Scene Semantic Reconstruction from Egocentric RGB-D-Thermal Videos

Rachel Luo, Ozan Sener, and Silvio Savarese
Stanford University
{rsluo, osener, ssilvio}@stanford.edu

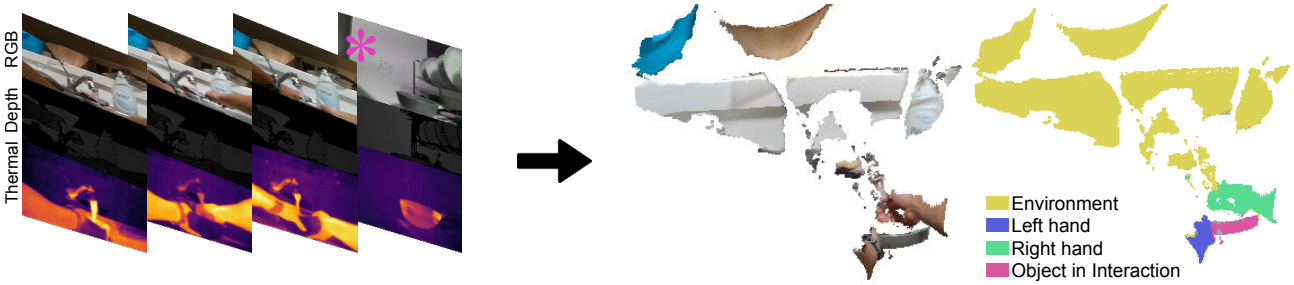


Figure 1: Scene Understanding. We propose a method for egocentric SLAM to gain geometric and semantic understanding of complex scenes where humans manipulate and interact with objects. The input of our system is an RGB-D-Thermal sensory stream from the perspective an operator (i.e. an egocentric view) (left panel). The desired output is a 3D reconstruction of the scene where the location and pose of the observer is detected (center panel) as well as a 3D semantic segmentation in terms of the elements that involve human interaction (for instance: left hand, right hand, object that the operator interacts with, remainder of the environment) (right panel).

Abstract

In this paper we focus on the problem of inferring geometric and semantic properties of a complex scene where humans interact with objects from egocentric views. Unlike most previous work, our goal is to leverage a multimodal sensory stream composed of RGB, depth, and thermal (RGB-D-T) signals and use this data stream as an input to a new framework for joint 6 DOF camera localization, 3D reconstruction, and semantic segmentation. As our extensive experimental evaluation shows, the combination of different sensing modalities allows us to achieve greater robustness in situations where both the observer and the objects in the scene move rapidly (a challenging situation for traditional methods for semantic reconstruction). Moreover, we contribute a new dataset that includes a large number of egocentric RGB-D-T videos of humans performing daily real-word activities as well as a new demonstration hardware platform for acquiring such a dataset.

1. Introduction

Consider a typical robotics scenario, in which a robot must understand a scene that includes humans, objects, and some human-object interactions. These robots will need to detect, track, and predict human motions [34, 42, 8], and relate them to the objects in the environment. For example, a kitchen robot helping a chef should understand which object the chef is reaching for on a very crowded and highly-occluded kitchen table in order to deduce and bring back the missing next ingredient from the refrigerator. This type of high-level reasoning requires the ability to infer rich semantics and geometry associated with both the humans (e.g. the position and pose of the chef's hands) and the objects (e.g. a pan, a cabinet, etc.) in the scene. We seek to solve this problem by introducing: i) a new egocentric dataset that integrates various sensing modalities including RGB, depth, and thermal; and ii) a framework that allows us to jointly extract critical semantic and geometric properties such as 3D location and pose.

Using egocentric videos in our dataset leads to a clear definition of semantics: each point can be labeled as part of either the human, the object in interaction, or the environment. This definition of semantics describes human-object interactions, and we define semantics and geometry as they relate to such interactions for the remainder of this paper. We include three critical modalities - RGB, depth, and thermal - in our raw data. We design an affordable hardware

to capture all of these modalities by mounting a structured-light camera (RGB-D) and a mobile thermal camera (RGB-Thermal) to a chest harness (Figure 2). We then develop a system to calibrate and time-synchronize these cameras. The resulting data is a 2D stream of RGB, depth, and thermal information. However, although a stream of RGB, depth, and thermal images includes the necessary semantic and geometric information, it is not structured in a way that is useful for scene understanding. We propose a new framework that can jointly infer the semantics (human vs. object vs. static environment) and the geometry (camera poses and 3D reconstructions) of the elements that are involved in an interaction (e.g. a hand, a plate that the hand is holding). The problem of jointly inferring semantics and scene geometry can be considered a form of a semantic SLAM problem whereby all three modalities (RGB, depth, and thermal) are considered in conjunction to increase robustness and accuracy. Moreover, our framework can naturally handle both static scene elements (e.g. a desk) and moving objects (a hand), unlike most SLAM or SFM methods, which assume that the entire 3D scene is static.

In summary, our contributions are: i) Designing an affordable, multi-modal data acquisition system for better scene understanding. ii) Sharing a large dataset of RGB-depth-thermal videos of egocentric scenes in which humans interact with the environment. Annotations (bounding boxes and labels) of the elements related to an interaction (e.g. a hand or a plate) are also provided. iii) An egocentric SLAM algorithm that can combine the three data modalities (RGB, depth, and thermal) and can handle both static and moving objects. We envision our proposed real-time SLAM architecture as a critical tool for modeling or discovering affordances or functional properties of objects from complex egocentric videos for robotics or co-robotics applications.

The rest of this paper is organized as follows. In Section 2, we discuss some related work. In Section 3, we describe the data acquisition system that we built as well as characteristics of the collected dataset. In Section 4, we explain the problem of solving SLAM for our egocentric videos to obtain semantic and geometric information. In Section 5, we show some of the outputs from our framework. Finally, Section 6 concludes the paper.

2. Related Work

Egocentric Scenes: A few previous works have studied egocentric scenes. For example, [39], [24], [15], and [22] look at first-person pose and activity recognition. [19] creates object-driven summaries for egocentric videos. However, none of these include large-scale publicly available datasets, and none of them include a thermal modality.

SLAM: SLAM is the problem of constructing a map of an unknown environment while tracking the location of a moving camera within it. Although there are visual odometry approaches [14, 23], explicit models of the map typically increase the accuracy of ego-motion estimation as well. Thus, SLAM has become an increasingly popular area of research, especially for robotics or virtual reality applications even when only the ego-motion is needed. Several early papers propose methods for monocular SLAM [4, 11]. More recently, ORB-SLAM proposes a sparse, feature-based monocular SLAM system [26]. LSD-SLAM is a dense, direct monocular SLAM algorithm for large-scale environments [7], and DSO is a sparse, direct visual odometry formulation [6].

Several stereo SLAM methods also exist for RGB-D settings, for example the dense visual method DVO-SLAM [16]. Kintinuous is another dense visual SLAM system that can produce large-scale reconstructions [46]. ElasticFusion is a dense visual SLAM system for room scale environments [47]. ORB-SLAM2 extends ORB-SLAM for monocular, stereo, and RGB-D cameras [27]. KinectFusion can map indoor scenes in variable lighting conditions [29], and BundleFusion estimates globally optimized poses [3].

All of the above algorithms are designed for static scenes from a more global perspective. Nevertheless, although most SLAM systems assume a static environment, a few methods have been developed with dynamic objects in mind. [41] builds a system that allows a user to rotate an object by hand and see a continuously-updated model. [48] presents a structured light technique that can generate a scan of a moving object by projecting a stripe pattern. More recently, DynamicFusion builds a dense SLAM system for reconstructing non-rigidly deforming scenes in real time with RGB-D information [28]. However, these methods reconstruct only single objects rather than entire scenes, and none consider the egocentric perspective.

Human-Object Interactions: One plausible application of scene understanding with human-object interactions is for robotics. Numerous attempts have been made over the past several decades to better understand human-object interactions. J.J. Gibson coined the term "affordance" as early as 1977 to describe the action possibilities latent in the environment [9]. Donald Norman later appropriated the term to refer to only the action possibilites that are perceivable by an individual [30].

More recently, [18] and [17] learn human activities by using object affordances. [35] teaches a robot about the world by having it physically interact with objects. [25] predicts long-term sequential movements caused by applying external forces to an object, which requires reasoning about the physics of the situation.

Several works also examine human-object interactions as they relate to hands or grasps. For instance, [40] and [2] both explore grasp classification in an attempt to understand hand-object manipulation. [12] studies the hands to discover a taxonomy of grasp types using egocentric videos.

Although we do not attack this problem directly, the output of our framework can be useful for learning about human-object interactions.

Hand Tracking: Hand tracking is another area of interest for human-object interactions, and it is another potential application of our framework. [21] and [20] perform pixel-wise hand detection for egocentric videos by using a dataset of labeled hands, and by posing the hand detection problem as a model recommendation task, respectively. [45] and [38] perform depth-based hand pose estimation from a third-person and an egocentric perspective. [44] simultaneously tracks both hands manipulating an object as well as the object pose using RGB-D videos. The closest to our work is [38]; however, it considers only images, lacks thermal information, and experiments only at small scale.

3. Dataset

In order to test our approach for obtaining a semantic reconstruction of complex egocentric scenes, we designed a multi-modal data acquisition system combining an RGB-D camera with a mobile thermal camera. We then used this setup to collect a large dataset of aligned multi-modal videos, and annotated semantically relevant information in these videos. In this section, we explain our process in detail and discuss the characteristics of the collected dataset. In Section 3.1, we describe the hardware that we used; in Section 3.2, we describe our method for geometrically calibrating our data; in Section 3.3, we describe our annotations; in Section 3.4, we discuss the collected data; and in Section 3.5, we discuss potential future applications of the dataset.

3.1. Hardware

Our data acquisition system includes two mobile cameras: one RGB-D (an Intel RealSense SR300 [36]) and one thermal (a Flir One Android [33]). We mounted both cameras on a GoPro chest harness and connected them through a single USB 3.0 cable to a lightweight laptop kept in the backpack of the data collector. We developed a GNU/Unix driver for the Flir One, since the camera was originally designed for Android mobile phones. We also time synchronized the cameras using the frame rate of the slower of the two cameras (the Flir One), resulting in a data acquisition rate of 8.33 FPS. (Because the US government allows only thermal cameras with frame rates lower than 9fps to be exported without a license, most consumer thermal cameras are limited to 9fps. However, our system still works well at this frame rate - by using a thermal camera with a higher frame rate, performance may further improve.)

The final developed camera setup and chest mount are shown in Figure 2. After calibrating (as described in Section 3.2), we considered the spatial area covered by both cameras and saved per-pixel RGB, depth, and temperature

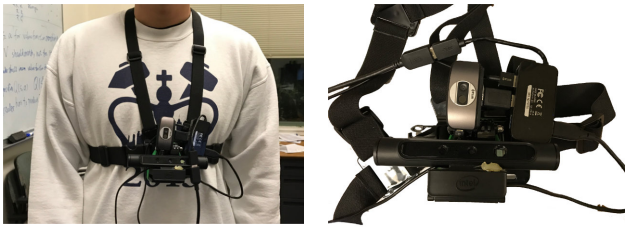


Figure 2. Developed multi-modal camera setup. We combined an RGB-D camera (Intel RealSense SR300) with a mobile thermal camera (Flir One) and attached both to a GoPro chest harness.

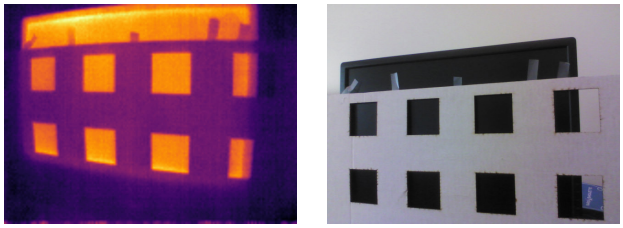


Figure 3. Multi-modal calibration board for geometric calibration of thermal and RGB cameras. The same scene is shown in both the thermal and RGB views. As evident from the figure, our calibration board results in very visible edges and corners in both modalites.

values. We relied on RGB values provided by the RGB-D camera and used nearest neighbor interpolation for pixel values that were missing due to the resolution mismatch between different modalities. We will open source all code and hardware designs.

3.2. Geometric Calibration

We geometrically calibrated the two cameras using a special checkerboard that we designed for this purpose. We laser cut a checkerboard pattern out of a piece of cardboard such that each square measured 2in. x 2in. We attached this piece of cardboard to an LCD monitor that had been used right before the calibration process but then turned off to darken the screen. Because the LCD monitor remained hot right after being turned off and the cardboard did not conduct heat, both a strong temperature difference and a large color difference were visible. After capturing multiple shots of the RGB and thermal images, we used the Caltech Camera Calibration Toolbox [1]. A sample image of the calibration board in both the thermal and RGB modalities is shown in Figure 3.

3.3. Annotations

The hands and the objects that the hands interact with comprise the key semantic information that we need. We annotated the location of each hand and each object in interaction for every fifth video frame and interpolated between them for the frames in between. Since the RGB-D and ther-

mal cameras are jointly calibrated, annotating the the locations in the RGB modality suffices; hence, annotations were done for the RGB channel of the RGB-D camera. Our hand and object annotations take the form of 2D bounding boxes. We also annotated a few segmentations for the hands and objects in interaction.

To obtain the ground truth camera poses, we chose approximately 10 uniformly distributed frames for 10 randomly chosen videos and manually provided a set of corresponding points in each pair of consecutive frames to calculate the relative camera pose. We used least square estimation with RANSAC using the camera calibration matrix provided by Intel.

3.4. Statistics and Visual Examples

Our dataset includes approximately 250 videos of people performing various activities. These activities are divided into four high-level categories - kitchen, office, recreation, and household - with 44 different types of action sequences distributed across these four categories. Some of the more common activity types are tabulated in Table 1. Videos are 3 minutes long on average, and there are over 450,000 frames in total. We observed 14 different people collecting data over more than 20 different environments.

We provide visual examples for some of the activities in Figure 4. One key property of our dataset is that all interactions are very natural, since we did not give the data collectors any specific instructions other than asking them to wear the camera while performing the high-level activity. Also, since we did not tell the data collectors what actions to perform, the distribution of the dataset reflects the natural distribution of activities that the data collectors do routinely. Furthermore, because we performed the data collection in various environments, our resulting dataset is highly diverse in terms of appearances, shapes, and interactions.

3.5. Possible Future Directions

Several interesting phenomena emerge from the different modalities of our dataset. One such phenomenon from the thermal modality is that an object that interacts with human hands usually gets warmer since the hand is warm. After the interaction is over, part of the object stays warmer than the rest of the object. We refer to this heat imprint from the hand as *thermal residue*, and it may act as an important indicator of properties such as where the hand touches the object, the amount of force applied to the object, and the object material. The temporal dissolution pattern of the thermal residue correlates with force and material properties. Qualitatively, touching an object with more force results in a longer and stronger thermal residue. Also, there is minimal thermal residue on surfaces with high thermal conductivity, but a long thermal residue on materials with low thermal conductivity. Thus, learning the force of an interaction or learning some material properties using this data are potentially interesting avenues for future research.

4. Egocentric SLAM with Semantics

In order to have a good understanding of a scene with human-object interactions, it is necessary to know about the humans, the objects in interaction, and their environment. Moreover, this knowledge must include both semantics and geometry. As discussed in Section 3, the multimodal dataset that we collected includes all of the required information in an unstructured form. In this section, we explain our proposed method for converting the raw videos into structured information. We first detect the hands using all modalities as an initialization step. Then, we learn the semantics of the scene by segmenting out the hands, the objects interacting with the hands, and the static environment points. Finally, we infer the geometry of the scene by learning the camera pose at each timestep and creating a 3D map of the scene.

There are two key problems that need to be solved to achieve this: i) labelling each 3D point with one of the semantic classes: *left hand, right hand, object in interaction, static environment*, and ii) aligning all frames in space in order to construct the geometric information. Since our videos are egocentric, the left and right hands describe the human component, while the object in interaction and the static environment describe the remaining components.

This problem can be seen as form of structure-from-motion (SfM) or simultaneous localization and mapping

Table 1. Distribution of collected videos over high level activities.

Activity	Number of Videos
Kitchen (Total)	**133**
Cutting vegetables	14
Using a microwave	14
Pouring a drink	12
Washing dishes	9
Wiping the counter	7
Office (Total)	**89**
Reading	18
Writing by hand	13
Using a laptop	9
Using scissors	9
Using tape	9
Recreation (Total)	**19**
Using a cell phone	5
Lifting weights	5
Playing piano	4
Household (Total)	**7**
Folding clothes	3
Sweeping	1
Vacuuming	1

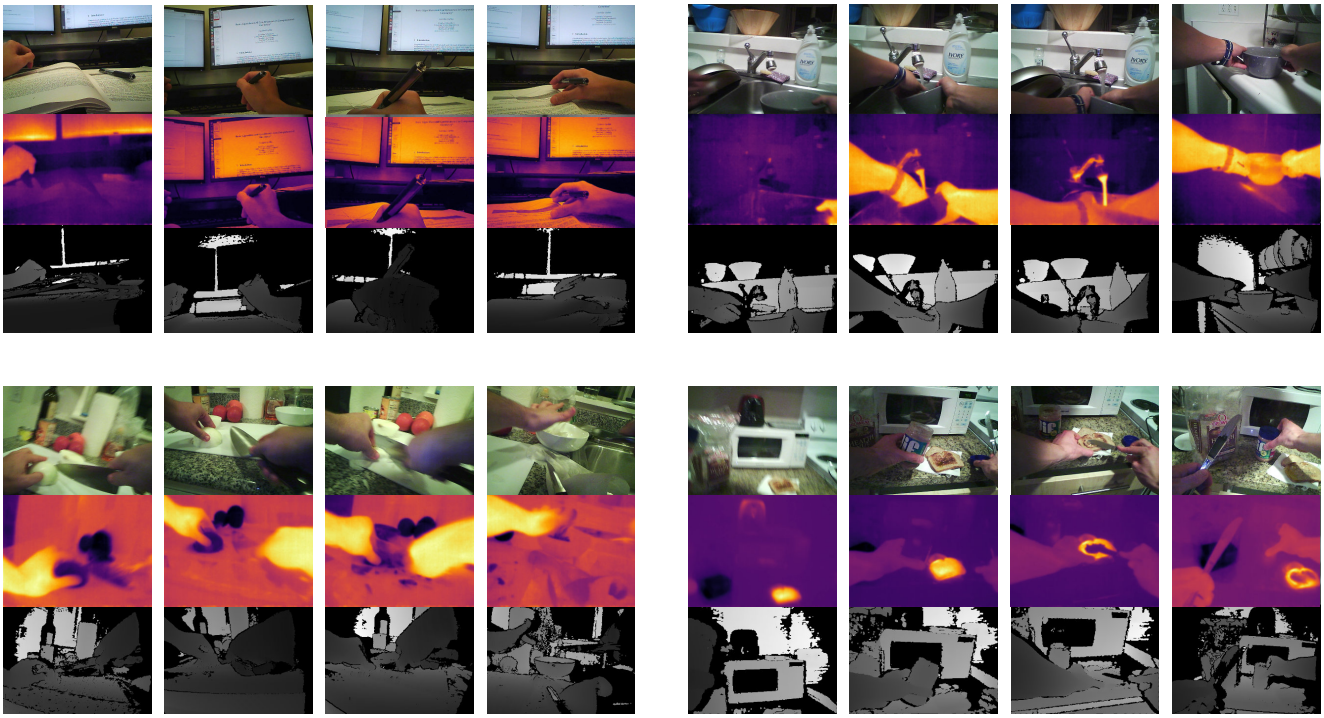


Figure 4. Visual examples from the dataset. We visualize various kitchen and office activities in this figure. The first row of each set shows the RGB images, the second row shows the thermal images, and the third row shows the depth images. Our dataset consists of subjects executing daily activities while wearing our cameras. Because we do not give any specific instructions to the subjects, all motions are natural. Actions are also performed in very different environments for high variability in the dataset.

(SLAM) problem depending on whether the geometric alignment is done in an online or offline fashion. Since our main application area is robotics, we choose to develop an online algorithm. Hence, the problem we address is *egocentric SLAM with semantics*. In the remaining part of this section, we will define and discuss the characteristics of the egocentric SLAM with semantics problem and then explain our algorithm in detail.

4.1. Problem Definition

The goal of our algorithm is to take in an aligned multi-modal video as input, and output a 3D point-cloud over time. Formally, the input to our algorithm is $\{\mathbf{z}_i^t, \mathbf{c}_i^t, d_i^t, \tau_i^t\}_{i\in\{1,\dots,W\times H\}}$ where $\mathbf{z}$ is the 2D-pixel location, $\mathbf{c}$ is the RGB color vector, d is the depth value, and τ is the temperature (normalized to $[0, 1]$). W and H are the width and height of the image. The desired output is a dense 3D point cloud as $\{\mathbf{x}_i^t, \mathbf{c}_i^t, \tau_i^t, s_i^t\}_{i\in\{1,\dots,N\}}$ where $\mathbf{x}$ is the 3D location in a global coordinate frame and s is a semantic label as $s \in \{\text{left hand}, \text{right hand}, \text{object in interaction}, \text{static}\}$. We call our problem "egocentric SLAM with semantics" since it has distinct properties when compared to a typical SLAM setup. We summarize these distinct properties as follows:

i) Structured dynamicity: Most SLAM algorithms assume a static environment and cannot handle dynamic objects. The dynamic SLAM methods that do exist are computationally too expensive and can only handle small motions. Our setup is a middle ground between fully static and fully dynamic scenes since all the dynamicity in the scene is caused by the person wearing the camera.

ii) Well-defined semantics: In our setup, we need to annotate each point with a well-defined class from *left hand, right hand, object in interaction, static environment*. These classes are well defined since they are caused by the geometry of the human-object interaction.

iii) High camera motion: Our camera is chest-mounted on a human with a harness, so it experiences a large amount of motion due to the non-linear movements of the humans and the elasticity of the harness. This setup is very different from the slowly-panning videos typically used in SLAM.

4.2. Approach

Our method iterates over camera localization, semantic segmentation, and sparse mapping for each frame. Given the multi-modal video frame with color, depth, and temperature information, we start by solving the pose of the camera as $\mathbf{R}^t$ and $\mathbf{t}^t$, the rotation and translation of the camera, respectively. Using the camera pose, we label each pixel in the input frame as one of *left hand, right hand, object in interaction, static environment*. Then, we use only the static environment points to update the internal sparse map.

Our camera localization and sparse mapping is based on extending an existing SLAM method. We carefully extend ORB-SLAM [27] for our setup since it is robust to camera motion and motion blur. We use the camera localization and sparse mapping submodules of ORB-SLAM [27] by masking our input to include only the static points; this step is necessary since ORB-SLAM handles only static scenes. Our key contribution is extending existing SLAM methods with the knowledge of static and dynamic regions resulting from our accurate and effective semantic segmentation algorithm. We skip the details of ORB-SLAM here and provide a short summary in the Supplementary Materials for the sake of completeness, as our contributions are general and can be applied to any SLAM algorithm.

4.3. Semantic Segmentation - Separating the Dynamic and Static Points

One of the key problems in our egocentric SLAM with semantics framework is the segmentation of the input frame into *left hand, right hand, object in interaction, static environment* classes. The importance of this segmentation is two-fold: i) removing the dynamic points from the input frame is critical for successful SLAM operation, and ii) these labels create the structure needed for good scene understanding.

We perform semantic segmentation after the initial camera localization. Although ORB-SLAM implicitly assumes a static scene, using the entire scene for an initial localization is reasonable since only the updated map, which includes only static points, is used for pose estimation. The camera pose is useful for semantic segmentation because having accurate pose information is instrumental in reasoning about moving objects.

Our semantic segmentation algorithm is based on a few priors: We have priors for the hands due to their consistent color model, high temperature, and distinct shape. Moreover, hand location is a prior for the location of the object in interaction since the object needs to be in contact with the hand. Therefore, we can perform semantic segmentation as a two-step process: first, we segment the left and right hands from the frame; then, we segment the object in interaction. Note that to distinguish between the right and left hands, we use the prior that at the beginning of the interaction, the right hand is at the right side of the image frame, and the left hand is at the left side. If we lose the hands, we reinitialize this process in the same way.

We use CRF based image segmentation to segment the hands, defining an energy minimization problem:

$$\min_{\alpha_i^t} \sum_i U(\alpha_i^t, \mathbf{y}_i^t) + \sum_i \sum_{j \in \mathcal{N}(i)} V(\mathbf{y}_i^t, \mathbf{y}_j^t) 1[\alpha_i^t \neq \alpha_j^t] \quad (1)$$

In this formulation, α_i^t is a binary variable that is 1 if pixel i is part of the hand at time t and 0 otherwise. $\mathcal{N}(i)$ is the set of pixels neighboring i, and $1(\cdot)$ is an indicator function. $\mathbf{y}$ is the concatenated vector of $\mathbf{z}, \mathbf{c}, d, \tau$.

$U(\alpha_i^t, \mathbf{y}_i^t)$ is the unary energy representing the likelihood that the i^{th} pixel is part of the hand. It is a weighted combination of the likelihood over the temperature (T), color (C), hand-detector outputs (S), and history over time (H).

$$\begin{aligned} U(\alpha_i^t, \mathbf{y}_i^t) &= w_T U^T(\alpha_i^t, \mathbf{y}_i^t) + w_C U^C(\alpha_i^t, \mathbf{y}_i^t) \\ &+ w_S U^S(\alpha_i^t, \mathbf{y}_i^t) + w_H \sum_i U(\alpha_i^{t-1}, \mathbf{y}_i^{t-1}) e^{-\Delta(\mathbf{y}_i^t, \mathbf{y}_i^{t-1})} \end{aligned} \quad (2)$$

where $\Delta(\cdot,\cdot)$ is the geodesic distance over rgb-thermal space between two voxels *(see the Supplementary Materials for formal definition)*. $V(\cdot,\cdot)$ is the binary consistency term defined over neighboring pixels as

$$V(\mathbf{y}_i^t, \mathbf{y}_j^t) = \exp\left(-\frac{|\mathbf{y}_i^t - \mathbf{y}_j^t|_2}{\gamma}\right) \quad (3)$$

where $\gamma = \frac{1}{N}\sum_i \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} |\mathbf{y}_i^t - \mathbf{y}_j^t|_2$, and N is the total number of pixels.

We define the each component of the unary energy as

$$\begin{aligned} U^T(\alpha_i^t, \mathbf{y}_i^t) &= \tau_i^t 1[\alpha_i^t = 1] + (1-\tau_i^t) 1[\alpha_i^t = 0] \\ U^C(\alpha_i^t, \mathbf{y}_i^t) &= p(\mathbf{c}_i^t | \alpha_i^t) \\ U^S(\alpha_i^t, z_i^t) &= \sum_{k \in \mathcal{H}} p_k e^{-\Delta(\mathbf{y}_i^t, \mathbf{y}_k)} 1[\alpha_i^t = 1] \end{aligned} \quad (4)$$

where $p(\mathbf{c}_i^t|\alpha_i^t)$ is an RGB-color model represented as a Gaussian Mixture Model (GMM) with five components and learned separately for the hand and the static scene from the training data. $\mathcal{H}$ is a collection of hand detections, each represented by a centroid $\mathbf{c}_k$ and a detection likelihood p_k. $\mathbf{y}_k$ is the color, position, depth and temperature of centroid of the detected hand.

All components of this energy function can be computed in log-linear time using bi-linear filters, and minimized using the min-cut/max-flow framework as explained in [43]. We use the open source code released by the authors of [43] and refer the readers to the original paper for details.

After the hands are segmented, we segment the rest of the image into static and dynamic object components. We use the same energy minimization framework after introducing an additional prior on motion and dropping the prior on color. The motion prior corresponds to the fact that the motion of the object in interaction is different from the camera motion and is defined as:

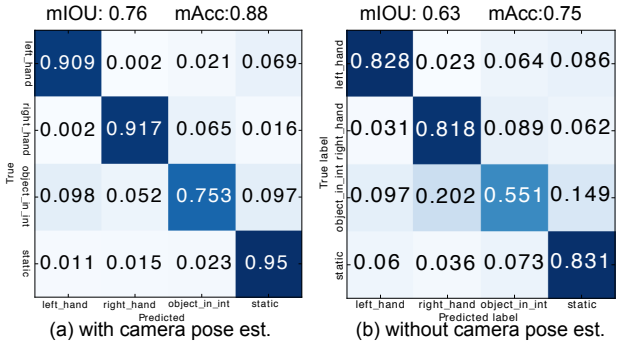
$$U^M(\alpha_i^t, \mathbf{y}_i^t) = \rho(|\mathbf{z}_i^t - \mathbf{z}^{t-1}_{\pi(\mathbf{R}^t \mathbf{x}_i^t + \mathbf{t}^t)}|) \quad (5)$$

where ρ is the Huber function, π is the pinhole projection, $\mathbf{R}, \mathbf{t}$ are the estimated camera pose, and $\mathbf{X}_i$ is the 3D position of i^{th} point in homogeneous coordinates. With some abuse of notation, α_i is a binary variable that is 1 if pixel i

is part of the object in interaction and 0 otherwise. The full formulation is included in the Supplementary Materials.

To learn the tradeoff parameters w_T, w_S, w_H, w_M, we use cross-validation and explain the implementation details in the Supplementary Materials.

Hand Detection: We use all three modalities to detect the hands with an algorithm based on the YOLO object detector [37] for real-time performance. Our analysis suggests that 2D bounding box detection using all three modalities results in higher accuracy than state-of-the-art, model-based RGB-D hand pose detection algorithms. In order to train the YOLO detector, we use features pre-trained on ImageNet [5] and train with the annotated bounding boxes in the dataset. Since the pre-training exists only for RGB images, we use knowledge distillation [10] for transferring pre-trained features to thermal and depth modalities. *(See the Supplementary Materials for details).*



(a) with camera pose est. (b) without camera pose est.

Figure 5. Confusion matrix for semantic segmentation.

5. Experimental Results

We perform several experiments to evaluate our design choices and demonstrate the effectiveness of our algorithm. We start by discussing qualitative results. Our algorithm is designed to generate the semantics and geometry of human-object interactions. As a visualization tool, we create a dense 3D map of the scene by directly transforming all frames into the coordinate system of the first frame and overlapping all points. To visualize both the intermediate steps and the final result, in Figure 6 we show the input RGB-D point clouds as well as their semantic segmentations at various time steps for two sequences.

As Figure 6 suggests, our algorithm results in accurate semantic segmentation, and the semantic reasoning results in accurate geometric information after the egocentric SLAM procedure is completed. Consider the input in the right column, for instance. It is difficult to segment even for a human; however, our algorithm accurately segments the hands and the object in interaction (a 3-hole punch). We will now look at some quantitative results for the semantic reasoning and the hand detection.

Semantic Reasoning: We believe that semantic reasoning is key for efficient camera pose estimation. Although there are a few existing methods for SLAM with dynamic objects [28, 13], none of them include open source code and therefore we cannot compare our methods with theirs. Nonetheless, both existing dynamic SLAM papers state that their methods apply only when the objects are large and the motion is very slow - which is certainly not the case for our data. Hence, these methods are not applicable.

In order to quantify the semantic segmentation quality, we annotate a subset of video frames with their semantic segmentations and compute the confusion matrix shown in Figure 5. Our confusion matrix suggests that our semantic segmentation is very accurate. In order to evaluate the effect of geometric reasoning, we also compare our semantic segmentation quality against a baseline not using camera pose estimation (we simply ignore the $U^M(\cdot)$ term in segmentation). As the result in Figure 5 suggests, the camera pose estimation is crucial for accurate segmentation. Hence, semantic segmentation and SLAM should be solved jointly.

Table 2. Hand detection accuracy. Our results suggest that thermal and RGB are the most important modalities.

Algorithm	Average Precision (AP)
FORTH [32] (rgb-d)	51.1
Deep Hand [31] (rgb-d)	68.7
YOLO [37]	66.3
YOLO* (rgb-d)	73.6
YOLO* (rgb-t)	86.3
YOLO* (full)	89.2

In order to further evaluate the effect of the semantic segmentation, we compare our egocentric SLAM against a version without semantic segmentation (vanilla ORB-SLAM) in Table 3 for camera localization accuracy. Results suggest that semantic segmentation is an integral part of our pipeline. For example, for the cases of filling a cup and reading, the vanilla ORB-SLAM loses the tracker and fails to produce any output without our semantic segmentation.

Hand Detection: The hand detection is another important part of our algorithm. We quantitatively analyze hand detection performance by comparing various state-of-the-art object detection and RGB-D hand detection algorithms for the hand detection problem, as shown in Table 2. We name our hand detection algorithm YOLO*, simply representing YOLO [37] combined with cross-modal distillation. We also compare the effect of each modality on the accuracy.

Our results suggest that the thermal information is the most useful modality. This result is unsurprising because a human hand is typically warmer than its surroundings. However, a thermal image by itself is not enough since there are usually other warm objects in the environment as well (computer monitors, etc.). Our simple 2D bounding box

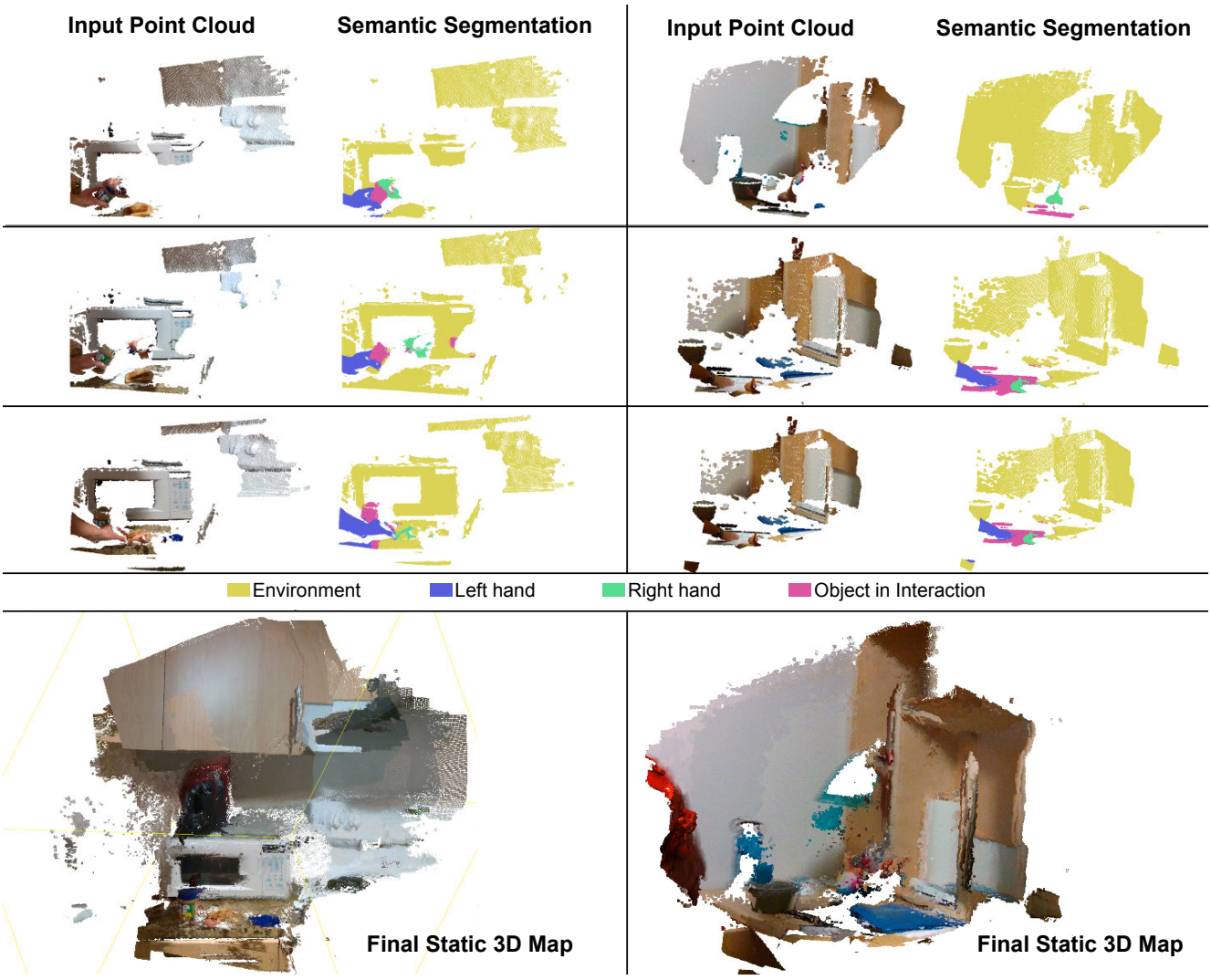


Figure 6. Qualitative results for our egocentric SLAM approach. We show various RGB-D point clouds as input, as well as their semantic segmentations. We also display the final 3D dense static map, reconstructed by combining all input point clouds.

Table 3. Comparison of our method against a version without semantic segmentation. Lost corresponds to videos where SLAM diverged.

		Can opener	Cutting vegetables	Filling a cup	Hole punch	Micro-wave	Peanut butter	Reading	Washing dishes	Average
rMSE Translation Error (cm)	No semantics	4.48	0.81	Lost	4.00	3.48	19.62	Lost	1.78	6.13
	Full method	**4.44**	**0.79**	**1.12**	**3.88**	**2.56**	**12.76**	**1.87**	**1.77**	**3.64**
rMSE Rotation Error ($0.01 \times rad$)	No semantics	4.19	0.96	Lost	2.32	3.81	9.59	Lost	2.20	3.84
	Full method	**4.09**	**0.93**	**0.89**	**2.06**	**3.48**	**7.66**	**1.60**	**2.17**	**2.86**

hand detection outperforms model-based articulated hand pose estimation methods like [32] and [31]; thus, a careful combination of the thermal, RGB, and depth modalities can act as a very powerful framework for hand detection.

6. Conclusion

We introduce a new data modality for better scene understanding. We also present a novel framework for simultaneously inferring geometric and semantic properties of a complex scene with human-object interactions, as seen from an egocentric point of view. We contribute a new large-scale dataset of egocentric RGB-D-T videos of everyday activities, along with an affordable camera setup for acquiring such a dataset. We then use the multi-modal sensory stream from our dataset as an input to our framework for 6 DOF camera localization, 3D reconstruction, and semantic segmentation. Our results show state-of-the-art performance for hand detection, 3D reconstruction, and pose estimation, and they provide rich information about human-object interactions, ready for use in robotics pipelines.

References

[1] J.-Y. Bouguet. Camera calibration toolbox for matlab. 2004.

[2] M. Cai, K. M. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. *Conference on Robotics Science and Systems (RSS)*, 2016.

[3] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using online surface re-integration. *arXiv preprint arXiv:1604.01093*, 2016.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[6] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. In *arXiv:1607.02565*, July 2016.

[7] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[8] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun. Real-time human pose tracking from range data. In *ECCV*, 2012.

[9] J. Gibson James. The theory of affordances. *Perceiving, Acting, and Knowing, Eds. Robert Shaw and John Bransford*, 1977.

[10] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *In Proc. Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] S. Holmes, G. Klein, and D. W. Murray. A square root unscented kalman filter for visual monoslam. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3710–3716. IEEE, 2008.

[12] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[14] A. Jaegle, S. Phillips, and K. Daniilidis. Fast, robust, continuous monocular egomotion computation. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 773–780. IEEE, 2016.

[15] H. Jiang and K. Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. *arXiv preprint arXiv:1603.07763*, 2016.

[16] C. Kerl, J. Sturm, and D. Cremers. Dense visual slam for rgb-d cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013.

[17] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[18] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *Robotics: Science and Systems (RSS)*, 2013.

[19] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.

[20] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2624–2631, 2013.

[21] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2013.

[22] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016.

[23] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012.

[24] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, page 3, 2013.

[25] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi. "what happens if..." learning to predict the effect of forces in images. *CoRR*, abs/1603.05600, 2016.

[26] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[27] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *arXiv preprint arXiv:1610.06475*, 2016.

[28] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[29] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. a. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*. IEEE, October 2011.

[30] D. A. Norman. *The psychology of everyday things*. Basic books, 1988.

[31] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015.

[32] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, number 2, page 3, 2011.

[33] F. One Android. http://www.flir.com/flirone/android/. *Accessed in*, 2017.

[34] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, Oct 2004.

[35] L. Pinto, D. Gandhi, Y. Han, Y. Park, and A. Gupta. The curious robot: Learning visual representations via physical interactions. *CoRR*, abs/1604.01360, 2016.

[36] I. RealSense. www.intel.com/realsense. *Accessed in*, 2017.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[38] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *Workshop at the European Conference on Computer Vision (ECCV)*, pages 356–371. Springer, 2014.

[39] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4333, 2015.

[40] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 3889–3897, Washington, DC, USA, 2015. IEEE Computer Society.

[41] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Trans. Graph.*, 21(3):438–446, July 2002.

[42] O. Sener and A. Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *Robotics: Science and Systems*. Citeseer, 2015.

[43] O. Sener, K. Ugur, and A. A. Alatan. Efficient mrf energy propagation for video segmentation via bilateral filters. *IEEE Transactions on Multimedia*, 16(5):1292–1302, Aug 2014.

[44] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[45] J. S. Supancic III, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: Data, methods, and challenges. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[46] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald. Real-time large scale dense rgb-d slam with volumetric fusion. In *IJRR*, 2014.

[47] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. Elasticfusion: Dense slam without a pose graph. In *RSS*, 2015.

[48] L. Zhang, B. Curless, and S. M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 24–36, June 2002.