In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (ICMI 2018), Boulder, Colorado, USA October 2018

Inferring User Intention using Gaze in Vehicles

Yu-Sian Jiang
The University of Texas at Austin
Austin, TX
sharonjiang@utexas.edu

Garrett Warnell
US Army Research Laboratory
Adelphi, MD
garrett.a.warnell.civ@mail.mil

Peter Stone
The University of Texas at Austin
Austin, TX
pstone@cs.utexas.edu

ABSTRACT

Motivated by the desire to give vehicles better information about their drivers, we explore human intent inference in the setting of a human driver riding in a moving vehicle. Specifically, we consider scenarios in which the driver intends to go to or learn about a specific point of interest along the vehicle's route, and an autonomous system is tasked with inferring this point of interest using gaze cues. Because the scene under observation is highly dynamic — both the background and objects in the scene move independently relative to the driver — such scenarios are significantly different from the static scenes considered by most literature in the eye tracking community. In this paper, we provide a formulation for this new problem of determining a point of interest in a dynamic scenario. We design an experimental framework to systematically evaluate initial solutions to this novel problem, and we propose our own solution called *dynamic interest point detection* (DIPD). We experimentally demonstrate the success of DIPD when compared to baseline nearest-neighbor or filtering approaches.

KEYWORDS

dynamic interest point detection; intent recognition; Markov random field; eye tracking; autonomous driving agent; vehicle intelligence

ACM Reference Format:
Yu-Sian Jiang, Garrett Warnell, and Peter Stone. 2018. Inferring User Intention using Gaze in Vehicles. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3243018

Figure 1: A use case of intention inference. In the setting of a human driver riding in a vehicle, a DMC captures the driver gaze, and a road camera captures the street view. Based on the captured images, an autonomous agent infers which point of interest in the street view that the driver is interested in going toward or obtaining more information about.

1 INTRODUCTION

In recent years, automakers have started adding bio-sensing technologies to vehicle interiors as a means by which to give vehicles better information about the occupants inside. Information from these sensors can be used to infer, e.g., driver distraction or drowsiness [3], driver identity [4], occupant type [21], occupant proximity [1], or occupant health and wellness [29]. This information not only has the potential to enable more useful vehicle functionality and better safety features, but also novel forms of human-machine interaction. These interactions will be inherently multimodal, with opportunities for visual and audio two-way communication, and might involve anything from touch sensors to interaction in free-form natural language.

One particularly interesting type of occupant information that can be obtained using bio-sensors is human 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 [5]. Therefore, we expect that providing automated agents with a similar ability will provide a better user experience in human-machine interaction. Indeed, several examples in the literature (e.g., [7, 10, 18, 19, 24, 32]) have demonstrated that an autonomous agent utilizing human gaze cues can better interpret the human's intent and thus make for a better partner. By understanding a human's attentiveness and intention, the vehicle may, for example, take additional safety measures if it finds the human's actions to be dangerous, or provide intelligent assistance by reacting to inferred human intent. This inferred human attentiveness and intention can then be combined with touch and/or voice commands and feedback with a larger multimodal system. For example, the driver may look at a store's sign and request information about the store's hours of operation.

Inspired by that literature, we seek here to explore the use of driver gaze information obtained using imaging sensors in intelligent vehicles. We are particularly concerned with the setting in which a human driver rides in a moving vehicle, and we assume that the driver'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 her intention, previous studies on theory of mind [2] have shown it to be highly correlated. We envision a two-camera system that is able to capture views of both the interior

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243018

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* (RC) 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 environment captured by the road camera, the vehicle's task is that of inferring which point of interest is associated with the human's intent, i.e., the *intended point of interest*. A representative illustration of such a system is shown in Figure 1. The DMC can be seen in the middle of the steering wheel, and the road camera is located behind the rear view mirror.

Inferring the driver's intended point of interest in this setting is a new problem, and it is challenging for many reasons. First, 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. For example, in a challenging driving scene as shown in Figure 2, the vehicle's motion changes the location and dimensions of the points of interest relative to the human within the vehicle. 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. 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. Such a highly dynamic environment leads to ambiguity in the meaning of shifts in the driver's gaze. These challenges are made more difficult due to multiple sources of noise in the gaze information coming from the DMC, e.g., eye blinks, misalignment, and vehicle shaking. Because most classical techniques in eye tracking were developed mainly for static scenes, the problem outlined above requires new solutions.

The contributions of this paper are twofold. First, we develop a new dynamic gaze tracking problem for detecting the user's point of interest in highly dynamic environments. We address this new problem and design an experiment through which it can be systematically investigated. We assume that the interior of the vehicle is equipped with a DMC that is able to capture the human driver's face, and that the vehicle is also equipped with a road camera to capture the scene outside of the vehicle. Under these conditions, the question of interest is: *given both the DMC and road camera videos, to which interest point in the road camera video is the person attending?* Secondly, we propose our own solution called *dynamic interest point detection* (DIPD), which seeks to address the above challenges in order to determine the driver's intended point of interest. As in other work [31], our own DIPD method takes as input the observations of gaze in the form of points in the environment. However, we do not necessarily assume that the raw gaze point aligns perfectly with the human's intended point of interest, as would be 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. Instead, DIPD treats the observed gaze points as noisy inputs into a more robust dynamic Markov random field (MRF) model that seeks to estimate the correct point of interest. We evaluate DIPD to the new problem mentioned above and quantify its benefit over baseline nearest-neighbor and filtering approaches.

2 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 relevant to the problem of identifying fixations, we review the literature in which eye movement data filtering and fixation detection has been previously studied. Second, since our overall goal is to infer the driver's intent for multimodal interaction, we also review the field of multimodal interaction using intent recognition. While there has been much work done in both areas, our work considers a unique situation and proposes a unique solution.

2.1 Filtering and Identifying Fixations in Eye Movement Data

In order to extract and analyze gaze information from eye movement data, previous work has employed filtering techniques that smooth and denoise the eye tracking data. Finite-impulse response (FIR) filtering is a technique that computes the weighted average of several latest gaze points, where different weighting functions may be applied. One study has shown that FIR filters with triangular or Gaussian kernel weighting functions outperform other real-time filters for HCI purposes [28]. In practice, a simple low-pass filter (LPF) that computes the weighted sum of the current raw gaze point and the previous filtered gaze point is commonly used for fast, real-time gaze point filtering [20]. Unfortunately, these conventional filtering techniques fail to remove the low-frequency noise caused by eye blinks and vehicle shaking. In dynamic environments where the eyes are tracking a moving interest point, conventional filter methods may even degrade inference success rate due to the delay introduced by the filters.

Human visual perception involves six types of eye movements: fixations, saccades, smooth pursuits, optokinetic reflex, vestibulo-ocular reflex, and vergence [17]. 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 [26]. 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. In prior literature, fixation detection is usually performed under the setting where a static object is presented to a human subject, and a fixation detection algorithm classifies the eye movement data into fixation and saccade types. Only a few recent works specifically address the problem of detecting smooth pursuit eye movement under the setting where a constant velocity target or a periodically moving target is presented. Examples of existing methods include using a three stage algorithm [16], a threshold-based algorithm, or a probabilistic-based algorithm [27] to distinguish smooth pursuit eye movement from fixation or saccade eye movement data. Importantly, the aforementioned works cannot tell which object is being fixated or pursued if there are multiple, and they assume that the background is static, the object is of constant size, there is only

Figure 2: 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. The points of interest in this figure are marked by magenta bounding boxes.

one moving object, or some combination of these. In this work, we instead deal with the problem of eye-tracking in the presence of multiple moving objects and a moving background scene, where the object sizes are time-varying, and further infer human's attention to one of the moving objects.

2.2 Intention Inference

An important aspect of successful human-machine interaction design is the autonomous agents' ability to infer the human agent's intent [6, 11, 30]. One 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 [12, 22, 34]. 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 [8, 15]. In a driving application, head motion has been used as an important cue for predicting a driver's intent to change lanes [9]. 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 [13].

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 [25]. 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 [10]. 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 [7]. In other application areas such as hand-eye coordination tasks and player-adaptive digital games, machine learning-based methods (e.g., SVM, kNN, LSTM, etc.) have been shown to be effective in predicting user intent from gaze observations [19, 24].

The above methods deal with coarse-grained regions of interest. In this paper, however, we are instead interested in inferring a human's point of interest at a fine-grained scale such as that of an object or a sign board in a driving scene. To the best of our knowledge, this type of eye tracking in the scenarios we have described has not been substantially explored in previous literature.

3 PROBLEM FORMULATION AND SOLUTION

In this section, we describe the formulation of the question we outlined in Section 1 (i.e., *given both the DMC and road camera videos, to which interest point in the road camera video is the person attending?*) and the method we developed that attempts to provide an answer to this question.

3.1 Dynamic Gaze Tracking

One factor that complicates the problem is the fact that the DMC and the road camera each capture video in a separate coordinate system. While geometrically rectifying this difference in coordinate systems is, in general, an important and interesting problem, our interest here is in a more abstract problem that exists even after the rectification has been done. Therefore, we instead formulate the problem on a projection plane **A**, where plane **A** is the vertical plane that is co-located with the DMC. As illustrated in Figure 3, we calculate the gaze point on plane **A**, and we assume that the scene captured by the road camera has the same field of view as what the human can see. Both the gaze point and the bounding box location of interest points in the 3D world can then be projected onto plane **A** for correlation, i.e., gaze points can be represented by two-dimensional coordinates $\mathbf{b} = (b^x, b^y)$ and the i^{th} interest point

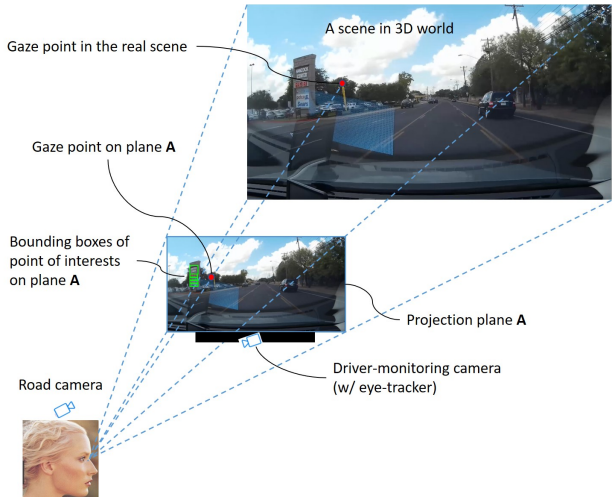


Figure 3: Illustration of the dynamic eye-tracking problem on a virtual plane.

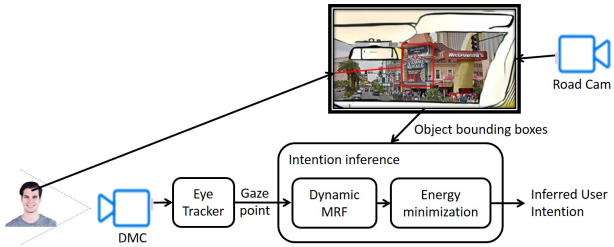


Figure 4: A system diagram of the DIPD method for inferring a human's point of interest in a driving scene. The intention inference engine obtains the gaze point of the human driver (from the DMC) and the object bounding boxes in the driving scene (from the road camera) to infer the user intent among finer-grain objects by using a dynamic MRF model and energy minimization.

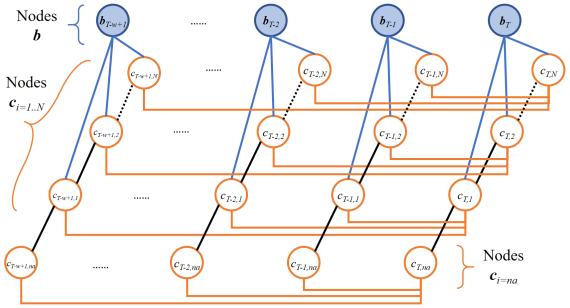


Figure 5: Illustration of the dynamic MRF model in our DIPD method for addressing dynamic eye-tracking challenges.

can be represented by a two-dimensional bounding box centered at $\mathbf{u}_i = (u_i^x, u_i^y)$.

We now make the problem even more explicit by describing a representative experiment we have designed for the purposes of evaluating the accuracy of using different eye-tracking methods that aim to solve the challenge set out above. The testing environment is a driving scene, where the possible points of interest are densely located and they move non-linearly in the road camera video due to the vehicle turning. In order to simulate a dynamic environment for a human participant, we set up a screen that displayed a street view video pre-recorded using a real RC in a real moving vehicle. The screen is used to define the two-dimensional coordinate system on the projection plane **A** in Figure 3. An imaging sensor taking the role of DMC should be co-located with the vertical plane of the screen. We use a standalone eye tracker mounted on the screen to compute the human's gaze points. Then, we calculate the user's gaze coordinates $\mathbf{b}$ in terms of two-dimensional coordinates on the screen. A computer connecting to the screen is configured to run the intention inference algorithms we would like to evaluate.

The experiment requires human subjects to perform individual trials in the above setting. For each trial, the subject is asked to find a specific point of interest in the street view video and fixate their gaze onto that point (i.e. the user's intent is cued by the experimenter). This 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 y_t is correct if $y_t = z_t$. The success rate of inferring user intention can then be defined as the ratio of the total number of correct inferences to the total number of frames when the users fixate their gaze onto the specified point of interest.

3.2 DIPD

We now propose our own novel solution to the dynamic gaze-tracking problem, and we call this solution *dynamic interest point detection* (DIPD). Figure 4 shows the system diagram for DIPD. 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 spans both space and time in order 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.

The 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 not only the gaze points at current time T but also the historical gaze points upto $w-1$ previous frames (i.e. we consider gaze points in a window w). The top layer nodes are denoted as $\{\mathbf{b}_t = (b_t^x, b_t^y) : t \in \mathbb{Z}, T-w+1 \le t \le T\}$ 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 $\{c_{t,i} : t \in \mathbb{Z}, T-w+1 \le t \le T; i = na, 1, 2, ..., N\}$ 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 Σ 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

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

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

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

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

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

We use iterated conditional modes (ICM) [14] 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 with respect to time. For example, the number of available interest points usually changes as the vehicle is moving along a street. Some interest points may be occluded by other scene objects and so they may disappear for a few frames during the window w. This may also be the case if the object recognition system fails to identify all the interest points in the scene. To handle these scenarios, DIPD constructs the dynamic MRF nodes for all interest points that appear in any frame within the window w and computes the likelihood for all of them. If an interest point was missing in a frame, DIPD simply assigns a zero probability to its corresponding node in the dynamic MRF model. The energy function corrects such outliers when we compute the inference results $\mathbf{c}_t^*$.

4 EXPERIMENTS

In this section, we describe our experiment design and the reasons behind the design. Then, we describe the mathematics behind the baseline approaches we compare to. Finally, we provide the experiment results of our human study, which shows that DIPD has a 28% better inference success rate than the baseline approaches.

4.1 Experimental Design

We performed a laboratory study to validate our proposed techniques since it allowed us to collect experiment data from multiple participants experiencing the same road scenes and same point of interests and compare different methods statistically. Further, it allowed us to more easily overcome low-level eye tracking and coordinate transformation problems and instead focus on the higher-level goals of DIPD.

The street view video used in our experiment is about 13 seconds long (404 frames).[1] Note that, although the study was done in a laboratory setting, it used real data in the sense that it was gathered using an in-car camcorder to pre-record a challenging scene with a cluster of high density signs and emulates the road scenes by a computer screen. We identified 3-5 interest points and their bounding boxes in each frame. A unique ID number between 0 and 4 was assigned to each of the interest points (i.e., Hancock, HEB, Fitness, Petco, and Sears). Due to occlusion by other vehicles, some interest points did not appear in all frames of the video. During the experiment, the street view video was displayed in full-screen view so that the screen pixel coordinates of the bounding boxes directly represent the two-dimensional coordinates of the bounding boxes on plane **A**.

We ran the experiment described in Section 3.1 for 70 trials collected over 4 human subjects. In our experiment, we set up a 15-inch laptop showing a street view video of the environment recorded by a road camera. A Tobii Eye Tracker 4C [33] mounted at the bottom of the screen was used to simulate the setup of a DMC with eye-tracking function for obtaining gaze points on the screen. The eye tracker computed the user's gaze coordinates **b** in terms

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

of two-dimensional coordinates on the screen. We used the gaze coordinates information to process the data discussed in Sections 4.2 and 3.2.

The objective of this experiment is to test whether the systems give correct inference results when the user is intended to the specified point of interest. The users are asked to search for the specified point of interest after video of the road camera starts. As a result, the system often reports a null of POI (i.e. N/A) in the first 2 seconds due to the fact that typically the participant is searching for the specific point of interest at the beginning of a trial. Therefore, the data of the first 60 frames are dropped when calculating the success rate.

4.2 Baseline Approaches

Since inferring a human's point of interest in dynamic environments is a new problem, we first propose two baseline approaches which are adapted from the classical eye-tracking literature. These baseline approaches will be described in detail below.

4.2.1 Nearest-neighbor-based Approach. Given gaze point coordinates and the bounding box locations of the interest points on plane $\mathbf{A}$, one intuitive way to infer which interest point is attended by the human is to use a nearest-neighbor (NN) approach. In this approach, the point of interest with nearest distance to the gaze point is considered to be the inferred point of interest.

More specifically, for every frame, we obtain a set of interest points indexed by $i \in \{1, \ldots, N\}$ that have been extracted by, e.g., and object detection algorithm. Each interest point is represented by a two-dimensional bounding box centered at $\mathbf{u}_i = (u_i^x, u_i^y)$. We also receive the coordinates of the current gaze point, denoted by $\mathbf{b} = (b^x, b^y)$. If no valid gaze point can be obtained from the eye tracker, we simply use the gaze point coordinates from the previous time step. The NN approach computes the inferred point of interest, i^*, as

$$i^* = \underset{i}{\operatorname{argmin}}\ d(\mathbf{b}, \mathbf{u}_i) \tag{4}$$

where $d(\mathbf{b}, \mathbf{u}_i) = ||\mathbf{b} - \mathbf{u}_i||$ is the Euclidean distance between gaze point and the center of the i^{th} interest point bounding box.

4.2.2 Filter-based Approach. Because the source of error in our dynamic gaze-tracking problem can be regarded as a form of noise, the second baseline approach we develop uses filtering in an attempt to reject this noise. This is similar to approaches used in the static eye-movement analysis problem. Here, we propose baseline approaches that are based on processing the gaze observations using two filtering techniques: low-pass filtering (LPF) and finite-impulse response (FIR) filtering.

In the LPF method, we fist apply an LPF to the observed gaze points $\mathbf{b}_t$ in order to compute a filtered gaze point, $\tilde{\mathbf{b}}_t^{LPF}$. We then calculate the distance from $\tilde{\mathbf{b}}_t^{LPF}$ to the center of each interest point, and then select the interest point with the shortest distance as the inferred point of interest. The LPF itself is implemented according to the following equation [20]:

$$\tilde{\mathbf{b}}_t^{LPF} = \lambda \tilde{\mathbf{b}}_{t-1}^{LPF} + (1 - \lambda)\mathbf{b}_t \tag{5}$$

where λ is a coefficient chosen to be between 0 and 1. Higher λ results in more rejection of high-frequency noise, but also slower system reaction to abrupt changes in the gaze point location.

In the FIR method, we first apply a FIR filter with a Gaussian kernel (weighting) function to the observed gaze points in order to calculate filtered gaze points $\tilde{\mathbf{b}}_t^{FIR}$. We then compute the distance from $\tilde{\mathbf{b}}_t^{FIR}$ to the center of each interest point, and select the interest point with the shortest distance as the inferred point of interest. To implement the FIR filter, we use a buffer to store the latest gaze points, and compute a weighted sum of each of these gaze points according to a Gaussian kernel function [28]. The Gaussian kernel function is expressed as follows:

$$W_i = e^{-\frac{i^2}{2\sigma^2}} \tag{6}$$

where the σ is a chosen parameter. The output value from the FIR filter is computed by the following equation:

$$\tilde{\mathbf{b}}_t^{FIR} = \frac{\sum_{i=0}^{N} W_i \times \mathbf{b}_{t-i}}{\sum_{i=0}^{M} W_i} \tag{7}$$

Here, higher σ means more previous frames are taken into account for filtering the current gaze point.

4.3 Experimental Results

Table 1: The success rates of inferring a driver's point of interest using the methods discussed in Sections 4.2 and 3.2.

ID	NN	LPF	FIR	DIPD @$w = 30$	DIPD @$w = 60$
0	0.97	0.94	0.85	**1.00**	**1.00**
1	0.91	0.89	0.83	**1.00**	0.97
2	0.71	0.69	0.65	0.89	**0.99**
3	0.86	0.85	0.84	0.91	**1.00**
4	0.83	0.83	0.84	0.88	**1.00**

Table 1 compares the success rate of our intention inference method (DIPD) to baseline approaches described in Section 4.2. 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).

For the LPF method, a value of 0.667 was used as the hyperparameter λ . For FIR filtering, the Guassian kernel width used was $\sigma = 10$. When calculating the Gaussian kernel function, we only take the latest M gaze points which correspond to $W_i \geq 0.05$ and ignore other gaze points which correspond to $W_i < 0.05$. For DIPD, in Eq. 2, the hyperparameter k, which represents the probability value of a gaze point not attending any point of interest, was set at $1/(N + 1)$. The hyperparameter α represents the assumed relative importance of each term, and we used a value of $\alpha = 1$ 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. Since the typical mean fixation duration of human gaze is 260-330 ms for scene perception and 180-275 ms for visual search [23] and the fps (frame-per-second) of our system is 30, we sweep the window size w from 6 to 60 in our experiment. 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. Figure 6 plots the success rate of our method for different

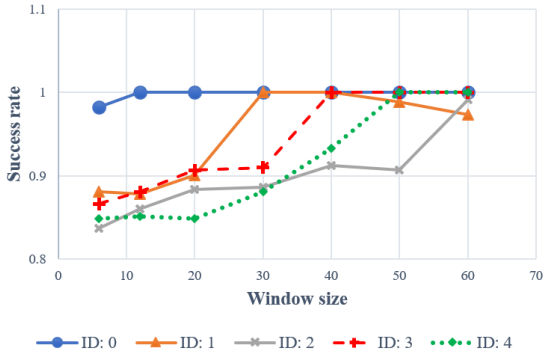


Figure 6: Success rate versus window size setting for DIPD.

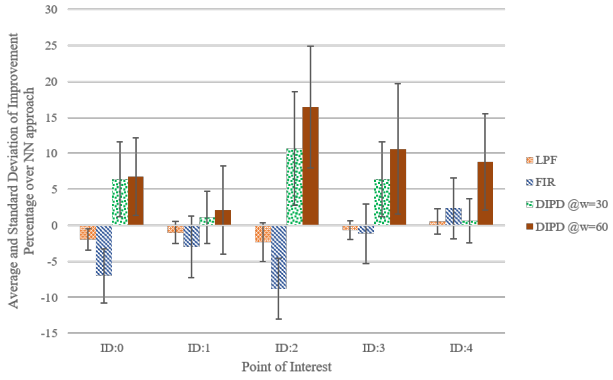
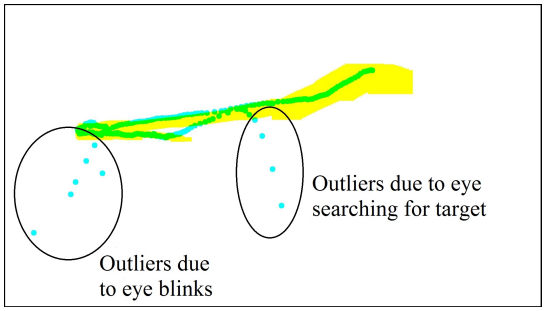


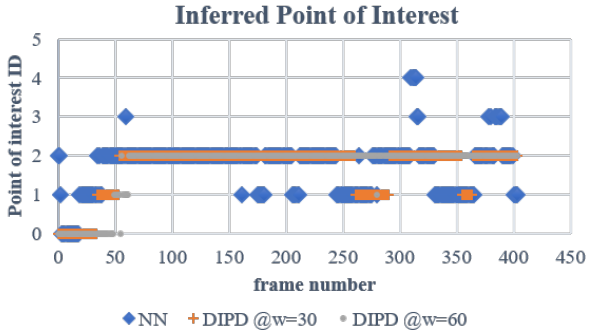
Figure 7: Average and standard deviation of the improvement percentage compared to the NN baseline for different methods. The filter-based approaches (i.e., LPF and FIR) generally perform much worse than PIPD.

window size settings w. The success rate is in general higher when setting the hyperparameter w to be 60 (i.e., 2 sec).

We calculate the inference success rate for each method evaluated on each ground truth point of interest as described in Section 3.1. We can see in Table 1 that DIPD has a 28% better inference success rate than the baseline NN approach, and even more for the filter-based approaches. Figure 7 shows the statistical results of filter-based approaches (i.e., LPF or FIR) and DIPD from 70 trials of user experiment. The effectiveness of each method is quantified by the improvement percentage of success rate compared to the NN baseline. We notice that conventional filter-based approaches perform much worse than DIPD, and most of them are even worse than pure NN. We posit that this is because these filters are linear filters, which cannot completely remove the outliers caused by eye blinks. Further, if larger filtering parameters (i.e., λ for the LPF and σ for the FIR filter) were selected to smooth the eye tracking signal further, it may introduce a longer delay in the system. This longer delay might degrade the system success rate since the system needs to be agile enough to follow the changing gaze location when the eyes are tracking a moving interest point.



(a)



(b)

Figure 8: 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 DIPD method can eliminate most of the glitches and outliers for better inference success (i.e., better alignment with the ground truth point of interest ID 2 in this case).

Our experiments show that the dynamic MRF model and energy function can help to tolerate poor eye-tracking accuracy or stability and can remove noise that arises due to blinking, a moving background, and vehicle shaking. To illustrate this, Figure 8a shows a shifting gaze point and intended object bounding box in a trial. The noise 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. DIPD successfully eliminates the noise and outliers and is able to achieve a better success rate. Figure 8b shows the point of interest inference results from the NN approach and DIPD approach in a time series. 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. The figure shows that inference results from the NN approach (i.e., blue dots) have more glitches and outliers than our DIPD method. 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, and our experimental results show that the DIPD method has significantly higher inference success rate for such challenging objects (for DIPD with $w = 60$, ID #2: $Mean = 16.45, S.D. = 8.45$, Others: $Mean = 7.11, S.D. = 7.46$; $z = 4.85, p < 0.00001$).

5 DISCUSSION AND FUTURE WORK

In this section, we highlight the unique problems DIPD resolves and provide a better context for our baseline methods. We also discuss the high-level difference between DIPD and NN, LPF/FIR approaches, how DIPD hyperparameters such as w, k, and α would affect the experiment results, the limitations of DIPD, and how they might be overcome (i.e., lessons learned).

The output of DIPD is an inferred user point of interest, which is fundamentally different (higher-level) than simply smoothed gaze points. In particular, it additionally requires rejecting dynamic environmental disturbances (e.g., vehicle shaking) and must deal with the size and variance of detected points of interest, which are challenges that are not addressed by simple gaze filtering techniques. After detecting and modeling moving interest points in a dynamic scene, we must then correlate the gaze points with a POI while rejecting noise from many sources. DIPD aims to infer user intention, which goes beyond smoothing gaze points. This is what motivates us to propose using intuitive extensions of gaze-filtering methods for a baseline, i.e., the NN method and higher-level filtering techniques. DIPD is effectively a special kind of filtering operation, whereas the NN approach does not have filtering over time. More specifically, DIPD uses graphical modeling in order to be robust to noise in high-speed tracking environments. This is in contrast to the LPF/FIR baseline approaches that filter over time, but do not leverage any higher-level information in order to accomplish the high-speed tracking needed in dynamic environments.

The DIPD algorithm has several hyperparameters that can be adjusted to make the algorithm workable in different dynamic environments. With respect to the DIPD window size in particular, it seems that, in general, setting a larger window size results in a larger success rate. Setting 30-60 frames as the window size in a 30fps system represents 1-2 seconds, which is based on the assumption that a driver being interested in an object would spend at least 1-2 seconds of attention on it. The window size is tuned empirically but is an adjustable hyperparameter. While using even larger window sizes (e.g., $w = 90$) may result in an even better success rate, it also requires more computing time for inference. The computing time grows about linearly with respect to the window size. Therefore, in practice, an upper bound on window size is desirable due to increased computation time. Regarding the DIPD hyperparameter k, i.e., the assumed probability that the driver is not attending to any of the interest points, we assumed this value to be 1/6 in our experiments. However, if k is set to 0, the inference result always "snaps" to one of the interest points (i.e. the inference result is always among one of the interest points even if the eye movement is in a saccade phase). If we set k to a higher value, this behavior is relaxed. The selection of this hyperparameter may be improved and is left for future work, where it may adapt to fixation/saccade probability of eye movement patterns.

In our experiment, we observed that the eye-tracker we used requires per-session calibration and that there are accuracy issues for freely moved/rotated heads. These flaws would limit the applicability of DIPD in vehicles unless more-robust low-level tracking is used. Moreover, an in-vehicle DIPD system would need to account for things like perspective transformation to translate the camera position of the road scene to the human's head position since the road camera is often set up near the rear-view mirror.

In our experiment, the DIPD method is applied to inferring the attended shop sign along the road, 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 the DIPD method in a real vehicle and investigating ways to improve the proposed energy function.

6 CONCLUSION

In this paper, we have proposed using multimodal interactions based on gaze with voice/touch commands to enable safer and more efficient human machine interaction for intelligent vehicles. We identified gaze analysis as an important component of such multimodal interactions, and introduced the problem of inferring a human's intended point of interest in a dynamic scene for scenarios in which human drivers are riding in moving vehicles. This is a new problem with many challenges beyond those considered in the classical eye-tracking literature, and requires solutions that work in highly dynamic environments and are robust to different sources of noise. We defined this new problem, designed an experiment to evaluate different inference methods, and proposed a new DIPD method for addressing these challenges. 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 DIPD experimentally and quantified its benefit over a NN method and other filtering based methods. The DIPD technique outperforms both the NN approach and the filtering-based baseline techniques, especially for small, challenging objects in congested scenarios. In light of the encouraging initial results from our DIPD method, we hope that the problem, new challenges, and experimental procedure developed here will spur research and new solutions in this area.

ACKNOWLEDGMENTS

The authors would like to thank Mike Huang from Mindtronic AI for his valuable inputs. 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

[1] Steven Ashley. 2014. Touch-less control coming to cars. *Automotive Engineering* (2014), 21.
[2] Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias. 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 (2009), 469–493.

[3] DS automotive. 2018. DS 7. http://www.dsautomobiles.co.uk/ds-models/ds-7-crossback/design/technology.

[4] Subaru automotive. 2018. Subaru Forester. https://www.subaruofniagara.ca/2166-2019-subaru-forester-introduces-facial-recognition/.

[5] Andrew J. Calder, Andrew D. Lawrence, Jill Keane, Sophie K. Scott, Adrian M. Owen, Ingrid Christoffels, and Andrew W. Young. 2002. Reading the mind from eye gaze. *Neuropsychologia* 40, 8 (2002), 1129–1138.

[6] Shan Chen, Zheng Chen, Bin Yao, Xiaocong Zhu, Shiqiang Zhu, Qingfeng Wang, and Yang Song. 2017. Adaptive robust cascade force control of 1-DOF hydraulic exoskeleton for human performance augmentation. *IEEE/ASME Transactions on Mechatronics* 22, 2 (2017), 589–600.

[7] In-Ho Choi, Sung Kyung Hong, and Yong-Guk Kim. 2016. Real-time categorization of driver's gaze zone using the deep learning techniques. In *International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 143–148.

[8] Vincent Delaitre, Josef Sivic, and Ivan Laptev. 2011. Learning person-object interactions for action recognition in still images. In *Advances in Neural Information Processing Systems (NIPS)*. 1503–1511.

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

[10] Luke Fletcher, Gareth Loy, Nick Barnes, and Alexander Zelinsky. 2005. Correlating driver gaze with the road scene for driver assistance systems. *Robotics and Autonomous Systems* 52, 1 (2005), 71–84.

[11] Eduardo A. Gonzalez and Fernando A. Auat Cheein. 2015. Strategy for collaborative navigation in assistive robotics: Reducing user workload. In *The 4th International Conference on Systems and Control (ICSC)*. IEEE, 430–435.

[12] Laura M. Hiatt, Anthony M. Harrison, and J. Gregory Trafton. 2011. Accommodating human variability in human-robot teams through theory of mind. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 22. 2066.

[13] Ashesh Jain, Hema S. Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. 2015. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 3182–3190.

[14] Josef Kittler and Janos Föglein. 1984. Contextual classification of multispectral pixel data. *Image and Vision Computing* 2, 1 (1984), 13–29.

[15] Hema Koppula and Ashutosh Saxena. 2013. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In *Proceedings of the International Conference on Machine Learning*. 792–800.

[16] Linnéa Larsson, Marcus Nyström, and Martin Stridh. 2014. Discrimination of fixations and smooth pursuit movements in high-speed eye-tracking data. In *The 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3797–3800.

[17] R. John Leigh and David S. Zee. 2015. *The neurology of eye movements*. Vol. 90. Oxford University Press, USA.

[18] Yoshio Matsumotot, Tomoyuki Ino, and Tsukasa Ogsawara. 2001. Development of intelligent wheelchair system with face and gaze based interface. In *Proceedings of the 10th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*. IEEE, 262–267.

[19] Wookhee Min, Bradford Mott, Jonathan Rowe, Robert Taylor, Eric Wiebe, Kristy Elizabeth Boyer, and James Lester. 2017. Multimodal goal recognition in open-world digital games. In *Proceedings of the 13th Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. 80–86.

[20] Pontus Olsson. 2007. *Real-time and offline filters for eye tracking*. Master's thesis. KTH, School of Electrical Engineering (EES), Automatic Control.

[21] Toby Perrett and Majid Mirmehdi. 2016. Cost-based feature transfer for vehicle occupant classification. In *Asian Conference on Computer Vision*. Springer, 405–419.

[22] Miquel Ramırez and Hector Geffner. 2011. Goal recognition over POMDPs: Inferring the intention of a POMDP agent. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*. 2009–2014.

[23] Keith Rayner and Monica Castelhano. 2007. Eye movements. *Scholarpedia* 2, 10 (2007), 3649.

[24] Yosef Razin and Karen M. Feigh. 2017. Learning to Predict Intent from Gaze During Robotic Hand-Eye Coordination. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 4596–4602.

[25] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems (NIPS)*. 199–207.

[26] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications*. ACM, 71–78.

[27] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. 2016. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research and Applications*. ACM, 163–170.

[28] Oleg Špakov. 2012. Comparison of eye movement filters used in HCI. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 281–284.

[29] Eliza Strickland. 2017. 3 Ways Ford Cars Could Monitor Your Health. https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/3-ways-ford-cars-could-monitor-your-health.

[30] Karim A. Tahboub. 2006. Intelligent human-machine interaction based on dynamic bayesian networks probabilistic intention recognition. *Journal of Intelligent and Robotic Systems* 45, 1 (2006), 31–52.

[31] Kentaro Takemura, Yoshio Matsumoto, and Tsukasa Ogasawara. 2005. Drive monitoring system based on non-contact measurement system of driver's focus of visual attention. *Nippon Kikai Gakkai Ronbunshu C Hen, Transactions of the Japan Society of Mechanical Engineers* 17, 2 (2005), 505–512.

[32] Martin Tall, Alexandre Alapetite, Javier San Agustin, Henrik H.T Skovsgaard, John Paulin Hansen, Dan Witzner Hansen, and Emilie Møllenbach. 2009. Gaze-controlled driving. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 4387–4392.

[33] Tobii. 2017. Tobii Eye tracker 4C. https://tobiigaming.com/eye-tracker-4c/.

[34] Kristina Yordanova, Samuel Whitehouse, Adeline Paiement, Majid Mirmehdi, Thomas Kirste, and Ian Craddock. 2017. What's cooking and why? Behaviour recognition during unscripted cooking tasks for health monitoring. In *IEEE International Conference on Pervasive Computing and Communications (PerCom) Workshops*. IEEE, 18–21.