Service humanoid robotics: a novel interactive system based on bionic-companionship framework

Jiaji Yang[1], Esyin Chew[1] and Pengcheng Liu[2]

[1] Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, Cardiff, United Kingdom
[2] The Department of Computer Science, University of York, York, United Kingdom

ABSTRACT

At present, industrial robotics focuses more on motion control and vision, whereas humanoid service robotics (HSRs) are increasingly being investigated and researched in the field of speech interaction. The problem and quality of human-robot interaction (HRI) has become a widely debated topic in academia. Especially when HSRs are applied in the hospitality industry, some researchers believe that the current HRI model is not well adapted to the complex social environment. HSRs generally lack the ability to accurately recognize human intentions and understand social scenarios. This study proposes a novel interactive framework suitable for HSRs. The proposed framework is grounded on the novel integration of Trevarthen's (2001) companionship theory and neural image captioning (NIC) generation algorithm. By integrating image-to-natural interactivity generation and communicating with the environment to better interact with the stakeholder, thereby changing from interaction to a bionic-companionship. Compared to previous research a novel interactive system is developed based on the bionic-companionship framework. The humanoid service robot was integrated with the system to conduct preliminary tests. The results show that the interactive system based on the bionic-companionship framework can help the service humanoid robot to effectively respond to changes in the interactive environment, for example give different responses to the same character in different scenes.

INTRODUCTION

Humanoid service robots (HSRs) have seen a sharp rise in adoption recently and are seen as one of the major technologies that will drive the service industries in the next decade (*Harris, Kimson & Schwedel, 2018*). An increasing number of researchers are committed to investigating HSRs to help humans complete repetitive or high-risk service and interactive tasks such as serving patients with infectious diseases, delivering meals and so on. Delivery robots, concierge robots, and chat robots have been increasingly used by travel and hospitality companies (*Ivanov, 2019*). Although the contribution of these achievements mainly comes from the rapid development of robotics engineering, *Ivanov et al. (2019)* indicated that future research focus will gradually shift from robotics

engineering to human-robot interaction (HRI), thus opening up interdisciplinary research directions for researchers.

In the early days, *Fong, Thorpe & Baur (2003)* proposed that in order to make robots perform better, the robot needs to be able to use human skills (perception, cognition, etc.) and benefit from human advice and expertise. This means that robots that rely solely on self-determination have limitations in performing tasks. The authors further propose that the collaborative work between humans and robots will be able to break this constraint, and research on human-robot interaction has begun to emerge. *Fong, Thorpe & Baur (2003)* believe that to build a collaborative control system and complete human-robot interaction, four key problems must be solved. (1) The robot must be able to detect limitations (what can be done and what humans can do), determine whether to seek help, and identify when it needs to be resolved. (2) The robot must be self-reliant and secure. (3) The system must support dialog. That is, robots and humans need to be able to communicate with each other effectively. However, dialog is restricted at present. Through collaborative control, dialog should be two-way and require a richer vocabulary. (4) The system must be adaptive. Although most of the current humanoid service robots already support dialog and can complete simple interactive tasks, as propounded in the research, such dialog in the present time remains limited and "inhuman." In the process of interacting with robots, humans always determine the state of the robot (the position of the robot or the action the robot is doing) through vision, and then communicate with the robot through a dialog system. However, HSR cannot perform this yet as they do not seem to fully satisfy the two-way nature of dialog. Therefore, this research responds to the current gap and attempts to differ from the current HRI research. This research attempts to introduce deep learning into the existing dialog system of HSR, thus advancing the field.

With the continuous development of humanoid robots, more and more humanoid robots are used in the service industry, especially the hospitality industry. Human-Robot Interaction (HRI) has become a hot potato by more and more researchers (*Yang & Chew, 2020*). However, with the deepening of research, some researchers found that when humans interact with humanoid service robots (HSRs), humans hope that HSRs should have the ability and interest to interact with the dynamic thoughts and enthusiasm of the partner's relationship, and can recognize the environment, blended with what others think is meaningful and the emotions to express sympathy (*Yang & Chew, 2020*). This coincides with *Trevarthen*'s *(2001)* companionship theory, so the concept of human robot companion (HRC) was proposed this research. The earlier concept of the robot companion is mentioned by *Dautenhahn et al. (2005)*: HSRs need to have a high degree of awareness and sensitivity to social environment. Through the review of the above literatures, it is proposed to establish an interactive and companion framework for HSRs using deep learning and neural image caption generation, thus advance the current field of HSRs to tackle with bionic-interactive tasks of the service industry and further evolve from conventional HRI to Human and Robot Companion (HRC) (See Table 1).

This study proposes that the introduction of visual data into the current HRI model of HSRs enables HSRs to have a high level of sensitivity to the social dynamic environment while interacting with humans, thereby enhancing the current HRI model to HRC. With

Table 1 Scenario based comparison for HRI and HRC.

Scenario:	HRI	HRC
Scenario 1: Hospitality: **The enhancement from HRI to HRC:** **(1) Compared with HRI, robots in HRC can recognize the environment (luggage) and changes in customers' appearance (red shirt), which is in line with the proposed concept of companion should become dynamic in the theory of companionship.** **(2) More enthusiastic and bionic interaction capabilities (automatically detect whether they are regular customers, and greet enthusiastically).** **(3) The robot in HRC remember the customers' past orders and provide meal recommendation for well-being.**	When you enter a hotel, you see a reception area dominated by robots. When you approach the reception area, the HSR will say "Welcome to Hotel XYZ, please follow the instructions to check-in on my display screen". After completing the check-in, the robot will tell you the room number and issue you a room card, you go to your room, change a suit and prepare to go downstairs to eat. When you go back to the reception, the robot says 'welcome, please follow the instructions to place an order on my display'. You choose a few dishes that look good on the screen of the robot, but when the food comes up you don't seem to be satisfied with the taste…	When you enter a hotel, you can a reception area thoughtfully served by robots. The robots also see you and wave to you, 'Welcome Jack, you have a nice luggage, I can help you to check-in. What else can I do for you?' After completing the check-in, the robot will tell you the room number and issues you a room card. You go to your room and change to a red shirt to go downstairs to eat. When you go back to the reception, the robot says, "Welcome Jack, you wear nice red shirt, what can I do for you?" You choose a few dishes that look good on the robot's screen, but the robot tells you that 'According to your past order and diet preferences, these meals may not be suitable for you. Feel free to change it to a less cholesterol dishes with special house promotion and I recommend you to take this quality wine as a treat to have a healthy eating while enjoying your stay with us.'
Scenario 2: Health care: **(1) The enhancement from HRI to HRC: Robots have dynamic thinking and real-time neural image captioning ability: able to deal with emergencies and a quick decision making from what it sees the environment in real-time.** **(2) Robots has been improved from conventional smart Q&A and interactions to new concept of bionic companionship.**	You bought a robot at home to monitor your health. The robot obtains some of your health indicators (such as temperature, blood pressure, etc.) through some external devices. When there is a problem with your indicators, the robot can give you corresponding suggestions or help you contact a doctor. A total of 1 day you suddenly fainted at home for some reasons, but because you did not aim at the detection device connected to the robot, the robot did not find your condition. Fortunately, your neighbor found you fainted at home. . .	You bought a home care robot to monitor your health. The robot obtains some of your health indicators (such as pulses, blood pressure, etc.) through some external devices. When there is a problem with your indicators, the robot can give you corresponding suggestions or help you contact a doctor. A total of 1 day you suddenly fainted at home for some reason. The robot discovered your real-time condition through the deep learning vision system and contacted the your family member or hospital in time, subject to what the robot sees, *e.g.*, fainted human with lots of blood or motion (call hospital for emergency); fainted human with conscious and free speech (call family members).

the continuous development of deep learning, some researchers have recently realized the transformation of static pictures or videos from conventional camera input into text descriptions (*Li et al., 2020*; *Hu et al., 2020*; *Luo, Hsu & Ye, 2019*). This deep learning algorithm model is called neural image capturing (NIC). This research attempts to adapt and integrate NIC into HSRs and propose a novel framework (bionic-companionship framework) to enhance the traditional HRI experience. This framework aims to improve the current HRI interaction mode in the field of HSRs to a higher level of HRC (*Yang & Chew, 2021*). The bionics in this research refers to the humanoid service robot imitating all the tastes of life, trying to adapt to the seven emotions of ancient human nature (joy, anger, sadness, fear, love, disgust, liking) and six biological wills (life, death, eyes, ears, mouth, nose) (*Chew et al., 2021*). The system proposed in this study combines visual intelligence and Speech Intelligence, and imitates human behavior in social activities, which is in line with the

concept of robot bionics proposed by researchers such as *Chew et al. (2021)*. Therefore, this study believes that the proposed system is a bionic system.

RELATED WORKS

With the continuous development of HRI research, industrial robots have been able to interact with humans accurately and self-adaptively. Some advanced control systems (*Zhang, Hu & Gow, 2020*) and algorithms (*Tang, Zhang & Hu, 2020*) have been proposed as Industrial robots provide reliable support for completing interactive tasks in an industrial environment. However, as HSRs began to enter the service industry, some research cases began to discover that there are still problems with the interaction of HSRs in the social environment. *Caleb-Solly et al. (2018)* believed that users can also help robots when robots help users; meanwhile, users can give feedback to optimize the system. The feedback reflects not only the optimization of the robot system but also the satisfaction of customers. *Chung & Cakmak (2018)* study indicated that hotels in the hospitality industry want to collect customer feedback in real-time to immediately disseminate positive feedback and respond to unsatisfactory customers while they are still on the scene. Guests want to inform their experience without affecting their privacy. Stakeholders in the hospitality industry hope that intelligent robots can interact more with users. Besides, *Rodriguez-Lizundia et al. (2015)* concluded that the optimal distance between users and robots is 69.58 cm. Specifically, interaction with a certain greeting mode can attract users to maintain a longer interaction time; robots with the active search are more attractive to participants. The interaction time is longer than that of passively searching robots, suggesting that robots should be designed to keep at a certain distance from humans and consider adding the ability to allow robots to actively identify customers and attract them.

Research suggests that the current interactive system used by HSRs lacks the ability to process and adapt to dynamic social environments. The dynamic social environment here refers to the same human behavior and language often expressing different meanings in different social situations, such as In different situations, the handshake may require two completely different interactive messages to respond. Therefore, this research proposes the concept of HRC to develop a new interactive mode to solve the current problems faced by HRI in the hospitality industry. For a more detailed comparison of HRI and HRC, please refer to the video in the appendix link (https://youtu.be/fZmV4MKeYtQ).

Review of neural image captioning

The challenge of generating natural language descriptions from visual data has been extensively researched in the field of computer vision. However, early research has mainly focused on generating natural language descriptions from video-type visual data (*Gerber & Nagel, 1996*; *Mitchell et al., 2012*). These systems convert complex visual data into natural languages using rule-based systems. However, because the rules are artificially designed, these systems are sufficiently robust, bionic, and have been shown to be beneficial in limited applications such as traffic scenarios (*Vinyals et al., 2015*). In the past decade, various researchers, inspired by the successful use of sequence-to-sequence

training with neural networks for machine translation, proposed a method for generating image descriptions based on recurrent neural networks (RNNs) (*Cho et al., 2014*; *Sutskever, Vinyals & Le, 2014*). In fact, this method of replacing the encoder in the encoder-decoder framework in machine translation with image features transforms the original complex task of generating image data caption into a simple process of "translating" the image into a sentence (*Cho et al., 2014*). Furthermore, *Donahue et al. (2014)* used long short-term memory (LSTM) for end-to-end large-scale visual learning processes. In addition to images, *Donahue et al. (2014)* also applied LSTM to videos, allowing their models to generate video descriptions. *Vinyals et al. (2015)* and *Kiros, Salahutdinov & Zemel (2014)* initially proposed the structure of a currently popular neural image generation algorithm based on the combination of a convolutional neural network (CNN) image recognition model and a natural language processing (NLP) structured model. Moreover, the neural image captioning algorithm based on the attention mechanism has also attracted extensive attention in the field of computer vision. *Denil et al. (2012)* proposed a real-time target tracking and attention recognition model driven by sight data. *Tang, Srivastava & Salakhutdinov (2014)* proposed an attention-generation model based on deep learning. From the perspective of visual neuroscience, the model requires object-centric data collection for model generation. Subsequently, *Mnih et al. (2014)* proposed a new recurrent neural network model, which can adaptively select specific areas or locations to extract information from images or videos and process the selected area at high resolution. As the algorithm has increasingly mature, the application of the algorithm in related fields has also been breaking through recently, such as the caption generation of car images (*Chen, He & Fan, 2017*), the description generation of facial expressions (*Kuznetsova et al., 2014*), and educational NAO robots driven by image caption generation for video Q&A games for children's education (*Kim et al., 2015*). Recent research on image caption generation also shows that the accuracy and reliability of the technology have increased (*Ding et al., 2019*). In addition, reinforcement learning to automatically correct image caption generation networks have also been proposed (*Fidler, 2017*). These deep learning–based studies have undoubtedly laid a foundation for the possible NIC integration with HSRs as proposed in this study. The novel integration led to the possibility for humanoid robots to interact with humans while recognizing the social environment in real time, thereby improving the interactive service quality of the HSRs.

Neural image caption generation algorithm 'crash into' robot

An increasing number of studies have been conducted on HRI combined with image caption generation algorithm. *Kim et al. (2015)* used the structure of a convolutional neural network (CNN) combined with RNN + deep concept hierarchies (DCH) to design and develop an educational intelligent humanoid robot system for play video games with children. In this study, CNN was used to extract and pre-process cartoons with educational features, and RNN and DCH were used to convert the collected video features into Q&A about cartoons. During the game, after watching the same cartoon, the child and the robot ask and answer questions based on the content of the cartoon. The research results

Yang et al. (2021), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.674

5/20

show that such a system can interact effectively with children. However, for HRIs, such simple and limited-structured Q&A conditions cannot satisfy all the interaction scenarios required. *Cascianelli et al. (2018)* used a gull-gated recurrent unit (GRU) encoder-decoder architecture to develop a human-robot interface that provides interactive services for service robots. This research solves a problem called natural language video description (NLVD). The authors also compared the performance when using LSTM and GRU with two different algorithms to solve these problems. They demonstrated that the GRU algorithm runs faster and consumes less memory. This type of model may be more suitable for HSRs. Although the research model is competitive on public datasets, the experimental results on the designed datasets show that the model suffers from significant overfitting. This proves that in the actual model training process, a specific training dataset for HSR interaction should be established, and other methods (such as transfer learning) should be considered to improve the generalization ability of the model for interactive tasks. *Luo et al. (2019)* created a description template to add various image features collected by the robot, such as face recognition and expression, to the generated description. Compared with the previous models, their interaction is slightly more natural and closer to the human description. However, *Luo et al. (2019)* use the model to provide limited services to industry managers, hard to generalize, and not for developing an entire HRI framework.

Like the research on robot vision language, research on robot vision action is in its infancy. *Yamada et al. (2016)* used RNNs to enable robots to learn commands online from humans and respond with corresponding behaviors. This research furthermore provides a reference and direction for humanoid robots to use deep learning to obtain online learning capabilities for human commands. Inspired by the above study, the rationale and hypothesis proposed in the present research are that the description generated by the neural image captions can drive HSRs to perform appropriate behaviors, and HSRs can even obtain online learning capabilities of interacting with surrounding people through studying and analyzing social environments. *Tremblay et al. (2018)* and *Nguyen et al. (2018)* believe that non-experts often lack the rationality of task descriptions when issuing instructions to robots. They use deep learning to allow robots to automatically generate human-readable instructions' descriptions according to the surrounding social environment. In addition, *Nguyen et al. (2018)* also used visual data to make humanoid robots imitate and learn human actions under corresponding commands so that the robot can learn how to complete the corresponding tasks only through visual data; however, social robots cannot complete precise control of movements when they imitate movements of visual data.

CONTRIBUTION TO THE KNOWLEDGE: THE BIONIC-COMPANIONSHIP FRAMEWORK WITH NIC FOR HSRS

The contribution of the present study is the novel investigation and design of the bionic-companionship framework for HSRs, adapting and integrating neural image caption generation algorithms and bionic humanoid robots, to be validated in a lab-controlled environment and real-life exploration. The new HRC framework is anticipated to enhance HRI to reach a new state, making it possible for HSRs to become bionic companions of humans.
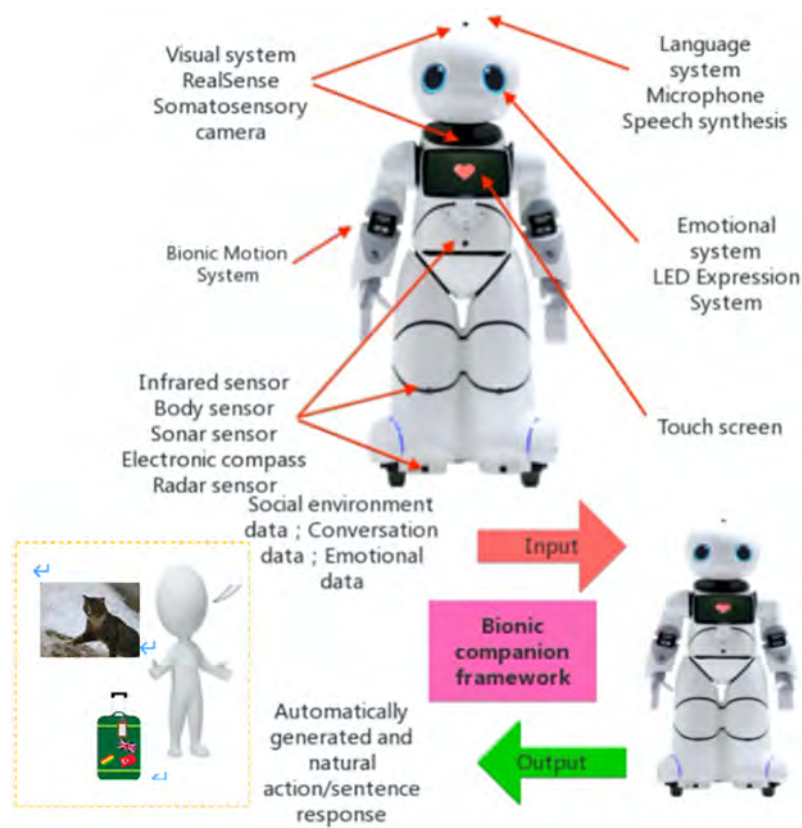
**Figure 1 HSR capabilities with the proposed high-level of HRC conceptual model.**
Full-size ☑ DOI: 10.7717/peerj-cs.674/fig-1

This study proposes adapting and integrating deep learning techniques to one of the world's most advanced HSRs so that robots can autonomously and in a timely fashion convert pictures or data information captured by robotic visions and sensors into texts or sentences in order to respond and communicate more naturally with humans. The conceptual model of the proposed system consists of various modules, as shown in Figs. 1, 2. The contributions of this research are summarized as follows:

1. In order to solve the current problems of HSRs in the hospitality industry, a new interactive concept-HRC is proposed.
2. A novel bionic interaction framework is designed based on the proposed HRC.
3. A system that can be used on HSRs is developed based on the bionic interaction framework, and the system has been tested and verified. The preliminary results prove that the system can enable HSRs to handle dynamic social environments.

Humanoid service robot used in research

The design and investigation of this HRC framework involves using the Canbot U05E humanoid robot (see Fig. 1 for the high-level design, Figs. 2–5 for further details) (*CANBOT, 2020*). The robot's 22-degree-of-freedom motion joints enable it to perform a
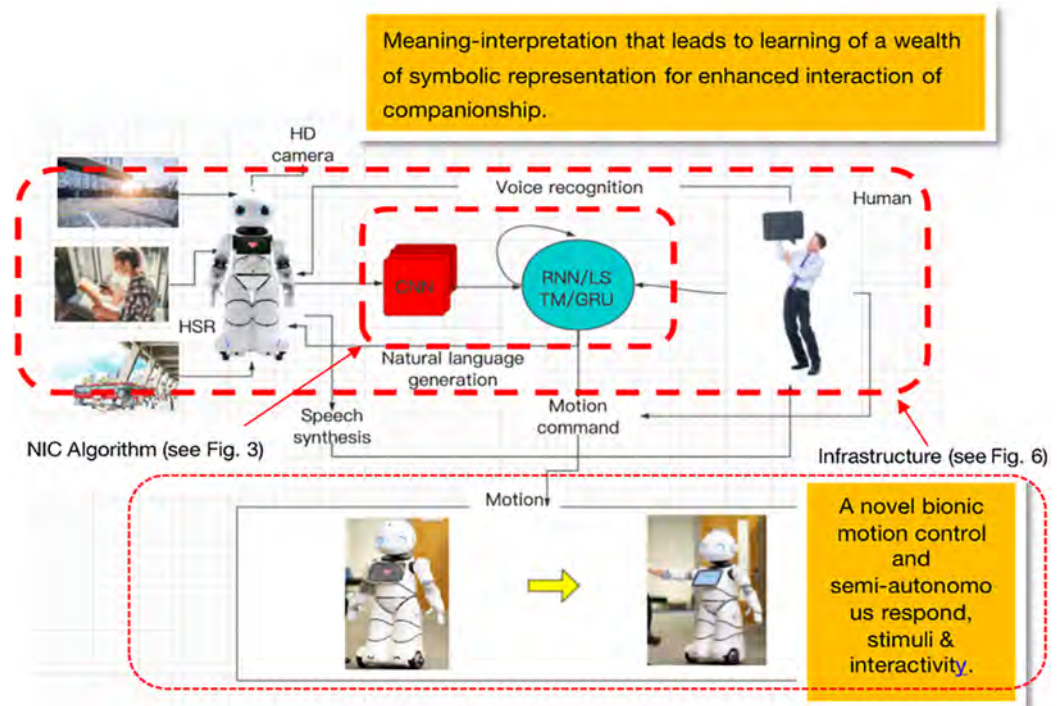
Figure 2 Bionic-companionship framework design. Full-size ⬛ DOI: 10.7717/peerj-cs.674/fig-2
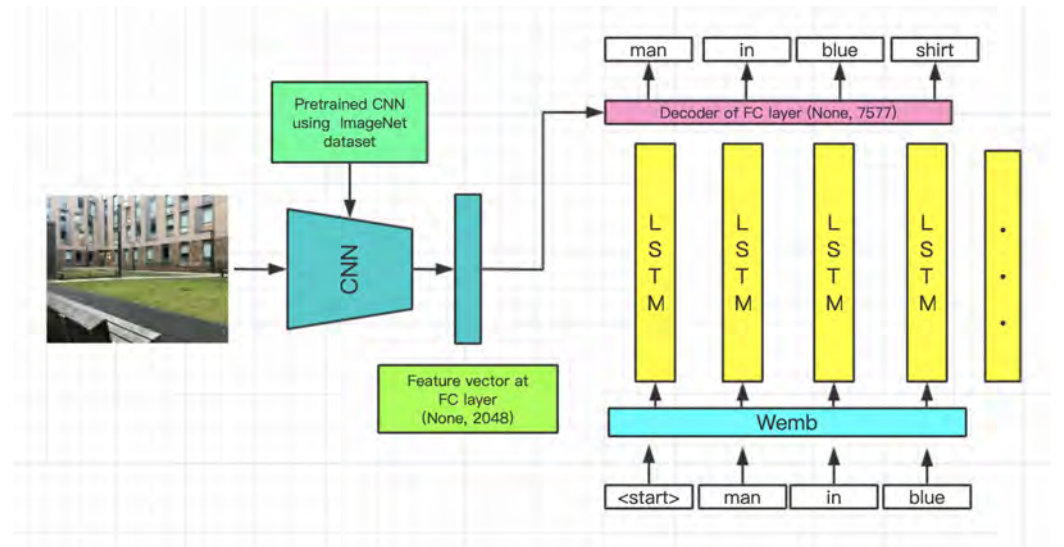
Figure 3 Neural image captioning model structure for HSR.
Full-size ⬛ DOI: 10.7717/peerj-cs.674/fig-3

variety of simulated movements, such as raising the head, turning the head, raising the arm, shaking the crank, shaking hands, leaning back, walking, and turning, and based on the proposed framework, it can acquire natural human behaviors and, as a result, efficiently interact with humans. In addition, Canbot U05E's advanced vision system and sensors can collect more complete environmental data for the proposed design and make

0. A black dog is running after a white dog in the snow .
1. Black dog chasing brown dog through snow
2. Two dogs chase each other across the snowy ground .
3. Two dogs play together in the snow .
4. Two dogs running through a low lying body of water .

0. A man and a dog are sitting on a park bench with several people walking down the path that runs alongside it .
1. A man and a dog on a park bench in the foreground , with a group of walking people in the distance .
2. A man sits on a bench in the park with a dog , while others walk nearby .
3. People sit on a park bench while others jog on a path .
4. The man sits with his dog on a bench in a park .

Figure 4 Samples Flickr 8k (*Rashtchian et al., 2010*) training data set.
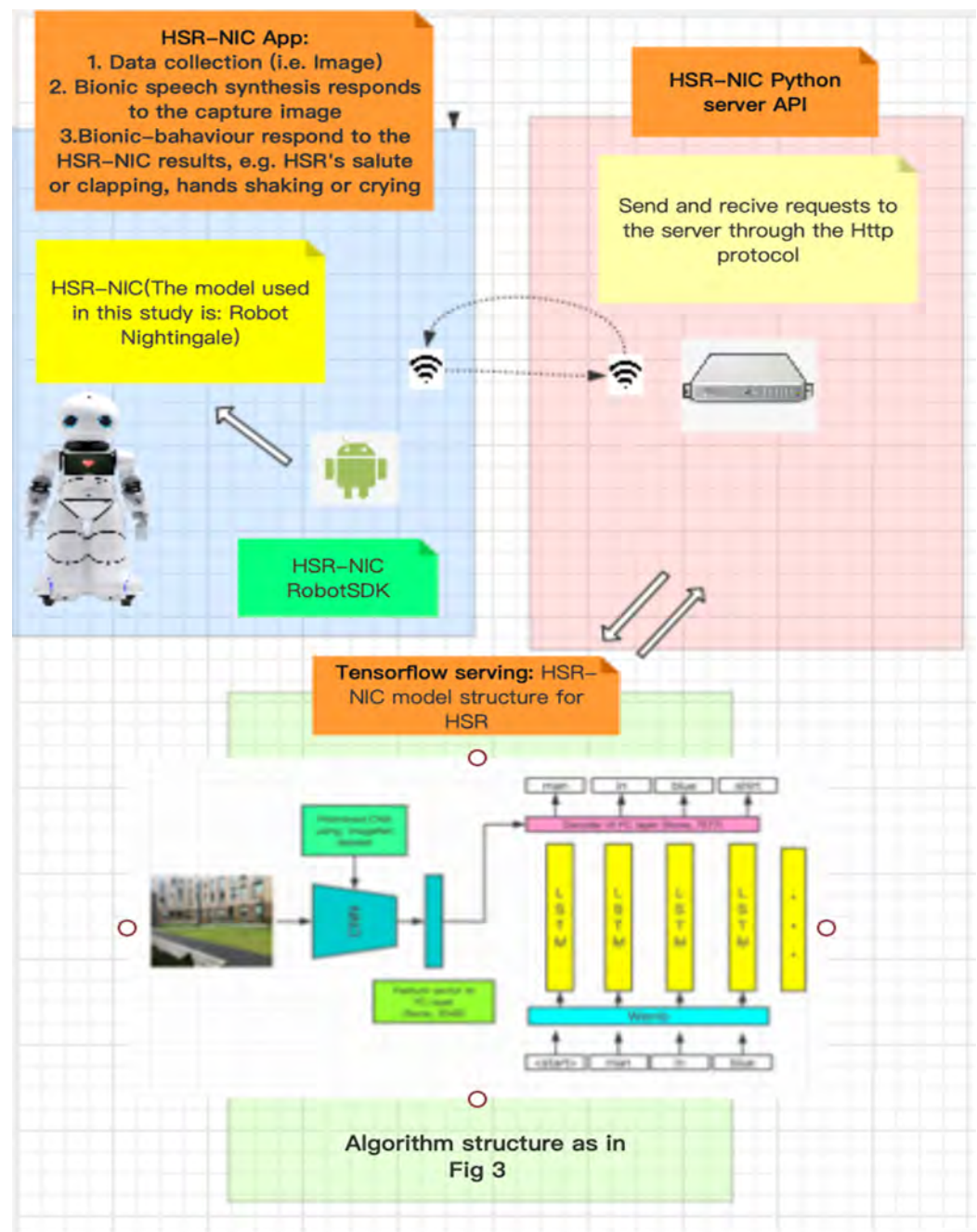Full-size ⬚ DOI: 10.7717/peerj-cs.674/fig-4

the novel framework more robust. The robot is designed to imitate the human's seven senses, providing strong support for the concept and implementation of the bionic partner designed in this study.

Bionic-companionship framework

In this study, we review the previous works on this topic and research gaps in the literature and describe a novel humanoid service robot and human interaction framework with neural image subtitles as its core (details are shown in Fig. 2). The framework uses the structure of the NIC algorithm to better realize the interaction of HSRs from HRI to the direction of bionic-companionship. According to the initial descriptions of robot companions, as in the studies by *Turkle (2006)* and (*Kim et al., 2015*), the proposed framework should provide HSRs with more natural interactions and a more sensitive understanding of the environment, and hence, the design of the framework is divided into two subsystems (see the dotted red).

Image/video description generation system

These subsystems are the core modules of the entire interactive framework. HSRs collect visual data of the surrounding environment through equipped visual sensors (such as HD or 3D cameras) and sensors (such as tactile and radar). The type of visual data collected depends on the complexity of the interactive task to be completed by HSRs. It is generally considered that more complex interactive tasks require the use of continuous images or real-time videos. The system uses the latest neural image generation algorithm structure and CNN to perform feature extraction on the pictures and video data of the surrounding social environment, and converts the data into feature vector sequences that

Figure 5 The infrastructure of the humanoid service robot generating neural image captions as part of the bionic companionship framework. Full-size ☒ DOI: 10.7717/peerj-cs.674/fig-5

can be used by RNN. Finally, the RNN completes the process of generating an interactive description from the visual data. HSRs use a speech synthesis system that converts these descriptions into voices to communicate with humans. This process is different from the past mode of using HSRs human sensing sensors and setting fixed interactive feedback; the innovation of this system is that HSRs can automatically and naturally generate interactive feedback. This means that the change in the scene during the interaction will

cause a continuous change in the interaction feedback, and this change is not preset by humans. In addition, in further conversation interactions, human voice response and social environment data will be coordinated by HSRs and produce continuous conversation interaction behavior.

Command-robot behavior system

For HSRs, simple conversation interactions are insufficient. HSRs should generate corresponding motions based on visual and human behavior data. For example, when humans wave to a robot, the robot should also actively respond. The hypothesis of this study is to classify or cluster description text generated from visual data and use these classified description texts to control the motions of HSRs in response to complex interactive tasks. For example, when the description generated by neural image captions is "Hello", then HSRs will automatically determine whether 'Hello' matches a category that requires interactive motion and performs corresponding motions such as waving.

PILOT TESTING, PRELIMINARY RESULTS, AND DISCUSSION

In the present study, we designed and integrated a classic NIC model on the HSR and performed a preliminary evaluation.

Introduction to HSR-NIC model

The structure of the HSR-NIC algorithm used in this study was adapted and enhanced from the model structure proposed by *Mao et al. (2014)* who used a classic encoder-decoder structure. In this study, the encoder uses the Xception pre-trained CNN to convert the input image into a feature vector. The word sequence is then input into the LSTM after a layer of word embedding layer, and finally, and add operation is performed on the word features output by the LSTM and the image features extracted by the trained CNN. These are then input into a decoder composed of a single-layer fully connected layer, which generates the probability distribution of the next word using a softmax layer. The LSTM introduced by the model can solve the long-term dependency problem in the traditional RNN, thereby improving the accuracy of the model. The dense representation of word embedding can reduce the amount of calculations involved in the model; it also enables the model to capture similar relationships between words. In addition, the model used in this study also introduces a dropout layer with a probability of 50% to increase the robustness of the model. The teacher forcing mechanism was used during model training to accelerate the model training process. The optimizer used in the research is Adam, which has the advantages of making the model converge more quickly and automatically adjusting the learning rate with learning. The variables of the model are updated by minimizing the cross-entropy loss between the probability distribution of the predicted result and the probability distribution of the true result and back-propagation. The model structure diagram as follow (Fig. 3).

Model forward propagation process

The training process of the image captioning task can be described as follows: For a picture in the training set, its corresponding description is a sequence that represents the words in the sentence. For model θ, given input image I from the HSR's vision, the probability of the model generating sequence is expressed as

$$P(S|I; \theta) = \Pi_{t=0}^{N} P(S_t | S_0, S_1, \ldots, S_{t-1}, I; \theta) \tag{1}$$

The logarithm of the likelihood function is used to obtain the log-likelihood function:

$$\log P(S|I; \theta) = \sum_{t=0}^{N} \log P(S_t | S_0, S_1, \ldots S_{t-1}, I; \theta) \tag{2}$$

The training objective of the model is to maximize the sum of the log-likelihoods of all training samples:

$$\theta^* = arg_{max_{\theta}} \sum_{(I,S)} \log P(S|I; \theta) \tag{3}$$

where *(I, S)* is the training sample. This method of maximum likelihood estimation is equivalent to empirical risk minimization using the log-loss function. Therefore, in the forward propagation process of this research model, the image feature vector Iv is extracted from the image using the CNN, and a two-dimensional vector of shape (batch size, 2048) is the output.

$$I_v = CNN_{\theta_c}(I) \tag{4}$$

The extracted image features need to be encoded by a fully connected layer into the context feature vector C that can be matched with word features. The word feature vector is the output O_t of the LSTM over the time step. The input word of LSTM passes through a word-embedding layer to generate a dense vector representation W(s).

$$C = W_{\theta}(I_v), \ O_t = LSTM_{\theta}(W(s)) \tag{5}$$

Finally, word feature *Ot* and context feature C are together input into a decoder composed of a single fully connected layer after the softmax calculation generates the probability distribution of the next word $P(S_i|I; \theta)$.

$$P(S_i|I; \theta) = softmax(W_{\theta}(C + O_t)) \tag{6}$$

The loss function is expressed as

$$L = \sum_{t=1}^{T} y^{(t)} \log p^{(t)} + (1 - y^t) \log(1 - p^t) \tag{7}$$

Training dataset

For the present study, we use Flickr 8k (*Rashtchian et al., 2010*) as the training dataset. This is a new benchmark collection for sentence-based image descriptions and searches. It consists of 8,000 images. Each image was paired with five different captions. These

captions provide content descriptions of the objects and events in the picture. The images do not contain any well-known people or locations but depict random scenes and situations. Examples of datasets are shown in Fig. 4. The Flickr 8k dataset not only contains images of animals and objects, but also of some social scenes. These data can help robots to better understand natural, day-to-day scenes.

The process of humanoid service robot generating image captions

To explore the feasibility of the bionic-companionship framework, preliminary tests were conducted on a real humanoid service robot (Canbot U05E). The process of generating image captions by a humanoid service robot is divided into four steps, as shown in Fig. 5.

Step 1. The HSR-NIC API is responsible for controlling the robot to call the high-definition camera to collect surrounding environment information (the data collection in this study is focused on HSR capture images). The collected data will be sent to the local host service program through the HTTP protocol and wait for a response from the HSR.

Step 2. The HSR-NIC localhost server program receives the data, and the requests perform preliminary processing and cleaning of the data (image) and send the data (image) to the HSR-NIC model server program to wait for the calculation result (the generated caption description).

Step 3. The HSR-NIC model server program analyzes the image data according to the training parameters saved before, generates the descriptive caption, and returns it to the local server.

Step 4. The HSR-NIC local server program sends the caption description to the robot application through the HTTP protocol, and the robot application controls the robot to respond according to the caption description, such as speech synthesis and motion control.

Preliminary test results and limitation

In this study, we conducted a preliminary test on a humanoid service robot integrated with the NIC algorithm. The results of the preliminary test were found to be promising.

With the discuss of the last chapter, the research will integrate the NIC into the HSRs to make the HSRs take advantage of the change of the surrounding environment interact with the human better. Therefore, the system proposed by this research will combine qualitative analysis and quantitative analysis to initially validate the performance of the system.

This study introduces the cross-entropy loss curve of the last 50 epochs of the model as the evaluation metric for quantitative analysis. As shown in the Fig. 6, the model finally converges to the minimum loss value of 2.65 in the training set and 2.71 in the validation set, which proves that the model has no over-fitting and under-fitting, and has generalization ability. Since the loss value is calculated from the sum of the difference between the probability value of each predicted word in the predicted description and the true value, the loss value will be affected by the sentence length of the predicted description. In related work, researchers (*Li et al., 2020*; *Hu et al., 2020*) used some more reliable evaluation methods to evaluate the performance of the model, including the
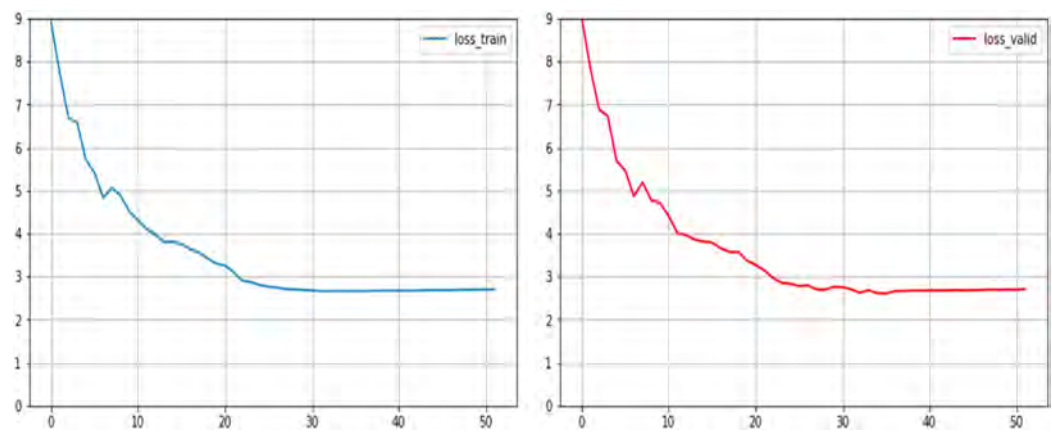
Figure 6 Loss curve of NIC model on training set and validation set.
Full-size ⬛ DOI: 10.7717/peerj-cs.674/fig-6

BLUE4 (*Papineni et al., 2002*) and CIDEr (*Vedantam, Lawrence Zitnick & Parikh, 2015*). These evaluation metrics are usually used in the field of machine translation instead of manual evaluation. Since the tasks handled by the NIC model can be regarded as translated from images/scenes into English, the evaluation metrics can also be applied to the evaluation of NIC. This study will use qualitative analysis to replace quantitative analysis of metrics such as BLUE4 and CIDEr, so as to further evaluate the preliminary performance of HSR after the integrated NIC model.

As shown in Figs. 7 and 8, the researcher conducted two sets of tests in three different scenarios with HSR. In the first set of tests, the researcher wore a hat and changed scenarios. In the second set of tests, the researcher did not wear a hat, and the scene switching method was the same as in the first set. It can be seen from the experimental results that the humanoid robot can complete the perception of scene switching through this algorithm and generate a rough description of the scene. In the first set of tests, most of the content described was accurate. The robot equipped with the NIC algorithm can effectively identify 'man', 'black shirt', and 'sitting on a bench'. However, in the second group of tests, there were many errors in the recognition results. This could be attributed to the researcher's long hair. Interestingly, researchers with long hair are easily identified as women or children. This indicates that the accuracy of the NIC algorithm still has room for improvement.

In addition, in order to test the performance of the system in a dynamic environment. The researcher conducted the test in a real environment (as Fig. 8). The researcher selected six real environments as the test data and let the robot generate interactive information. Among the six real interactive environments, there are three scenes that can be more accurately recognized by the robot and produce corresponding descriptions. The description information can correspond to the test environment, and the corresponding part of the description has been highlighted with the same color in the Fig. 8. Some of the objects, facilities, and human movements in these scenes can be accurately predicted, such as sidewalk, traffic, bench, building, building, etc. However, in the other three environments, the robot did not give an accurate description. The researchers believe that
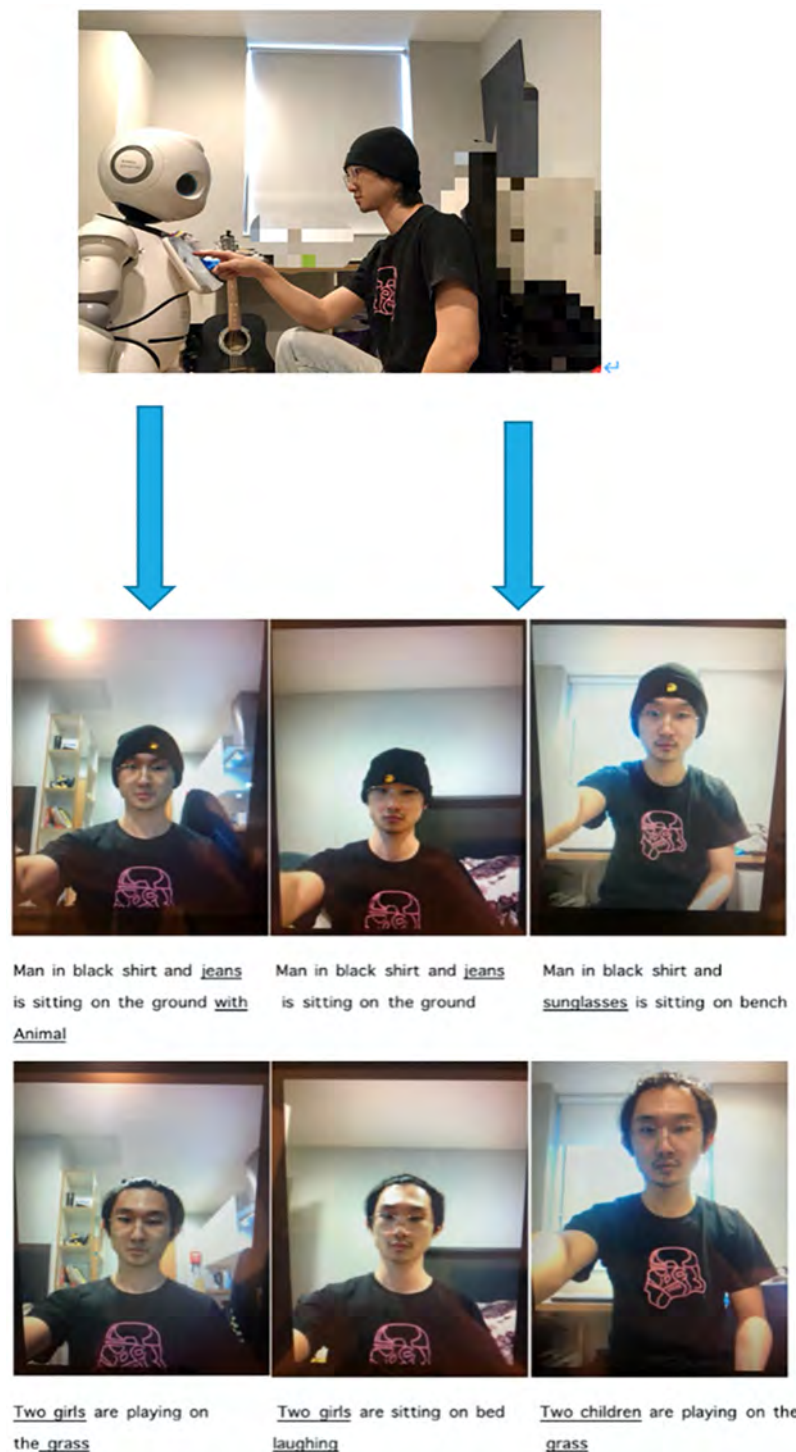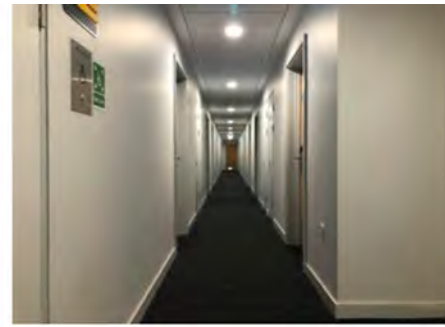
Figure 7 **A series of preliminary testing results captured from Canbot U05E and bionic-companionship preliminary framework.** Full-size ◩ DOI: 10.7717/peerj-cs.674/fig-7

this may be due to the fact that the training set does not contain objects in these three environments, causing the model to fail to learn how to express the 'unfamiliar environment'.
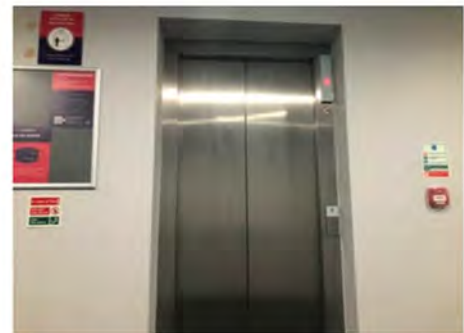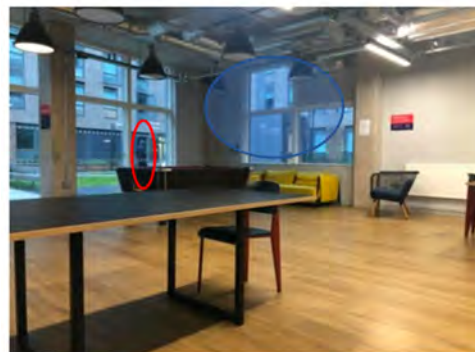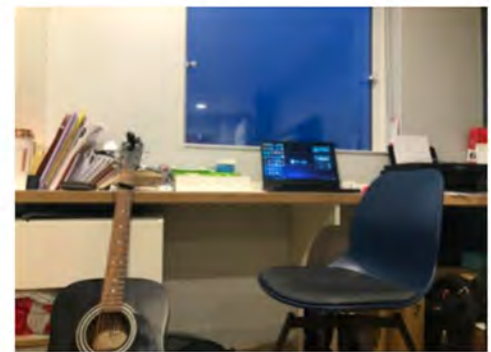
Figure 8 **Real social environment test examples.** Full-size 🖼 DOI: 10.7717/peerj-cs.674/fig-8

In general, as per the results of the to two experimental sets, it was proven that the robot equipped with the NIC algorithm can capture the changes in the surrounding environment and generate different feedbacks according to the changes. The results also demonstrate the feasibility of the proposed bionic-companionship framework. Although there is still a gap between the prediction results of the algorithm and the real communication scene, the researcher believes that special data collection for some specific interaction scenarios and model training for these specific data can be effective in addressing this gap. Future

research directions will mainly focus on improving the accuracy of algorithms and achieving more human-like interactions. (The detailed process is shown in the HSR-NIC demo video.) In addition, the researcher believes that the scene understanding of static images is the basis for dealing with dynamic environments. Some researches have mentioned that the introduction of related algorithms of object detection into NIC can identify and generate descriptions of scenes in dynamic environments. This is also the current research limitation of this research and the research challenges that will be faced in the future.

CONCLUSIONS

This study presents a review of neural image generation algorithms and application cases in the field of robotics, and proposes a novel humanoid service robot and human interaction framework based on the bionic-companionship theory. The subsystems of the bionic-companionship framework are designed and introduced in detail. Preliminary tests also initially proved that the framework could increase the sensitivity of HSRs to changes in the surrounding environment. The proposed framework will contribute to further development from HRI to HRC. Future work will focus on implementing each of the subsystems in the framework and applying the framework to HSRs to verify its performance.

ADDITIONAL INFORMATION AND DECLARATIONS

Author Contributions
- Jiaji Yang conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Esyin Chew conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Pengcheng Liu conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability
The following information was supplied regarding data availability:
   The research code is available in the Supplemental File. The training dataset for the system is Flickr 8k and is available at Kaggle: https://www.kaggle.com/adityajn105/flickr8k.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.674#supplemental-information.

REFERENCES

CANBOT. 2020. Intelligent Robot AI Makers Laboratory. *Available at* https://www.canbotrobots.com/html/yy-.

Caleb-Solly P, Dogramadzi S, Huijnen CA, Heuvel HVD. 2018. Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance. *The Information Society* **34(3)**:153–165 DOI 10.1080/01972243.2018.1444255.

Cascianelli S, Costante G, Ciarfuglia TA, Valigi P, Fravolini ML. 2018. Full-GRU natural language video description for service robotics applications. *IEEE Robotics and Automation Letters* **3(2)**:841–848 DOI 10.1109/LRA.2018.2793345.

Chen L, He Y, Fan L. 2017. Let the robot tell: describe car image with natural language via LSTM. *Pattern Recognition Letters* **98(8)**:75–82 DOI 10.1016/j.patrec.2017.09.007.

Chew E, Lee PL, Hu S, Yang J. 2021. Investigating the first robotic nurses: humanoid robot nightingale and partners for COVID-19 preventive design. In: *The Seventh Edition of the International Workshop on New Trends in Medical and Service Robots Conference*. Basel, Switzerland: University Hospital Basel.

Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *EMNLP*.

Chung MJY, Cakmak M. 2018. How was your stay?: Exploring the use of robots for gathering customer feedback in the hospitality industry. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Piscataway: IEEE, 947–954.

Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay KL, Werry I. 2005. What is a robot companion-friend, assistant or butler? In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Piscataway: IEEE, 1192–1197.

Denil M, Bazzani L, Larochelle H, de Freitas N. 2012. Learning where to attend with deep architectures for image tracking. *Neural computation* **24(8)**:2151–2184 DOI 10.1162/NECO_a_00312.

Ding S, Qu S, Xi Y, Sangaiah AK, Wan S. 2019. Image caption generation with high-level image features. *Pattern Recognition Letters* **123(1)**:89–95 DOI 10.1016/j.patrec.2019.03.021.

Donahue J, Hendrikcs LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. 2014. Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389v2. *Available at* https://arxiv.org/abs/1411.4389.

Fidler S. 2017. Teaching machines to describe images with natural language feedback. In: *Advances in Neural Information Processing Systems*. 5068–5078.

Fong T, Thorpe C, Baur C, C. Collaboration. 2003. Dialogue, human-robot interaction. In: *Robotics Research*. Berlin, Heidelberg: Springer, 255–266.

Gerber R, Nagel N-H. 1996. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In: *Proceedings of 3rd IEEE International Conference on Image Processing*. **2**:805–808.

Harris K, Kimson A, Schwedel A. 2018. Why the automation boom could be followed by a bust, Harvard business review (March 13). *Available at* https://hbr.org/2018/03/why-the-automation-boom-could-be-followed-by-a-bust.

Hu X, Yin X, Lin K, Wang L, Zhang L, Gao J, Liu Z. 2020. Vivo: surpassing human performance in novel object captioning with visual vocabulary pre-training. arXiv preprint arXiv:2009.13682. *Available at https://arxiv.org/abs/2009.13682.*

Ivanov S. 2019. Ultimate transformation: how will automation technologies disrupt the travel, tourism and hospitality industries? *Zeitschrift für Tourismuswissenschaft* **11(1)**:25–43 DOI 10.1515/tw-2019-0003.

Ivanov S, Gretzel U, Berezina K, Sigala M, Webster C. 2019. Progress on robotics in hospitality and tourism: a review of the literature. *Journal of Hospitality and Tourism Technology* **9074(3)**:75 DOI 10.1108/JHTT-08-2018-0087.

Kiros R, Salahutdinov R, Zemel R. 2014. Multimodal neural language models. In: *International Conference on Machine Learning.* 595–603.

Kim KM, Nan CJ, Ha JW, Heo YJ, Zhang BT. 2015. Pororobot: a deep learning robot that plays video Q&A games. In: *AAAI Fall Symposium Series.*

Kuznetsova P, Ordonez V, Berg TL, Choi Y. 2014. TREETALK: composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics* **2(1)**:351–362 DOI 10.1162/tacl_a_00188.

Luo RC, Hsu YT, Ye HJ. 2019. Multi-modal human-aware image caption system for intelligent service robotics applications. In: *IEEE 28th International Symposium on Industrial Electronics (ISIE).* Piscataway: IEEE, 1180–1185.

Luo RC, Hsu Y, Wen Y, Ye H. 2019. Visual image caption generation for service robotics and industrial applications,. In: *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), Taipei, Taiwan.* 827–832.

Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y, Gao J. 2020. Oscar: object-semantics. aligned pre-training for vision-language tasks. In: *European Conference on Computer Vision.* Cham: Springer, 121–137.

Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint. *Available at https://arxiv.org/abs/1412.6632.*

Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daumé H III. 2012. Midge: generating image descriptions from computer vision detections. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 747–756.

Mnih V, Hees N, Graves A, Kavukcuoglu K. 2014. Recurrent models of visual atten-tion. In: *NIPS.*

Nguyen A, Kanoulas D, Muratore L, Caldwell DG, Tsagarakis NG. 2018. Translating videos to commands for robotic manipulation with deep recurrent neural networks. In: *IEEE International Conference on Robotics and Automation (ICRA).* Piscataway: IEEE, 1–9.

Papineni K, Roukos S, Ward T, Zhu WJ. 2002. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* 311–318.

Rodriguez-Lizundia E, Marcos S, Zalama E, Gómez-García-Bermejo J, Gordaliza A. 2015. A bellboy robot: study of the effects of robot behaviour on user engagement and comfort. *International Journal of Human-Computer Studies* **82(1)**:83–95 DOI 10.1016/j.ijhcs.2015.06.001.

Rashtchian C, Young P, Hodosh M, Hockenmaier J. 2010. Collecting image annotations using amazon's mechanical turk. In: *Proceedings of the NAACL HLT, 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* 139–147.

Sutskever I, Vinyals O, Le QV. 2014. Sequence to sequence learning with neural networks. In: *NIPS*. 3104–3112.

Tang X, Zhang Q, Hu L. 2020. An EKF-based performance enhancement scheme for stochastic nonlinear systems by dynamic set-point adjustment. *IEEE Access* **8**:62261–62272 DOI 10.1109/ACCESS.2020.2984744.

Tang Y, Srivastava N, Salakhutdinov RR. 2014. Learning generative models with vi-sual attention. In: *NIPS*. 1808–1816.

Tremblay J, To T, Molchanov A, Tyree S, Kautz J, Birchfield S. 2018. Synthetically trained neural networks for learning human-readable plans from real-world demonstrations. In: *IEEE International Conference on Robotics and Automation (ICRA)*. Piscataway: IEEE, 1–5.

Trevarthen C. 2001. Intrinsic motives for companionship in understanding: their origin, development, and significance for infant mental health. *Infant Mental Health Journal: Official Publication of the World Association for Infant Mental Health* **22(1–2)**:95–131 DOI 10.1002/(ISSN)1097-0355.

Turkle S. 2006. A nascent robotics culture: new complicities for companionship. American Association for Artificial Intelligence Technical Report Series AAAI.

Vedantam R, Lawrence Zitnick C, Parikh DC. 2015. Consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 4566–4575.

Vinyals O, Toshev A, Bengio S, Erhan D. 2015. Show and tell: a neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 3156–3164.

Yang J, Chew E. 2020. A systematic review for service humanoid robotics model in hospitality, springer international journal for social robotics. *International Journal of Social Robotics* **22(2)**:253 DOI 10.1007/s12369-020-00724-y.

Yang J, Chew E. 2021. The Design Model for Robotic Waitress. *International Journal of Social Robotics* 1–11 DOI 10.1007/s12369-021-00745-1.

Yamada T, Murata S, Arie H, Ogata T. 2016. Dynamical integration of language and behavior in a recurrent neural network for human-robot interaction. *Frontiers in Neurorobotics* **10**:5 DOI 10.3389/fnbot.2016.00005.

Zhang QC, Hu L, Gow J. 2020. Output feedback stabilization for mimo semi-linear stochastic systems with transient optimisation. *International Journal of Automation and Computing* **17(1)**:83–95 DOI 10.1007/s11633-019-1193-8.