Inferring Human Attention by Learning Latent Intentions

Ping Wei[1,2], **Dan Xie**[2], **Nanning Zheng**[1], **Song-Chun Zhu**[2]
[1]Xi'an Jiaotong University, Xi'an, China
[2]University of California, Los Angeles, Los Angeles, USA
{pingwei.pw, xiedwill}@gmail.com, nnzheng@mail.xjtu.edu.cn, sczhu@stat.ucla.edu

Abstract

This paper addresses the problem of inferring 3D human attention in RGB-D videos at scene scale. 3D human attention describes where a human is looking in 3D scenes. We propose a probabilistic method to jointly model attention, intentions, and their interactions. Latent intentions guide human attention which conversely reveals the intention features. This mutual interaction makes attention inference a joint optimization with latent intentions. An EM-based approach is adopted to learn the latent intentions and model parameters. Given an RGB-D video with 3D human skeletons, a joint-state dynamic programming algorithm is utilized to jointly infer the latent intentions, the 3D attention directions, and the attention voxels in scene point clouds. Experiments on a new 3D human attention dataset prove the strength of our method.

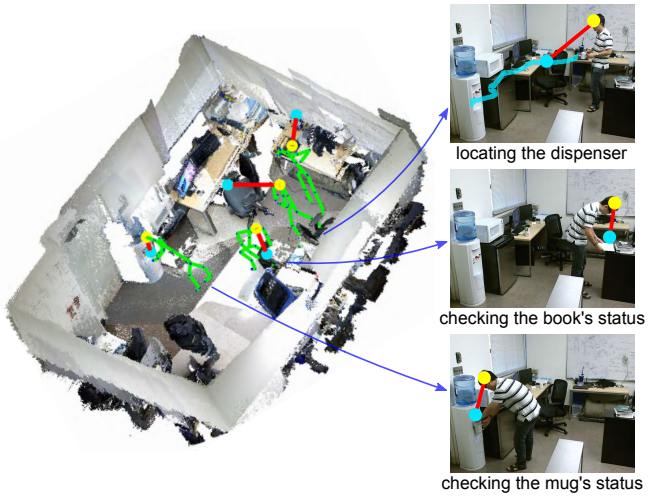

Figure 1: Human attention and intentions in a 3D scene.

1 Introduction

Inferring 3D human attention is an important issue in many applications. For example, in a task of human-robot collaboration, perceiving where the human is looking in the 3D scene is crucial for the robot to infer the human's intentions, and therefore to communicate or interact with the human.

Inferring 3D human attention at scene scale is a challenging problem. First, in 3D space, human attention has weak observable features but huge degrees of freedom. As Figure 1 shows, at the scale of daily-activity scenes, it is hard to obtain effective features of eyes or faces that are directly related to the human attention. Moreover, the human activity sequence data captured by RGB-D sensors are noisy. Different human activities present various poses, motions, and views, which makes it hard to precisely estimating the attention across different activities.

Human attention is related to human intentions [Land *et al.*, 1999]. The attention driven by different intentions presents different observation features and motion patterns. Land *et al.* [Land *et al.*, 1999] divided the roles of human fixations into four categories: *locating objects*, *directing hands*, *guiding an object to approach another*, and *checking an object's status*. As Figure 1 shows, when the person's intention is to *locate the dispenser*, his attention sweeps from the table to the dispenser; while fetching water from the dispenser, his intention is to *check* if the mug is full and his attention steadily focuses on the mug.

The driving rules of intentions acting on attention can be independent of activity categories. For example, in Figure 1, the attention driven by the intention *checking status* always presents as *steadily focusing*, even in different activities. This phenomenon makes it possible to infer the attention with the same rules across different activities. However, these driving rules are hidden and should be learned from data.

This paper proposes a probabilistic method to infer 3D human attention by jointly modeling attention, intentions, and their interactions. The attention and intention are represented with features extracted from human skeletons and scene voxels. Human intentions are taken as latent variables which guide the motions and forms of human attention. Conversely, the human attention reveals the intention features. Attention inference is modeled as a joint optimization with latent human intentions.

We adopt an EM-based [Bishop, 2006] approach to learn the model parameters and mine the latent intentions. Given an RGB-D video with human skeletons captured by the Kinect camera, a joint-state dynamic programming algorithm

is utilized to jointly infer the latent intention, the 3D attention direction, and the attention voxel in each video frame.

We collected a new dataset of 3D human attention. This dataset includes 14 categories of human activities and 150 RGB-D videos with 3D human skeletons. The experimental results on this dataset prove the strength of our method.

1.1 Related Work

Attention in psychology and cognition. Human attention has been intensively studied in psychology and cognitive science [Yarbus, 1967; Land *et al.*, 1999; Scholl, 2001; Yu and Smith, 2015]. Some studies indicate that human attention is related to human intentions and objects [Land *et al.*, 1999; Scholl, 2001]. Land *et al.* [Land *et al.*, 1999] defined four roles of human fixations. Scholl [Scholl, 2001] presented that attention was object-based. These works inspire us to study computational models for attention inference.

2D Attention. To model attention in images or videos, bottom-up and top-down cues are utilized [Itti *et al.*, 1998; Hou and Zhang, 2008; Liu *et al.*, 2011; Damen *et al.*, 2016; Recasens *et al.*, 2015; Duan *et al.*, 2013; Mnih *et al.*, 2014; Benfold and Reid, 2009; Marin-Jimenez *et al.*, 2014; Zhang *et al.*, 2015; Li *et al.*, 2013; Borji *et al.*, 2012; Fathi *et al.*, 2012; Belardinelli *et al.*, 2015]. Visual saliency focuses on the attention of humans who look at the image [Itti *et al.*, 1998; Hou and Zhang, 2008; Liu *et al.*, 2011]. Recasens *et al*. [Recasens *et al.*, 2015] combined saliency maps and human head information to follow gazes in 2D images. Fathi et al. [Fathi *et al.*, 2012] jointly modeled gazes and actions in videos.

In some cases, incorporating top-down information, like human activities, into attention modeling might be ineffective. This is because a human who is performing an activity does not necessarily look at the targets that are related to the activity. Our model incorporates latent intentions into human attention and mines the hidden driving rules to infer attention.

3D Attention. Many works have been done on 3D human attention [Sugano *et al.*, 2014; Jeni and Cohn, 2016; Funes-Mora and Odobez, 2016; Mansouryar *et al.*, 2016; Chen *et al.*, 2008; Lanillos *et al.*, 2015]. Funes-Mora and Odobez [Funes-Mora and Odobez, 2016] estimated gaze directions based on head poses and eye images. Chen *et al.* [Chen *et al.*, 2008] estimated 3D gaze directions with eye features.

Many studies estimate 3D gazes based on eye or face features. In daily-activity scenes where the captured videos are with low resolution, the eye or face features are hard to be obtained. Our method utilizes human skeleton and scene features to infer not only attention directions but also the attention voxels in scenes. It does not need eye or face features and therefore can be applied in larger scenes.

2 Attention and Intention Representation

Each video frame includes RGB-D data and a 3D human skeleton, which are recorded by a Kinect camera. The 3D human skeleton is a collection of 3D location coordinates of the body joints, as shown in Figure 2 (a). The scene point clouds defined by the scene depth data are converted into voxels, as shown in Figure 2(b). A voxel is a cube in 3D point clouds and it is like a pixel in 2D images. We define attention and intention features based on the 3D human skeletons and the scene voxels.

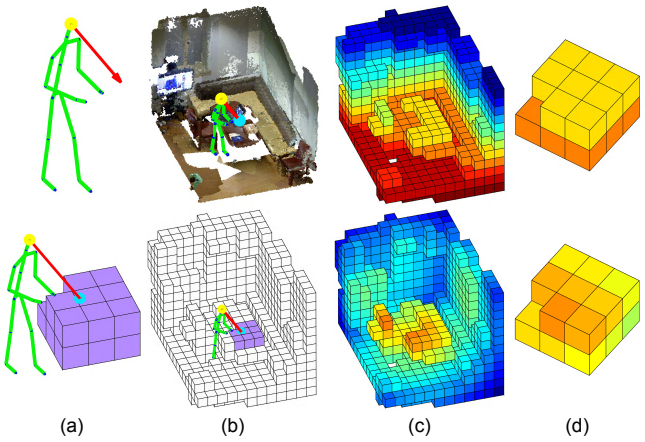

Figure 2: Attention and intention representation. (a) The attention direction and voxel. (b) Scene point clouds and voxels. (c) The voxel height and distance. (d) The voxel features.

2.1 Attention

In 3D space, human attention includes two attributes: the direction and the voxel, as shown in Figure 2(a). The attention direction is a 3D vector with unit length which describes the sight line direction from the human head to what is looked at. In the attention direction, the voxel at which the sight line intersects with the scene point clouds is the attention voxel.

In daily activities, the directions of human body parts imply the attention directions. For example, when a human is manipulating an object with the hands, the directions from the head to the hands strongly signal the attention direction. We define the observation features of attention directions with eight directions extracted from human skeletons, such as the normal vector of the head and shoulder plane, the directions from the head to the hands, etc.

To normalize the data, all human skeletons are aligned to a reference skeleton with similarity transformation. The eight observation directions are defined on the aligned skeletons. The encapsulation of the eight normalized direction vectors is the observation feature of the attention.

2.2 Intention

In our work, intentions are discrete latent variables and describe the human attention motivation. The observation feature of an intention is the encapsulation of the attention feature and the voxel feature. The attention feature is defined in Section 2.1. It characterizes the attention direction patterns in intentions.

The voxel feature is defined with the attention voxel and its neighbouring voxels, as shown in Figure 2(c) and Figure 2(d). The voxel feature is composed of the height part and the distance part. Around the attention voxel, we define a $Nx \times Ny \times Nz$ cubic grid of voxels, where Nx, Ny, and Nz are voxel numbers along the axis X, Y, and Z, respectively. The height feature is a $Nx \times Ny \times Nz$-dimensional vector

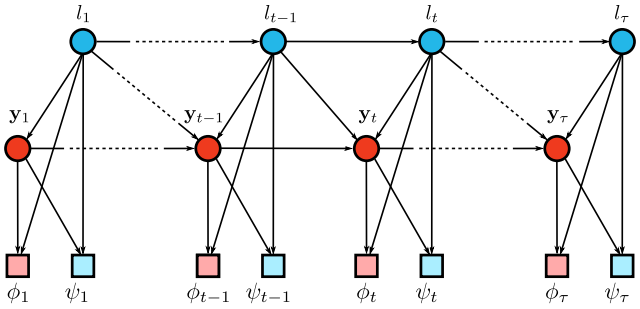

Figure 3: Joint probabilistic model of human attention and latent intentions.

whose entries correspond to the $Nx \times Ny \times Nz$ voxels in the cubic grid. The value of each entry is the height of the corresponding voxel relative to the floor. The distance feature is defined in a similar way but the vector entry value is the distance from the voxel to the human head.

The height feature reflects the spatial configuration of the attention voxels. The distance feature characterizes the human-scene interaction.

3 Model

Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_\tau)$ be a video sequence of length τ. Each video frame $\mathbf{x}_t$ includes a 3D human skeleton and the scene voxels. Given $\mathbf{X}$, the goal is to infer the attention direction and the attention voxel in each video frame. Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_\tau)$ be the attention direction sequence, where $\mathbf{y}_t$ denotes the attention direction in frame $\mathbf{x}_t$.

In each frame, we introduce a latent variable l_t to represent the latent intention. $\mathbf{l} = (l_1, ..., l_\tau)$ denotes the intention sequence of all the frames in $\mathbf{X}$.

We use a probabilistic model to jointly represent $\mathbf{X}$, $\mathbf{l}$, $\mathbf{Y}$, and their relations in time and 3D space, as shown in Figure 3. The joint probability is

$$\begin{aligned} p(\mathbf{X}, \mathbf{l}, \mathbf{Y}|\boldsymbol{\theta}) &= p(l_1) \prod_{t=1}^{\tau} p(\psi(\mathbf{x}_t)|l_t, \mathbf{y}_t) \prod_{t=2}^{\tau} p(l_t|l_{t-1}) \\ &\cdot p(\mathbf{y}_1|l_1) \prod_{t=1}^{\tau} p(\phi(\mathbf{x}_t)|\mathbf{y}_t, l_t) \prod_{t=2}^{\tau} p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1}). \end{aligned} \tag{1}$$

$\boldsymbol{\theta}$ is the set of model parameters. $\psi(\mathbf{x}_t)$ and $\phi(\mathbf{x}_t)$ are the intention and attention features, respectively, extracted from frame $\mathbf{x}_t$ as defined in Section 2. They are abbreviated as ψ_t and ϕ_t below. $p(\psi_t|l_t, \mathbf{y}_t)$ represents the intention identification and $p(\phi_t|\mathbf{y}_t, l_t)$ is the attention observation probability.

$p(l_t|l_{t-1})$ and $p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1})$ describe transition relations of intentions and attention in two successive frames, respectively. $p(l_1)$ and $p(\mathbf{y}_1|l_1)$ characterize the initial states of the intention and the attention, respectively.

As Figure 3 shows, our model is a joint representation of the intention and the attention. The intentions guide not only the attention observations but also the attention transitions. Conversely, the intention observation features depend on the voxels observed by the human.

Our model is similar but different from the switching dynamic models [Kim, 1994; Ghahramani and Hinton, 2000]. In our model, the latent variables of attention and intentions have different observation features.

3.1 Attention Model

We model human attention under the framework of the linear dynamic system (LDS) [Bishop, 2006]. Different from the conventional LDS, we introduce an additional layer of latent variables to control the observation and motion patterns.

Initial attention $\mathbf{y}_1$ is modeled as:

$$\begin{aligned} \mathbf{y}_1 &= \boldsymbol{\mu}_{l_1} + u, \\ u &\sim \mathcal{N}(0, \mathbf{V}_{l_1}), \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{l_1}$ is the prior value of $\mathbf{y}_1$ conditioned on intention l_1. u is the noise which follows Gaussian distribution with mean 0 and covariance $\mathbf{V}_{l_1}$. The initial attention probability is

$$p(\mathbf{y}_1|l_1) = \mathcal{N}(\mathbf{y}_1|\boldsymbol{\mu}_{l_1}, \mathbf{V}_{l_1}). \tag{3}$$

Attention observation describes the generation relation of the attention and the observation, which is formulated as:

$$\begin{aligned} \phi_t &= \mathbf{C}_{l_t}\mathbf{y}_t + \mathbf{v}, \\ \mathbf{v} &\sim \mathcal{N}(0, \boldsymbol{\Sigma}_{l_t}), \end{aligned} \tag{4}$$

where $\mathbf{v}$ is the noise which follows Gaussian distribution with mean 0 and covariance $\boldsymbol{\Sigma}_{l_t}$. The generation matrix $\mathbf{C}_{l_t}$ is governed by the intention l_t, which reflects the intention constraints on the attention observations. The attention observation probability is

$$p(\phi_t|\mathbf{y}_t, l_t) = \mathcal{N}(\phi_t|\mathbf{C}_{l_t}\mathbf{y}_t, \boldsymbol{\Sigma}_{l_t}). \tag{5}$$

Attention transition describes the temporal relations between attention in successive frames, which is formulated as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}_{l_{t-1},l_t}\mathbf{y}_{t-1} + \mathbf{w}, \\ \mathbf{w} &\sim \mathcal{N}(0, \boldsymbol{\Gamma}_{l_{t-1},l_t}), \end{aligned} \tag{6}$$

where $\mathbf{w}$ is the noise which follows Gaussian distribution with mean 0 and covariance $\boldsymbol{\Gamma}_{l_{t-1},l_t}$. The transition matrix $\mathbf{A}_{l_{t-1},l_t}$ is related to the intentions in successive frames, which reflects the intention constraints on the attention motions. The transition probability model is

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, l_t, l_{t-1}) = \mathcal{N}(\mathbf{y}_t|\mathbf{A}_{l_{t-1},l_t}\mathbf{y}_{t-1}, \boldsymbol{\Gamma}_{l_{t-1},l_t}). \tag{7}$$

3.2 Intention Model

Intention model is composed of three parts: initial intention, intention identification, and intention transition.

Initial intention describes the prior knowledge about the intention in the first frame. It is formulated as:

$$p(l_1 = i) = \boldsymbol{\lambda}^i, \tag{8}$$

where $\boldsymbol{\lambda}$ is a discrete probability vector, and its ith entry $\boldsymbol{\lambda}^i$ represents the probability of the ith intention category.

Intention identification is formulated as

$$p(\psi_t|l_t, \mathbf{y}_t) \propto p(l_t|\psi_t, \mathbf{y}_t, \boldsymbol{\omega}). \tag{9}$$

$p(l_t|\psi_t, \mathbf{y}_t, \boldsymbol{\omega})$ is the posterior probability and $\boldsymbol{\omega}$ is the parameter of a linear classifier. The classifier is trained with

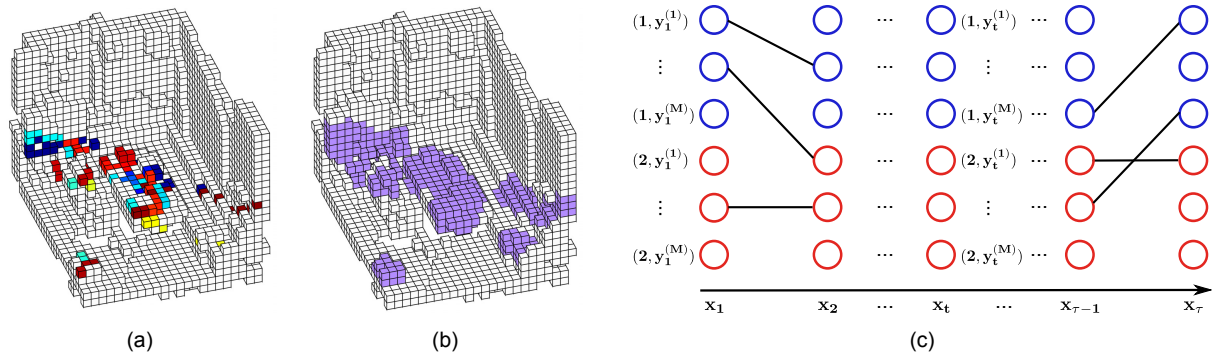

Figure 4: Joint-state dynamic programming. (a) Seed voxels in a video. The warmer colors indicate more recent time. (b) Candidate voxels. (c) Inference on a sequence, where two intention states are used to illustrate the algorithm.

Support Vector Machine and the scores output by the classifier are converted to probabilities [Chang and Lin, 2011].

The intention observation is dependent on the attention voxels related to the attention direction $\mathbf{y}_t$, which reflects the joint relations between the intentions and the attention.

Intention transition describes the relations of intentions in two successive frames, which is represented as

$$p(l_t = j | l_{t-1} = i) = \mathbf{\Lambda}^{ij}, \tag{10}$$

where $\mathbf{\Lambda}$ is the transition matrix. The entry $\mathbf{\Lambda}^{ij}$ in the ith row and jth column is the probability of the transition from the ith intention category to the j intention category.

4 Inference

Given an RGB-D video $\mathbf{X}$, we aim to infer the 3D human attention in each video frame, which is formulated as

$$\mathbf{Y}^* = \arg\max \; p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}), \tag{11}$$

where

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \sum\nolimits_{\mathbf{l}} p(\mathbf{l}, \mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}).$$

Dynamic programming is one of the most widely-used algorithms to interpret temporal sequences [Forney, 1973]. However, the attention and intentions are correlated, which means the conventional dynamic programming is inapplicable in our task.

We adopted a joint-state dynamic programming method to solve Equation (11). The general procedures of the algorithm include: 1) in each video frame, a seed voxel is proposed, as shown in Figure 4 (a); 2) the seed voxel generates candidate attention voxels in a cube around the seed, as shown in Figure 4 (b); 3) the candidate voxels and all intentions are combined to form joint states; a joint state includes an attention voxel (direction) and an intention; 4) the dynamic programming [Forney, 1973] is performed on these joint states to produce the attention voxels (directions) and the latent intentions, as shown in Figure 4 (c).

In each frame, we use attention features extracted from human skeletons to propose possible attention directions, which intersect with the scene to produce the seed voxels. Around the seed voxel, we define a cube containing M neighbouring voxels as candidate attention voxels. Connecting the human head and these candidate voxels generates a set of candidate directions $\mathcal{Y}_t = \{\mathbf{y}_t^{(1)}, ..., \mathbf{y}_t^{(M)}\}$. In each frame, the joint state space is formed with $\mathcal{Y}_t$ and all possible intentions.

5 Learning

Let $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \mathbf{V}_i, \mathbf{C}_i, \mathbf{\Sigma}_i, \mathbf{A}_{ij}, \mathbf{\Gamma}_{ij}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{\Lambda}\}$ be all the parameters of the model. The subscripts i, j indicate parameters of different intentions. Given N videos and their attention sequences $\{(\mathbf{X}^1, \mathbf{Y}^1), ..., (\mathbf{X}^N, \mathbf{Y}^N)\}$, the goal is to learn $\boldsymbol{\theta}$ from the N samples by maximizing the likelihood function,

$$\boldsymbol{\theta}^* = \arg\max \sum_{n=1}^{N} \ln p(\mathbf{X}^n, \mathbf{Y}^n|\boldsymbol{\theta}), \tag{12}$$

where

$$p(\mathbf{X}^n, \mathbf{Y}^n|\boldsymbol{\theta}) = \sum_{\mathbf{l}^n} p(\mathbf{X}^n, \mathbf{l}^n, \mathbf{Y}^n|\boldsymbol{\theta}).$$

$\mathbf{l}^n$ is the latent intention sequence of the nth video sample.

Inspired by the general EM algorithm [Bishop, 2006], we optimize Equation (12) with the following steps.

1) Initialize $\mathbf{l}^n$ for each training sequence $(n = 1, ..., N)$ and compute corresponding $\boldsymbol{\theta}^{\text{old}}$ with Equation (14).
2) Compute the optimal latent intention sequence $\mathbf{l}^{n^*}$ for each training sequence $(n = 1, ..., N)$,

$$\mathbf{l}^{n^*} = \arg\max p(\mathbf{l}^n|\mathbf{X}^n, \mathbf{Y}^n, \boldsymbol{\theta}^{\text{old}}). \tag{13}$$

3) Compute new parameter $\boldsymbol{\theta}^{\text{new}}$ by optimizing

$$\boldsymbol{\theta}^{\text{new}} = \arg\max \sum\nolimits_{n=1}^{N} \ln p(\mathbf{X}^n, \mathbf{l}^{n^*}, \mathbf{Y}^n|\boldsymbol{\theta}) \tag{14}$$

4) If the convergence condition is met, stop and output the results; else set $\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{\text{new}}$ and return to step 2).

In step 1), we use k-means to cluster the intention features and produce the initial intention labels. In step 2), we compute the optimal latent intention sequence $\mathbf{l}^{n^*}$ with the dynamic programming. In step 3), Equation (14) is optimized by computing derivatives of the log likelihood function with respect to the parameters.

	Direction Error	Voxel Error
Multi-Reg	0.63	0.84
LDS-KF	0.66	0.89
Our Method	**0.60**	**0.79**

Table 1: The overall performance comparison of different methods on attention prediction.

6 Experiment

6.1 3D Attention Dataset

We collected a 3D attention dataset. Volunteers freely performed daily activities in several scenes. A Kinect camera was used to capture the RGB-D videos and 3D human skeletons. We also scanned and synthesized the point clouds of the whole scenes where the activities were performed. The groundtruth of 3D attention locations were manually annotated in the synthesized scenes.

This dataset includes a total of 150 RGB-D videos with 3D human skeletons and 14 activity categories: *drink water with mug*, *drink water from fountain*, *mop floor*, *fetch water from dispenser*, *fetch object from box*, *write on whiteboard*, *move bottle*, *write on paper*, *watch TV*, *throw trash*, *use computer*, *use elevator*, *use microwave*, and *use refrigerator*.

6.2 Evaluation

Our evaluations include predictions of attention directions and attention voxels. We use the average distance between the predicted value and the groundtruth value as the prediction error. Each attention direction vector is normalized to unit length. For attention voxels, the distance is defined between the centers of the predicted voxel and the groundtruth voxel. Sequence data is transformed into the whole synthesized scene. The voxel features of intentions are computed in the synthesized scenes.

We compare our method with two methods: Multivariate Regression (Multi-Reg) and LDS with Kalman Filter (LDS-KF) [Arulampalam *et al.*, 2002]. Multi-Reg estimates the attention directions with a multivariate linear model. Kalman Filter estimates the attention directions in a linear dynamic system. The comparison methods use the same attention features with our method.

Table 1 shows the overall performance of different methods on attention predictions. Our method achieves better results than the other methods. Compared to the Kalman method, our method achieves impressive improvement in performance by introducing latent intentions.

Table 2 and Table 3 show the performance on different activity categories. In many activity categories, our method achieves better performance than other methods. These results prove that the proposed method can be used in different activity categories.

Figure 5 visualizes examples of the attention voxel prediction. The human attention in video 1 moves a large range while in video 2 it focuses on a small area. Though the two videos present different attention patterns, our method can reasonably predict the attention.

Figure 6 visualizes examples of the attention direction prediction. Those skeletons are noisy, contorted, and present various poses, such as *bend down, sit, stand, raise leg*, etc. Despite the challenges, our method achieves reasonable predictions. These examples also show that our method estimates the human attention independent of specific activity categories, which is a favorable characteristic in real applications.

7 Conclusion

This paper presents a method to infer 3D human attention in RGB-D videos. We model the attention inference as a joint optimization with latent intentions. An EM-based learning algorithm is utilized to learn the latent intentions and the model parameters. We adopt a joint-state dynamic programming algorithm to infer the latent intentions and the 3D human attention. We collected a new dataset of 3D human attention in RGB-D videos. Experimental results prove the strength of our method. Our future work will focus on the related issues of human mind modeling and human-robot collaboration.

Acknowledgments

This research was supported by the grants DARPA SIMPLEX project N66001-15-C-4035, a MURI project N00014-16-1-2007, NSF IIS-1423305, and NSFC 61503297.

References

[Arulampalam *et al.*, 2002] M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[Belardinelli *et al.*, 2015] Anna Belardinelli, Oliver Herbort, and Martin V. Butz. Goal-oriented gaze strategies afforded by object interaction. *Vision Research*, 106:47–57, 2015.

[Benfold and Reid, 2009] Ben Benfold and Ian Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, pages 1–11, 2009.

[Bishop, 2006] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[Borji *et al.*, 2012] Ali Borji, Dicky N. Sihite, and Laurent Itti. Probabilistic learning of task-specic visual attention. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 470–477, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Chen *et al.*, 2008] Jixu Chen, Yan Tong, Wayne Gray, and Qiang Ji. A robust 3d eye gaze tracking system using noise reduction. In *Proceedings of the Symposium on Eye Tracking Research & Applications*, pages 189–196, 2008.

[Damen *et al.*, 2016] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
Multi-Reg	**0.50**	**0.73**	0.68	0.67	0.40	**0.68**	**0.77**	0.47	0.59	0.53	0.48	0.80	0.56	**0.66**
LDS-KF	0.52	0.76	0.72	0.73	**0.37**	0.72	0.85	0.54	0.69	**0.52**	**0.45**	0.91	0.57	0.74
Our Method	0.56	0.76	**0.64**	**0.66**	**0.37**	0.70	0.84	**0.46**	**0.56**	0.53	0.47	**0.63**	**0.49**	**0.66**

Table 2: Prediction errors of attention directions on different activities, where 'A' denotes 'activity'.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
Multi-Reg	**0.61**	1.32	1.27	0.99	0.67	**0.61**	1.31	**0.50**	0.81	0.77	0.44	1.63	0.58	**0.71**
LDS-KF	0.66	**1.31**	1.30	**0.98**	**0.65**	0.63	1.25	0.80	0.86	**0.75**	**0.43**	1.65	0.67	0.83
Our Method	0.74	1.39	**1.16**	**0.98**	**0.65**	0.62	**1.23**	0.58	**0.67**	0.76	0.44	**1.10**	**0.52**	0.76

Table 3: Prediction errors of attention voxels on different activities.

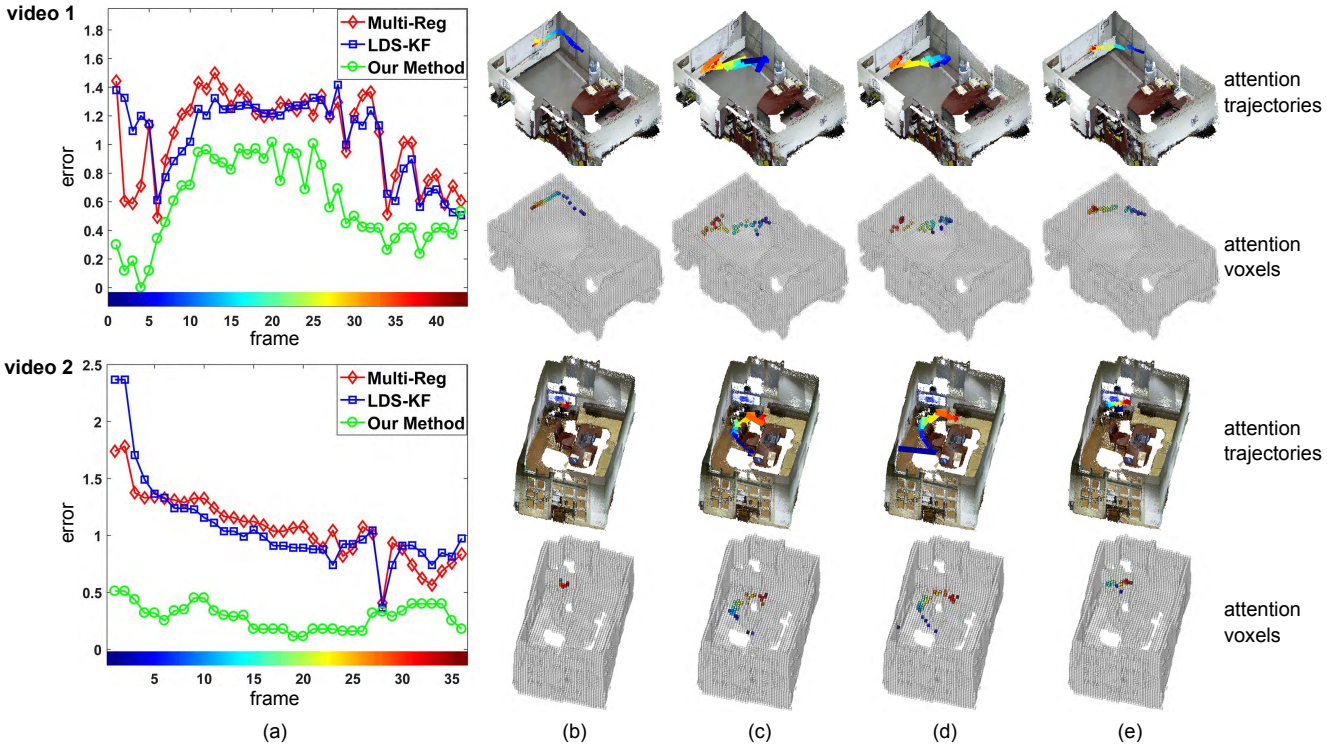

Figure 5: Attention voxel visualization. Attention voxel and trajectory colors denote temporal orders. Warmer colors mean more recent time. (a) Voxel error curves. (b) Groundtruth. (c) Multi-variable regression. (d) Kalman filter. (e) Our method.

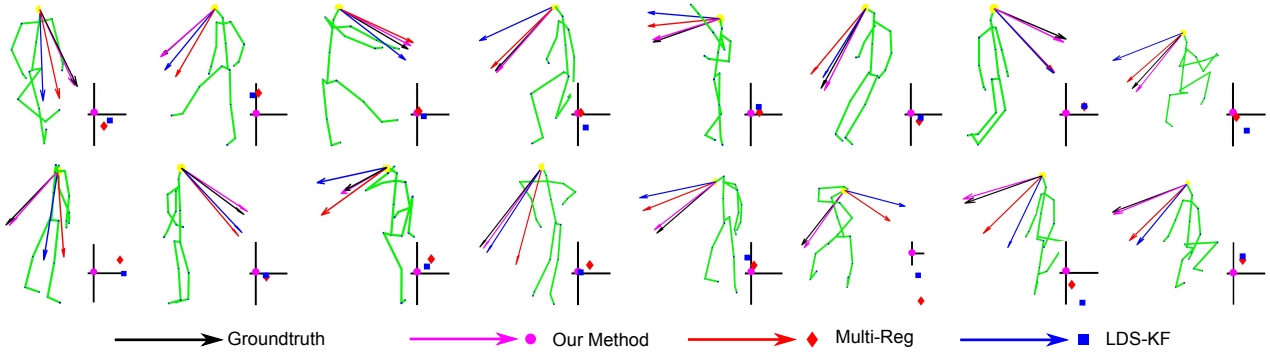

Figure 6: Attention direction visualization. The dots in the 2D coordinate systems are intersection points of attention directions with the plane orthogonal to the grundtruth attention direction. The coordinate system origins are groundtruth direction points.

[Duan *et al.*, 2013] Dingrui Duan, Lu Tian, Jinshi Cui, Li Wang, Hongbin Zha, and Hamid Aghajan. Gaze estimation in childrens peer-play scenarios. In *Proceedings of the 2nd Asian Conference on Pattern Recognition*, pages 760–764, 2013.

[Fathi *et al.*, 2012] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the 12th European Conference on Computer Vision*, pages 314–327, 2012.

[Forney, 1973] G. David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[Funes-Mora and Odobez, 2016] Kenneth A. Funes-Mora and Jean-Marc Odobez. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118(2):194–216, 2016.

[Ghahramani and Hinton, 2000] Zoubin Ghahramani and Geoffrey E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864, 2000.

[Hou and Zhang, 2008] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *Proceedings of the 21th International Conference on Neural Information Processing Systems*, pages 681–688, 2008.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[Jeni and Cohn, 2016] László A. Jeni and Jeffrey F. Cohn. Person-independent 3d gaze estimation using face frontalization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 792–800, 2016.

[Kim, 1994] Chang-Jin Kim. Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1):1 – 22, 1994.

[Land *et al.*, 1999] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

[Lanillos *et al.*, 2015] Pablo Lanillos, João Filipe Ferreira, and Jorge Dias. Multisensory 3d saliency for artificial attention systems. In *Proceedings of 3rd Workshop on Recognition and Action for Scene Understanding, 16th International Conference of Computer Analysis of Images and Patterns*, pages 1–13, 2015.

[Li *et al.*, 2013] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013.

[Liu *et al.*, 2011] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[Mansouryar *et al.*, 2016] Mohsen Mansouryar, Julian Steil, Yusuke Sugano, and Andreas Bulling. 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 197–200, 2016.

[Marin-Jimenez *et al.*, 2014] Manuel J. Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2204–2212, 2014.

[Recasens *et al.*, 2015] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Proceedings of International Conference on Neural Information Processing Systems*, pages 199–207, 2015.

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

[Sugano *et al.*, 2014] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014.

[Yarbus, 1967] Alfred L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.

[Yu and Smith, 2015] Chen Yu and Linda B. Smith. Linking joint attention with hand-eye coordination a sensorimotor approach to understanding child-parent social interaction. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2763–2768, 2015.

[Zhang *et al.*, 2015] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015.