Where and Why Are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks

Ping Wei[1,2], Yang Liu[2], Tianmin Shu[2], Nanning Zheng[1], and Song-Chun Zhu[2]

[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2]Center for Vision, Cognition, Learning, and Autonomy, University of California, Los Angeles
pingwei@xjtu.edu.cn,{yangliu2014,tianmin.shu}@ucla.edu, nnzheng@xjtu.edu.cn, sczhu@stat.ucla.edu

Abstract

This paper addresses a new problem - jointly inferring human attention, intentions, and tasks from videos. Given an RGB-D video where a human performs a task, we answer three questions simultaneously: 1) where the human is looking - attention prediction; 2) why the human is looking there - intention prediction; and 3) what task the human is performing - task recognition. We propose a hierarchical model of human-attention-object (HAO) which represents tasks, intentions, and attention under a unified framework. A task is represented as sequential intentions which transition to each other. An intention is composed of the human pose, attention, and objects. A beam search algorithm is adopted for inference on the HAO graph to output the attention, intention, and task results. We built a new video dataset of tasks, intentions, and attention. It contains 14 task classes, 70 intention categories, 28 object classes, 809 videos, and approximately 330,000 frames. Experiments show that our approach outperforms existing approaches.

1. Introduction

While recognizing what a human is doing in videos has been extensively studied over the past decades, inferring what a human is thinking is a rarely-investigated but important problem. For example, in a scene of human-robot collaboration, a human standing still is looking around without any body actions. To collaborate with the human, the robot needs to know what the human is thinking, e.g. *is the human searching for anything or checking any object's state?*

Answering these questions involves inferring human attention and intentions in tasks. A task is a complex goal-driven human activity [18] and performing a task is a process of eye-hand coordination [23], as the task *mop floor* shown in Fig. 1. Human attention describes where a human is looking [38]. It includes the attributes of 3D location, 3D direction, and 2D location, as shown in Fig. 1.

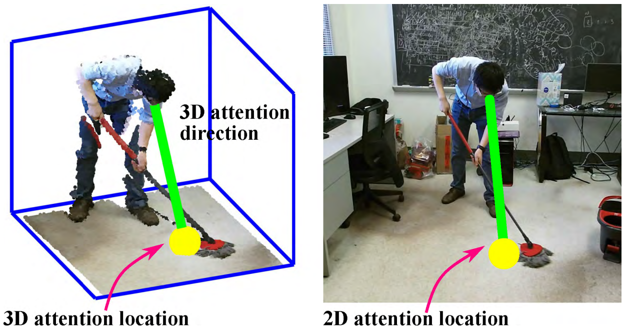


Figure 1: Human attention and intention in the task *mop floor*. While mopping the floor, the person is looking at the floor and his intention is checking if the floor has been cleaned or not.

Human intentions in our work describe the mental motivation why a human is looking at a place. In cognitive studies, Land *et al.* [16] defined four basic types of human fixation roles - *locate*, *direct*, *guide*, and *check*. As shown in Fig. 2, we extend the four fixation roles to explain human intentions in complex tasks: 1) *locate* is to identify the location of an object in a scene; 2) *direct* means a human directs the hands to something or to do something; 3) *guide* means a human guides an object to approach another; 4) *check* is to check the object states. With different compositions of objects and actions in various tasks, the four basic types can be expanded into numerous categories, such as *locate mop*, *locate coffee jar*, etc. We define these expanded categories as human intentions in tasks. Intention prediction is to label each video frame with one of the intention categories.

As the saying goes, *'eyes are the windows to the soul'*. Human attention and intentions are closely related to each other in a task. By perceiving where a human is looking, we can infer the human's intentions. For example, in the task

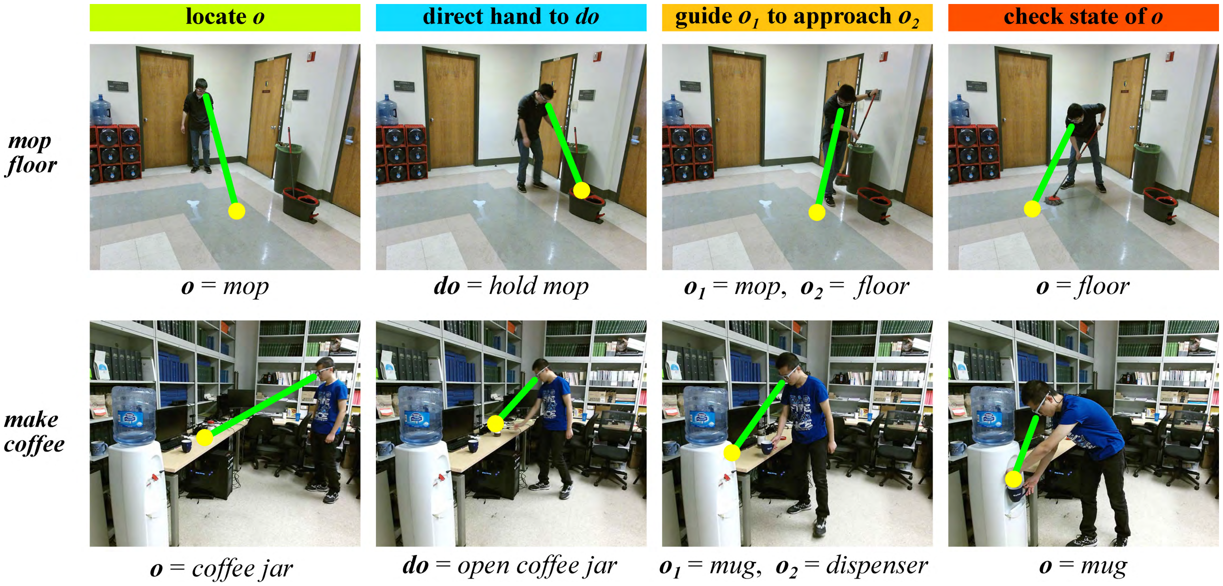


Figure 2: Four basic types of intentions when humans perform tasks.

make coffee shown in Fig. 2, while fetching water from the dispenser, the person's attention focuses on the mug and his intention is to check the mug's state (full or not). On the other hand, human intentions drive human attention, which makes attention present different characteristics in different intentions [38]. For example, in Fig. 2, when the person's intention is to check the mug's state, his attention focuses on the mug; when the person's intention is to locate the mug, his attention rapidly moves on the desk.

In this paper, we propose a hierarchical graph model of human-attention-object (HAO) to jointly represent and infer human attention, intentions, and tasks in videos. A task is represented as sequential intentions which transition to each other. An intention is composed of the human pose, the human attention, and the intention-related objects. The attention bridges the human and objects in both spatial and temporal domains. For an RGB-D video, we adopt a beam search algorithm to jointly infer the task label, the intention, the 3D attention direction, the 2D and 3D attention locations in each video frame. We collected a new large-scale video dataset of tasks, intentions, and attention (TIA). Experimental results prove the strength of our method.

This paper makes three major contributions:

1) It studies a new problem and develops video understanding from recognizing what a human is doing to inferring what a human is thinking.
2) It proposes a hierarchical model to represent tasks as transitional intentions which are described with human poses, attention, and objects.
3) It presents an RGB-D video dataset of tasks, intentions, and attention.

1.1. Related Work

Human Intention and Mind. Intentions can be roughly divided into action intentions [31, 24, 32] and mind intentions [36, 12, 27, 3, 16, 23, 6, 41, 40]. Action intentions describe subsequent actions. Mind intentions describe invisible motivations or motions in human minds [16, 36]. Such intentions cannot be directly perceived from visual features but only can be inferred from spatial-temporal cues. Moreover, mind intentions usually occur before action intentions since what humans are thinking drives their subsequent actions. The intentions in our work belong to mind intentions.

Human Attention and Gaze. Visual saliency [13] describes image regions which attract the attention of observers outside the image. Inside-data attention describes where a human inside the image is looking [25, 11]. The attention in our work belongs to the inside-data attention.

Eye or face features are often used to estimate human gazes [25, 35, 42, 22, 33, 14, 10, 19, 7, 11]. However, in large-scale daily-activity scenes, it is hard to obtain usable eye or face features due to low resolution. In this case, human body feature is an alternative to infer gazes [21, 38, 40].

Some studies model gazes with object or action information [28, 34, 5, 4, 20, 2, 9, 17, 40]. However, attention is also driven by intentions. In a task, the human does not necessarily look at the related objects all the time. It is necessary to jointly model attention, intentions, tasks, and objects.

Action and Task. Traditional action recognition is concerned with what humans are doing in images or videos [29, 37, 39]. Actions are interpreted with visible features and lay less stress on goals. Tasks are goal-driven activities with more complex spatial-temporal structures [12, 18].

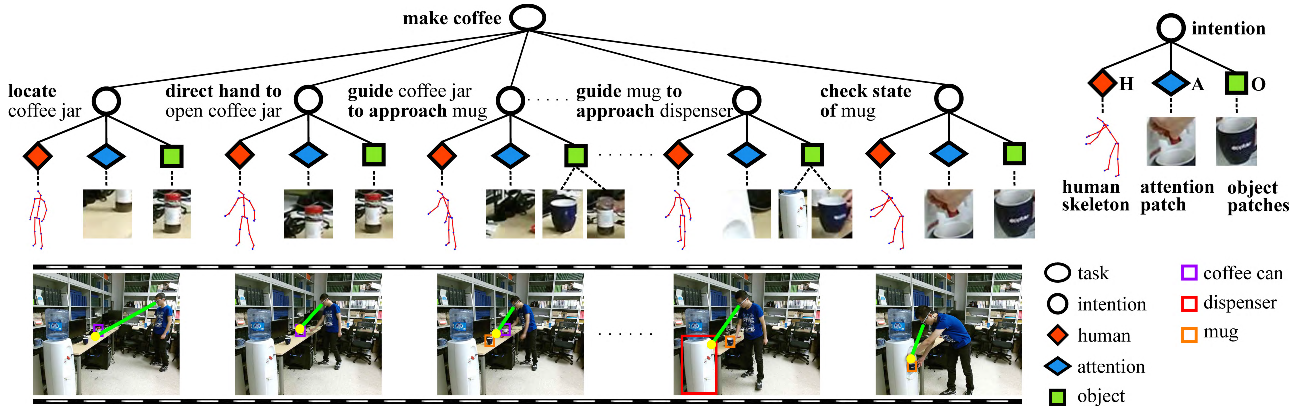


Figure 3: Human-attention-object (HAO) graph. The image patch under the attention node is the attention area where the human looks.

2. Model

We propose a hierarchical human-attention-object (HAO) graph to represent tasks, human intentions, and attention, as shown in Fig.3. The graph contains four layers which correspond to the task, intentions, attention-bridged human body and objects, and the video, respectively.

A task is divided into several intentions in time domain. As shown in Fig. 3, the task *make coffee* is composed of eight sequential intentions, such as *locate coffee jar*, *guide mug to approach dispenser*, *check state of mug*, etc. These intentions can transition to each other.

Intentions are revealed by cues of human bodies, human attention, and objects. Therefore, an intention is decomposed into the human pose, the human attention, and the intention-related objects, as shown in Fig. 3. The human attention bridges the human body and the objects.

2.1. Representation and Formulation

We use RGB-D videos recorded by motion capture technology like Kinect as inputs. Each frame includes an RGB image, a depth image, and a 3D human skeleton composed of 3D joint locations.

Let $\mathbf{I} = \{I_t | t = 1, ..., \tau\}$ be an input RGB-D video with length τ. I_t is the RGB-D frame at time t.

$\mathbf{H} = \{(\mathbf{h}_t, \mathbf{x}_t) | t = 1, ..., \tau\}$ is the human pose feature sequence. $\mathbf{h}_t$ and $\mathbf{x}_t$ are the appearance and geometric features extracted from the 3D skeleton at time t, respectively.

S is the task label of the input video. $\mathbf{L} = \{l_t | t = 1, ..., \tau\}$ is the human intention sequence of the video, where l_t is the intention label of the frame at time t.

$\mathbf{Y} = \{y_t | t = 1, ..., \tau\}$ is the human attention sequence. $\mathbf{y}_t$ is the *3D attention direction* in the t-th frame. It is defined as a unit 3D vector starting from the human head. The intersection point of the 3D attention direction and the scene point cloud is the *3D attention location*. With depth data, the 3D attention point is projected onto the image to form the *2D attention location*.

In the t-th RGB frame, we define a square image patch centered at the 2D attention point to extract the attention appearance feature $\mathbf{a}_t$. This image patch is like a central area where the human is looking, as shown in Fig. 3.

In the t-th frame, suppose $\mathbf{o}_t = (o_t^1, ..., o_t^m)$ is a bounding box collection of m intention-related objects, such as *mug* and *coffee jar* in the intention *guide coffee jar to approach mug*. These bounding boxes are proposed by the Faster R-CNN [26] object detectors. With depth values of the RGB-D data, the 2D centers of object bounding boxes are projected onto the 3D space to form the objects' 3D locations $\mathbf{z}_t = (z_t^1, ..., z_t^m)$.

The energy that the input video is labeled with the task S, the intention $\mathbf{L}$, and the attention $\mathbf{Y}$ is defined as

$$\mathcal{E}(\mathbf{Y}, \mathbf{L}, S | \mathbf{I}, \mathbf{H}) = \underbrace{\sum_{t=1}^{\tau} \Phi(\mathbf{h}_t, \mathbf{a}_t, \mathbf{o}_t, l_t)}_{\text{feature matching}} + \underbrace{\sum_{t=1}^{\tau} \Psi(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, l_t)}_{\text{HAO geometric relation}} + \underbrace{\sum_{t=2}^{\tau} \Gamma(\mathbf{y}_{t-1}, \mathbf{y}_t, l_{t-1}, l_t)}_{\text{attention and intention transition}} . \quad (1)$$

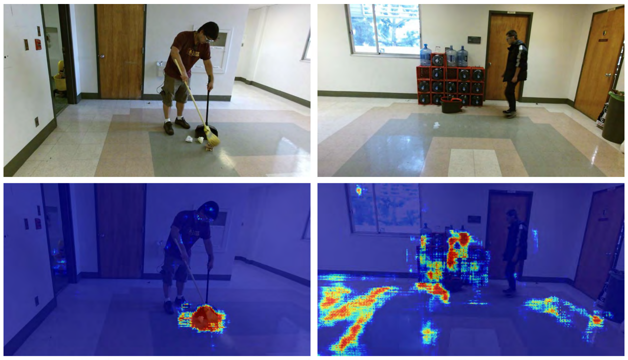
$\Phi(\cdot)$ is the feature matching energy; $\Psi(\cdot)$ describes the relations among the human body, attention, and objects; $\Gamma(\cdot)$ represents the temporal transitions of attention and intention. Since the relation between a task and its intentions is a hard constraint, we omit S in the right side of Eq.(1).

2.2. Feature Matching of HAO

The feature matching term is written as

$$\Phi(\mathbf{h}_t, \mathbf{a}_t, \mathbf{o}_t, l_t) = \phi_1(\mathbf{h}_t, l_t) + \phi_2(\mathbf{a}_t, l_t) + \phi_3(\mathbf{o}_t, l_t). \quad (2)$$

Human pose matching $\phi_1(\mathbf{h}_t, l_t)$ describes the compatibility of the pose feature $\mathbf{h}_t$ and the intention l_t. With the 3D skeleton, we compute the differences between each joint



check state of floor

locate mop

Figure 4: Attention map. Each map pixel value is the probability that the human looks at the pixel with the intention shown below.

and other joints [37], and concatenate the difference vector of each joint to form $\mathbf{h}_t$. Using pose features of all intention classes, we train a classifier with logistic regression [8] for pose classification. The probability output by the classifier is used as $p(l_t|\mathbf{h}_t)$. The energy is

$$\phi_1(\mathbf{h}_t, l_t) = -\log p(l_t|\mathbf{h}_t). \tag{3}$$

Attention feature matching $\phi_2(\mathbf{a}_t, l_t)$ describes the compatibility between the attention feature $\mathbf{a}_t$ and the intention l_t. We train a CNN classifier with the VGG16 model [30] on the square attention patch samples. The score output from the network is used as the attention patch labeling probability $p(l_t|\mathbf{a}_t)$. Fig. 4 shows two examples of the probability maps. The attention matching energy is

$$\phi_2(\mathbf{a}_t, l_t) = -\log p(l_t|\mathbf{a}_t). \tag{4}$$

Object matching represents the compatibility between the object features in the video frame and the object classes related to the intention. $(o_t^1, ..., o_t^m)$ is the object bounding boxes related to the intention l_t. We fine-tune Faster R-CNN models [26] on our training data to detect objects in each frame. The score output from the Faster R-CNN detector is used as an object's probability $p(o_t^i)$. The energy of all related objects in the frame is

$$\phi_3(\mathbf{o}_t, l_t) = -\frac{1}{m}\sum\nolimits_{i=1}^{m} \log p(o_t^i). \tag{5}$$

2.3. Geometric Relations of HAO

The human attention bridges the human body and the objects. The geometric relation term $\Psi(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, l_t)$ describes the location and direction constraint of the human pose, attention, and objects. It is written as

$$\Psi(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, l_t) = \psi_1(\mathbf{x}_t, \mathbf{y}_t, l_t) + \psi_2(\mathbf{z}_t, \mathbf{y}_t, l_t). \tag{6}$$

Human pose and attention relation $\psi_1(\mathbf{x}_t, \mathbf{y}_t, l_t)$ describes the constraint between the 3D attention direction and the human pose. In daily-activity scenes, the body part directions imply the attention directions [38]. For example, when a human manipulates objects with hands, the direction from the head to the hands implies the attention direction.

We adopt a similar method to the work [38] to model the pose and attention relations. Eleven 3D vectors are extracted from the 3D human skeleton, such as the normal vector of the head and shoulder plane, the direction from the head to the hands, etc. These 3D vectors are concatenated as the attention direction feature $\mathbf{x}_t$,

We train a regression model from the attention direction feature to the 3D attention direction with a 3-layer fully-connected neural network f. For an attention feature $\mathbf{x}_t$, the network f estimates a hypothesized 3D attention direction $f(\mathbf{x}_t)$. The relation between the human attention direction $\mathbf{y}_t$ and $f(\mathbf{x}_t)$ is defined as

$$\begin{aligned} \mathbf{y}_t &= f(\mathbf{x}_t) + \mathbf{w}_{l_t}, \\ \mathbf{w}_{l_t} &\sim \mathcal{N}(\boldsymbol{\mu}_{l_t}, \mathbf{\Sigma}_{l_t}), \end{aligned} \tag{7}$$

where $\mathbf{w}_{l_t}$ is a noise variable following Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{l_t}, \mathbf{\Sigma}_{l_t})$. The geometric energy is written as

$$\psi_1(\mathbf{x}_t, \mathbf{y}_t, l_t) = -\log \mathcal{N}(\mathbf{y}_t | f(\mathbf{x}_t) + \boldsymbol{\mu}_{l_t}, \mathbf{\Sigma}_{l_t}). \tag{8}$$

The intention l_t in $\boldsymbol{\mu}_{l_t}$ and $\mathbf{\Sigma}_{l_t}$ suggests different geometric relations in different intentions, which reflects the constraints of intentions on attention.

Attention and object relation $\psi_2(\mathbf{z}_t, \mathbf{y}_t, l_t)$ describes the constraint between the human attention location and the object locations in 3D space. The attention location is closely related to the object location, but not necessarily the same. For example, in the intention *locate mug*, the attention location shifts from the nearby areas to the mug.

Suppose $\tilde{\mathbf{y}}_t$ is the 3D attention location. It is the intersection point of the 3D attention direction $\mathbf{y}_t$ and the scene point cloud. The relation between the attention location $\tilde{\mathbf{y}}_t$ and the object bounding box o_t^i is formulated as

$$\begin{aligned} \mathbf{z}_t^i &= \tilde{\mathbf{y}}_t + \mathbf{v}_{l_t, \tilde{o}_t^i}, \\ \mathbf{v}_{l_t, \tilde{o}_t^i} &\sim \mathcal{N}(\lambda_{l_t, \tilde{o}_t^i}, \mathbf{\Lambda}_{l_t, \tilde{o}_t^i}), \end{aligned} \tag{9}$$

where $\tilde{o}_t^i$ is the object class label of the box o_t^i. $\mathbf{z}_t^i$ is the object's 3D location. $\mathbf{v}_{l_t, \tilde{o}_t^i}$ is a noise variable following Gaussian distribution $\mathcal{N}(\lambda_{l_t, \tilde{o}_t^i}, \mathbf{\Lambda}_{l_t, \tilde{o}_t^i})$. The subscripts $l_t, \tilde{o}_t^i$ in $\lambda_{l_t, \tilde{o}_t^i}$ and $\mathbf{\Lambda}_{l_t, \tilde{o}_t^i}$ suggests that the attention-object relations are different for different intentions and object classes.

The relation energy of multiple objects in the frame is

$$\psi_2(\mathbf{z}_t, \mathbf{y}_t, l_t) = -\frac{1}{m}\sum_{i=1}^{m} \log \mathcal{N}(\mathbf{z}_t | \tilde{\mathbf{y}}_t + \lambda_{l_t, \tilde{o}_t^i}, \mathbf{\Lambda}_{l_t, \tilde{o}_t^i}). \tag{10}$$

2.4. Temporal Transition of Attention and Intention

$\Gamma(\mathbf{y}_{t-1}, \mathbf{y}_t, l_{t-1}, l_t)$ represents the transitions of attention and intention in time domain. It is written as

$$\Gamma(\mathbf{y}_{t-1}, \mathbf{y}_t, l_{t-1}, l_t) = \gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t) + \gamma_2(l_{t-1}, l_t). \quad (11)$$

Attention transition $\gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t)$ describes the temporal relations between attention directions in two successive frames. It is formulated as a linear dynamic system [1, 38]:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Q}_{l_{t-1},l_t}\mathbf{y}_{t-1} + \mathbf{u}_{l_{t-1},l_t}, \\ \mathbf{u}_{l_{t-1},l_t} &\sim \mathcal{N}(0, \mathbf{\Upsilon}_{l_{t-1},l_t}), \end{aligned} \quad (12)$$

where $\mathbf{Q}_{l_{t-1},l_t}$ is the transition matrix. $\mathbf{u}_{l_{t-1},l_t}$ is a noise variable following Gaussian distribution $\mathcal{N}(0, \mathbf{\Upsilon}_{l_{t-1},l_t})$. The attention transition energy is

$$\gamma_1(\mathbf{y}_{t-1}, \mathbf{y}_t) = -\log \mathcal{N}(\mathbf{y}_t | \mathbf{Q}_{l_{t-1},l_t}\mathbf{y}_{t-1}, \mathbf{\Upsilon}_{l_{t-1},l_t}). \quad (13)$$

$\mathbf{Q}_{l_{t-1},l_t}$ and $\mathbf{\Upsilon}_{l_{t-1},l_t}$ are both related to the intentions l_{t-1} and l_t, which reflects the fact that the motion patterns of human attention are constrained by human intentions.

Intention transition $\gamma_2(l_{t-1}, l_t)$ represents the transition relations between different intentions. We model the transition as a Markov process. $p(l_t = j | l_{t-1} = i) = d_{ij}$ is the transition probability between two intentions in successive frames. The transition energy is defined as

$$\gamma_2(l_{t-1} = i, l_t = j) = -\log p(l_t = j | l_{t-1} = i). \quad (14)$$

3. Inference

Given an input RGB-D video $\mathbf{I}$ with 3D human skeletons $\mathbf{H}$, we aim to jointly output: 1) the human intention in each frame; 2) the 3D attention direction in each frame; and 3) the task label of the video. This problem is formulated as

$$(\mathbf{Y}, \mathbf{L}, S)^* = \arg\min \ \mathcal{E}(\mathbf{Y}, \mathbf{L}, S | \mathbf{I}, \mathbf{H}). \quad (15)$$

We use an algorithm similar to beam search [39] to solve Eq. (15), as shown in Fig. 5. It includes three procedures.

1) Proposing hypothesized attention points. The possible attention points on RGB images are proposed according to human poses. As introduced in Sec. 2.3, with the pose feature $\mathbf{x}_t$, a hypothesized 3D attention direction $f(\mathbf{x}_t)$ is computed with the network f. A 3D attention point derived from $f(\mathbf{x}_t)$ is projected onto the image plane to form a 2D location. Around this location, we propose a group of possible 2D attention points, as shown in Fig. 5. The point range and step are empirically defined. Each 2D point is attached a probability vector of all possible intentions computing with the attention matching model in Eq. 4.

2) Proposing hypothesized objects. We use Faster R-CNN [26] to detect all possible objects related to all the tasks and intentions in each frame, as shown in Fig. 5. Each detected box has the probabilities of all object classes.

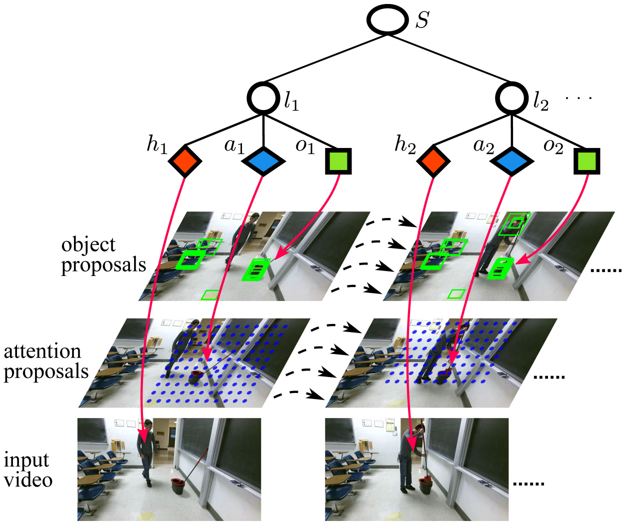


Figure 5: Inference algorithm. For clarity, only parts of the proposed object boxes and attention points are visualized.

3) Graph-guided optimization. With the hypothesized attention points and objects, the goal is to select optimal attention points, objects, intentions, and the task label in each video frame to minimize $\mathcal{E}(\mathbf{Y}, \mathbf{L}, S | \mathbf{I}, \mathbf{H})$.

From training samples, we construct HAO graphs for each task category. These graphs specify the intentions, related objects, the geometric and temporal relations. Let $\mathbf{I}_t$ be the video clip from time 1 to t. The graph-guided optimization is summarized as follows:

i) In frame I_t, all possible combinations of attention points, object bounding boxes, and intention labels for each task category are generated according to the HAO graph structure. Each of such combination is taken as one hypothesized joint label of frame I_t.

ii) The union of one joint label of I_t and one joint label sequence of the past video $\mathbf{I}_{t-1}$ forms a hypothesized joint label sequence of the video $\mathbf{I}_t$. The energy of the hypothesized joint label sequence is computed with Eq. 1. At time t, all hypothesized joint label sequences are sorted according to their energies. The J joint label sequences with lowest energies are kept and others are pruned.

iii) The step i) and step ii) are iterated frame by frame until the video ends. The joint label sequence with the lowest energy is the output result, which includes the task label, human attention and intentions for each frame.

4. Experiment

We evaluate our method with three experiments: intention prediction, attention prediction, and task recognition. Intention prediction accuracy is defined as the ratio of the

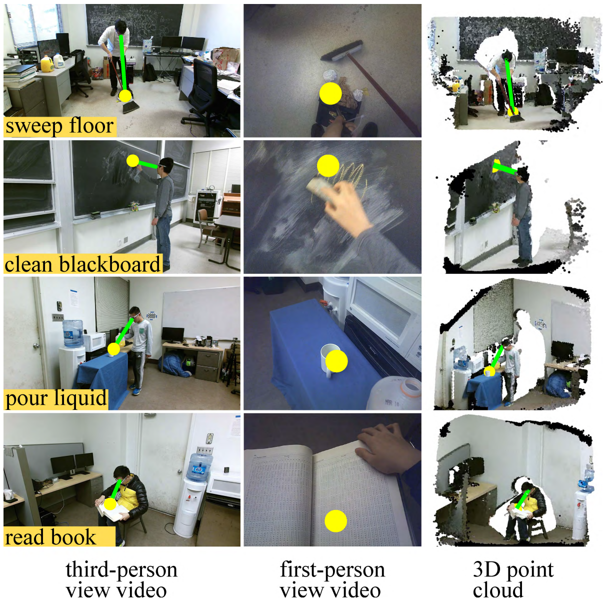


Figure 6: Samples in TIA dataset. Each row is a task.

correctly labeled frame number to all testing frame number. Attention prediction error is defined as the average distance between the predicted values and ground-truth values in all testing frames. Task recognition accuracy is the ratio of the correctly labeled video number to all testing video number.

4.1. TIA Dataset

We built a large-scale dataset of tasks, intentions, and attention (TIA). Fig. 6 shows some frame examples. The data was captured with two types of cameras simultaneously. A Kinect camera was fixed in scenes to capture RGB-D videos of human activities from a third-person view. 14 volunteers freely perform and independently accomplish different tasks in various scenes.

An eye-tracking camera was worn on volunteers' heads to capture egocentric videos with human gaze points in each frame. The egocentric videos and gaze points are used for annotating the ground-truth attention points in third-person view videos, not for training or testing in our experiment.

We manually annotated the task labels, intention labels, 2D attention points, object labels and bounding boxes in each video frame. In total, the dataset contains 809 videos and approximately 330,000 frames. Each frame includes four types of data: the RGB image at resolution of 1920 × 1080, the depth image, the 3D human skeleton, and the egocentric RGB image at resolution of 1280 × 960.

The dataset contains 14 classes of tasks: *sweep floor, mop floor, write on blackboard, clean blackboard, use elevator, pour liquid from jug, make coffee, read book, throw trash, microwave food, use computer, search drawer, move bottle to dispenser*, and *open door*. It contains 70 categories of human intentions, such as *locate broom, direct hand to hold mop, check state of microwave*, etc, and 28 classes of objects, such as *broom, mop, chalk, coffee jar, drawer*, etc.

Methods	Accuracy
SVM-JF	0.26
NN-JF	0.29
RGB Frame CNN [30]	0.17
H (human pose)	0.34
A (attention patch)	0.07
O (object feature)	0.18
H + Relation of H and A	0.36
A + Relation of H and A	0.28
A + Relation of A and O	0.25
O + Relation of A and O	0.28
Our HAO	**0.40**

Table 1: Comparison of overall inttention prediction accuracy.

4.2. Implementation Details

We divide the 809 video samples into training, validation, and testing sets with the video number ratios of 0.5, 0.25 and 0.25, respectively.

For the pose matching model in Eq. (3), we extract joint features [37] from 3D skeletons. A classifier is trained with a L2-regularized logistic regression [8].

For the attention matching model in Eq. (4), we crop attention patches with a 64×64 size centered at the ground-truth attention points. With these image patches, we train a CNN classifier with the VGG16 model [30]. The learning rate and batch size are 0.0001 and 64, respectively.

For the object matching model in Eq. (5), we fine-tune Faster R-CNN model [26] on our training data with VGG16 features [30]. The non-maximum suppression threshold and the confidence threshold are 0.6 and 0.5, respectively.

4.3. Intention Prediction

Intention prediction is to label each video frame with an intention. Table 1 shows the overall prediction accuracy of 70 intention categories. Fig. 7 shows some examples.

We compare our HAO method with other approaches, as shown in Table 1. The methods SVM-JF and NN-JF use the joint features (JF) [37] extracted from aligned 3D human skeletons. With these features, SVM-JF trains a classifier with support vector machines and NN-JF trains a three-layer fully-connected neural network. SVM-JF and NN-JF predict intentions in all testing frames with single frame features. The method RGB Frame CNN uses whole RGB frames as inputs. It trains a classifier based on VGG16 model [30]. The learning rate and batch size are 0.0001 and 64, respectively.

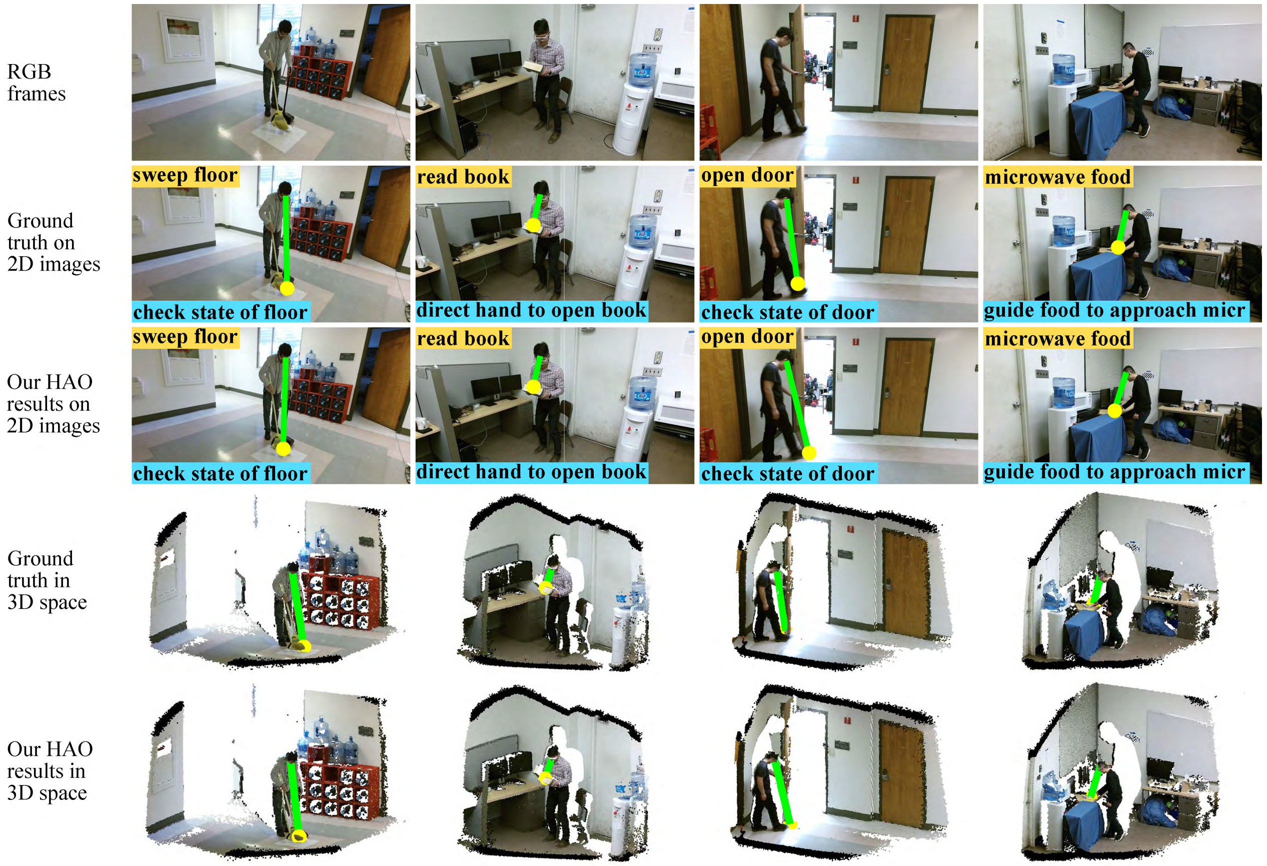


Figure 7: Visualization of intention prediction, attention prediction, and task recognition results. The texts on the RGB frames are the task label and intention label, respectively.

Our model combines the different information terms together. To diagnose the effect of each term, we compute the performance of the methods that use the information of human poses (H), attention patches (A), objects (O), and the geometric relations between them. All the diagnosis methods adopt the same model parameters and inference algorithm with HAO but only different information terms.

Tabel 1 shows that our HAO outperforms other approaches by a considerable margin. The human body features, like joint features [37] used in NN-JF and NN-JF, describe human action information. The experimental results show that it is difficult to distinguish human intentions only relying on the action features.

The RGB Frame CNN [30] method uses whole frames as inputs. The frames contain much scene and background information. Such information is valid for object and scene understanding, but less effective to distinguish human intentions, and therefore leads to a lower performance.

Our HAO exploits the joint information of human poses, attention patches, objects, and their interacting relations. Thus, it achieves better results. This is also reflected in the diagnosis results of Table 1. Using pure H, A, or O information is ineffective to predict intentions. When incorporating the relations among them, the performance improves greatly. Our HAO further improves the performance by incorporating all the information into a unified framework.

Fig. 7 shows that our HAO can reasonably predict intentions even if humans do not have obvious actions. For example, in the task *sweep floor* where a human stands, our HAO predicts that the human's intention is *check state of floor* according to the objects and the attention location.

4.4. Attention Prediction

Attention prediction is to predict the 3D attention directions, 3D and 2D attention locations in each frame. Table 2 shows the prediction errors. Fig. 7 visualizes some attention prediction results. The 3D location error with the unit of meter is defined in scene point clouds. For 3D attention directions, we normalize all the attention directions so that all direction vectors start from the 3D origin with a norm 1. The 2D location error with the unit of pixel is defined in images at resolution of 960×540.

We compare our method with Multivariate Regression (Mv-Reg), Linear Dynamic System with Kalman Filter

Methods	3D Location	3D Direction	2D Location
Mv-Reg	0.656	0.543	99
LDS-KF [1]	0.656	0.562	98
NN-Reg	0.655	0.540	97
Our HAO	**0.628**	**0.475**	**93**

Table 2: Comparison of average attention prediction errors.

(LDS-KF) [1], and a neural network regression (NN-Reg). The NN-Reg adopts a fully-connected regression network with 3 layers. All the three approaches use the same input skeleton features with our HAO.

Table 2 shows our HAO outperforms other comparison approaches. Compared to Mv-Reg, LDS-KF [1], and NN-Reg, our method jointly utilizes the information of human poses, attention patches, and objects, which impressively improves the performance both in 2D and 3D.

4.5. Task Recognition

Task recognition is to label each video with a task. We compare our HAO with several methods: 4DHOI [39], Frame CNN [15], and Two-Stream CNN [29]. 4DHOI [39] jointly uses human poses, interacting objects, and the human-object relations to label videos. Frame CNN [15] labels videos by voting based on the frame classification with CNN. Two-Stream CNN [29] combines the RGB and optical flow features with convolutional neural network to label videos. Table 3 shows the overall recognition accuracy comparison. Fig. 7 visualizes some examples.

Similar to intention prediction, we also compute the performance of methods which use human poses (H), attention patches (A), objects (O), and the relations. By analyzing the performance of these methods, we can diagnose the effects of different factors on task recognition.

Table 3 shows that our HAO method outperforms other methods. Traditional activity recognition methods mainly rely on the human appearance and motions to label videos. However, a complex task video are often very long and contains many different forms of actions, which make it hard to distinguish the task only by appearance and motion information. Our HAO decomposes tasks into intention processes, which is more flexible. It jointly takes advantages of human, attention, and object information to recognize tasks, and therefore achieves better results. The diagnosis experiment results also show the advantages of our joint model.

5. Conclusion

In this paper, we study a new problem of jointly inferring intentions, attention, and tasks from RGB-D videos. Our work develops video understanding from recognizing what humans are doing to inferring what humans are thinking. We propose a human-attention-object (HAO) graph to jointly represent tasks, attention, and intentions in videos. A task is temporally decomposed into intentions, and an intention is decomposed into the human pose, the human attention, and the related objects. Given an RGB-D video, a beam search algorithm is used to jointly infer the task labels, the intentions, and the attention. We presented a new large-scale video dataset of tasks, intentions, and attention. Experiments on intention prediction, attention prediction, and task recognition prove the strength of our approach.

Methods	Accuracy
4DHOI [39]	0.62
Frame CNN [15]	0.39
Two-Stream CNN [29]	0.54
H (human pose)	0.58
A (attention patch)	0.20
O (object feature)	0.50
H + Relation of H and A	0.61
A + Relation of H and A	0.46
A + Relation of A and O	0.54
O + Relation of A and O	0.66
Our HAO	**0.73**

Table 3: Comparison of overall task recognition accuracy.

The experiments show that human attention play significant roles on human intention and task modeling. In the future work, we will study mind modeling in robotics.

Acknowledgments

P. Wei and N.N. Zheng thank the support of grants: NSFC 61503297, NSFC 61773312, NSFC 61751401, and NSFC 91748208. Y. Liu, T.M. Shu, and S.-C. Zhu thank the support of grants: DARPA XAI N66001-17-2-4029 and ONR MURI N00014-16-1-2007.

References

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE TSP*, 50(2):174–188, 2002.

[2] A. Belardinelli, O. Herbort, and M. V. Butz. Goal-oriented gaze strategies afforded by object interaction. *Vision Research*, 106:47–57, 2015.

[3] S. Blakemore1 and J. Decety. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, 2:561–567, 2001.

[4] A. Borji, D. N.Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *CVPR*, pages 470–477, 2012.

[5] A. Borji, D. N. Sihite, and L. Itti. What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE TSMCS.*, 44(5):523–538, 2014.

[6] S. A. Butterfill and I. A. Apperly. How to construct a minimal theory of mind. *Mind and Language*, 28(5):606–637, 2013.

[7] J. Chen, Y. Tong, W. Gray, and Q. Ji. A robust 3d eye gaze tracking system using noise reduction. In *ACM Symposium on Eye Tracking Research & Applications*, pages 189–196, 2008.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, pages 1226–1233, 2012.

[10] K. A. Funes-Mora and J.-M. Odobez. Gaze estimation in the 3d space using rgb-d sensors. *IJCV*, 118(2):194–216, 2016.

[11] D. Harari, T. Gao, N. Kanwisher, J. B. Tenenbaum, and S. Ullman. Measuring and modeling the perception of natural and unconstrained gaze in humans and machines. *CoRR*, abs/1611.09819, 2016.

[12] C.-M. Huang, S. Andrist, A. Saupp, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6:1049, 2015.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

[14] L. A. Jeni and J. F. Cohn. Person-independent 3d gaze estimation using face frontalization. In *CVPR Workshop*, pages 792–800, 2016.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[16] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–1328, 1999.

[17] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, pages 3216–3223, 2013.

[18] Y. Liu, P. Wei, and S.-C. Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *ICCV*, pages 2943–2951, 2017.

[19] M. Mansouryar, J. Steil, Y. Sugano, and A. Bulling. 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. In *ACM Symposium on Eye Tracking Research & Applications*, pages 197–200, 2016.

[20] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014.

[21] P. Moors, F. Germeys, I. Pomianowska, and K. Verfaillie. Perceiving where another person is looking: the integration of head and body information in estimating another persons gaze. *Frontiers in Psychology*, 6:909, 2015.

[22] D. Parks, A. Borji, and L. Itti. Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision Research*, 116:113–126, 2015.

[23] J. Pelz, M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139:266277, 2001.

[24] H. C. Ravichandar, A. Kumar, and A. Dani. Bayesian human intention inference through multiple model filtering with gaze-based priors. In *International Conference on Information Fusion*, pages 2296–2302, 2016.

[25] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *NIPS*, pages 199–207, 2015.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[27] B. M. Scassellati. Foundations for a theory of mind for a humanoid robot. *MIT PhD Thesis*, 2001.

[28] B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80(12):1–46, 2001.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576. 2014.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[31] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic. Predicting human intention in visual observations of hand/object interactions. In *ICRA*, pages 1608–1615, 2013.

[32] F. Stulp, J. Grizou, B. Busch, and M. Lopes. Facilitating intention prediction for humans by optimizing robot motions. In *IROS*, pages 1249–1255, 2015.

[33] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *CVPR*, pages 1821–1828, 2014.

[34] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113:766–786, 2006.

[35] S. Ullman, D. Harari, and N. Dorfman. From simple innate biases to complex visual concepts. *PNAS*, 109(44):1821518220, 2012.

[36] C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, pages 2997–3005, 2016.

[37] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE TPAMI*, 36(5):914–927, 2014.

[38] P. Wei, D. Xie, N. Zheng, and S.-C. Zhu. Inferring human attention by learning latent intentions. In *IJCAI*, pages 1297–1303, 2017.

[39] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *TPAMI*, 39(6):1165–1179, 2017.

[40] D. Xie. Inferring the intentions and attentions of agents from videos. *UCLA PhD Thesis*, 2016.

[41] C. Yu and L. B. Smith. Linking joint attention with hand-eye coordination a sensorimotor approach to understanding child-parent social interaction. In *CogSci*, pages 2763–2768, 2015.

[42] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, pages 4511–4520, 2015.