# Towards Environment Aware Social Robots using Visual Dialog

**Aalind Singh**[1*], **Manoj Ramanathan**[2], **Ranjan Satapathy**[2], **Nadia Magnenat-Thalmann**[2,3]
[1]Vellore Institute of Technology, India [3]MIRALab, University of Geneva
[2]Institute for Media Innovation, Nanyang Technological University

**Abstract**

State of the art social robots are limited in their ability to understand their environment and have a meaningful conversation based on it. Visual Dialog is a research field that combines computer vision and natural language processing techniques to achieve visual awareness. In this paper, we employ a Visual Dialog system in a humanoid social robot, Nadine, to improve its social interaction and reasoning capabilities. The Visual Dialog module consists of a neural encoder-decoder model, namely a memory network encoder and a discriminative decoder. The ability to carry out audio-visual scene-aware dialog with a user not only augments the human-robot interaction, but also lets the user to think of the robot of more than just a mere entity piece. Moreover, we also discuss the various applications and the psychological impacts this can have in the formation of human-robot bonds.

## Introduction

Figure 1: Nadine humanoid social robot.

Visual awareness and the ability to hold dialog, based on the understanding of the surrounding environment and previous conversation history is what lacks the most in the present day social robots (Gockley et al. 2005). Although there is an abundance of robots which can perform simple repetitive tasks based on natural language, but we still are far away from achieving human-like interaction in robots based on the changing environment around the user, taking into consideration events from the past, with all of this happening seamlessly in real time. With this work we want to pave way for intelligent social robots with these capabilities (Breazeal 2002),(S. Cross, Hortensius, and Wykowska 2019), which can interact or assist humans with applications such as helping elderly/blind, medical bots or as service workers (Vishwanath et al. 2019).

A theoretical framework for intelligent and expressive social robots was proposed in (Breazeal 2003) to carry out human-like behavior. Although a lot of research is being conducted on Visual Dialog (Das et al. 2017a), audio-visual scene-aware dialog systems (AlAmri et al. 2019), embodied question-answering (Das et al. 2018), but we are still lacking in developing actual robots who will be able to carry out these functions. In this paper, we incorporate Visual Dialog in a humanoid robot, Nadine[1], and then test it out under changing environment variables and different conversational settings. Our objective is to develop a fully aware anthropomorphic social robot which is able to interact and engage in a natural conversation with a human, taking into account the environment and events from the past. The robot's camera is used to provide images to perceive its environment and human-robot conversation history as text are combined to develop a Visual Dialog system. Due to this the robot will be able to hold a more natural and interesting dialog with the user that ultimately leads to a more satisfying human-robot interaction (HRI) experience (Breazeal 2004).

In the coming subsections we discuss in detail the work that has been done in the direction of achieving visually aware dialog and its applications in the field of social robotics.

### VQA and Visual Dialog

(Agrawal et al. 2017) proposed a model for Visual Question Answering (VQA), in which given an image and a natural language question about the image, a relevant answer was generated. The questions can encompass various regions of the image and can be related to the background or

[*] Work done while interning at Institute for Media Innovation, Nanyang Technological University, Singapore.

[1] https://en.wikipedia.org/wiki/Nadine_Social_Robot

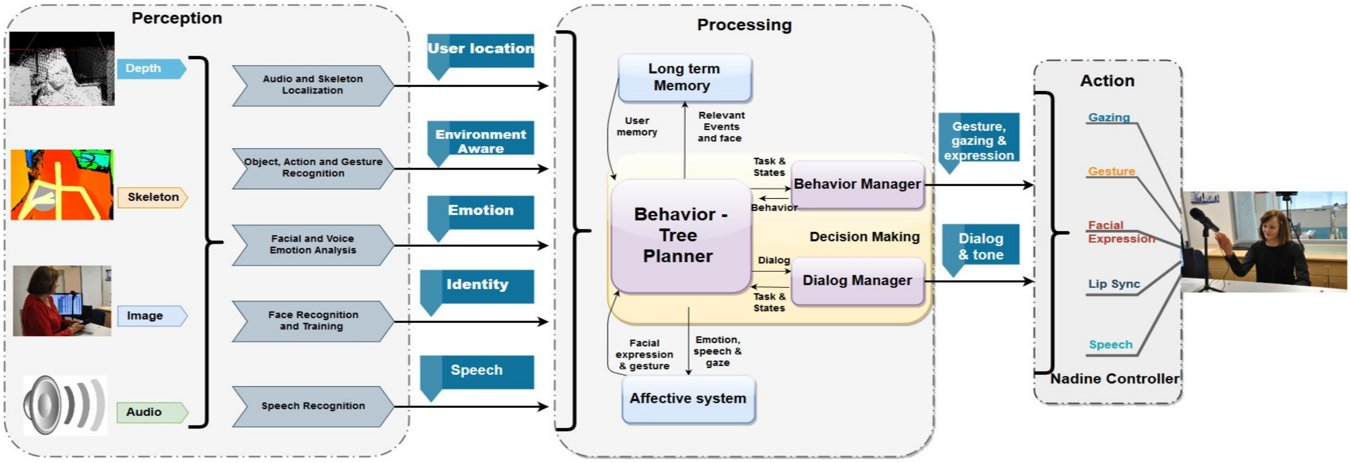


Figure 2: Humanoid Robot: Nadine's Architecture

the underlying context. Their model uses a two layer LSTM to encode the questions and VGGNet for encoding the images. This combined image and question embedding is then passed through a Multi-Layer Perceptron followed by a softmax layer to obtain a distribution over answers. Their model achieved state-of-art results on the VQA dataset proposed in the same paper. However the model had limitations, as it was able to generate only single word answers instead of natural complete sentences.

Das et al.(2017a) extended the above concept by also taking conversation history into account. So given an image, the dialog history and a question about the image, the model has to ground the question into the image, infer the context from history and answer the question accurately. In their paper, the authors also propose various novel encoder-decoder models for this task. Their model is capable of generating natural complete sentences in contrast to the previous VQA model. Moreover, richer models that incorporate deep-reinforcement learning have been explored by (Das et al. 2017b), in a later work.

Similar works have been performed by (De Vries et al. 2017), (Mostafazadeh et al. 2017) for achieving image-grounded conversations, with the former one even introducing a new two player guessing game dataset for the same. Moving forwards to a little bit practical aspect, embodied question answering Das et al.(2018) was introduced wherein a virtual agent can be asked to explore a 3D environment based on language commands. Although a lot of research is being conducted in these areas, we are still lagging behind in solving real-world problems encompassing these domains.

## Applications in Social Robotics

One of the practical applications of this work was shown in (Cho, Lee, and Kim 2017), which implemented a VQA network for answering questions related to the users environment and integrated it with a robotic head. The author used Dynamic Memory Networks (DMN) (Xiong, Merity, and Socher 2016), a deep learning network for VQA and a robotic head with three Degrees of Freedom (DOF) to achieve this task. The robot was tested out in a kitchen environment to answer normal questions such as (Q: *Where is the cup?*, A: *Shelf*) and it was able to correctly answer most of the questions in real time, with a lag of few seconds. Although the system engaged the user in a natural conversation flow, it had some limitations as it was only able to give single word answers (a typical VQA limitation) to most of the questions. Another limitation was that their model was trained to answer questions related to only kitchen environment.

# Proposed Framework

This section describes our framework to achieve Visual Dialog on Nadine social robot.

## Humanoid Platform

Nadine[1] (Ramanathan, Mishra, and Thalmann 2019) is a socially intelligent humanoid robot, and its architecture is based on a typical Perception-Decision-Action model as shown in Figure 2. The perception layer recognizes various cues from the surrounding environment using modules such as face recognition, gesture recognition, 3D hand pose estimation (Ge et al. 2017), and intent identification from the social situations. With regards to decision, a behaviour tree planner is used to trigger the other sub-modules such as emotion and memory models, as well as social attention coupled with a dialog manager. Finally, the action/interaction layer consists of a dedicated robot controller which includes emotion expression, lip synchronization, and gaze generation.

Nadine has a realistic human appearance and, a natural skin and hair. The humanoid has 27 degrees of freedom, making her able to coordinate body movements effectively (Magnenat-Thalmann et al. 2016) (Xiao et al. 2014). Nadine can also recognize people that she has met before and engage in a flowing conversation. Nadine can be seen as a part of human assistive technology (Magnenat-Thalmann and Zhang 2014), as she can assist people for continuous stretches without any intervals.

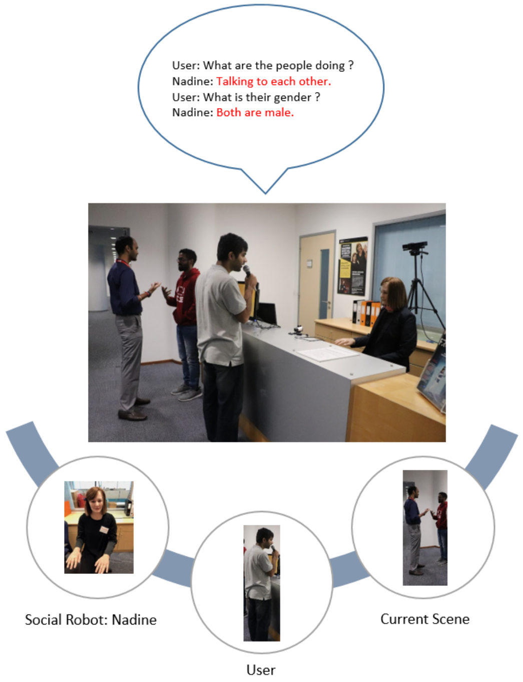

User: What are the people doing ?
Nadine: Talking to each other.
User: What is their gender ?
Nadine: Both are male.
Social Robot: Nadine
User
Current Scene

Figure 3: Experimental setup.

## Visual Dialog Model

This section details the dataset and the model used for achieving Visual Dialog in the social robot. The model is inspired from the work by Das et al.(2017a).

**Dataset** We use VisDial v1.0[2], which is a publicly available dataset containing 10-round dialogs on around 123$k$ training images, and 2$k$ validation images. This dataset was crowd-sourced on COCO (Lin et al. 2014) images and was constructed with a live chat interface, where two humans were asked to carry out dialog among themselves about a given image. One already knew about the image, while the other was asked to imagine the image given only the image caption and ask questions to the other human related to it.

**Model** In this section, we discuss about the module we used for Visual Dialog. The model uses a Memory Network (MN) encoder and a Discriminative decoder as described below.

**Memory Network Encoder**: The question is encoded using a LSTM and the image with a VGG-16 CNN. The representations are concatenated and are followed by a fully-connected layer and a $tanh$ non-linearity for obtaining a query vector. Each question-answer (QA) pair in the dialog history is separately encoded by an LSTM with shared weights. The query vector is then used to compute attention over the QA pairs using the inner product. The convex combination of attended history vectors is then passed through a fully connected layer and $tanh$ non-linearity successively and in the end the response is decoded.

**Discriminative Decoder**: It computes the dot product similarity between the input encoding and the LSTM encoding for each of the answer options following which the dot products are fed into a softmax to compute the posterior probabilities over the options.

### Environment Setting

As shown in Fig. 3 the environment consists of a social robot (Nadine), a user who is interacting with the robot, and current environment scene captured by the robot's camera.

During the conversation, the person is allowed to ask questions in third person about the environment to Nadine. This is due to design of dataset used for training described previously. The aim here is to test if the robot can comprehend the scene properly and engage in a natural flowing real time conversation with the person in complete sentences as shown in Fig. 4.

### Experiments

For capturing the real time visual data, we use a high dimensional 3D camera and the voice input is taken using a microphone. These inputs are obtained from perception layer of the robot architecture shown in Fig 2. The Visual Dialog module is incorporated as a part of the processing layer, where it can process the above inputs to generate dialog. Finally the action layer outputs the generated dialog, coupling it with a suitable facial expression, proper gaze and gesture.

The robot can give replies based on different environment settings as can be seen in Fig 4, where three different situations have been depicted along with a small human-robot conversation snippet based on them. The model's mean reciprocal rank (MRR)[3] was approximately 0.621, so it can be expected that for some questions the robot's replies will be a bit vague or incorrect as can be seen in Fig 4 (Human: *What is he doing?*, Nadine: *He is reading a book.*).

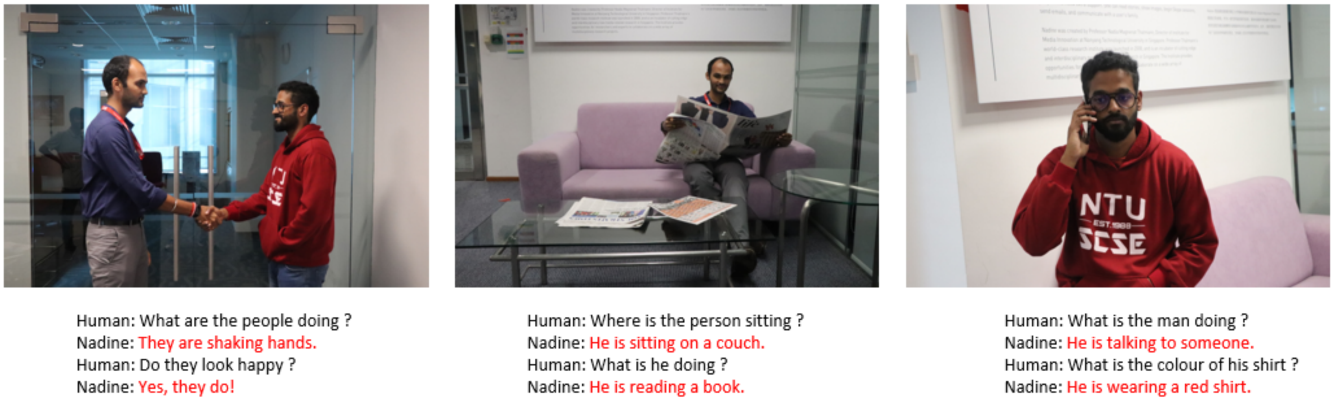
## Applications

Our experiment of having a social robot to carry out a vision based communication lays grounds to various applications, some of which are as listed below:

- Elderly/Blind Companion: Assist visually impaired users, elderly people in understanding their surroundings better. These robots (Broekens, Heerink, and Rosendal 2009) can serve as companions guiding people whenever necessary. Eg: (Human: *How many people are standing there?*, Robot: *I can see two people standing.*)
- Medical Psychotherapy: In the near future, robots with an accurate understanding of the environment can be used in clinical psychotherapy or other health applications (Breazeal 2011). Eg. treating patients with psychotic ailments such as *Schizophrenia, Dementia*, which cause effects such as hallucinations, visual delusions in normal people.

[2] https://visualdialog.org/data

[3] Evaluation metrics for VisDial dataset defined in Das et al.(2017a)



Human: What are the people doing ?
Nadine: They are shaking hands.
Human: Do they look happy ?
Nadine: Yes, they do!

Human: Where is the person sitting ?
Nadine: He is sitting on a couch.
Human: What is he doing ?
Nadine: He is reading a book.

Human: What is the man doing ?
Nadine: He is talking to someone.
Human: What is the colour of his shirt ?
Nadine: He is wearing a red shirt.

Figure 4: Conversation examples.

- Teleoperation: Major use-case in situations where the user operates the robot remotely using language commands. Eg: Fire fighter robots in search & rescue (SAR) operations (Human: *Do you see smoke coming out of any room?*, Robot: *Yes, from one room.*, Human: *Go there and look for people.*)

## Conclusion

This work presents an interesting application, where we extend a social robot's capabilities by allowing it to carry out a conversation with a human about visual content, a novel application still missing in most of the current social robots. We are still working on this project to make the framework much more better and to integrate it seamlessly for making it adaptable to different set of applications. There are some limitations to this work, such as the robot may sometimes give vague or incorrect replies which can have a psychological effect on the user. This is due to uncanny valley effect (Mori, MacDorman, and Kageki 2012), in which people expect humanoid robots to behave and come up with accurate natural replies in all situations, just like normal human beings. One possible solution is for the user to provide the correct answers when robot makes a mistake and then repeat the interaction process as the model can learn from previous conversation. Later on, the model can also be re-trained based on these errors. As indicated above, this kind of HRI will open a new avenue of potential applications that would require models to handle limitations identified here.

## ACKNOWLEDGMENT

This research is supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

## References

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. Vqa: Visual question answering. *Int. J. Comput. Vision* 123(1):4–31.

AlAmri, H.; Cartillier, V.; Das, A.; Wang, J.; Lee, S.; Anderson, P.; Essa, I.; Parikh, D.; Batra, D.; Cherian, A.; Marks, T. K.; and Hori, C. 2019. Audio-visual scene-aware dialog. *CoRR* abs/1901.09107.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge, MA, USA: MIT Press.

Breazeal, C. 2003. Emotion and sociable humanoid robots. *Int. J. Hum.-Comput. Stud.* 59(1-2):119–155.

Breazeal, C. 2004. Social interactions in hri: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34(2):181–186.

Breazeal, C. 2011. Social robots for health applications. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5368–5371.

Broekens, J.; Heerink, M.; and Rosendal, H. 2009. Assistive social robots in elderly care: A review. *Gerontechnology* 8:94–103.

Cho, S.; Lee, W.; and Kim, J. 2017. Implementation of human-robot vqa interaction system with dynamic memory networks. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 495–500.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017a. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1080–1089.

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2951–2960.

Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

De Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5503–5512.

Ge, L.; Liang, H.; Yuan, J.; and Thalmann, D. 2017. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5679–5688.

Gockley, R.; Bruce, A.; Forlizzi, J.; Michalowski, M.; Mundell, A.; Rosenthal, S.; Sellner, B.; Simmons, R.; Snipes, K.; Schultz, A. C.; and Jue Wang. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1338–1343.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. *CoRR* abs/1405.0312.

Magnenat-Thalmann, N., and Zhang, Z. 2014. Social robots and virtual humans as assistive tools for improving our quality of life. In *2014 5th International Conference on Digital Home*, 1–7.

Magnenat-Thalmann, N.; Yuan, J.; Thalmann, D.; and You, B., eds. 2016. *Context Aware Human-Robot and Human-Agent Interaction*. Human-Computer Interaction Series. Springer.

Mori, M.; MacDorman, K. F.; and Kageki, N. 2012. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine* 19(2):98–100.

Mostafazadeh, N.; Brockett, C.; Dolan, B.; Galley, M.; Gao, J.; Spithourakis, G.; and Vanderwende, L. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 462–472. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Ramanathan, M.; Mishra, N.; and Thalmann, N. M. 2019. Nadine Humanoid Social Robotics Platform. In *In: Gavrilova M., Chang J., Thalmann N., Hitzer E., Ishikawa H. (eds) Advances in Computer Graphics. CGI 2019. Lecture Notes in Computer Science, vol 11542. Springer, Cham*. Springer.

S. Cross, E.; Hortensius, R.; and Wykowska, A. 2019. From social brains to social robots: applying neurocognitive insights to human-robot interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 374.

Vishwanath, A.; Singh, A.; Victoria, C. Y. H.; Dauwels, J.; and Magnenat-Thalmann, N. 2019. Humanoid co-workers: How is it like to work with a robot?

Xiao, Y.; Zhang, Z.; Beck, A.; Yuan, J.; and Thalmann, D. 2014. Human-robot interaction by understanding upper body gestures. *Presence: Teleoper. Virtual Environ.* 23(2):133–154.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2397–2406.