

Neural Policy Style Transfer

Raul Fernandez-Fernandez*, Juan G. Victores, Jennifer J. Gago, David Estevez,
Carlos Balaguer

All of the authors are members of the Robotics Lab research group within the Department of Systems Engineering and Automation, Universidad Carlos III de Madrid (UC3M), Madrid, Leganes, Av. Universidad 30, 28911, Spain.

Abstract

Style Transfer has been proposed in a number of fields: fine arts, natural language processing, and fixed trajectories. We scale this concept up to control policies within a Deep Reinforcement Learning infrastructure. Each network is trained to maximize the expected reward, which typically encodes the goal of an action, and can be described as the *content*. The expressive power of deep neural networks enables encoding a secondary task, which can be described as the *style*. The Neural Policy Style Transfer (NPST)¹ algorithm is proposed to transfer the *style* of one policy to another, while maintaining the *content* of the latter. Different policies are defined via Deep Q-Network architectures. These models are trained using demonstrations through Inverse Reinforcement Learning. Two different sets of user demonstrations are performed, one for *content* and other for *style*. Different *styles* are encoded as defined by user demonstrations. The generated policy is the result of feeding a *content* policy and a *style* policy to the NPST algorithm. Experiments are performed in a catch-ball game inspired by the Deep Reinforcement Learning classical Atari games; and a real-world painting scenario with a full-sized humanoid robot, based on previous works of the authors. The implementation of three different Q-Network architectures (Shallow, Deep and Deep Recurrent Q-Network) to encode the policies within the NPST framework is proposed and the results obtained in the experiments with each of these architectures compared.

Keywords: Style Transfer, Deep Reinforcement Learning, Robotics, Deep Learning.

1. Introduction

The concept behind Style Transfer is to transform the *style* of a certain input without changing its original *content*. This *content* is often referred as the *what*, whereas the *style* as the *how*. Style Transfer has been applied essentially to three main fields: computer vision with fine arts, relating objects and shapes with painting techniques; natural language processing, relating the meaning of a text with the specific selection

¹NPST: Neural Policy Style Transfer

*Corresponding author: rauferna@ing.uc3m.es

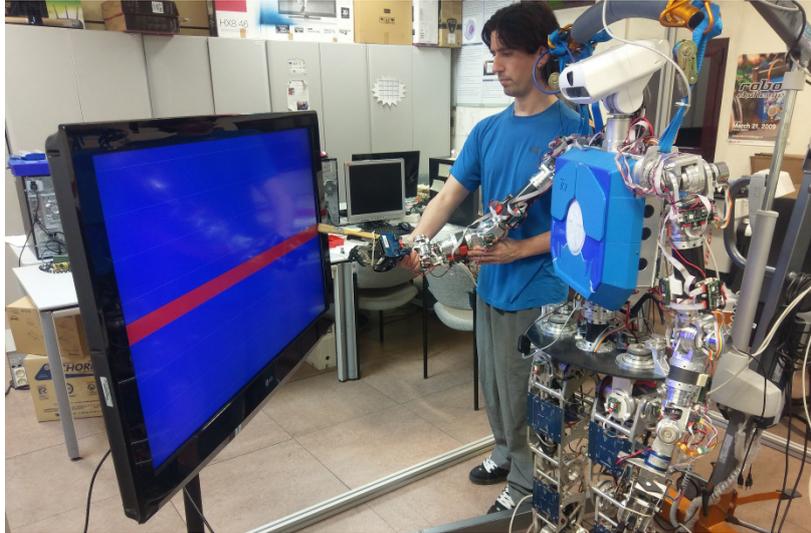


Figure 1: Neural Policy Style Transfer training: “Grid-world paint” scenario with TEO the humanoid robot.

of words; and, finally, fixed trajectories for animation, relating the final motion with the manner or emotion with which the trajectory is performed.

Early works in computer graphics, specifically in the area of creating trajectories for animating figures, apply “moods” or “emotions” (*style*) to “base motion” actions (*content*) by a weighted addition in the frequency-phase domain [21]. This method generates fixed trajectories, which are a kind of actions, but do not involve feedback with respect to the state of an agent. Additionally, it relies on periodicity, which in the human domain limits actions to locomotion tasks. Further literature includes the incorporation of signal processing techniques [2], and Hidden Markov Model representations as well as statistical modelling [4].

The first works to coin the terms *style* and *content* are in the context of computer vision, concretely in the area of optical character recognition. An explicit bilinear model was used to separate the *style*, which corresponds to the used font or calligraphy, while the *content* is given by the actual letters or graphemes [20]. Further works include multilinear modelling techniques, which employ the N-mode singular value decomposition (SVD) tensor extension of the conventional matrix SVD to transform collections of images into spaces where the same bilinear model can be applied.

The model known as AlexNet [13] demonstrated the potential of Deep Neural Networks (DNN) for computer vision image classification tasks, and a large variety of alternative DNN models spawned:

On the one hand, VGG [19] became particularly popular for image classification. The VGG-19 network, pretrained on the ImageNet dataset, was used by Gatys et al [10] to develop Style Transfer for images, which marked the beginning of its application to fine arts. In this case, Style Transfer is developed as an optimization process where the pretrained weights of the VGG-19 network remain constant throughout all the iterations. Three images are used as the input of the network: These are the *content* image, the *style*

image, and the generated image. The generated image is the result of optimizing a loss function consisting on a weighted sum of *content* loss and *style* loss. The *content* loss is obtained as the difference between the outputs of the high level features of the network corresponding to the *content* and the generated image. The *style* loss is defined using the “Gram matrix” of the *style* and generated image, which is based on monitoring the activation values across channels and through different layers. Following the ideas proposed by Gatys et al, several works have been proposed in different areas. Dumoulin et al. [6] proposed the introduction of a parametric generalizable DNN for Style Transfer in images able to encode multiple styles. Fu et al. [9] proposed two different algorithms using Autoencoders to introduce Style Transfer in natural language processing without using parallel data. In Lee et al. [14], a Style Transfer method encoding the *style* as a noise introduced in the text is proposed. Two neural models are trained, one to introduce the style and other to remove the style and generate the clean text. Other applications include: the introduction and merging of new elements in artistic images [15]; portrait Style Transfer using soft masks [24] [25]; and fixed trajectory generation for animation as proposed by [11].

On the other hand, Deep Reinforcement Learning (DRL) [16] also arose as part of the trend, applying DNN to Reinforcement Learning, where agents learn control policies to maximize perceived rewards. Initial applications were video-game oriented, and the rewards were equivalent to the obtained scores. Temporal aspects of the actions were taken into consideration using as input a concatenation of k images corresponding to the last k time steps. DRL marks the beginning of DNN applied to control policies, using the network to represent the Q-value function of the Reinforcement Learning problem statement. Modern DRL approaches include Deep Q-Networks (DQN), Trust Region Policy Optimization, Generative Adversarial Imitation Learning, and Asynchronous Advantage Actor-Critic [3].

Drawing inspiration from the two main previously stated sources, emerged from the outbreak of DNN, this paper presents a Neural Policy Style Transfer (NPST) algorithm. NPST is proposed with the goal to improve the generalization capabilities of robots with the introduction of *styles* as a way to achieve action adaptation. The same base action can be adapted to be used in different contexts with the introduction of different recorded *styles*. In addition to this, a wide range of applications can be derived from the contributions proposed by this paper and summarized in the following three points:

- The introduction of Inverse Reinforcement Learning (IRL) algorithms to encode the *content* and *style* using DQN as defined by human demonstrations.
- Proposal of the Neural Policy Style Transfer (NPST) algorithm to perform Style Transfer between policies, applicable to Q-value functions expressed as DQN.
- Two different experimental scenarios: the “Catch-ball” scenario inspired by the Deep Reinforcement Learning classical Atari games, where a ball must be caught by a side-by-side moving paddle; and

the ‘‘Grid-world paint’’ scenario, a painting scenario with a full-sized humanoid robot, which builds upon previous experiments in evolutionary cognitive robotics [7] [8] and is equivalent to a grid-world problem in the real world.

The remainder of this paper is organized as followed: Section 2 provides a Reinforcement Learning background. Section 3 describes the proposed Style Transfer algorithm. Sections 4, 5, and 6 depict the experiments performed in this paper. Finally, conclusions are drawn in Sections 7 and 8.

2. Background

Let a Markov Decision Process (MDP) be defined using the tuple $M = \{S, A, T, \gamma, D, R\}$, where S represents the state space, A the action space, $T = \{P_{sa}\}$ the transition probabilities defined by the environment, where P_{sa} is the state transition distribution upon taking action a in state s , $\gamma \in [0, 1]$ is the discount factor, D is the initial state distribution, and $R : S \rightarrow A$ the reward function. An optimal policy $\pi^*(s)$ can be found such as the reward obtained by the agent is maximized over a full execution of the problem. This policy defines the behavior of the agent mapping actions with states. A Q-learning approach [22] defines this policy $\pi^*(s)$ using the Q-value function defined in Eq. 1.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, A) - Q(S_t, A_t)] \quad (1)$$

where t can be any time step, α is the learning rate, and $Q(S_t, A_t)$ defines the expected reward obtained over a full episode starting from the state s with an action a and following a greedy policy. In DQN, this Q-function is encoded using a DNN.

The definition of a proper reward function is a critical step of Reinforcement Learning architectures that defines the behavior of the obtained agent. Inverse Reinforcement Learning (IRL) is proposed as a way to define the reward function R using a set of m expert demonstrations $E = \{s_0^{(i)}, s_1^{(i)}, \dots\}_{i=1}^m$.

The IRL algorithm, as defined by Abbeel et al [1], assumes that a k -dimensional feature vector $\phi(s) \in [0, 1]$ exists in S such that R can be defined as $R(s) = w \cdot \phi(s)$, where $w \in \mathbb{R}^k[0, 1]$ is defined as the weight feature vector to optimize. While the feature vector $\phi(s)$ can be hand-crafted by the designer, some approaches have been proposed for its selection to be automated [23].

The expert feature expectation $\hat{\mu}_E$ can be defined as in Eq. 2:

$$\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^m \sum_{t=0}^{\infty} \gamma^t \phi(s_t^{(i)}) \quad (2)$$

This expert feature expectation $\hat{\mu}_E$ depicts a measure of the degree of desirability of the different features for the demonstrator. This provides an intuition of which features are related to larger and lower rewards and, therefore, allows the estimation of the reward as a function of these features.

Given an MDP where the reward function $R(s)$ is unknown, the objective of the IRL algorithm is to find an optimal policy π^* , defined by $R(s)$, such that it satisfies Eq. 3:

$$\|\mu(\pi^*) - \hat{\mu}_E\|_2 < \epsilon \quad (3)$$

where $\mu(\pi^*)$ is defined as in Eq. 4:

$$\mu(\pi^*) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \middle| \pi^* \right] \in \mathbb{R}^k \quad (4)$$

Based on these premises, several approaches have been proposed to solve the IRL problem. In this paper, we use the Maximum Entropy IRL approach [26], where the IRL algorithm is reduced to the maximization of a likelihood distribution defined as in Eq. 5:

$$\mathcal{L}(w) = \log P(E, w | R) = \mathcal{L}_E + \mathcal{L}_w \quad (5)$$

where \mathcal{L}_E and \mathcal{L}_w are, respectively:

$$\mathcal{L}_E = \log P(E | R) \quad (6)$$

$$\mathcal{L}_w = \log P(w) \quad (7)$$

Wulfmeier et al [23] adapted the distribution defined by Eq. 5 to work with DNN by defining the gradient of the reward function with respect to the weights obtained using backpropagation as in the Eq. 9 extracted from Eq. 8:

$$\frac{\delta \mathcal{L}}{\delta w} = \frac{\delta \mathcal{L}_E}{\delta w} + \frac{\delta \mathcal{L}_w}{\delta w} \quad (8)$$

where $\frac{\delta \mathcal{L}_E}{\delta w}$ is given by Eq. 9:

$$\frac{\delta \mathcal{L}_E}{\delta w} = \frac{\delta \mathcal{L}_E}{\delta R} \cdot \frac{\delta R}{\delta w} = (\hat{\mu}_E - \frac{1}{m} \sum_{i=1}^m \sum_{t=0}^{\infty} P(s_t^{(i)} | R)) \cdot \frac{\delta R}{\delta w} \quad (9)$$

The gradient of the expert demonstration term \mathcal{L}_E with respect to the model parameters of a linear function is equal to the feature expectation difference along the expert trajectories [26]. The used DNN model and the Style Transfer algorithm is described in the following section.

3. Neural Policy Style Transfer

Let \mathcal{C} and \mathcal{S} be two different DQN encoding two different actions defined via user demonstrations, which in turn define their control policies π_c and π_s . The input must correspond to the observation or state space, and the output must express the Q-value function.

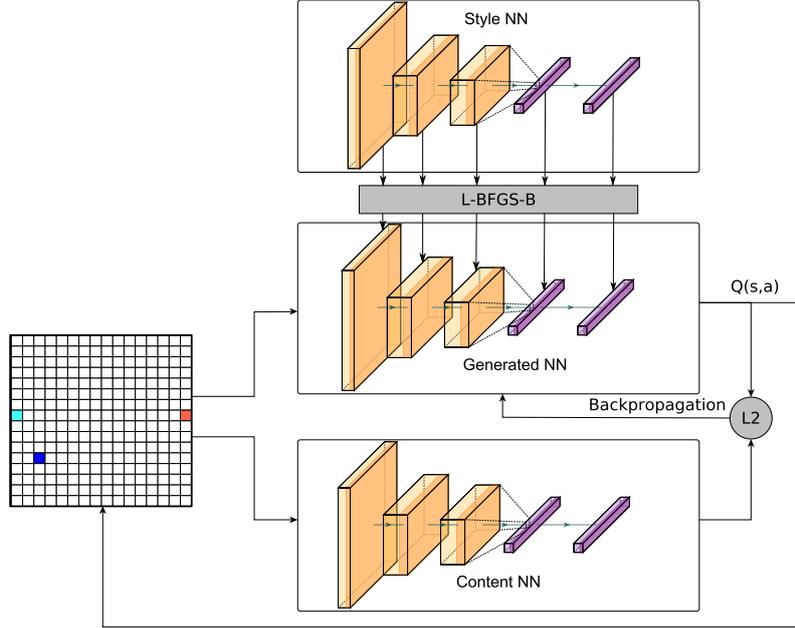


Figure 2: Neural Policy Style Transfer framework. The framework is composed by three Deep Q-Networks with the same architecture. The content DQN is trained using the content demonstrations. The style DQN is trained using the style demonstrations. The Generated DQN is generated using the output of the Content DQN and the weights of the Style DQN.

Let \mathcal{G} be the DQN that encodes the generated action, the output of the Neural Policy Style Transfer (NPST) algorithm, which in turn defines the control policy π_g . \mathcal{G} can be defined as a combination of the *content* and *style* defined by the two base DQN, \mathcal{C} and \mathcal{S} . The *content* can be defined as the high level features of \mathcal{C} encoding the goal of the action. The *style* can be defined as the weights of \mathcal{S} defining the “mood” or “emotion” as defined by user demonstrations.

An approach analogous to the one proposed by Gatys et al [10] would involve optimizing to minimise a weighted sum of a *content* loss and a *style* loss. However, there are significant differences in the meaning of each element involved. Caution must be taken so at least one of these losses is clearly separated from the other, to avoid straying into a simple *mix* of actions.

The *content* transfer is performed by training \mathcal{G} based on the high-level features of \mathcal{C} , corresponding to the Q-value output. A backpropagation step is performed over \mathcal{G} using the output of \mathcal{C} as the true label. A Mean Squared Error (MSE) function is used as the loss function for the backpropagation algorithm. This is equivalent to defining the *content* loss $\mathcal{L}_{content}$ for a single output as in Eq. 10:

$$\mathcal{L}_{content}(\mathcal{G}, \mathcal{C}) = \|q_G - q_C\|_2 \quad (10)$$

Where q_G and q_C are the high level features of \mathcal{G} and \mathcal{C} . These are the DQN outputs of \mathcal{G} and \mathcal{C} for a given state.

Its counterpart, the *style* transfer of the algorithm is implemented via an optimization that depends on the *style* loss \mathcal{L}_{style} , which is defined as in Eq. 11:

$$\mathcal{L}_{style}(\mathcal{G}, \mathcal{S}) = \|w_{\mathcal{G}} - w_{\mathcal{S}}\|_2 \quad (11)$$

where $w_{\mathcal{G}}$ and $w_{\mathcal{S}}$ correspond to the weights of the models that represent \mathcal{G} and \mathcal{S} , respectively.

The assumption here is that by introducing the *style* transfer through the weights of the network, the framework is giving preference to the outputs preferred by the *content*. The *style* is introduced as a secondary task in the *content* execution.

Algorithm 1 Neural Policy Style Transfer (NPST)

```

1: procedure NPST( $\mathcal{C}, \mathcal{S}, env, N$ )

2:   Initialise:
3:    $\mathcal{G} \leftarrow \mathcal{S}$ 
4:    $w_{\mathcal{S}} \leftarrow \mathcal{S}.get\_weights()$ 
5:    $env.init()$ 
6:    $state \leftarrow env.observe()$ 

7:   for  $n=1:N$  do
8:     Update Environment:
9:      $q_{\mathcal{G}} \leftarrow \mathcal{G}.predict(state)$ 
10:     $a_{\mathcal{G}} \leftarrow \operatorname{argmax}_a(q_{\mathcal{G}})$ 
11:     $env.step(a_{\mathcal{G}})$ 
12:     $state \leftarrow env.observe()$ 

13:    Content Transfer:
14:     $q_{\mathcal{C}} \leftarrow \mathcal{C}.predict(state)$ 
15:     $q_{\mathcal{G}} \leftarrow \mathcal{G}.predict(state)$ 
16:     $\mathcal{G}.backprop(q_{\mathcal{G}}, q_{\mathcal{C}})$ 

17:    Style Transfer:
18:     $w_{\mathcal{G}} \leftarrow \mathcal{G}.get\_weights()$ 
19:     $w_{\mathcal{G}} \leftarrow \text{L-BFGS-B}(\|w_{\mathcal{G}} - w_{\mathcal{S}}\|_2)$ 
20:     $\mathcal{G}.set\_weights(w_{\mathcal{G}})$ 
21:  end for

22: end procedure

```

The DQN \mathcal{G} is initialized as a copy of \mathcal{S} , both in model as in pre-computed weights. The environment is initialized and its state is observed. \mathcal{G} is updated for N iterations:

1. The environment is updated using the action defined by \mathcal{G} .

2. The values q_C and q_G , which represent the high-level features of C and G respectively, are obtained for the current state. A backpropagation step is then performed over the model of G .
3. The weights of G are updated using a box-constrained limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) algorithm to minimise the *style* loss as presented in Eq. 11.

The full algorithm is shown in Alg. 1.

4. Experiments

Three different neural network architectures are introduced in the experiments to measure the performance of the NPST algorithm. The first architecture is a DQN with the same architecture as the one proposed by Mnih et al [17]. The second architecture, referred as Shallow Q-Network (SQN), is a smaller version of the first DQN architecture. In SQN, the second and third Convolutional Layers (CL) of the DQN architecture are removed. The same Fully Connected (FC) layers are used for both architectures. The last architecture introduced in the experiments is a Deep Recurrent Q-Learning Network (DRQN), referred as DRQN, with the same architecture as the one proposed by Brejl et al. [5]. The size of all the architectures is the same and corresponds to the size of the layers defined by Mnih et al. [17]. The Long Short-Term Memory (LSTM) layer defined with the DRQN architecture is composed by 256 nodes.

Two different experimental scenarios are proposed to measure the performance of these three architectures with the NPST algorithm. The “Catch-ball” experimental scenario is inspired by classical DRL scenarios using Atari games. The “Grid-world” paint scenario is designed to work with a humanoid robot and based on previous works of the authors [7] [8].

5. “Catch-ball” experiment

The first experiment consists in a “Catch-ball” game scenario, similar to the Pong arcade game. A ball is released from a random location from the top of the screen, and falls vertically. The agent can move a paddle horizontally at the bottom of the screen. The agent wins if it catches the ball with the paddle before it falls off the screen.

5.1. Experimental setup

Three different sets of five expert demonstrations are performed in these experiments. These sets correspond to three different actions: one *content* action, and two different *style* actions (“nervous” and “fall”). The *content* action aims to win the game by fulfilling the goal of catching the falling ball by using the paddle. Both demonstrated *style* actions ignore the position of the ball. The first *style* imitates a “nervous” behaviour or mood, tending to perform small moves around the same position. The second *style* imitates a “fall” movement, always tending towards the left side of the screen.

The reward functions that define these actions are obtained using the Maximum Entropy Deep IRL algorithm presented in section 2. These reward functions are used to train the networks that are introduced in the NPST algorithm. For the training of the IRL algorithm, a hand-crafted feature vector $\phi(s)$ and latent state space s are selected for the three sets of demonstrations. In the case of the *content* demonstrations, s is defined as a function of ball and paddle positions, and $\phi(s)$ classifies if the paddle and ball positions are aligned. In the case of the *style* demonstrations, the same latent state space s was used for both *styles*. This space is defined using the last three paddle positions, corresponding to the last three time steps of the paddle. The feature vector $\phi(s)$ used for the *styles* encodes the spatial-temporal information of the paddle. In the case of the “nervous” *style*, $\phi(s)$ classifies if a movement was performed with the same starting and ending position. In the case of the “fall” *style*, $\phi(s)$ classifies if a movement to the left was performed.

For the *content* demonstrations, 5 iterations of the IRL algorithm were performed, while 2 iterations were used for the *style* demonstrations. The IRL discount factor (γ) used was 0.9 with a learning rate of 0.01 for all the actions. Three reward functions $R(s)$ are obtained corresponding to the three base actions (*content*, “nervous” and “fall”). For each of these base actions, three neural networks, corresponding to the three architectures proposed, are trained using the same $R(s)$. These networks are trained using Q-learning and referred in the experimental results as the Vanilla neural networks. These Vanilla neural networks are the base neural networks that will later be used to define the *content* and style in the NPST framework. Experimental results obtained with these Vanilla networks are added as a baseline. The same Vanilla Content neural networks are introduced for the transferring of both of the *styles*. The input of the networks is the raw 80x80 pixel image of the screen, and the outputs correspond to the Q-values assigned to the three possible actions (stay still, go left, and go right).

The hyperparameters used for training the networks and performing the NPST algorithm are depicted in Table 1. The results for the generated NPST actions depicted in the next section are the average of 10 repetitions of the NPST algorithm.

5.2. Results

The results obtained with the NPST algorithm are depicted in Table 2 for the case of transferring the “nervous” *style* and Table 3 for the case of transferring the “fall” *style*. The $\mathcal{L}_{content}$ and \mathcal{L}_{style} are computed following Eq. 10 and Eq. 11 respectively and are the average of the full NPST execution. Results between different architectures in terms of obtained losses may not be comparable due to differences in the architecture (i.e. the number of nodes). The *Nervous Moves* parameter measures the number of times the agent changed its direction. This parameter was introduced as a way to measure the dynamic behavior of the agent to differentiate the “fall” and “nervous” *styles*. Finally, the *Wins(%)* parameter measures the percentage of wins obtained by the agent.

A clear increase in the number of *Nervous Moves* performed by the agent was measured when trans-

Table 1: Hyperparameters for the “Catch-ball” scenario.

Hyperparameter	Setting
<i>Shared between architectures</i>	
Activation	ReLU [18]
Initialization	Normal distribution
<i>DQN architecture</i>	
Layers	(CL, CL, CL, FC, FC)
Layers Configuration (Size, Kernel, Strides)	((32,8,4),(64,4,2),(64,3,1),(512,-,-),(3,-,-))
<i>SQN architecture</i>	
Layers	(CL, FC, FC)
Layers Configuration (Size, Kernel, Strides)	((32,8,4),(512,-,-),(3,-,-))
<i>DRQN architecture</i>	
Layers	(CL, CL, CL, LSTM, FC)
Layers Configuration (Size, Kernel, Strides)	((32,8,4),(64,4,2),(64,3,1),(256,-,-),(3,-,-))
<i>Q-learning and NPST</i>	
Image size	80x80
Number of input time steps	4 (1 for DRQN)
Optimizer	Adam [12]
Loss function	Mean Squared Error
Number of actions	3
Discount (γ)	0.99
Experience Replay size	5000
<i>Q-learning</i>	
Learning Rate	1e-6
Initial Epsilon	0.1
Final Epsilon	1e-5
Epsilon gradient	Lineal
Batch size	32
Exploration Epochs	100
Training Epochs	1000
<i>NPST</i>	
Learning Rate	0.01
Number of iterations (N)	One full Catch-ball episode
Batch size	100
L-BFGS-B internal iterations	1

ferring the “nervous” *style* with respect the Vanilla Content and Vanilla Fall networks. At the same time, the number of wins decreased when introducing the *styles* with respect the Vanilla Content networks but increased with respect the Vanilla Style networks. The resulting generated control policy combines the results obtained by the *content* and *style* policies.

Fig. 3 and Fig. 4 depict the number of times each position was visited by the paddle. The depicted results are the cumulative results of the ten repetitions. These figures depict the preferred states of the agent during the experiments. The background image of the game is introduced as reference. The “nervous” *style* transferred policies show a preference for the positions in the center. The “fall” *style* transferred policies show a preference for the positions in the borders. This tendency is clear in the case of the DQN architecture

Table 2: Experimental results for the “Catch-ball” action introducing the “nervous” *style*.

Actions	$\mathcal{L}_{content}$	\mathcal{L}_{style}	<i>Nervous Moves</i>	<i>Wins</i> (%)
Vanilla Content DQN	—	32857.59	75	100
Vanilla Content SQN	—	16805.58	73	100
Vanilla Content DRQN	—	33766.86	28	30
Vanilla Nervous Style DQN	1.51	—	280	20
Vanilla Nervous Style SQN	1.70	—	298	20
Vanilla Nervous Style DRQN	1.77	—	45	20
NPST Nervous Generated DQN	0.17	0.97	160	60
NPST Nervous Generated SQN	0.32	3.03	166	40
NPST Nervous Generated DRQN	0.17	1.61	57	30

Table 3: Experimental results for the “Catch-ball” action introducing the “fall” *style*.

Actions	$\mathcal{L}_{content}$	\mathcal{L}_{style}	<i>Nervous Moves</i>	<i>Wins</i> (%)
Vanilla Content DQN	—	16816.59	73	100
Vanilla Content SQN	—	32837.50	82	50
Vanilla Content DRQN	—	33753.42	28	30
Vanilla Fall Style DQN	1.21	—	2	30
Vanilla Fall Style SQN	2.19	—	14	0
Vanilla Fall Style DRQN	1.33	—	10	0
NPST Fall Generated DQN	0.20	0.94	48	50
NPST Fall Generated SQN	0.34	3.81	63	50
NPST Fall Generated DRQN	0.16	2.54	44	20

where the most visited position is at the left of the screen. This corresponds with the behavior shown by the Vanilla Fall networks.

Fig. 5 depicts intermediate time steps defined by the NPST generated policy. The screenshots, in the first and third row of the figure, depict the state of the game corresponding to some randomly chosen time steps. These game states were generated following the generated policy. The same time steps were chosen for both *styles*. The bar graphs compare the Q-values obtained using the Vanilla Content DQN and the NPST Generated DQN for each of the intermediate states and styles.

6. “Grid-world paint” experiment

The second set of experiments consists in a “Grid-world paint” scenario. The physical agent is the humanoid robot TEO, as shown in Fig. 1, which can move its end-effector vertically and horizontally in the Cartesian space. A large monitor emulates painting by changing large pixel colors upon presence of the robot end-effector. These pixels form a 16x16 grid. The goal is to reach a target pixel defined in the vertical center at the right side of the monitor.

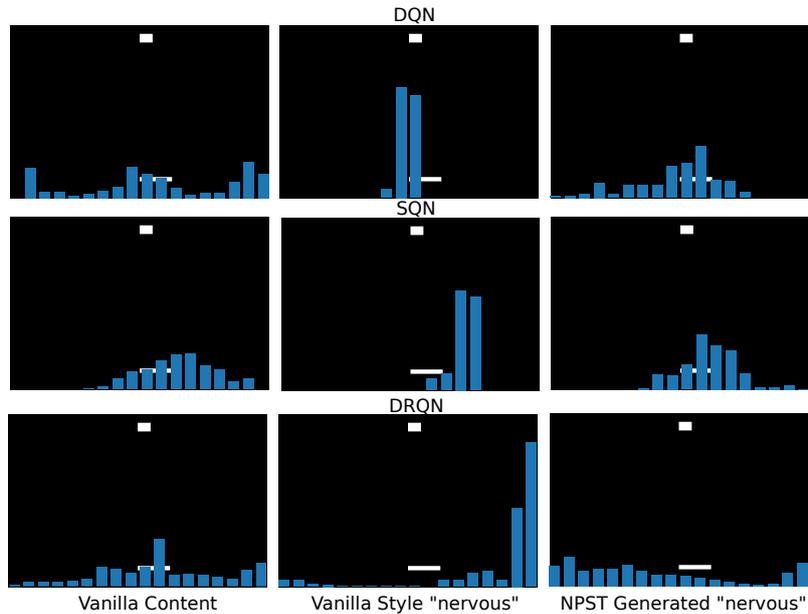


Figure 3: Paddle positions histograms for the case of transferring the “nervous” *style*. Each row corresponds to a different network architecture. Each column corresponds to a different action. The Y-axis depicts the number of times each position is visited. All Y-axis are scaled in the range [0, 250]. The X-axis depicts the possible paddle positions within the game environment.

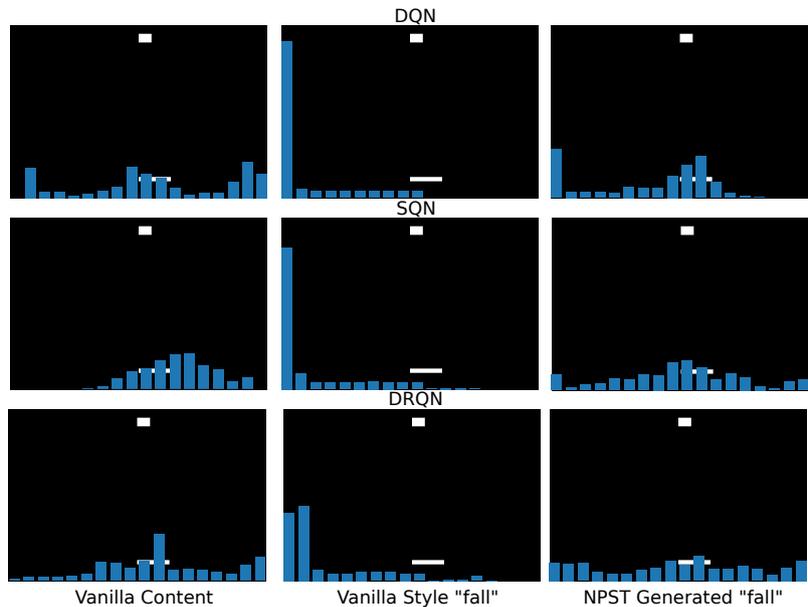


Figure 4: Paddle positions histograms for the case of transferring the “fall” *style*. Each row corresponds to a different network architecture. Each column corresponds to a different action. The Y-axis depicts the number of times each position is visited. All Y-axis are scaled in the range [0, 250]. The X-axis depicts the possible paddle positions within the game environment.

6.1. Experimental setup

Three different sets of five expert demonstrations are performed. These sets correspond to three different actions: one *content* action, and two different *style* actions (“nervous” and “fall”). The *content* action aims

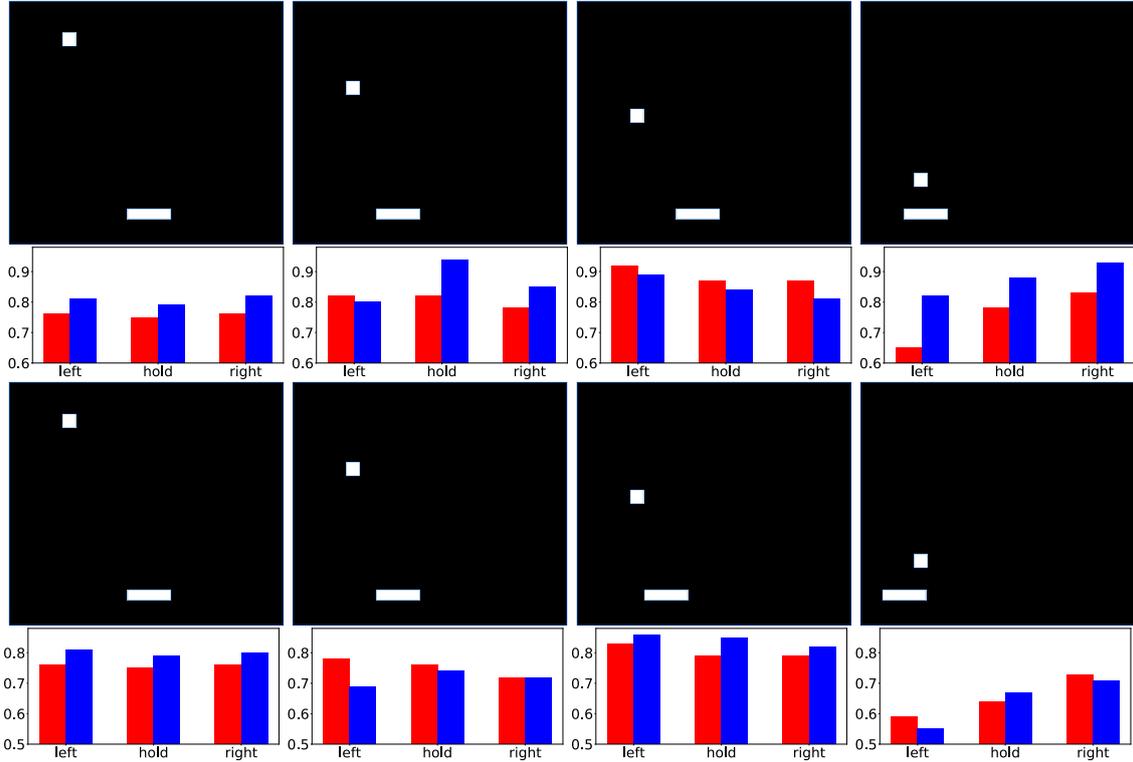


Figure 5: Results generated by the NPST algorithm with “nervous” *style* (top) and “fall” *style* (bottom). Red bars represent the outputs given by the original *content* network. Blue bars represent the output given by each of the networks obtained with the NPST algorithm. The game screenshots are the results of executing the NPST generated policy.

to win the game by fulfilling the goal of reaching the target pixel. The first *style* imitates a “nervous” behaviour or mood, tending to perform moves of one pixel around the same position. The second *style* imitates a “fall” movement, always tending towards the bottom side of the monitor.

The Maximum Entropy Deep IRL algorithm presented in section 2 is again used to obtain the reward functions that define these actions. These reward functions are then used to train the networks that are used with the NPST algorithm. For the IRL algorithm, a hand-crafted feature vector $\phi(s)$ and latent state space s are selected. In the case of the *content* demonstrations, s is a 256 element vector resulting from performing a flatten operation over the 16x16 pixel grid, and $\phi(s)$ classifies if the current pixel is the target pixel. The same latent state space s is used with the “fall” *style* demonstrations. Here, $\phi(s)$ classifies if the pixel belongs to the bottom row of the grid. In the case of the “nervous” *style* demonstrations, the latent state space s is a 16^3 element vector, where 16 is the number of possible vertical positions of the pixels, and 3 is the last three time steps. The feature vector $\phi(s)$ classifies if a movement was performed with the same starting and ending pixel.

For each of the three sets of demonstrations, 5 iterations of the IRL algorithm are executed. An IRL discount factor (γ) of 0.9 is used with a learning rate of 0.01. Three reward functions $R(s)$ are obtained

corresponding to the three base actions proposed (*content*, “nervous” and “fall”). For each of these base actions, three different neural networks, corresponding to the three architectures proposed, are trained using the same $R(s)$. These networks are trained using Q-learning and referred in the experiments as the Vanilla neural networks. These Vanilla neural networks are the base neural networks that will later be used to define the *content* and *style* in the NPST step. Experimental results obtained with these Vanilla networks are added as a baseline. The same Vanilla Content neural networks are introduced for the transferring of both of the *styles*. The input of the networks is the raw 16x16 pixel grid of the monitor, and the outputs correspond to the Q-values assigned to the three possible actions (stay still, go left, and go right).

Table 4: Hyperparameters for the “Grid-world paint” scenario.

Hyperparameter	Setting
<i>Shared between architectures</i>	
Activation	ReLU [18]
Initialization	Normal distribution
<i>DQN architecture</i>	
Input shape	(16, 16, 4)
Layers	(CL, CL, CL, FC, FC)
Layers configuration (Size, Kernel, Strides)	((32,8,4),(64,4,2),(64,3,1),(512,-,-),(4,-,-))
<i>SQN architecture</i>	
Input shape	(16, 16, 4)
Layers	(CL, FC, FC)
Layers configuration (Size, Kernel, Strides)	((32,8,4),(512,-,-),(4,-,-))
<i>DRQN architecture</i>	
Input shape	(16, 16, 1)
Layers	(CL, CL, CL, LSTM, FC)
Layers configuration (Size, Kernel, Strides)	((32,8,4),(64,4,2),(64,3,1),(256,-,-),(4,-,-))
<i>Q-learning and NPST</i>	
Grid-map size	16x16
Number of input time steps	4 (1 for DRQN)
Optimizer	Adam [12]
Loss function	Mean Squared Error
Number of actions	4
Discount (γ)	0.99
Experience Replay size	50000
<i>Q-learning</i>	
Learning Rate	1e-6
Initial Epsilon	0.9
Final Epsilon	0.01
Epsilon gradient	Lineal
Batch size	32
Exploration Epochs	100
Training Epochs	5000
<i>NPST Algorithm</i>	
Learning Rate	0.01
Number of iterations (N)	One full Grid-world episode
Batch size	100
L-BFGS-B internal iterations	1

The hyperparameters used for training the neural networks and performing the NPST algorithm are depicted in Table 4. The results for the generated NPST actions depicted in the following section are the average of 10 repetitions of the NPST algorithm.

Table 5: Experimental results for the “Grid-world paint” action introducing the “nervous” *style*.

Actions	$\mathcal{L}_{content}$	\mathcal{L}_{style}	<i>Nervous Moves</i>	<i>Average Steps</i>	<i>Wins</i> (%)	<i>Partial Wins</i> (%)
Vanilla Content DQN	—	1064.41	0	15	100	100
Vanilla Content SQN	—	1370.71	0	37	50	100
Vanilla Content DRQN	—	2205.27	0	15	100	100
Vanilla Nervous Style DQN	20.41	—	580	60	0	0
Vanilla Nervous Style SQN	31.34	—	299	60	0	0
Vanilla Nervous Style DRQN	22.67	—	599	60	0	0
NPST Nervous Generated DQN	127.08	3.20	46	58	20	80
NPST Nervous Generated SQN	14.50	3.75	28	60	0	60
NPST Nervous Generated DRQN	1.66	5.16	19	55	20	100

Table 6: Experimental results for the “Grid-world paint” action introducing the “fall” *style*.

Actions	$\mathcal{L}_{content}$	\mathcal{L}_{style}	<i>Nervous Moves</i>	<i>Average Steps</i>	<i>Wins</i> (%)	<i>Partial Wins</i> (%)
Vanilla Content DQN	—	1068.48	0	15	100	100
Vanilla Content SQN	—	1380.58	0	37	50	100
Vanilla Content DRQN	—	2209.30	0	15	100	100
Vanilla Fall Style DQN	115.50	—	0	60	0	0
Vanilla Fall Style SQN	78.63	—	0	60	0	0
Vanilla Fall Style DRQN	72.28	—	0	60	0	0
NPST Fall Generated DQN	14.03	3.05	43	60	0	40
NPST Fall Generated SQN	10.14	3.47	46	58	20	50
NPST Fall Generated DRQN	2.60	4.43	10	51	40	100

6.2. Results

The results obtained with the NPST algorithm are depicted in Table 5 for the case of transferring the “nervous” *style* and Table 6 for the case of transferring the “fall” *style*. In these experiments, the *Nervous Moves* parameter measures the number of times the agent performed a full vertical direction swap (i.e. going upwards then downwards then upwards again). The *Average Steps* parameter measures the average number of steps performed by the agent per episode. Finally, the *Partial Wins*(%) parameter measures the number of times the agent ended the episode in the same column as the target. This parameter was introduced to complement the *Wins*(%) parameter due to the low tolerance of error defined by the environment.

The number of Nervous Moves performed was increased when introducing the “nervous” *style* in the DQN and DRQN architectures with respect the Vanilla Content and Vanilla Fall networks. One hypothesis behind the high number of *Nervous Moves* obtained in the “fall” networks is due to the combination of the

“fall” *style* and *content*. The “fall” *style* tries to constantly reach the bottom while the *content* tries to bring up the agent to the middle position where the target is located. This provokes multiple vertical direction changes. Similar to what happened in the “Catch-ball” game scenario, the number of wins decreased when introducing the *styles* with respect the Vanilla Content networks but increased with respect the Vanilla Style networks. The $\mathcal{L}_{content}$ obtained with the NPST algorithm using the DQN architecture was unusually high due to some outliers produced in the first steps of the NPST execution with unusually high $\mathcal{L}_{content}$ values. As in the “Catch-ball” scenario, the results obtained with the generated control policy are a combination of the *content* and *style* policies.

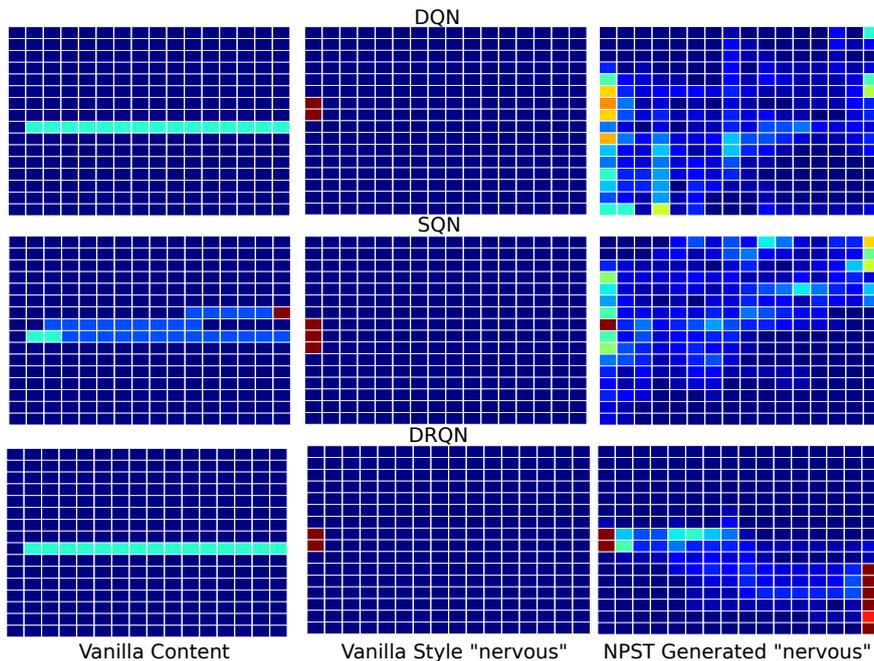


Figure 6: Robot end-effector heatmap results for transferring the “nervous” *style*. Each row corresponds to a different network architecture. Each column corresponds to a different action. Warm colors represent monitor pixels that are recurrently visited. Cold colors represent the less visited pixels. The graphs show the cumulative results over the 10 repetitions performed. The color scale of the heatmap goes from 0 to 50. Pixels with values higher than 50 are capped to this value

Fig. 6 and Fig. 7 are heatmaps representing the most visited states for each of the generated policies. Warmest colors depict the most recurrently visited pixels, and the coldest colours represent the least visited pixels. As expected, in these figures, the policies transferred with the “fall” *style* show a clear preference for the bottom states in the case of the DQN and DRQN architecture. The same preference is also depicted for the DRQN architecture while transferring the “nervous” *style*. The rest of the policies do not depict any relevant preference in terms of vertical position.

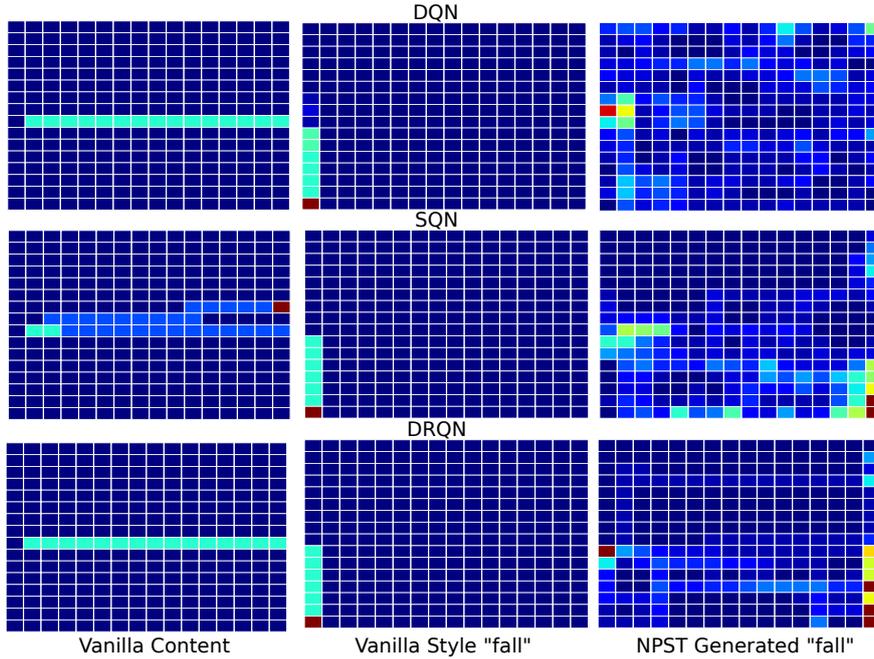


Figure 7: Robot end-effector heatmap results for transferring the “fall” *style*. Each row corresponds to a different network architecture. Each column corresponds to a different action. Warm colors represent monitor pixels that are recurrently visited. Cold colors represent the less visited pixels. The graphs show the cumulative results over the 10 repetitions performed. The color scale of the heatmap goes from 0 to 50. Pixels with values higher than 50 are capped to this value.

7. Discussion

One of the first ideas that has to be taken in account when studying the results proposed in this paper is that by introducing Style Transfer a new optimization problem is defined. With Style Transfer, the goal is not to find the optimal policy for the execution of the content action, but to find the optimal policy that is able to execute the content action with the selected *style*. This is relevant in the case of the results obtained in the experiments proposed. Here, the percentage of wins was reduced with the introduction of the two proposed styles. This reduction, however, came with a decrease in the total Style Transfer loss depicting a more optimal policy for the Style Transfer problem. The reduction in the percentage of wins can be therefore depicted as an expected or even desired behavior. Players with some emotional bias are expected to have a lower performance when executing the action that players without it.

The qualitative and quantitative results obtained in this paper show how NPST is able to transfer and combine the behavior of different policies into new ones. This allows the introduction of future applications introducing *styles* in robotic actions. For social robot applications, these *styles* can be defined as different human emotions. More pragmatic applications may define these *styles* as a way to improve the performance of the robot using only a set of demonstrated expert trajectories that can be generalized to the full range of robotic actions.

The formulation of the NPST algorithm presents the limitation of requiring to perform a new execution

each time a new trajectory has to be generated. The generated policy is defined by the output of the NPST execution. At the same time, the NPST algorithm presents some relevant advantages with respect other State of the Art methods. One critical advantage is that the NPST algorithm works without requiring an additional loss network. The same network is used to generate the trajectory and perform the loss extraction. This reduces computational times and the overall complexity of the Style Transfer process. In addition to this, the NPST framework is designed to work with robots. Within the authors knowledge, this is the first time Style Transfer is introduced within a Reinforcement Learning framework for the generation of *style* control policies and its execution with robotic actions.

8. Conclusions

Style Transfer aims to transform a certain input, adapting it via a certain *style* without changing the original *content*. It has been proposed and successfully introduced in a number of fields (fine arts, natural language processing, and fixed trajectories). By means of this work, a Neural Policy Style Transfer (NPST) algorithm has been proposed to perform Style Transfer with control policies.

The control policies are defined by neural networks that express Reinforcement Learning Q-value functions. The generated action is initialized to have the same model and number of parameters as the *style* action. The input must correspond to the observation or state space, and the output must express the Q-value function. The base neural networks are trained via Maximum Entropy Deep IRL algorithms that learn reward functions, taught by human demonstrators and NPST-ready.

Two sets of experiments were performed in this paper, each corresponding to the two scenarios presented. The “Catch-ball” game is inspired by the Deep Reinforcement Learning classical Atari games; and the “Grid-world paint” scenario includes a full-sized humanoid robot, equivalent to a grid-world problem in the real world, based on previous works of the authors. In both sets of experiments, the results show a clear influence in the execution of the policies after transferring each of the *styles*. Three different architectures were introduced to test the NPST algorithm. The results show a clear influence of the transferred *style* in the generated action while keeping the *content* goal. The resulting control policy introduces elements of both the proposed *style* and the defined *content*.

9. Acknowledgment

The research leading to these results has received funding from RoboCity2030-DIH-CM Madrid Robotics Digital Innovation Hub (“Robótica aplicada a la mejora de la calidad de vida de los ciudadanos. fase IV”; S2018/NMT-4331), funded by “Programas de Actividades I+D en la Comunidad de Madrid” and cofunded by Structural Funds of the EU. The authors thank Bartek Lukawski for his valuable collaboration.

References

- [1] Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first international conference on Machine learning*, number 346 in ICML '04, page 1.
- [2] Amaya, K., Bruderlin, A., and Calvert, T. (1996). Emotion from Motion. In *Graphics interface '96, Canadian Human-Computer Communications Society, GI '96*, pages 222–229, Toronto, Ont., Canada, Canada. Canadian Information Processing Society.
- [3] Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *IEEE Signal Processing Magazine, Special Issue on Deep Learning for Image Understanding*.
- [4] Brand, M. and Hertzmann, A. (2000). Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*, pages 183–192.
- [5] Brejl, R. K., Purwins, H., and Schoenau-Fog, H. (2018). Exploring deep recurrent q-learning for navigation in a 3d environment. *EAI Endorsed Transactions on Creative Technologies*, 5(14).
- [6] Dumoulin, V., Shlens, J., and Kudlur, M. (2017). A learned representation for artistic style. In *International Conference on Learning Representations 2017*.
- [7] Fernandez-Fernandez, R., Victores, J. G., Estevez, D., and Balaguer, C. (2018a). Real evaluations tractability using continuous goal-directed actions in smart city applications. *Sensors*, 18(11).
- [8] Fernandez-Fernandez, R., Victores, J. G., Estevez, D., and Balaguer, C. (2018b). Robot Imitation through Vision, Kinesthetic and Force Features with Online Adaptation to Changing Environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6546–6551, Madrid. IEEE.
- [9] Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- [10] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423. IEEE.
- [11] Holden, D., Habibie, I., Kusajima, I., and Komura, T. (2017). Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications*, 37(4):42–49.
- [12] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- [13] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- [14] Lee, J., Xie, Z., Wang, C., Drach, M., Jurafsky, D., and Ng, A. Y. (2019). Neural Text Style Transfer via Denoising and Reranking. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 74–81, Minneapolis, Minnesota. Association for Computational Linguistics.
- [15] Luan, F., Paris, S., Shechtman, E., and Bala, K. (2018). Deep Painterly Harmonization. *Computer Graphics Forum*, 37(4):95–106.
- [16] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv preprint*, pages 1–9.
- [17] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. a., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- [18] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, page 807–814. Omnipress.
- [19] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*.
- [20] Tenenbaum, J. and Freeman, W. (1997). Separating style and content. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press.
- [21] Unuma, M. and Takeuchi, R. (1991). Generation of human motion with emotion. In *Computer Animation'91*, pages 77–88. Springer.
- [22] Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.
- [23] Wulfmeier, M., Ondruska, P., and Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. *arXiv preprint*.
- [24] Zhao, H., Zheng, J., Wang, Y., Yuan, X., and Li, Y. (2020a). Portrait style transfer using deep convolutional neural networks and facial segmentation. *Computers & Electrical Engineering*, 85:106655.
- [25] Zhao, H.-H., Rosin, P. L., Lai, Y.-K., and Wang, Y.-N. (2020b). Automatic semantic style transfer using deep convolutional neural networks and soft masks. *The Visual Computer*, 36(7):1307–1324.
- [26] Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438.