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

Jiaji Yang [Corresp., 1] , **Esyin Chew** [1] , **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

Corresponding Author: Jiaji Yang
Email address: JYang@cardiffmet.ac.uk

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.

Service Humanoid Robotics: A Novel Interactive System Based on Bionic-Companionship Framework

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

1 EUREKA Robotics Lab, Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, CF5 2YB, United Kingdom.

2 The Department of Computer Science, University of York, York YO10 5GH, United Kingdom.

Corresponding Author:

Jiaji Yang[1]

Western Avenue, Cardiff, CF5 2YB

Email address: JYang@cardiffmet.ac.uk

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.

30   Compared to previous research a novel interactive system is developed based on the bionic-
31   companionship framework. The humanoid service robot was integrated with the system to
32   conduct preliminary tests. The results show that the interactive system based on the bionic-
33   companionship framework can help the service humanoid robot to effectively respond to changes
34   in the interactive environment, for example give different responses to the same character in
35   different scenes.

36   ## Introduction

37   Humanoid service robots (HSRs) have seen a sharp rise in adoption recently and are seen as one
38   of the major technologies that will drive the service industries in the next decade (Harris et al.,
39   2018). An increasing number of researchers are committed to investigating HSRs to help humans
40   complete repetitive or high-risk service and interactive tasks such as serving patients with infectious
41   diseases, delivering meals and so on. Delivery robots, concierge robots, and chat robots have been
42   increasingly used by travel and hospitality companies (Ivanov, 2019). Although the contribution
43   of these achievements mainly comes from the rapid development of robotics engineering, Ivanov
44   et al. (2019) indicated that future research focus will gradually shift from robotics engineering to
45   human-robot interaction (HRI), thus opening up interdisciplinary research directions for
46   - researchers.
47

48   In the early days, Fong et al. (2003) proposed that in order to make robots perform better, the
49   robot needs to be able to use human skills (perception, cognition, etc.) and benefit from human
50   advice and expertise. This means that robots that rely solely on self-determination have
51   limitations in performing tasks. The authors further propose that the collaborative work between
52   humans and robots will be able to break this constraint, and research on human-robot interaction
53   has begun to emerge. Fong et al. (2003) believe that to build a collaborative control system and
54   complete human-robot interaction, four key problems must be solved. (1) The robot must be able
55   to detect limitations (what can be done and what humans can do), determine whether to seek help,
56   and identify when it needs to be resolved. (2) The robot must be self-reliant and secure. (3) The
57   system must support dialog. That is, robots and humans need to be able to communicate with each
58   other effectively. However, dialog is restricted at present. Through collaborative control, dialog
59   should be two-way and require a richer vocabulary. (4) The system must be adaptive. Although
60   most of the current humanoid service robots already support dialog and can complete simple
61   interactive tasks, as propounded in the research, such dialog in the present time remains limited
62   and "inhuman." In the process of interacting with robots, humans always determine the state of
63   the robot (the position of the robot or the action the robot is doing) through vision, and then
64   communicate with the robot through a dialog system. However, HSR cannot perform this yet as
65   they do not seem to fully satisfy the two-way nature of dialog. Therefore, this research responds
66   to the current gap and attempts to differ from the current HRI research. This research attempts to
67   introduce deep learning into the existing dialog system of HSR, thus advancing the field.
68

69   With the continuous development of humanoid robots, more and more humanoid robots are used
70   in the service industry, especially the hospitality industry. Human-Robot Interaction (HRI) has
71   become a hot potato by more and more researchers (Yang & Chew, 2020). However, with the
72   deepening of research, some researchers found that when humans interact with humanoid service
73   robots (HSRs), humans hope that HSRs should have the ability and interest to interact with the
74   dynamic thoughts and enthusiasm of the partner's relationship, and can recognize the
75   environment, blended with what others think is meaningful and the emotions to express sympathy
76   (Yang & Chew, 2020). This coincides with Trevarthen Companionship Theory (Trevarthen,
77   2001), so the concept of human robot companion (HRC) was proposed this research. The earlier
78   concept of the robot companion is mentioned by Dautenhahn et al. (2005): HSRs need to have a
79   high degree of awareness and sensitivity to social environment. Through the review of the above
80   literatures, it is proposed to establish an interactive and companion framework for HSRs using
81   deep learning and neural image caption generation, thus advance the current field of HSRs to
82   tackle with bionic-interactive tasks of the service industry and further evolve from conventional
83   HRI to Human and Robot Companion (HRC) (See Tab. 1).

84

85   This study proposes that the introduction of visual data into the current HRI model of HSRs
86   enables HSRs to have a high level of sensitivity to the social dynamic environment while
87   interacting with humans, thereby enhancing the current HRI model to HRC. With the continuous
88   development of deep learning, some researchers have recently realized the transformation of static
89   pictures or videos from conventional camera input into text descriptions (Li et al., 2020; Hu et al.,
90   2020; Luo, 2019). This deep learning algorithm model is called neural image capturing (NIC).
91   This research attempts to adapt and integrate NIC into HSRs and propose a novel framework
92   (bionic-companionship framework) to enhance the traditional HRI experience. This framework
93   aims to improve the current HRI interaction mode in the field of HSRs to a higher level of HRC
94   (Yang & Chew, 2021). The bionics in this research refers to the humanoid service robot imitating
95   all the tastes of life, trying to adapt to the seven emotions of ancient human nature (joy, anger,
96   sadness, fear, love, disgust, liking) and six biological wills (life, death, eyes, ears, mouth, nose)
97   (Chew et al.,2021). The system proposed in this study combines visual intelligence and Speech
98   Intelligence, and imitates human behavior in social activities, which is in line with the concept of
99   robot bionics proposed by researchers such as Chew et al (2021). Therefore, this study believes
100   that the proposed system is a bionic system.

Related works

102   With the continuous development of HRI research, industrial robots have been able to interact
103   with humans accurately and self-adaptively. Some advanced control systems (Zhang et al., 2020)
104   and algorithms (Tang et al., 2020) have been proposed as Industrial robots provide reliable
105   support for completing interactive tasks in an industrial environment. However, as HSRs began to
106   enter the service industry, some research cases began to discover that there are still problems with
107   the interaction of HSRs in the social environment. Caleb Solly (2018) believed that users can also

108   help robots when robots help users; meanwhile, users can give feedback to optimize the system.
109   The feedback reflects not only the optimization of the robot system but also the satisfaction of
110   customers. Chung's (2018) study indicated that hotels in the hospitality industry want to collect
111   customer feedback in real-time to immediately disseminate positive feedback and respond to
112   unsatisfactory customers while they are still on the scene. Guests want to inform their experience
113   without affecting their privacy. Stakeholders in the hospitality industry hope that intelligent robots
114   can interact more with users. Besides, Rodriguez (2015) concluded that the optimal distance
115   between users and robots is 69.58 cm. Specifically, interaction with a certain greeting mode can
116   attract users to maintain a longer interaction time; robots with the active search are more attractive
117   to participants. The interaction time is longer than that of passively searching robots, suggesting
118   that robots should be designed to keep at a certain distance from humans and consider adding the
119   ability to allow robots to actively identify customers and attract them.

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

128   ***Review of Neural Image Captioning***
129

130   The challenge of generating natural language descriptions from visual data has been extensively
131   researched in the field of computer vision. However, early research has mainly focused on
132   generating natural language descriptions from video-type visual data (Gerber, 1996; Mitchell et
133   al, 2012). These systems convert complex visual data into natural languages using rule-based
134   systems. However, because the rules are artificially designed, these systems are sufficiently
135   robust, bionic, and have been shown to be beneficial in limited applications such as traffic
136   scenarios (Vinyals et al., 2015). In the past decade, various researchers, inspired by the successful
137   use of sequence-to-sequence training with neural networks for machine translation, proposed a
138   method for generating image descriptions based on recurrent neural networks (RNNs) (Cho et al.,
139   2014; Sutskever et al., 2014). In fact, this method of replacing the encoder in the encoder-decoder
140   framework in machine translation with image features transforms the original complex task of
141   generating image data caption into a simple process of "translating" the image into a sentence
142   (Cho et al., 2014). Furthermore, Donahue et al. (2014) used long short-term memory (LSTM) for
143   end-to-end large-scale visual learning processes. In addition to images, Donahue et al. (2014) also
144   applied LSTM to videos, allowing their models to generate video descriptions. Vinyals et al.
145   (2015) and Kiros et al. (2014) initially proposed the structure of a currently popular neural image
146   generation algorithm based on the combination of a convolutional neural network (CNN) image
147   recognition model and a natural language processing (NLP) structured model. Moreover, the

148   neural image captioning algorithm based on the attention mechanism has also attracted extensive
149   attention in the field of computer vision. Denil et al. (2012) proposed a real-time target tracking
150   and attention recognition model driven by sight data. Tang et al. (2014) proposed an attention-
151   generation model based on deep learning. From the perspective of visual neuroscience, the model
152   requires object-centric data collection for model generation. Subsequently, Mnih et al. (2014)
153   proposed a new recurrent neural network model, which can adaptively select specific areas or
154   locations to extract information from images or videos and process the selected area at high
155   resolution. As the algorithm has increasingly mature, the application of the algorithm in related
156   fields has also been breaking through recently, such as the caption generation of car images (Chen
157   et al., 2017), the description generation of facial expressions (Kuznetsova et al., 2014), and
158   educational NAO robots driven by image caption generation for video Q&A games for children's
159   education (Kim, 2015). Recent research on image caption generation also shows that the accuracy
160   and reliability of the technology have increased (Ding et al., 2019). In addition, reinforcement
161   learning to automatically correct image caption generation networks have also been proposed
162   (Fidler, 2017). These deep learning–based studies have undoubtedly laid a foundation for the
163   possible NIC integration with HSRs as proposed in this study. The novel integration led to the
164   possibility for humanoid robots to interact with humans while recognizing the social environment
165   in real time, thereby improving the interactive service quality of the HSRs.
166
167   ***Neural Image Caption Generation Algorithm 'Crash Into' Robot***
168
169   An increasing number of studies have been conducted on HRI combined with image caption
170   generation algorithm. Kim et al. (2015) used the structure of a convolutional neural network
171   (CNN) combined with RNN + deep concept hierarchies (DCH) to design and develop an
172   educational intelligent humanoid robot system for play video games with children. In this study,
173   CNN was used to extract and pre-process cartoons with educational features, and RNN and DCH
174   were used to convert the collected video features into Q&A about cartoons. During the game,
175   after watching the same cartoon, the child and the robot ask and answer questions based on the
176   content of the cartoon. The research results show that such a system can interact effectively with
177   children. However, for HRIs, such simple and limited-structured Q&A conditions cannot satisfy
178   all the interaction scenarios required. Cascianelli et al. (2018) used a gull-gated recurrent unit
179   (GRU) encoder-decoder architecture to develop a human-robot interface that provides interactive
180   services for service robots. This research solves a problem called natural language video
181   description (NLVD). The authors also compared the performance when using LSTM and GRU
182   with two different algorithms to solve these problems. They demonstrated that the GRU algorithm
183   runs faster and consumes less memory. This type of model may be more suitable for HSRs.
184   Although the research model is competitive on public datasets, the experimental results on the
185   designed datasets show that the model suffers from significant overfitting. This proves that in the
186   actual model training process, a specific training dataset for HSR interaction should be
187   established, and other methods (such as transfer learning) should be considered to improve the
188   generalization ability of the model for interactive tasks. Luo et al. (2019, June) created a

189 description template to add various image features collected by the robot, such as face recognition
190 and expression, to the generated description. Compared with the previous models, their interaction
191 is slightly more natural and closer to the human description. However, Luo et al. use the model to
192 provide limited services to industry managers, hard to generalize, and not for developing an entire
193 HRI framework.
194
195 Like the research on robot vision language, research on robot vision action is in its infancy.
196 Yamada et al. (2016) used RNNs to enable robots to learn commands online from humans and
197 respond with corresponding behaviors. This research furthermore provides a reference and
198 direction for humanoid robots to use deep learning to obtain online learning capabilities for
199 human commands. Inspired by the above study, the rationale and hypothesis proposed in the
200 present research are that the description generated by the neural image captions can drive HSRs to
201 perform appropriate behaviors, and HSRs can even obtain online learning capabilities of
202 interacting with surrounding people through studying and analyzing social environments.
203 Tremblay et al. (2018) and Nguyen et al. (2018) believe that non-experts often lack the rationality
204 of task descriptions when issuing instructions to robots. They use deep learning to allow robots to
205 automatically generate human-readable instructions' descriptions according to the surrounding
206 social environment. In addition, Nguyen et al. (2018) also used visual data to make humanoid
207 robots imitate and learn human actions under corresponding commands so that the robot can learn
208 how to complete the corresponding tasks only through visual data; however, social robots cannot
209 complete precise control of movements when they imitate movements of visual data.

Contribution to the Knowledge: The Bionic-Companionship Framework with NIC for HSRs

212 The contribution of the present study is the novel investigation and design of the bionic-
213 companionship framework for HSRs, adapting and integrating neural image caption generation
214 algorithms and bionic humanoid robots, to be validated in a lab-controlled environment and real-
215 life exploration. The new HRC framework is anticipated to enhance HRI to reach a new state,
216 making it possible for HSRs to become bionic companions of humans.
217
218 This study proposes adapting and integrating deep learning techniques to one of the world's most
219 advanced HSRs so that robots can autonomously and in a timely fashion convert pictures or data
220 information captured by robotic visions and sensors into texts or sentences in order to respond and
221 communicate more naturally with humans. The conceptual model of the proposed system consists
222 of various modules, as shown in Figures 1, 2. The contributions of this research are summarized
223 as follows:
224 (1) In order to solve the current problems of HSRs in the hospitality industry, a new interactive
225 concept - HRC is proposed.
226 (2) A novel bionic interaction framework is designed based on the proposed HRC.

227 (3) A system that can be used on HSRs is developed based on the bionic interaction framework,
228 and the system has been tested and verified. The preliminary results prove that the system can
229 enable HSRs to handle dynamic social environments.
230

231 ***Humanoid service robot used in research***
232

233 The design and investigation of this HRC framework involves using the Canbot U05E humanoid
234 robot (see Fig. 1 for the high-level design, Figs 2–5 for further details) (Canbot, 2021). The
235 robot's 22-degree-of-freedom motion joints enable it to perform a variety of simulated
236 movements, such as raising the head, turning the head, raising the arm, shaking the crank, shaking
237 hands, leaning back, walking, and turning, and based on the proposed framework, it can acquire
238 natural human behaviors and, as a result, efficiently interact with humans. In addition, Canbot
239 U05E's advanced vision system and sensors can collect more complete environmental data for the
240 proposed design and make the novel framework more robust. The robot is designed to imitate the
241 human's seven senses, providing strong support for the concept and implementation of the bionic
242 partner designed in this study.

243

244 ***Bionic-Companionship Framework***
245

246 In this study, we review the previous works on this topic and research gaps in the literature and
247 describe a novel humanoid service robot and human interaction framework with neural image
248 subtitles as its core (details are shown in Fig. 2). The framework uses the structure of the NIC
249 algorithm to better realize the interaction of HSRs from HRI to the direction of bionic-
250 companionship. According to the initial descriptions of robot companions, as in the studies by
251 Turkle (2006) and (Kim et al., 2015), the proposed framework should provide HSRs with more
252 natural interactions and a more sensitive understanding of the environment, and hence, the design
253 of the framework is divided into two subsystems (see the dotted red).
254

255 ***Image/Video Description Generation System***
256

257 These subsystems are the core modules of the entire interactive framework. HSRs collect visual
258 data of the surrounding environment through equipped visual sensors (such as HD or 3D cameras)
259 and sensors (such as tactile and radar). The type of visual data collected depends on the
260 complexity of the interactive task to be completed by HSRs. It is generally considered that more
261 complex interactive tasks require the use of continuous images or real-time videos. The system
262 uses the latest neural image generation algorithm structure and CNN to perform feature extraction
263 on the pictures and video data of the surrounding social environment, and converts the data into
264 feature vector sequences that can be used by RNN. Finally, the RNN completes the process of
265 generating an interactive description from the visual data. HSRs use a speech synthesis system

266 that converts these descriptions into voices to communicate with humans. This process is different
267 from the past mode of using HSRs human sensing sensors and setting fixed interactive feedback;
268 the innovation of this system is that HSRs can automatically and naturally generate interactive
269 feedback. This means that the change in the scene during the interaction will cause a continuous
270 change in the interaction feedback, and this change is not preset by humans. In addition, in further
271 conversation interactions, human voice response and social environment data will be coordinated
272 by HSRs and produce continuous conversation interaction behavior.
273

274 ***Command-Robot Behavior System***
275

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

Pilot testing, Preliminary Results, and Discussion

285

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

288

289 ***Introduction to HSR-NIC Model***
290

291 The structure of the HSR-NIC algorithm used in this study was adapted and enhanced from the
292 model structure proposed by Mao et al. (2014) who used a classic encoder-decoder structure. In
293 this study, the encoder uses the Xception pre-trained CNN to convert the input image into a
294 feature vector. The word sequence is then input into the LSTM after a layer of word embedding
295 layer, and finally, and add operation is performed on the word features output by the LSTM and
296 the image features extracted by the trained CNN. These are then input into a decoder composed of
297 a single-layer fully connected layer, which generates the probability distribution of the next word
298 using a softmax layer. The LSTM introduced by the model can solve the long-term dependency
299 problem in the traditional RNN, thereby improving the accuracy of the model. The dense
300 representation of word embedding can reduce the amount of calculations involved in the model; it
301 also enables the model to capture similar relationships between words. In addition, the model used
302 in this study also introduces a dropout layer with a probability of 50% to increase the robustness

303   of the model. The teacher forcing mechanism was used during model training to accelerate the
304   model training process. The optimizer used in the research is Adam, which has the advantages of
305   making the model converge more quickly and automatically adjusting the learning rate with
306   learning. The variables of the model are updated by minimizing the cross-entropy loss between
307   the probability distribution of the predicted result and the probability distribution of the true result
308   and back-propagation. The model structure diagram as follow (Fig. 3):

309   ***Model Forward Propagation Process***
310

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

315
316
$$P(\mathrm{S}\,|\,I;\theta) = \Pi_{t\,=\,0}^{N}P(S_t|S_0,S_1,...,S_{t-1},I;\theta) \quad (1)$$
317

318   The logarithm of the likelihood function is used to obtain the log-likelihood function:
319
320
$$log\,\mathrm{P}(S\,|\,I;\theta) = \sum_{t\,=\,0}^{N}log\,P(S_t|S_0,S_1,...S_{t-1},I;\theta) \quad (2)$$
321

322   The training objective of the model is to maximize the sum of the log-likelihoods of all training
323   samples:
324
325
$$\theta^{*} = \arg\max_{\theta}\sum_{(I,S)}log\,P(S|I;\theta) \quad (3)$$
326

327   where *(I, S)* is the training sample. This method of maximum likelihood estimation is equivalent
328   to empirical risk minimization using the log-loss function. Therefore, in the forward propagation
329   process of this research model, the image feature vector Iv is extracted from the image using the
330   CNN, and a two-dimensional vector of shape (batch size, 2048) is the output.
331
332
$$I_v = CNN_{\theta_c}(I) \quad (4)$$
333

334   The extracted image features need to be encoded by a fully connected layer into the context
335   feature vector C that can be matched with word features. The word feature vector is the output 0_t
336   of the LSTM over the time step. The input word of LSTM passes through a word-embedding
337   layer to generate a dense vector representation W(s).
338
339
$$C = W_{\theta}(I_v), O_t = LSTM_{\theta}(W(s)) \quad (5)$$
340

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

344

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

346

347 The loss function is expressed as

348

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

350

351 ***Training Dataset***
352

353 For the present study, we use Flickr 8k (Rashtchian et al., 2010) as the training dataset. This is a
354 new benchmark collection for sentence-based image descriptions and searches. It consists of
355 8,000 images. Each image was paired with five different captions. These captions provide content
356 descriptions of the objects and events in the picture. The images do not contain any well-known
357 people or locations but depict random scenes and situations. Examples of datasets are shown in
358 Fig. 4. The Flickr 8k dataset not only contains images of animals and objects, but also of some
359 social scenes. These data can help robots to better understand natural, day-to-day scenes.
360

361 ***The Process of Humanoid Service Robot Generating Image Captions***

362 To explore the feasibility of the bionic-companionship framework, preliminary tests were
363 conducted on a real humanoid service robot (Canbot U05E). The process of generating image
364 captions by a humanoid service robot is divided into four steps, as shown in Fig 5.
365

366 **Step 1.** The HSR-NIC API is responsible for controlling the robot to call the high-definition
367 camera to collect surrounding environment information (the data collection in this study is
368 focused on HSR capture images). The collected data will be sent to the local host service program
369 through the HTTP protocol and wait for a response from the HSR.
370

371 **Step 2.** The HSR-NIC localhost server program receives the data, and the requests perform
372 preliminary processing and cleaning of the data (image) and send the data (image) to the HSR-
373 NIC model server program to wait for the calculation result (the generated caption description).
374

375 **Step 3.** The HSR-NIC model server program analyzes the image data according to the training
376 parameters saved before, generates the descriptive caption, and returns it to the local server.
377

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

381

382 ***Preliminary Test Results and Limitation***
383

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

386

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

391

392 This study introduces the cross-entropy loss curve of the last 50 epochs of the model as the
393 evaluation metric for quantitative analysis. As shown in the Fig. 6, the model finally converges to
394 the minimum loss value of 2.65 in the training set and 2.71 in the validation set, which proves that
395 the model has no over-fitting and under-fitting, and has generalization ability. Since the loss value
396 is calculated from the sum of the difference between the probability value of each predicted word
397 in the predicted description and the true value, the loss value will be affected by the sentence
398 length of the predicted description. In related work, researchers (Li et al., 2020; Hu et al., 2020)
399 used some more reliable evaluation methods to evaluate the performance of the model, including
400 the BLUE4 (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015). These evaluation metrics
401 are usually used in the field of machine translation instead of manual evaluation. Since the tasks
402 handled by the NIC model can be regarded as translated from images/scenes into English, the
403 evaluation metrics can also be applied to the evaluation of NIC. This study will use qualitative
404 analysis to replace quantitative analysis of metrics such as BLUE4 and CIDEr, so as to further
405 evaluate the preliminary performance of HSR after the integrated NIC model.

406

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

418

419 In addition, in order to test the performance of the system in a dynamic environment. The
420 researcher conducted the test in a real environment (As Fig. 8). The researcher selected six real
421 environments as the test data and let the robot generate interactive information. Among the six
422 real interactive environments, there are three scenes that can be more accurately recognized by the
423 robot and produce corresponding descriptions. The description information can correspond to the
424 test environment, and the corresponding part of the description has been highlighted with the
425 same color in the Fig. 8. Some of the objects, facilities, and human movements in these scenes can
426 be accurately predicted, such as sidewalk, traffic, bench, building, building, etc. However, in the
427 other three environments, the robot did not give an accurate description. The researchers believe
428 that this may be due to the fact that the training set does not contain objects in these three
429 environments, causing the model to fail to learn how to express the 'unfamiliar environment' .

430

431 In general, as per the results of the to two experimental sets, it was proven that the robot equipped
432 with the NIC algorithm can capture the changes in the surrounding environment and generate
433 different feedbacks according to the changes. The results also demonstrate the feasibility of the
434 proposed bionic-companionship framework. Although there is still a gap between the prediction
435 results of the algorithm and the real communication scene, the researcher believes that special
436 data collection for some specific interaction scenarios and model training for these specific data
437 can be effective in addressing this gap. Future research directions will mainly focus on improving
438 the accuracy of algorithms and achieving more human-like interactions. (The detailed process is
439 shown in the HSR-NIC demo video). In addition, the researcher believes that the scene
440 understanding of static images is the basis for dealing with dynamic environments. Some
441 researches have mentioned that the introduction of related algorithms of object detection into NIC
442 can identify and generate descriptions of scenes in dynamic environments. This is also the current
443 research limitation of this research and the research challenges that will be faced in the future.

444

Conclusions

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

References

455 CANBOT. (2020). Retrieved 4 June 2020, from https://www.canbotrobots.com/html/yy-
456   detail.html
457 Caleb-Solly P, Dogramadzi S, Huijnen CA, Hvd Heuvel (2018) Exploiting ability for human
458   adaptation to facilitate improved human-robot interaction and acceptance. Inf Soc
459   34(3):153–165
460 Cascianelli, S., Costante, G., Ciarfuglia, T. A., Valigi, P., & Fravolini, M. L. (2018). Full-GRU
461   natural language video description for service robotics applications. IEEE Robotics and
462   Automation Letters, 3(2), 841-848.
463 C. Trevarthen (2017). Play with infants: The impulse for human story-telling,In, Tina Bruce,
464   Pentti Hakkarainen and Milda Bredikyte (Eds.)The Routledge International Handbook of
465   Play in Early Childhood.Abingdon: Taylor & Francis/Routledge, Chapter 15.
466   http://www.becera.org.uk/BECERA%202017/CT%20ON%20PLAYRoutledge%20Handboo
467   k%202017.pdf
468 Chen, L., He, Y., & Fan, L. (2017). Let the robot tell: describe car image with natural language
469   via LSTM. Pattern Recognition Letters, 98, 75-82.
470 Chew E., Lee PL., Hu S. & Yang J. (2021). Investigating the First Robotic Nurses: Humanoid
471   Robot Nightingale and Partners for COVID-19 Preventive Design, the Seventh edition of the
472   International Workshop on New Trends in Medical and Service Robots Conference,
473   University Hospital Basel, Basel, Switzerland, 7-9 June 2021. https://www.mesrob2021.org
474   (Springer-indexed)
475 Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger,
476   and Bengio, Yoshua. (2014). Learning phrase representations using RNN encoder-decoder
477   for statistical machine translation. In EMNLP
478 Chung MJY, Cakmak M (2018) "how was your stay?" : Exploring the use of robots for
479   gathering customer feedback in the hospitality industry. In: 2018 27th IEEE International
480   Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, pp 947–
481   954
482 Dautenhahn, K ; Woods, S ; Kaouri, C ; Walters, M.L ; Kheng Lee Koay & Werry, I., (2005).
483   What is a robot companion - friend, assistant or butler? (2005) IEEE/RSJ International
484   Conference on Intelligent Robots and Systems, pp.1192–1197.
485 Denil, M., Bazzani, L., Larochelle, H., & de Freitas, N. (2012). Learning where to attend with
486   deep architectures for image tracking. Neural computation, 24(8), 2151-2184.
487 Ding, S., Qu, S., Xi, Y., Sangaiah, A.K. and Wan, S., 2019. Image caption generation with high-
488   level image features. Pattern Recognition Letters, 123, pp.89-95.
489 Donahue, Jeff, Hendrikcs, Lisa Anne, Guadarrama, Segio, Rohrbach, Marcus, Venugopalan,
490   Subhashini, Saenko, Kate, and Darrell, Trevor. (2014). Long-term recurrent convolutional
491   networks for visual recognition and description. arXiv:1411.4389v2.
492 Fidler, S. (2017). Teaching machines to describe images with natural language feedback. In
493   Advances in Neural Information Processing Systems (pp. 5068-5078).

494 Fong, T., Thorpe, C., & Baur, C. Collaboration (2003) dialogue, human-robot interaction.
495   In Robotics Research, pp. 255-266. Springer, Berlin, Heidelberg.
496 Gerber, R. & Nagel, N.-H., (1996). Knowledge representation for the generation of quantified
497   natural language descriptions of vehicle traffic in image sequences. Proceedings of 3rd
498   IEEE International Conference on Image Processing, 2, pp.805–808 vol.2.
499 Gui, L. Y., Zhang, K., Wang, Y. X., Liang, X., Moura, J. M., & Veloso, M. (2018, October).
500   Teaching robots to predict human motion. In 2018 IEEE/RSJ International Conference on
501   Intelligent Robots and Systems (IROS) (pp. 562-567). IEEE.
502 Harris, Karen, Austin Kimson, and Andrew Schwedel (2018), "Why the Automation Boom Could
503   Be Followed by a Bust," Harvard Business Review (March 13),
504   https://hbr.org/2018/03/why-the-automation-boom-could-be-followed-by-a-bust.
505 Hu, X., Yin, X., Lin, K., Wang, L., Zhang, L., Gao, J., & Liu, Z. (2020) Vivo: Surpassing human.
506   performance in novel object captioning with visual vocabulary pre-training. arXiv preprint
507   arXiv:2009.13682.
508 Ivanov, S. (2019). Ultimate transformation: How will automation technologies disrupt the travel,
509   tourism and hospitality industries? Zeitschrift für Tourismuswissenschaft, 11 (1), 25–43.
510 Ivanov, S., Gretzel, U., Berezina, K., Sigala, M., & Webster, C. (2019). Progress on robotics in
511   hospitality and tourism: A review of the literature. Journal of Hospitality and Tourism
512   Technology. https://doi.org/10.1108/JHTT-08-2018-0087.
513 Kim, K. M., Nan, C. J., Ha, J. W., Heo, Y. J., & Zhang, B. T. (2015, September). Pororobot: A
514   deep learning robot that plays video Q&A games. In 2015 AAAI Fall Symposium Series.
515 Luo, R. C., Hsu, Y. T., & Ye, H. J. (2019, June). Multi-Modal Human-Aware Image Caption
516   System for Intelligent Service Robotics Applications. In 2019 IEEE 28th International
517   Symposium on Industrial Electronics (ISIE) (pp. 1180-1185). IEEE.
518 Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020) Oscar: Object-semantics.
519   aligned pre-training for vision-language tasks. In European Conference on Computer Vision,
520   pp. 121-137. Springer, Cham.
521 Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T.,
522   Stratos, K., & Daumé III, H. (2012). Midge: Generating image descriptions from
523   computer vision detections. In Proceedings of the 13th conference of the European chapter
524   of the   association for computational linguistics(pp. 747–756). Association for
525   Computational Linguistics.
526 Nguyen, A., Kanoulas, D., Muratore, L., Caldwell, D. G., & Tsagarakis, N. G. (2018, May).
527   Translating videos to commands for robotic manipulation with deep recurrent neural
528   networks. In 2018 IEEE International Conference on Robotics and Automation (ICRA)
529   (pp. 1-9). IEEE.[19]
530 Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation
531   of machine translation. In Proceedings of the 40th annual meeting of the Association for
532   Computational Linguistics, pp. 311-318.

533 P. Kuznetsova, V. Ordonez, T.L. Berg, Y. Choi, (2014), TREETALK: Composition and
534   compression of trees for image descriptions, Trans. Assoc. Comput.Ling. 2 (1) 351–362.
535 R. Kiros, R. Salahutdinov, R. Zemel, (2014), Multimodal neural Language models, in:
536   International Conference on Machine Learning, pp. 595–603.
537 Robert K. Yin (2014) Case Study Research Design and Methods (5th ed.). Thousand Oaks, CA:
538   Sage. 282 pages.
539 Rodriguez-Lizundia E, Marcos S Zalama (2015) A bellboy robot: Study of the effects of robot
540   behaviour on user engagement and comfort. Int J Human-Comp Stud 82:83–95
541 Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010) Collecting image annotations
542   using amazon's mechanical turk. In Proceedings of the NAACL HLT 2010 Workshop on
543   Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139-147.
544 R. C. Luo, Y. Hsu, Y. Wen and H. Ye, "Visual Image Caption Generation for Service Robotics
545   and Industrial Applications," (2019) IEEE International Conference on Industrial Cyber
546   Physical Systems (ICPS), Taipei, Taiwan, 2019, pp. 827-832, doi:
547   10.1109/ICPHYS.2019.8780171.
548 Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. (2014). Sequence to sequence learning with
549   neural networks. In NIPS, pp. 3104– 3112
550 Tang, X., Zhang, Q., & Hu, L. (2020). An EKF-based performance enhancement scheme for
551   stochastic nonlinear systems by dynamic set-point adjustment. IEEE Access, 8, 62261-62272.
552 Tremblay, J., To, T., Molchanov, A., Tyree, S., Kautz, J., & Birchfield, S. (2018, May).
553   Synthetically trained neural networks for learning human-readable plans from real-world
554   demonstrations. In 2018 IEEE International Conference on Robotics and Automation
555   (ICRA) (pp. 1-5). IEEE.
556 Trevarthen, C. (2001). Intrinsic motives for companionship in understanding: Their origin,
557   development, and significance for infant mental health. Infant Mental Health Journal:
558   Official Publication of The World Association for Infant Mental Health, 22(1‐2), 95-131.
559
560 Turkle, S. (2006). A nascent robotics culture: New complicities for companionship. American
561   Association for Artificial Intelligence Technical Report Series AAAI.
562 Vedantam, R., Lawrence Zitnick, C., & Parikh, D. Cider (2015) Consensus-based image
563   description evaluation. In Proceedings of the IEEE conference on computer vision and
564   pattern recognition pp. 4566-4575.
565 V. Mnih, N. Hees, A. Graves, K. Kavukcuoglu, (2014), Recurrent models of visual atten- tion,
566   NIPS
567 Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption
568   generator. In Proceedings of the IEEE conference on computer vision and pattern
569   recognition (pp. 3156-3164).
570 Vedantam, R., Lawrence Zitnick, C., & Parikh, D. Cider (2015) Consensus-based image.
571   description evaluation. In Proceedings of the IEEE conference on computer vision and
572   pattern recognition pp. 4566-4575.

573   Yang, J. & Chew, E. (2020) A Systematic Review for Service Humanoid Robotics Model in
574   Hospitality, Springer International Journal for Social Robotics.
575   https://doi.org/10.1007/s12369-020-00724-y (Impact Factor 3.168; indexed by JCSR &
576   Scopus, within the top 20 of Google Robotics publication)
577   Yang, J. & Chew, E. (2021) The Novel Design Model for Robotic Waitress, The International
578   Journal for Social Robotics (Impact Factor 3.168; WoS & Scopus-indexed, ranked within the
579   top 20 of the Top Publications Google Scholar citations)
580   Y. Tang, N. Srivastava, R.R. Salakhutdinov, (2014), Learning generative models with vi- sual
581   attention, in: NIPS, pp. 1808–1816.
582   Yamada, T., Murata, S., Arie, H., & Ogata, T. (2016). Dynamical integration of language and
583   behavior in a recurrent neural network for human–robot interaction. Frontiers in
584   neurorobotics, 10, 5.
585   Zhang, Q. C., Hu, L., & Gow, J. (2020). Output feedback stabilization for mimo semi-linear
586   stochastic systems with transient optimisation. International Journal of Automation and
587   Computing, 17(1), 83-95.

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: **The enhancement from HRI to HRC:** 1) **Robots have dynamic thinking and real-time neural image captioning ability: able to deal with**	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. One day	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. One day you suddenly fainted at home for some reason. The robot discovered your real-

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 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. . .	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).

Figure 1

HSR capabilities with the proposed high-level of HRC conceptual model

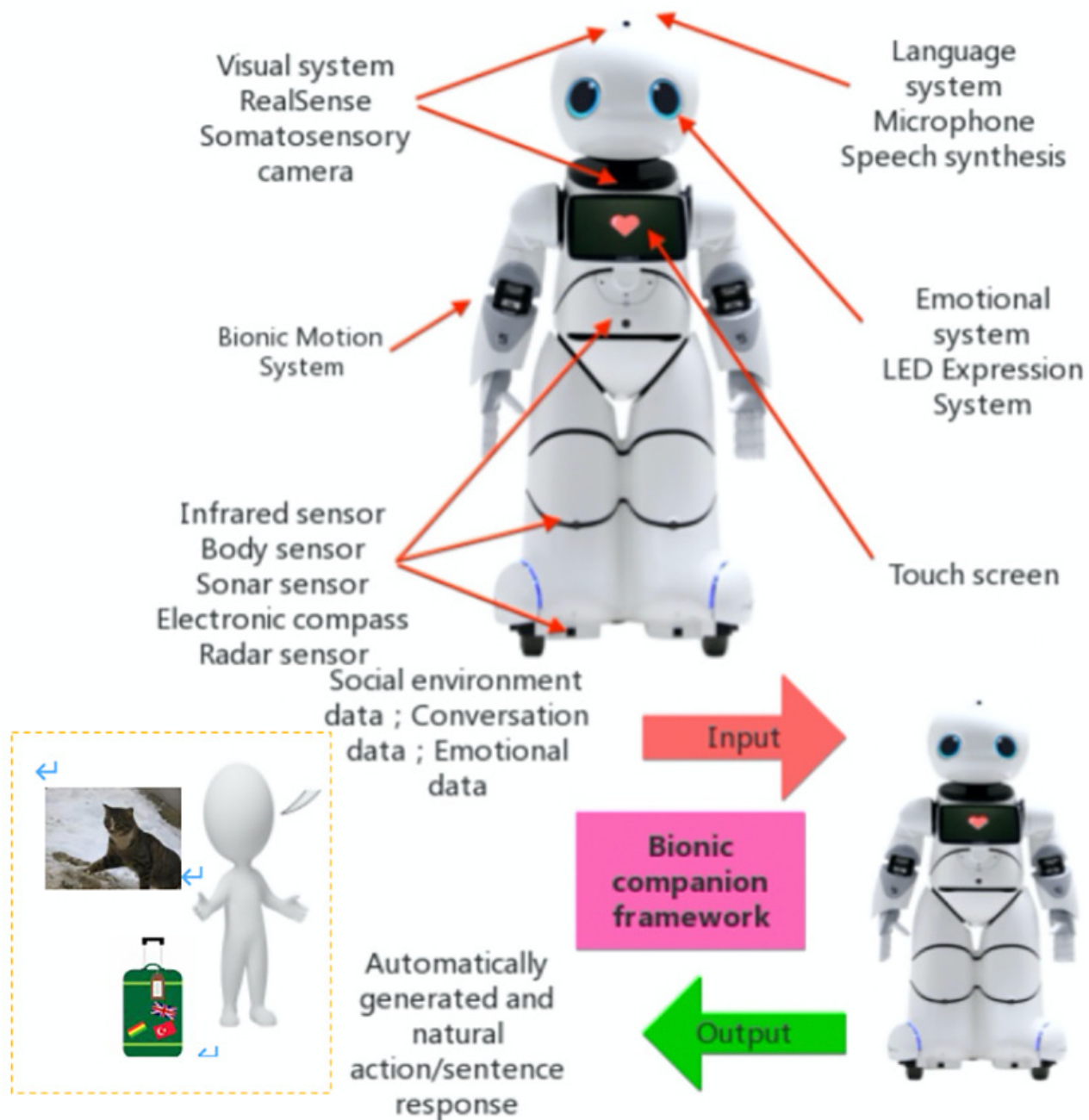



Figure 2

Bionic-companionship framework design

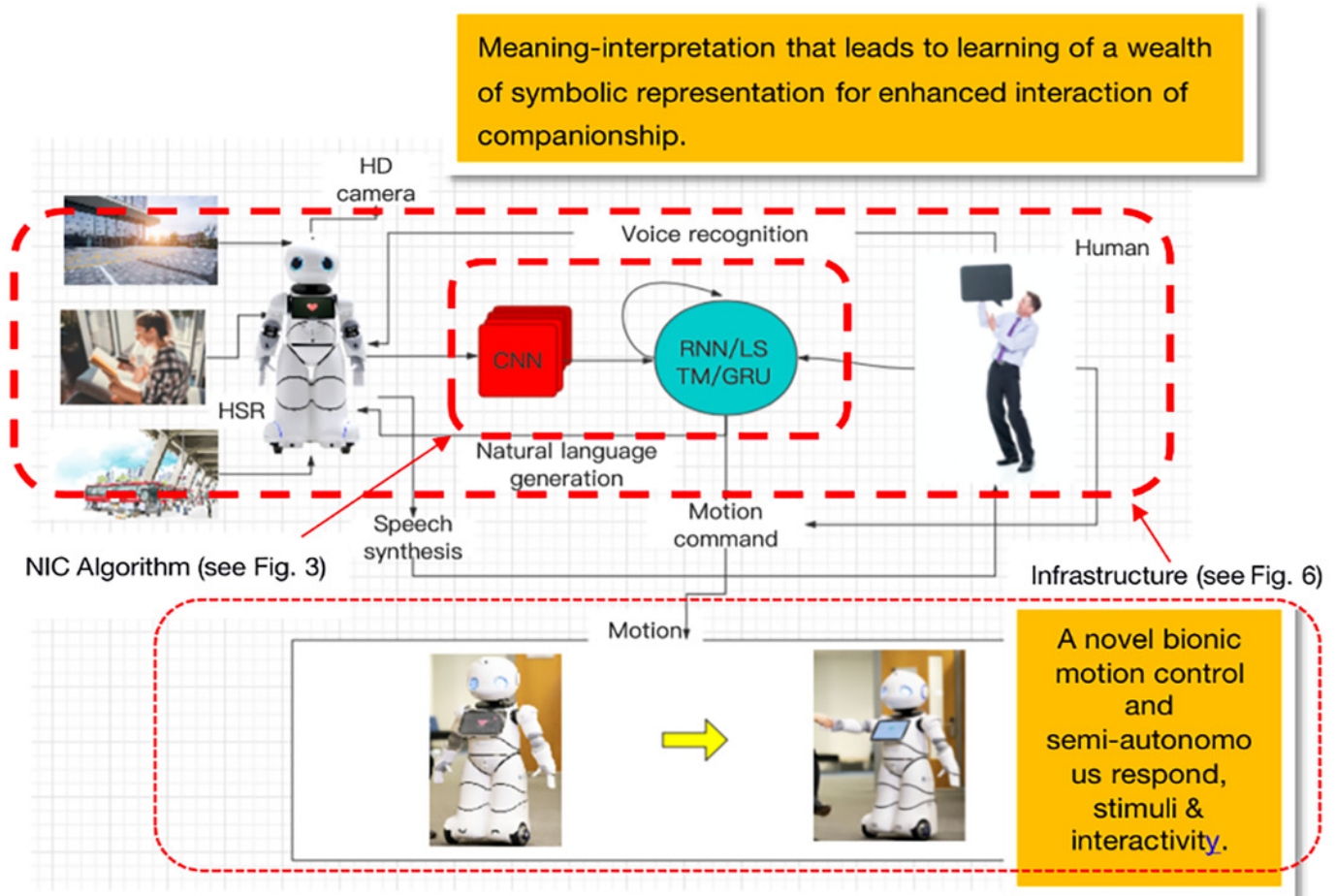



Figure 3

Neural image captioning model structure for HSR

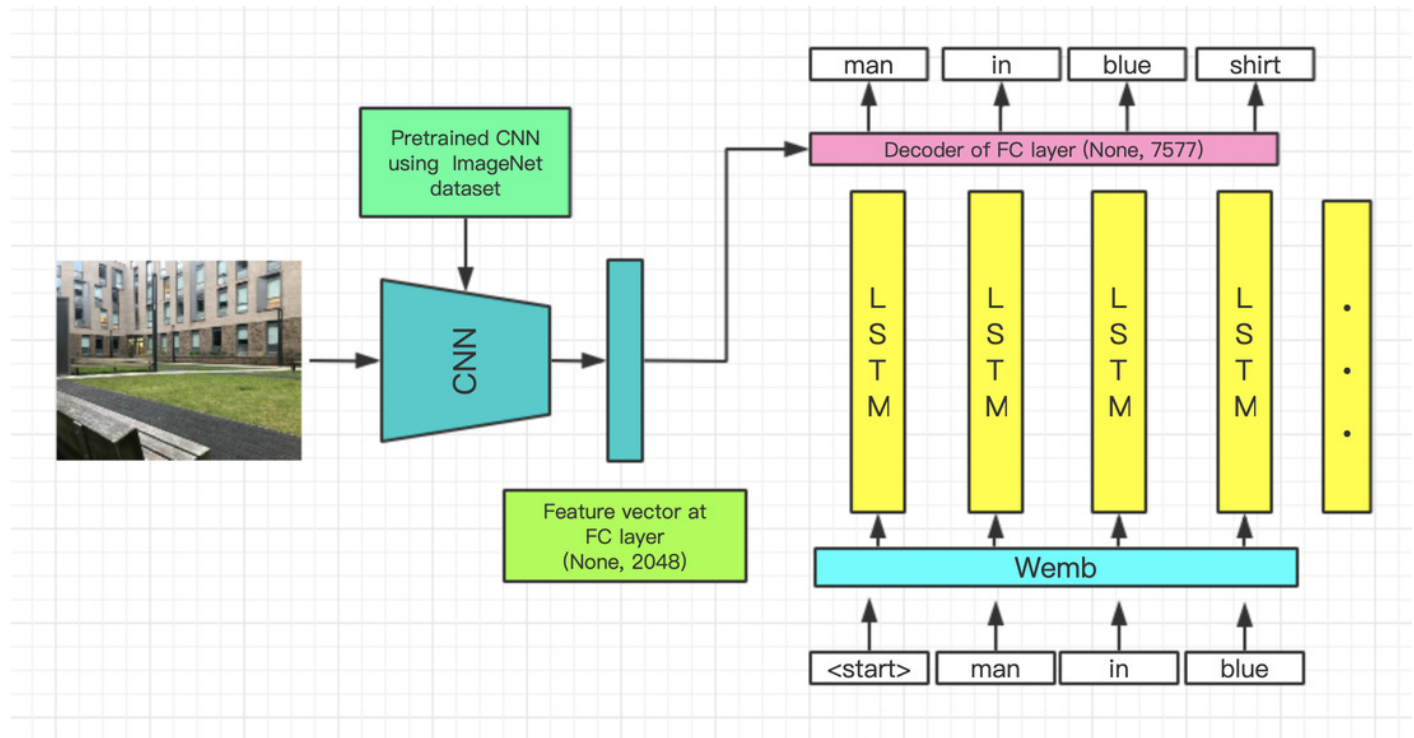



Figure 4

Samples Flickr 8k (Rashtchian et a.l, 2010) training data set

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 5

The Infrastructure of the Humanoid Service Robot Generating Neural Image Captions with as part of the Bionic Companionship Framework

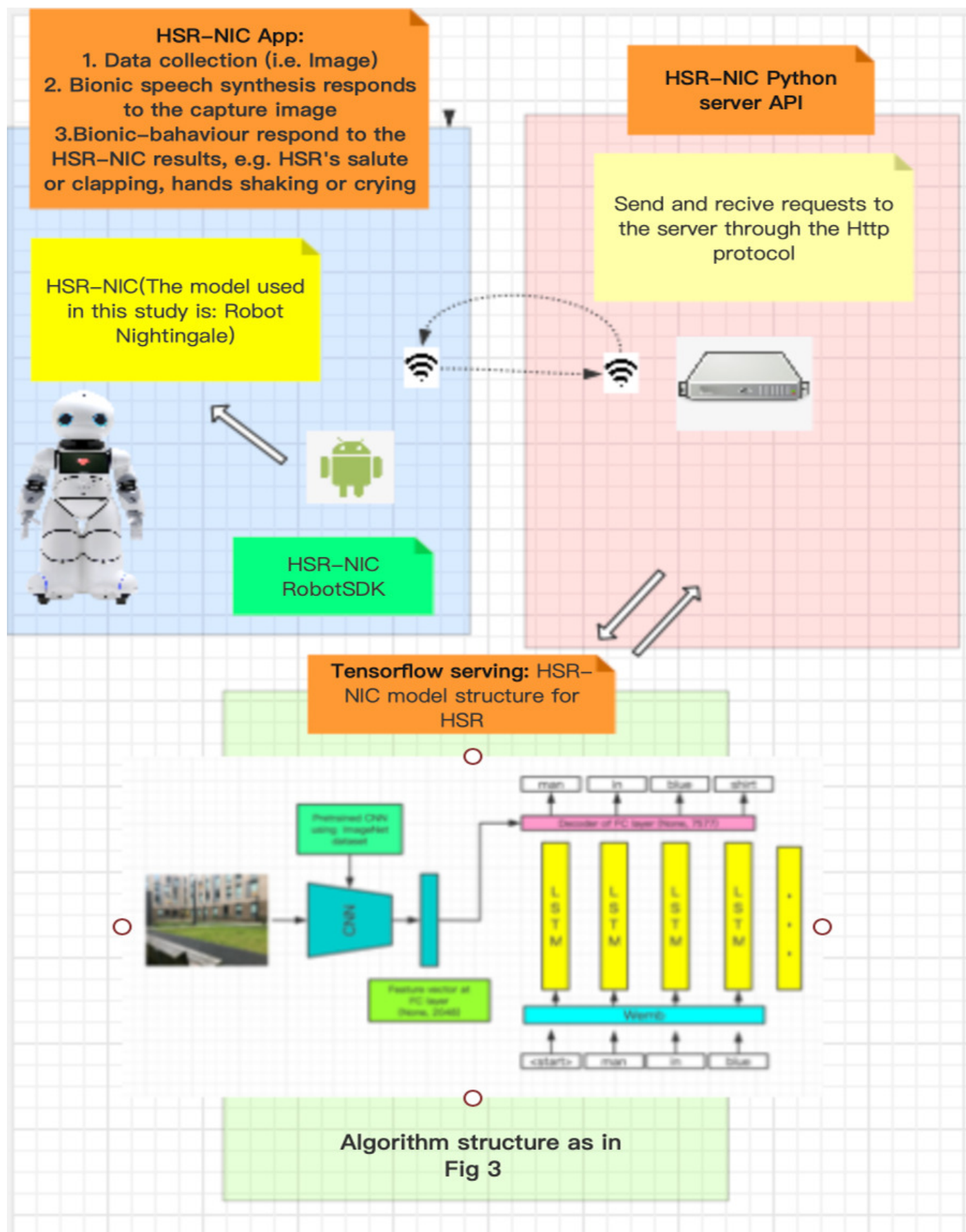



Figure 6

Loss curve of NIC model on training set and validation set

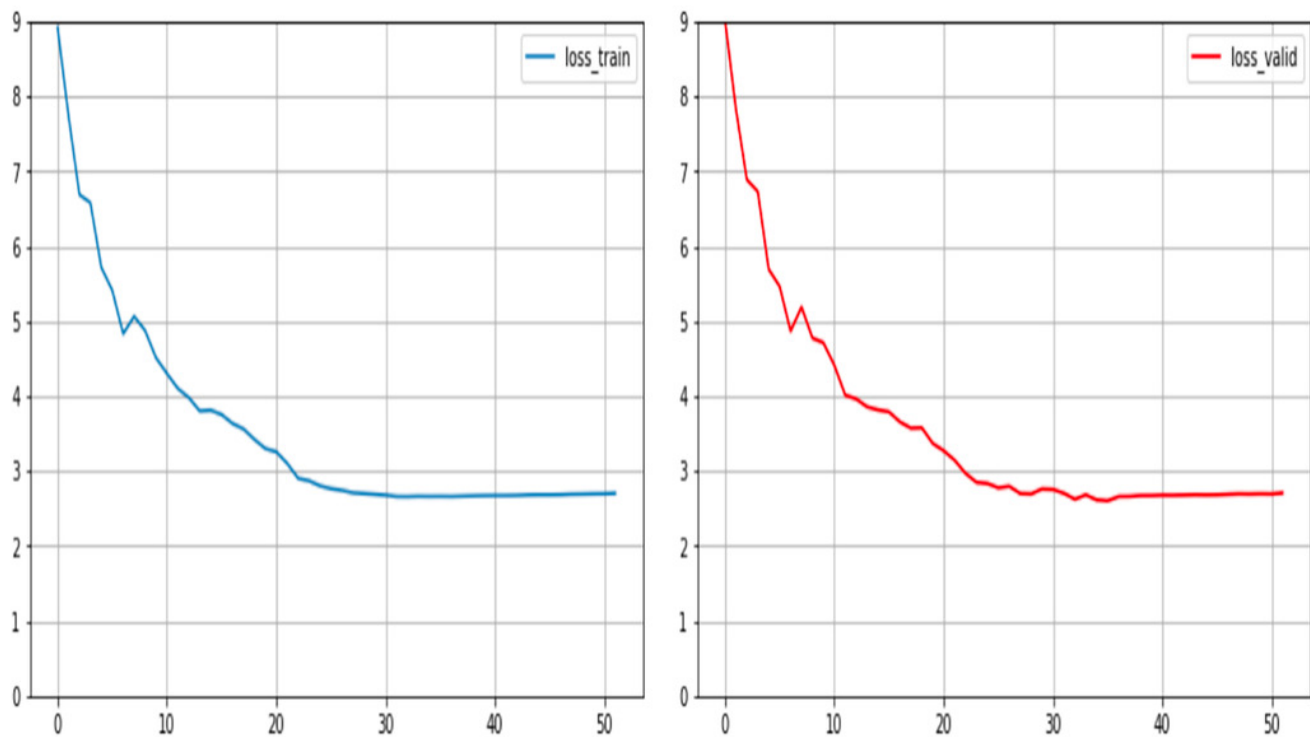
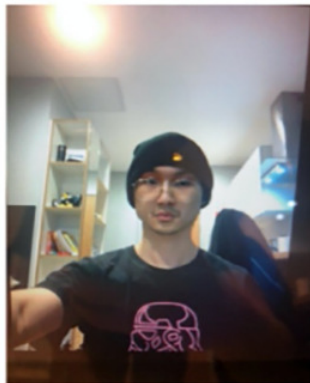
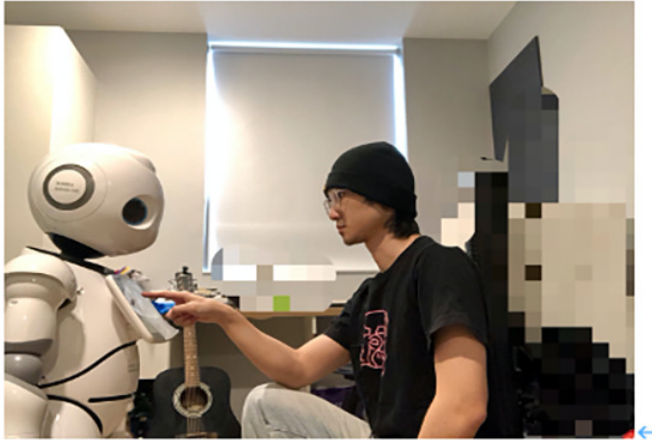
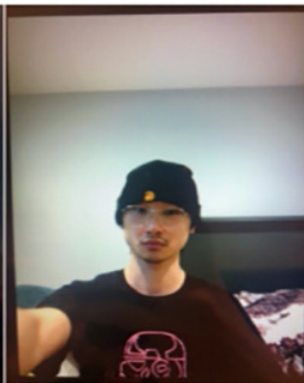
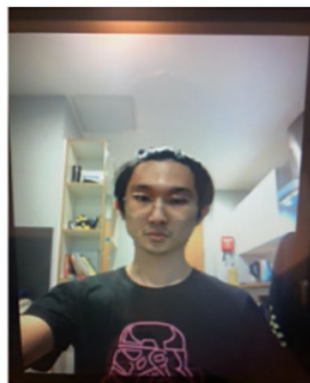



Figure 7

A Series of Preliminary Testing Results Captured from Canbot U05E and Bionic-Companionship Preliminary Framework

Man in black shirt and jeans is sitting on the ground with Animal

Man in black shirt and jeans is sitting on the ground

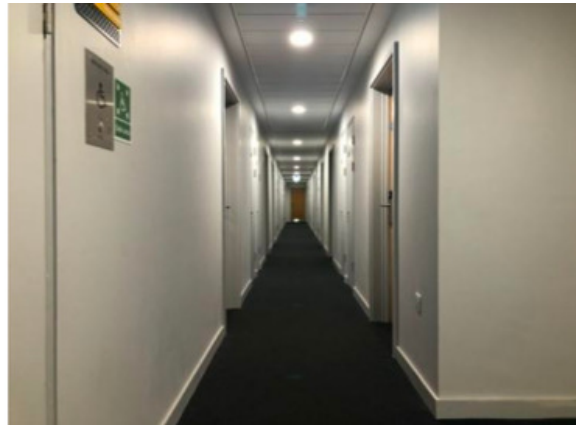
Man in black shirt and sunglasses is sitting on bench

Two girls are playing on the grass

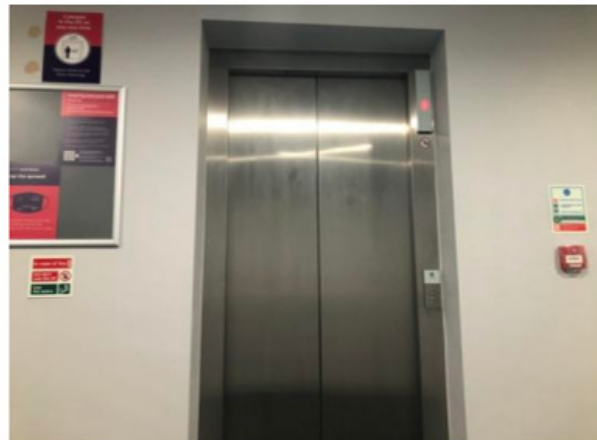
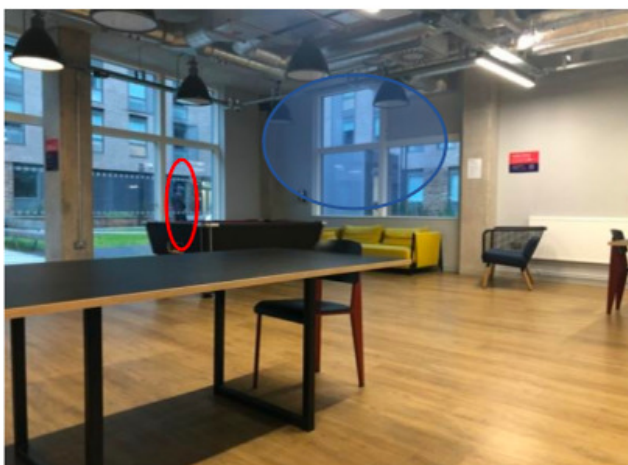
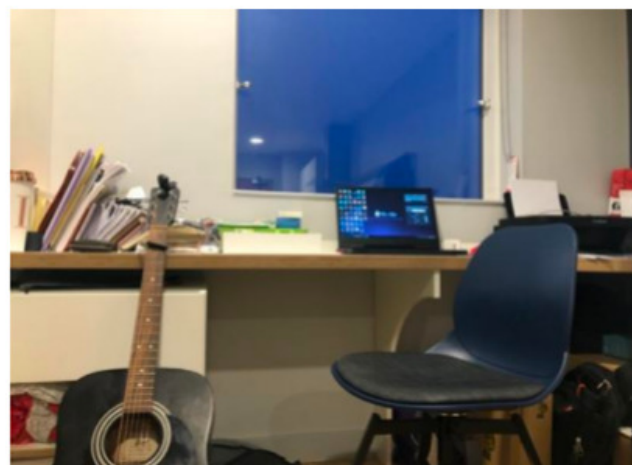
Two girls are sitting on bed laughing

Two children are playing on the grass

Figure 8

Real social environment test examples

man in helmet is sitting on the
sidewalk with traffic

man is sitting on bench with his feet in the air

man in blue shirt is sitting on bench
with his feet up in front of building

man sitting on the ground with his feet up in the air

man in black shirt is standing
in front of large building

man is standing in front of crowd