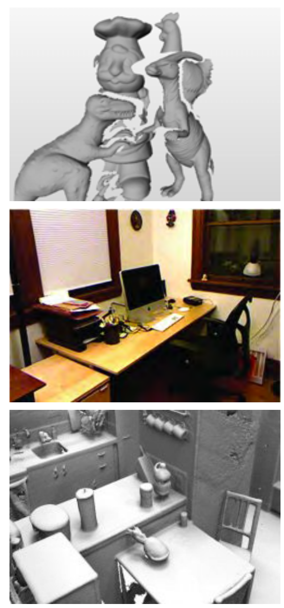
RGBD Datasets: Past, Present and Future

Michael Firman
University College London
http://www.michaelfirman.co.uk/RGBDdatasets/

Abstract

Since the launch of the Microsoft Kinect, scores of RGBD datasets have been released. These have propelled advances in areas from reconstruction to gesture recognition. In this paper we explore the field, reviewing datasets across eight categories: semantics, object pose estimation, camera tracking, scene reconstruction, object tracking, human actions, faces and identification. By extracting relevant information in each category we help researchers to find appropriate data for their needs, and we consider which datasets have succeeded in driving computer vision forward and why.

Finally, we examine the future of RGBD datasets. We identify key areas which are currently underexplored, and suggest that future directions may include synthetic data and dense reconstructions of static and dynamic scenes.



(a) **Past**

Before the Microsoft Kinect, most depth datasets were small and captured in the laboratory.

Image from [77]

(b) **Present**

We now enjoy RGBD data from dynamic and static scenes from the real world, with a range of labeling and capture conditions.

Image from [93]

(c) **Future**

We can anticipate scans of static and dynamic scenes as fused geometry, exploiting improvements in reconstruction algorithms.

Image from [20]

Figure 1. The past, present and future of RGBD datasets.

1. Introduction

Before the Microsoft Kinect was launched in November 2010, collecting images with a depth channel was a cumbersome and expensive task. Researchers built custom active stereo setups [12] and made use of 3D scanners costing tens of thousands of dollars [77, 19]. Many of these early datasets captured static images of objects in isolation, as the sensors used did not transport easily (Fig 1a).

Early Kinect datasets also focused on static images, often of single objects or small scenes. As the field matures we see research being put to effect in creating larger and more ambitious RGBD datasets, and the quantity released each year shows no sign of decreasing (Figure 2). Semantic labels have been propagated through videos [112], dense reconstruction has been exploited to capture the surfaces of whole objects [21] and generative scene algorithms have been used to create plausible synthetic data [43]. We also see new labels applied to existing data [41] and previous releases being recompiled into new offerings [95].

In spite of the current availability of sensors, though, collecting RGBD data is still not trivial. Researchers using the Kinect have built battery devices [93, 95], written drivers [95] and developed custom data formats [34]. Publicly available RGBD datasets can, at the most basic level, remove the need to repeat data capture. More importantly, they provide transparency in the presentation of results and allow for scores to be compared on the same data by different researchers. This in turn can drive competition for better-performing algorithms. Finally, a dataset can help draw research towards previously under-explored directions.

Our primary contribution is to give a snapshot of public RGBD datasets, allowing researchers to easily select data appropriate for their needs (Section 2). We are more comprehensive than earlier efforts, describing 101 datasets compared with the 14 in [9], 19 in [42][1] and the 44 action datasets in [117]. We secondly identify areas where there is opportunity for new data to facilitate novel areas of research (Section 3). We hypothesize that we can expect datasets to continue to move away from single images, to dense reconstructions of static and dynamic scenes (Figure 1c).

[1] [42] references more than 19 datasets, but most are not RGBD

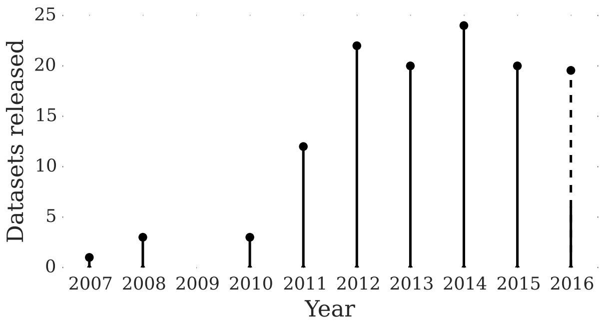


Figure 2. Our estimate of the number of depth datasets released each year, where projected releases in 2016 are shown as a dashed line. The Kinect was first released in November 2010.

2. State-of-the-art in RGBD datasets

Here we review state-of-the-art datasets across eight categories. Some fall into more than one category, and the difference between categories depends as much on the labeling as it does the image content.

We include datasets which have been captured with an active capture devices such as time-of-flight or structured light, but exclude data from passive stereo. We also exclude Lidar datasets, focusing instead on data from the separate world of commodity depth capture. Following the mantra that 'data is cheap, information is expensive', we focus on data which has some form of human labeling applied. We exclude very small datasets, and those which have been produced mainly to demonstrate an acquisition method.

With these exceptions, we aim to be comprehensive and correct. Please flag omissions and errors to m.firman@cs.ucl.ac.uk so this document can be updated. We also maintain a web-based version[2].

We first look at datasets of objects in isolation, before moving on to datasets for camera tracking, scene reconstruction and then datasets where the pose of objects is to be inferred. Semantic, and then tracking datasets come next, before videos for action and gesture recognition. We finish with two more categories involving humans: faces and identity recognition.

2.1. Objects in isolation

Following earlier stereo setups such as [79], RGBD turntable datasets offer multiple unoccluded views of the same object from different angles (Table 1).

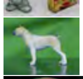
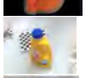
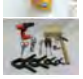
The 2011 RGB-D Object Dataset [62] is a well-used dataset with 300 objects, but does not contain accurate camera poses. This was rectified by more recent datasets such as BigBIRD [94]. While a smaller dataset, BigBIRD is captured with calibrated Kinects and DSLRs.

Turntable datasets have been exploited in 'natural' scenes for tasks such as object detection [63] and discovery [31]. In many ways, though, they are limited by their deviation from real-world data. Without occlusion, lighting changes or varying distances to objects these datasets sit in a different domain to the real-world scenes which we ultimately aim to understand.

Choi et al. [21] exploit improvements in camera tracking to form a dataset of individual objects scanned in the real world. With 10,000 items ranging in size from books to cars, this is the largest dataset of real-life objects by two orders of magnitude.

[2] http://www.michaelfirman.co.uk/RGBDdatasets/

Table 1. Datasets capturing single objects in isolation

		Device[a]	# objects	Camera pose?[b]	Year
	RGBD Object Dataset [62]	Kinect v1	300	-	'11
	KIT object database [54]	Minolta Vi-900 and stereo pair	>100	✓✓	'12
	A dataset of Kinect-based 3D scans [24]	Kinect and Minolta Vi-900	59	✓✓	'13
	MV-RED [68]	Kinect v1	505	-	'14
	BigBIRD dataset [94]	Asus Xtion Pro, DSLR	125	✓✓	'14
	YCB Object and Model Set[c] [15]	Asus Xtion Pro, DSLR	88	✓✓	'15
	A large dataset of object scans [21]	PrimeSense Carmine	>10,000	✓	'16

[a] The Kinect v1, Asus Xtion Pro and PrimeSense Carmine have almost identical internals and can be considered to give equivalent data.
[b] ✓= camera pose computed from RGBD data; ✓✓= camera pose from calibration system.
[c] Captured using the same turntable setup as the BigBIRD dataset.

2.2. Camera tracking and scene reconstruction

Arguably some of the main advances brought by consumer depth cameras have been in camera tracking and dense reconstruction. Ground truth camera poses are necessary to validate these algorithms, and these are difficult to acquire as they require external hardware.

For **camera tracking**, the TUM benchmark [99] has become a de-facto standard for evaluation, with ground truth data from a motion tracking system and a range of scenes and camera motions. We summarize this and similar datasets in Table 2.

Some datasets [91, 74, 120, 30] use manually verified tracking from the Kinect itself as a ground truth pose. This data is only suitable for tasks an order of magnitude harder than tracking, such as camera relocalization [91] or voxel occupancy prediction [30].

The difficulties involved with acquiring ground truth data can be circumvented with synthetic data. The ICL-NUIM dataset [44] provides 8 camera trajectories for two synthetic indoor scenes, with camera paths taken from real hand-held camera trajectories. While synthetic datasets may not be a perfect representation of our world, they allow users to more carefully control aspects such as motion blur and texture levels to gain introspection into their algorithm (see Section 3.1 for further discussion).

Scene reconstruction is rarely evaluated directly, as good camera tracking usually corresponds to good reconstruction and camera paths are easier to obtain as ground truth than dense surfaces. The synthetic ICL-NUIM dataset [44] is suitable for reconstruction evaluation, especially with additional camera paths provided by [20]. More recently Wasenmüller et al. [109] created a dataset containing ground truth camera motions and scene reconstructions from a laser scanner. This is the only real-world dataset we are aware of with both these data, though the scenes are less diverse than [99].

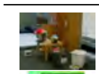
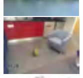
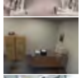
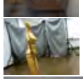
Firman et al. [30] have a dataset of tabletop objects scanned so every visible surface is observed in the reconstruction. This provides ground truth for the task of estimating the unobserved voxel occupancy from a depth image.

2.3. Object pose estimation

The problem of inferring the 6-DoF pose of an object is again a task which has been aided by the absolute scale provided by depth cameras. Given *a priori* a 3D model of an object, the aim is to find the transformation which best aligns it into the scene. As with camera tracking it is hard to get ground truth for this type of challenge, which requires both a 3D model of the object and its pose in each image. One solution has been to fix the target objects to a calibration board to allow for ground-truth tracking using the RGB channels [45], while [87] and [85] have the poses manually aligned.

These datasets, summarized in Table 3, feature tabletop-sized objects. Acquiring 3D models, and ground truth poses, for larger objects is difficult, so works that have attempted this problem on a room scale typically find an alternative method of evaluation or rely on human annotations as an approximate ground truth [95]. Synthetic data could be an avenue worth exploring here.

Table 2. Datasets for camera pose and scene reconstruction

		Device[a]	# videos	Camera pose[b]	Ground truth surface	Notes	Year
	IROS 2011 Paper Kinect Dataset [84]	Kinect v1	27	✓✓		-	’11
	KinectFusion for Ground Truth [74]	Kinect v1		✓	✓	Lidar surface ground truth for some scenes	’12
	TUM benchmark [99]	Kinect v1	47	✓✓		-	’12
	Indoor RGB-D Dataset [88]	Kinect v1	4	✓✓		Collected from a robot	’13
	Microsoft 7-scenes [91]	Kinect v1	>14	✓		Designed for camera relocalization tasks	’13
	Robust Reconstruction Datasets [120]	Asus Xtion Pro	8	✓		-	’13
	ICL-NUIM Dataset [44]	Synthetic	8	✓✓	✓	Camera paths from [20] allow for reconstruction evaluation	’14
	CoRBS Dataset [109]	Kinect v2	20	✓✓	✓	Surface ground truth from fixed structured light scanner	’16
	Voxel Occupancy Prediction [30]	Asus Xtion Pro	90	✓		Densely captures full visible surface	’16

[a] The Kinect v1, Asus Xtion Pro and PrimeSense Carmine have almost identical internals and can be considered to give equivalent data.
[b] ✓= approximated camera pose from Kinect tracking. ✓✓= ground truth camera pose from external system.

Realism: ●○○

Laboratory scenarios, with a limited set of objects arranged by hand.

Image from [103]

Realism: ●●○

Real-world scenes, but with furniture or objects artificially arranged.

Image from [62]

Realism: ●●●

Real-world scenes with no interference by researchers.

Image from [93]

Figure 3. Semantic datasets described in Table 4 view the world in various levels of 'realism', which we discretise into three categories.

2.4. Semantic labeling

Semantic labeling of images and videos moves us to a more general understanding of the world. Datasets with labels which could be used for semantic understanding are listed in Table 4. We give an indication of the 'realism' of each dataset as a score out of three, explained in Figure 3. Note that a low score here does not correspond to a worse or less useful dataset, as datasets with specially constructed scenarios can be vital for proving concepts, and they can often provide higher quality ground truth than fully natural scenes.

The 1449-frame subset of the NYUv2 dataset [93] with dense semantic labels has become a de-facto standard for indoor scene labeling. The quality and variety of labels on this real-world dataset has helped make it one of the most highly used in the literature. The SUN3D dataset [112] counters the single, static-frame modality of NYUv2 with object labels propagated through Kinect videos. However, in spite of their effort, there are only 8 annotated sequences.

We note that all these semantic datasets, even those with videos, depict a static world. This contrasts with our *dynamic* world, an area which is explored by datasets designed for tracking.

2.5. Tracking

Tracking datasets feature videos of *dynamic* worlds, where the aim is to detect where an object is in each frame. We know of only four datasets explicitly designed for this purpose, all of which use bounding boxes as annotations. The Princeton Tracking Benchmark [96] contains 100 videos of moving objects, such as dogs and toy cars. The RGB-D people dataset [97, 70], the Kinect Tracking Precision dataset [81] and the RGBD Pedestrian Dataset [7] all track humans.

Other datasets contain labels appropriate for tracking: two semantic scene datasets [112, 62] have static objects labeled through video as the camera moves, while the 6-DOF object pose annotations in [45] could also be useful.

2.6. Activities and gestures

Given the original use case of the Kinect as a sensor designed for human interaction, it is inevitable that much research would focus on recognizing gestures and activities from videos. See Table 5 for an overview of the large number of datasets in this area, and we refer the interested reader to [117] for a more detailed survey of this field.

Actions being performed include sign language [59], Italian hand gestures [26] and common daily actions such as standing up, drinking and reading [100, 58, 82, 106, 65, 16]. Three datasets of humans falling over [60, 39, 38] reflect an interest in use of RGBD sensors for monitoring vulnerable humans in their daily lives. Others are more niche: 50 Salads [98] features over 4 hours of people preparing mixed salads. Four datasets stand out for capturing humans with a full MoCap setup [23, 34, 83, 51], while the Manipulation Action Dataset [1] is unique in providing semantic segmentation of objects as they are manipulated. By far the largest gesture and action datasets are the ChaLearn gesture challenge [49] and NTU RGB+D [90], each with around 50,000 videos.

Many of these datasets suffer from being filmed in the confines of an office or laboratory, with researchers performing the actions. Filming real people at work and home would help prevent dataset bias and provide a more believable baseline for activity and gesture recognition.

Table 3. Datasets for object pose estimation

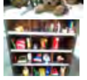
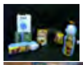
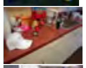
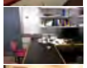
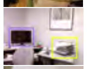
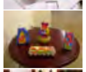
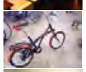
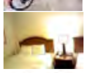
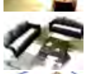
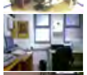
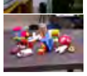
		Device	# objects	# frames	Notes	Year
	Cluttered scenes dataset [77]	Minolta Vivid 910	5	48	Manual ground truth alignment	'06
	LINEMOD RGBD dataset [45]	Kinect v1	15	>18,000	Ground truth from calibration board	'12
	SHOT dataset [87]	Kinect v1	6	16	-	'14
	Rutgers APC RGB-D Dataset [85]	Kinect v1	24	10,368	Semi-manual ground truth alignment	'16

Table 4. Datasets for semantic reasoning and segmentation

		Size	Video?	Realism[a]	Labeling	Year
	RGB-D Semantic Segmentation Dataset [103]	16 frames		●○○	Dense pixel labeling	'11
	RGBD Scenes dataset [62]	8 scenes	✓	●●○	Bounding box labeling of objects from the RGBD Objects dataset	'11
	Cornell-RGBD-Dataset [57]	52 scenes	✓	●●●	Semantic segmentation of reconstructed point cloud into 17 classes	'11
	NYUv1 [92]	2283 frames	-[b]	●●●	Dense pixel labeling	'11
	Berkeley 3-D Object Dataset [52]	848 frames		●●●	Bounding box annotation	'11
	Object segmentation dataset [86]	111 frames		●○○	Per-pixel segmentation into objects; no semantics	'12
	MPII Multi-Kinect dataset [101]	2240 frames total from 4 Kinects		●○○	Polygon segmentation of objects arranged on kitchen worktop	'12
	Willow garage dataset [2]	~160 frames		●○○	Dense pixel labeling	'12
	Object Disappearance for Object Discovery [73]	1231 frames	✓	●●○	Ground truth object segmentations of objects of interest	'12
	NYUv2 [93]	1449 frames from 464 scenes	-[b]	●●●	Dense pixel labeling. A synthetic re-creation of the 3D scenes also exists [41]	'12
	RGBD Dataset for Category Modeling [119]	900 frames		●●○	Which of 7 categories the dominant object in each image is in	'13
	SUN3D [112]	8 scenes	✓	●●●	Polygon labels. 8 scenes labeled, though full dataset has more	'13
	RGBD Scenes dataset v2 [61]	14 scenes	✓	●●○	Items from the RGBD Objects dataset labeled on reconstructed point cloud	'14
	SUN RGB-D [95]	10,335 frames[c]		●●●	3D object bounding boxes, and polygons on 2D images	'15
	ViDRILO [72]	22454 frames from 5 scenes	✓	●●●	Semantic category of frame, plus which objects are visible in each frame	'15
	Toy dataset [50]	449 frames		●○○	Per-pixel segmentation into objects; no semantics	'16

[a] See Figure 3
[b] Extended version of dataset has video, but labels are only present in subset described here.
[c] Combines new Kinect v2 frames with new labels on existing datasets [93, 52, 112]

2.7. Faces

Early face datasets focused on the method of acquisition (e.g. [118]) or tended to be quite small (e.g. [13]). The field has now expanded to include datasets for identity recognition [25], pose regression [12, 28], and those where the expressions or emotions are to be inferred [27, 78]. We summarize these in Table 6, and more details on some of these datasets can be found in [3]. As front-facing depth cameras become installed in laptops and tablets we expect this area of research to continue to gain attention.

2.8. Recognition

Like datasets of actions, datasets designed for human recognition (Table 7) typically film people performing activities such as walking. However, the aim now is to recognize the identity, gender or other attributes about the subjects, rather than the activity they are performing.

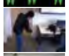
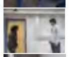
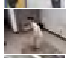
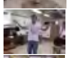
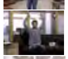
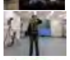
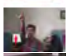
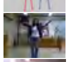
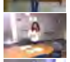
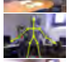
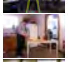
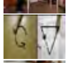
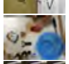
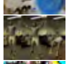
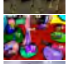
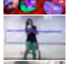
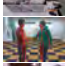
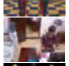
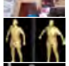
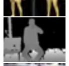
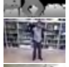
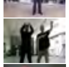
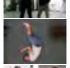
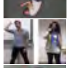
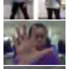
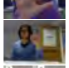
3. Future areas for datasets

In Section 2 we reviewed the past and the present of RGBD datasets. We now look to the future, and identify underexplored 'gaps in the market'.

3.1. Synthetic data

Aside from a few examples such as ICL-NUIM [44] and SceneNet [43], synthetic data has received relatively little attention for vision problems with depth cameras. Yet such artificial data can offer many advantages. Ground truth for tasks such as segmentation, reconstruction, tracking and camera or object pose is perfect and available with no requirement for expensive human labeling. Sequences can be recaptured with carefully adjusted parameters, e.g. motion blur and lighting changes, for algorithm introspection. It is also possible to create scenarios difficult to capture in real life, for example car crashes.

Table 5. Datasets representing activities and gestures

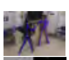
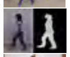
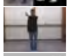
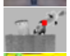
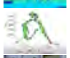
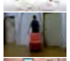
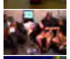
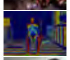
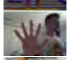
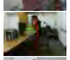
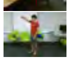
		# Subjects	# Actions	# Videos	Skeleton[a]	Examples of actions	Year
	MSR Action3D [65]	10	20	567	✓	e.g. *high arm wave, side kick, jogging*	’10
	RGBD-HuDaAct [82]	30	12	1189		e.g. *get up, enter room, stand up, mop the floor*	’11
	SBU Kinect Interaction Dataset [116]	7	8	300	✓	Two people interacting e.g. *approaching, departing*	’12
	ACT4^2 [18]	24	14	6844		4 Kinects filming. Actions: e.g. *collapse, reading*	’12
	UTKinect-Action [111]	10	10	200	✓	e.g. *walk, sit down, stand up, carry, clap hands*	’12
	MSRDailyActivity3D [106]	10	16	320	✓	e.g. *drink, eat, read book*	’12
	G3D Gaming Action Dataset [11]	10	20	600	✓	Typical gaming actions	’12
	MSRC-12 Kinect gesture [33]	30	12	594	✓	Arm gestures	’12
	MSRGesture3D [59]	10	12	336		American Sign Language	’12
	ChaLearn Gesture Challenge [49]	20	850	50000		Many, e.g. *diving signals and mudras*	’12
	Senior Activity Recognition (RGBD-SAR) [114]	30	9	810	✓	Older people performing activities e.g. *sit down, eat, walk, stand up*	’13
	K3HI [48]	15	8	320	✓	Two humans interacting e.g. *approaching, punching*	’13
	UPCV action dataset [102]	20	10	400	✓	e.g. *walk, wave, scratch head, phone, cross arms*	’13
	DML-SmartAction [5]	16	12	932		Continuous recording. e.g. *writing, sit down, walk, clean table, stand up*	’13
	Florence 3D actions dataset [89]	10	9	215	✓	e.g. *wave, drinking, answer phone, clap, stand up*	’13
	Cornell activty 60/120 [100, 58]	4	12/10	60/120	✓	e.g. *brushing teeth, drinking, talking on couch*	’13
	Sheffield KInect Gesture (SKIG) [69]	6	10	1080		Hand gestures e.g. *circle, up-down, comehere*	’13
	50 Salads [98]	25	2	50		Each person prepares two salads. Accelerometer on utensils	’13
	Berkeley Multimodal Human Action [83]	12	11	660	✓✓	e.g. *jumping, bending, punching*	’13
	Manipulation Action Dataset [1]	5	28	140		Manipulation actions e.g. *cutting*, plus sequences of actions. Semantic segmentation of frames.	’14
	Composable activities dataset [66]	14	16	693	✓	e.g. *throw, talk on phone, walk, wave, crouch, punch*	’14
	TUM Morning Routine Dataset [53]	1	-	-[b]	✓	Typical morning routine activities	’14
	ShakeFive [105]	37	2	100	✓	Hand shake or high-five between two individuals	’14
	Office activity dataset [108]	>10	20	1180		e.g. *mopping, sleeping, finding-objects, chatting*	’14
	Human3.6M [51]	11	17	-[b]	✓✓	e.g. *Discussion, smoking, taking photo*	’14
	MSR 3D Online Action [115]	24	7	-[b]		e.g. *drinking, eating, using laptop*	’14
	Northwestern-UCLA Multiview Action 3D [107]	10	10	-[b]	✓	Three Kinects filming. Actions: e.g. *stand up, throw*	’14
	G3Di Gaming Interaction Dataset [10]	12	17	-[b]	✓	Humans interacting with computer game	’14
	UR Fall Detection [60]	?	1	70		Humans falling over. Two Kinects. Accelerometer from human	’14
	Montalbano Gesture [26]	27	20	13858	✓	Italian hand gestures	’14
	LaRED Hand Gesture Dataset [47]	10	27	810		Modified American Sign Language	’14
	LTTM MS Kinect and Leap Motion [71]	14	10	1400		American Sign Language, recorded using Kinect and the Leap Motion	’14
	TJU dataset [67]	22	22	1936	✓	e.g. *boxing, one hand wave, forward bend, sit down*	’15

Continued overleaf ↓

[a] ✓ = 2D skeleton joint positions labeled on video frames; ✓✓ = 3D skeleton joint positions acquired from MoCap system

[b] These datasets feature continuous footage, so the discrete number of videos is less meaningful here.

↪ Continued from previous page

	M²I dataset [113]	22	22	1760	✓	Two people interacting, e.g. *walk together*	’15
	Multi-view TJU [67]	20	22	7040	✓	Front and side view Kinects. Actions as TJU dataset	’15
	UTD Multimodal Human Action [16]	8	27	861	✓	Accelerometer data. Actions: e.g. *wave, boxing*	’15
	TST Fall Detection ver. 1/ver. 2 [39, 38]	4/11	2	20/111	✓	Humans falling over	’15
	TST TUG [22]	20	?	60	✓	Timed Up and Go tests	’15
	TST Intake Monitoring ver 1/ver 2 [37]	35	?	35/60		Humans simulating eating	’15
	Life activities with occlusions [23]	1	-	12	✓✓	No specific actions	’15
	Background activity dataset [34]	52	4	-[b]	✓✓	Humans natually interacting in semi-natural environment	’15
	K3Da [64]	53	13	?	✓	To assess human health, e.g. *leg jump, walking*	’15
	LTTM Creative Senz3D [76]	4	11	1320		Hand gestures e.g. *‘OK’*	’15
	Watch-n-Patch [110]	7	21	458		A sequence of actions e.g. *making drink*	’15
	NTU RGB+D [90]	40	60	56,000	✓	e.g. *drinking, eating, sneezing, staggering, punching, kicking*	’16

While sensor noise can be emulated [44, 40, 29, 75], it can be very difficult for synthetic scenes to capture the true properties of the real world. One option is to use existing 3D assets. The synthetic Sintel dataset [14], for example, has been used for RGB tasks such as optical flow. With depth channels now available this may yet find a use in the RGBD community. Another route is to use generative models of scenes, following work on scene synthesis [32, 43].

3.2. Full voxel occupancy

Most existing semantic datasets view the world as a 2.5D image, where only surfaces directly viewed from one static camera position are visible (Figure 4a). Even datasets with videos (e.g. SUN3D [112]) tend to fail to capture the full surface geometry of scenes (Figure 4b). Full surface geometry is captured on an object level by [21] and on tabletop scenes by [30] (Figure 4c), but capturing and reconstruct-

Table 6. Datasets of faces for pose and recognition

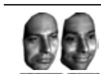
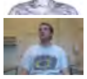
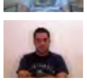
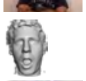
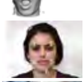
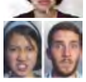
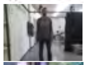
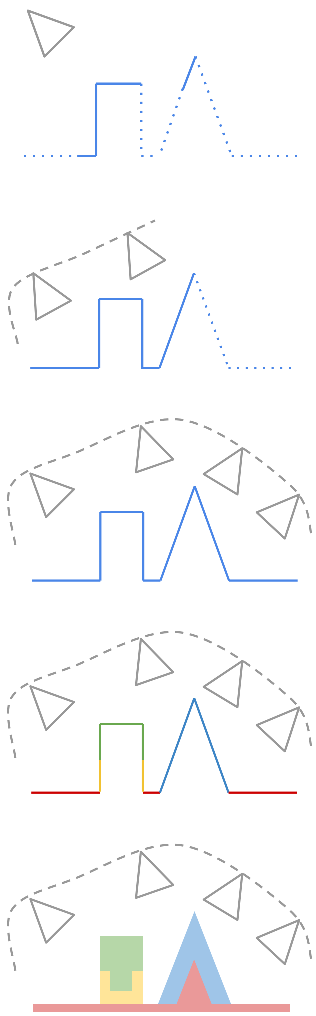
		Subjects	Sensor	Description	Labeling	Year
	Human Face [13]	1	Structured light scanner	15 expressions performed by one face	-	’07
	CASIA 3D Face Database [19]	123	Minolta Vivid 910	4624 images of various expressions, poses and lighting	Expression performed	’08
	Bosphorus Database [4]	105	Inspeck Mega Capturor II 3D	Faces performing expressions at different rotations	Expression and pose	’08
	ETH Face Pose Range Image Data Set [12]	20	Custom active stereo setup	Videos of face in various poses	Nose position and coordinate frame at the nose	’08
	B3D(AC)^2 [27]	14	Custom active stereo setup	Recordings of humans speaking	Perceived emotions. Audio labeled with phonemes	’10
	Biwi Kinect Head Pose Database [28]	20	Kinect v1	People moving their heads in different directions	3D position of the head and its rotation	’11
	VAP RGB-D Face Database [46]	31	Kinect v1	1581 images of people doing different poses in front a camera	Which person is in shot, and a discretised gaze direction	’12
	3D Mask Attack Dataset [25]	17	Kinect v1	Some frames are of person with a face mask of someone else	Person’s identity, and if ‘spoofing’ is occurring. Eye positions	’13
	Face Warehouse [17]	150	Kinect v1	People performing expressions	Which of 20 expressions, plus 74 landmarks and meshes	’14
	Eurecom Kinect Face Dataset [78]	52	Kinect v1	Faces with different expressions, occlusion and illumination	Expression type, and six facial landmark locations	’14
	VT-KFER [3]	32	Kinect v1	7 facial expressions labeled, in scripted and unscripted scenarios	Percieved expression	’15

Table 7. Datasets for human recognition

		Subjects	Description	Labeling	Year
	RGB-D Person Re-identification [8]	79	Humans walking, where subjects change clothes between sessions	2D skeleton positions. Which human is in each video	’12
	IAS-Lab RGBD-ID Dataset [80]	11	Humans walking, where subjects change clothes (or room) between sessions	2D skeleton positions. Which human is in each video	’14
	BIWI RGBD-ID Dataset [80]	50	Humans moving, where subjects change clothes (or room) between sessions	2D skeleton positions. Which human is in each video	’14
	UPCV Gait dataset [55]	30	Each human walks down corridor multiple times	Identity and gender of each person	’15



(a) Early RGBD datasets focussed on single images of scenes, representing them in 2½D.

Examples: [93, 52, 86]

(b) As reconstruction algorithms improved, datasets have used videos to capture more of the scene. These still miss the backs of many objects.

Example: [112]

(c) Few datasets capture the full visible surface geometry.

Examples: [21] captures objects, and [30] captures tabletop scenes.

(d) No datasets, to our knowledge, capture the full surface geometry of scenes *and* provide semantic labeling on the observed surface.

(e) Furthermore, our world extends beyond the visible surface. Dense *volumetric* labeling of scenes would enable a deeper level of understanding and interaction.

Figure 4. Datasets progress to include more 3D information.

ing a dataset of large, real-world scenes is left as an open challenge.

Labeling the surfaces of such dense reconstructions (Figure 4d) would allow for semantic segmentation on a *mesh* level. Many opportunities would be afforded by datasets which provide labeled on this form of dense reconstruction rather than on images or videos.

Furthermore, we can imagine the benefits of an algorithm which could segment or semantically label a scene on a *voxel* level, following works such as [56]. To train and validate such a system we would require a dataset containing semantic labeling of each voxel in a scene (Figure 4e). The difficulty of applying such labeling by hand may make synthetic data necessary for this problem.

3.3. Geometry of dynamic scenes

Aside from a single sequence from [6], we know of no RGBD datasets captured from *dynamic* scenes with ground truth dense geometry. One option is to use deformable meshes provided for face datasets [36, 104] or fabrics [35], which can be synthetically re-rendered to give dense correspondences between frames (e.g. Zollhöefer et al. [121] re-render data from [104]). Datasets of humans with motion capture data (Section 2.6) also give a very sparse dense geometry with correspondences.

The open challenge for the field of dense reconstruction is to directly capture an RGBD dataset of deforming objects with ground truth geometry and correspondences between frames.

4. Conclusion

We have discovered a considerable quantity of RGBD datasets available for researchers to use. While some overlap in their scope, overall the field is promisingly diverse which suggests that depth information is useful in many different sectors.

Most datasets we reviewed have been captured as single frames or videos from static cameras. We are now entering an era where the *collection* and *labeling* of datasets requires state-of-the-art computer vision research. For example, capturing a dense dataset such as [21] would not have been possible when the Kinect was first launched. As reconstruction and labeling algorithms for RGBD data improve, the community has a massive opportunity to create and share new datasets of 3D reconstructions of static, and ultimately dynamic scenes.

Acknowledgements

I am extremely grateful to Gabriel Brostow and his group for their relentless support, and to Lourdes Agapito for her helpful discussions. A big thanks also goes out to everyone who has released their datasets. Keep them coming!

References

For references which refer to a dataset we give a URL to the project page from which the data can be downloaded.

[1] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter. Model-free incremental learning of the semantics of manipulation actions. *Robotics and Autonomous Systems*, 2014. http://www.dpi.physik.uni-goettingen.de/~eaksoye/dataset.html.

[2] A. Aldoma and A. Richtsfeld. The Willow Garage Object Recognition Challenge, 2012. http://www.acin.tuwien.ac.at/forschung/v4r/mitarbeiterprojekte/willow/.

[3] S. Aly, A. Trubanova, A. L. Abbott, S. W. White, and A. E. Youssef. VT-KFER: A Kinect-based RGBD+time dataset for spontaneous and non-spontaneous facial expression recognition. In *International Conference on Biometrics*, 2015. http://sufficiency.ece.vt.edu/VT-KFER/.

[4] A. S. N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In *COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008)*, 2008. http://bosphorus.ee.boun.edu.tr/Home.aspx.

[5] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, and V. C. Leung. Non-intrusive human activity monitoring in a smart home environment. In *Conference on e-Health Networking, Application and Services (IEEE Healthcom 2013)*, 2013. http://dml.ece.ubc.ca/data/smartaction/.

[6] A.Varol, M. Salzmann, P. Fua, and R. Urtasun. A constrained latent variable model. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. http://cvlab.epfl.ch/data/dsr.

[7] T. Bagautdinov, F. Fleuret, and P. Fua. Probability occupancy maps for occluded depth images. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. http://cvlab.epfl.ch/data/rgbd-pedestrian.

[8] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with RGB-D sensors. In *First International Workshop on Re-Identification*, 2012. http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html.

[9] K. Berger. The role of RGB-D benchmark datasets: an overview. *arXiv:1310.2053*, 2013.

[10] V. Bloom, V. Argyriou, and D. Makris. G3Di: A gaming interaction dataset with a real time detection and evaluation framework. In *European Conference on Computer Vision (ECCV) Workshops*, 2014. http://dipersec.king.ac.uk/G3D/.

[11] V. Bloom, D. Makris, and V. Argyriou. G3D: A gaming action dataset and real time action recognition evaluation framework. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2012. http://dipersec.king.ac.uk/G3D/.

[12] M. D. Breitenstein, D. Kuettel, T. Weise, L. van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Computer Vision and Pattern Recognition (CVPR)*, 2008. http://www.vision.ee.ethz.ch/datasets/headposeCVPR08/.

[13] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Calculus of non-rigid surfaces for geometry and texture manipulation. *Transactions on Visualization and Computer Graphics*, 2007. http://tosca.cs.technion.ac.il/book/resources_data.html.

[14] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012. http://sintel.is.tue.mpg.de/.

[15] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *Conference on Advanced Robotics (ICAR)*, 2015. http://rll.eecs.berkeley.edu/ycb/.

[16] C. Chen, R. Jafari, and N. Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *International Conference on Image Processing (ICIP)*, 2015. http://www.utdallas.edu/~kehtar/UTD-MHAD.html.

[17] C. Chen, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: a 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics*, 2014. http://gaps-zju.org/facewarehouse/.

[18] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *European Conference on Computer Vision (ECCV) workshops*, 2012. http://vipl.ict.ac.cn/rgbd-action-dataset.

[19] Chinese Academy of Sciences' Institute of Automation (CASIA). CASIA-3D FaceV1, 2008. http://biometrics.idealtest.org/.

[20] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. http://redwood-data.org/indoor/dataset.html.

[21] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun. A large dataset of object scans. *arXiv:1602.02481*, 2016. http://redwood-data.org/3dscan.

[22] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, J. Wahslen, I. Orhan, and T. Lindh. Time synchronization and data fusion for RGB-Depth cameras and wearable inertial sensors in AAL applications. In *ICC2015 - Workshop on ICT-Enabled Services and Technologies for eHealth and Ambient Assisted Living*, 2015. http://www.tlc.dii.univpm.it/blog/databases4kinect.

[23] A. Dip and F. Charpillet. Pose estimation for a partially observable human body from RGB-D cameras. In *Intelligent Robots and Systems (IROS)*, 2015. https://team.inria.fr/larsen/software/datasets/.

[24] A. Doumanoglou, S. Asteriadis, D. S. Alexiadis, D. Zarpalas, and P. Daras. A dataset of Kinect-based 3D scans. In *3D Image/Video Technologies and Applications*, 2013. http://vcl.iti.gr/3d-scans/.

[25] N. Erdogmus and S. Marcel. Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. *Biometrics: Theory, Applications and Systems*, 2013. https://www.idiap.ch/dataset/3dmad.

[26] S. Escalera, X. Baró, J. González, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. ChaLearn looking at people challenge 2014: Dataset and results. In *European Conference on Computer Vision (ECCV) Workshops*, 2014. http://gesture.chalearn.org/2013-multi-modal-challenge/data-2013-challenge.

[27] G. Fanelli, J. Gall, H. Romsdorfer, T.Weise, and L. V. Gool. A 3-D audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 2010. http://www.vision.ee.ethz.ch/~gfanelli/head_pose/head_forest.html#db.

[28] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, 2011. http://www.vision.ee.ethz.ch/datasets/b3dac2.en.html.

[29] M. Firman and S. Julier. 'Misspelled' visual words in unsupervised range data classification: The effect of noise on classification performance. In *Intelligent Robots and Systems (IROS)*, 2011.

[30] M. Firman, O. Mac Aodha, S. Julier, and G. Brostow. Structured prediction of unobserved voxels from a single depth image. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. (To appear). http://visual.cs.ucl.ac.uk/pubs/depthPrediction/.

[31] M. Firman, D. Thomas, S. Julier, and A. Sugimoto. Learning to discover objects in RGB-D images using correlation clustering. In *Intelligent Robots and Systems (IROS)*, 2013.

[32] M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan. Example-based synthesis of 3D object arrangements. *ACM Transactions on Graphics*, 2012.

[33] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *Human Factors in Computing Systems (CHI)*, 2012. http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/.

[34] D. Freeman, R. Jota, D. Vogel, D. Wigdor, and R. Balakrishnan. A dataset of naturally occurring, whole-body background activity to reduce gesture conflicts. *arXiv:1509.06109*, 2015. http://www.dgp.toronto.edu/%CB%9Cdustin/backgroundactivity/.

[35] R. Garg, A. Roussos, and L. Agapito. Robust trajectory-space TV-L1 optical flow for non-rigid sequences. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2011. http://www0.cs.ucl.ac.uk/staff/lagapito/subspace_flow/.

[36] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, 2013. http://gvv.mpi-inf.mpg.de/projects/MonFaceCap/.

[37] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, and F. F. Revuelta. Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using Kinect. In *Technologies for Active and Assisted Living (TechAAL)*, 2015. http://www.tlc.dii.univpm.it/blog/databases4kinect.

[38] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, J. Wahslen, I. Orhan, and T. Lindh. Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In *ICT Innovations 2015, Workshop ELEMENT*, 2015. http://www.tlc.dii.univpm.it/blog/databases4kinect.

[39] S. Gasparrini, E. Cippitelli, S. Spinsante, and E. Gambi. A depth-based fall detection system using a Kinect® sensor. *Sensors*, 2014. http://www.tlc.dii.univpm.it/blog/databases4kinect.

[40] M. Gschwandtner, R. Kwitt, and A. Uhl. BlenSor: Blender sensor simulation toolbox. In *Advances in Visual Computing: 7th International Symposium, (ISVC)*, 2011.

[41] R. Guo and D. Hoiem. Support surface prediction in indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2013. http://aqua.cs.uiuc.edu/site/projects/scenemodel.html.

[42] Y. Guo, J. Zhang, M. Lu, J. Wan, and Y. Ma. Benchmark datasets for 3D computer vision. In *Industrial Electronics and Applications (ICIEA)*, 2014.

[43] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. SceneNet: Understanding real world indoor scenes with synthetic data. *arXiv:1601.05511*, 2016. http://robotvault.bitbucket.org/.

[44] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *International Conference on Robotics and Automation (ICRA)*, 2014. http://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html.

[45] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In *Asian Conference on Computer Vision (ACCV)*, 2012. http://campar.in.tum.de/Main/StefanHinterstoisser.

[46] R. Høg, P. Jasek, C. Rofidal, K. Nasrollahi, T. Moeslund, and G. Tranchet. An RGB-D database using Microsofts Kinect for Windows for face detection. In *International Conference on Signal Image Technology & Internet Based Systems*, 2012. http://www.vap.aau.dk/rgb-d-face-database/.

[47] Y. Hsiao, J. Sanchez-Riera, T. Lim, K. Hua, and W. Cheng. LaRED: a large RGB-D extensible hand gesture dataset. In *Multimedia Systems Conference*, 2014. http://mclab.citi.sinica.edu.tw/dataset/lared/lared.html.

[48] T. Hu, X. Zhu, W. Guo, and K. Su. Efficient interactions recognition through positive action based representation. *Mathematical Problems in Engineering*, 2013. http://www.lmars.whu.edu.cn/prof_web/zhuxinyan/DataSetPublish/dataset.html.

[49] I.Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the ChaLearn gesture challenge 2012. In *Advances in Depth Image Analysis and Applications*, 2012. http://gesture.chalearn.org/data.

[50] A. Ikkala, J. Pajarinen, and V. Kyrki. Benchmarking RGB-D segmentation: Toy dataset of complex crowded scenes. In *Computer Vision Theory and Applications (VISAPP)*, 2016. http://irobotics.aalto.fi/software-and-data/toy-dataset.

[51] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Pattern Analysis and Machine Intelligence (PAMI)*, 2014. http://vision.imar.ro/human3.6m.

[52] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-D object dataset: Putting the Kinect to work. In *International Conference on Computer Vision (ICCV)Workshop on Consumer Depth Cameras in Computer Vision*, 2011. http://kinectdata.com/.

[53] M. Karg and A. Kirsch. A human morning routine dataset. In *Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2014. Extended Abstract http://tinyurl.com/zzsbr4j.

[54] A. Kasper, Z. Xue, and R. Dillmann. The KIT object models database: An object model database for object recognition, localization and manipulation in service robotics. *International Journal of Robotics Research*, 2012. http://his.anthropomatik.kit.edu/objectmodels/ [Link broken as of March 2016].

[55] D. Kastaniotis, I. Theodorakopoulos, C. Theoharatos, G. Economou, and S. Fotopoulos. A framework for gait-based recognition using Kinect. *Pattern Recognition Letters*, 2015. http://www.upcv.upatras.gr/personal/kastaniotis/datasets.html.

[56] B.-s. Kim, P. Kohli, and S. Savarese. 3D scene understanding by voxel-CRF. In *International Conference on Computer Vision (ICCV)*, 2013.

[57] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3D point clouds for indoor scenes. In *Neural Information Processing (NIPS)*, 2011. http://pr.cs.cornell.edu/sceneunderstanding/data/data.php.

[58] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. *International Journal of Robotics Research (IJRR)*, 2013. http://pr.cs.cornell.edu/humanactivities/data.php.

[59] A. Kurakin, Z. Zhang, and Z. Liu. A real-time system for dynamic hand gesture recognition with a depth sensor. In *EUSIPCO*, 2012. http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/.

[60] B. Kwolek and M. Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 2014. http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html.

[61] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3D scene labeling. In *International Conference on Robotics and Automation (ICRA)*, 2014. http://rgbd-dataset.cs.washington.edu/dataset/rgbd-scenes-v2/.

[62] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *International Conference on Robotics and Automation (ICRA)*, 2011. http://rgbd-dataset.cs.washington.edu/dataset/rgbd-scenes/.

[63] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labelling in 3D scenes. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[64] D. Leightley, M. Yap, J. Coulson, Y. Barnouin, and J. McPhee. Benchmarking human motion analysis using Kinect One: an open source dataset. In *,IEEE International Conference by Asia-Pacific Signal and Information Processing Association*, 2015. http://k3da.leightley.com/.

[65] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Workshop on CVPR for Human Communicative Behavior Analysis*, 2010. http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/.

[66] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. http://web.ing.puc.cl/~ialillo/ActionsCVPR2014/.

[67] A. Liu, W. Nie, Y. Su, L. Ma, T. Hao, and Z. Yang. Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing*, 2015. http://media.tju.edu.cn/tju_dataset.html.

[68] A. Liu, Z. Wang, W. Nie, and Y. Su. Graph-based characteristic view set extraction and matching for 3D model retrieval. *Information Sciences*, 2015. http://media.tju.edu.cn/mvred/dataset1.html. Under review.

[69] L. Liu and L. Shao. Learning discriminative representations from RGB-D video data. In *International Joint Conference on Artificial Intelligence*, 2013. http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm.

[70] M. Luber, L. Spinello, and K. O. Arras. People tracking in RGB-D data with on-line boosted target models. In *Intelligent Robots and Systems (IROS)*, 2011. http://www2.informatik.uni-freiburg.de/~spinello/RGBD-dataset.html.

[71] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and Kinect devices. In *International Conference on Image Processing (ICIP)*, 2014. http://lttm.dei.unipd.it/downloads/gesture/index.html.

[72] J. Martínez-Gómez, I. García-Varea, M. Cazorla, and V. Morell. ViDRILO: The visual and depth robot indoor localization with objects information dataset. *International Journal of Robotics Research*, 2015. http://www.rovit.ua.es/dataset/vidrilo/.

[73] J. Mason, B. Marthi, and R. Parr. Object disappearance for object discovery. In *Intelligent Robots and Systems*

(IROS), 2012. http://wiki.ros.org/Papers/IROS2012_Mason_Marthi_Parr.

[74] S. Meister, S. Izadi, P. Kohli, M. Hämmerle, C. Rother, and D. Kondermann. When can we use KinectFusion for ground truth acquisition? In *Intelligent Robots and Systems (IROS) Workshop on Color-Depth Camera Fusion in Robotics*, 2012. http://hci.iwr.uni-heidelberg.de//Benchmarks/document/kinectFusionCapture/.

[75] S. Meister, R. Nair, and D. Kondermann. Simulation of time-of-flight sensors using global illumination. In *Vision, Modeling, and Visualization Workshop*, 2013.

[76] A. Memo, L. Minto, and P. Zanuttigh. Exploiting silhouette descriptors and synthetic data for hand gesture recognition. In *STAG: Smart Tools & Apps for Graphics*, 2015. http://lttm.dei.unipd.it/downloads/gesture/index.html.

[77] A. Mian, M. Bennamoun, and R. Owens. 3D model-bsed object recognition and segmentation in cluttered scenes. *Pattern Analysis and Machine Intelligence (PAMI)*, 2006. http://www.csse.uwa.edu.au/~ajmal/recognition.html.

[78] R. Min, N. Kose, and J.-L. Dugelay. KinectFaceDB: A Kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 2014. http://rgb-d.eurecom.fr/.

[79] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. In *International Conference on Computer Vision (ICCV)*, 2005.

[80] M. Munaro, A. Basso, A. Fossati, L. V. Gool, and E. Menegatti. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *International Conference on Robotics and Automation (ICRA)*, 2014. http://robotics.dei.unipd.it/reid/index.php/8-dataset/5-overview-iaslab.

[81] M. Munaro, F. Basso, and E. Menegatti. People tracking within groups with RGB-D data. In *Intelligent Robots and Systems (IROS)*, 2012. http://www.dei.unipd.it/~munaro/KTP-dataset.html.

[82] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE Workshop on Consumer Depth Cameras for Computer Vision in conjunction with ICCV*, 2011. http://adsc.illinois.edu/sites/default/files/files/ADSC-RGBD-dataset-download-instructions.pdf.

[83] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *Winter Conference on Applications of Computer Vision (WACV)*, 2013. http://tele-immersion.citris-uc.org/berkeley_mhad.

[84] F. Pomerleau, S. Magnenat, F. Colas, M. Liu, and R. Siegwart. Tracking a depth camera: Parameter exploration for fast ICP. In *Intelligent Robots and Systems (IROS)*, 2011. http://projects.asl.ethz.ch/datasets/doku.php?id=Kinect:iros2011Kinect.

[85] C. Rennie, R. Shome, K. E. Bekris, and A. F. D. Souza. A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 2016. http://www.pracsyslab.org/rutgers_apc_rgbd_dataset.

[86] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze. Segmentation of unknown objects in indoor environments. In *Intelligent Robots and Systems (IROS)*, 2012. http://www.acin.tuwien.ac.at/?id=289.

[87] S. Salti, F. Tombari, and L. D. Stefano. SHOT: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 2014. http://www.vision.deis.unibo.it/research/80-shot.

[88] A. Schmidt, M. Fularz, M. Kraft, A. Kasiski, and M. Nowicki. An indoor RGB-D dataset for the evaluation of robot navigation algorithms. In *Advanced Concepts for Intelligent Vision Systems*. Springer, 2013. http://www.vision.put.poznan.pl/?p=70.

[89] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Computer Vision and Pattern Recognition (CVPR) workshops*, 2013. http://www.micc.unifi.it/resources/datasets/florence-3d-actions-dataset/.

[90] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[91] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. http://research.microsoft.com/en-us/projects/7-scenes/.

[92] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *International Conference on Computer Vision (ICCV) Workshops*, 2011. http://cs.nyu.edu/~silberman/datasets/nyu_depth_v1.html.

[93] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, 2012. http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

[94] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel. BigBIRD: A large-scale 3D database of object instances. In *International Conference on Robotics and Automation (ICRA)*, 2014. http://rll.berkeley.edu/bigbird/.

[95] S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. http://rgbd.cs.princeton.edu/.

[96] S. Song and J. Xiao. Tracking revisited using RGBD camera: Unified benchmark and baselines. In *International Conference on Computer Vision (ICCV)*, 2013. http://tracking.cs.princeton.edu/dataset.html.

[97] L. Spinello and K. O. Arras. People detection in RGB-D data. In *Intelligent Robots and Systems (IROS)*, 2011.

http://www2.informatik.uni-freiburg.de/~spinello/RGBD-dataset.html.

[98] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp*, 2013. http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/.

[99] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Intelligent Robots and Systems (IROS)*, 2012. http://vision.in.tum.de/data/datasets/rgbd-dataset.

[100] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011. http://pr.cs.cornell.edu/humanactivities/data.php.

[101] W. Susanto, M. Rohrbach, and B. Schiele. 3D object detection with multiple Kinects. In *European Conference on Computer Vision (ECCV)*, 2012. http://tinyurl.com/hlmwga7.

[102] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos. Pose-based human action recognition via sparse representation in dissimilarity space. *J. Vis. Commun. Image R.*, 2013. http://www.upcv.upatras.gr/personal/kastaniotis/datasets.html.

[103] F. Tombari, L. D. Stefano, and S. Giardino. Online learning for automatic segmentation of 3D data. In *Intelligent Robots and Systems (IROS)*, 2011. http://vision.deis.unibo.it/fede/kinectDataset.html.

[104] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*, 2012.

[105] C. van Gemeren, R. T. Tan, R. Poppe, and R. C. Veltkamp. Dyadic interaction detection from pose and flow. In *European Conference on Computer Vision (ECCV)*, 2014. http://www.projects.science.uu.nl/shakefive/.

[106] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/.

[107] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. http://users.eecs.northwestern.edu/~jwa368/my_data.html.

[108] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3D human activity recognition with reconfigurable convolutional neural networks. In *ACM International Conference on Multimedia*, 2014. http://vision.sysu.edu.cn/projects/3d-activity/.

[109] O. Wasenmüller, M. Meyer, and D. Stricker. CoRBS: Comprehensive RGB-D benchmark for SLAM using Kinect v2. In *Winter Conference on Applications of Computer Vision (WACV)*, 2016. http://corbs.dfki.uni-kl.de/.

[110] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-Patch: Unsupervised understanding of actions and relations. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. http://watchnpatch.cs.cornell.edu/.

[111] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition (CVPR) workshops*, 2012. http://cvrc.ece.utexas.edu/KinectDatasets/HOJ3D.html.

[112] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *International Conference on Computer Vision (ICCV)*, 2013. http://sun3d.cs.princeton.edu/.

[113] N. Xu, A. Liu, W. Nie, Y. Wong, F. Li, and Y. Su. Multi-modal & multi-view & interactive benchmark dataset for human action recognition. In *ACM International Conference on Multimedia*, 2015. http://media.tju.edu.cn/m2i.html.

[114] Z. Yang, L. Zicheng, and C. Hong. RGB-depth feature for 3D human activity recognition. *China Communications*, 2013. http://www.uestcrobot.net/en/?q=download.

[115] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision (ACCV)*, 2014. http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/.

[116] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Computer Vision and Pattern Recognition (CVPR) workshops*, 2012. http://www3.cs.stonybrook.edu/~kyun/research/kinect_interaction/.

[117] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tanga. RGB-D-based action recognition datasets: A survey. *arXiv:1511.07041v2*, 2015. http://robotvault.bitbucket.org/.

[118] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime Faces: High-resolution capture for modeling and animation. In *ACM SIGGRAPH Proceedings*, 2004. http://grail.cs.washington.edu/projects/stfaces/.

[119] Q. Zhang, X. Song, X. Shao, H. Zhao, and R. Shibasaki. Category modeling from just a single labeling: Use depth information to guide the learning of 2D models. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. http://shiba.iis.u-tokyo.ac.jp/song/?page_id=343.

[120] Q.-Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics*, 2013. http://qianyi.info/scenedata.html.

[121] M. Zollhöfer, M. Nießner, S. Izadi, C. Rhemann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (TOG)*, 2014.