Attention Sensing through Multimodal User Modeling in an Augmented Reality Guessing Game

Felix Putze
felix.putze@uni-bremen.de
University of Bremen
Bremen, Bremen

Dennis Küster
dkuester@uni-bremen.de
University of Bremen
Bremen, Bremen

Timo Urban
urbant@uni-bremen.de
University of Bremen
Bremen, Bremen

Alexander Zastrow
azastrow@uni-bremen.de
University of Bremen
Bremen, Bremen

Marvin Kampen
mkampen@uni-bremen.de
University of Bremen
Bremen, Bremen

ABSTRACT

We developed an attention-sensitive system that is capable of playing the children's guessing game "I spy with my litte eye" with a human user. In this game, the user selects an object from a given scene and provides the system with a single-sentence clue about it. For each trial, the system tries to guess the target object. Our approach combines top-down and bottom-up machine learning for object and color detection, automatic speech recognition, natural language processing, a semantic database, eye tracking, and augmented reality. Our evaluation demonstrates performance significantly above chance level, and results for most of the individual machine learning components are encouraging. Participants reported very high levels of satisfaction and curiosity about the system. The collected data shows that our guessing game generates a complex and rich data set. We discuss the capabilities and challenges of our system and its components with respect to multimodal attention sensing.

CCS CONCEPTS

• **Human-centered computing** → **Usability testing**; **Mixed / augmented reality**; *Natural language interfaces*; *Empirical studies in HCI*.

KEYWORDS

attention, Augmented Reality, top-down and bottom-up modeling, gamification

ACM Reference Format:
Felix Putze, Dennis Küster, Timo Urban, Alexander Zastrow, and Marvin Kampen. 2020. Attention Sensing through Multimodal User Modeling in an Augmented Reality Guessing Game. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3382507.3418865

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '20, October 25–29, 2020, Virtual event, Netherlands
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418865

1 INTRODUCTION

Attention modelling is an important research field. In smart-homes, robotics, in-car speech assistants, or in augmented/ virtual reality, systems need to process information from multiple sources to determine the center of the user's attention. Especially in human-robot interaction, it is important to establish a joint attention, e.g., for the robot to know when a person can be interrupted, to know which item in the scene a person is referring to, or to anticipate objects or associated tasks for which to robot can provide assistance. Such anticipatory and context-dependent behavior will lead to more natural and efficient interactions.

However, single cues tend to be too ambiguous to determine a person's attention target. For example, gaze detection does not work if objects are too close to the user, and semantic information of the user's voice input fails if there are multiple objects of the same kind in the scene. To understand the whole scene, at which the user is looking, we combine multiple cues about their attention and use artificial intelligence to evaluate these cues.

To generate rich data on attention tracking, we study the game "I Spy with my little eye", which follows these rules: The player looks at a scene with different objects (see Fig. 2), chooses one of these objects secretly and tells the system a clue sentence about it. This sentence starts with the phrase "I spy", "I see" or "I spy with my little eye". For example: "I spy something, that is used to eat dinner with".

Our system analyzes which object the player gazed at via an eye tracker and detects which objects are present in the scene. In addition, the clue sentence of the player is processed by automatic speech recognition (ASR) and natural language processing to extract a semantic representation of the clue. The result is forwarded to a semantic database to find connections between the objects of the scene and the clue sentence. The Merger component then combines all information to predict which object was chosen.

2 RELATED WORK

In human interaction, gaze-based biometrics such as fixations have recently received substantial attention for the detection of hidden knowledge during concealed recognition. For example, [4] examined scan paths and fixations towards individuals in criminal lineups. Even simply observing others' shared attention has been shown to modulate gaze following, as demonstrated by faster

responses to objects appearing at a cued location [3]. Likewise, establishing shared attention between a technical system and a human has been studied in broad range of areas. Examples include interactions with artificial agents [18], robot-initiated handovers of objects to humans [17], the design of educational HRI [13], as well as more casual interactions like playing a simple game of whack-a-mole [6]. Importantly, eye-tracking allows the study of shared attention with respect to specific objects [7], thus paving the way for further insights into underlying mechanisms. For example, as shown by [22], gaze and descriptive language can be used to establish joint attention directed at one of multiple objects on a table.

In contrast to efforts at deception detection in human interaction, most scenarios in HRI still assume a highly co-operative user who helps the robot identify the shared attention. For example, modules have been developed for automatic directed gaze, pointing gestures, question-answer responses, and backchanneling of brief verbal responses (e.g., "uh huh") to generate engagement [10]. However, the initial novelty of the interaction experience could also wear off rather quickly [14]. Therefore, systems aiming to maintain joined attention and user engagement over many repetitions of a game or task should not simply take shared attention for granted. Instead, the system has to infer attention and engagement from a number of ambiguous cues in behavior, language, and environment of the observed human.

We argue that intelligent attention tracking may help to reap the benefits of shared attention for a large number of novel applications. For example, the use of eye tracking for the detection of attentional shifts may help to better understand and support the learning process in serious games and multimodal virtual reality [9]. Another area of interest are Augmented Reality interfaces. Here, the interaction between user and the system is more implicit and the AR device needs to establish joint attention for the appropriate display of virtual content from ambiguous cues in the user's behavior. Vortmann et al. [24] investigated how EEG and eye tracking can be used to discriminate internal and external attention in an AR setting and how this information can be used to create an attention-adaptive interface [25].

The novel contributions of our work are: 1) We investigate a combination of ambiguous bottom-up and top-down information sources to determine attention of a *non-cooperative* person, 2) we only rely on data from "inside-out" sensing components worn by the person, 3) we collect data on attention in a gaming context that generates rich, complex data, and 4) we show how the model can be integrated in an end-to-end augmented reality interface that leads to a engaging player experience.

3 SYSTEM OVERVIEW

The system aims to guess the correct objects when playing "I spy with my little eye". All inputs and outputs are managed through an Augmented Reality (AR) headset worn by the player. For each turn, the system processes a spoken clue sentence about the object, eye-tracking based fixations, and images recorded via the eye tracker's scene camera. Gaze patterns are recorded while the player selects the object, and analyses of fixations are combined with object detection performed on the image data. The clue sentence is transcribed through automatic speech recognition (ASR) and natural language processing (NLP), and then sent to a semantic database (ConceptNet [21]) to examine semantic links to the detected objects. The player's input is further combined with object color detection. Finally, the merger component combines all information sources to generate a ranked list of the objects indicating the likelihood of each object being the target. The result is presented to the player as a projection in the scene through the Augmented Reality interface.

Based on these steps, we have organized the system into a set of loosely coupled modules as seen in Figure 1. For this purpose, all components are connected through the LabStreamingLayer middleware (LSL)[1], which also takes care of the temporal synchronization and recording for later playback of sessions. In the following, we will describe all components in detail.

Visual Processing: The visual processing component has two major tasks. First, the objects must be detected from the camera image of the AR headset to identify the type and location of potential target objects. Second, we need to identify visual attributes of the objects which are referenced through the given clues.

Changing view angles, light and shadows or over- and underexposure are common problems in image processing. While the employed models for object and color detection provide some invariance to such effects, a pre-processing of the images from the AR camera was beneficial to the performance. For example, the images as provided by the eye tracker were often too dark for color detection. We therefore employed a two step normalization algorithm: Step one normalized the histogram to adapt the brightness and contrast automatically. To enhance the contrast further, 1% of the histogram ends were cut off. In step two, a light bilateral filter was used to smooth the image and reduce noise while minimizing a blurring of the edges.

Our object detection was based on the YOLOv3 [19] neural network model, which was pretrained on the MS COCO data set [16], with a Non-Maximal Suppression (nms) of 0.45. After testing different thresholds, we chose a minimum threshold of 0.10 to ignore objects detected with a confidence below 0.10. On a pilot data set, this value provided the highest average recall rate (0.89; precision = 0.71).

As objects were sampled at a rate of 1 Hz by the AR camera, we tracked and identified the same object across several subsequent images. While many object tracking frameworks could, in principle, be used for this purpose (e.g., Simple Object Realtime Tracking; [2]), we employed a custom tracking approach based on the calculation of the euclidean distance between the centers of the object bounding boxes (i.e., an approach suitable for low sampling rates). Furthermore, we considered the detected object class and compared the bounding boxes per object class.

As objects are often described by naming individual attributes, the system must be able to infer such attributes from the scene. In our current implementation, we do this for the example of object color, which is traditionally one of the most important clue types in the game.

Before applying the color detection, we first remove the background from the detected objects using GrabCut [20]. For each object, this algorithm extracts the image foreground via a bounding

[1] https://github.com/sccn/labstreaminglayer

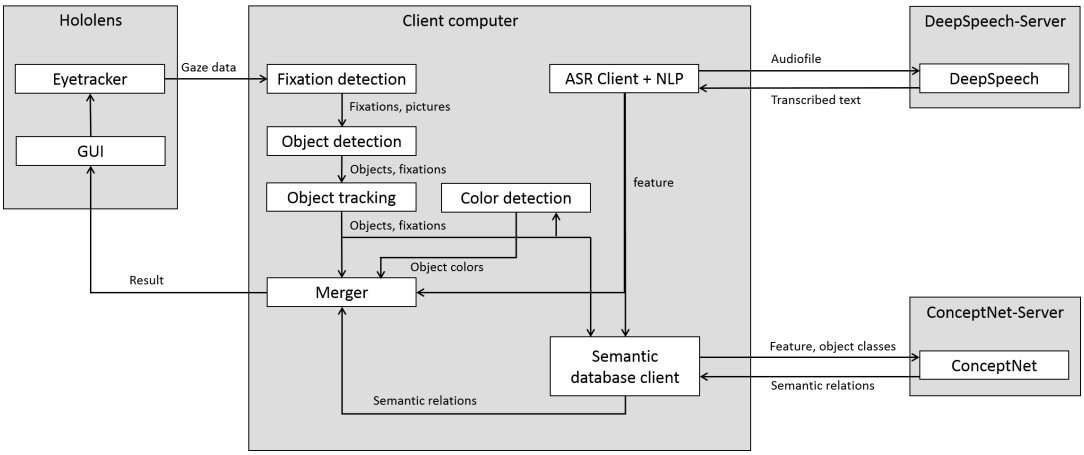

Figure 1: The system architecture of our attention tracking.

box, and interprets the remainder of the image as the background. Next, a voting system is used to achieve higher resilience. Every object is divided into 100 cells (10x10 grid), with k-means clustering applied to each cell. For each cell, we search for 3 centroids (=clusters), each representing an RGB vector.

After finding a color value for each cell, it is compared to the colors defined in Table 1 using the CIE ΔE 2000 color space [12] (other color spaces such as L*A*B can be used). The overall color is then determined as the closest match, with further matches ranked in descending order. For the further analysis, we consider the best two matches to achieve a higher confidence, as well as to process objects with two dominant colors.

Table 1: Defined color space as RGB vectors.

color	value	color	value
black	(0, 0, 0)	pink	(255, 192, 203)
blue	(0, 0, 255)	purple	(128, 0, 128)
brown	(165, 42, 42)	red	(255, 0, 0)
grey	(143, 142, 141)	white	(186, 182, 183)
green	(127, 134, 3)	yellow	(248, 177, 2)
orange	(254, 152, 3)		

Fixation detection: We employed a Pupil Labs eye-tracker to capture fixations on individual object. For a given fixation, the eye hardly moves, and players can usually be expected to fixate on the chosen object at least once. The detection of fixations via eye-tracking can therefore help to identify the focus of a player's visual attention during a trial, and contributes an important clue to the decision making of the system.

We implemented a Random Forest Classifier for fixation detection based on Zemblys et al. [27]. For each frame, the player's fixation view point is classified based on twelve features. All features are computed in a 100 ms rolling window, centered on the frame. Key features include the root mean square, and the standard deviation of the gaze position. Further, the bivariate contour ellipse area, dispersion, velocity, acceleration, and the differences between the sample and median, mean, root mean square, standard deviation and bivariate contour ellipse area are used. Our classifier was trained on the MPIIEgoFixation dataset (five subjects, ca. 9000 frames) provided by [23]. On the test set, our classifier achieved an accuracy of 93.79%. We excluded fixations shorter than 200 ms.

Speech Processing: Being able to speak clues aloud helps to create a similar playing experience as the original game. Our system used automatic speech recognition to transcribe the spoken clues, based on the open source solution DeepSpeech [5]. We recorded 16 kHz, mono-channel WAVE audio files, and used the provided pre-trained acoustic and language models for the English language. Next, we extracted the semantically meaningful clue terms to query the semantic database. Towards this aim, we converted all clue-terms into lowercase and removed punctuation [8], followed by the SpaCy NLP framework [11] to tokenize the input. We further used the part-of-speech tags (PoS) of the tokens, and the dependencies of each token, to identify semantically relevant tokens. We excluded tokens from a custom stoplist.

Semantic Database: The idea behind the semantic database (semDB) component is the assumption that the search term given by the user is semantically connected to the target item. The primary goal of the semDB is therefore an estimation of the semantic connectedness between the clue term and the scene objects. To achieve this, we used ConceptNet as an underling source of semantic relationships. *"ConceptNet is a knowledge graph that connects words and phrases of natural language (terms) with labeled, weighted edges (assertions)."* [21]. In this graph, connectedness can be defined as the shortest path length between the clue term and a scene object. This distance measure thus reflects how well the clue term is semantically connected to a given object.

The clue term can have two kinds of structures: First, it can be a single word whereby the class of word does not matter, or secondly a compound of two or more words, what enables a more accurate description of the searched item, e.g., "drinking wine" instead of "drinking" as a search term for a wine glass. However, such compound word search terms only work if they exist as a concept in ConceptNet. The main limitation of the semDB is that the search term, in contrast to the object instance and its context,

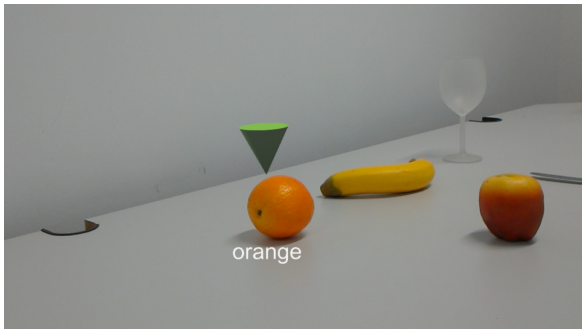

Figure 2: This is the view of a HoloLens user. The orange was detected and a marker put above it in 3D space.

has to be typical for the object class of the searched object. Search terms related to the color, condition, placement, structure, or form of an individual object that can be inferred via common sense as a possible instance, but that are not highly typical for the object class as such (e.g., a green banana), will generally lead to poor results [15].

Augmented Reality User Interface: We used a Microsoft HoloLens to create an immersive and intuitive interface and to display the guesses of the system. The device registers the player's position in relation to their environment, and places a 3D marker together with a describing text box near the identified target object. Figure 2 shows the interface from the player's perspective.

Merger: The merger is the final component required to generate an answer. For this, it calculated a ranked list for every object detected during the trial. We included the following features: The distance between object and fixation, the duration of the fixation, the score from the semantic database, and the detected color. Each feature yielded a rank-score for every detected object.

Features were ranked by the merger as follows. First, distance rankings were calculated as the inverse of the Euclidean distance between the center of the object's bounding box, and the center of the fixation's bounding box. We normalized (0-1) all distance rankings of the objects in one trial. Second, we computed the SemDB scores as the inverse of a clue-term's distance within the semantic network - which we then normalized (0-1). Finally, color was taken into account, if the color flagged by the object detection matched the label indicated by the NLP. If the color of an object matched the NLP-label, a fixed value was added to the final ranking. The final ranking was then computed as the sum of these three scores.

4 EVALUATION

We recruited 20 participants (2 female, 18 male) to evaluate the system performance and game experience. All participants were recruited from the local university, provided written consent, and were reimbursed with 10€. All data collected during that study is available to the research community at https://osf.io/7n3s9/.

Procedure: On arrival, participants were introduced to the game, and donned the HoloLens. The eye tracker was calibrated, and participants performed a short training segment to understand the user interface. The task consisted of a kitchen table scene with 10 objects representing 8 different object classes. To ensure that all items were visible without head movements, players stood in front of the table to view the entire scene at a comfortable angle. They then played the guessing game for all 10 items, in a self-determined order.

For each trial, participants stood at the same (marked) location to choose a new object, and spoke a single-sentence clue. Some objects, such as apples or spoons, were present twice in the scene to ensure that target objects could not be identified solely via the semantic analysis. Instead, guesses made by the system incorporated also spatial- and color-detection performed on the objects. Item positions were varied between participants to ensure that recognition difficulties did not depend on a fixed item location. Participants were instructed to provide a useful, but potentially ambiguous, hint on the selected item. To ensure variability of clue-types beyond simple colors, we provided participants with a number of potential examples. We aimed to set reasonable expectations, but also to show the spectrum of possible semantic clues, including references to color, material, or function of the object. Further, as the ASR processing of the spoken clues yielded some errors (non-native English speakers), we included a manual fallback for entering the clue-sentence by the experimenter. To provide direct feedback to players, the most likely target item (i.e., the system's guess) was marked through the AR display. The experiment continued to the next trial if the guess was confirmed as correct by the participant, or after three guesses. After all 10 trials were completed, participants filled out the subjective evaluation questionnaire, and the experiment was concluded. In total, the data collection lasted 16 minutes on average, with a duration of ca. 25 minutes.

Overall Performance: Figure 3 shows the overall system performance compared to the baseline (random guessing). We evaluated performance after the initial guess, the second guess, and the final guess - and compared that to probabilities from a hypergeometric cumulative distribution function (sampling without replacement) with 1, 2, or 3 draws, respectively. Results show that the system significantly outperformed the baseline for all three conditions (all ps < 0.001). Accuracy of the first guess correlated significantly ($r = -0.67$, $p = 0.03$) with the position in the sequence, i.e., participants tended to choose the easier objects (e.g., easier to describe, more isolated) first. On average, the spoon was chosen early ($M_{\text{position}} = 3.7$), while the milky glass was chosen relatively late ($M_{\text{position}} = 7.0$).

Regarding the impact of the individual feature types, the eye tracking (and object recognition) module provided the most accurate information in 52% of all cases, the language input (and semantic database) provided the most accurate information in 25% of all cases, and both are on par in the remainder of all cases. Thus, both types of information need to be combined for an optimal result. We argue that multimodal data will be important for the development of more believable, robust, and ecologically valid interaction systems. We further expect that the relative contribution of individual components will change for different play behaviors. In the following, we look in more detail at the results for the individual components.

Object Tracking: Object detection and tracking were critical system components because only objects detected during a trial could become possible system guesses. Figure 4 shows the average classification accuracy by object type. The overall object detection

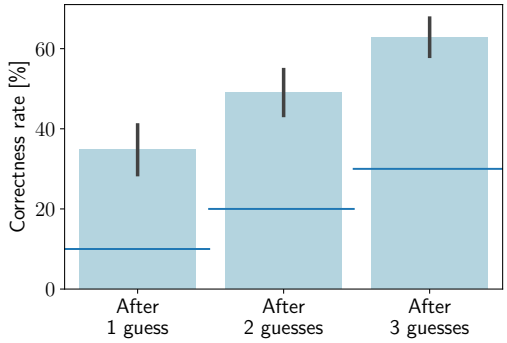

Figure 3: Overall system performance, measured as classification accuracy, for first, second, or third guess. Whiskers indicate the standard deviation (*SD*). Horizontal lines indicate the expected accuracy for random guessing.

rate was 95.5% on average, and above 80% for the most difficult target. Furthermore, the system was not artificially restricted to the objects on the table, but could also detect other objects in the periphery of the scene. I.e., the testing environment was designed to include also a number of potential distractor items as they might occur in a natural environment "at home". Among the detected objects, the most frequent ones were “pottedplant” (189 trials), “chair” (71), “handbag” (12), which were all indeed present in the scene at some point. In total, 14 items that were not part of the experiment composition were detected at least once.

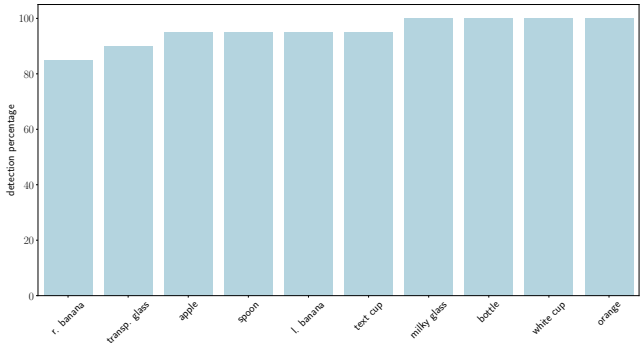

Figure 4: Detection rate for each object in the scene.

Color Detection: As color is the most iconic clue to give in the game of “I Spy”, it also played an important role in the performance of the system. To measure color detection performance, we analyzed the 56 trials that included a color as part of the user input. We then treated each of these colors as a potential label for the associated object class. In some cases, this led to multiple color labels derived for one object type. Specifically, the labels were green (11x) (Apple); yellow (19x) and white (1x) (Banana); orange (9x) (Orange); white (13x) and black (2x) (Cup); white (1x) (Wine Glass). For the other object classes, no color clues were used. With these labels, we evaluated the detected colors of all objects in all trials, including the ones in which no color was mentioned.

Figure 5 shows the color detection accuracy for the five relevant object categories. As the system took the two most prevalent colors into account, we report performance for the top-1 and top-2 results separately. Overall, the first (i.e., highest ranked) color was a correct color label in 61% of all cases, and either the first or the second color were among the valid color labels in 91% of cases. The wine glass appeared to be the most difficult object to detect its color. This might be due to the blue table cloth showing through, resulting in a "milky" rather than truly "white" appearance for participants as well as for the system.

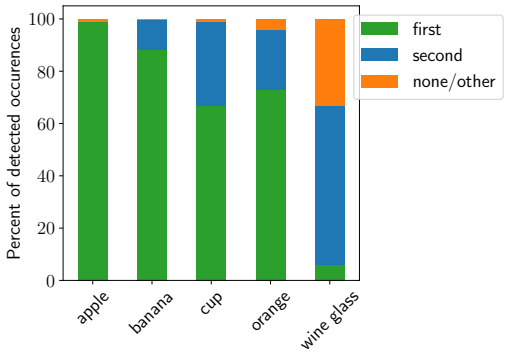

Figure 5: Color detection performance. First: Percentage of color-clues correctly detected as most frequent; second: Percentage of color-clues flagged as the next most frequent (i.e., second) color; none/other: Percentage of no color, or another (wrong) color, being falsely detected as most frequent.

Eye Tracking: To evaluate the performance of the eye tracking component, we considered all detected fixations in a trial. As the decision making component of the system takes the closeness of the fixation center to the object into account, we differentiated between three different levels of precision: “On object”, if the target object was closest to a fixation, “near object” if the target object was the second or third closest to a fixation, or “off-target”. Overall, we find that in 76.4% of all trials, a fixation on or close to the target object was detected. In the remaining 23.6% of the trials, no relevant fixation was found. In some cases, this lack of a fixation could be due to a mis-calibration of the eye tracker. However, we assume that a large proportion of these “error” trials may instead have been due to the participant not having looked at the target object for a sufficient amount of time. Such trials may reflect a typical disguise strategy in the game of “I Spy”, that is frequently used by players to confuse the other (human) players. An indicator supporting this conclusion is that for only a single participant, more than 50% of all trials missed a target fixation; a erroneous calibration would have let to a stronger concentration of misses for the affected participants. Misses are also distributed evenly across time for all participants, further supporting this hypothesis.

Language Processing & Semantic Database: As a final component, we evaluated the language processing and semantic database retrieval. Participants spoke their input naturally, but none of the participants were native English speakers. Further, they used short phrases from a large potential set of clues, leading to challenging conditions for speech recognition. The word error rate (WER)

of the speech recognition was 58% on average, with high variability between speakers, ranging from a maximum WER of 100% to a minimum of 20%. Therefore, the clue phrase was transcribed manually to examine system performance when controlling for this type of error.

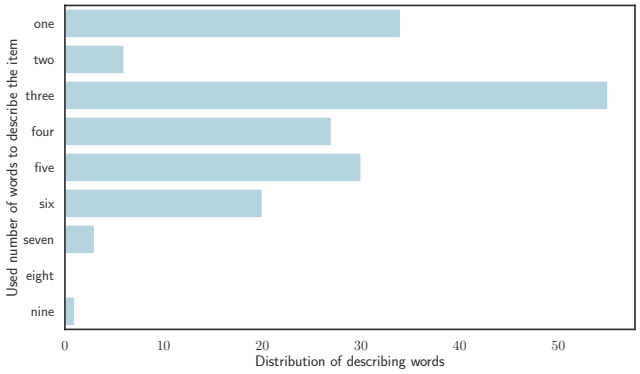

Figure 6: Number of unique clue terms employed to describe the different objects.

To study the richness of the used clues, we investigate the clue term lengths and find that while the most frequent clue length is three, more than 25% of clues have five or more words, up to a maximum of nine. Our results suggest that participants did not restrict themselves to the shortest possible phrase constructs, but frequently explored more complex phrases to describe objects.

Table 2: Categories of clue terms used by participants.

Category	Example	Frequency
function	*used for drinking*	26%
visual (w/o color)	*transparent*	18%
color	*yellow*	15%
shape	*round*	13%
part	*with signs on it*	13%
material	*made out of glass*	6%
other	origin, type, taste	9%

Figure 6 further demonstrates that participants used a large variety of unique clue phrases to describe the objects. The milky glass yielded the most varied descriptions, with 75% of all participants using unique clue words not used by anyone else. The used clue terms cover a range of description categories, showing the broad range of used descriptors. Table 2 summarizes the categories and there frequencies.

Together, these analyses show that the descriptions provided by participants, even in this restricted experimental scenario, were highly flexible and of non-trivial complexity. Thus, the game may be an interesting way to collect large amounts of creative item descriptions. Furthermore, such a gamified approach covers a wider range of relevant semantic relationships with the target object than what might be typically be achieved using conventional crowd sourcing methods (e.g., via Amazon Mechanical Turk). On the downside, however, this of course also means that some of the more creative clue phrases are not part of the semantic database utilized in the present study.

Next, we evaluated the semantic analysis of the spoken clues. For this purpose, we mimicked the way that the merger component took the distance between input term and each object into account when identifying the target object. Specifically, we differentiated between three different categories of outcomes: If an edge existed in semDB between the searched object and input word, we considered this as a distance of 1. If a connection between searched object and input word existed over two edges, we considered this as a distance of 2. If no connection existed, or if the minimum path length was greater than 2, we treated this as a distance of 3.

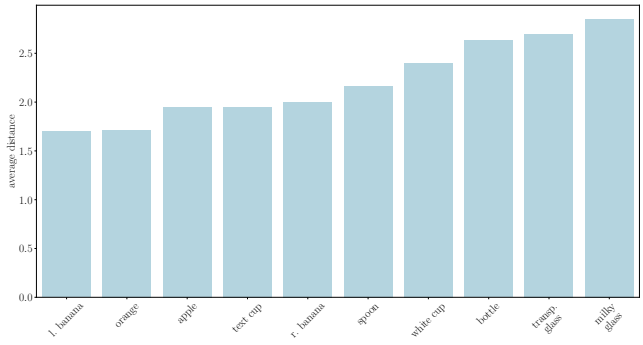

Figure 7: Average distance of semantic database for each object over all 200 trials.

Figure 7 shows the average distances (across all sessions) for the different objects. The results suggest that the retrieval worked well for some of the items, e.g., "banana" and "orange", for which clue terms like "round" or "yellow" were used successfully. For other items, the retrieval in the semantic database was less successful, resulting in higher average differences. For example, many participants used terms like "plastic" and "recyclable" as clue terms for the bottle; however, the semantic database only supports canonical information about item categories, and these generally do not refer to specific material properties of the presented item. This shows that future versions of the system could benefit from more components like the color detection that allow the identification of object attributes (such as form, material, etc.) independently from the semantic database. Conversely, a semantically more flexible database might be better equipped to extract also more loosely associated properties of objects.

Subjective Evaluation: To evaluate the overall experience of "I Spy" as a game, participants answered the 10-item curiosity and enjoyment questionnaire by indicating their agreement in the form of 7-point Likert scale items (1 "not at all" to 7 "very much"). The questionnaire was adapted from the curiosity questionnaire by [26] to match the experimental setup, as well as 3 additional items to assess overall enjoyment and the wish to play again (last items in Figure 8). Figure 8 summarizes the results. Across all items, the responses show that participants enjoyed the game experience and considered it fun. Furthermore, the session was also considered as engaging and motivating. This suggests that a game of this type

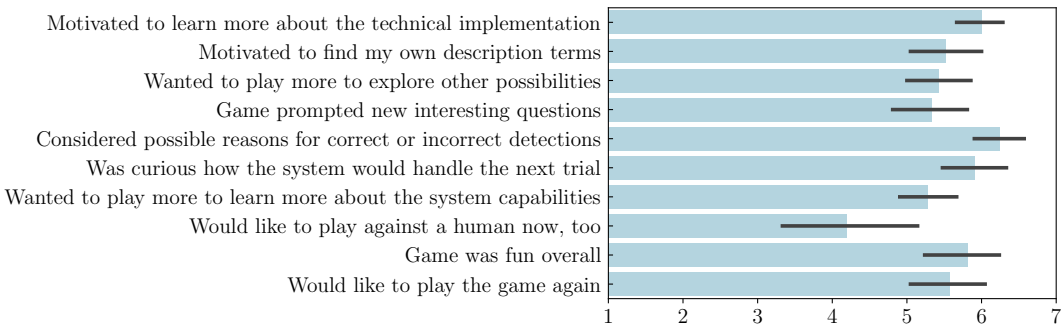

Figure 8: Average agreement to Likert scale items in subjective evaluation.

can be a good way to interest people in new technology, exploring capabilities and limitations of such technology, and as a way to collect data in settings that are intrinsically motivating. The comparably neutral response to the item "I would like to play against a human now, too" stresses that the reported engagement is likely not primarily a result of the familiar game mechanics as such. Instead, we speculate that this could relate to overall curiosity about new technical possibilities, as well as to being curious about how and why such a complex system sometimes gets things right - while at other times it still makes funny mistakes. In this latter sense, playing "I Spy" against our system might be more similar to playing against a novice human player, e.g., a child, than it is to playing against another adult.

5 CONCLUSION

We presented our system for "I Spy" and examined its performance at both overall and component-based levels. Results show that is possible to create an artificial player that reliably outperforms a random baseline and leads to rich data and an engaging gaming experience. Attention sensing is important in human communication. It allows us to detect when someone is ready to listen to us, and can alert us about important events in the environment that we might not yet have detected ourselves. In our case, sensing the other player's focus of attention is a key component of detecting which one of several possible targets a given cue is most likely to refer to. This can contribute to building more believable interactive systems, e.g., social robots. We argue that multimodal data becomes critical whenever any individual channel might fail, or be easily manipulated. I-Spy allows us to investigate deliberately ambiguous cues. In particular, our system addresses the challenge of detecting someone's focus of attention in a relatively rich and naturalistic context. Our approach can help improve a broad range of experiences in HCI and HRI, rendering them more believable and robust. It can be used for multimodal data collection to generate creative examples and increasingly challenging clues. In particular, a gamified approach may come closer to how humans make sense of each other's multimodal signals. Here, we expect a flexible learning of mutually compensatory cues that would be difficult to emulate in a rote labeling schema (e.g., on mTurk).

However, there are still a number of limitations: First, the system currently supports only one clue at a time, i.e., the user cannot help the system after an incorrect guess by providing another clue. Second, our system thus far only performs a rule-based merging. Here, a ranking algorithm based on machine learning might improve results further. With such improvements, it is our hope that future versions of "I Spy" could (1) compete with another human player, and (2) that the analysis of semantic and behavioral cues will help obtain further insights into how humans track each other's attention when one partner is actively trying to be ambiguous.

Another challenge concerns the addition of further object-specific attributes, i.e., non-prototypical properties that are not represented in the semantic database. Currently, color is the only such attribute that can be handled by the system. However, While color was the single most important object-specific clue given by participants, others were frequently used as well, as we showed in our analysis. Following the approach used in Visual Question Answering (VQA) systems, input from the video image and the player's verbal clues could be used to extract further object specific attributes. For example, [1] extracted textual representations of images, combined them with information from an external knowledge base, and generated an answer to a question about the image via an LSTM model. In return, VQA systems can also benefit from the gamified "I Spy" setup, which appears to be a very promising data collection approach that builds on curiosity to learn more about increasingly subtle and ambiguous semantic connections between natural language and physical objects. Additionally, our system goes one step further than VQAs by considering also biological data in the application, as well as in feeding back results to the human in the loop via AR.

ACKNOWLEDGEMENT

The work is partially supported by DFG project "Plasticity of the minimal self in healthy aging – how virtual and real-life changes in sensorimotor experiences shape perception of body ownership, and agency" under Grant No. 402779631.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2425–2433. https://doi.org/10.1109/ICCV.2015.279

[2] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple Online and Realtime Tracking. *CoRR* abs/1602.00763 (2016). arXiv:1602.00763 http://arxiv.org/abs/1602.00763

[3] Anne Böckler, Günther Knoblich, and Natalie Sebanz. 2011. Observing shared attention modulates gaze following. *Cognition* 120, 2 (Aug. 2011), 292–298. https://doi.org/10.1016/j.cognition.2011.05.002

[4] Virginio Cantoni, Mirto Musci, Nahumi Nugrahaningsih, and Marco Porta. 2018. Gaze-based biometrics: An introduction to forensic applications. *Pattern Recognition Letters* 113 (Oct. 2018), 54–57. https://doi.org/10.1016/j.patrec.2016.12.006

[5] Mozilla Corporation. 2019. Project DeepSpeech. Website. https://github.com/mozilla/DeepSpeech.

[6] Lee J. Corrigan, Christina Basedow, Dennis Kuster, Arvid Kappas, Christopher Peters, and Ginevra Castellano. 2015. Perception matters! Engagement in task orientated social robotics. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Kobe, Japan, 375–380. https://doi.org/10.1109/ROMAN.2015.7333665

[7] Lee J. Corrigan, Christopher Peters, Dennis Küster, and Ginevra Castellano. 2016. Engagement Perception and Generation for Social Robots and Virtual Agents. In *Toward Robotic Socially Believable Behaving Systems - Volume I*, Anna Esposito and Lakhmi C. Jain (Eds.). Vol. 105. Springer International Publishing, Cham, 29–51. https://doi.org/10.1007/978-3-319-31056-5_4

[8] Li Deng and Yang Liu. 2018. *Deep Learning in Natural Language Processing* (1st ed.). Springer Publishing Company, Incorporated.

[9] Shujie Deng, Julie A. Kirkby, Jian Chang, and Jian Jun Zhang. 2014. Multimodality with Eye tracking and Haptics: A New Horizon for Serious Games? *International Journal of Serious Games* 1, 4 (Oct. 2014). https://doi.org/10.17083/ijsg.v1i4.24

[10] Aaron Holroyd, Charles Rich, Candace L. Sidner, and Brett Ponsler. 2011. Generating connection events for human-robot collaboration. In *2011 RO-MAN*. IEEE, Atlanta, GA, USA, 241–246. https://doi.org/10.1109/ROMAN.2011.6005245

[11] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* (2017).

[12] Garrett M Johnson and Mark D Fairchild. 2003. A top down description of S-CIELAB and CIEDE2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 28, 6 (2003), 425–435.

[13] Aidan Jones, Dennis Küster, Christina Anne Basedow, Patrıcia Alves-Oliveira, Sofia Serholt, Helen Hastie, Lee J. Corrigan, Wolmet Barendregt, Arvid Kappas, Ana Paiva, and Ginevra Castellano. 2015. Empathic Robotic Tutors for Personalised Learning: A Multidisciplinary Approach. In *Social Robotics (Lecture Notes in Computer Science)*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Springer International Publishing, Cham, 285–295. https://doi.org/10.1007/978-3-319-25554-5_29

[14] Iolanda Leite, Carlos Martinho, Andre Pereira, and Ana Paiva. 2009. As Time goes by: Long-term evaluation of social presence in robotic companions. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, Toyama, Japan, 669–674. https://doi.org/10.1109/ROMAN.2009.5326256

[15] Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. 2016. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *AI*IA 2016 Advances in Artificial Intelligence*, Giovanni Adorni, Stefano Cagnoni, Marco Gori, and Marco Maratea (Eds.). Vol. 10037. Springer International Publishing, Cham, 435–449. https://doi.org/10.1007/978-3-319-49130-1_32 Series Title: Lecture Notes in Computer Science.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[17] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zeng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*. ACM Press, Bielefeld, Germany, 334–341. https://doi.org/10.1145/2559636.2559656

[18] Christopher Peters, Stylianos Asteriadis, and Kostas Karpouzis. 2010. Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces* 3, 1-2 (March 2010), 119–130. https://doi.org/10.1007/s12193-009-0029-1

[19] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[20] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, Vol. 23. ACM, 309–314.

[21] Robert Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI Conference on Artificial Intelligence*. http://arxiv.org/abs/1612.03975

[22] Maria Staudte and Matthew W. Crocker. 2009. Visual attention in spoken human-robot interaction. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 77–84. https://doi.org/10.1145/1514095.1514111 ISSN: 2167-2148.

[23] Yusuke Sugano and Andreas Bulling. 2015. Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 363–372. https://doi.org/10.1145/2807442.2807445

[24] Lisa-Marie Vortmann, Felix Kroll, and Felix Putze. 2019. EEG-Based Classification of Internally- and Externally-Directed Attention in an Augmented Reality Paradigm. *Frontiers in Human Neuroscience* 13 (2019). https://doi.org/10.3389/fnhum.2019.00348

[25] Lisa-Marie Vortmann and Felix Putze. 2020. Attention-Aware Brain Computer Interface to avoid Distractions in Augmented Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, USA.

[26] Pieter Wouters, Herre van Oostendorp, Rudy Boonekamp, and Erik van der Spek. 2011. The role of Game Discourse Analysis and curiosity in creating engaging and effective serious games by implementing a back story and foreshadowing. *Interacting with Computers* 23, 4 (July 2011), 329–336. https://doi.org/10.1016/j.intcom.2011.05.001

[27] Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2018. Using machine learning to detect events in eye-tracking data. 50, 1 (2018), 160–181. http://dx.doi.org/10.3758/s13428-017-0860-3