Face Analysis and Body Language Understanding from Egocentric Cameras

Ronja Möller[1], Antonino Furnari[1], Sebastiano Battiato[1], Aki Härmä[2], and Giovanni Maria Farinella[1]

[1] University of Catania, V. Andrea Doria, 95125 Catania, Italy
ronja.moller@unict.it, furnari@dmi.unict.it
battiato@dmi.unict.it, gfarinella@dmi.unict.it
[2] Philips Research, High Tech Campus 34, 5656 North Brabant, Netherlands
aki.harma@philips.com

Abstract. The objective of the project described in this position paper is to develop and evaluate algorithms that enable a mobile agent, e.g., a robot, to observe a user during his/her day to day activities and infer relevant information which could help improve human-machine interaction. To achieve this goal we will first explore intelligent navigation strategies. The overall focus will be on visual data, analysing the user's action, face and body language. Once the algorithms run on the robot, they can be used to log user activity/emotional states and support them during daily activities. The collected information of the users will be useful for further analysis by healthcare professionals or assistive applications. In addition to the mentioned domains, attention will also be paid to speech analysis and synthesis to ensure natural interaction with the user. The algorithms will be able to infer age, gender, emotions, activity and body language of the user. Lastly, information obtained by First Person Vision Systems worn by a user will be considered as an external source of data to make more accurate inferences and explore possible correlations.

Keywords: Machine Learning · Computer Vision · Human Robot Interaction

1 Motivation

In recent years, the academic community has seen rapid advancements in the fields of machine learning, computer vision and human-robot-interaction. Deep Learning approaches, in particular, have proven to be very effective at solving tasks like object classification and detection in the image domain, object tracking, pose estimation and action recognition in the video domain, and speech processing and synthesis in the domain of sound [16]. In parallel to this development the capabilities of corresponding hardware in terms of the volume of data that can be handled as well as the speed at which it can be processed have increased drastically.

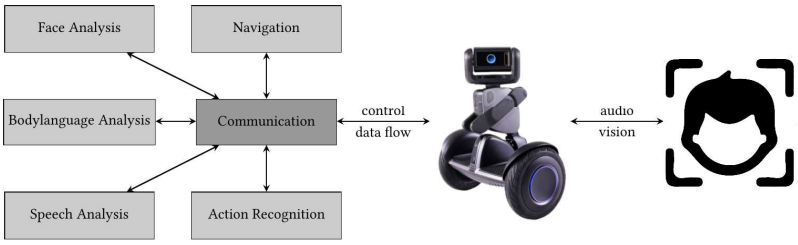

Fig. 1. Overview of the proposed architecture of our demonstrator.

With the widespread availability of AI-capable hardware and software we are seeing a growing acceptance of technology in day to day scenarios: starting from Smartphones, Smartwatches via Amazon Echo, Google Home to autonomous cars. This opens up new possibilities in entertainment, education and organisation but also in healthcare applications. Since there are various moral and legal concerns associated with indiscriminately collecting medical data, we see a lot of potential in tailor-made solutions for assistive (see also [17] and [19]) and natal care that are built upon personalised data acquisition.
Given the increasing average age of citizens in developed countries in particular, it is only a matter of time before certain aspects of healthcare will be automated - if they haven't been already. To minimize stress, inconvenience and biased behaviour of customers a personalised robot that can operate at home is an appealing choice. In addition, it doubles as a perfect test and demonstration setup. To meet the previously outlined demand, we will develop algorithms for an egocentric agent (in our case the segway Loomo [3]) that can interact with the user in a useful way and unobtrusively collect relevant data.

2 Related Work

2.1 Robotic Navigation

Navigation in robotics is a popular and well-explored field. A variety of solutions have been proposed depending on the sensors of the robot, the environment and the given task [5], [11], [15], [22], [24], [29] . Our robot will operate in indoor environments. It will operate alone and our algorithms will mostly make use of the visual sensors of the Loomo robot. Vision-based navigation is not a new topic but it tends to be more complex than approaches that utilise only distance sensors or GPS. This complexity is further compounded by the fact that our robot has to operate in mapless settings. Due to the high data demand of complex (deep) vision-based navigation, approaches will have to utilise simulated environments and use real data mostly for fine-tuning and qualitative testing.

[3] http://eu-en.segway.com/products/have-something-that-does-anything

The existing bigger platforms for robotic navigation (AI Habitat [4], AI2Thor,[5], Matterport,[6], Gibson,[7]) are well-documented and -implemented but they tend to focus on navigation without humans in the simulated scene.

2.2 Human Robot Interaction

Evaluating the current state of the art in Human-Robot-Interactions (HRI), it is apparent that there is a big interest in natural everyday interactions but the focus tends to be on social aspects and functional interaction tends to be a priority only for specific industrial use cases [1], [12], [23], [28]. In addition, HRI publications often use an abstract representation of the state of the robot and the human (for example only positions in a grid of 1x1 m grid and similarly sparse camera angles), resulting in solutions that are either impossible to implement in the real world or need a continuous transformation of the real world into this abstracted state.

2.3 Activity/Action Recognition and Anticipation

Activity and Action Recognition have been explored in a multitude of ways. Existing approaches utilise single RGB frames, videos, depth information, IMU data or more abstract representations that can be extracted from those modalities such as joint positions [10], [20], [27]. Simple RGB videos are favoured because recording depth and IMU information is still vastly more expensive and time consuming than just recording a video. Moreover, only a limited number of devices are able to collect all the different modalities synchronised.

With the recent advances in the area of Deep Learning for object detection and recognition and the demonstrated usefulness of recurrent network architectures for detecting temporal events, it is a natural progression to apply Deep Recurrent Networks to Activity and Action Recognition. An extension to the Action Recognition problem has recently been proposed - analogous to how humans operate by reasoning about future events - which consists in training networks to anticipate actions [8].

2.4 Face, Body and Speech Analysis

The task of detecting, re-identifying [6] and analysing faces has long been of academic interest and we will draw on a multitude of approaches for this project such as emotion estimation [2], [13] which will be exploited to gauge the user's emotional state.

With the availability of reliable Human Pose Estimation algorithms, the community has proposed several promising approaches to Body Language Analysis

[4] https://aihabitat.org/

[5] https://ai2thor.allenai.org/

[6] https://niessner.github.io/Matterport/

[7] http://svl.stanford.edu/gibson2/

such as the ones introduced in [3], [9], [26]. A more general overview can be found in [21].
Speech Analysis, like the two other topics, is a popular and well-researched area that we plan on exploiting for our demonstrator. In addition to the multitude of solutions for generalised Speech Recognition, we plan on exploring approaches that focus on emotion detection and healthcare, such as [18] and [25].

2.5 Approaches using the Loomo Platform

Using the Loomo robot as platform for Machine Learning applications has been explored in other settings such as socially aware navigation [4] and as an assistant for the elderly or disabled [8]. This encourages us to explore the platform further.

3 Architecture

In this section, we will give a brief overview of the general structure of our proposed demonstrator. The architecture is designed to be very modular to facilitate the development of semi-independent components that can easily be exchanged for better alternatives during the course of the project as can be seen in Figure 1.

3.1 Software Modules

The core functionalities of the software will be described in the following paragraphs. As previously discussed, the final goal of the project is to create an agent that can collect useful information from body language, facial information and speech. Since the robot is a mobile agent, we first have to solve the task of finding the best possible way to position (and continuously reposition) the robot in order to create the best conditions for information gathering. This positioning task is closely related to the action recognition task since different actions will influence the "best" position for the robot.
All the modules will naturally interact with each other and all information streams will be used to improve the performance of the individual tasks.

Intelligent Navigation In addition to tackling standard navigation tasks such as obstacle avoidance and planning, our goal is to enable the robot to interact with humans in a way that is efficient and intelligent. It should not obstruct the user's path or inconvenience them, all the while maintaining a position that is optimal for observing the user to gain useful insights. We will explore useful behaviour policies to ensure that the robot continuously maximises the information that can be gained.

[8] https://davidgollasch.com/a3bot-project/

Action Recognition The robot will be able to infer what the user is doing and when possible predict what they are going to in order to log user activity and if necessary assist or interfere. Depending on the type of action that is detected a decision has to be made about where the optimal position for the robot is and if it is even feasible to reach as well as maintain it with the constraints that the robot faces in the wild.
This module will utilize joint positions as well as body and face orientation and interact naturally with the navigation model and vice versa. Further work will also explore contextual cues and information about objects that are present in the scene.

Body Language Analysis As an extension of the action recognition task that will utilize the position of the user's joints to infer what they are doing, this module will evaluate the social cues that can be taken from the user's body language. In other words, the Action Recognition module will determine what the user is doing and this module will evaluate how the user is feeling. The robot will observe body language to infer emotional states, attention levels or other relevant information such as deviation from baseline movement.

Face Analysis As an extension of the body language analysis soft biometrics like age/gender/mood can be used to further refine the previously extracted information about the user's emotional state. This information can also be used to inform other models when applicable, for example to include heuristics about age-specific actions. Face Re-Identification [7] could be also considered as priors to understand user behaviour.

Speech Analysis The robot provides the options of both processing and synthesizing speech and, given its mobile nature and the small size of its display, this will be the main avenue of communication between the user and the robot. Since the target demographic for this demonstrator explicitly includes elderly and disabled persons, ensuring natural and easy interaction with the robot is a must - therefore a speech analysis module will be implemented.

3.2 Hardware

We will be using the segway Loomo robot to run our algorithms. The robot allows for a certain amount of computation onboard but we plan on running some functionalities on an external processing unit that will also handle the logging/saving of the gathered information.

Segway Loomo Robot The Loomo Segway is a personal robot built with the goal to be mobile and autonomous, it can be used both indoors and outdoors and in almost all types of terrain. Loomo has many features that are of interest to us, such as a fish-eye camera, Intel RealSense, ultrasonic sensors, infrared distance

sensors, touch sensors, encoders and IMUs that can be used for algorithms that allow the robot to recognize objects, faces and voices. Loomo also comes with an Android API for development.

External Processing Unit The Loomo platform provides a certain amount of diskspace (50GB) and processing power but for the more data intensive applications that we have planned, we will need an external processing unit. In the demonstrator phase that this project will cover this will be most likely a PC that communicates wirelessly with the robot.

4 Our Plan

In this section we will outline our plan for the next steps that will be taken in the project. As described in Section 3.1 we aim to first solve the navigation task and the action recognition task before implementing the modules for the analysis of facial information, body language and speech.

4.1 Navigation Task

Following the current state of the art approaches for navigation such as the one proposed in [14] we will employ Deep Reinforcement Learning (DRL) to learn useful behaviour. We will draw inspiration from existing solutions and plan on exploring popular architectures and training strategies.
DRL has the same drawback that all Deep Learning approaches share, namely the necessity of a dataset that is big and varied enough to realistically represent the environment the robot will operate in. An ongoing topic of interest in the Reinforcement Learning community is the possibility of training behaviour in a simulated environment and then transferring to the real world. This transformation can vary a lot in complexity depending on the chosen simulation. In our case, it is impossible to collect enough data both in terms of amount as well as in terms of variation. We cannot limit ourselves to real data, so simulations will have to be explored.

4.2 Activity Recognition Task

Going from popular current approaches such as the ones introduced in Section 2.3, we will start with using basic joint information to infer current and possible future actions and gradually extend our algorithms to also include facial information and gestures. Furthermore, we plan on including information about relevant objects and the scene to achieve a holistic solution that takes all possible streams of information into account. In addition, our focus will be on exploiting the mobile nature of the robot since this is a feature that sets our approach apart from existing solutions.

5 Conclusion

We have outlined the main goals of our project, some related work that already exists in the field and how we plan on implementing the features of our demonstrator. Considering the trends in related fields, the recent developments in terms of algorithms and hardware as well as the general social climate, we believe that it is the right time to attempt this ambitious project and we are optimistic about our chances of success. Our demonstrator will naturally fit into the progression towards automated and personalised assistive and healthcare technologies and serve as a testbed for further developments.

Acknowledgements

This research is part of the PhilHumans[9] project supported by Marie Sklodowska-Curie Innovative Training Networks - European Industrial Doctorates.

References

1. Akalin, N., Kiselev, A., Kristoffersson, A., Loutfi, A.: Enhancing social human-robot interaction with deep reinforcement learning. pp. 48–50 (07 2018). https://doi.org/10.21437/AI-MHRI.2018-12
2. Awiszus, M., Grasshof, S., Kuhnke, F., Ostermann, J.: Unsupervised features for facial expression intensity estimation over time. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1199–11998 (2018)
3. Bhattacharya, U., Roncal, C., Mittal, T., Chandra, R., Bera, A., Manocha, D.: Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping (2019)
4. Chen, C., Liu, Y., Kreiss, S., Alahi, A.: Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning (2018)
5. Chen, X., Ghadirzadeh, A., Folkesson, J., Jensfelt, P.: Deep reinforcement learning to acquire navigation skills for wheel-legged robots in complex environments (2018)
6. Farinella, G.M., Farioli, G., Battiato, S., Leonardi, S., Gallo, G.: Face re-identification for digital signage applications. vol. 8811, pp. 40–52 (08 2014)
7. Farinella, G., Farioli, G., Battiato, S., Leonardi, S., Gallo, G.: Face re-identification for digital signage applications. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8811**, 40–52 (2014)
8. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention (2019)
9. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture (2019)
10. Hosseini, B., Montagne, R., Hammer, B.: Deep-aligned convolutional neural network for skeleton-based action recognition and segmentation (2019)
11. Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., Hutter, M.: Learning agile and dynamic motor skills for legged robots. Science Robotics **4**(26), eaau5872 (Jan 2019). https://doi.org/10.1126/scirobotics.aau5872

[9] http://www.philhumans.eu

12. Jiang, Y., Yang, F., Zhang, S., Stone, P.: Integrating task-motion planning with reinforcement learning for robust decision making in mobile robots (2018)
13. Kollias, D., Sharmanska, V., Zafeiriou, S.: Face behavior à la carte: Expressions, affect and action units in a single network (2019)
14. Kulhanek, J., Derner, E., de Bruin, T., Babuska, R.: Vision-based navigation using deep reinforcement learning. 2019 European Conference on Mobile Robots (ECMR) (Sep 2019). https://doi.org/10.1109/ecmr.2019.8870964
15. Kutbi, M., Chang, Y., Sun, B., Mordohai, P.: Learning to navigate robotic wheelchairs from demonstration: Is training in simulation viable? In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–44 (05 2015). https://doi.org/10.1038/nature14539
17. Leo, M., Furnari, A., Medioni, G.G., Trivedi, M., Farinella, G.M.: Deep learning for assistive computer vision. In: The European Conference on Computer Vision (ECCV) Workshops (September 2018)
18. Lian, Z., Li, Y., Tao, J., Huang, J.: Speech emotion recognition via contrastive loss under siamese networks. Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data - ASMMC-MMAC'18 (2018). https://doi.org/10.1145/3267935.3267946
19. Marco, L., Farinella, G.M.: Computer Vsision for Assistive Healthcare. Academic Press (05 2018)
20. Ng, Y.B., Fernando, B.: Human action sequence classification (2019)
21. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition (2018)
22. Paraicu, I., Leordeanu, M.: Learning navigation by visual localization and trajectory prediction (2019)
23. Qureshi, A.H., Nakamura, Y., Yoshikawa, Y., Ishiguro, H.: Robot gains social intelligence through multimodal deep reinforcement learning (2017)
24. Saeed, R., Recupero, D.R., Remagnino, P.: A boundary node method for path planning of mobile robots. Robotics and Autonomous Systems **123**, 103320 (2020)
25. Salekin, M.S., Zamzmi, G., Paul, R., Goldgof, D., Kasturi, R., Ho, T., Sun, Y.: Harnessing the power of deep learning methods in healthcare: Neonatal pain assessment from crying sound (2019)
26. Sun, M., Mou, Y., Xie, H., Xia, M., Wong, M., Ma, X.: Estimating emotional intensity from body poses for human-robot interaction (2019)
27. Tadesse, G.A., Cavallaro, A.: Visual features for ego-centric activity recognition: a survey. In: WearSys '18 (2018)
28. Thabet, M., Patacchiola, M., Cangelosi, A.: Sample-efficient deep reinforcement learning with imaginary rollouts for human-robot interaction (2019)
29. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning (2016)