditions of the environment (one desk, several desks and whole room) and illumination conditions (weak and strong) keep changing during the recording process. Example frames from this dataset are depicted in Fig. 19. The latest version of this dataset is extended to include dynamic sequences, longer trajectories and sequences captured by a mounted Kinect on a wheeled robot. The sequences are labeled with 6-DOF ground truth from a motion capture system having 10 cameras. Six research publications about evaluating ego-motion estimation and SLAM over TUM Benchmark dataset are [21, 38, 45, 84, 86, 87].

3.4.2 ICL-NUIM dataset

ICL-NUIM dataset [34] is a benchmark for experimenting the algorithms devoted to RGB-D visual odometry, 3D reconstruction and SLAM. It is founded in 2014 by the researchers from Imperial College London and National University of Ireland Maynooth. Unlike the previous presented datasets that only focus on pure two-view disparity estimation or trajectory estimation (i.e., Sintel, KITTI, TUM RGB-D), ICL-NUIM dataset combines realistic RGB and depth information together with a full 3D geometry scene and the trajectory ground truth. The camera view field is 90 degrees and the image resolution is with 640×480 pixels. This dataset collects image data from two different environments: the living room and the office room. The four RGB-D videos from the office room environment contain trajectory data but do not have any explicit 3D models. Therefore, it can only be used for benchmarking camera trajectory estimation. However, the four synthetic RGB-D video sequences from the living room scene have camera pose information associated with a 3D polygonal model (ground truth). Thus, they can be used to benchmark both camera



Fig. 19 Sample images from TUM Benchmark dataset [21]



Fig. 20 Sample images of the office room scene taken at different camera poses from ICL-NUIM dataset [34]

trajectory estimation and 3D reconstruction. In order to mimic the real-world environment, artifacts such as specular reflections, light scattering, sunlight, color bleeding and shadows are added into the images. More details about ICL-NUIM dataset can be found in [34]. Sample images of the living room and the office room scene taken at different camera poses can be found in Figs. 20 and 21.

If we compare TUM Benchmark dataset with ICL-NUIM dataset, it becomes clear that the former is more popular, because it has much more citations. It may be partially due to the fact that the former one is earlier than the later one. Apart from it, this dataset is more challenging and realistic since it covers large areas of office space and the camera motions are not restricted. The related performance comparisons between TUM Benchmark dataset and ICL-NUIM dataset are shown in [22] and [75].

3.5 Hand gesture analysis

In recent years, the research of hand gesture analysis from RGB-D sensors develops quickly, because it can facilitate a wide range of applications in human computer interaction, human robot interaction and pattern analysis. Compared to human activity analysis, hand gesture analysis does not need to deal with the dynamics from other body parts but only focuses on the hand region. On the one hand, the focus on the hand region only helps to increase the

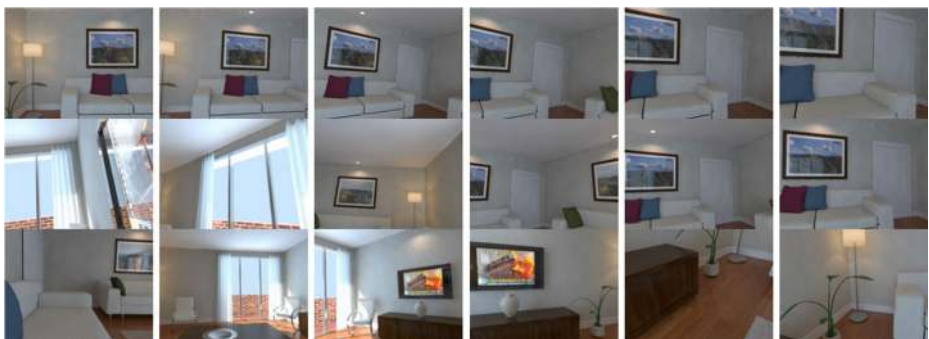


Fig. 21 Sample images of the living room scene taken at different camera poses from ICL-NUIM dataset [34]

analysis accuracy. On the other hand, it also alleviates the complexity of the system, thus enabling real-time applications. Basically, a hand gesture analysis system involves three components: hand detection and tracking, hand pose estimation and gesture classification. In the past years, the research is restrained due to the fact that it is so hard to solve the problems, like occlusions, different illumination conditions and skin color. However, the research in this field is triggered again after the invention of RGB-D sensor, because this new image representation is resistant to the variations mentioned above. In just a few years, we have found several RGB-D gesture dataset available.

3.5.1 Microsoft research Cambridge-12 kinect gesture dataset

Microsoft Research Cambridge created MSRC-12 Gesture dataset [26] in 2012, which includes relevant gestures and their corresponding semantic labels for evaluating gesture recognition and detection systems. This dataset consists of 594 sequences of human skeletal body part gestures, which are totally 719,359 frames with a duration over 6 hours and 40 min at a sample rate of 30Hz. During the collection procedure, there are 30 participants (18 males and 12 females) performing two kinds of gestures. One is called as iconic gestures, e.g., crouching or hiding (500 instances), putting on night vision goggles (508 instances), shooting a pistol (511 instances), throwing an object (515 instances), changing weapons (498 instances) and kicking (502 instances). The other one is referred to metaphoric gestures such as starting system/music/raising volume (508 instances), navigating to next menu/moving arm right (522 instances), winding up the music (649 instances), taking a bow to end music session (507 instances), protesting the music (508 instances) and moving up the tempo of the song/beat both arms (516 instances). All the sequences are recorded in front of a white and simple background so that all body movements are within the frame. Each video sequence is labeled with gesture performance and motion tracking of human body joints. An application oriented case study about this dataset can be found in [23].

3.5.2 Sheffield kinect gesture dataset (SKIG)

SKIG is a hand gesture dataset which was supported by the University of Sheffield since 2013. It is first introduced in [57] and applied to learn discriminative representations. This dataset includes totally 2016 hand-gesture video sequences from six people, 1080 RGB sequences and 1080 depth sequences, respectively. In this dataset, there are 10 categories of gestures: triangle (anti-clockwise), circle (clockwise), right and left, up and down, wave, hand signal “Z”, come here, cross, pat and turn around. All these sequences are extracted through a Kinect sensor and the other two synchronized cameras. In order to increase the variety of recorded sequences, subjects are asked to perform three kinds of hand postures: fist, flat and index. Furthermore, three different backgrounds (i.e., wooden board, paper with text and white plain paper) and two illumination conditions (light and dark) are used in SKIG. Therefore, there are 360 different gesture sequences accompanied by hand movement annotation for each subject. Figure 22 shows some frames in this dataset.

3.5.3 50 salads dataset

School of Computing in University of Dundee created 50 Salads dataset [88] that involves manipulative objects in 2013. The intention of this well-designed dataset is to stimulate the research on wide range of recognition gesture problems including the applications of automated supervision, sensor fusion, and user-adaptation. This dataset records 25 people,

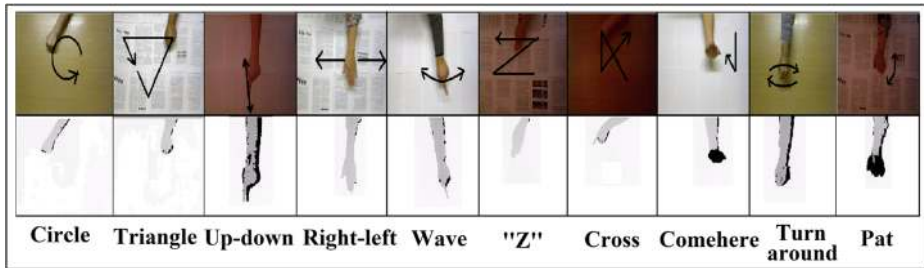


Fig. 22 Sample frames from Sheffield Kinect gesture dataset and the descriptions of 10 different categories [57]

each cooking 2 mixed salads. The RGB-D sequence length is over 4 hours and with the resolution of 640×480 pixels. Additionally, 3-axis accelerometer data are attached to cooking utensils (mixing spoon, knife, small spoon, glass, peeler, pepper dispenser and oil bottle) simultaneously.

The collection process of this dataset can be described as follows. The Kinect sensor is mounted on the wall in order to cover the whole view of the cooking place. 27 persons from different age groups and different cooking levels were making a mixed salad twice, thus resulting in totally 54 sequences. These activities for preparing the mixed salad were annotated continuously, which include adding oil (55 instances), adding vinegar (54 instances), adding salt (53 instances), adding pepper (55 instances), mix dressing (61 instances), peel cucumber (53 instances), cutting cucumber (59 instances), cutting cheese (55 instances), cutting lettuce (61 instances), cutting tomato (63 instances), putting cucumber into bowl (59 instances), putting cheese into bowl (53 instances), putting lettuce into bowl (61 instances), putting tomato into bowl (62 instances), mixing ingredients (64 instances), serving salad onto plate (53 instances) and adding dressing (44 instances). Meanwhile, each activity is split into pre-phase, core-phase and post-phase which were annotated respectively. As a result, there are 518,411 video frames and 966 activity instances that are annotated in 50 Salads dataset. Figure 23 shows example snapshots from this dataset. It is worth noting that task orderings given to the participants are randomly sampled from a statistical recipe model. Some published papers using this dataset can be found in [89, 102].

Among above three RGB-D datasets, the most popular dataset is MSRC-12 Gesture dataset which has nearly 100 citations. Since the RGB-D videos from MSRC-12 Gesture dataset not only contain the gesture information but also the whole person information, it is still a challenging dataset for classification problem. The state-of-the-art classification rate about this dataset is from [100] (72.43 %). Therefore, more research efforts are needed in order to achieve a better result on this dataset. Compared to MSRC-12 Gesture dataset, the challenge of SKIG and 50 Salads dataset is simpler. Because the RGB-D sensors only shoot the gestures of the participants, these two datasets only include the information of gestures. The latest classification performance of SKIG is 95.00 % [19]. The state-of-the-art result of 50 Salads dataset is mean precision (0.62 ± 0.05) and mean recall (0.64 ± 0.04) [88].

3.6 Discussions

In this section, the comparison of RGB-D datasets is conducted from several aspects. For easy access, all the datasets are ordered alphabetically in three tables (from 4 to 6). If the dataset name starts with a digital number, it is ranked numerically following all the datasets which starts with English letters. For more comprehensive comparisons, besides these 20

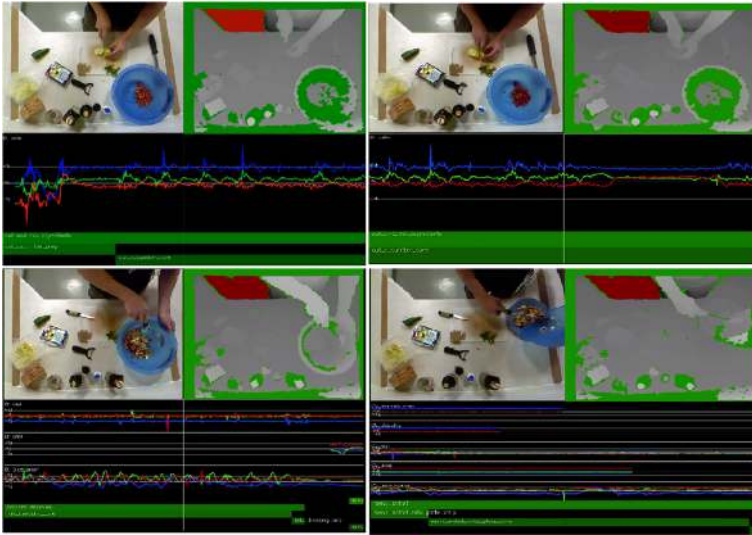


Fig. 23 Example snapshots from 50 Salads dataset¹⁴, from top left to bottom right is the chronological order from the video. The curves under the images are the accelerometer data at 50 Hz of devices attached to the knife, the mixing spoon, the small spoon, the peeler, the glass, the oil bottle, and the pepper dispenser

mentioned datasets above, another 26 extra RGB-D datasets for different applications are also added into the tables: Birmingham University Objects, Category Modeling RGB-D [104], Cornell Activity [47, 92], Cornell RGB-D [48], DGait [12], Daily Activities with occlusions [1], Heidelberg University Scenes [63], Microsoft 7-scenes [78], MobileRGBD [96], MPII Multi-Kinect [93], MSR Action3D Dataset [97], MSR 3D Online Action [103], MSRGesture3D [50], DAFT [31], Paper Kinect [70], RGBD-HuDaAct [68], Stanford Scene Object [44], Stanford 3D Scene [105], Sun3D [101], SUN RGB-D [82], TST Fall Detection [28], UTD-MHAD [14], Vienna University Technology Object [2], Willow Garage [99], Workout SU-10 exercise [67] and 3D-Mask [24]. In addition, we name those datasets without original names by means of creation place or applications. For example, we name the dataset in [63] as Heidelberg University Scenes.

Let us now explain these tables. The first and second columns in the tables are always the serial number and the name of the dataset. Table 4 shows some features including the authors of the datasets, the year of the creation, the published papers describing the dataset, the related devices, data size and number of references related to datasets. The author (the third column) and the year (the fourth column) are collected directly in the datasets or are found in the oldest publication related to the dataset. The cited references in the fifth column contain the publications which elaborate the corresponding dataset. Data size (the seventh column) refers to the size of all information, such as the RGB and depth information, camera trajectory, ground truth and accelerometer data. For a scientific evaluation about these datasets, the comparison of number of citation is added into Table 4. A part of these statistical numbers are derived from the number of papers which use related dataset as benchmark. The rest is from the papers which do not directly use these datasets but mention these datasets in their published papers. It is noted that the numbers are roughly estimated. It can

¹⁴<http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/>.

Table 4 The characteristics of the selected 46 RGB-D datasets

No.	Name	Author	Year	Description	Device	Datasez	Number of citation
1	Big BIRD	Arjun Singh et al.	2014	[80]	Kinect v1 and DSLR	≈ 74G	Unknown
2	Birmingham University objects	Krzysztof Walas et al.	2014	No	Kinect v2	Unknown	Unknown
3	Biwi head pose	Fanelli et al.	2013	[25]	Kinect v1	5.6G	88
4	B3DO	Allison Janoch et al.	2011	[41]	Kinect v1	793M	96
5	Category modeling RGB-D	Quanshi Zhang et al.	2013	[104]	Kinect v1	1.37G	4
6	Cornell activity	Jaeyong Sung et al.	2011	[92]	Kinect v1	44G	> 100
7	Cornell RGB-D	Abhishek Anand et al.	2011	[48]	Kinect v1	≈ 7.6G	60
8	DAFT	David Gossow et al.	2012	[31]	Kinect v1	207M	2
9	Daily activities with occlusions	Abdallah DIB et al.	2015	[1]	Kinect v1	6.2G	0
10	DGait	Ricard Borrs et al.	2012	[12]	Kinect v1	9.2G	7
11	Heidelberg University scenes	Stephan Meister et al.	2012	[63]	Kinect v1	3.3G	24
12	HRRFaceD	Tomás Mantecón et al.	2014	[62]	Kinect v2	192M	Unknown
13	ICL-NUIM	A. Handa et al.	2014	[34]	Kinect v1	18.5G	3
14	Kinect FaceDB	Rui Min et al.	2012	[64]	Kinect v1	Unknown	1
15	Microsoft 7-scenes	Antonio Criminisi et al.	2013	[78]	Kinect v1	20.9G	10
16	MobileRGBD	Dominique Vaufreydaz et al.	2014	[96]	Kinect v2	Unknown	Unknown
17	MPII multi-kinect	Wandi Susanto et al.	2012	[93]	Kinect v1	15G	11
18	MSRC-12 gesture	Simon Fothergill et al.	2012	[26]	Kinect v1	165M	83
19	MSR Action3D Dataset	Jiang Wang et al.	2012	[97]	Similar to Kinect	56.4M	> 100

Table 4 (continued)

No.	Name	Author	Year	Description	Device	Dataseize	Number of citation
20	MSRDailyActivity3D	Zicheng Liu et al.	2012	[97]	Kinect v1	3.7M	> 100
21	MSR 3D online action	Gang Yu et al.	2014	[103]	Kinect v1	5.5G	9
22	MSRGesture3D	Alexey Kurakin et al.	2012	[50]	Kinect v1	28M	94
23	NYU Depth V1 and V2	Nathan Silberman et al.	2011	[79]	Kinect v1	520G	> 100
24	Object RGB-D	Kevin Lai et al.	2011	[52]	Kinect v1	84G	> 100
25	Object discovery	Julian Mason et al.	2012	[60]	Kinect v1	7.8G	8
26	Object segmentation	A. Richtsfeld et al.	2012	[72]	Kinect v1	302M	28
27	Paper kinect	F. Pomerleau et al.	2011	[70]	Kinect v1	2.6G	32
28	People	L. Spinello et al.	2011	[83]	Kinect v1	2.6G	> 100
29	Person re-identification	B. I. Barbosa, M et al.	2012	[5]	Kinect v1	Unknown	37
30	PTB	Shuran Song et al.	2013	[81]	Kinect v1	10.7G	12
31	RGBD-HuDaAct	Bingbing Ni et al.	2011	[68]	Kinect v1	Unknown	> 100
32	SKIG	L. Liu et al.	2013	[57]	Kinect v1	1G	35
33	Stanford scene object	Andrej Karpathy et al.	2014	[44]	Xtion Pro live	178.4M	29
34	Stanford 3D scene	Qian-Yi Zhou et al.	2013	[105]	Xtion Pro live	≈ 33G	15
35	Sun3D	Jianxiong Xiao et al.	2013	[101]	Xtion Pro live	Unknown	16
36	SUN RGB-D	S. Song et al.	2015	[82]	Kinect v1, Kinect v2, etc.	6.4G	8
37	TST fall detection	S. Gasparrini et al.	2015	[28]	Kinect v2	12.1G	25
38	TUM	J. Sturm et al.	2012	[85]	Kinect v1	50G	> 100
39	TUM texture-less	S Hinterstoisser et al.	2012	[36]	Kinect v1	3.61G	26
40	UR fall detection	Michal Kepski et al.	2014	[46]	Kinect v1	≈ 5.75G	2

Table 4 (continued)

No.	Name	Author	Year	Description	Device	Dataseize	Number of citation
41	UTD-MHAD	Chen Chen et al.	2015	[14]	Kinect v1 and Kinect v2	≈ 1.1G	3
42	Vienna University Technology object	Aitor Aldoma et al.	2012	[2]	Kinect v1	81.4M	19
43	Willow garage	Aitor Aldoma et al.	2011	[99]	Kinect v1	656M	Unknown
44	Workout SU-10 exercise	F Negin et al.	2013	[67]	Kinect v1	142G	13
45	3D-Mask	N Erdognus et al.	2013	[24]	Kinect v1	Unknown	18
46	50 salads	Sebastian Stein et al.	2013	[88]	Kinect v1	Unknown	4

Table 5 The characteristics of the selected 46 RGB-D datasets

No.	Name	Intended applications	Label information	Data modalities	Number of categories
1	Big BIRD	Object and scene recognition	Masks, ground truth poses, registered mesh	Color, depth	125 objects
2	Birmingham University objects	Object detection and tracking	The model into the scene	Color, depth	10 to 30 objects
3	Biwi head pose	Human activity analysis	3D position and rotation	Color, depth	20 objects
4	B3DO	Object and scene recognition	Bounding box labeling at a class level	Color, depth	50 objects and 75 scenes
5	Category modeling RGB-D	Object and scene recognition	Edge segments	Color, depth	900 objects and 264 scenes
6	Cornell activity	Human activity analysis	Skeleton, joint position and orientation on each frame	Color, depth, skeleton	120+ activities
7	Cornell RGB-D	Object and scene recognition	Per-point object-level labeling	Color, depth, accelerometer	24 office scenes and 28 home scenes
8	DAFT	SLAM	Camera motion type, 2D homographies	Color, depth	Unknown
9	Daily activities with occlusions	Human activity analysis	Position markers of the 3D joint	Color, depth, skeleton	Unknown
10	DGait	Human activity analysis	Location from a MoCap system Subject, gender, age and an entire walk cycle	Color, depth	11 activities
11	Heidelberg University scenes	SLAM	Frame-to-frame transformations and LiDAR ground truth	Color, depth	57 scenes
12	HRRFaceD	Object and scene recognition	No	Color, depth	22 subjects

Table 5 (continued)

No.	Name	Intended applications	Label information	Data modalities	Number of categories
13	ICL-NUIM	SLAM	Camera trajectories for each video. Geometry of the scene	Color, depth	2 scenes
14	Kinect FaceDB	Object and scene recognition	The position of six facial landmarks	Color, depth	52 objects
15	Microsoft 7-scenes	SLAM	6DOF ground truth	Color, depth	7 scenes
16	MobileRGBD	SLAM	speed and trajectory	Color, depth	1 scene
17	MPI Multi-Kinect	Object detection and tracking	Bounding box and polygons	Color, depth	10 objects and 33 scenes
18	MSRC-12 gesture	Hand gesture analysis	Gesture, motion tracking of human joint locations	Color, depth, skeleton	12 gestures
19	MSR Action3D dataset	Human activity analysis	Activity being performed and 20 joint locations of skeleton positions	Color, depth, skeleton	20 actions
20	MSRDailyActivity3D	Human activity analysis	Activity being performed and 20 joint locations of skeleton positions	Color, depth, skeleton	16 activities
21	MSR 3D online action	Human activity analysis	Activity in each video	Color, depth, skeleton	7 activities
22	MSRGesture3D	Hand gesture analysis	Gesture in each video	Color, depth	12 activities
23	NYU depth V1 and V2	Object and scene recognition	Dense multi-class labeling	Color, depth, accelerometer	528 scenes
24	Object RGB-D	Object detection and tracking	Auto-generated masks	Color, depth	300 objects and scenes
25	Object discovery	Object detection and tracking	Ground truth object segmentations	Color, depth	7 objects
26	Object Segmentation	Object detection and tracking	Per-pixel segmentation	Color, depth	6 categories
27	Paper Kinect	SLAM	6DOF ground truth	Color, depth	3 scenes

Table 5 (continued)

No.	Name	Intended applications	Label information	Data modalities	Number of categories
28	People	Object detection and tracking	Bounding box annotations and a 'visibility' measure	Color, depth	Multiple people
29	Person re-identification	Object and scene recognition	Foreground masks, skeletons, 3D meshes and an estimate of the floor	Color, depth	79 people
30	PTB	Object detection and tracking	Bounding box covering target object	Color, depth	3 types and 6 scenes
31	RGBD-HuDaAct	Human activity analysis	Activities being performed in each sequence	Color, depth	12 activities
32	SKIG	Hand gesture analysis	The gesture is performed	Color, depth	10 gestures
33	Stanford Scene Object	Object detection and tracking	Ground truth binary labeling	Color, depth	58 scenes
34	Stanford 3D Scene	SLAM	Estimated camera pose	Color, depth	6 scenes
35	Sun3D	Object detection and tracking	Polygons of semantic class and instance labels	Color, depth	254 scenes
36	SUN RGB-D	Object and scene recognition	Dense semantic	Color, depth	19 scenes
37	TST fall detection	Object detection and tracking Human activity analysis	Activity performed, acceleration data and skeleton joint locations	Color, depth, skeleton, accelerometer	2 categories
38	TUM	SLAM	6DOF ground truth	Color, depth, accelerometer	2 scenes

Table 5 (continued)

No.	Name	Intended applications	Label information	Data modalities	Number of categories
39	TUM texture-less	Object detection and tracking	6DOF pose	Color, depth	15 objects
40	UR Fall Detection	Human activity analysis	Accelerometer data	Color, depth, accelerometer	66 falls
41	UTD-MHAD	Human activity analysis	Accelerometer data with each video	Color, depth, skeleton, accelerometer	27 actions
42	Vienna University Technology object	Object and scene recognition	6DOF GT of each object	Color, depth	35 objects
43	Willow garage	Object detection and tracking	6DOF pose, per-pixel labelling	Color, depth	6 categories
44	Workout SU-10 exercise	Human activity analysis	Motion Files	Color, depth, skeleton	10 activities
45	3D-mask	Object and scene recognition	Manually labeled eye positions	Color, depth	17 people
46	50 salads	Hand gesture analysis	Accelerometer data and labeling of steps in the recipes	Color, depth, accelerometer	27 people

Table 6 The characteristics of the selected 46 RGB-D datasets

No.	Name	Camera movement	Multi-Sensors	Conditions required	Link
1	Big BIRD	Yes	Yes	Yes	http://rll.berkeley.edu/bigbird/
2	Birmingham University objects	No	No	Yes	http://www.cs.bham.ac.uk/~walask/SHREC2015/
3	Biwi head pose	No	No	No	https://data.vision.ee.ethz.ch/cvl/gfaneli/head_pose/head_forest.html#
4	B3DO	No	No	No	http://kinectdata.com/
5	Category modeling RGB-D	No	No	No	http://sdrv.ms/Z4px7u
6	Cornell activity	No	No	No	http://pr.cs.cornell.edu/humanactivities/data.php
7	Cornell RGB-D	Yes	No	No	http://pr.cs.cornell.edu/sceneunderstanding/data/data.php
8	DAFT	Yes	No	No	http://ias.cs.tum.edu/people/gossow/rgb-d
9	Daily activities with occlusions	No	No	NO	https://team.inria.fr/larsen/software/datasets/
10	DGait	No	No	No	http://www.cvc.uab.es/DGaitDB/Download.html
11	Heidelberg University scenes	No	No	Yes	http://hci.iwr.uni-heidelberg.de/Benchmarks/document/kinectFusionCapture/
12	HRRFaceD	No	No	No	https://sites.google.com/site/hrrfaced/
13	ICL-NUIM	Yes	No	No	http://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html
14	Kinect FaceDB	No	No	Yes	http://rgb-d.eurecom.fr/
15	Microsoft 7-scenes	Yes	No	Yes	http://research.microsoft.com/en-us/projects/7-scenes/
16	MobileRGBD	Yes	No	Yes	http://mobilergb-d.inrialpes.fr/#RobotView
17	MPII multi-kinect	No	Yes	No	https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/object-recognition-and-scene-understanding/mpii-multi-kinect-dataset/

Table 6 (continued)

No.	Name	Camera movement	Multi-Sensors	Conditions required	Link
18	MSRC-12 Gesture	No	No	No	http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/
19	MSR Action3D dataset	No	No	No	http://research.microsoft.com/en-us/um/people/zliu/actionreconrc/
20	MSRDailyActivity3D	No	No	No	http://research.microsoft.com/en-us/um/people/zliu/actionreconrc/
21	MSR 3D online action	No	No	No	http://research.microsoft.com/en-us/um/people/zliu/actionreconrc/
22	MSRGesture3D	No	No	No	http://research.microsoft.com/en-us/um/people/zliu/actionreconrc/
23	NYU Depth V1 and V2	Yes	No	No	http://cs.nyu.edu/~silberman/datasets/nyu_depth_v1.html
24	Object RGB-D	No	No	No	http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
25	Object discovery	Yes	No	No	http://rgbd-dataset.cs.washington.edu/
26	Object segmentation	No	No	No	http://wiki.ros.org/Papers/IROS2012_Mason_Martha_Parr
27	Paper kinect	Yes	No	No	http://www.acin.tuwien.ac.at/?id=289
28	People	No	Yes	No	http://projects.asi.ethz.ch/datasets/doku.php?id=Kinect:iros2011Kinect
29	Person re-identification	No	No	Yes	http://www2.informatik.uni-freiburg.de/~spinello/RGBD-dataset.html
30	PTB	Yes	No	No	http://www.iiit.it/en/datasets-and-code/datasets/rgbdid.html
31	RGBD-HuDaAct	No	No	Yes	http://tracking.cs.princeton.edu/dataset.html
					http://adsc.illinois.edu/sites/default/files/files/ADSC-RGBD-dataset-download-instructions.pdf
32	SKIG	No	No	No	http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm
33	Stanford scene object	NO	No	No	http://cs.stanford.edu/people/karpathy/discovery/
34	Stanford 3D Scene	Yes	No	No	https://drive.google.com/folderview?id=0B6qjzcYeERgaW5zRWZc2FuRDg&usp=sharing

Table 6 (continued)

No.	Name	Camera movement	Multi-Sensors	Conditions required	Link
35	Sun3D	Yes	No	No	http://sun3d.cs.princeton.edu/
36	SUN RGB-D	No	No	No	http://rgbd.cs.princeton.edu
37	TST fall detection	No	Yes	No	http://www.tlc.dii.univpm.it/blog/databases4 Kinect
38	TUM	Yes	Yes	No	http://vision.in.tum.de/data/datasets/rgbd-dataset
39	TUM texture-less	No	No	No	http://campar.in.tum.de/Main/StefanHinterstoisser
40	UR fall detection	No	Yes	No	http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html
41	UTD-MHAD	No	No	No	http://www.utdallas.edu/~kehtar/UTD-MHAD.html
42	Vienna University Technology object	No	No	No	http://users.acin.tuwien.ac.at/aaldoma/datasets/ECCV.zip
43	Willow garage	No	No	No	http://www.acin.tuwien.ac.at/forschung/v4r/mitarbeiterprojekte/willow/
44	Workout SU-10 exercise	No	No	Yes	http://vpa.sabanciuniv.edu.tr/phpBB2/vpa_views.php?s=31&serial=36
45	3D-Mask	NO	NO	Yes	https://www.widiap.ch/dataset/3dmask
46	50 salads	No	No	Yes	http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/

be easily seen from the table that the datasets with longer history [48, 52, 85] always have more related references than those of new datasets [44, 104]. Particularly, Cornell Activity, MSR Action3D Dataset, MSRDailyActivity3D, MSRGesture3D, Object RGB-D, People, RGBD-HuDaAct, TUM and NYU Depth V1 and V2 all have more than 100 citations. However, it does not necessarily mean that the old datasets are better than the new ones.

Table 5 presents the following information: the intended applications of the datasets, label information, data modalities and the number of the activities or objects or scenes along with the datasets. The intended applications (the third column) of the datasets are divided into five categories. However, each dataset may not be limited to one specific application only. For example, object RGB-D can be used in detection as well. The label information (the fourth column) is valuable because it aids in the process of annotation. The data modalities (the fifth column) include color, depth, skeleton and accelerometer, which are helpful for researchers to quickly identify the datasets especially when they work on multi-modal fusion [15, 16, 58]. Accelerometer data is able to indicate the potential impact of the object and starts an analysis of depth information, at the same time, it simplifies complexity of the motion feature and increases its reliability. The number of the activities or objects or scenes is connected closely with the intended application. For example, if the application is SLAM, we focus on the number of the scenes in the dataset.

Table 6 concludes the information, such as whether the sensor moves during the collection process, whether it enables multi-sensor or not, whether it is download restricted, and the web link of the dataset. Camera movement is another important information when the algorithm selects the datasets for its evaluation. The rule in this survey is as follows: if the camera is still all the time in the collection procedure, it is marked “No”, otherwise “Yes”. The fifth column is related to the license agreement requirement. Most of the datasets can be downloaded directly from the web. However, downloading data from some datasets may need to fill in a request form. Moreover, few datasets are not public. The link to each dataset is also provided which can better help the researchers in related research areas. It needs to pay attention that some datasets are updating while some dataset webs may change.

4 Conclusion

There is a great number of RGB-D datasets created for evaluating various computer vision algorithms since the low-cost sensor such as Microsoft Kinect has been launched. The growing number of datasets actually increases the difficulty in selecting appropriate dataset. This survey tries to cover the lack of a complete description of the most popular RGB-D test sets. In this paper, we have presented 46 existing RGB-D datasets, where 20 more important datasets are elaborated but the other less popular ones are briefly introduced. Each dataset above falls into one of five categories defined in this survey. The characteristics, as well as the ground truth format of each dataset, are concluded in the tables. The comparison of different datasets belonging to the same category is also provided, indicating the popularity and also the difficult level of the dataset. The ultimate goal is to guide researchers in the election of suitable datasets for benchmarking their algorithms.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

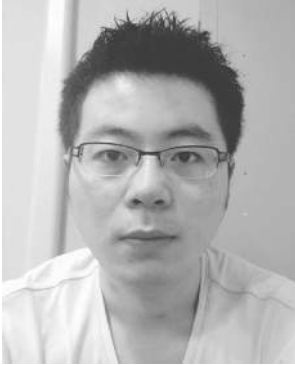
References

1. Abdallah D, Charpillet F (2015) Pose estimation for a partially observable human body from rgb-d cameras. In: International Conference on Intelligent Robots and Systems, p 8
2. Aldoma A, Tombari F, Di Stefano L, Vincze M (2012) A global hypotheses verification method for 3d object recognition. In: European Conference on Computer Vision, pp 511–524
3. Aggarwal JK, Cai Q (1997) Human motion analysis: A review. In: Nonrigid and Articulated Motion Workshop, pp 90–102
4. Baltrusaitis T, Robinson P, Morency L (2012) 3d constrained local model for rigid and non-rigid facial tracking. In: Conference on Computer Vision and Pattern Recognition, pp 2610–2617
5. Barbosa BI, Cristani BI, Del Bue A, Bazzani L, Murino V (2012) Re-identification with rgb-d sensors. In: First International Workshop on Re-Identification, pp 433–442
6. Berger K The role of rgb-d benchmark datasets: an overview. arXiv:1310.2053
7. Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C (2014) Learning 6d object pose estimation using 3d object coordinates. In: European Conference on Computer Vision, pp 536–551
8. Bourke A, O'Brien J, Lyons G (2007) Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. *Gait & posture* 26(2):194–199
9. Bo L, Ren X, Fox D (2011) Depth kernel descriptors for object recognition. In: International Conference on Intelligent Robots and Systems, pp 821–826
10. Bo L, Lai K, Ren X, Fox D (2011) Object recognition with hierarchical kernel descriptors. In: Conference on Computer Vision and Pattern Recognition, pp 1729–1736
11. Bo L, Ren X, Fox D (2013) Unsupervised feature learning for rgb-d based object recognition. In: Experimental Robotics, pp 387–402
12. Borràs R, Lapedriza À, Igual L (2012) Depth information in human gait analysis: an experimental study on gender recognition. In: Image Analysis and Recognition, pp 98–105
13. Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery. *Pattern Recogn Lett* 34(15):1995–2006
14. Chen C, Jafari R, Kehtarnavaz N (2015) Utd-mad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: IEEE International conference on image processing
15. Chen C, Jafari R, Kehtarnavaz N (2015) Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Transactions on Human-Machine Systems* 45(1):51–61
16. Chen C, Jafari R, Kehtarnavaz N A real-time human action recognition system using depth and inertial sensor fusion
17. Cruz L, Lucio D, Velho L (2012) Kinect and rgbd images: Challenges and applications. In: Conference on Graphics, Patterns and Images Tutorials, pp 36–49
18. Chua CS, Guan H, Ho YK (2002) Model-based 3d hand posture estimation from a single 2d image. *Image and Vision computing* 20(3):191–202
19. De Rosa R, Cesa-Bianchi N, Gori I, Cuzzolin F (2014) Online action recognition via nonparametric incremental learning. In: British machine vision conference
20. Drouard V, Ba S, Evangelidis G, Deleforge A, Horaud R (2015) Head pose estimation via probabilistic high-dimensional regression. In: International conference on image processing
21. Endres F, Hess J, Engelhard N, Sturm J, Cremers D, Burgard W (2012) An evaluation of the RGB-D SLAM system. In: International Conference on Robotics and Automation, pp 1691–1696
22. Endres F, Hess J, Sturm J, Cremers D, Burgard W (2014) 3-D mapping with an rgb-d camera. *IEEE Trans Robot* 30(1):177–187
23. Ellis C, Masood SZ, Tappen MF, Laviola JJ Jr, Sukthankar R (2013) Exploring the trade-off between accuracy and observational latency in action recognition. *Int J Comput Vis* 101(3):420–436
24. Erdogmus N, Marcel S (2013) Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect:1–6
25. Fanelli G, Dantone M, Gall J, Fossati A, Van Gool L (2013) Random forests for real time 3d face analysis. *International Journal on Computer Vision* 101(3):437–458
26. Fothergill S, Mentis HM, Kohli P, Nowozin S (2012) Instructing people for training gestural interactive systems. In: Conference on Human Factors in Computer Systems, pp 1737–1746
27. Garcia J, Zalevsky Z (2008) Range mapping using speckle decorrelation. United States Patent 7 433:024
28. Gasparrini S, Cippitelli E, Spinsante S, Gambi E A depth-based fall detection system using a kinect@sensor. *Sensors* 14(2):2756–2775

29. Gao J, Ling H, Hu W, Xing J (2014) Transfer learning based visual tracking with gaussian processes regression. In: European Conference on Computer Vision, pp 188–203
30. Geng J (2011) Structured-light 3d surface imaging: a tutorial. *Adv Opt Photon* 3(2):128–160
31. Gossow D, Weikersdorfer D, Beetz M (2012) Distinctive texture features from perspective-invariant keypoints. In: International Conference on Pattern Recognition, pp 2764–2767
32. Gupta S, Girshick R, Arbeláez P, Malik J (2014) Learning rich features from rgb-d images for object detection and segmentation. In: European Conference on Computer Vision, pp 345–360
33. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. *IEEE Transactions on Cybernetics* 43(5):1318–1334
34. Handa A, Whelan T, McDonald J, Davison A (2014) A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: International Conference on Robotics and Automation, pp 1524–1531
35. Helmer S, Meger D, Muja M, Little JJ, Lowe DG (2011) Multiple viewpoint recognition and localization. In: Asian Conference on Computer Vision, pp 464–477
36. Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K, Navab N (2012) Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes, pp 548–562
37. Hinterstoisser S, Cagniard C, Ilic S, Sturm P, Navab N, Fua P, Lepetit V (2012) Gradient response maps for real-time detection of texture-less objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(5):876–888
38. Hornung A, Wurm KM, Bennewitz M, Stachniss C, Burgard W (2013) Octomap: an efficient probabilistic 3d mapping framework based on octrees. *Auton Robot* 34(3):189–206
39. Hu G, Huang S, Zhao L, Alempijevic A, Dissanayake G (2012) A robust rgb-d slam algorithm. In: International Conference on Intelligent Robots and Systems, pp 1714–1719
40. Huynh O, Stanculescu B (2015) Person re-identification using the silhouette shape described by a point distribution model. In: IEEE Winter Conference on Applications of Computer Vision, pp 929–934
41. Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T (2013) A category-level 3d object dataset: Putting the kinect to work. In: Consumer Depth Cameras for Computer Vision, Research Topics and Applications, pp 141–165
42. Jhuo IH, Gao S, Zhuang L, Lee D, Ma Y Unsupervised feature learning for rgb-d image classification
43. Jin L, Gao S, Li Z, Tang J (2014) Hand-crafted features or machine learnt features? together they improve rgb-d object recognition. In: IEEE International Symposium on Multimedia, pp 311–319
44. Karpathy A, Miller S, Fei-Fei L (2013) Object discovery in 3d scenes via shape analysis. In: International Conference on Robotics and Automation (ICRA), pp 2088–2095
45. Kerl C, Sturm J, Cremers D (2013) Robust odometry estimation for rgb-d cameras. In: International Conference on Robotics and Automation, pp 3748–3754
46. Kepski M, Kwolek B (2014) Fall detection using ceiling-mounted 3d depth camera. In: International Conference on Computer Vision Theory and Applications, vol 2, pp 640–647
47. Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research* 32(8):951–970
48. Koppula HS, Anand A, Joachims T, Saxena A (2011) Semantic labeling of 3d point clouds for indoor scenes. In: Advances in Neural Information Processing Systems, pp 244–252
49. Kumatani K, Arakawa T, Yamamoto K, McDonough J, Raj B, Singh R, Tashev I (2012) Microphone array processing for distant speech recognition: Towards real-world deployment. *Asia Pacific Signal and Information Processing Association Conference*:1–10
50. Kurakin A, Zhang Z, Liu Z (2012) A real time system for dynamic hand gesture recognition with a depth sensor. In: European Signal Processing Conference (EUSIPCO), pp 1975–1979
51. Kwolek B, Kepski M (2014) Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput Methods Prog Biomed* 117(3):489–501
52. Lai K, Bo L, Ren X, Fox D (2011) A large-scale hierarchical multi-view rgb-d object dataset. In: International Conference on Robotics and Automation, pp 1817–1824
53. Lai K, Bo L, Ren X, Fox D (2013) Rgb-d object recognition: Features, algorithms, and a large scale benchmark. In: Consumer Depth Cameras for Computer Vision, pp 167–192
54. Lee TK, Lim S, Lee S, An S, Oh SY (2012) Indoor mapping using planes extracted from noisy rgb-d sensors. In: International Conference on Intelligent Robots and Systems, pp 1727–1733
55. Leibe B, Cornelis N, Cornelis K, Van Gool L (2007) Dynamic 3d scene analysis from a moving vehicle. In: Conference on Computer Vision and Pattern Recognition, pp 1–8
56. Leroy J, Rocca F, Mancaçs M, Gosselin B (2013) 3d head pose estimation for tv setups. In: Intelligent Technologies for Interactive Entertainment, pp 55–64

57. Liu L, Shao L (2013) Learning discriminative representations from rgb-d video data. In: International joint conference on Artificial Intelligence, pp 1493–1500
58. Liu K, Chen C, Jafari R, Kehtarnavaz N (2014) Fusion of inertial and depth sensor data for robust hand gesture recognition. *Sensors Journal* 14(6):1898–1903
59. Luber M, Spinello L, Arras KO (2011) People tracking in rgb-d data with on-line boosted target models. In: International Conference on Intelligent Robots and Systems, pp 3844–3849
60. Mason J, Marthi B, Parr R (2012) Object disappearance for object discovery. In: International Conference on Intelligent Robots and Systems, pp 2836–2843
61. Mason J, Marthi B, Parr R (2014) Unsupervised discovery of object classes with a mobile robot. In: International Conference on Robotics and Automation, pp 3074–3081
62. Mantecon T, del Bianco CR, Jaureguizar F, Garcia N (2014) Depth-based face recognition using local quantized patterns adapted for range data. In: International Conference on Image Processing, pp 293–297
63. Meister S, Izadi S, Kohli P, Hämmerle M, Rother C, Kondermann D (2012) When can we use kinectfusion for ground truth acquisition? In: Workshop on color-depth camera fusion in robotics
64. Min R, Kose N, Dugelay JL (2014) Kinectfacedb: a kinect database for face recognition. *IEEE Transactions on Cybernetics* 44(11):1534–1548
65. Nathan Silberman PK, Hoiem D, Fergus R (2012) Indoor segmentation and support inference from rgb-d images. In: European Conference on Computer Vision, pp 746–760
66. Narayan KS, Sha J, Singh A, Abbeel P Range sensor and silhouette fusion for high-quality 3d scanning, *sensors* 32(33):26
67. Negin F, Özdemir F, Akgül CB, Yüksel KA, Erçil A (2013) A decision forest based feature selection framework for action recognition from rgb-depth cameras. In: *Image Analysis and Recognition*, pp 648–657
68. Ni B, Wang G, Moulin P (2013) Rgb-d-hudaact: A color-depth video database for human daily activity recognition. In: *Consumer Depth Cameras for Computer Vision*, pp 193–208
69. Oikonomidis I, Kyriazis N, Argyros AA (2011) Efficient model-based 3d tracking of hand articulations using kinect. In: *British Machine Vision Conference*, pp 1–11
70. Pomerleau F, Magnenat S, Colas F, Liu M, Siegwart R (2011) Tracking a depth camera: Parameter exploration for fast icp. In: International Conference on Intelligent Robots and Systems, pp 3824–3829
71. Rezik A, Ben-Hamadou A, Mahdi W (2013) 3d face pose tracking using low quality depth cameras. In: International Conference on Computer Vision Theory and Applications, vol 2, pp 223–228
72. Richtsfeld A, Morwald T, Prankl J, Zillich M, Vincze M (2012) Segmentation of unknown objects in indoor environments. In: International Conference on Intelligent Robots and Systems, pp 4791–4796
73. Richtsfeld A, Mörwald T, Prankl J, Zillich M, Vincze M (2014) Learning of perceptual grouping for object segmentation on rgb-d data. *Journal of visual communication and image representation* 25(1):64–73
74. Rusu RB, Cousins S (2011) 3d is here: Point cloud library (pcl). In: International Conference on Robotics and Automation, pp 1–4
75. Salas-Moreno RF, Glocken B, Kelly PH, Davison AJ (2014) Dense planar slam. In: *IEEE International Symposium on Mixed and Augmented Reality*, pp 157–164
76. Satta R (2013) Dissimilarity-based people re-identification and search for intelligent video surveillance. Ph.D. thesis
77. Shao T, Xu W, Zhou K, Wang J, Li D, Guo B (2012) An interactive approach to semantic modeling of indoor scenes with an rgb-d camera. *ACM Trans Graph* 31(6):136
78. Shotton J, Glocker B, Zach C, Izadi S, Criminisi A, Fitzgibbon A (2013) Scene coordinate regression forests for camera relocalization in rgb-d images. In: *Conference on Computer Vision and Pattern Recognition*, pp 2930–2937
79. Silberman L, Fergus R (2011) Indoor scene segmentation using a structured light sensor. In: International Conference on Computer Vision - Workshop on 3D Representation and Recognition, pp 601–608
80. Singh A, Sha J, Narayan KS, Achim T, Abbeel P (2014) Bigbird: A large-scale 3d database of object instances. In: International Conference on Robotics and Automation, pp 509–516
81. Song S, Xiao J (2013) Tracking revisited using rgb-d camera: Unified benchmark and baselines. In: International Conference on Computer Vision, pp 233–240
82. Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: A rgb-d scene understanding benchmark suite. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 567–576
83. Spinello L, Arras KO (2011) People detection in rgb-d data. In: International Conference on Intelligent Robots and Systems, pp 3838–3843

84. Sturm J, Magnenat S, Engelhard N, Pomerleau F, Colas F, Burgard W, Cremers D, Siegwart R (2011) Towards a benchmark for rgb-d slam evaluation. In: RGB-D workshop on advanced reasoning with depth cameras at robotics: Science and systems conference, vol 2
85. Sturm J, Engelhard N, Endres F, Burgard W, Cremers D (2012) A benchmark for the evaluation of rgb-d slam systems. In: International Conference on Intelligent Robot Systems, pp 573–580
86. Sturm J, Burgard W, Cremers D (2012) Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark. In: Workshop on color-depth camera fusion in international conference on intelligent robot systems
87. Steinbruecker D, Sturm J, Cremers D (2011) Real-time visual odometry from dense rgb-d images. In: Workshop on Live Dense Reconstruction with Moving Cameras at the International Conference on Computer Vision, pp 719–722
88. Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: International joint conference on Pervasive and ubiquitous computing, pp 729–738
89. Stein S, McKenna SJ (2013) User-adaptive models for recognizing food preparation activities. In: International workshop on Multimedia for cooking & eating activities, pp 39–44
90. Sutton MA, Orteu JJ, Schreier H (2009) Image correlation for shape, motion and deformation measurements: basic concepts theory and applications
91. Sun M, Bradski G, Xu BX, Savarese S (2010) Depth-encoded hough voting for joint object detection and shape recovery. In: European Conference on Computer Vision, pp 658–671
92. Sung J, Ponce C, Selman B, Saxena A Human activity detection from rgb-d images., plan, activity, and intent recognition 64
93. Susanto W, Rohrbach M, Schiele B (2012) 3d object detection with multiple kinects. In: European Conference on Computer Vision Workshops and Demonstrations, pp 93–102
94. Tao D, Jin L, Yang Z, Li X (2013) Rank preserving sparse learning for kinect based scene classification. *IEEE Transactions on Cybernetics* 43(5):1406–1417
95. Tao D, Cheng J, Lin X, Yu J Local structure preserving discriminative projections for rgb-d sensor-based scene classification, *Information Sciences*
96. Vaufreydaz D, Nègre A (2014) Mobilergbd, an open benchmark corpus for mobile rgb-d related algorithms. In: International conference on control, Automation, Robotics and Vision
97. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1290–1297
98. Wang J, Liu Z, Wu Y, Yuan J (2014) Learning actionlet ensemble for 3d human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(5):914–927
99. Wohlkinger W, Aldoma A, Rusu RB, Vincze M (2012) 3dnet: Large-scale object class recognition from cad models. In: International Conference on Robotics and Automation, pp 5384–5391
100. Wu D, Shao L (2014) Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 724–731
101. Xiao J, Owens A, Torralla A (2013) Sun3d: A database of big spaces reconstructed using sfm and object labels. In: International Conference on Computer Vision, pp 1625–1632
102. Yang Y, Guha A, Fermueller C, Aloimonos Y (2014) Manipulation action tree bank: A knowledge resource for humanoids:987–992
103. Yu G, Liu Z, Yuan J (2015) Discriminative orderlet mining for real-time recognition of human-object interaction. In: Asian Conference on Computer Vision, pp 50–65
104. Zhang Q, Song X, Shao X, Shibusaki R, Zhao H (2013) Category modeling from just a single labeling: Use depth information to guide the learning of 2d models. In: Conference on Computer Vision and Pattern Recognition, pp 193–200
105. Zhou Q-Y, Koltun V (2013) Dense scene reconstruction with points of interest. *ACM Trans Graph* 32(4):112–117



Ziyun Cai received the B.Eng. degree in telecommunication and information engineering from Nanjing University of Posts and Telecommunications, Nan jing, China, in 2010. He is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K. His current research interests include RGB-D human action recognition, RGB-D scene and object classification and computer vision.



Jungong Han is currently a Senior Lecturer with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle, UK. He received his Ph.D. degree in Telecommunication and Information System from Xidian University, China, in 2004. During his Ph.D study, he spent one year at Internet Media group of Microsoft Research Asia, China. Previously, he was a Senior Scientist (2012-2015) with Civolution Technology (a combining synergy of Philips Content Identification and Thomson STS), a Research Staff (2010-2012) with the Centre for Mathematics and Computer Science (CWI), and a Senior Researcher (2005-2010) with the Technical University of Eindhoven (TU/e) in Netherlands. Dr. Han's research interests include multimedia content identification, multisensor data fusion, computer vision and multimedia security. He has written and co-authored over 80 papers. He is an associate editor of Elsevier Neurocomputing and an editorial board member of Springer Multimedia Tools and Applications. He has edited one book and organized several special issues for journals such as IEEE T-NNLS and IEEE T-CYB.



Li Liu received the B.Eng. degree in electronic information engineering from Xi'an Jiaotong University, Xi'an, China, in 2011, and the Ph.D. degree in the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., in 2014. Currently, he is a research fellow in the Department of Computer Science and Digital Technologies at Northumbria University. His research interests include computer vision, machine learning and data mining.



Ling Shao is a Professor with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle upon Tyne, U.K. Previously, he was a Senior Lecturer (2009-2014) with the Department of Electronic and Electrical Engineering at the University of Sheffield and a Senior Scientist (2005-2009) with Philips Research, The Netherlands. His research interests include Computer Vision, Image/Video Processing and Machine Learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology.