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.

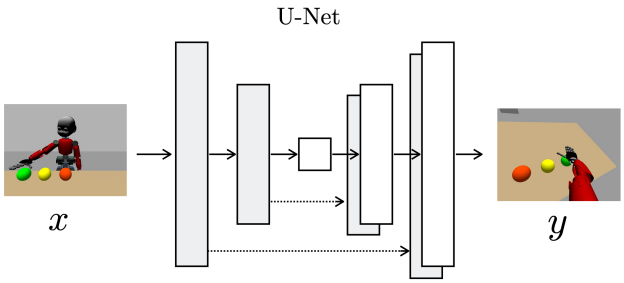

Fig. 1. The proposed generative model for view translation. The network is based on a U-net, it learns to transform source video frames x into other target frames y. Our model extracts the features by means of convolutional layers and max-pooling layers until obtaining an embedded representation of each 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.