Multi-Channel Sensory and Cognitive Modeling for Robotic Pet Context Recognition

Ninghan Zhong[1]

Abstract— **Pets are increasingly popular. Studies have shown that human-pet interactions can offer a range of positive physiological, physical, and social benefits. However, Owning a pet might be infeasible for many potential pet owners due to personal situations. In such cases, robotic pets serve as a good potential alternative, offering a comparable pet experience with lower time and financial demands. The foundation of a rewarding pet experience roots in human-pet interactions, and such fulfilling interactions rely on pets' capability of perceiving and recognizing the context of interactions. For instance, a dog might offer solace by staying close with its owner when it detects owner's sadness, or might initiate play behaviors when it senses owner's engagement intention. In this work, a context-aware human-robot-pet interaction model is developed. In particular, multi-modal sensory data, including audio and vision, are used to extract multi-channel context cues, such as the owner's posture, gesture, and audio keywords. A rule-based cognitive model combines the context cues and estimates the appropriate interaction context, such as engagement or following, and the pet robot behaves accordingly. The context detection model is evaluated in a sequence of human-robot-pet interactions and evaluation metrics are reported.**

I. Introduction

Pets are increasingly popular in people's life. Many companion-pet owners develop strong emotional bonds with their pet counterparts [1]. Studies have shown that such emotional bonds between human and pets offer human with a range of positive physiological, physical, and social benefits [2]. Nonetheless, owning a pet could be a time demanding task as many animals require dedicated interaction time with their owners, and can also lead to long-term financial burden. Other factors such as allergies and traveling plans further limit people from owning a desired pet. In response, robotic pets have emerged as a potential alternative, providing a comparable experience to living pets but with fewer demands on the owners' time and resources.

In the past decade, robotics and artificial intelligence has made unprecedented breakthrough, rendering the idea of owning robotic pet increasingly practical. It has been shown that children are able to establish long-term bonding with a robot pet [3]. Onofrio *et al.* [4] also shows that pet robots receive a good acceptance among elder communities and have beneficial effects on mental and physical interactions.

The foundation of a deep bond between humans and pets is rooted in their interactions. The establishment of these rewarding interactions relies on the pets' capability of perceiving and recognizing the context of interactions. For instance, a dog might offer comfort by staying close when it detects sadness in its owner, or initiate playful behaviors by wagging its tail and hopping when it senses the owner's desire to engage in fun activities. In another example, Cooper *et al.* [5] have shown that domestic dogs are likely to avoid a person with a book blocking the person's eyes, because such a status indicates the person might not want to be disturbed. Therefore, integrating context-aware behaviors is a fundamental aspect of designing intelligent robotic pets to provide an authentic pet experience.

Pets' capability of context detection is interleaved with their umwelt. Pets perceive context through a combinations of sensory experiences, including sound signals that contain the owner's vocal tones, and vision signals that capture the owner's gestures or actions. For instance, a dog might be able to sense the tone in from the owner's speech, recognizing whether the owner is pleased or angry towards it. A cat might welcome a gentle stroke on its head when it sees a soft approach by the owner's hand, yet may escape quickly if the owner does the same action abruptly. Similarly, when a dog sees its owner going toward the dog-food cabinet, it probably knows that dinnertime is coming. Following these observations, a multi-channel human-robot-pet interaction context detection model is proposed.

Multi-modal sensory inputs, including audio and vision sensing, are used to extract multi-channel context cues, such as human posture, gesture, and speech. The context detection model deploys a voting scheme to fuse the context cues, estimating the most likely interaction context. The context detection model is incorporated into a larger human-pet-robot interaction system such that the pet robot behaves in accordance with the detected context. The context detection quality is validated based on a sequence of human-robot-pet interactions, showing promising evaluation results. Additional visual illustrations can be found in the supplemental.

The rest of the manuscript is organized as follows. Section II introduces the related works in the areas of robotic pets and context detection. Section III describes the proposed system. Section IV discusses the evaluation and results. Section V concludes the work with future direction highlights.

II. Related Work

A. Robotic Pets

The development of robotic pet started when Sony released the AIBO robot. The robot [7] is designed to be dog-like and has limited basic functions of interactions. Later Petoi developed the Nybble [8] and Bittle [9] pet robots. While palm-sized robotic pets possess high structural variability due to their quadruped architectures and robust locomotion

[1]Ninghan Zhong is with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. n5zhong@uwaterloo.ca

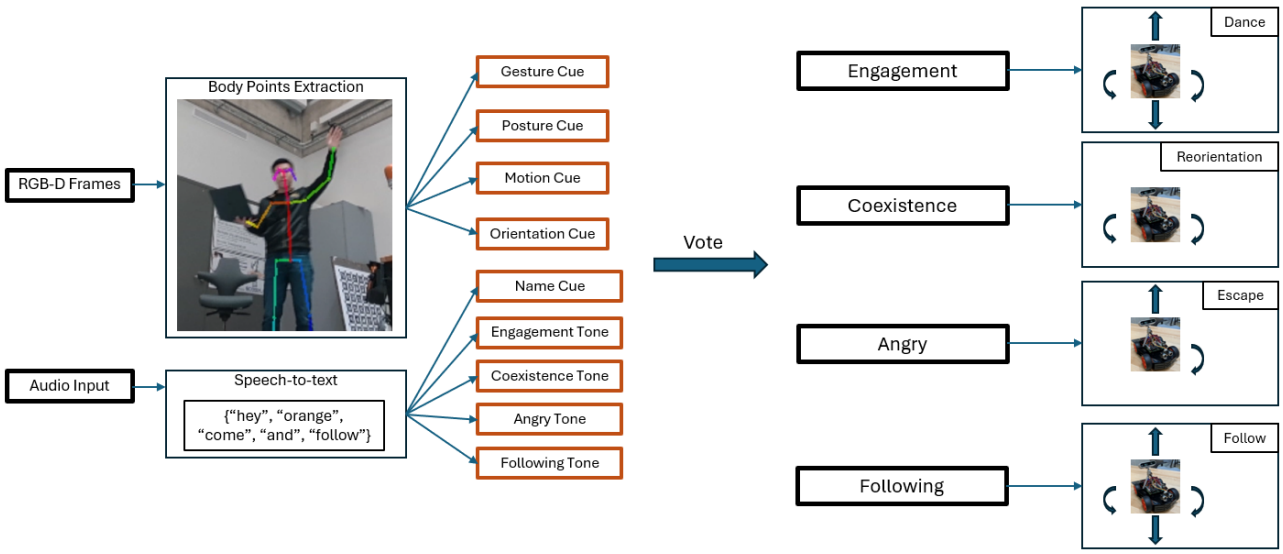

Fig. 1: Diagram of the proposed context-aware human-pet-robot interaction system. The system takes two streams of raw inputs – RGB-D frames from the depth camera, and audio input from the on-board microphone. The RGB-D frames are pre-processed by skeleton tracking [6] and raw audio input undergoes speech-to-text recognition. The processed signals are used to extract context cues, which then cast votes to determine the most likely interaction context. The robot executes a behavior that aligns with the predicted context.

controls, their perturbation bandwidth is limited due to the lack of pre-programmed interaction strategies. Another smart robotic pet is the MarsCat created by the elephant robotics company. MarsCat is the first bionet robotic pet cat and is able to recognize face, sense touch and hear sounds. However, MarsCat has limited flexibility and low moving speed, resulting in a limited structural variability [10]. Paro, the robotic baby seal developed by created by Paro Takanori Shibata is regarded as one of the most well-established robotic pet. While Paro possess a range of sensors such as touch, thermal and light [10], it lacks vision sensing, which is a crucial sensory capability in most pet animals.

More recently, Taleb *et al.* [11] developed CoFiBot V2, a low-price quadruped robot pet with home monitoring capabilities such as fire detection. While CoFiBot V2 has a simple gesture detection combined with some pet-like behavior implementations such as sit and trot gait, the robot pet lacks context detection based on a simulated umwelt of a pet. Gao *et al.* [12] developed a quadruped pet robot OM-C01 with visual learning and few-shot learning capability, and the closest work to this proposal. OM-C01 is capable of visual and auditory interactions, and is able to detect the owner's facial expression. However, visual and auditory signals are used for visual object learning, and only the owner's facial expression is used for generating the robot pet behaviors.

B. Context Aware Human-Robot Interactions

Context prediction is an active research area in human-robot interactions, particularly for assistive and social robots. Engagement detection is a popular approach of context estimation. Zhang *et al.* [13] proposed a learning-based engagement detection model for multi-person scenarios for social robots. In [13], multiple indicative signals from the surrounding persons, such as distance, moving speed, and body orientation are extracted from vision inputs. A CNN-LSTM network is used to fuse the signals and output a probability estimation for engagement intention. Abdelrahman *et al.* [14] takes a multi-modal approach to engagement predictions robotic arms. In [14], neural network models are applied to extract multi-modal engagement cues from humans, such as gaze and face location. A rule-based model is developed to fuse the cues and predict engagement level.

Context estimation can also be formulated as a multi-class intent detection. Wang *et al.* [15] proposed a multi-modal intent that integrates audio and gesture (i.e. vision) channels. The model is mainly designed for understanding elderly individuals' intent, such as grabbing and placing, for assistive robots. The model employs speech-to-text recognition and gesture detection to precisely capture elders' intents. On the other hand, *et al.* [16] formulate context as environment affordances, and proposed a framework for social friendly assistive robot navigation by not interrupting people in an indoor environment. In [16], a multi-layer system is developed to perform event detection, scene detection, and object detection. With such semantic information, an environment affordance map is constructed for social friendly path planning.

The study by Luo *et al.* [17] bears resemblances to the current work. In [17], context is classified into environment and interaction contexts. Environment context emcompasses obstacle avoidance and object recognition, whereas interaction context pertains to the responses to users' emotions and movements. Nonetheless, the work is exclusively developed for VR pets.

In general, multi-channel context detection is extensively studied in assistive robots, companionship robots, and social robots, but limited works investigate context detection in pet robots. Further, to the best of my knowledge, this work

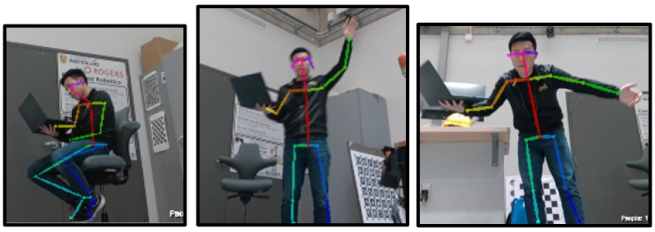

Fig. 2: Gesture context cue

is the first attempt to study multi-channel context detection particularly for pet robots.

III. Context-Aware Interaction Model

Although there exists a wide range of pet animals, this work concentrates on designing a robotic pet dog, given pet dogs' status as one of the most prevalent pets and the comparatively extensive observation of their behaviors. The structure of the proposed multi-channel context recognition model is presented in Fig.1. The model continuously takes as input the RGD-D images viewed by the robot's camera, and the audio signals captured from the robot's mini-microphone.

A. Multi-Modal Data Pre-Prosessing

Raw RGB-D vision inputs are pre-processed by OpenPose [6] real-time skeleton detection with a ROS wrapper[1] to extract useful body points. The body-point information is heavily used in this work, as existing studies [18], [19] have suggested that dogs are able to recognize protruding human body parts and use that information for communication. Thus, using such body-point information could potentially mirror dogs' umwelt.

Raw audio inputs are pre-processed by real-time speech-to-text recognition[2]. Since dogs cannot reason about human language, the semantic meanings of the transcribed text will not be used. However, studies [20], [21] have suggested that dogs can be sensitive to certain keywords conditioned through daily lives. Thus the transcribed text will be used to search for sensitive keywords.

B. Multi-Channel Context Cues Acquisition

1) Visiual Context Cues: The skeleton tracking information from the camera frames is used to detect visual context cues, as shown in the upper four orange boxes in Fig. 1. Specifically, with the extracted body points, context cues from four channels – gesture, posture, motion, and orientation – are estimated using rule-based approaches.

- Gesture: dogs' ability to comprehend human gestures has long been recognized as aiding their understanding of human intentions [22], [23]. Therefore, the use of gesture context cues is a critical component in the proposed model. Gestures are detected using wrist, neck, and eye locations. Possible gestures considered in this work include *wave*, *hug*, and *none*. Specifically, *wave* is detected by measuring the distance of a wrist above both eyes and *hug* is determined by measuring the horizontal distance between the two wrists. If no gesture is detected, the gesture context cue is *none*. Fig. 2 provides a working illustration of gesture context cue detection.
- Posture: posture is detected using hip, knee, and ankle positions. Possible postures considered in this work include *sit*, *stand*, as shown in Fig. 3. In particular, using the hip, knee, and ankle locations, the angles formed by the thighs and thins from both legs are computed. If the angles are lower than a threshold, the person is likely to be sitting or crouching down, and so the posture is *sit*. Otherwise, the posture is *stand*.

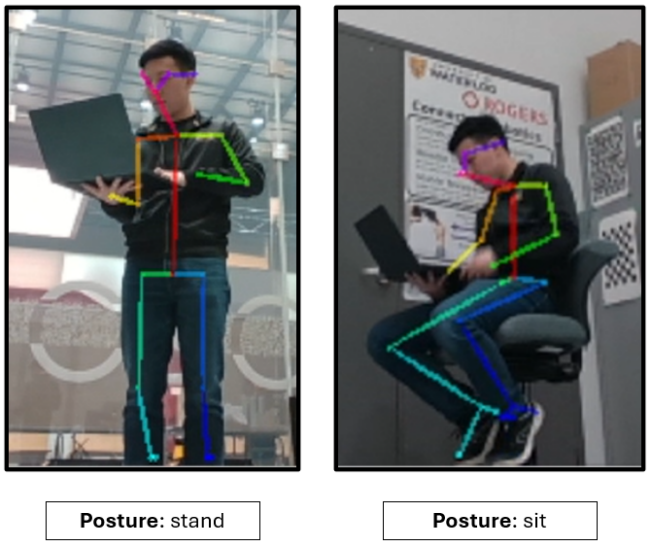

Fig. 3: Posture context cue

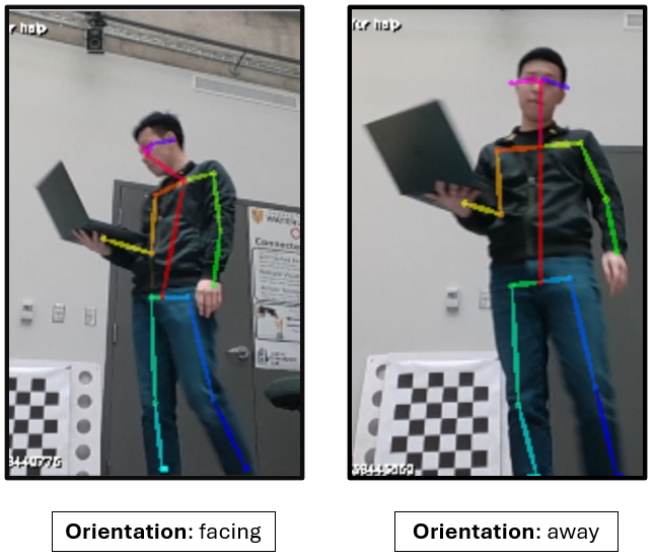

Fig. 4: Orientation context cue

[1] https://github.com/ravijo/ros_openpose
[2] https://github.com/Uberi/speech_recognition

Posture	Engagement
Name mentioned	{"Orange"}
Engagement tone	{"play", "come", "dance"}
Coexistence tone	{"stay", "sit", "shush", "stop"}
Angry tone	{"angry", "out", "away"}
Following tone	{"come", "follow"}

TABLE I: Keyword lists for audio context cues. The robot pet's name is Orange.

- Motion: relative motion of the target human (i.e. the pet owner) is detected using the depth information of the human head point within a sliding window. In this work, motion status includes *approaching*, *leaving*, *none*. If the currently computed distance differs from the previous distance more than a threshold, the motion status becomes *approaching* or *leaving* depending on whether the new distance is closer or further away. If the new distance is only slightly different from the previous distance, the motion status is *none*. To account for sensor noise, both the current and previous distances are computed by taking the averages from sliding windows
- Orientation: studies have shown that dogs are excellent at recognizing humans' attantion [24], [5]. Further, Wynee [18] and other related studies have suggested that dogs are capable of recognizing whether a human is facing or seeing the dog. Thus, detecting the orientation is important in reproducing the dogs' capability in attention cognition. Correspondingly, this work detects orientation status as *facing* and *away*, as shown in Fig. 4. Specifically, the confidence scores of the detected left and right ears are compared. If the confidence scores are comparable, the person is likely to be facing the robot pet. On the other hand, a large difference between the scores suggests the person is facing away from the robot pet.

2) Audio Context Cues: With the transcribed text from audio data, audio context cues from five audio channels – *name mentioned*, *engagement tone*, *coexistence tone*, *angry tone* and *following tone*, are estimated. Each of the audio channels is a binary channel, where true indicates the corresponding audio context cue is found in the person's speech. For instance, *engagement tone* being true indicates that the human's speech contains an engagement tone, while *angry tone* being false means the human's speech does not contain an angry tone. Similarly, *name mentioned* would be true if the pet's name is mentioned in the human speech.

To determine the output for each audio channel, each channel contains a set of keywords (See Table I). The transcribed text is first converted to lowercase letters, and is compared with each list. If one of the keywords from a list is found from the transcribed text, the audio channel of that list outputs true. Otherwise the channel outputs false.

C. Context Detection

The core of this work lies in the proposed multi-channel context detection model (Fig. 1), which fuses the multi-channel context cues and estimates the context.

Gesture	Engagement	Coexistence	Angry	Following
Hug	3	0	0	0
Wave	0	0	0	2
None	0	1	0	0

TABLE II: Voting rules for gesture channel context cues

Posture	Engagement	Coexistence	Angry	Following
Stand	1	0	0	2
Sit	0	2	0	0

TABLE III: Voting rules for posture channel context cues

The set of context in this work is defined as $S = \{\text{'engagement'}, \text{'coexistence'}, \text{'angry'}, \text{'following'}\}$. "Engagement" denotes a scenario where the owner is inclined to participate in enjoyable activities with the pet. "Coexistence" reflects a situation where the owner is preoccupied with personal tasks, such as reading or watching television. "Angry" suggests a condition where the owner is possibly upset with the pet, possibly due to the pet damaging something valuable. "Following" indicates that the pet should accompany the owner, for example, during walks. Downstream motor commands are tailored to align with the predicted context.

The context detection model deterministically outputs the most likely context. Specifically, an efficient voting-based architecture is used, where the context cue from each channel casts one or several votes to the related context, and the context with the maximum votes is determined. For instance, gesture channel would cast three votes for 'Engagement' if the gesture is *hug*, and posture channel would cast two votes for 'Coexistence' if the posture is *sit*. Details about the voting schemes for each context cue channel are presented in tables II, III, IV, V and VI. Note that these voting schemes are adjustable at deployment time to fit specific scenarios.

Further, note that in the real world, multiple interaction contexts could potentially exist at the same time (e.g., 'coexistence' and 'angry'), and we leave such scenarios to future work.

D. Motor Controls

A set of action primitives are first implemented using the control codes provided by the robot platform (see Sec. III-E), including forward(m), backward(m), turnLeft(θ), and turnRight(θ), where m is the distance unit in meter and θ is the angle unit in degree. A motor control scheme is designed for each of the possible contexts. When a context is detected, the corresponding motor control scheme is deployed.

Motion	Engagement	Coexistence	Angry	Following
Leaving	0	1	0	1
Approach	1	0	0	0

TABLE IV: Voting rules for motion channel context cues

Orientation	Engagement	Coexistence	Angry	Following
Facing	0	1	0	0
Away	0	3	0	0

TABLE V: Voting rules for orientation channel context cues

Fig. 5: Diagram showing the Coexistence Control Scheme. In a coexistence context, the pet robot watches its owner saliently without any interaction motions. When the owner is getting out of the pet robot's field of view, the pet robot reorients itself by performing a turn to ensure it can always see its owner.

Audio Cues	Engagement	Coexistence	Angry	Following
Name	1	0	1	1
Engage tone	1	0	0	0
Coexist tone	0	1	0	0
Angry. tone	0	0	1	0
Follow tone	0	0	0	1

TABLE VI: Voting rules for all audio context cues

1) Engagement Control Scheme: In an engagement context, the pet robot tries to engage in fun activities with the owner. It does this by performing a series of actions that represent a dancing behavior. Such a series of actions includes moving forward and backward by a small amount twice, followed by turning left and right by a small amount. Specifically, $\{\text{forward}(0.2), \text{backward}(0.2), \text{forward}(0.2), \text{backward}(0.2), \text{turnLeft}(4), \text{turnRight}(4)\}$. Such a sequence of controls is repeatedly executed until a new context is predicted.

2) Coexistence Control Scheme: A coexistence context implies that the owner does not want to be interrupted, so the pet robot simply silently watches the owner without any other activities. Specifically, the pet robot re-orients itself such that the owner is always within the pet robot's field of view by performing either turnLeft or turnRight. When the owner is approaching the left boundary of the robot's camera field of view, the pet robot first computes the angle between the camera center line and the owner's position and performs a turnLeft by that amount such that the owner is again in the center of the camera frame. The Coexistence control scheme and the reorientation are shown in Fig. 5.

3) Angry: In an angry context, the owner is angry toward the pet robot, so the pet robot will try to escape. This is done by first performing a 180-degree turn and then a forward move by 2 meters. Note that due to the limited time frame, this work does not include obstacle avoidance, so it is assumed that the environment has at least 2 meters of free space for the pet robot to turn and escape.

4) Following Control Scheme: The following context implies that the pet robot should follow the owner. The pet robot follows its owner by repeatedly reorienting (i.e. turnLeft or turnRight) itself such that the owner is at the camera frame center (similar to Coexistence Control Scheme) and moving forward or backward to maintain a short linear distance between the robot and the owner. As the owner moves away, the pet robot will move forward to follow the owner; as the owner moves too close toward the pet robot, the pet robot will move backward to maintain a short distance.

E. Hardware Setup

Fig. 6 shows the hardware setup of the robot pet. SunFounder PiCar-4wd[3] is used as the mobile robot platform for simulating the artificial pet. The mobile robot is capable of performing basic movements such as moving forward, backward, and turning. The onboard computer is a Raspberry Pi Model 4B with 8 GB of RAM. Intel RealSense D435i is mounted on the robot pet to provide RGB-D frames for vision sensing. The robot pet is also equipped with a mini-microphone for audio signals. An ultrasonic range sensor is attached to a servo at the front of the robot pet, currently serving a purely decorative function, though it has potential for obstacle avoidance and range detection applications. The skeleton tracking, audio keyword extraction, and context detection are running on an off-board computer with an Intel Core I7. The robot and the computer communicate wirelessly through Robot Operating System (ROS) Noetic.

At runtime, the context cue acquisition (Sec III-B.1) runs at 10 Hz, context detection (Sec III-C) is performed at 5 Hz, and context-based controls (Sec III-D) are updated at 10 Hz. This study prioritizes the behavior and sensory emulation of a pet animal, leaving the development of the robot pet's physical design beyond the scope of this work.

IV. Pilot Study Evaluation

A. Evaluation Setup

The core of this work lies in the proposed multi-channel context detection model, so the pilot study is centered around context detection quality at deployment time.

The pilot study is conducted at the University of Waterloo RoboHub. In the current version of this work, the speech-to-text implementation for audio context cue cannot target a specific person's voice, hence it has to be assumed that the experimenter is only person speaking in the environment. Unfortunately, at the time of the pilot study, there are merely opportunities where the experimenter is the only person at the RoboHub, rendering evaluating the audio channel infeasible. As a further consequence, the 'angry' context is never detected as in the current formulation, 'angry' context is only detected from speech. As a result, the pilot study focuses on the remaining three context, namely 'Engagement',

[3]https://docs.sunfounder.com/projects/picar-4wd/en/latest/

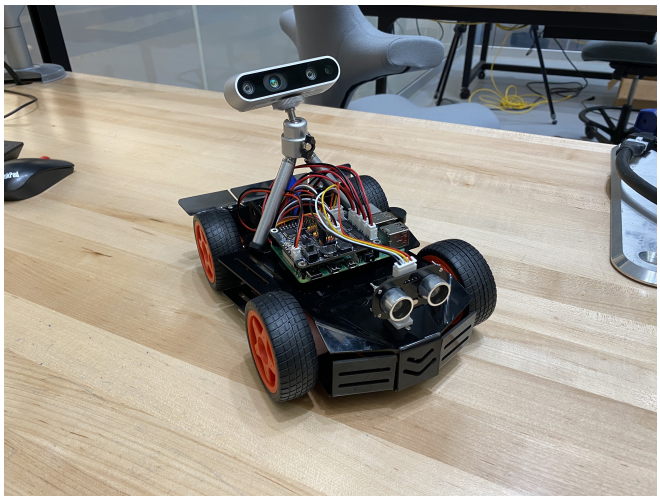

Fig. 6: Robot pet platform with sensors

'Coexistence', and 'Following', and uses the four channels of context cues from the vision data, namely gesture, posture, motion, and orientation.

Nonetheless, a working illustration of the audio channel context cues is included in the supplemental video. The illustration is performed in a quiet room but in a standalone fashion (i.e. with audio data only). The reason is that the onboard Intel RealSense D435i camera is owned by the RoboHub, and could not be removed from the RoboHub premises.

B. Evaluation Procedure

A total of seven sequences of context is randomly generated, serving as the ground truth. Each sequence consists of five context that are randomly generated under the constraint that adjacent context are non-repeating. In each trial, the experimenter interacts with the pet robot by following a sequence of context among the seven sequences, giving a total of seven trials. Each trial of the experiment is video recorded, and the exact timesteps of context changes are marked offline by the experimenter to serve as ground truth. The predicted context per frame is recorded to be compared with the ground truth.

The central premise of this work is that by modeling the the multi-channel sensory and cognitive experiences of a pet animal, a robotic pet's capability to understand context can be improved. To evaluate the effectiveness of using multi-channel context cues, the proposed model is compared against two baseline models. The first baseline, denoted No-Posture (NP), does not leverage the posture channel, and only relies on the gesture, motion, and orientation channels. The second baseline, denoted No-Gesture (NG), does not leverage the gesture channel, and only relies on the posture, motion, and orientation channels. In the evaluation, the proposed model is denoted as Full-Vision (FV).

For each of the three models, the experimenter performs seven trials of experiment using the same seven sequences of context. To analyze the results, for each context under each model, accuracy, precision, recall, and F1 scores are reported.

C. Evaluation Results and Discussion

The overall prediction accuracy is presented in table VII. Tables VIII, IX, and X display the per-context prediction quality of the three evaluated models.

The Full-Vision model provides the highest overall accuracy, while the No-Gesture model performs the worst. This is expected because the Full-Vision model leverages all channels of context cues. Context detection heavily relies on gesture cues, which could potentially explain the worst performance from the No-Gesture model.

Comparing the per-context performances, the Full-Vision model also gives the highest prediction quality and the No-Gesture model performs the worst. The No-Posture model performs better the the No-Gesture model. However, the No-Posture model is still outperformed by the Full-Vision model across all metrics in all context (except Engagement precision) by a smaller margin. This indicates that integrating human posture and gesture detection is effective in improving context detection. The performance margin between Full-Vision and No-Posture being smaller than the performance margin between Full-Vision and No-Gesture potentially indicates that while both gesture and posture are indicative context cues, gesture is a comparatively more reliable feature for context detection. Such observation also aligns with the conclusion from prior studies [22], [23] that dogs are excellent at comprehending human gestures.

Further, note that for both Full-Vision and No-Gesture models, the 'Coexistence' context receives the lowest accuracy. Also, the 'Coexistence' context has a low precision and a high recall in both modals. This observation indicates that both models overestimate 'Coexistence' likelihood during the experiment. In addition, observing that for both models, 'Following' context receives the lowest recall, indicating that both models underestimates 'Following' likelihood. The two observations potentially suggest that the both Full-Vision and No-Posture models sometimes confuses 'Following' with 'Coexistence'. The author hypothesize that this aspect could be potentially improved with additional audio context cues, which are not tested in this pilot study.

Overall, the comparisons indicate that the Full-Vision context detection model offers the best performance. Such result suggests that leveraging context cues from a wider range of channels, or in other words, increasing the perturbation bandwidth in which the the pet robot interact with the environment, could benefit the pet robot's context understanding capability.

	Full-Vision	No-Posture	No-Gesture
Accuracy	0.83	0.77	0.60

TABLE VII: Overall Accuracy for each model

V. Conclusion and Future Work

In this work, a novel multi-channel context-aware human pet robot interaction model is presented. By leveraging

Context	Accuracy	Precision	Recall	F1 Score
Engagement	0.95	0.93	0.81	0.86
Coexistence	0.84	0.68	0.93	0.79
Following	0.88	0.96	0.77	0.86

TABLE VIII: Full-Vision Model context-wise performance

Context	Accuracy	Precision	Recall	F1 Score
Engagement	0.94	0.94	0.77	0.85
Coexistence	0.80	0.66	0.92	0.77
Following	0.83	0.95	0.66	0.78

TABLE IX: No-Posture (NP) Model context-wise performance

multi-modal sensor data, rule-based methods are develop to extract multi-channel context cues, ranging from human posture, gesture, and audio tone. A context detection model is developed to fuse the multi-channel context cues and estimate the most likely interaction context. Downstream motor controls are implemented to align with each predicted context. A pilot study evaluation is performed to evaluate the proposed context detection model, and the result suggests that leveraging a wider range of context cues could be beneficial for pet robot designs.

The current work leaves rooms for improvement on several aspects in future work. First, the current speech detection system is unable to target a specific person's voice, rendering the formal evaluation of audio context cues infeasible during the project time frame. It would be beneficial to examine how much improvement would incorporating the audio context bring to the context detection quality. Further, inability to target a specific person's voice does not align with real pet dogs' audio cognition ability, as dogs are able to recognize their owner's voice. Hence, it would be nice to incorporate speaker recognition in addition to speech recognition.

Second, the current work does not incorporate obstacle detection and avoidance, which is a native and crucial capability for dogs. Extending the current work with obstacle avoidance could not only make the pet robot behaves more similarily to a real dog, but can also ensure the robot's safety.

Third, the current work detects interaction context solely based on the information from the person (i.e. the owner), namely the body points and speech. However, semantic information from the environment also provides rich cues for context prediction. For instance, a person holding a book is less likely to start an interaction than a person holding a Frisbee. From such aspects, scene understanding and object recognition might offer additional improvements.

In conclusion, the current work shows that multi-channel sensory and cognitive modeling could be beneficial for robotic pet context understanding, but further improvements can be made in both cognitive modeling and system control aspects.

Context	Accuracy	Precision	Recall	F1 Score
Engagement	0.75	0.48	0.76	0.59
Coexistence	0.76	0.68	0.92	0.78
Following	0.69	0	0	0

TABLE X: No-Gesture (NG) Model context-wise performance

REFERENCES

[1] P. Martens, M.-J. Enders-Slegers, and J. K. Walker, "The emotional lives of companion animals: attachment and subjective claims by owners of cats and dogs. anthrozoös 29 (1): 73–88," 2016.

[2] E. K. Crawford, N. L. Worsham, and E. R. Swinehart, "Benefits derived from companion animals, and the use of the term "attachment"," *Anthrozoös*, vol. 19, no. 2, pp. 98–112, 2006.

[3] J. M. K. Westlund, H. W. Park, R. Williams, and C. Breazeal, "Measuring young children's long-term relationships with social robots," in *Proceedings of the 17th ACM conference on interaction design and children*, 2018, pp. 207–218.

[4] G. D'Onofrio, L. Fiorini, H. Hoshino, A. Matsumori, Y. Okabe, M. Tsukamoto, R. Limosani, A. Vitanza, F. Greco, A. Greco *et al.*, "Assistive robots for socialization in elderly people: results pertaining to the needs of the users," *Aging clinical and experimental research*, vol. 31, pp. 1313–1329, 2019.

[5] J. J. Cooper, C. Ashton, S. Bishop, R. West, D. S. Mills, and R. J. Young, "Clever hounds: social cognition in the domestic dog (canis familiaris)," *Applied Animal Behaviour Science*, vol. 81, no. 3, pp. 229–244, 2003.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[7] M. Fujita, "On activating human communications with pet-type robot aibo," *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1804–1813, 2004.

[8] E. Ackerman, E. Guizzo, and F. Shi, "Video friday: Open source robotic kitten, and more," Oct 2018. [Online]. Available: https://spectrum.ieee.org/video-friday-open-source-robotic-kitten

[9] ——, "Video friday: Bittle is a palm-sized robot dog now on kickstarter," Sep 2020. [Online]. Available: https://spectrum.ieee.org/video-friday-bittle-robot-dog

[10] S. Taleb, L. C. FOURATI, and M. FOURATI, "Study on communicative robots assisting elderly persons," in *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE, 2023, pp. 209–214.

[11] C. W. Lumoindong and E. Sitompul, "A prototype of an iot-based pet robot with customizable functions (cofibot v2)," *International Journal of Mechanical Engineering and Robotics Research*, vol. 10, no. 9, pp. 510–518, 2021.

[12] F. Gao, C. Lei, X. Long, J. Wang, and P. Song, "Design and development of an intelligent pet-type quadruped robot," in *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2021, pp. 366–371.

[13] Z. Zhang, J. Zheng, and N. M. Thalmann, "Engagement intention estimation in multiparty human-robot interaction," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 117–122.

[14] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, "Multimodal engagement prediction in multiperson human–robot interaction," *IEEE Access*, vol. 10, pp. 61 980–61 991, 2022.

[15] Y. Wang, Z. Feng, and H. Wang, "Multimodal intent understanding and interaction system for elderly-assisted companionship," *CCF Transactions on Pervasive Computing and Interaction*, pp. 1–16, 2023.

[16] P.-T. Wu, C.-A. Yu, S.-H. Chan, M.-L. Chiang, and L.-C. Fu, "Multilayer environmental affordance map for robust indoor localization, event detection and social friendly navigation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2945–2950.

[17] Y. Luo, F. Liu, Y. She, and B. Yang, "A context-aware mobile augmented reality pet interaction model to enhance user experience," *Computer Animation and Virtual Worlds*, vol. 34, no. 1, p. e2123, 2023.

[18] C. D. Wynne, "What is special about dog cognition?" *Current Directions in Psychological Science*, vol. 25, no. 5, pp. 345–350, 2016.

[19] J. Kaminski and M. Nitzschner, "Do dogs get the point? a review of dog–human communication ability," *Learning and Motivation*, vol. 44, no. 4, pp. 294–302, 2013.

[20] D. Mills, "What's in a word? a review of the attributes of a command affecting the performance of pet dogs. anthrozoös 18 (3): 208–221," 2005.

[21] M. Murg, "Why do dogs come when you call their name,"

Mar 2018. [Online]. Available: https://wagwalking.com/behavior/why-do-dogs-come-when-you-call-their-name

[22] B. Agnetta, B. Hare, and M. Tomasello, "Cues to food location that domestic dogs (canis familiaris) of different ages do and do not use," *Animal cognition*, vol. 3, pp. 107–112, 2000.

[23] M. A. Udell, R. F. Giglio, and C. D. Wynne, "Domestic dogs (canis familiaris) use human gestures but not nonhuman tokens to find hidden food." *Journal of comparative psychology*, vol. 122, no. 1, p. 84, 2008.

[24] Z. Virányi, J. Topál, M. Gácsi, Á. Miklósi, and V. Csányi, "Dogs respond appropriately to cues of humans' attentional focus," *Behavioural processes*, vol. 66, no. 2, pp. 161–172, 2004.