Video action recognition and prediction architecture for a robotic coach*

Nino Cauli[1][0000−0002−9611−0655] and Diego Reforgiato Recupero[2][0000−0001−8646−6183]

[1] University of Catania, Piazza Università, 2, Catania, 95131, Italy
nino.cauli@unict.it
[2] HRI Lab, University of Cagliari, Via Ospedale, 72, Cagliari, 09124, Italy
diego.reforgiato@unica.it

Abstract. In this paper we introduce a novel architecture to recognise and to predict human actions from video sequences. Specifically, this architecture will be part of a larger system meant to promote elders' active ageing. The system will consist of a robotic coach able to schedule daily exercises, listening to patients' requests, monitoring the exercises, and correcting the errors in the execution. Using a monocular RGB camera video stream as input, the proposed architecture will be able to recognise the movement performed by the elder and to predict the next expected visual (camera frames) and proprioceptive (encoders) sensory inputs. In order to keep track of past frames, a Convolutional Neural Network (CNN) with both standard and recurrent convolutional layers (ConvLSTM or ConvGRU) has been chosen. Based on the Predictive Coding paradigm, the network will recognise the actions and predict the future visuo-proprioceptive stimuli using a single architecture. The full robotic coach system will be implemented on an affordable humanoid robot, the NAO.

Keywords: Action prediction · Recurrent neural networks · Deep learning · Predictive coding · Human robot interaction · Active ageing.

1 Introduction

The fast increase in percentage of elder population is an important issue of modern society[3]. This older slice of the population needs daily assistance and monitoring in order to live a safe and productive life. Several researchers focused their work on assistive technologies and monitoring systems for elders. Technologies promoting active ageing appear to be the right solution to increase

* The research leading to these results has received funding from the EU's Marie Curie training network PhilHumans - Personal Health Interfaces Leveraging HUman-MAchine Natural interactionS under grant agreement 812882 and from the project PON AIM1893589 promoting the attraction of researchers back to Italy.

[3] https://ec.europa.eu/social/main.jsp?catId=1062&langId=en

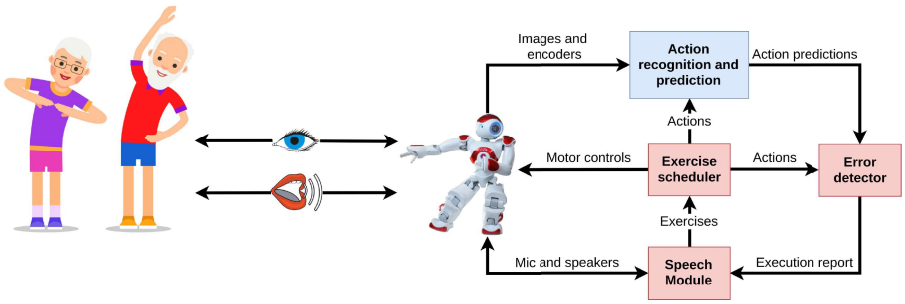

Fig. 1. Proposed architecture for a robotic coach

elders' independence and wellbeing [14]. In order to guarantee that people stay in charge of their own lives for as long as possible, regular physical activity is critical. The elder must perform a daily training schedule regularly and correctly. Usually, this is not the case and the presence of a caregiver to monitor the elder activities and progress is needed. This solution is not scalable and researchers are trying to automatise the teaching and monitoring processes [14]. Virtual coaches able to propose exercises, monitor the execution (possibly correcting errors) and send the results to a doctor are an appealing solution to the problem (see [2] or Robot-Era[4] and AHA[5] projects). Even if a system composed by a computer, a screen, and a camera is enough to implement a virtual coach, humanoid robots are better suited for this task. Interacting with a robot able to move and to show emotions is more engaging than staring at a computer screen. Also, a humanoid robot is able to show and perform by it-self the exercise, making the movement easier to understand by the elder.

To succeed in its task, a robotic coach needs to understand the vocal directives from the users, schedule their activities, monitor their exercises using onboard sensors, and correct their mistakes. In this paper we will present an architecture to address the monitoring tasks using out of the shelf RGB cameras. RGB cameras are cheap and powerful sensors to use in action recognition.

Video action recognition can be split into two main steps: action representation and action classification [11]. Traditionally, handcrafted features are used to represent the actions [8, 18], and standard classifiers are used to recognise the action (e.g. SVN, k-means). With the improvements in computational power and the increase in quality and size of video action recognition datasets [1, 10, 13], Convolutional Neural Networks (CNNs) are frequently used to extract features for action representation, achieving state-of-the-art results [7, 16, 19]. In order to use the time information of video sequences, researchers use different approaches: 3D convolutions to take into consideration space and time [7], multi-stream networks using both optical flow and RGB images as input [16], and hybrid networks that fuse together CNNs with recurrent neural networks (RNNs) architectures [19].

[4] http://www.robot-era.eu/robotera/
[5] http://welcome.isr.tecnico.ulisboa.pt/aha-project/

A more complex task is the prediction of future actions. The most recent action/motion prediction systems tend to use the combination of CNNs and RNNs [12]. Some researchers are even able to predict future frames based on the action to be performed and past frames [5, 9].
Until now we presented action recognition, prediction, and generation as separate problems, but there is a strong interconnection between them. In the Predictive Coding [15] cognition theory, the brain is constantly predicting the sensory outcome (top-down process) and comparing it with the actual one. At the same time the error between predicted and actual sensory stimuli is back-propagated to the highest layers (bottom-up process) in order to revise and update the internal predictive models (a similar idea applied to robot control was studied under the name of Expected Perception [3, 4]). Jun Tani implemented on robotics platforms several models based on the Predictive Coding paradigm [17]. One of the most recent is the Predictive Visuo-Motor Deep Dynamic Neural Network (P-VMDNN) [6]. This Deep-RNN model can be used both to predict visual and proprioceptive (that are produced and perceived within an organism/robot) stimuli, and to recognise an action performed by a human placed in front of the robot.
Comparing the predicted motion with the one performed by the elder, it is possible to detect mistakes in the execution of the exercise and warn the elder in case of a predicted fall. The comparison between predicted and actual state can also be used by a robot to spot the physical limitations of the elder and the predictive model can be updated accordingly.
The architecture presented in this paper is based on the P-VMDNN introduced by Hwang et al. in [6]. Structural changes are made in order to train and validate the model in a challenging real world scenario.
To summarize, the contributions of this paper are:

1. The introduction of a novel robotic coach system to promote elders' active life.
2. The design of a recurrent CNN architecture to recognise and to predict human actions from video sequences meant to be validated on real clinical studies.

We also introduce the procedure for the creation of a new video dataset containing sets of exercises performed by elders that will be augmented in simulation. This dataset will be created in collaboration with the faculty of medicine and surgery of Cagliari. Their help will be fundamental in the definition of the exercises, and for providing patients and infrastructure for the collection of the dataset and clinical studies. The remainder of this paper is organized as it follows. Section 2 introduces the robotic coach system we propose. Section 3 discusses about the module to recognize and to predict the action performed by the elder. Section 4 gives some hints on how we are going to collect data whereas Section 5 ends the paper.

2 Robotic coach system

Before describing in more details the action recognition and prediction architecture, the whole robotic coach system will be briefly introduced. The elders and the robotic coach interact through visual and vocal communication. The elders ask the robot about their daily exercise routine, giving relevant information about their health status (*e.g.* pains, blood pressure measurements, tiredness). The robot listens to the information and creates an ad-hoc exercise schedule. The robotic coach will then explain vocally the exercises, and will demonstrate how to perform them. After viewing the demonstration, the elders will perform the exercises monitored by the robotic coach. During the monitoring process, the robot will give a live feedback, pointing out errors and praising a correct execution by the elders. Fig. 1 shows the proposed architecture for a robotic coach. The whole system can be divided into four modules:

Speech module: This is the module in charge of the vocal interaction between the robot and the elders. A Natural Language Processing (NLP) sub-module interprets the elders' instructions received via robot's embedded microphones. The instructions are used to select the proper exercise to be sent to the Exercise scheduler. A speech generator sub-module informs the elders, through the robot's speakers, on how well they are performing the exercises and what they need to improve in their execution.

Exercise scheduler: The role of this module is to break the selected exercise into atomic actions, and send these actions to the Error detector and Action recognition and prediction modules. Moreover, this module contains a controller that translates the actions in motor commands to send to the robot.

Error detector: The Error detector module analyses the results coming from the Action recognition and prediction module based on the required actions received by the Exercise scheduler. After evaluating the performed action, the module sends an evaluation report to the speech module in order to inform the elder.

Action recognition and prediction: This module has the dual task of recognising the action performed by the elder using the RGB videos coming from its embedded camera, and predicting the future viso-proprioceptive stimuli based on the action that is being performed. The predicted frames and encoders' positions are sent to the Error detector module to be analyzed. This module is the main focus of this paper and it will be described in details in section 3.

The entire system will run on external computational resources (cloud services or external computer) connected via wireless to the robot.

3 Action recognition and prediction

As the name suggests, the Action recognition and prediction module must perform two distinct tasks:

1. recognize the action performed by the elder, and
2. predict the future visuo-proprioceptive sensory stimuli produced by the movement.

While it is possible to design two different models to handle each one of these tasks, a single architecture able to recognize and predict the action performed is able to exploit the shared information needed by both tasks. The Action recognition and prediction architecture proposed in this paper is based on the P-VMDNN network introduced by Hwang et al. in [6]. In the next sub-section the core ideas behind the P-VMDNN network are introduced, while sub-section 3.2 describes the actual integration of the network inside the proposed architecture.

3.1 P-VMDNN

P-VMDNN network is able to learn and recognise viso-proprioceptive patterns, and at same time generate visuo-proprioceptive predictions with a given intention. The network consists of a visual and a proprioceptive pathway. Each pathway is organized in a hierarchical fashion with multiple levels, starting from the lower sensory input/output levels to end up with the higher abstract levels. The two pathways are reciprocally connected at each level, allowing a bidirectional exchange of information. Moreover, each level possesses bidirectional connection with the adiacent ones, giving to the network the ability to output visual-prorpioceptive signals with the same dimensions of the input ones. In order to encode temporal dynamics of the input signal, each layer possesses different dynamics (faster for the lowest levels and slower for the highest ones) imposed with different spatio-temporal constraints on the internal state updates. Both pathways are implemented with recurrent layers, convolutional for the visual pathway and fully connected for the proprioceptive one. The inputs and the outputs are RGB video frames for the visual pathway and encoder values for the propriocetive one.
The forward dynamics of the network are of two kinds: open loop and closed loop. In the open loop case, the network receives each step the input stimuli and predicts the visuo-proprioceptive stimuli of the immediate next timestep. On the other end, running closed loop the visuo-proprioceptive prediction generated at the current time step by the network is used as input for the next one. In the closed loop case, the network does not need a continuous input from the environment, being able to predict a sequence of visuo-prorpioceptive stimuli from its initial state.
The optimization of the network parameters can be divided in two phases: offline training and online prediction error (PE) minimization. First the network is trained offline in a supervised manner. Weights, biases and initial states of the recurrent layers are optimized to minimize the error between the onestep look-ahead visuo-proprioceptive prediction and the training visuo-proprioceptive sequences. After the offline training the network is ready to be used. The PE minimization is performed online while observing the action performed by the elder. A temporal window size W is selected. Each time step t the network is

executed in a closed loop manner starting from step $t - W$ until step t. Then, the PE between the desired and the predicted visuo-proprioceptive sequence is calculated and back propagated to optimize the internal state at time $t - W$. This process is repeated k times each step. Thanks to this process, the network is able to adapt online to the observed action.

3.2 Network integration

Until now the P-VMDNN model has been tested on simulation in trivial experiments aimed to show the network potential. Our goal is to design an architecture for a robotic coach to be used in real life clinical studies. In order to successfully integrate the model in the proposed architecture, few improvements need to be done.
The robotic coach will be used in many different test scenarios, with changes in illumination, background and subjects being monitored. P-VMDNN is a complex recurrent network, rich in number of parameters. In order to successfully train the network, a large and varied dataset is needed. For this reason, we are planning to augment through simulation a dataset originally recorded with clinical patients. The data collection and augmentation strategy are described in details in section 4.
Convergence during training of recurrent neural networks is generally not easy to obtain, due to the vanishing of the gradient problem. LSTM and GRU, on the contrary, use a gated structure to improve convergence during training and are able to memorize longer time dependencies. For this reason we propose to replace each recurrent layer in P-VMDNN model with a GRU or ConvGRU one (depending on the fact whether the layer is in the visual or proprioceptive pathway). Fig. 2 depicts the proposed modified model.
In order to integrate the network inside the full architecture the following steps are necessary. First the network is trained offline on the augmented dataset as described in section 3.1. After training, the initial network internal states for each sequence in the training set are clustered based on the action that they represent, obtaining an initial internal state for each action. These internal states are used to initialize the network before monitoring the actions. While the elder is performing the exercise, the network runs adapting online to the video received by the camera of the robot using the PE minimization. During execution, the immediate next timestep predicted images and encoder values are passed to the Error detector module where they are compared with the desired ones.
Running the network several steps ahead the current one, it would be possible for the Error detector module to identify dangerous motions (resulting in falls or injuries) and warn the elder to stop before getting hurt.

4 Dataset

There will be a close collaboration with the faculty of medicine and surgery of Cagliari. Their help will be fundamental in the definition of the exercises, and

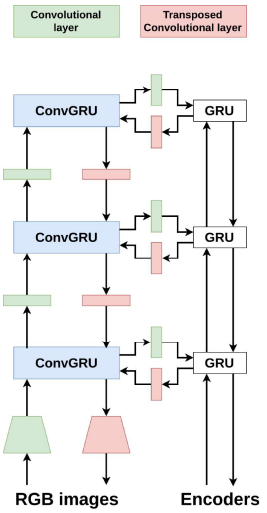

Fig. 2. Implementation of the P-VMDNN model using Conv-GRU and GRU as recurrent layers. On the left is the full convolutional visual pathway and on the right the proprioceptive one.

for providing patients and infrastructure for the collection of the dataset and clinical studies. A dataset of a few hundreds of RGB-D videos of a set of patients performing the exercises will be recorded. The patients' motion will be labelled using a motion capture system (*e.g.* Vicon[6]).
The recorded dataset would not be various and big enough to train the Action recognition and prediction module presented in section 3. In order to better generalize and successfully converge during training, we propose an augmentation method based on simulated data generated using a game engine (*e.g.* Unity[7] or Unreal Engine[8]). The motion caption labels, extracted for the original dataset, will be used to move photo-realistic avatars inside simulated 3D environments. Avatars, 3D environments, lights and camera positions and parameters, will be randomized to create an augmented dataset several time bigger than the original one.

5 Conclusions

In this paper we presented a robotic coach system to promote the elders active ageing. The proposed robotic coach will be able to vocally interact with the users, to propose a proper training schedule and to monitor the execution of the exercises. We focus our attention on the Action recognition and prediction module,

[6] https://www.vicon.com/
[7] https://unity.com/
[8] https://www.unrealengine.com/

based on the P-VMDNN model. We proposed an integration of the P-VMDNN model with the robotic coach system to monitor the proper execution of the exercises and detect future failures. Moreover, we introduced a novel augmentation strategy for a training dataset that will be recorded in collaboration with the faculty of medicine and surgery of Cagliari. We believe that with the proposed P-VMDNN implementation and the novel augmented dataset, the system will be robust enough to be employed in real clinical studies.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Albaina, I.M., Visser, T., Van Der Mast, C.A., Vastenburg, M.H.: Flowie: A persuasive virtual coach to motivate elderly individuals to walk. In: 2009 3rd International Conference on Pervasive Computing Technologies for Healthcare. pp. 1–7. IEEE (2009)
3. Barrera, A., Laschi, C.: Anticipatory visual perception as a bio-inspired mechanism underlying robot locomotion. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 3206–3209. IEEE (2010)
4. Cauli, N., Falotico, E., Bernardino, A., Santos-Victor, J., Laschi, C.: Correcting for changes: expected perception-based control for reaching a moving target. IEEE Robotics & Automation Magazine **23**(1), 63–70 (2016)
5. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Advances in neural information processing systems. pp. 64–72 (2016)
6. Hwang, J., Kim, J., Ahmadi, A., Choi, M., Tani, J.: Dealing with large-scale spatio-temporal patterns in imitative interaction between a robot and a human by using the predictive coding framework. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2018)
7. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
8. Jia, K., Yeung, D.Y.: Human action recognition using local spatio-temporal discriminant embedding. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
9. Jung, M., Matsumoto, T., Tani, J.: Goal-directed behavior under variational predictive coding: Dynamic organization of visual attention and working memory. arXiv preprint arXiv:1903.04932 (2019)
10. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
11. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230 (2018)
12. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 336–345 (2017)

13. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, Y., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence (2019)
14. Parra, C., Silveira, P., Far, I.K., Daniel, F., De Bruin, E.D., Cernuzzi, L., D'Andrea, V., Casati, F., et al.: Information technology for active ageing: A review of theory and practice. Foundations and Trends® in Human–Computer Interaction **7**(4), 351–448 (2014)
15. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience **2**(1), 79 (1999)
16. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
17. Tani, J.: Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press (2016)
18. Yuan, C., Wu, B., Li, X., Hu, W., Maybank, S., Wang, F.: Fusing $\mathcal{R}$ features and local features with context-aware kernels for action ecognition. International Journal of Computer Vision **118**(2), 151–171 (2016)
19. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)