Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 186–200

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Situated language understanding for a spoken dialog system within vehicles[☆]

Teruhisa Misu[a,*], Antoine Raux[a,1], Rakesh Gupta[a], Ian Lane[b]

[a] Honda Research Institute USA, 425 National Avenue, Mountain View, CA 94043, United States
[b] Carnegie Mellon University, NASA Ames Research Park, Moffett Field, CA 93085, United States

Received 29 May 2014; received in revised form 3 February 2015; accepted 14 February 2015
Available online 23 February 2015

Abstract

In this paper, we address issues in situated language understanding in a moving car, which has the additional challenge of being a rapidly changing environment. More specifically, we propose methods for understanding user queries regarding specific target buildings in their surroundings. Unlike previous studies on physically situated interactions, such as interactions with mobile robots, the task at hand is very time sensitive because the spatial relationship between the car and target changes while the user is speaking. We collected situated utterances from drivers using our research system called Townsurfer, which was embedded in a real vehicle. Based on this data, we analyzed the timing of user queries, the spatial relationships between the car and the targets, the head pose of the user, and linguistic cues. Based on this analysis, we further propose methods to optimize timing and spatial distances and to make use of linguistic cues. Finally, we demonstrate that our algorithms improved the target identification rate by 24.1% absolute.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Spoken dialog systems; Situated dialog; Language understanding; Reference resolution

1. Introduction

Recent advances in sensing technologies have enabled researchers to explore applications requiring a clear awareness of a system's dynamic context and physical surroundings. Such applications include multi-participant conversation systems (Bohus and Horvitz, 2009) and human robot interaction systems (Tellex et al., 2011; Sugiura et al., 2011). The general problem of understanding and interacting with human users in such environments is referred to as *situated interaction*.

In this current study, we address another environment in which situated interactions take place, i.e., a moving car. In our previous work, we collected over 60 h of in-car human–human interactions in which drivers interacted with an expert co-pilot sitting next to them in the vehicle (Cohen et al., 2014). One of the insights from the analysis on this corpus is that drivers frequently use referring expressions about their surroundings (e.g., *What is <u>that big building</u> on*

[☆] This paper has been recommended for acceptance by R.K. Moore.

[*] Corresponding author. Tel.: +1 6503140416.
E-mail address: teruhisa.misu@gmail.com (T. Misu).

[1] Currently with Lenovo.

http://dx.doi.org/10.1016/j.csl.2015.02.002

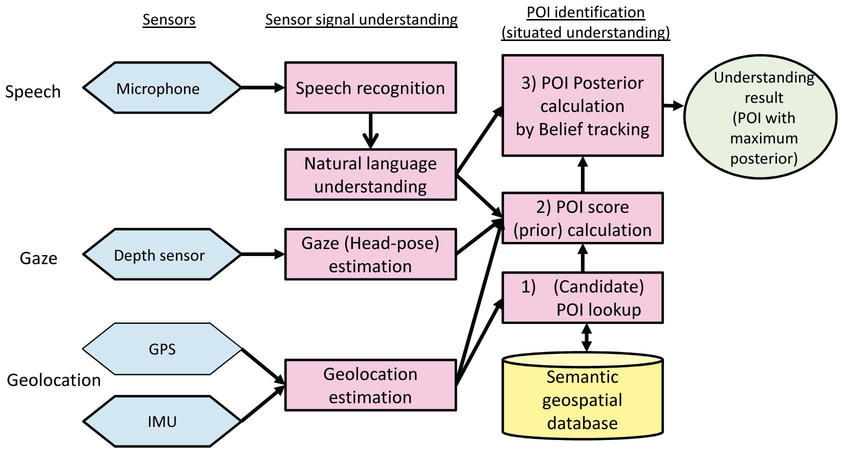
Sensors
Sensor signal understanding
POI identification (situated understanding)
Speech
Microphone
Speech recognition
Natural language understanding
Gaze
Depth sensor
Gaze (Head-pose) estimation
Geolocation
GPS
IMU
Geolocation estimation
3) POI Posterior calculation by Belief tracking
2) POI score (prior) calculation
1) (Candidate) POI lookup
Semantic geospatial database
Understanding result (POI with maximum posterior)

Fig. 1. System overview of Townsurfer (POI identification part).

the right?). Based on this insight, we developed Townsurfer (Lane et al., 2012; Misu et al., 2013), a situated in-car intelligent assistant. Using geolocation information, the system can answer user queries/questions that contain object references regarding points-of-interest (POIs) in their surroundings. We use driver (user) face orientation to understand their queries and provide the requested information regarding the POI they are looking at. We previously demonstrated and evaluated this system in a simulated environment (Lane et al., 2012). In this paper, we evaluate its utility in real driving situations.

Compared to conventional situated dialog tasks, query understanding in our task is expected to be more time sensitive, due to the rapidly changing environment while driving. Typically, a car moves 10 m in one second while driving at 25 mi/h (40 km/h). Therefore, timing can be a crucial factor. In addition, it is not well understood what kind of linguistic cues are naturally provided by drivers, and their contributions to situated language understanding in such an environment. To the best of our knowledge, this is the first study that tackles the issue of situated language understanding in rapidly moving vehicles.

In addition to this introductory section, we present an overview of the Townsurfer in-car spoken dialog system in Section 2. Based on our data collection using this system, we analyze user behavior while using the system with a focus on language understanding; our findings are presented in Section 3. More specifically, we address the following research questions regarding the task and the system through data collection and analysis:

1 Is timing an important factor of situated language understanding?
2 Does head pose play an important role in language understanding? Or is spatial distance information enough?
3 What is the role of linguistic cues in this task? What kinds of linguistic cues do drivers naturally provide?

Based on the hypotheses obtained from the analysis of these questions, we propose methods to improve situated language understanding in Section 4, and then analyze their contributions based on the collected data in Sections 5 and 6. We then clarify our research contributions through discussion in Section 7 and compare our work with related studies in Section 8. Finally, we conclude our work in Section 9.

2. Architecture and hardware of Townsurfer

The Townsurfer system uses three main input modalities, namely speech, geolocation, and head pose. Speech is the primary input modality and is used to trigger interactions with the system. User speech is recognized, and requested concepts and values are then extracted. Geolocation and head pose information are used to understand the target POI of the user query. An overview of the system and its process flow is illustrated in Fig. 1; an example dialog with the system is shown in Table 1.

In our research, we address issues in identifying user-intended POIs; this is a form of reference resolution using multi-modal information sources. That is, we handle information request about specific POIs, such as U1 and U3 in

Table 1
Example dialog with Townsurfer.

U1:	What is *that place*? (POI in gaze)
S1:	This is Specialty Cafe, a midscale coffee shop that serves sandwiches.
U2:	What is *its* (POI in dialog history) rating?
S2:	The rating of Specialty Cafe is above average.
U3:	How about *that one* on the left? (POI located on the left)
S3:	This is Roger's Deli, a low-priced restaurant that serves American food.

Table 1. We do not currently address issues in language understanding related to dialog history and query type, for example, requests regarding specific properties of POIs, such as U2 in the table. The POI identification process consists of the three steps summarized below (cf. Fig. 1). Note that these steps are similar to, but different from our previous work on landmark-based destination setting (Ma et al., 2012).

(1) The system lists candidate POIs based on geolocation at the time of a driver query. Relative positions of POIs to the car are calculated based on geolocation and the heading of the car.
(2) Based on spatial linguistic cues in the user's speech (e.g., *to my right, on the left*), a 2D scoring function is selected to identify areas in which the target POI is likely to be present. This function takes into account the position of the POI relative to the car, as well as the driver's head pose. Scores for all candidate POIs are calculated.
(3) Posterior probabilities of each POI are calculated using the score of step 2 as prior, and non-spatial linguistic information (e.g., POI categories, building properties) as observations. This posterior calculation is computed using our Bayesian belief tracker, DPOT (Raux and Ma, 2011).

Further details are provided in Section 4.

The system hardware consists of a 3D depth sensor (PrimeSense Carmine 1.09), a USB GPS (BU-353S4), an inertial measurement unit (IMU) sensor (3DM-GX3-25), and a close-talking microphone (Plantronics Voyager Legend UC). We installed these consumer-grade sensors in our experimental Honda Pilot. We used the Point Cloud Library[2] (PCL) for face direction estimation, with the model included in the library. The geolocation was estimated based on an Extended Kalman filter-based algorithm using GPS and gyro information as input at 1.5 Hz. The system was implemented based on the Robot Operating System (ROS) (Quigley et al., 2009). Each component was implemented as a node of the ROS, and the communication between nodes was performed using standard message passing mechanisms of ROS.

3. Data collection and analysis

3.1. Collection setting

We collected data using a test route that passed through **downtown** Mountain View and **residential areas** around the Honda Research Institute. We assumed that a POI is downtown when it is located within the rectangle formed by two geolocation coordinates. We manually constructed a database that contained 250 POIs including businesses, restaurants, and other companies in the area. Each database entry (i.e., POI) had a name, geolocation, category and property information (as explained in Section 3.4). POI geolocation was represented as a latitude–longitude pair (e.g., 37.4010, −122.0539). The size and shape of buildings were not taken into account. The route took approximately 30 min to drive, with major difference between the residential areas and downtown in the POI density. While each POI located downtown had on average 7.2 other POIs within 50 m, in residential areas, POIs had only 1.9 such neighbors. Speed limits also differed between the two (35 mi/h and 25 mi/h, respectively).

We collected data from 14 subjects. Each participant was asked to drive the test route and make queries about surrounding businesses. We showed a demo video of the system to the users before starting the data collection. We also told them that the objective was data collection for a situated spoken dialog system rather than the evaluation

[2] http://pointclouds.org/.

Table 2
Example user utterances.

- What is that blue restaurant on the right?
- How about this building to my right with outside seating?
- What is that Chinese restaurant on the left?
- Orange building to my right.
- What kind of the restaurant is that on the corner?
- The building on my right at the corner of the street.
- What about the building on my right with the woman with a jacket in front?
- Do you know how good is this restaurant to the left?
- Townsurfer, there is an interesting bakery; what is that?
- Is this restaurant on the right any good?

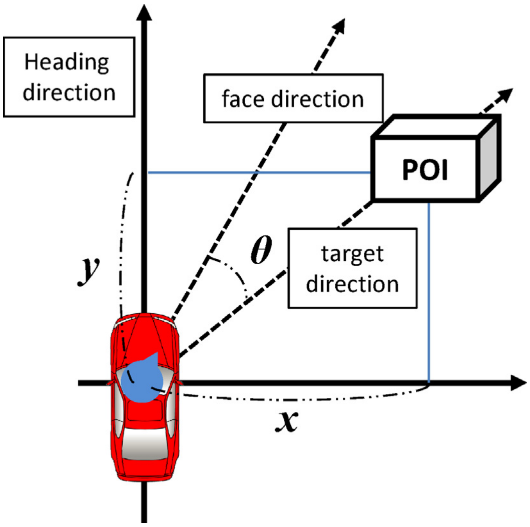

Fig. 2. Parameters used to calculate POI score (prior).

of the system. We asked subjects to include the full description of the target POI within a single utterance to avoid queries whose understanding required dialog history information.[3] Although the system answered based on the baseline strategy explained in Section 4.1, we asked subjects to ignore system responses.[4]

We collected 399 queries with valid target POIs. Example user utterances are shown in Table 2. Queries regarding businesses that do not exist in our database (typically a vacant store) were excluded. The data contain 171 and 228 POIs in the downtown and residential areas, respectively. The queries were transcribed and user-intended POIs were manually annotated by confirming the intended target POI with the subjects after data collection based on a video taken during the drive.[5]

3.2. Analysis of spatial relationship of POIs and head pose

We first analyzed the spatial relationship between position cues (right/left) and the position of the user-intended target POIs. Out of the collected 399 queries, 237 (59.4%) contained either right or left position cues (e.g., *What is that on the left?*). The relationship between the position cues (cf. Fig. 2) and POI positions at the start-of-speech timing[6] is plotted in Fig. 3. In the figure, the X-axis is the lateral distance (i.e., the distance in the direction orthogonal

[3] Including dialog history information is earmarked as part of our future work.
[4] We thought that giving system responses to the users will create better engagement for them.
[5] This means that there is a possibility of annotation errors.
[6] Specifically, the latest GPS and face direction information at that timing is used.

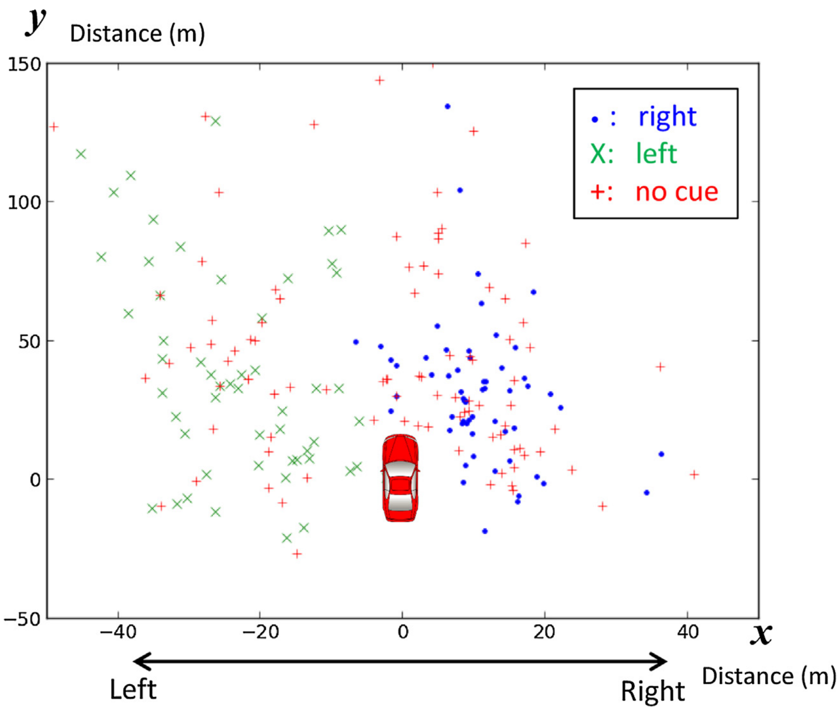

Fig. 3. Target POI positions.

to the heading with a positive value indicating the right) and the *Y*-axis is the axial distance (i.e., the distance in the heading direction with a negative value indicating that the POI is in back of the car). The most obvious finding from the scatterplot is that right and left are certainly powerful cues for the system in identifying target POIs.[7] We also observed that the POI position distribution has a large standard deviation. This is partly because the route has multiple sites from both downtown and the residential areas. While the average distance to the target POI downtown was 37.0 m, that in the residential areas was 57.4 m. POI positions per site are illustrated in Fig. 4.

We also analyzed the relationship between face direction and POI positions. Fig. 5 plots the relationship between the axial distance and angular difference θ (i.e., between the user face direction and the target POI direction) (cf. Fig. 2). The scatterplot suggests that the angular differences for distant target POIs are often small. For close target POIs, the angular differences are larger and have a larger variance.[8]

3.3. *Analysis of timing*

Referring expressions such as “the building on the right” must be resolved with respect to the context in which the user intended; however, in a moving car, such a context (i.e., the position of the car and the situation in the surroundings) can be very different between the time when the user starts speaking the sentence and the time they finish speaking it. Therefore, situated understanding must be very time sensitive.

To confirm and investigate this issue, we analyze the difference in the POI positions between the time the ASR result is output vs. the time the user actually started speaking. The hypothesis is that the latter yields a more accurate context in which to interpret the user sentence. In contrast, our baseline system uses the more straightforward approach of

[7] There were some cases that points which were geometrically to the left of the car were described as being to the right meaning that a driver was speaking while making a left turn. This is because there is one intersection with left turn in the test route that does not have a traffic signal (cf. Appendix 3rd maneuver), and the subjects entered the intersection without slowing down. On the other hand, all right turns in the test route were made in the intersections with traffic signals, thus the subjects stopped and made queries. We think this is the reason why the corresponding error for left did not occur.

[8] We will discuss the reason for this in Section 6.2.

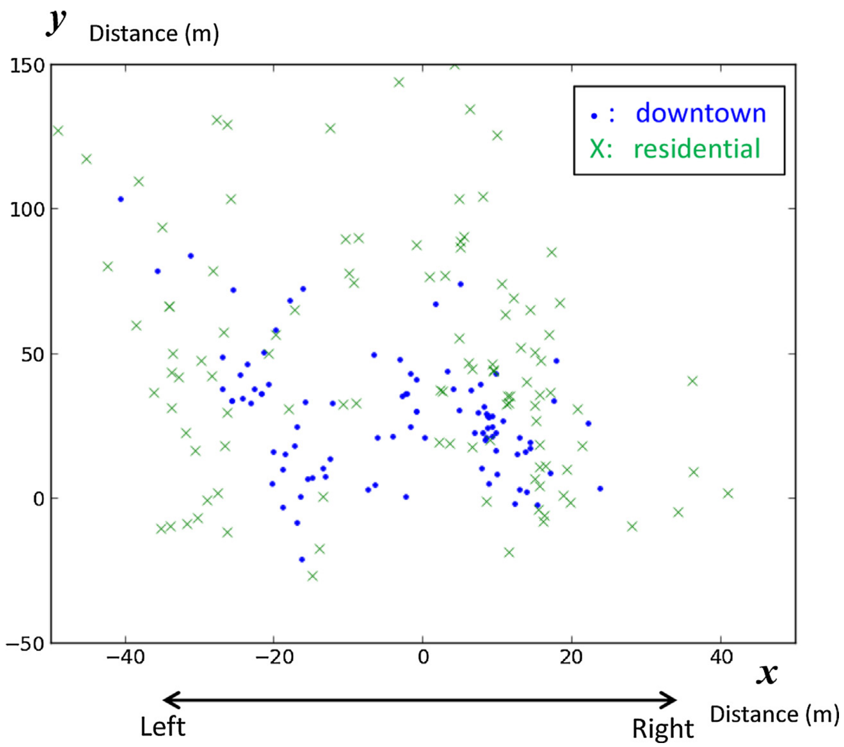

Fig. 4. Relation of POI positions (downtown vs. residential area).

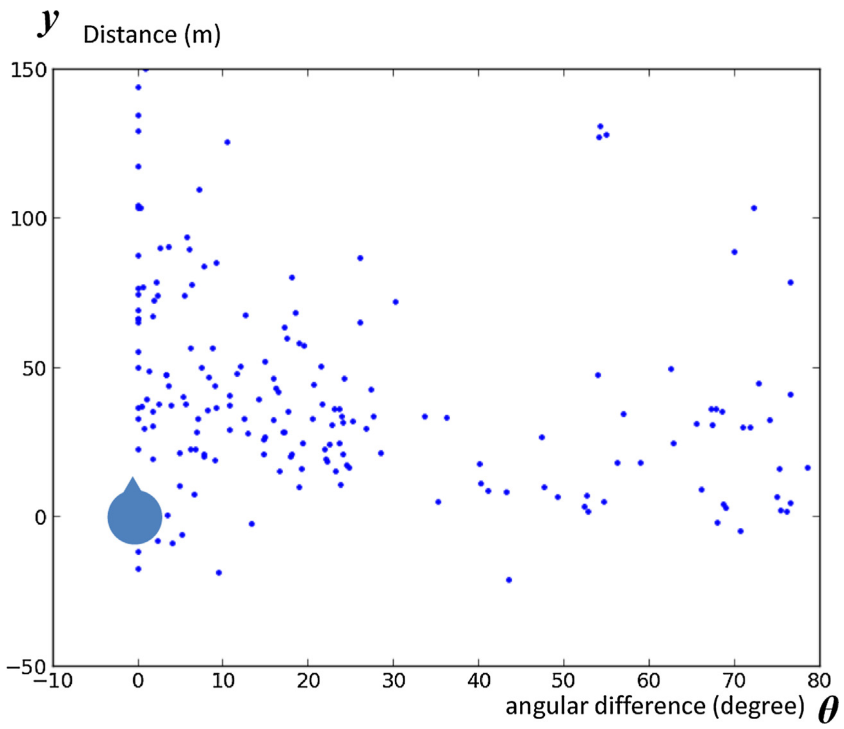

Fig. 5. Relation between POI positions and head pose.

Table 3
Comparison of average and standard deviation of distance (m) of POI from the car.

Position cue	Site	ASR result timing		Start-of-speech timing	
		Ave dist.	Std dist.	Ave dist.	Std dist.
Right/left	Downtown	17.5	31.0	31.9	28.3
	Residential	22.0	36.3	45.2	36.5
No right/left cue	Downtown	17.4	27.8	31.1	26.5
	Residential	38.3	45.9	52.3	43.4

Table 4
User-provided linguistic cues.

Category of linguistic cue	Percentage used (%)
Relative position to the car (right/left)	59.4
Business category (e.g., restaurant, cafe)	31.8
Color of the POI (e.g., green, yellow)	12.8
Cuisine (e.g., Chinese, Japanese, Mexican)	8.3
Equipments (e.g., awning, outside seating)	7.2
Relative position to the road (e.g., corner)	6.5

resolving expressions using the context at the time of resolution, i.e., whenever the ASR/NLU has finished processing an utterance (hereafter “ASR results timing”).

Specifically, we compare the average axial distance to the target POIs and its standard deviation between these two timings; Table 3 lists these figures as per the query types of position cues and sites. The average axial distance from the car to the target POIs is small at the ASR result timing; however the standard deviation is generally small at the start-of-speech timing.[9] This indicates that the target POI positions at the start-of-speech timing are more consistent across users and sentence lengths than that at the ASR result timing. This result indicates the possibility of a better POI likelihood function using the context (i.e., car position and orientation) at the start-of-speech timing than using the ASR result timing.

3.4. *Analysis of linguistic cues*

We analyzed the linguistic cues provided by the users. Here, we focus on objective and stable cues. We excluded subjective cues (e.g., *big, beautiful, colorful*) and cues that might change in a short period of time (e.g., *with a woman dressed in green in front*).[10] We categorized the linguistic cues used to describe the target POIs; Table 4 lists the cue types and the percentage of user utterances containing each cue type.

The cues that the users most often provided concern POI positions relative to the car (right and left). Nearly 60% of the queries included this type of cue, and every subject provided it at least once. The second most frequent cue is the category of business, especially downtown. Users also provided colors of POIs; other cues include cuisine, equipment, and relative position to the road (e.g., *on the corner*).

Another interesting finding is that users provided more linguistic cues as the number of candidate POIs in their field of view increased. More specifically, users provided an average of 1.51 categories per query downtown, and an average of only 1.03 categories in the residential areas, though the difference was not statistically significant (cf. POI density in Section 3.2: 7.2 vs. 1.9). This indicates that users provided cues based in part on environment complexity.

[9] The results are not statistically significant.

[10] Percentages of user utterances containing these cues are 3.3% and 0.5%, respectively.

4. Methods for situated language understanding

4.1. Baseline strategy

We used our previous version (Misu et al., 2013) as a baseline system for situated language understanding. The baseline strategy is summarized in the three paragraphs below, which correspond to processes (1), (2), and (3) in Section 2 and Fig. 1.

The system performs a POI lookup based on the geolocation information at the time the ASR result is obtained. The search range of candidate POIs was within the range (i.e., relative geolocation of POIs against the car location) of −50 to 200 m in the traveling direction and 100 m to both the left and right in the lateral direction. The ASR result timing was also used to measure distances to the candidate POIs.

POI priors were calculated based on the distance from the car (the axial distance) following the “the closer to the car the more likely” principle. We used a likelihood function inversely proportional to the distance. Furthermore, we used position cues simply to remove POIs from a list of candidates. For example, the “right” position cue was used to remove candidate POIs located to the left (with a lateral distance less than zero). When no right or left cue was provided, POIs outside 45° from the face direction were removed from the list of candidates.

No linguistic cues except right and left were used to calculate POI posterior probabilities. Therefore, the system selected the POI with the highest prior (POI score) as the language understanding result.

4.2. Strategies toward better situated language understanding

To achieve better situated language understanding (and therefore POI identification) based on our analysis in Section 3, we modified steps (1), (2), and (3) as follows:

1 Use start-of-speech timing for the POI prior calculation
2 Use a Gaussian mixture model (GMM)-based POI probability (prior) calculation
3 Use linguistic cues for the posterior calculation.

We used start-of-speech timing instead of the time the ASR result is output; because the standard deviations of the POI distances were small (cf. Section 3.2), we expect that a better POI probability score estimation with the POI positions at this timing will be achieved in subsequent processes. The POI lookup range was the same as the baseline.

We applied a GMM with diagonal covariance matrices over the input parameter space (Bishop, 2006). The POI probability (prior) $p(\mathbf{x})$ was given by the following equation.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

where $\mathbf{x}$ is a d-dimensional data vector (input parameters), $\boldsymbol{\pi}_k$, $k = 1, \ldots, K$, are the mixture weights, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are the component Gaussian densities. Each component density is a d-variate Gaussian function of the form,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\boldsymbol{\pi})^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\},$$

with mean vector $\boldsymbol{\mu}_k$ and diagonal covariance matrix $\boldsymbol{\Sigma}_k$. The mixture weights sum up to 1 ($\boldsymbol{\Sigma}_{k=1}^{K}\boldsymbol{\pi}_k = 1$). These parameters are estimated using EM algorithm (Bishop, 2006) based on data. We used two input parameters ($d = 2$), i.e., the lateral and axial distances for queries with right or left cue; for queries without right or left cues, we used three parameters ($d = 3$), i.e., the lateral and axial distances, and the difference in degree between the target and head pose directions (The effect of these parameters is discussed in Section 6.2). We empirically set the number of Gaussian components to 2 ($K = 2$). An example GMM fitting to the POI positions for queries with right and left cues is illustrated in Fig. 6 in which the center of the ellipse is the mean of the Gaussian.

We used the five linguistic cue categories of Section 3.4 for the posterior calculation by the belief tracker (Raux and Ma, 2011). In our system, because we do not use contextual information the likelihood is given by agreement between the NLU result and system properties. In the following experiments, we used either 1 or 0 as the likelihoods of

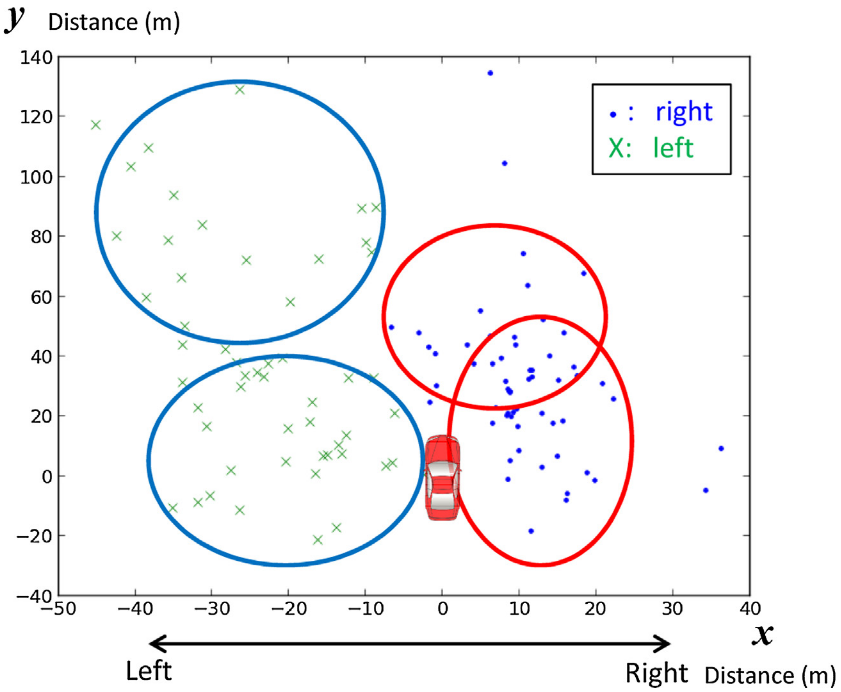

Fig. 6. Example GMM fitting.

the natural language understanding (NLU) observations. The likelihood of a category value is 1 if a user query (NLU result) contains the target value or a user query does not contain any target value; otherwise, 0. The likelihood of a query is given by the product of the likelihoods of these categories. This corresponds to a strategy of simply removing candidate POIs that do not have category values specified by the user, and using the priors of the remaining candidate POIs as their posteriors. Here, we assumed a clean POI database with all properties manually annotated.

5. Experiments

We used manual transcriptions and natural language understanding results of user queries to focus our evaluations on the issues listed in Section 1. We evaluated the situated language understanding (POI identification) performance based on cross-validation. We used the data from 13 users to train the GMM parameters and defined a set of possible linguistic values; the data from the remaining user was used for evaluation. We trained the model parameters of the GMM ($\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) using the EM algorithm (Bishop, 2006). Knowledge about the sites (downtown or residential) was not used in the training (the performance was better when the knowledge was not used).

We did not set a rejection threshold. We judged that the system successfully understood a user query when the posterior of the target (user-intended) POI was the highest. The chance rate, given by the average of the inverse number of candidate POIs in the POI lookup, was 10.0%.

6. Analysis of the results

We first analyzed the effect of our three methods described in Section 4.2; the results are listed in Table 5.

Simply using the POI positions at the start-of-speech timing instead of those at the ASR result timing did not lead to any improvements. This result is reasonable because the distances to target POIs were often smaller at the ASR result timing, as shown in Table 3 above; however, we achieved an improvement (7.5% over the baseline) by combining this modification with GMM-based prior calculations. The results support our hypothesis (Section 3.3) that the POI position is less dependent on users and scenes at the start-of-speech timing. The linguistic cues were the most powerful information for this task. The improvement over the baseline was 11.5%. By using these three methods together, we

Table 5
Comparison of POI identification rate.

Method	Success rate (%)
Right/left linguistic cues, the-closer-the-likely likelihood, ASR result timing (**Baseline**)	43.1
1) Start-of-speech timing	42.9
2) GMM-based prior	47.9
3) Linguistic cues	54.6
1) + 2)	50.6
1) + 3)	54.4
2) + 3)	62.2
1) + 2) + 3)	67.2

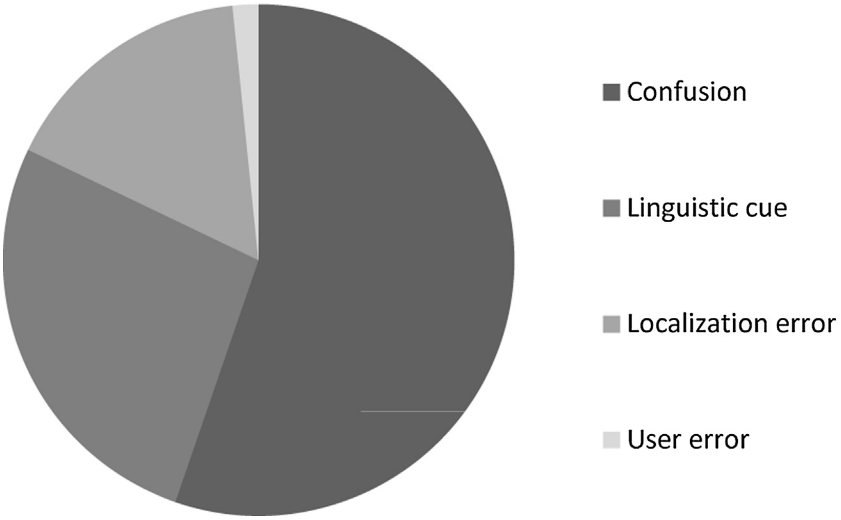

Fig. 7. Breakdown of error causes.

obtained a more-than-additive improvement of 24.1% in the POI identification rate over the baseline.[11] The success rates per site were 60.8% downtown and 71.9% in the residential areas.

6.1. Error analysis

To analyze the causes of the remaining errors, we categorized the errors into the following four categories:

1. **Ambiguous references**: There were multiple POIs that matched the user query (e.g., another *yellow building* was positioned next to the target).
2. **Linguistic cues**: The driver used undefined linguistic cues such as subjective expressions or dynamic references to other objects (e.g., optometrist, across the street, colorful).
3. **Localization errors**: Errors in estimating geolocation or the heading of the car.
4. **User errors**: There were errors in user descriptions (e.g., the user misunderstood the neighbor POI's outside seating as belonging to the target POI).

The distribution of error causes is illustrated in Fig. 7. More than half of the errors were related to reference ambiguity. These errors are expected to be resolved through clarification dialogs[12] (e.g., asking the user, "*Did you*

[11] For reference, the performances of "1) + 2) + 3)" were 62.9%, 67.2%, 66.1%, 67.2%, and 66.2% when the number of Gaussian components was 1, 2, 3, 4, and 5, respectively.
[12] The current system does not have any clarification dialog. It will be a part of our future work.

Table 6
Relation between the parameters used for the POI identification and success rates (%).

Parameters used	Query type	
	Right/left	No cue
Lateral (x) distance	58.6	51.2
Axial (y) distance	59.5	53.7
Face direction	43.3	44.4
Lateral + axial ($x+y$)	73.8	54.3
Lateral (x) + face direction	57.8	48.1
Axial (y) + face direction	59.1	54.9
Lateral + axial + face	68.4	57.4

Table 7
Effect of linguistic cues.

Linguistic cue category used	Success rate (%)
No linguistic cues (*)	50.6
(*) + Business category (e.g., cafe)	59.1
(*) + Color of the POI (e.g., green)	57.6
(*) + Cuisine (e.g., Chinese)	54.1
(*) + Equipments (e.g., awning)	53.9
(*) + Relative position (e.g., corner)	51.4

mean the one in front or back?”). Linguistic errors might be partly resolved by using a better database with more detailed category information. For dynamic references and subjective cues, the use of image processing techniques may help. Localization errors can be solved by using high-quality GPS and IMU sensors. Finally, user errors were rare and only made in downtown where potential POIs were dense.

6.2. Breakdown of effect of the spatial distance and head pose

Next, we evaluated the features (input parameters used as $\mathbf{x}$ in Section 4.2) used for the POI prior calculation to investigate the effect of the input parameters, i.e., the lateral and axial distances and the difference in degree between the target and user face direction angles. Table 6 lists the relationships between the parameters used for the GMM-based prior calculation and the POI identification performance.[13]

Results indicated that the axial distance is the most important parameter. We achieved a slight improvement by using face direction information for queries without right or left cues, but the improvement was insignificant. Conversely, the use of face direction information for the right and left queries clearly degraded the POI identification performance; this might have occurred because users finished looking at the POI and again faced front when they started speaking, thus they explicitly provided right or left information to the system.

6.3. Breakdown of the effect of linguistic cues

We evaluated the effect of linguistic cues for each category. Table 7 lists the relationships between the categories used for the posterior calculation and the success rates. From the table, we observe that there is a strong correlation between the frequency of the cues used (cf. Table 4) and their contributions to the improvement in success rate. For example, business category information contributed the most, boosting performance by 8.5%.

[13] Note that we first determined the function to calculate POI scores (priors) based on position cues, and then calculated scores with the selected function.

Another observation here is that the contribution of the business and cuisine categories is large. Even though other categories (e.g., color) were not readily available in public POI databases (e.g., Google Places API,[14] Yelp API[15]), we obtained reasonable performance without using a special database or image processing techniques.

We also observed that linguistic cues were especially effective downtown. While the improvement[16] was 20.0% downtown, that for the residential areas was 14.4%, primarily because users provided more linguistic cues downtown in consideration of the difficulty of the task.

6.4. Using speech recognition results

We evaluated the degradation due to the use of ASR results. We used Google ASR[17] and the Julius speech recognition system (Kawahara et al., 2004) with a language model trained from 38,000 example sentences generated from a grammar. An acoustic model trained from the Wall Street Journal (WSJ) speech corpus was used. Note that these are not necessarily the best corpora for this domain. Google ASR uses a general language model for dictation and Julius uses a mismatched acoustic model in terms of the noise condition.

The query success rate was 56.3% for Julius and 60.3% for Google ASR and achieved ASR accuracies were 77.9% and 80.4%, respectively. We believe that the performance will improve when N-best hypotheses with confidence scores are used in the posterior calculations using the belief tracker, though parameter tuning for handling confidence score will be a non-trivial issue.

7. Discussion

The major limitation of our work is the small amount of data we were able to collect. It is unclear how our results would be generalize to other sites, POI densities, velocities, and sensor performances. Furthermore, results might depend on experimental conditions such as weather, hours, and seasons. Hyperparameters, such as the optimal number of Gaussian components, might have to be adapted to different situations. We therefore acknowledge that the route and settings we experimented with are only a limited case of daily driving activities; however, the methods we proposed are general-purpose and our findings should be verifiable without loss of generality by collecting more data and using more input parameters for the POI prior calculations.

Moreover, much future work is required for realizing natural interactions with the system, such as incorporating dialog history and selecting optimal system responses. Conversely, we believe that this is one of the best platforms for investigating situated interactions. The major topics that we will tackle in our future work include the following:

1. Dialog strategy: Dialog strategy and system prompt generation for situated environments are important research topics, especially for clarifying the target when there is ambiguity as mentioned in Section 6.1. The topic will include an adaptation of system utterances (entrainment) to the user (Hu et al., 2014).
2. Eye tracker: Although we believe that the head pose is good enough for estimating user intentions because we are trained to move our head in driving schools to look around to confirm safety, we would like to confirm any differences between face direction and eye-gaze direction.
3. POI identification using face direction trajectory: Our analysis showed that the use of face direction sometimes degraded the POI identification performance; however, we believe that using a trajectory of face direction will improve results.
4. Database: We assumed a clean and perfect database; however, we will evaluate the performance when a noisy database is used (e.g., a database based on image recognition results or user dialog logs).
5. Feedback: Koller et al. (2012) demonstrated that referential resolution is enhanced by providing gaze information feedback to the user. We would like to analyze the effect of such feedback with an automotive augmented reality environment using our 3D head-up display (Ng-Thow-Hing et al., 2013).

[14] http://developers.google.com/places/.

[15] http://www.yelp.com/developers/.

[16] 1) + 2) vs. 1) + 2) + 3).

[17] Although it is not realistic to use a cloud-based speech recognition system considering the current latency, we use this as a reference system.

8. Related work

There are many commercial spoken dialog systems for in-car interactions (e.g., DragonDrive[18] and G-BOOK[19]); however, their use of situated information is limited. They only use geolocation information to select from reply candidates (e.g., select the closest branch of a business). They do not handle information regarding the immediate surroundings of the car as seen by the driver, or distinguish input timing differences at a fine level of granularity (i.e., on the order of one second or less).

Other related studies include a landmark-based navigation system that handles landmarks as information for a dialog. Similar system concepts have been provided for pedestrian navigation situations (Janarthanam et al., 2013; Hu et al., 2014), though they do not handle a rapidly changing environment.

Several works have used timing to enhance natural interaction with systems. Rose and Kim (2003) and Raux and Eskenazi (2009) used timing information to detect user barge-ins. Studies on incremental speech understanding and generation (Skantze and Hjalmarsson, 2010; Dethlefs et al., 2012) have proved that real-time feedback actions have potential benefits for users. Komatani et al. (2012) used user speech timing against a user's previous and a system's utterances to understand the intentions of user utterances. While the above studies have handled timing issues by focusing on (para-)linguistic aspects, our work handles timing issues in relation to the user's physical surroundings.

Recent advancements in gaze and face direction estimation have led to better user behavior understanding. There are a number of studies that have analyzed the relationship between gaze and user intention, such as user focus (Yonetani et al., 2010), preference (Kayama et al., 2010), and reference expression understanding (Koller et al., 2012), and between gaze and turn-taking (Jokinen et al., 2010; Kawahara, 2012). Nakano et al. (2013) used face direction for addressee identification. Previous studies most related to our current work include reference resolution methods by Chai and Prasov (2010), Iida et al. (2011) and Kennington et al. (2013). They confirmed that the system's reference resolution performance is enhanced by taking the user's eye fixation into account; however, their results were not directly applicable to an interaction in a rapidly changing environment while driving in which eye fixations are unusual activities.

Marge and Rudnicky (2010) analyzed the effect of space and distance for spatial language understanding in a human–robot interaction. Our task differs from this because we handle a rapidly changing environment. We believe we can improve our understanding performance based on their findings.

9. Conclusion

In this paper, we addressed situated language understanding in a moving car. We focused on issues in understanding user language regarding timing, spatial distance, and linguistic cues. Based on our analysis of collected user utterances, we proposed methods of using start-of-speech timing for the POI prior calculation, GMM-based POI probability (prior) calculation, and linguistic cues for the posterior calculation to improve the accuracy of situated language understanding. The effectiveness of our proposed methods was confirmed by achieving a significant improvement in a POI identification task.

Acknowledgements

The authors would like to thank Yi Ma at Ohio State University for his contributions to the development of HRItk.

Appendix A.

Test route and the maneuvers are displayed in Fig. 8.

Test route (Google Map):

`https://www.google.com/maps/preview/dir/Honda+Research+Institute,+425 +National+Ave+%23100,+Mountain+View,+CA+94043/37.4009909,-122.0518957/`

[18] http://www.nuance.com/for-business/mobile-solutions/dragon-drive/index.htm.

[19] http://g-book.com/.

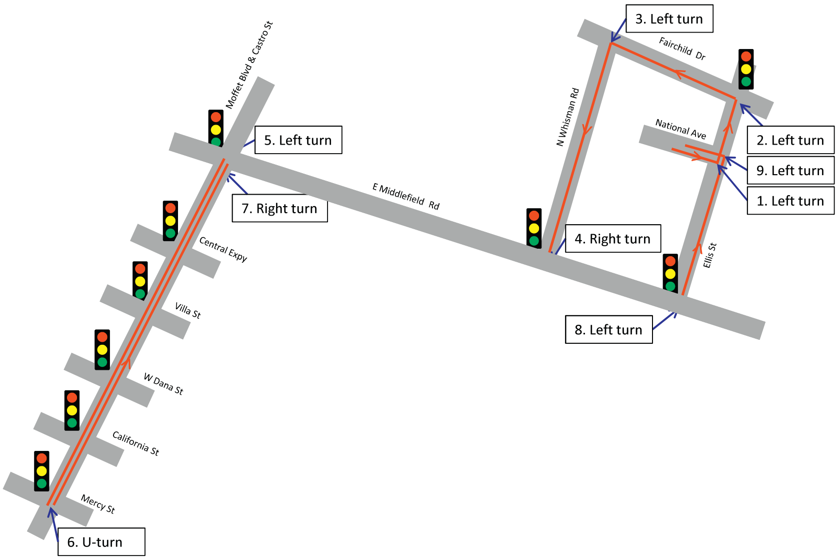

Fig. 8. The route and the maneuvers.

```
37.4052337,-122.0565795/37.3973374,-122.0595982/37.4004787,-
122.0730021/Wells+Fargo/37.4001639,-122.0729708/37.3959193,-
122.0539449/37.4009821,-122.0540093/@37.3999836,-122.0792529,14z/data=!4m21!
4m20!1m5!1m1!1s0x808fb713c225003d:0xcf989a0bb230e5c0!2m2!1d-
122.054006!2d37.401016!1m0!1m0!1m0!1m0!1m5!1m1!1s0x0:0x86ca9ba8a2f15150!2m2!
1d-122.082546!2d37.388722!1m0!1m0!1m0!3e0.
```

References

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Bohus, D., Horvitz, E., 2009. Models for multiparty engagement in open-world dialog. In: Proc. SIGDIAL, pp. 225–234.

Chai, J., Prasov, Z., 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric reference in situated dialogue. In: Proc. EMNLP.

Cohen, D., Chandrashekaran, A., Lane, I., Raux, A., 2014. The HRI-CMU corpus of situated in-car interactions. In: Proc. IWSDS, pp. 201–212.

Dethlefs, N., Hastie, H., Rieser, V., Lemon, O., 2012. Optimising incremental dialogue decisions using information density for interactive systems. In: Proc. EMNLP, pp. 82–93.

Hu, Z., Halberg, G., Jimenez, C., Walker, M., 2014. Entrainment in pedestrian direction giving: How many kinds of entrainment? In: Proc. IWSDS, pp. 90–101.

Iida, R., Yasuhara, M., Tokunaga, T., 2011. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In: Proc. IJCNLP, pp. 84–92.

Janarthanam, S., Lemon, O., Liu, X., Bartie, P., Mackaness, W., Dalmas, T., 2013. A multithreaded conversational interface for pedestrian navigation and question answering. In: Proc. SIGDIAL, pp. 151–153.

Jokinen, K., Nishida, M., Yamamoto, S., 2010. On eye-gaze and turn-taking. In: Proc. EGIHMI.

Kawahara, T., 2012. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In: Proc. SIGDIAL.

Kawahara, T., Lee, A., Takeda, K., Itou, K., Shikano, K., 2004. Recent progress of open-source LVCSR Engine Julius and Japanese model repository. In: Proc. ICSLP, Vol. IV.

Kayama, K., Kobayashi, A., Mizukami, E., Misu, T., Kashioka, H., Kawai, H., Nakamura, S., 2010. Spoken dialog system on plasma display panel estimating user's interest by image processing. In: Proc. 1st International Workshop on Human-Centric Interfaces for Ambient Intelligence (HCIAmi).

Kennington, C., Kousidis, S., Schlangen, D., 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. In: Proc. SIGDIAL.

Koller, A., Garoufi, K., Staudte, M., Crocker, M., 2012. Enhancing referential success by tracking hearer gaze. In: Proc. SIGDIAL, pp. 30–39.

Komatani, K., Hirano, A., Nakano, M., 2012. Detecting system-directed utterances using dialogue-level features. In: Proc. Interspeech.

Lane, I., Ma, Y., Raux, A., 2012. AIDAS – Immersive Interaction within Vehicles. In: Proc. SLT.

Ma, Y., Raux, A., Ramachandran, D., Gupta, R., 2012. Landmark-based location belief tracking in a spoken dialog system. In: Proc. SIGDIAL, pp. 169–178.

Marge, M., Rudnicky, A., 2010. Comparing spoken language route instructions for robots across environment representations. In: Proc. SIGDIAL, pp. 157–164.

Misu, T., Raux, A., Lane, I., Devassy, J., Gupta, R., 2013. Situated multi-modal dialog system in vehicles. In: Proc. Gaze in Multimodal Interaction, pp. 25–28.

Nakano, Y., Baba, N., Huang, H., Hayashi, Y., 2013. Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In: Proc. ICMI, pp. 35–42.

Ng-Thow-Hing, V., Bark, K., Beckwith, L., Tran, C., Bhandari, R., Sridhar, S., 2013. User-centered perspectives for automotive augmented reality. In: Proc. ISMAR.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A., 2009. ROS: an open-source Robot Operating System. In: Proc. ICRA Workshop on Open Source Software.

Raux, A., Eskenazi, M., 2009. A finite-state turn-taking model for spoken dialog systems. In: Proc. HLT/NAACL, pp. 629–637.

Raux, A., Ma, Y., 2011. Efficient probabilistic tracking of user goal and dialog history for spoken dialog systems. In: Proc. Interspeech, pp. 801–804.

Rose, R., Kim, H., 2003. A hybrid barge-in procedure for more reliable turn-taking in human–machine dialog systems. In: Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 198–203.

Skantze, G., Hjalmarsson, A., 2010. Towards incremental speech generation in dialogue systems. In: Proc. SIGDIAL, pp. 1–8.

Sugiura, K., Iwahashi, N., Kawai, H., Nakamura, S., 2011. Situated spoken dialogue with robots using active learning. Adv. Robotics 25 (17), 2207–2232.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., Roy, N., 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In: Proc. AAAI.

Yonetani, R., Kawashima, H., Hirayama, T., Matsuyama, T., 2010. Gaze probing: Event-based estimation of objects being focused on. In: Proc. ICPR, pp. 101–104.