Implicit Human Intention Inference through Gaze Cues for People with Limited Motion Ability

Songpo Li, Xiaoli Zhang IEEE Member
Department of Mechanical Engineering
Colorado School of Mines
Golden, Colorado, 80401, USA
{soli, & xlzhang}@mines.edu

Abstract – The promising assistive technologies bring the hope that enlightens the independent daily living for the elderly and disabled people. However, most modern human-machine communication means is not affordable to those people with very limited motion ability to effectively express their service requests. In the paper, we presented a novel interaction framework which can facilitate the communication between human and assistive devices. In the framework, human intention is inferred implicitly by monitoring the gaze movements. The advantage of this framework is that gaze-based communication requires very little effort from the user and most elderly and disabled people with motion impairment retain the visual capability. The architecture of the presented framework and its effectiveness were introduced and validated. The relationship between human intentions with gaze behaviors was further discussed. This work is expected to simplify the human-machine interaction, consequently enhancing the adoption of assistive technologies and the user’s independence in daily living.

Index terms – human intention, gaze, assistive technology, independent living.

I. INTRODUCTION

Human’s motion ability got impaired and even severely lost due to aging and critical injury or diseases such as spinal traumas, paralysis, stroke, amputation, and Parkinson’s. For those people, it will be difficult or impossible to independently carry out normal activities in daily living, like standing-up, dressing, cooking, eating etc.

Although the promising assistive technologies, such as assistive devices [1] [2], smart home systems [3] [4], bring the hope that enlightens the independent daily living for the elderly and disabled people, one obstacle lying between a human user and assistive machines that may considerably undermine the acceptance of assistive technologies. This obstacle is the lack of a suitable communication means that allows the elderly and disabled to effectively and naturally interact with the assistive system. This problem will become even more severe with the development of more and more assistive systems with higher function capabilities. Nowadays, the user must generate explicit service requests, which mostly involve large motor motions, to control an assistive system. Some existing communication modalities include speech [5] [6] [7], joystick [8], gesture [9] [10] [11] [12], physical contact [13] [14] [15], and electromyography (EMG) [16] [17]. All these interfaces require multiple muscles strong enough to cooperatively work together, which are often unaffordable to the elderly and disabled with limited motion ability. Another promising interaction channel is the brain-signal-based communication interface [18]. However, it is still in its infancy stage. To achieve effective and natural human machine interaction, new techniques, which is easy to learn and affordable for users with very limited motion capability, are needed.

Gaze movements require very little effort from the human, and the capability of controlling the gaze remains in most of the elderly and disabled people with motion impairments, which make the gaze a very promising means to interact with an assistive system for the elderly and disabled people. Gaze tracking is a technique that persistently estimates where a person is looking (gaze) in real time based on monitoring his/her eyes’ movements. In the light of movement monitoring hardware, it can be classified into three categories: contact lens [19], electrooculogram (EOG) [20], and optical method [21] [22] [23] [24]. The optical method, based on a video camera capturing eye images, has gained wide acceptance and usage, due to its non-invasiveness, unobtrusiveness and inexpensiveness. Gaze has been used in psychology/mental studies, Human-Computer Interaction (HCI), and Human-Robot Interaction (HRI). For HCI, gaze is generally used as a pointing device, to select buttons [25] or text through an onscreen keyboard [26] on a computer. For HRI, the employment of gaze remains at indicating the moving direction of a robot for location [27] or object approaching [28].

Implicit intention understanding using subtle gaze cues recently started to gain researchers’ attention. Gaze data are linked to a number of cognitive processes, such as attention. This link to cognitive processes makes gaze a distinct source of information on understanding a person’s intention implicitly. However, using gaze to understand human intention is challenging. This is because the eyes are a perceptual organ not means for motor controlling tasks, such that the usage of gaze should not disturb the user’s natural perception behaviors. Another problem is that the link from gaze to human intention is recessive. The gaze can indicate a person’s visual attention and this visual attention can be used to infer his/her intention on that attention. However, the relation between visual attentions and human intention is ambiguous. For example, when a person looks at a cup, he/she may want to drink water or to wash the cup. The gaze-based intention inference has been investigated to augment human cognition capability by reminding the user the objects related to his/her intention [29], and to promote the communication of disabled people with human helpers through a HCI interface [30].

In this paper, a novel gaze-based interaction framework for human and assistive technologies is introduced. This framework allows the elderly and disabled, with very limited
motion ability, to express their service requests naturally, intuitively, and effortlessly. Meanwhile, using this framework, the assistive systems are endowed with the capability of understanding the user’s intention implicitly by monitoring his/her visual attention. Moreover, this framework will not hinder the user’s normal visual behaviors but allow the user to express various complex intentions. It is expected to simplify the human-machine interaction, consequently enhancing the adoption of assistive technologies and the user’s independence in daily living and the quality of living.

II. GAZE-BASED INTENTION INFERENCE FRAMEWORK

The overall framework of human intention inference is illustrated in Fig. 1. The human intention inference framework is constituted by four components, including attention extraction, object identification, intention knowledge base, and intention inference. In the framework, the user looks at a scene image while the user’s visual attentions are extracted from tracking his/her gaze. The objects that attracts the user’s visual attention then can be extracted. From those attention objects, which the user pays attention to particularly, the intention of the user can be inferred. In this framework, a particular knowledge to describe each human intention to the correlated objects is required. This knowledge can be created from substantial observation and surveys.

A. Attention Extraction

Human’s attention could be extracted from analyzing his/her gaze. Gaze is the estimation of where the human is looking at one instantaneous point of time from his/her eye movements. The raw gaze data is superimposed with noises, which are caused by the small involuntary, jerk-like rolling and drift of eyes. The noise will be filtered first before the extraction of visual attentions.

Generally, a person’s attention is indicated by a cluster of gaze points. The reason for this is that, when a person effectively perceives the information of one location, the gaze roughly maintains still at that location for a while. To confirm the current gaze cluster is the user’s attention, methods including prolonged dwell time, intentional blink, EMG signal, and physical button have been investigated, and in our study the prolonged dwell time is used.

The visual attention in this study is defined as one circle area that can cover all the gaze points in the cluster with the minimal radius. Whereas in most of the existing applications, a visual attention is considered as a single point, without dimensions. The main advantage of using a circle to represent a virtual attention is that it can reduce the mental burden on the user, as it will not require the user intentionally focus his/her gaze on one single point, such that the user can perceive external information naturally. Secondly as all the objects in the environment have dimensions, a circle area is more suitable to indicate an object. Ideally, the attention object should be fully or partially in the circle area.

B. Intention Knowledge Base

From the objects that the user has paid special attention, many information related to the internal intention can be inferred, as the accomplishment of human’s intention significantly depends on the essential objects. It is a very natural behavior that the human looks at the related objects first when he/she has one intention in mind. Such that, by monitoring the objects the user looks at, his/her intention can be inferred. Take drinking water as a simple example, if the user looks at a cup, a bottle of water, we can straightforwardly conclude that the user wants to drink water.

To be able to recognize the intention from the objects, the system has to possess a certain degree of knowledge about how the intention is related to the objects, or how the objects reflect the intention. This knowledge base is created by experiment observations and surveys about what objects have been involved when the human wants to accomplish various intentions.

C. Intention Inference

To perform the intention inference, one inference engine needs to be created and provided with the intention knowledge. The advanced machine learning algorithms for classification could be used to create the engine. The procedure that provides the inference engine with the prior knowledge is called training of the inference engine. After that, the inference engine is ready to infer human intentions. The objects that a subject gazes upon will be sent to the inference engine, and the result will be the system’s best guess about the user’s intention.

An inference engine based on Multi-Class Support Vector Machine (MCSVM) [31] will be created in our study. MCSVM is a supervised learning algorithm that has been widely used for pattern recognition techniques. The inference engine will be implemented in MATLAB [32] [33].

To correctly interpret human intention from the attention objects is challenging, as there are many uncertainties caused by the user and the environment. For different users, they have different preferences. For example, someone likes using a bowl but others like a cup for milk; for an intention involves five objects, a person might just look at three or four of the five objects. In addition, when a person is searching for one object, his/her attention can be easily distracted by irrelevant things. The inference engine should be provided with a large number of training data for all-scale intention knowledge base that embraces individual differences and preferences and have great tolerance on the distraction.

III. EXPERIMENT

To validate the effectiveness of our gaze-based human intention inference framework, experiments have been performed. Five subjects participated the experiments. Two
experiments were performed. One was to establish the intention knowledge and the other one was to test the intention inference engine.

A. Eye Tracking

The gaze data of the participants in the experiments were tracked by GP3 Eye Tracker. It is a commercialized video-based eye tracking system, which can track the user’s eyes remotely from 0.5-0.8m. GP3 is easy to be set up and after a simple user calibration it can report eye movement data at 60Hz with the accuracy of 0.5° - 1° for the visual angle. Although it may not be suitable for the elderly and disabled in the real settings, GP3 is adequate for validating the effectiveness of using gaze to infer human intention.

B. Build the Knowledge Base

A small-scale knowledge base was created, involving four service-related intentions and 12 manipulatory objects in a virtual kitchen environment. The 12 manipulatory objects include a cup, a coffee pot, a milk box, a medicine container, a tap, a kettle, a microwave oven, spoons, bowls, a box of oatmeal, a box of spaghetti, and a bottle of dishwashing liquid. Not all the objects are related to the tasks. Some of them are only used as distractions. Each object was assigned an identification number and they are summarized in Table 1.

<table>
<thead>
<tr>
<th>Object</th>
<th>Cup</th>
<th>Coffee Pot</th>
<th>Milk Box</th>
<th>Medicine container</th>
<th>Tap</th>
<th>Dishwashing liquid</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>O01</td>
<td>O02</td>
<td>O03</td>
<td>O04</td>
<td>O05</td>
<td>O06</td>
</tr>
<tr>
<td>Object</td>
<td>Spoon</td>
<td>Bowl</td>
<td>Oatmeal box</td>
<td>Spaghetti</td>
<td>Kettle</td>
<td>Microwave oven</td>
</tr>
<tr>
<td>ID</td>
<td>O07</td>
<td>O08</td>
<td>O09</td>
<td>O10</td>
<td>O11</td>
<td>O12</td>
</tr>
</tbody>
</table>

And the four potential intentional tasks are:

T1: “Prepare a cup of coffee”;
T2: “Prepare oatmeal for breakfast”;
T3: “Take medicines”;
T4: “Wash a washable target (it could be a cup, coffee pot, spoon, or bowl).”

In the experiment, the participants sat in front of a monitor and looked at the kitchen scene image, shown as the Fig. 2 and Fig. 3. They were asked to express the four intentions in a random order by looking at the objects freely. Their gaze data was recorded simultaneously when they were observing the scene image. The gaze data was analyzed off-line to extract the visual attentions and identify the gazed objects, directly using the location information of each object in the image. After each experiment, the participant was asked to report his/her actual intentions. The reported actual intentions and the corresponding gazed objects were then used to build the knowledge base. They were organized as intention-objects pairs, for example:

$$\{T_i \leftrightarrow o_1, o_2, o_3, \cdots, o_j, \cdots, o_{12}\},$$

where the $$T_i$$ is $$i^{th}$$ intention, and $$o_j$$ indicates whether the object $$j$$ was intentionally viewed by the participant. To collect more information about human intention, one additional question was given to the participants after the experiments. The question was “What object did you come up first in your mind when you had this intention?”

C. Intention Inference

In the second experiment, the experiment setup was the same as the previous one. However this time, the participants freely chose three or four intentions and expressed them using their gaze one by one. The gaze data was also recorded for off-line analysis. After the experiment, their actual intentions were reported and compared with the intentions inferred by the intention engine.

IV. RESULTS

From the first experiment, 19 validated intention-objects pairs were collected, which were used to train the inference engine and here they were called trainData pairs. One example of how the subject looked at the scene image was shown in the Fig. 4 (It was from the second participant). When the participant
was trying to express the intention “Wash the cup”. She looked at the cup, dishwashing liquid and tap consecutively.

The statistical results of the trainData pairs are shown in Fig. 5 A-D, which shows how the participants expressed their intentions by looking at related objects. It clearly shows that different participants looked at different objects when they were trying to convey the same intention, and it varied more significantly in the complex tasks, such as the kettle, spoon, bowel and tap for the intention 1. However, there were still common points shared by different people. For example in the intention 1, coffee were viewed by all the participants.

From the statistical results, we found in each intention several objects appeared with a high frequency, and those objects were called dominant objects of that intention. For example, in the intention 1, “Prepare a cup of coffee”, the dominant objects were coffee pot, cup and milk, which had a strong correlation with the intention 1. The dominant objects for each intention were summarized in Table 2.

<table>
<thead>
<tr>
<th>Intention</th>
<th>Dominant objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intention 1</td>
<td>Coffee pot; Cup; Milk</td>
</tr>
<tr>
<td>Intention 2</td>
<td>Oatmeal; Bowl; Kettle</td>
</tr>
<tr>
<td>Intention 3</td>
<td>Medicine container</td>
</tr>
<tr>
<td>Intention 4</td>
<td>Washable target; Tap; Dishwashing liquid</td>
</tr>
</tbody>
</table>

During the first experiment, we recorded the first object $O_{mind}$ that each subject came up in his/her mind when he/she had an intention. This object was compared with the object $O_{looked}$ that each subject firstly looked at in the kitchen image, as shown in Table 3. It showed that, there were nine sets (out of 19) of $O_{mind}$ and $O_{looked}$ matched, which meant the subjects firstly looked at the object that came up in their mind. Whereas
there were 10 sets did not match. From this comparison, we found that when the subject first had an intention, he/she would try to find the object that came up into his/her mind first. While it was also affected by the environment, as some other related objects might be more outstanding in the scene or closer to the subject’s current visual focus.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mind</td>
<td>O02</td>
<td>O01</td>
<td>O01</td>
<td>O01</td>
</tr>
<tr>
<td></td>
<td>Looked</td>
<td>Sink</td>
<td>O01</td>
<td>O08</td>
<td>O02</td>
</tr>
<tr>
<td>12</td>
<td>Mind</td>
<td>O08</td>
<td>O09</td>
<td>O09</td>
<td>O09</td>
</tr>
<tr>
<td></td>
<td>Looked</td>
<td>O09</td>
<td>O05</td>
<td>O09</td>
<td>O08</td>
</tr>
<tr>
<td>13</td>
<td>Mind</td>
<td>O04</td>
<td>O04</td>
<td>O04</td>
<td>Water</td>
</tr>
<tr>
<td></td>
<td>Looked</td>
<td>O04</td>
<td>O04</td>
<td>O04</td>
<td>O05</td>
</tr>
<tr>
<td>14</td>
<td>Mind</td>
<td>Sink</td>
<td>O06</td>
<td>WT</td>
<td>WT</td>
</tr>
<tr>
<td></td>
<td>Looked</td>
<td>O06</td>
<td>WT</td>
<td>WT</td>
<td>O06</td>
</tr>
</tbody>
</table>

Another comparison was performed between the $O_{\text{mind}}$ and $O_{\text{looked}}$ with the dominant objects of each intention. It was found that, there were 18 (out of 19) $O_{\text{mind}}$ belonged to the dominant objects. The only exception was a sink (not involved in the 12 objects in this study) that subject thought first in the fourth intention, which can be considered equivalent to the dominant object Tap. On the other hand, there were 15 $O_{\text{looked}}$ (out of 19) that belonged to the dominant objects. From this statistical results, it demonstrated that when a person wanted to express his/her intention, the first thing came up in the mind was strongly related to the intention he/she wanted to finish. And for the object that a person firstly looked at might not match his/her original thought, but it was still strongly related to the intention. Such that, in the research and development of the inference engine, a particular weight might be added to the first viewed object, as it had a strong inference to the desired intention.

From the second experiment, 15 intention-objects pairs were collected, which were used to test the trained inference engine and here they were named as testData pairs. The inference engine was trained using the trainData pairs. And 100% correct rate was achieved when the testData was applied. To further examine the robustness of the inference engine, noise was added to the original testData, by randomly adding one distractive object. The correct rate dropped to 93.33%. Then, with the incomplete intention testData, for example one dominant object missing from the intention-objects pair (nothing changed for the third intention, as there was only one dominant object), the correct rate was 86.67%.

From this investigation, it proved that the inference engine based on MCSVM was robust on inferring human intention implicitly based on the gaze, even with some visual distractions in the environment. The accuracy might drop significantly if the user missed one dominant object. To improve the robustness, more training data or modification on the MCSVM may be needed.

V. CONCLUSION AND DISCUSSIONS

In the paper, one novel human interaction inference framework was presented, in which human intention can be intuitively inferred by monitoring what the user has looked and is looking at. It is expected to make significant contributions to increase the living independence and quality of the elderly and disabled people with very limited motor capability. This interaction framework requires very little physical effort from the user which makes it affordable for the elderly and disabled. Moreover, this framework was validated from the simulated experiments. In the experiment, one small-scale intention knowledge was created, and based on these knowledge the participants’ intentions were recognized correctly from the objects they viewed.

In the future work, much effort will be needed to make this intention framework practical in the real living environment. A complete object database is essential to enable the system to recognize all the objects that the user has viewed. Furthermore, a large-scale intention knowledge is needed. With the heavily inflated object database and intention knowledge, the speed and correctness of the recognition may drop significantly. Such that, more investigation on the framework and the algorithm are expected in the future.

References


