

3D Datasets

Thiago João Miranda Baldivieso

Email: thiagojmb@ime.eb.br

Date: 21/11/2019

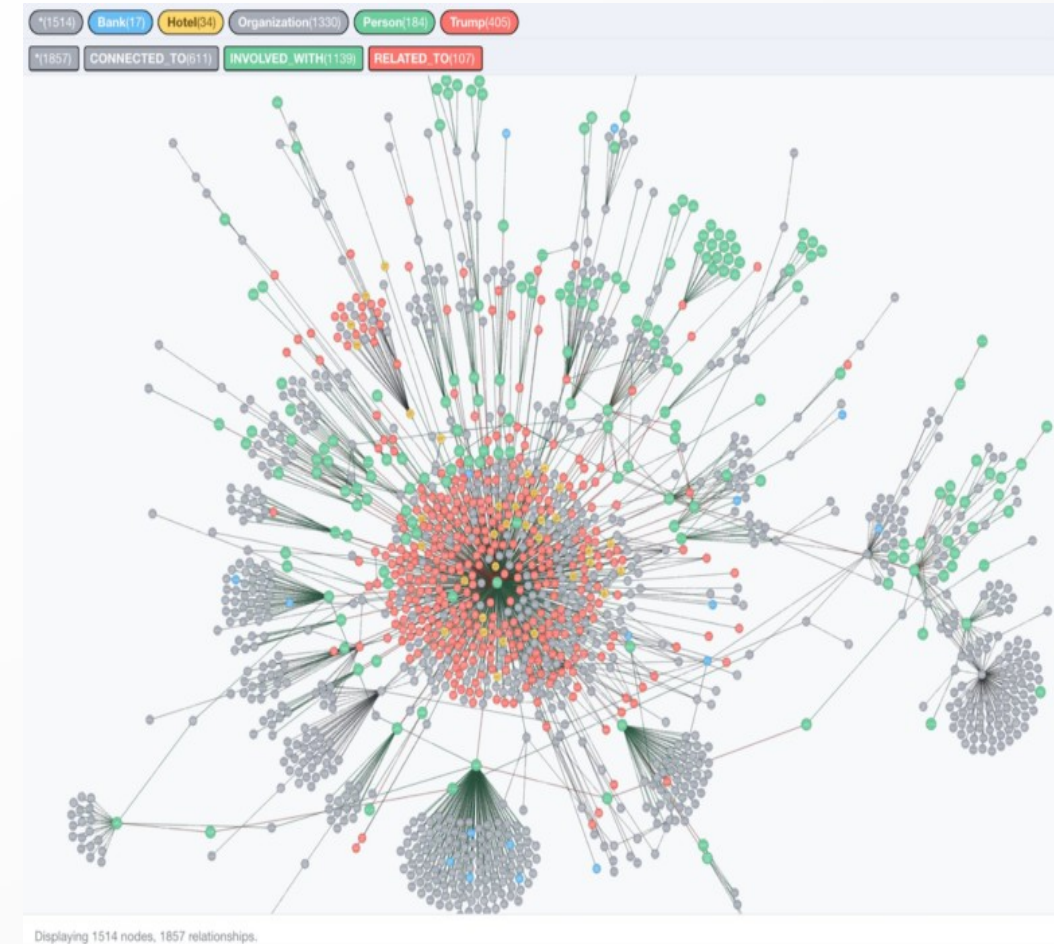
Course: Image Processing 2019 | Prof: Luiz Velho

Sumário e Referências

- Principal paper: RGBD Datasets: Past, Present and Future, 2016, Firman Michael [[source](#)]
- Introdução
- Aquisição 3D
- Evolução de Datasets
- Tipos de Hardware RGB-D
- Exemplos de Aplicação em vídeo: Object Labeling in 3D Scenes, Unsupervised Feature Learning for 3D Scene Labeling , SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels
- Rastreamento de Câmera | Reconstrução de cenas / objetos
- Problema de inferir pose 6GDL
- Datasets de objetos | Rotulação semântica | Tracking
- Datasets de atividades / gestos / faces / ações
- Futuro
- Conclusão

O que é um dataset?

- Trata-se de um conjunto de dados relacionados composto por elementos separados mas podem ser manipulados como uma unidade por um computador.
- Ele é finito e tem uma função e característica específica.
- Datasets necessitam ser autocontidos, ou seja, todas as informações necessárias para responder questões de análise devem estar presentes.



TIMELINE

Each circle represents a thesis in order of publication
The color means the degree

Master thesis (red)
Ph.D. thesis (blue)

AUTHOR'S NETWORK

Each circle represents a different person and the lines connecting people represent their interaction

Color represents a role

Student (white)
Advisor (yellow)

The size represents the number of contributions

only master thesis (small white)
master and Ph.D. thesis (medium white)
advised many works (large yellow)

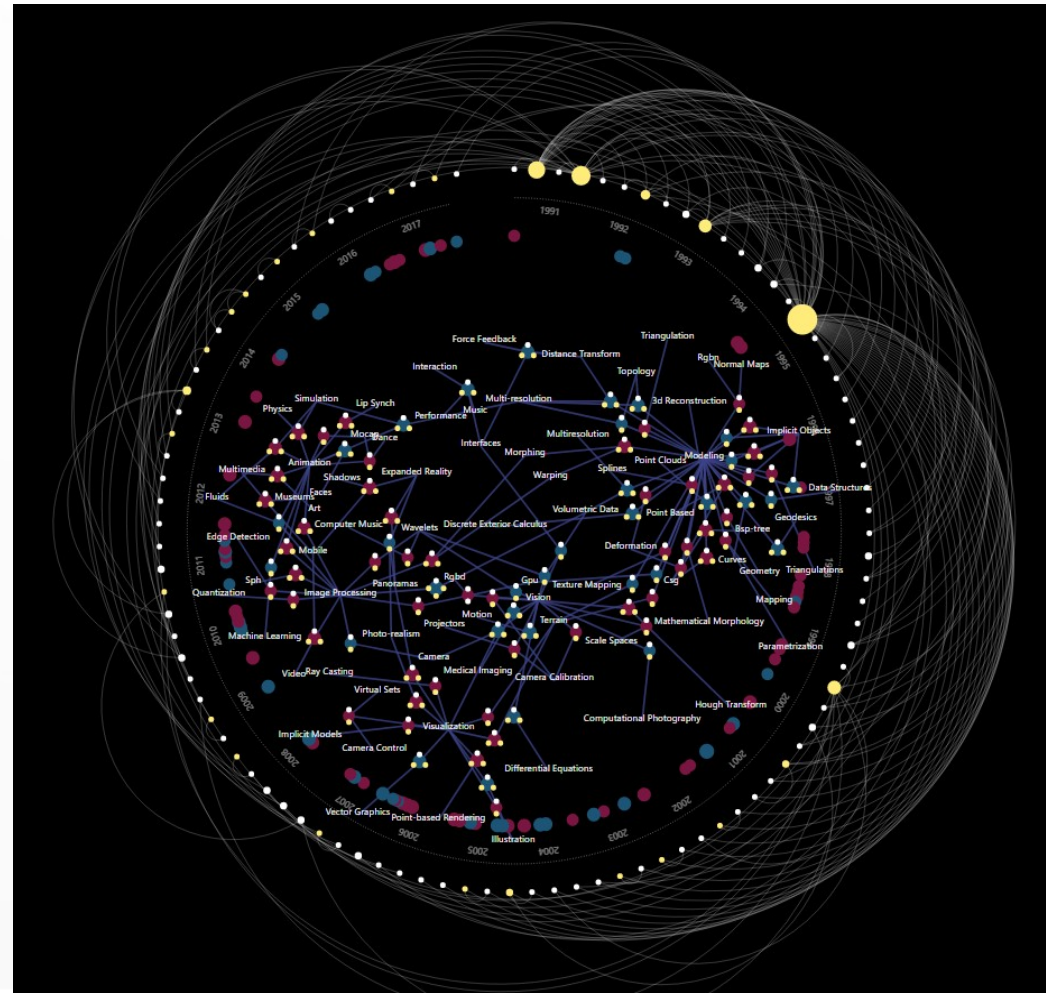
CLUSTER

Each circle represents a thesis clustered by area of subject.

The color means the degree and the dots around the circle the number of people involved

Master thesis with one advisor (red circle with 1 dot)
Ph.D. thesis with two advisors (blue circle with 2 dots)

Topology
Modeling
Vision



Aquisição 3D

- **Triangulação à laser:** O dispositivo emite um padrão de laser sobre o objeto e um sensor ótico, calibrado com o emissor de laser, identifica a posição desse padrão e calcula a informação de profundidade por simples triangulação. (+ Custo, Preciso, probl. refletância)
- **Tempo de percurso:** O sistema calcula o tempo de ida e volta dos impulsos de luz emitida pelo dispositivo para determinar a distância ao longo da superfície da cena. (Dados grandes, Necess. Retificação, Ruídos)
- **Visão estéreo:** O sistema é composto por dois sensores óticos calibrados que captura imagens simultaneamente. Distância entre os pontos correspondentes em ambas as imagens indicam o quão distante estão do ponto de vista do sistema. (- Custo, textura, - resolução e precisão em ambientes não controlados, possib. tempo-real)
- **Forma a partir do movimento:** O sistema gera a informação 3D a partir de uma sequência de imagens de um mesmo objeto, geralmente com uma câmera de vídeo. Features são rastreadas ao longo de frames, a diferença na movimentação dos pontos baseia a criação de um espaço tridimensional que determina a posição relativa entre o objeto e a câmera.

Aquisição 3D

- **Forma a partir da silhueta:** Com uma coleção de imagens de um objeto com diferentes pontos de vista, a geometria do objeto é deduzida de sua silhueta.
- **Forma a partir da sombra:** Esta classe de métodos deduz a geometria de um objeto a partir de uma coleção de imagens sob iluminação variável.
- **Fotometria:** Semelhante ao método de forma a partir da sombra, mas assume um ambiente com fontes de luz e câmeras pré-calibrados. (Usando ambiente e equip. especiais obtém precisão submilimétrica parecido ao scanner laser)
- **Forma a partir do foco:** O sistema requer uma coleção de imagens a partir do mesmo ponto de vista da captura com diferentes ajustes de foco. (Necess. Lentes especiais ou microscópio profissional)

Aquisição 3D

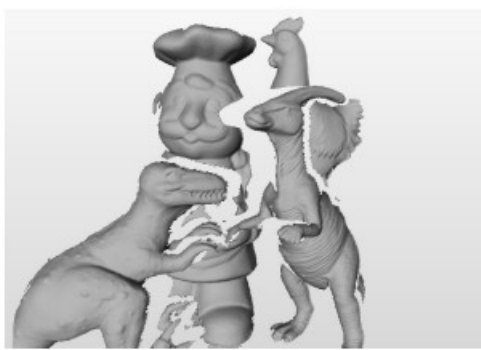
- **Digitalização por contato:** Esta classe de método adota geralmente um braço robótico com variado grau de liberdade. O braço passa sobre a superfície do objeto tocando com a ponta de uma agulha que tem sua posição constantemente monitorada. (+ tempo, + preciso, contato com o objeto, limitado a áreas acessíveis)
- **Topografia:** O sistema é composto de uma estação total geodésica e é adequado para objetos / ambientes de grande escala. (+ tempo)
- **Luz estruturada:** O dispositivo projeta um conjunto específico de padrões de luz e extrai a geometria das distorções destes padrões ao longo da superfície digitalizada. Os dispositivos nesta categoria têm diferentes tecnologias. (Captura textura; probl. refletância e transparência pode afetar medições; geralmente sensíveis a iluminação ambiente)

Tabela comparativa de métodos de aquisição

Tecnologias	Qualidade	Preço	Portabilidade	Tempo de aquisição
Triangulação à laser	+1	-1	0	0
Tempo de percurso	-1	+1	+1	+1
Luz estruturada(RGB-D)	-1	+1	+1	+1
Visão estéreo	-1	0	+1	+1
Forma a partir do movimento	-1	+1	+1	+1
Forma a partir da silhueta	0	+1	+1	0
Forma a partir da sombra	0	+1	+1	0
Fotometria	+1	0	-1	0
Forma a partir do foco	+1	-1	-1	-1
Digitalização por contato	+1	0	-1	-1
Topografia	-1	-1	-1	-1

- Tabela classifica as tecnologias de aquisição segundo as características de Remondino et al. +1 significa que a tecnologia é favoravelmente comparada com as demais tecnologias, 0 significa que ela é neutra, e -1 significa um comparativo desfavorável.

RGBD Dataset: Passado, Presente e Futuro



- **Passado:** Antes da Microsoft Kinect, a maioria dos conjuntos de dados de profundidade eram pequenos e capturados em laboratório.
- **Presente:** Agora desfrutamos de dados RGBD de cenas dinâmicas e estáticas do mundo real, com uma gama de condições de rotulação e de captura.
- **Futuro:** Vamos poder antecipar escaneamentos de cena estática e cenas dinâmicas como a fusão de geometrias, explorando melhorias nos algoritmos de reconstrução.

Evolução de datasets

- Inicialmente os datasets eram focados em imagens estáticas, muitas vezes de objetos únicos ou pequenas cenas.
- À medida que o campo amadurece, vemos pesquisas sendo implementadas na criação de datasets RGBD maiores e mais ambiciosos, e a quantidade liberada a cada ano não mostra sinais de diminuição.
- Rótulos semânticos foram propagados através de vídeos.
- Reconstrução densa foi explorada para capturar as superfícies de objetos inteiros.
- Algoritmos de cena generativos foram usados para criar dados sintéticos plausíveis.
- Também vemos novos rótulos aplicados aos dados existentes e versões anteriores sendo recompiladas em novas empreitadas.

Por que RGB-D?

Remove a necessidade de repetir a captura de dados. Mais importante, eles fornecem transparência na apresentação de resultados e permitir que as pontuações sejam comparadas no mesmo dados por diferentes pesquisadores.

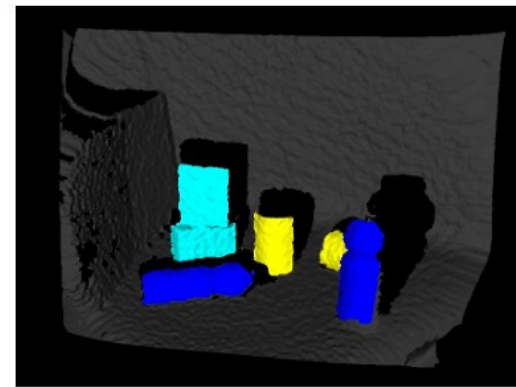
O registro de nuvens de pontos capturadas por sensores de profundidade é uma importante etapa em aplicações de reconstrução 3D. Em diversos casos como localização e mapeamento para robótica ou realidade aumentada para entretenimento, o registro deve ser realizado não só com precisão estrita, como também na frequência de dados de aquisição do sensor.



Scene (2D)



Scene (3D)



GroundTruth (3D)

Figura obtida de [\[source\]](#)

Por que RGB-D?

- Dispositivos similares ao Kinect são equipamentos leves, flexíveis e com baixo custo quando comparado a sistemas laser scanning terrestre e câmeras de distância 3D.
- Três sensores: dois sensores CMOS (Complementary Metal-Oxide-Semiconductor) que registram energia eletromagnética na faixa do espectro correspondente ao visível (RGB) e infravermelho próximo (IR); e um emissor LASER (Light Amplification by Stimulated Emission of Radiation) infravermelho.
- Os sensores RGB e IR capturam cenas com 640 x 480 pixels em uma taxa de 30 quadros por segundo (fps), sendo que cada quadro (frame) capturado pode conter até 300.000 pontos.

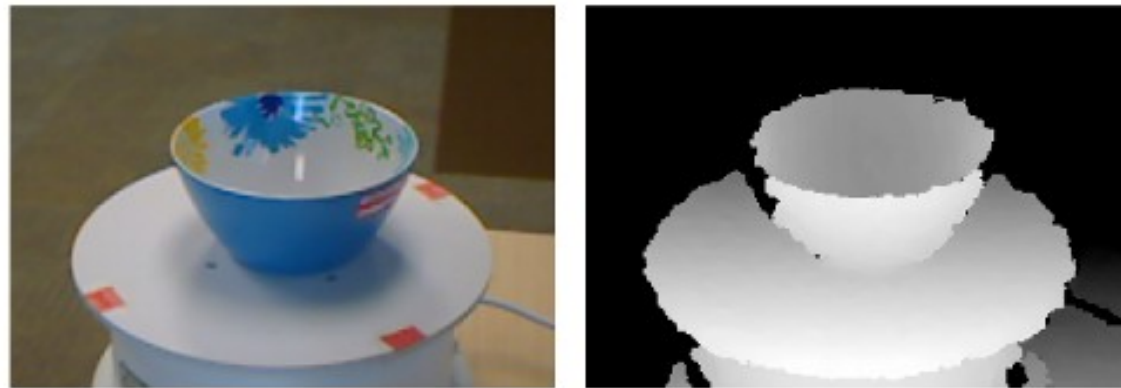
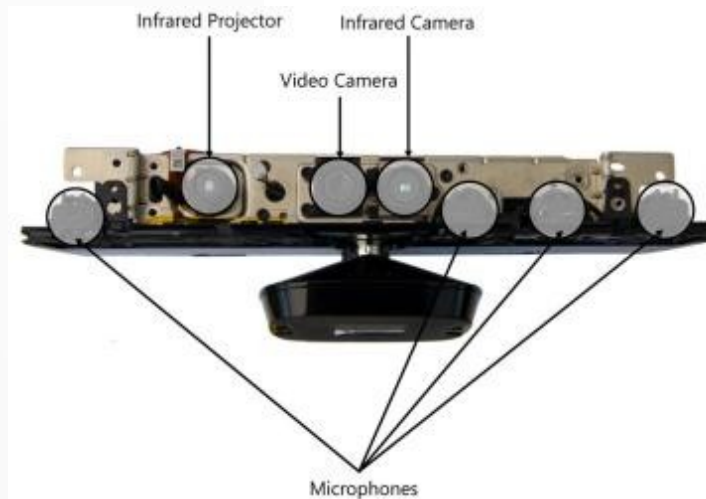


Figura obtida de [[source](#)]

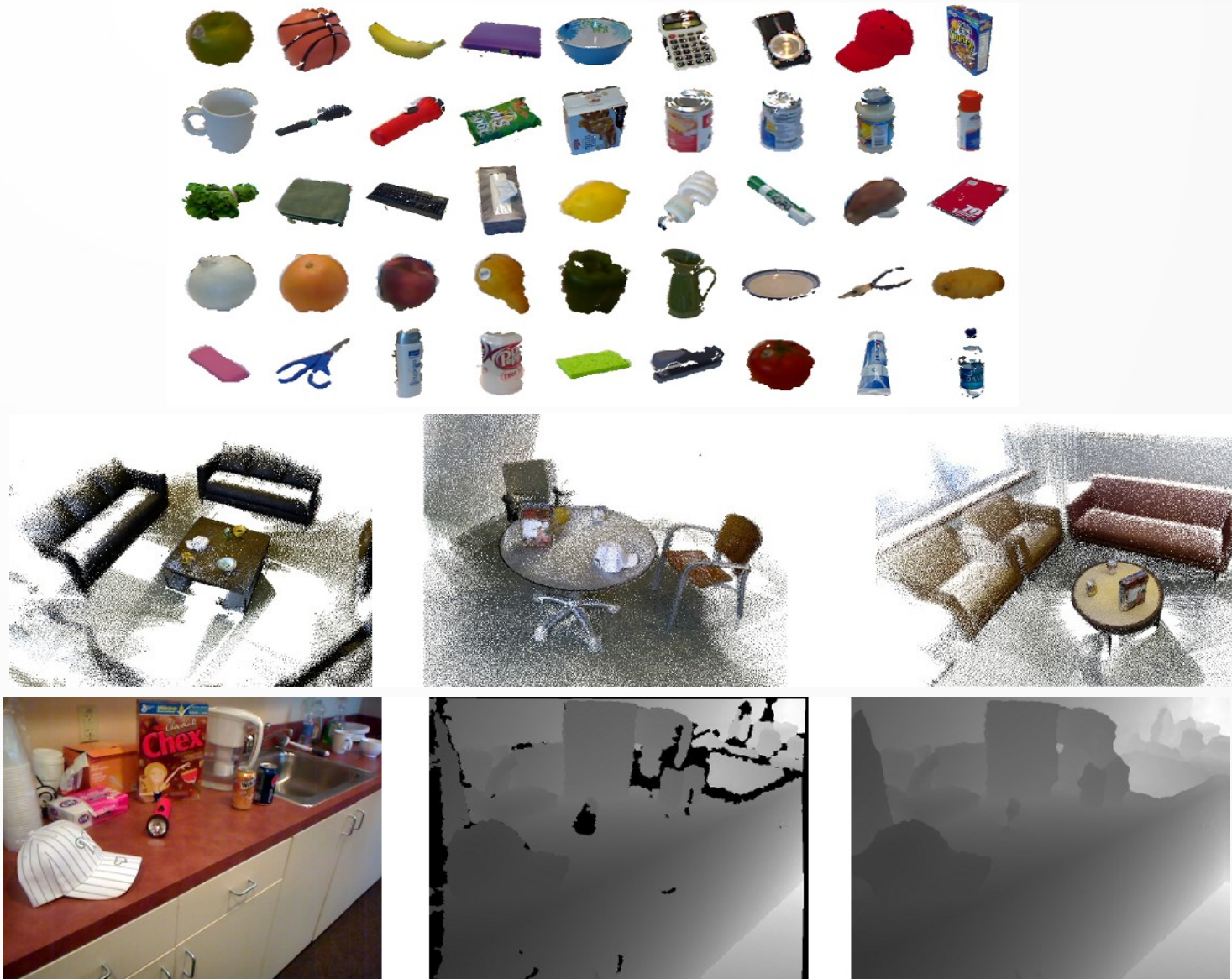
Hardware RGB-D

- Kinect v1, Kinect v2, Intel RealSense, Asus Xtion Live Pro



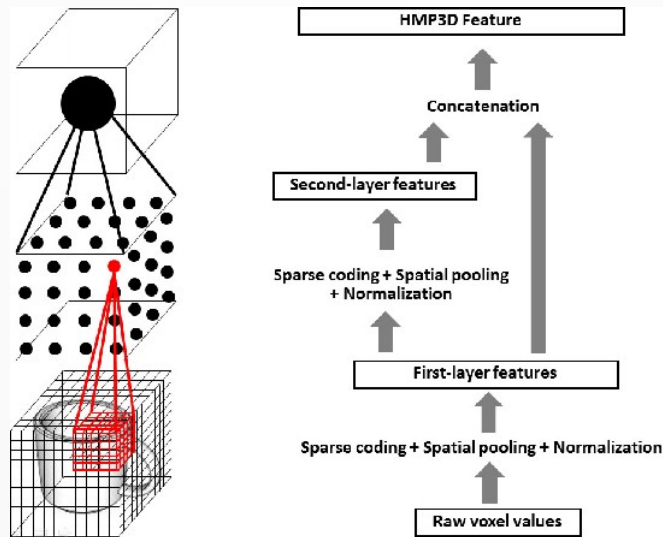
Dataset de captura de objetos simples

- RGBD Object dataset
- Apresentado: ICRA 2011
- Hardware: Kinect v1
- Descrição: 300 instâncias de objetos domésticos, 51 categorias. 250,000 frames no total
- Rotulação: Categorias e instância do rótulo. Incluindo máscaras auto-geradas, mas sem dados exatos da pose 6-GDL.



Aplicação com RGB-D

- **Vídeo 01** – Abordagem baseada em visualização para rotular objetos em cenas 3D reconstruídas a partir de vídeos RGB-D (Kinect).
- **Vídeo 02** – Deep-learning - Demonstra a combinação de *sliding window* HMP e HMP3D *voxel features* em uma estrutura MRF (Markov Random Fields) para rotular objetos em cenas 3D reconstruídas a partir de vídeos RGB-D.

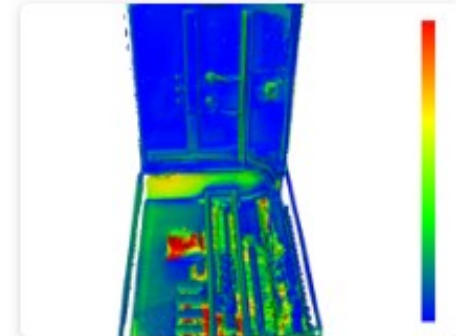
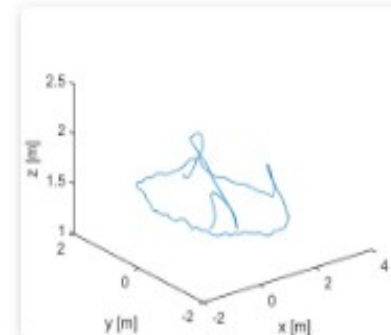


Rastreamento de câmera

- Um dos **principais avanços** trazidos pelas câmeras de profundidade para consumidor final foi o **rastreamento de câmeras e na reconstrução densa**. O ground truth da poses de câmera é necessário para validação desses algoritmos e é difícil de adquirir, pois exige hardware externo.
- Para rastreamento de câmera, o *benchmark* TUM tornou-se um padrão de fato para avaliação, como *ground truth* de um sistema de rastreamento de movimento e uma variedade de cenas e movimentos de câmera.
- Alguns datasets utilizam a verificação de pose de câmera manualmente usando o próprio kinect, porém esses dados são indicados apenas para tarefas cuja ordem de magnitude é mais difícil do que o rastreamento, como relocalização de câmera ou previsão de ocupação de voxel.

Reconstrução de cenas

- A reconstrução de cena raramente é avaliada diretamente, pois um bom rastreamento da câmera geralmente corresponde a uma boa reconstrução e os caminhos da câmera são mais fáceis de obter como *ground truth* que superfícies densas.
- **Wasenmuller et al.** criaram um conjunto de dados contendo movimentos de câmera ground truth e reconstruções de cena a partir de um scanner a laser. Este é um conjunto de dados do mundo real que possui com esses dois dados, embora as cenas sejam menos diversas do que Firman et al.



Reconstrução de Objetos

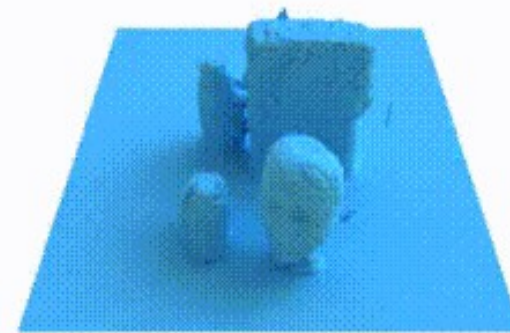
- Firman et al. têm um conjunto de dados de objetos de mesa digitalizados para que todas as superfícies visíveis sejam observadas na reconstrução. Isso fornece uma base sólida para a tarefa de estimar a ocupação não observada de voxel a partir de uma imagem de profundidade.



Input RGB



Visible Depth



Our Depth Completion

Problema de inferir pose do objeto 6-GDL

- Tarefa é auxiliada pela escala absoluta fornecida por câmeras de profundidade.
- Dado a priori um modelo 3D de um objeto, o objetivo é encontrar a transformação que melhor alinha-o à cena.
- Como no rastreamento da câmera, é difícil obter *ground truth* para esse tipo de desafio, que exige um modelo 3D do objeto e sua pose em cada imagem.
- Uma solução foi fixar os objetos de destino em uma placa de calibração para permitir o rastreamento do *ground truth* usando o canal RGB, enquanto em alguns trabalhos têm as poses manualmente alinhadas.
- Em alguns *datasets*, apresentam objetos com tamanho de tablet. Adquirindo modelos 3D e *ground truth* poses básicas, para objetos maiores é difícil; portanto, trabalhos que tentaram esse problema em um ambiente de escala geralmente encontram um método alternativo de avaliação ou contam com anotações humanas como *ground truth* do terreno.

Dataset de captura de objetos simples

BigBIRD: (Big) Berkeley Instance Recognition Dataset A Large-Scale 3D Database of Object Instances

Arjun Singh, James Sha, Karthik Narayan, Tudor Achim, Pieter Abbeel
bigbird@lists.eecs.berkeley.edu

This is the website for the dataset introduced in the ICRA 2014 publication "A Large-Scale 3D Database of Object Instances." Specifically, for each of (currently) **125 objects**, we provide:

- 600 12 megapixel images, sampling the viewing hemisphere
- 600 registered RGB-D point clouds from a Carmine 1.09 sensor
- Pose information for each of the above images and point clouds
- Segmentation masks for each of the above images (and segmented point clouds)
- Merged point clouds consisting of data from all 600 viewpoints
- Reconstructed meshes from the merged point clouds

Note that some objects, depending on their properties (e.g. transparency) may not have point clouds or meshes. We include them to enable others to develop methods to reconstruct point clouds and/or meshes.

We plan to continuously collect and upload objects and test scenes. As researchers use the data, we will list results and benchmarks here. If you have any results on the data that you would like listed here, please contact us (bigbird@lists.eecs.berkeley.edu).

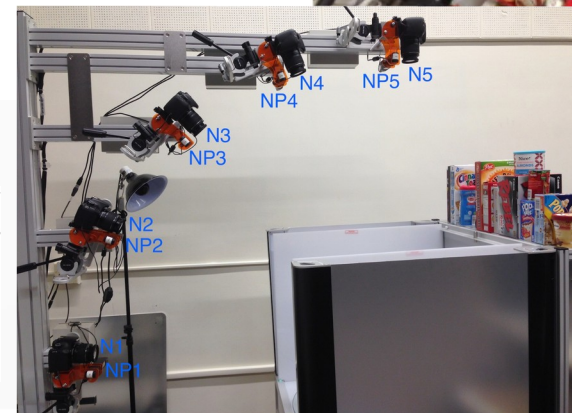

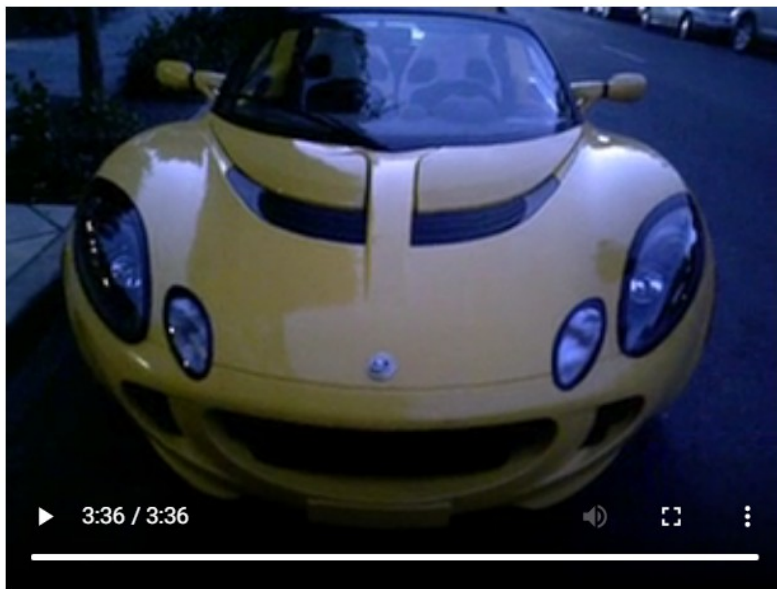


Image	Name	Raw RGB-D	Raw high resolution	Processed
	3m_high_tack_spray_adhesive	RGB-D (.tgz)	High res (.tgz)	Processed (.tgz)

Dataset de captura de objetos simples

- A large dataset of object scans
- Hardware: Apple - PrimeSense Carmine
- Descrição: Mais de 10,000 objetos densamente escaneados e reconstruídos. Dados capturados do mundo real por operadores não técnicos.
- Rotulação: Objetos presentes em cada escaneamento.



Rotulação Semântica



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]

A rotulação semântica nos dá uma compreensão mais geral do mundo.

Observe que pontuação baixa aqui não corresponde a um conjunto de dados pior ou menos útil, pois conjuntos de dados com cenários especialmente construídos podem ser vitais para a comprovação de conceitos, e muitas vezes podem fornecer um *ground truth* de qualidade superior às cenas totalmente naturais.

O subconjunto de 1449 quadros do conjunto de dados **NYUv2** com rótulos semânticos densos tornou-se um padrão para a rotulação de cenas *indoor*. A qualidade e a variedade de etiquetas neste conjunto de dados do mundo real ajudaram a torná-lo um dos mais usados na literatura.

O conjunto de dados SUN3D contraria a modalidade de quadro estático único do NYUv2 com etiquetas de objetos propagadas pelos vídeos do Kinect.

Observamos que todos esses conjuntos de dados semânticos, mesmo aqueles com vídeos, representam um mundo estático. Isso contrasta com o nosso mundo dinâmico, uma área que é explorada por conjuntos de dados projetados para rastreamento.

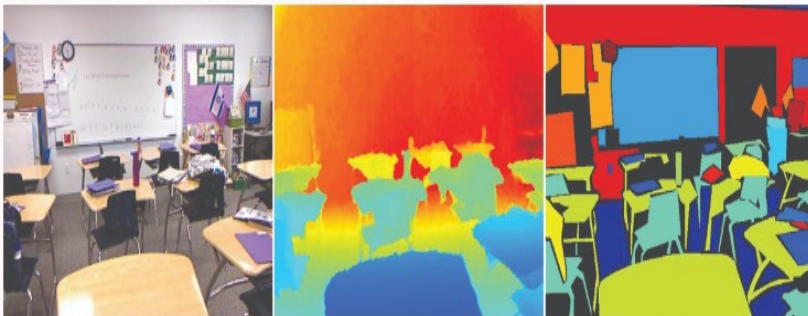


















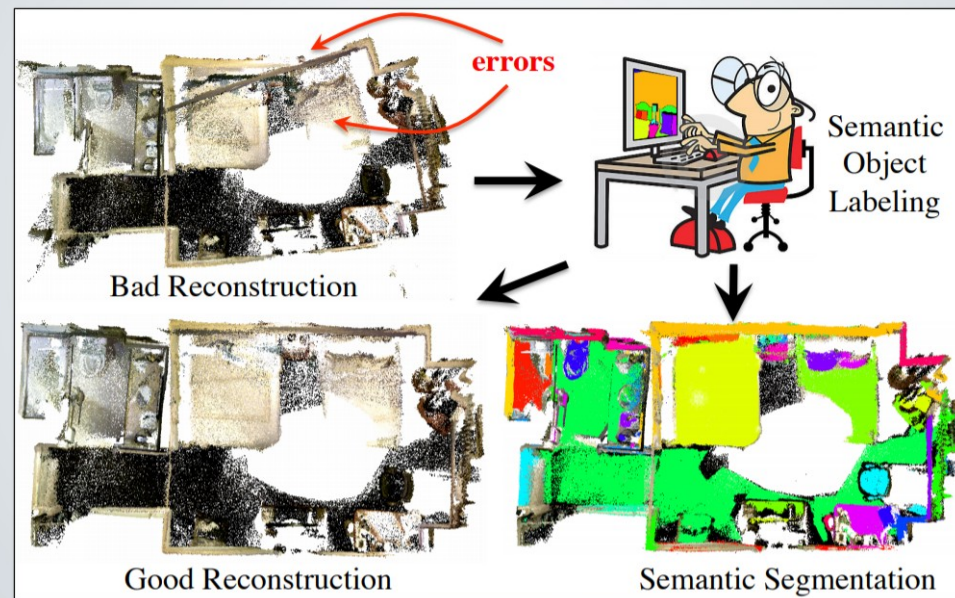
Tabela com Datasets com segmentação semântica

		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

Dataset com rotulação semântica SUN3D

Vídeo 03 -SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels – É feito a rotulação manualmente a cada frame do vídeo, logo após verifica-se a otimização da reconstrução.

Semantic Bundle Adjustment



Tracking - Rastreamento

Datasets de rastreamento apresentam vídeos de ambientes dinâmicos, onde o objetivo é detectar onde um objeto está em cada quadro.

O Princeton Tracking Benchmark contém cenas do mundo real, mas com móveis ou objetos dispostos artificialmente. Possui 100 RGBD videos de movimento de objetos assim como humanos bolas e carros.



Datasets de atividades e gestos
























Dado o caso de uso original do Kinect como um sensor projetado para a interação humana, é inevitável que muitas pesquisas se concentrem no reconhecimento de gestos e atividades de vídeos.













As ações que executadas incluem linguagem de sinais, gestos italianos das mãos e ações diárias comuns, como levantar, beber e ler. Três conjuntos de dados de seres humanos caindo sobre refletem um interesse no uso de sensores RGBD para monitorar humanos vulneráveis em suas vidas diárias.

Os maiores conjuntos de dados de gestos e ações são o desafio do gesto ChaLearn e NTU RGB + D, cada um com cerca de 50.000 vídeos.

Tabela com datasets de gestos

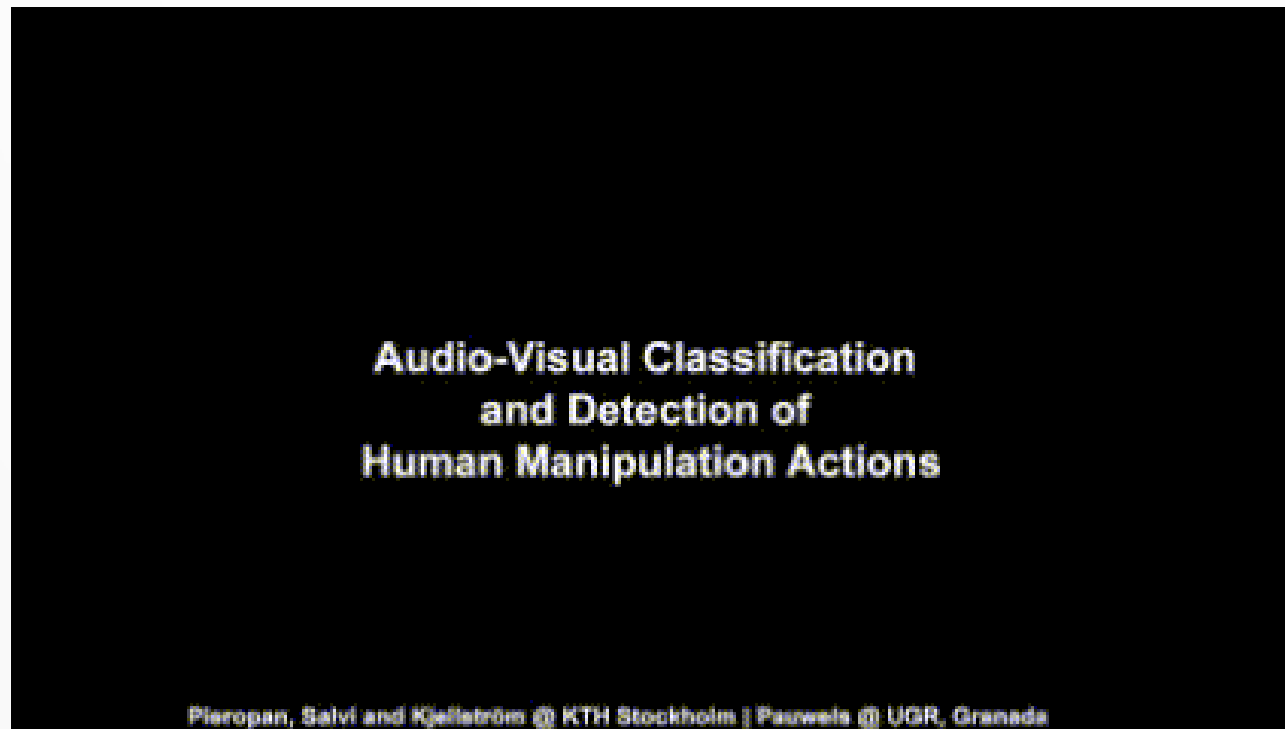
Table 5. Datasets representing activities and gestures

	# Subjects	# Actions	# Videos	Skeleton*	Examples of actions	Year
 MSR Action3D [65]	10	20	567	✓	e.g. <i>high arm wave, side kick, jogging</i>	'10
 RGBD-HuDaAct [82]	30	12	1189		e.g. <i>get up, enter room, stand up, mop the floor</i>	'11
 SBU Kinect Interaction Dataset [116]	7	8	300	✓	Two people interacting e.g. <i>approaching, departing</i>	'12
 ACT4 ² [18]	24	14	6844		4 Kinects filming. Actions: e.g. <i>collapse, reading</i>	'12
 UTKinect-Action [111]	10	10	200	✓	e.g. <i>walk, sit down, stand up, carry, clap hands</i>	'12
 MSRDailyActivity3D [106]	10	16	320	✓	e.g. <i>drink, eat, read book</i>	'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. <i>diving signals and mudras</i>	'12
 Senior Activity Recognition (RGBD-SAR) [114]	30	9	810	✓	Older people performing activities e.g. <i>sit down, eat, walk, stand up</i>	'13
 K3HI [48]	15	8	320	✓	Two humans interacting e.g. <i>approaching, punching</i>	'13
 UPCV action dataset [102]	20	10	400	✓	e.g. <i>walk, wave, scratch head, phone, cross arms</i>	'13
 DML-SmartAction [5]	16	12	932		Continuous recording. e.g. <i>writing, sit down, walk, clean table, stand up</i>	'13
 Florence 3D actions dataset [89]	10	9	215	✓	e.g. <i>wave, drinking, answer phone, clap, stand up</i>	'13
 Cornell activity 60/120 [100, 58]	4	12/10	60/120	✓	e.g. <i>brushing teeth, drinking, talking on couch</i>	'13
 Sheffield Kinect Gesture (SKIG) [69]	6	10	1080		Hand gestures e.g. <i>circle, up-down, come here</i>	'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. <i>jumping, bending, punching</i>	'13
 Manipulation Action Dataset [1]	5	28	140		Manipulation actions e.g. <i>cutting</i> , plus sequences of actions. Semantic segmentation of frames.	'14
 Composable activities dataset [66]	14	16	693	✓	e.g. <i>throw, talk on phone, walk, wave, crouch, punch</i>	'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. <i>mopping, sleeping, finding-objects, chatting</i>	'14
Human3.6M [51]	11	17	- ^b	✓✓	e.g. <i>Discussion, smoking, taking photo</i>	'14

 MSR 3D Online Action [115]	24	7	- ^b		e.g. <i>drinking, eating, using laptop</i>	'14
 Northwestern-UCLA Multiview Action 3D [107]	10	10	- ^b	✓	Three Kinects filming. Actions: e.g. <i>stand up, throw</i>	'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. <i>boxing, one hand wave, forward bend, sit down</i>	'15
 M2I dataset [113]	22	22	1760	✓	Two people interacting, e.g. <i>walk together</i>	'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. <i>wave, boxing</i>	'15
 TST Fall Detection ver. 1/ver. 2 [39, 38]	4/11	2	20/11	✓	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 naturally interacting in semi-natural environment	'15
 K3Da [64]	53	13	?	✓	To assess human health, e.g. <i>leg jump, walking</i>	'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. <i>making drink</i>	'15
 NTU RGB+D [90]	40	60	56,000	✓	e.g. <i>drinking, eating, sneezing, staggering, punching, kicking</i>	'16

Exemplo de Dataset de Atividades

- Manipulation Action Dataset é único no fornecimento de segmentação semântica de objetos à medida que são manipulados.



Dataset de Faces

Os datasets focados no método de aquisição de faces tendem a ser pequenos. O campo foi expandido para incluir conjuntos de dados para reconhecimento de identidade, regressão de pose e aqueles onde as expressões de emoções devem ser inferidas. Como o **ETH Face Pose Range Image DataSet** que possui mais de 10 mil imagens de 20 pessoas, e rotula a posição do nariz e coordena o frame pelo nariz, usa setup de stereo ativo.

À medida que as câmeras de profundidade voltadas para a frente são instaladas em laptops e tablets, espera-se que essa área de pesquisa continue ganhando atenção.



Datasets de ações

Como conjuntos de dados de ações, os datasets são projetados para reconhecimento de ações humanas tipicamente filmam pessoas executando atividades como caminhar, pular, dançar, entre outras.

No entanto, o objetivo agora é reconhecer a identidade, gênero ou outros atributos sobre assuntos, em vez da atividade que eles estão realizando.

Futuro

“Lacunas no mercado” pouco exploradas:

- **Dados Sintéticos:** pouca atenção para problemas de visão com câmeras de profundidade. No entanto, dados artificiais podem oferecer muitas vantagens.
 - *Ground Truth* para tarefas como segmentação, reconstrução, rastreamento e pose de câmera ou objeto é perfeito e está disponível sem necessidade de rotulagem humana dispendiosa.
 - As sequencias podem ser recapturadas com parâmetros cuidadosamente ajustados, por exemplo, desfoque de movimento e alterações de iluminação, para introspecção de algoritmos.
 - Possibilidade de criar cenários difíceis de capturar na vida real, por exemplo acidentes.

Ocupação total do voxel

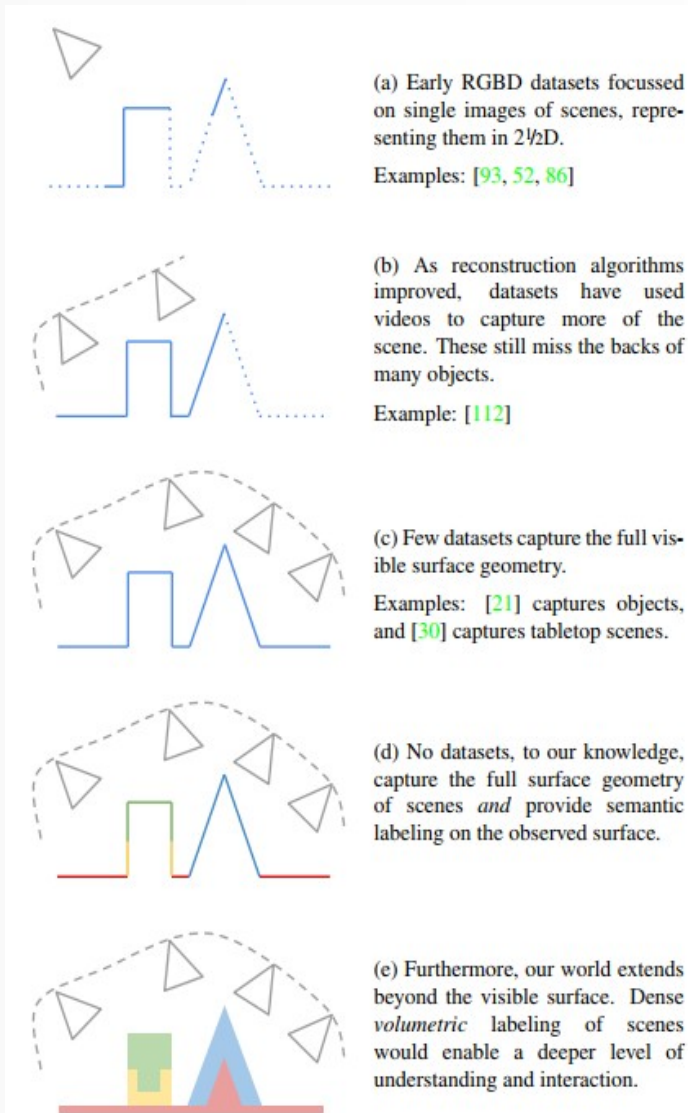


Figure 4. Datasets progress to include more 3D information.

A maioria dos conjuntos de dados semânticos existentes vê o mundo como uma imagem 2.5D, onde apenas as superfícies diretamente vistas de uma posição estática da câmera são visíveis.

A **criação de um conjunto de dados de grandes cenas do mundo real** é deixada como um **desafio aberto**.

A **identificação das superfícies dessas reconstruções densas** permitiria a **segmentação semântica no nível da malha**. Muitas oportunidades seriam oferecidas por conjuntos de dados que fornecem etiquetas nesta forma de reconstrução densa, em vez de imagens ou vídeos.

Além disso, podemos imaginar os benefícios de um algoritmo que poderia segmentar ou rotular semanticamente uma cena no nível de voxel. Para treinar e validar esse sistema, precisaríamos de um conjunto de dados contendo rotulagem semântica de cada voxel em uma cena.

A dificuldade de aplicar essa rotulagem manualmente pode tornar os dados sintéticos necessários para esse problema.

Geometria de cenas dinâmicas

Além de uma única sequência de, não conhecemos conjuntos de dados RGBD capturados de cenas dinâmicas com geometria densa no solo.

Uma opção é usar **malhas deformáveis** fornecidas para conjuntos de dados de faces ou tecidos, que podem ser re-renderizados sinteticamente para fornecer correspondências densas entre os quadros e a re-renderização dados. Conjuntos de dados de humanos com dados de captura de movimento também fornecem uma geometria densa muito esparsa com correspondências.

O desafio aberto para o campo da reconstrução densa é **capturar diretamente um conjunto de dados RGBD de objetos deformados** com o *Ground Truth* da geometria e correspondências entre os quadros.

Conclusão

Descobrimos uma quantidade considerável de conjuntos de dados RGBD disponíveis para uso dos pesquisadores. Embora exista alguma sobreposição em seu escopo, no geral o campo é promissoramente diverso, o que sugere que informações detalhadas são úteis em muitos setores diferentes.

A maioria dos conjuntos de dados que analisamos foram capturados como quadros únicos ou vídeos de câmeras estáticas.

Agora, estamos entrando em uma era em que a coleta e a rotulagem de conjuntos de dados exigem pesquisa avançada em visão computacional.

Por exemplo, capturar um conjunto de dados denso como não teria sido possível quando o Kinect foi lançado. À medida que os algoritmos de reconstrução e rotulagem para dados RGBD melhoram, a comunidade tem uma grande oportunidade de criar e compartilhar novos conjuntos de dados de reconstruções 3D de cenas estáticas e, finalmente, dinâmicas.

Referências

- RGBD Datasets: Past, Present and Future, 2016, Firman Michael [source]
- Database obtido de <https://www.visgraf.impa.br/vizdb/all.html>
- Reconstrução 3D de acervos culturais usando câmeras RGB-D: solução de compromisso entre precisão e tempo aplicada ao projeto Aleijadinho Digital, 2016, Gomes Leonardo, tese de doutorado UFPR [source].
- Remondino, F. and Rizzi, A. (2010). Reality-based 3D documentation of natural and cultural heritage sites: techniques, problems, and examples. *Applied Geomatics*, 2(3):85–100.
- 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.
- 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/>.
- O. Wasenmuller, 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/>.
- Outras referências e links de datasets se encontram no canto inferior direito de cada slide na caixa [source].

Obrigado!