Embodied Generative Third Person View Translation and Encoding

Luca Garello
Istituto Italiano di Tecnologia
Genoa, Italy
luca.garello@iit.it

Francesco Rea
Istituto Italiano di Tecnologia
Genoa, Italy
francesco.rea@iit.it

Alessandra Sciutti
Istituto Italiano di Tecnologia
Genoa, Italy
alessandra.sciutti@iit.it

Nicoletta Noceti
Machine Learning Genoa Center (MaLGa)
DIBRIS, Università degli Studi di Genova
Genoa, Italy
nicoletta.noceti@unige.it

***Abstract*—We propose a self-supervised generative model for addressing the perspective translation problem. In particular we focus on third-person to first-person view translation as primary and more common form of perspective translation in human robot interaction. Evidences show how this skill is developed in children since the very first months of life. In nature, this skill has been also found in many animal species. Endowing robots with perspective translation would be an important contribution to the research fields such as imitation learning and action understanding. We trained our model on simple RGB videos representing actions seen from different perspectives, specifically the first person (ego-vision) and third person (allo-vision). We demonstrate that the learned model generates results that are visually consistent. We also show that our solution automatically learns an embedded representation of the action that can be useful for tasks like action/scene recognition.**

***Index Terms*—View translation, generative adversarial network, U-net, imitation learning, embedding**

I. Introduction

The recent development of artificial intelligence algorithms paves the way to new ways of learning in the field of robotics. Deep learning and reinforcement learning proved to be promising techniques to endow robots with the ability to perform motor control tasks. Efficient algorithms that learn quickly from a small amount of data are a key topic for modern research. The ability of learning in a human way is an ambitious target, indeed we have the ability to learn from a small amount of data and we are able to use the acquired knowledge to solve new (different) problems. In particular, imitation learning plays a key role in our development from the early months of our life [1]. In fact, by observing expert demonstrators we are able to learn new skills. Humans are not the only living beings that learn by imitation, it has been proven that in nature a large variety of animals learn from imitation [2]. For this reason the idea of having robots able to learn new tasks by using demonstration policies is the subject of an increasing number of research.

With this work we propose a perceptual algorithm based on Generative Adversarial Networks (GANs) which facilitates the robot to handle the view perspective shift between demonstrator and imitator. This is possible by learning from video demonstration data without any explicit reward function.

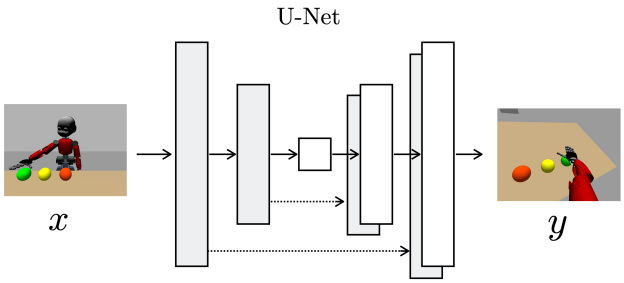

Fig. 1. The proposed generative model for view translation. The network is based on a U-net, it learns to transform source video frames x into other target frames y. Our model extracts the features by means of convolutional layers and max-pooling layers until obtaining an embedded representation of each frame. Then, starting from the embeddings, the images are upsampled in order to reconstruct the equivalent representations in first person.

II. Related Work

Imitation learning is the ability to learn how to perform a certain task using information generated by another expert agent performing that same task. The specific problem of imitation from observation (IfO) has recently gained the attention of the robotics community. To address the viewpoint difference, researchers share the idea of embedding actions by representing them with a numerical description.YuXuan Liu et al.[3] focus on learning a context translation model that can convert a demonstration from one context (e.g., a third person viewpoint and a human demonstrator) to another context (e.g., a first person viewpoint and a robot). By training a model to perform this conversion, they acquire a feature representation that is suitable for tracking demonstrated behavior . Sermanet et al. propose an encoding approach based on temporal consistency, forcing the encoder to map simultaneous viewpoints of the same action in the same embedding space [4]. Sharma and Smith implement a conditional GAN network with the aim of predicting at each time step the next state of the replicating robot [5][6] . In this work we focus on GAN approaches, trying to evaluate different architectures and assessing their capabilities in the specific task of view-translation.

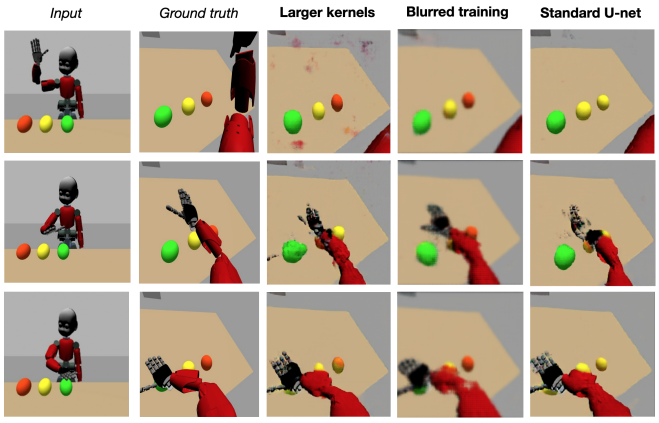

Fig. 2. Different architectures induce different quality of results. In this task the proposed U-nets outperform the standard U-net. In particular the training on blurred images leads to better, less noisy reconstructions.

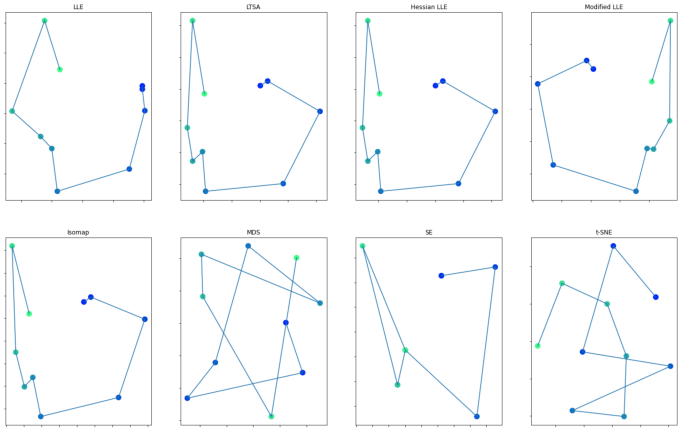

Fig. 3. The manifold analysis shows how an action is embedded in the latent space. The embeddings are generated with a temporal consistency, this pattern is not forced during the training phase, but autonomously learnt.

III. Approach

The goal of this research is to have an iCub robot that first observes the video demonstration of a human or a humanoid robot performing the task in front of it, then it is able to generate a first person (ego) representation of the same action. It is important to underline that during this process the robot learns to handle the view shift with the only input signal of a 2D raw video, iCub has not access to the explicit joint configuration of the demonstrator. Our approach aim at learning a viewpoint-invariant representations of objects and other agents in the environment, in a similar way to the functioning of "mirror neurons". In our test we explored different variants of GANs based on the U-net architecture. Similarly to autoencoders, U-Nets have the ability to learn an embedded representation using a bottleneck. Differently to autoencoders, which are feed forward architectures, U-nets includes skip connections. Skip connections copy the information from one compressing layer to another expanding layer, this helps to keep the higher level features of the image during the reconstruction phase. An intuitive graphical explanation is given in Fig [1].

IV. Experiments

For this preliminary study we considered a reaching task in which the robot reaches one colored ball at a time. For each reaching task the positions of the three balls (Orange, Green and Yellow) are randomly assigned in a interval of ±5 cm for both x and y coordinates. We trained the U-nets on 18 reaching actions, for a total of 222 images. We investigated several variants of the U-net, in particular we tested: (a) the standard U-net version as proposed by Isola et al. [8]. (b) a version with larger convolutional kernels, so as to favour the learning of spatial relationships. (c) again a version with large kernels but trained on a blurred dataset, this with the intention of reducing the noise during training.

Fig[2] shows the generated frames using actions from our Test Set. The proposed variants of U-net outperform the original U-net, especially in the colour rendering. In addition to this, we also analyzed the embedding generated by the networks Fig[3]. For this analysis we decided to consider the U-Net with Larger Convolutional Kernels, trained on the blurred dataset. The manifold analysis gives us a 2D representation of our 512-dimensional embeddings.

V. Conclusion

We show that generative models are an effective tool for addressing the view variance. Our approach not only considers the pixel domain but also deals with an embedded representation of the action. This embedded representation is learned automatically and is temporal consistent, even though during the training phase we did refrain from forcing a specific embedding pattern. Future works will better explore the capabilities of our network.

References

[1] Piaget, Jean. Play, dreams and imitation in childhood. Vol. 25. Routledge, 2013.

[2] Huber, L., Range, F., Voelkl, B., Szucsich, A., Viranyi, Z., Miklosi, A. (2009). The evolution of imitation: what do the capacities of non-human animals tell us about the mechanisms of imitation?. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1528), 2299-2309.

[3] Liu, Y., Gupta, A., Abbeel, P., Levine, S. (2017). Imitation from observation: Learning to imitate behaviors from raw video via context translation. arXiv preprint arXiv:1707.03374.

[4] Sermanet, P., Lynch, C., Hsu, J., Levine, S. (2017, July). Time-contrastive networks: Self-supervised learning from multi-view observation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 486-487). IEEE.

[5] Sharma, P., Pathak, D., Gupta, A. (2019). Third-person visual imitation learning via decoupled hierarchical controller. In Advances in Neural Information Processing Systems (pp. 2597-2607).

[6] Smith, L., Dhawan, N., Zhang, M., Abbeel, P., Levine, S. (2019). Avid: Learning multi-stage tasks via pixel-level translation of human videos. arXiv preprint arXiv:1912.04443.

[7] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.