DIPD: Gaze-Based Intention Inference in Dynamic Environments

Yu-Sian Jiang, Garrett Warnell, and Peter Stone
University of Texas at Austin, TX 78705
Computational and Information Sciences Directorate, US Army Research Laboratory, Adelphi, MD 20783
sharonjiang@utexas.edu, garrett.a.warnell.civ@mail.mil, pstone@cs.utexas.edu

Abstract

The ability of an autonomous system to understand something about a human's intent is important to the success of many systems that involve both humans and autonomous agents. In this work, we consider the specific setting of a human passenger riding in an autonomous vehicle, where the passenger intends to go to or learn about a specific point of interest along the vehicle's route. In this setting, we seek to provide the vehicle with the ability to infer this point of interest using real-time gaze information. This is a difficult problem in that the inference must be designed in the context of the moving vehicle, i.e., in a dynamic environment with dynamic interest points. We propose here a solution to this problem via a novel methodology called Dynamic Interest Point Detection (DIPD) for inferring the point of interest corresponding to the human's intent using gaze tracking data and a dynamic Markov Random Field (MRF) model. The energy function we develop allows the algorithm to successfully filter out noise from the eye tracker, such as eye blinks, high-speed tracking misalignment, and other sources of error. We demonstrate the success of this DIPD technique experimentally and show that it achieves up to a 28% increase in inference success compared to a nearest-neighbor approach.

Introduction

The ability of an autonomous system to understand something about a human's intent is important to the success of many systems that involve both humans and agents. Making inferences about human intent, for example, a robot can collaborate with a human more safely and efficiently (Mainprice, Hayne, and Berenson 2015); an intelligent human-computer interface can provide assistance to a user without an explicit user request (Yu, Ballard, and Zhu 2002); and a Driver Assistance System (DAS) can compensate for dangerous circumstances or cooperate with the driver (Doshi and Trivedi 2011; Bengler et al. 2014). While there are many forms of human intent, we choose here to focus on those that can be associated with a certain spatial location. For instance, in human-robot collaboration tasks, it is sometimes assumed that the human intends to interact with particular objects on the table (Ravichandar, Kumar, and Dani 2016). In this paper, we will assume that the human's intent is of this form and, therefore, that the intent inference problem is simply that of inferring the correct spatial location associated with the intent.

As a means to perform this inference, we focus on the human's eye gaze. Neurophsychology studies have suggested that, by observing a partner's gaze, humans can infer their partner's intention or goal towards a particular object (Calder et al. 2002). Therefore, we expect that enabling automated agents with a similar ability will provide a better user experience in human-machine interaction. Indeed, several examples in the literature have demonstrated that an autonomous agent utilizing human gaze cues can better interpret the human's intent and thus make for a better partner (Fletcher et al. 2005; Choi, Hong, and Kim 2016; Tall et al. 2009; Matsumotot, Ino, and Ogsawara 2001; Razin and Feigh 2017; Min et al. 2017).

We are particularly concerned with the setting in which a human passenger rides in an autonomous vehicle, and we assume that the passenger's intent is to go to or learn about a specific point of interest along the vehicle's route. Although a human's point of interest may not fully align with their intention, previous studies on Theory of Mind (Asteriadis et al. 2009) have shown it to be highly correlated. We envision a two-camera system that is able to capture views of both the interior and exterior of the vehicle, where we refer to the interior-facing camera as a *driver-monitoring camera* (DMC) that captures images of the human's head and face and the exterior-facing camera as a *road camera* that captures images of the surrounding environment. By correlating the information about the human's gaze captured by the DMC with the information about the exterior environment captured by the exterior-facing camera, we aim to infer which point of interest is associated with the human's intent, i.e., the *intended point of interest*. A representative figure of such a system is shown in Figure 1.

Inferring the driver's intended point of interest in this setting is challenging for many reasons. First, as in the case of shopping centers, many potential points of interest may be clustered together in a relatively small area, causing confusion regarding which one the driver is concerned with. Second, the vehicle's motion changes the location of the points of interest relative to the human within the vehicle, which causes ambiguity in the meaning of shifts in the driver's gaze. These challenges are made more difficult due to mul-

Figure 1: **A use case of intention inference:** In the setting of a human driver riding in a vehicle, a Driver Monitoring Camera captures the head-pose/eyes of the driver and a road camera captures the street view. Based on the captured images, an automated agent infers which point of interest in the street view that the driver is interested in going toward or obtaining more information about.

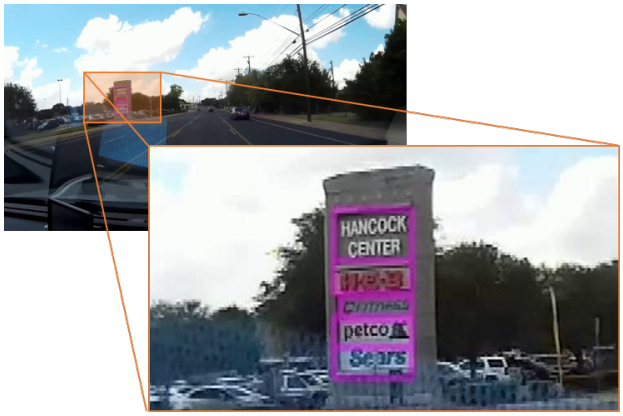


Figure 2: **A snapshot of clustered points of interest in a challenging driving scene:** The potential points of interest are shop signs highlighted in magenta. These interest points are hard to distinguish as they are clustered in a small region, causing confusion regarding which one the driver is concerned with. Our goal is to infer which one is the driver's intended point of interest.

tiple sources of noise in the gaze information coming from the DMC, e.g., eye blinks and vehicle shaking. Therefore, we seek here a systematic and robust way to address these challenges and determine the driver's intended point of interest.

As in previous work (Takemura et al. 2003), we make observations of gaze in the form of points in the environments. However, we do not necessarily assume that the raw gaze point aligns perfectly with the human's intended point of interest, as is done in a nearest-neighbor (NN) approach where the point of interest with nearest distance to the gaze point is considered to be the inferred point of interest. Such techniques are highly susceptible to the noise described above. Therefore, in our Dynamic Interest Point Detection (DIPD) algorithm we instead treat the observed gaze points as probabilistic inputs into a more-robust dynamic Markov Random Field (MRF) model that seeks to estimate the correct point of interest.

We evaluate our technique in a challenging driving scene as shown in Figures 2 and 3, where the points of interest in the road camera view are densely located and move non-linearly due to vehicle turning. The experimental results show that the success rate of our DIPD method may be improved by 28% compared to using NN method.

The contributions of our work are:

1. We formulate a dynamic MRF model with an energy function designed to be robust to noise for the problem of intention inference in dynamic environment.

2. We provide a solution to the above problem and quantify its benefit over a simpler, NN-based approach.

To the best of our knowledge, this is the first work that uses both gaze and a dynamic MRF model in inferring human intent.

Related Work

In this section we review prior work in two specific related areas. First, since the problem of inferring the driver's intended point of interest using gaze information is very similar to the problem of identifying fixations, we review the literature in which fixation detection has been previously studied. Second, since our overall goal is to infer the passenger's intent, we also review the field of intent recognition. While there has been much work done in both areas, our work has considered a unique situation and proposes a unique solution.

Identifying Fixations in Eye Movement Data

Human visual perception involves six types of eye movements: fixations, saccades, smooth pursuits, optokinetic reflex, vestibulo-ocular reflex, and vergence (Leigh and Zee 2015). Algorithms to identify the two most important types of eye movements, fixations and saccades, are usually based on velocity, acceleration, or area-based thresholding of the eye tracking data (Salvucci and Goldberg 2000). A common algorithm for fixation and saccade detection is the I-DT (dispersion-threshold identification) algorithm, which assumes that fixation points tend to cluster closely together as they have low velocity, and identifies fixations as groups of consecutive points within a particular dispersion. Recent research identified another type of event called glissades when analyzing the gaze tracking data (Nyström and Holmqvist 2010). Glissades are the undershoot/overshoot events between the transitions from saccades to fixations. An adaptive algorithm that detects glissades along with fixations and saccades can obtain more reasonable results in fixation and saccade durations compared to velocity-based or I-DT algorithms.

In prior literature, fixation detection is usually performed by determining whether a person's eyes are fixating at a

Figure 3: **In a challenging driving scene, the vehicle's motion changes the location and dimensions of the points of interest relative to the human within the vehicle:** As the vehicle moves along the street and takes a left turn, the size and position of points of interest change dynamically and non-linearly in the road camera view. This causes ambiguity in the meaning of shifts in the driver's gaze. Our DIPD method can infer the driver's point of interest under such dynamic and noisy environments.

static object from eye movement data. Detecting gaze fixation on a moving object is called smooth pursuit. Many works are concerned with modeling the smooth pursuit eye movement (SPEM) of the human visual system. An example of the smooth pursuit eye movement model was designed based on optimal control theory (Shibata et al. 2005). The model includes a recurrent neural network (RNN), which predicts the current or future target velocity, and a forward model of the target motion by online learning. Such biologically inspired SPEM models can match the human visual system quite well. Only a few recent works specifically address the problem of detecting smooth pursuit eye movement from eye tracking data. Examples of existing methods include using a three stage algorithm (Larsson et al. 2014), a threshold-based algorithm, or a probabilistic-based algorithm (Santini et al. 2016) to detect smooth pursuit eye movement for moving dot stimuli.

Importantly, the aforementioned work assumes the background is static, and the dynamic stimulus is a moving dot in the field of view. In this work, we instead propose a method for detecting object fixations in the presence of multiple moving objects and a moving background scene, where the object sizes are time-varying.

Intention Inference

An important aspect of a successful system which involves coexistence of both human and autonomous agents is the autonomous agent's ability to infer human agent's intent. A line of intention inference work relies on knowledge-based models which allow the autonomous agent to reason about human's actions and goals from current state information (Yordanova et al. 2017; Hiatt, Harrison, and Trafton 2011; Ramırez and Geffner 2011). Since our work focuses on utilizing bio-sensing data to infer a human's intent, we now review the literature related to these data-driven approaches.

A human's physical status (e.g., pose, action, and other physiological signals) and their interaction with the surrounding environment can sometimes reveal their intent. Therefore, intention inference can be partially achieved by analyzing one or more of these physical statuses. For example, some works have shown that modeling the relationship between human poses and objects in an image can be used to infer the person's next activity (Koppula and Saxena 2013; Delaitre, Sivic, and Laptev 2011). In a driving application, head motion has been used as an important cue for predicting a driver's intent to change lanes (Doshi and Trivedi 2008). Further, employing multi-modal data including GPS, speed, street maps, and driver's head movement can allow ADASs (advanced driver assistance system) to anticipate the driver's future maneuvers (Jain et al. 2015).

Gaze cues, which implicitly include head pose information, can help to infer human intent as it pertains to finer-grained points of interest (e.g., shop signs far away from a driver). A deep learning based method was proposed for doing so from a single image that combines gaze and saliency maps predicted using convolutional neural networks (CNNs) in order to form a predicted gaze direction (Recasens et al. 2015). The method was shown to be useful in both surveillance and human-robot teaming as a means by which to understand a person's intention from a third party perspective. In cases where the person's face and gaze targets were captured by different cameras, one needs to correlate the gaze tracking data from the face camera with the objects from the scene camera. Prior work on DAS has shown how to correlate a diver's gaze with road signs in the environment (Fletcher et al. 2005). The system calculates the disparity between the scene camera and gaze angles for the sign, and then uses this disparity to determine whether or not the driver sees the road sign. Another approach is to divide the scene into several regions and train a classifier on a dataset which contains the face images with annotated regions to predict the region of user attention. For example, nine gaze zones in the vehicle such as driver's front, rear view mirror, passenger's front, etc., were defined and a CNN classifier was trained to categorize the face images into the predefined fixed nine gaze zones so as to recognize the point of driver's attention (Choi, Hong, and Kim 2016). In other application areas such as hand-eye coordination tasks and player-adaptive digital games, machine learning-based methods (e.g., SVM, kNN, LSTM, ...) have been shown to be effective in predicting user intent from gaze observations (Razin and Feigh 2017; Min et al. 2017).

These methods assume that the gaze observations are noise-free. In contrast, our method treats the observed gaze points as probabilistic inputs and infers user intent among finer-grain objects in a dynamic environment.

System Overview

Figure 4 shows the system diagram for our DIPD algorithm to infer a human's intended gaze point. The system receives gaze data points from the eye tracker and object bounding boxes from an object-detection algorithm applied to the images from the road camera. The bounding boxes provide the position and dimension information for the possible gaze points in the scene. The observed gaze point is treated as a probabilistic input into a dynamic MRF model, which is constructed in temporal space to take into account gaze points in previous frames. An energy function associated with the dynamic MRF model is then minimized to infer the driver's intended point of interest.

Methodology

Our DIPD method performs inference using an MRF model. For each frame, we build a new MRF model as in Figure 5. This model takes into account the observed gaze location at the current time T and its location in a window of previous frames. The top layer nodes are denoted as $\{\mathbf{b}_t = (b_t^x, b_t^y) : t \in \mathbb{Z}, T-w+1 \leq t \leq T\}$ to represent the observed gaze pixel coordinates during this window. The window size w may be adapted to the camera frame rate and the speed of the moving objects. The bottom layer nodes are denoted as $\{c_{t,i} : t \in \mathbb{Z}, T-w+1 \leq t \leq T; i = na, 1, 2, ..., N\}$ to represent the points of interest in the scene, where N is the number of interest points in the scene, and where $i = na$ represents the case that the human is not attending to any of the points of interest. Each gaze point node $\mathbf{b}_t$ is connected to all the interest point nodes $c_{t,i}$ in every time frame. The array of $c_{t,i=na,1,..,N}$ under a gaze point node $\mathbf{b}_t$ is a one-hot vector, which consists of 0s in all elements with the exception of a single 1 used uniquely to identify the attended interest point. To infer $c_{t,i}$ from $\mathbf{b}_t$, nodes $\mathbf{b}_t$ and nodes $c_{t,i}$ are related by an energy potential that represents the likelihood of $\mathbf{b}_t$ given $c_{t,i}$. The nodes in the model are dynamically changed based on the number of available interest points in the dynamic environment.

We assume that the likelihood of the gaze point $\mathbf{b}_t$ given a point of interest $c_{t,i}$ is attended follows a Gaussian function centered at the interest point's bounding box center $\mathbf{u}_{t,i}$ with a covariance matrix $\mathbf{\Sigma}$ related to the bounding box dimensions (i.e., width and height). Therefore, the likelihood of $\mathbf{b}_t$ given $c_{t,i} = 1$ can be written as

$$P(\mathbf{b}_t|c_{t,i}=1) \propto \exp[-\frac{1}{2}(\mathbf{b}_t - \mathbf{u}_{t,i})^T \mathbf{\Sigma}^{-1}(\mathbf{b}_t - \mathbf{u}_{t,i})] \ . \tag{1}$$

Next, we formulate an energy function that can remove the undesirable effects caused from eye blinks and moving/shaking environment, and use this energy function to derive the most probable point of interest that is attended by the user (i.e., the user's point of interest). Assuming the inference results $\mathbf{c}_t$ will be highly correlated with the probability value $P(\mathbf{b}_t|c_{t,i})$, we form a "tracking" energy term as $-\sum_{i=1}^{N} c_{t,i} \cdot P(\mathbf{b}_t|c_{t,i}=1)$. This energy term will be lower when the likelihood of the gaze point b_t given a point of interest $c_{t,i}$ is attended is higher. Therefore, the location of the high (1) bit in the one-hot vector $\mathbf{c}_t$ will have the tendency to align the point of interest i which corresponds to the highest probability value $P(\mathbf{b}_t|c_{t,i}=1)$. In addition, we assume that the likelihood of a gaze point not attending any of the interest points is uniformly distributed in the space of all possible gaze point locations. We denote the probability value of this case as a constant k, and form an additional energy term $-k \cdot c_{t,i=na}$. Finally, we assume that people typically fixate their eye gaze at their point of interest for a while when they perceive it, and so the inferred point of interest should be fairly steady during this time period. Therefore, we form a "time-consistency" energy term that contains $\sum_{t'=T-w+1}^{T} |c_{t,i} - c_{t',i}|$ so that the energy is lower if the inference results are consistent over the window w. The complete energy function for the dynamic MRF model then takes the form

$$\begin{aligned} E(\mathbf{c}_t; \mathbf{b}_t, \mathbf{c}_{t'=T-w+1..T,i}) = &-\sum_{i=1}^{N} c_{t,i} \cdot P(\mathbf{b}_t|c_{t,i}=1) \\ &-k \cdot c_{t,i=na} + \alpha \sum_{i} \sum_{t'=T-w+1}^{T} \frac{1}{w} |c_{t,i} - c_{t',i}| \end{aligned} \tag{2}$$

where α is a positive constant. The first two terms essentially act as a high-pass filter that tracks moving location of the interest points, and the last term essentially acts as a low-pass filter that removes spikes and outliers due to eye blinks and moving/shaking effects.

The inference results $\mathbf{c}_t$ can be found by optimizing the energy function. That is, we would like to solve

$$\begin{aligned} \mathbf{c}_t^* = \underset{\mathbf{c}_t}{\operatorname{argmin}} \ & E(\mathbf{c}_t; \mathbf{b}_t, \mathbf{c}_{t'=T-w+1...T,i}) \\ \text{s.t. } & c_{t,i} \in \{0,1\}, \ \sum_i c_{t,i} = 1 \end{aligned} \tag{3}$$

We use Iterated Conditional Modes (ICM) (Kittler and Föglein 1984) to find the $\mathbf{c}_t^*$ that minimizes the total energy in the MRF model. The inference results $\mathbf{c}_t^*$ are typically obtained after a few iterations. The node with $c_{t,i} = 1$ corresponds to the inferred user's point of interest, denoted by y_t.

In practice, the number of available interest points may change dynamically among different frames. For example, the number of available interest points usually differs as the vehicle is moving along the street. Some interest points may be occluded by objects such as other vehicles in the scene so that they disappear in a few frames during the window w. In some cases, the number of available interest points varies because the object recognition system fails to identify all interest points in the scene. As such, our system will build up the dynamic MRF nodes for all interest points that appear in any frame within the window w and then compute the likelihood for all of them. If an interest point was missing in a frame, we may simply assign zero probability to its corresponding node in the dynamic MRF model. The energy function will correct such outliers when we compute the inference results $\mathbf{c}_t^*$.

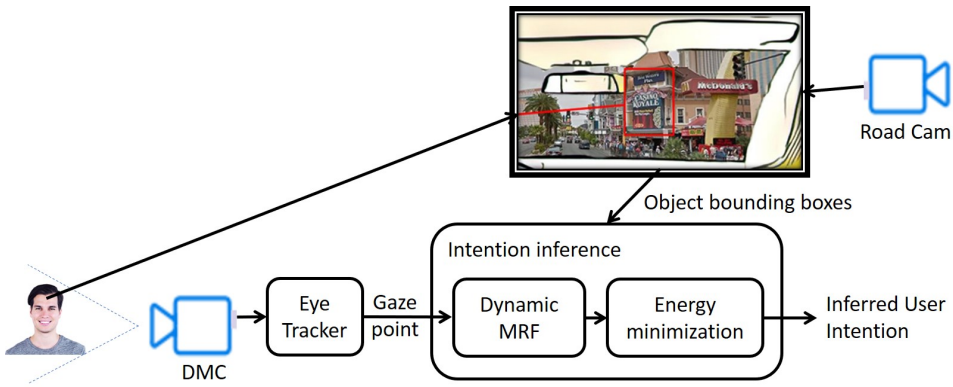


Figure 4: **A system diagram of the intention inference system that infers a human driver's point of interest in a driving scene:** The system obtains the gaze point of the human driver (from the DMC) and the object bounding boxes in the driving scene (from the Road Cam) to infer the user intent among finer-grain objects in a dynamic environment.

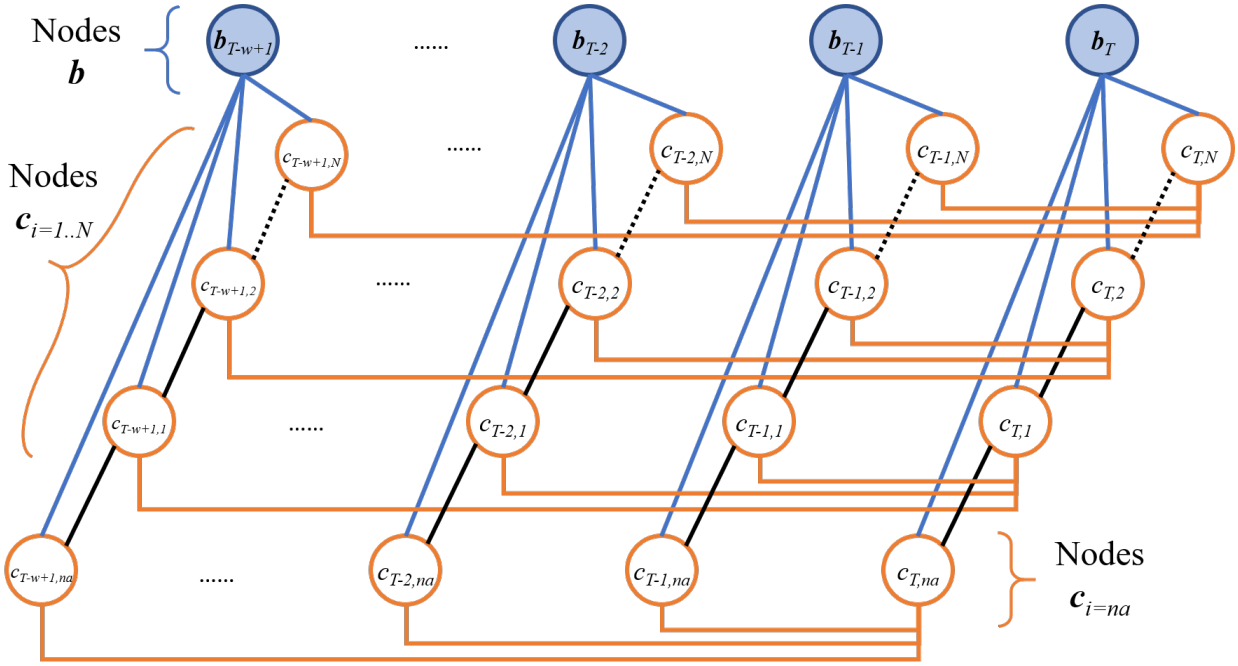


Figure 5: **The dynamic MRF model in our intention inference system for increasing the inference success rate:** Nodes $\mathbf{b}$ represent the observed gaze pixel coordinates, nodes $\mathbf{c}_{i=1..N}$ represent the points of interest in the scene, and nodes $\mathbf{c}_{i=na}$ represent the cases that the human is not attending to any of the points of interest. Nodes $\mathbf{b}$ and nodes $\mathbf{c}$ are related by an energy potential that represents the likelihood of $\mathbf{b}_t$ given $c_{t,i} = 1$. The nodes in the model are dynamically changed based on the number of available points of interest in the dynamic environment.

Experimental Results

We performed an experiment to verify the accuracy of our system for inferring a human's intended point of interest in a dynamic and challenging setting. In this section, we will explain our experiment setup and results.

Experiment Setup

We evaluated our DIPD algorithm in a challenging driving scene, where the possible points of interest are densely located and they move non-linearly in the exterior-facing camera view due to vehicle turning. We set up a 15-inch laptop showing a street view video of the environment recorded by a road camera. For gaze observations, we used a Tobii Eye Tracker 4C (Tobii 2017) mounted at the bottom of the screen. The eye tracker captured the eye images of the user and calculated the location of the user's gaze point on the screen. Based on the gaze point and the bounding boxes of each of the interest points, we infer the user's intended point of interest using both a baseline and our proposed method.

The street view video used in our experiment is about 13

seconds long (404 frames)[1]. We identify 3-5 interest points and their bounding boxes in each frame. An ID ranges from 0 to 4 is assigned to each of the interest points (i.e., Hancock, HEB, Fitness, Petco, and Sears). Due to occlusion by other vehicles, some objects do not exist for all frames in the video. We collected 70 experiment trials from 4 subjects, where in each trial we asked the user to find a specific point of interest in the street view video and fixate their gaze onto the point. The specified point of interest is the ground truth intended point of interest, denoted by z_t. For each time frame, the inferred point of interest is correct if $y_t = z_t$. The success rate of inferring user intention is defined as the ratio of the total number of correct inferences to the total number of frames. We drop the data of the first 60 frames when calculating the success rate since typically the participant is searching for the specific point of interest at the beginning of a trial.

Experiment Results

Table 1: **A comparison of the success rates of inferring user's point of interest using our intention inference method (DIPD) and an NN approach (i.e., baseline):** The results of our method for two different window size settings $w = 30$ and $w = 60$ (equivalent to 1 sec and 2 sec, respectively) are shown. The improvement percentages of DIPD compared to the NN baseline are shown in parentheses. Our intention inference method achieves a much better inference success rate.

ID	Baseline	DIPD @$w = 30$		DIPD @$w = 60$	
0	0.97	**1.00**	**(3.20%)**	**1.00**	**(3.20%)**
1	0.91	**1.00**	**(8.72%)**	0.97	(6.10%)
2	0.71	0.89	(17.44%)	**0.99**	**(27.91%)**
3	0.86	0.91	(5.23%)	**1.00**	**(14.24%)**
4	0.83	0.88	(4.94%)	**1.00**	**(16.86%)**

Table 1 shows the results of our experiment using our intention inference method (DIPD) versus an NN method (i.e., the baseline). Each row contains the results of an experiment in which the user's point of interest is specified in the first column (i.e., ground truth point of interest). Our DIPD method is computed based on the above methodology, whereas in the NN method we calculate the distance from the gaze point to the center of each interest point, and select the interest point with the shortest distance as the inferred point of interest. The improvement percentages of DIPD compared to the NN baseline are shown in parentheses.

The hyperparameter k of our method, which represents the probability value of a gaze point not attending any point of interest, is selected as $1/(N+1)$. The hyperparameter α represents the assumed relative importance of each term. We select $\alpha = 1$ for optimizing the energy function so that the tracking energy term and the time-consistency energy term are equally weighted. By using ICM, we vary the value of each node individually subject to the constraint in Equation (3) to find the values $\mathbf{c}_t^*$ that minimize the local potentials. Experimental results for our DIPD method using two different window size settings $w = 30$ and $w = 60$ (equivalent to 1 sec and 2 sec, respectively) in the dynamic MRF model are shown in Table 1.

[1]The video is available at `http://www.cs.utexas.edu/~larg/index.php/Gaze_and_Intent`

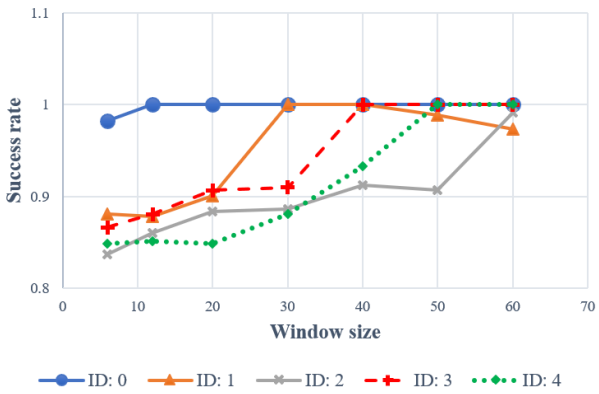


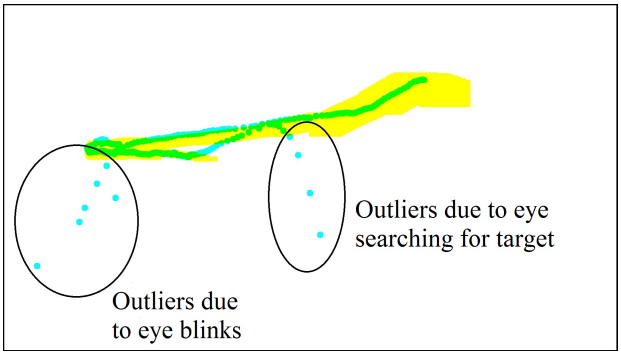
Figure 6: **Success rate versus window size setting:** In general, the success rate increases with the window size of the dynamic MRF filter.

We calculate the inference success rate for each method evaluated on each ground truth point of interest as described in Experiment Setup. The improvement of our method over the NN baseline is also reported. We can see that DIPD improves the inference success rate up to 28% over the baseline NN approach.

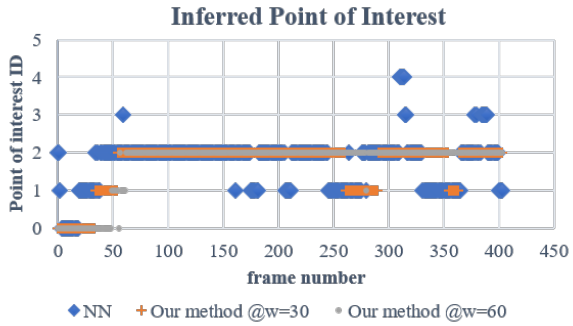
Since the typical mean fixation duration of human gaze is 260-330 ms for scene perception and 180-275 ms for visual search (Rayner and Castelhano 2007) and the fps (frame-per-second) of our system is 30, we sweep the window size w from 6 to 60 in our experiment. Figure 6 plots the success rate of our method for different window size settings w. The success rate is in general higher when setting the hyperparameter w to be 60 (i.e., 2 sec).

Discussion and Future Work

Our experiment shows that our dynamic MRF model and energy function can help to tolerate poor eye-tracking accuracy or stability and cancel the glitches due to blinks, moving background, and vehicle shaking. Figure 7a shows a shifting gaze point and intended object bounding box in a trial. The glitches and outliers are mainly caused by eye blinks, high-speed tracking misalignment, and the eyes searching for the specified point of interest at the beginning of the video, which can be observed in all trials. Our method can eliminate those glitches and outliers so as to achieve a better success rate (Figure 7b). In our experiment, ID #2 (Fitness) is the most difficult one since it is the smallest interest point located in the middle of the interest point clusters. Our experimental results show that the improvement for inference success rate is especially significant for such challenging objects (For DIPD with $w = 60$, ID #2: $Mean = 16.45, S.D. = 8.45$, Others: $Mean = 7.11, S.D. = 7.46$; $z = 4.85, p < 0.00001$).



(a)



(b)

Figure 7: **Selected experiment data:** (a) Traces of gaze points (in cyan) and intended object bounding boxes (in yellow) in a trial. The figure compacts the moving sequence of the gaze point and the object bounding box during the whole trial into one image. (b) Inferred point of interest for different methods. The ground truth point of interest ID is 2 for this case. The results of our DIPD method for two different window size settings $w = 30$ and $w = 60$ (in frames) are shown. Our method can eliminate most of the glitches and outliers for better inference success.

In general, setting a larger window size results in a better success rate. The results make sense because enlarging the window in the dynamic MRF model is like making the cutoff frequency of the low-pass filter lower to filter out more high frequency glitches. We also observe that setting higher window size (e.g., $w = 90$) can result in even better improvement in our experiment; however, it requires more computing time for inference. In practice, an upper bound on window size is desirable due to increased computation time.

The hyperparameter k is an assumed probability when the driver gazes none of the interest points. The inference result is not sensitive to the selection of k given a reasonable value is set. If k is set to 0, the inference result always "snap" to one of the interest points. As k increases, such behavior will be relaxed. This term is left for future work where it may adapt to fixation/saccade probability.

We have considered computing gaze fixations as a subcategory of the intention inference problem. Most previous research in gaze fixation detection has focused on analyzing still images, whereas our work considers this problem in the context of object and background motion. In our experiment, we demonstrated our technique on shop signs in a dynamic and noisy environment, though it can be applied to other inference applications as well, such as other vehicles on the road, third party objects in a human-robot interaction task, and the holograms in a mixed-reality world. Interesting directions for future work include deploying our system in a real vehicle and investigating ways to improve the proposed energy function.

Conclusion

In this paper, we presented a DIPD method for inferring a user's intended point of interest from eye-tracking data in a dynamic environment. The DIPD method utilizes a dynamic MRF model with an energy function designed to be robust to noise caused by eye blinks, vehicle shaking, and eyes and gaze tracker inaccuracy. We evaluated our technique experimentally and quantified its benefit over a simpler, NN-based approach. In general, our technique outperforms the NN approach. The improvement is especially significant for small, challenging objects in congested scenarios.

Acknowledgments

The author would like to thank Mike Huang from Mindtronic AI for useful discussion. This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (CNS-1305287, IIS-1637736, IIS-1651089, IIS-1724157), The Texas Department of Transportation, Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

Asteriadis, S.; Tzouveli, P.; Karpouzis, K.; and Kollias, S. 2009. Estimation of behavioral user state based on eye gaze and head pose – application in an e-learning environment. *Multimedia Tools and Applications* 41(3):469–493.

Bengler, K.; Dietmayer, K.; Farber, B.; Maurer, M.; Stiller, C.; and Winner, H. 2014. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine* 6(4):6–22.

Calder, A. J.; Lawrence, A. D.; Keane, J.; Scott, S. K.; Owen, A. M.; Christoffels, I.; and Young, A. W. 2002. Reading the mind from eye gaze. *Neuropsychologia* 40(8):1129–1138.

Choi, I.-H.; Hong, S. K.; and Kim, Y.-G. 2016. Real-time categorization of driver's gaze zone using the deep learning techniques. In *Big Data and Smart Computing (BigComp), 2016 International Conference on*, 143–148. IEEE.

Delaitre, V.; Sivic, J.; and Laptev, I. 2011. Learning person-object interactions for action recognition in still images. In *Advances in neural information processing systems*, 1503–1511.

Doshi, A., and Trivedi, M. 2008. A comparative exploration of eye gaze and head motion cues for lane change intent prediction. In *Intelligent Vehicles Symposium, 2008 IEEE*, 49–54. IEEE.

Doshi, A., and Trivedi, M. M. 2011. Tactical driver behavior prediction and intent inference: A review. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, 1892–1897. IEEE.

Fletcher, L.; Loy, G.; Barnes, N.; and Zelinsky, A. 2005. Correlating driver gaze with the road scene for driver assistance systems. *Robotics and Autonomous Systems* 52(1):71–84.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating human variability in human-robot teams through theory of mind. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, 2066.

Jain, A.; Koppula, H. S.; Raghavan, B.; Soh, S.; and Saxena, A. 2015. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision*, 3182–3190.

Kittler, J., and Föglein, J. 1984. Contextual classification of multispectral pixel data. *Image and Vision Computing* 2(1):13–29.

Koppula, H., and Saxena, A. 2013. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *International Conference on Machine Learning*, 792–800.

Larsson, L.; Nystro, M.; Stridh, M.; et al. 2014. Discrimination of fixations and smooth pursuit movements in high-speed eyetracking data. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 3797–3800. IEEE.

Leigh, R. J., and Zee, D. S. 2015. *The neurology of eye movements*, volume 90. Oxford University Press, USA.

Mainprice, J.; Hayne, R.; and Berenson, D. 2015. Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, 885–892. IEEE.

Matsumotot, Y.; Ino, T.; and Ogsawara, T. 2001. Development of intelligent wheelchair system with face and gaze based interface. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, 262–267. IEEE.

Min, W.; Mott, B.; Rowe, J.; Taylor, R.; Wiebe, E.; Boyer, K. E.; and Lester, J. 2017. Multimodal goal recognition in open-world digital games.

Nyström, M., and Holmqvist, K. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods* 42(1):188–204.

Ramırez, M., and Geffner, H. 2011. Goal recognition over pomdps: Inferring the intention of a pomdp agent. In *IJCAI*, 2009–2014. IJCAI/AAAI.

Ravichandar, H.; Kumar, A.; and Dani, A. 2016. Bayesian human intention inference through multiple model filtering with gaze-based priors. In *Information Fusion (FUSION), 2016 19th International Conference on*, 2296–2302. IEEE.

Rayner, K., and Castelhano, M. 2007. Eye movements. *Scholarpedia* 2(10):3649. revision #126973.

Razin, Y., and Feigh, K. M. 2017. Learning to predict intent from gaze during robotic hand-eye coordination. In *AAAI*, 4596–4602.

Recasens, A.; Khosla, A.; Vondrick, C.; and Torralba, A. 2015. Where are they looking? In *Advances in Neural Information Processing Systems*, 199–207.

Salvucci, D. D., and Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, 71–78. ACM.

Santini, T.; Fuhl, W.; Kübler, T.; and Kasneci, E. 2016. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 163–170. ACM.

Shibata, T.; Tabata, H.; Schaal, S.; and Kawato, M. 2005. A model of smooth pursuit in primates based on learning the target dynamics. *Neural Networks* 18(3):213–224.

Takemura, K.; Ido, J.; Matsumoto, Y.; and Ogasawara, T. 2003. Drive monitoring system based on non-contact measurement system of driver's focus of visual attention. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, 581–586. IEEE.

Tall, M.; Alapetite, A.; San Agustin, J.; Skovsgaard, H. H.; Hansen, J. P.; Hansen, D. W.; and Møllenbach, E. 2009. Gaze-controlled driving. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, 4387–4392. ACM.

Tobii. 2017. *Tobii Eye tracker 4C*.

Yordanova, K.; Whitehouse, S.; Paiement, A.; Mirmehdi, M.; Kirste, T.; and Craddock, I. 2017. What's cooking and why? behaviour recognition during unscripted cooking tasks for health monitoring. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, 18–21. IEEE.

Yu, C.; Ballard, D. H.; and Zhu, S. 2002. Attentional object spotting by integrating multimodal input. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, 287. IEEE Computer Society.