




Article

Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios

Niki Efthymiou ^{1,*} , Panagiotis Paraskevas Filntisis ¹, Gerasimos Potamianos ² and Petros Maragos ¹

¹ School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Greece; filby@central.ntua.gr (P.P.F.); maragos@cs.ntua.gr (P.M.)

² Department of Electrical and Computer Engineering, University of Thessaly, 38221 Volos, Greece; gpotam@ieee.org

* Correspondence: nefthymiou@central.ntua.gr

Abstract: This paper proposes a novel lightweight visual perception system with Incremental Learning (IL), tailored to child–robot interaction scenarios. Specifically, this encompasses both an action and emotion recognition module, with the former wrapped around an IL system, allowing novel actions to be easily added. This IL system enables the tutor aspiring to use robotic agents in interaction scenarios to further customize the system according to children’s needs. We perform extensive evaluations of the developed modules, achieving state-of-the-art results on both the children’s action BabyRobot dataset and the children’s emotion EmoReact dataset. Finally, we demonstrate the robustness and effectiveness of the IL system for action recognition by conducting a thorough experimental analysis for various conditions and parameters.

Keywords: visual perception; visual learning; incremental learning; action recognition; emotion recognition; child–robot interaction



Citation: Efthymiou, N.; Filntisis, P.P.; Potamianos, G.; Maragos, P. Visual Robotic Perception System with Incremental Learning for Child–Robot Interaction Scenarios. *Technologies* **2021**, *9*, 86. <https://doi.org/10.3390/technologies9040086>

Academic Editor: Fillia Makedon

Received: 16 October 2021
Accepted: 10 November 2021
Published: 15 November 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The robotic systems’ perception is considered fundamental for their development, especially when we refer to social robots. Apart from that, for a robot to interact with others, react to them and socially communicate, it needs to be aware of its actions. Consequently, since robots with social skills are incrementally employed in a great variety of applications—such as healthcare [1,2], companionship [3], education [4], and entertainment [5]—there is a need to design more advanced perception systems for them.

An exciting and challenging field to develop intelligent and robust recognition systems is that of Child–Robot Interaction (CRI). Children’s behavior and natural characteristics, for example, articulation, spontaneity, and body height differ from those of adults, and perception systems need to be developed specifically for children in tasks such as action and speech recognition [6,7]. This difference, along with the lack of children-related big data for training recognition algorithms turn usual recognition tasks into a seriously challenging problem.

Due to the great variety of robotic applications, CRI, apart from engineers, appeals also to a broad scientific interdisciplinary sector, including therapists, psychologists, educators, and teachers. Therapeutic purposes are among the most frequent uses of robotic agents with children, that is, autism [8], pediatric rehabilitation [9], and diabetes management [10]. Children’s mental and cognitive development and CRI’s effects on them have been studied extensively [11,12]. Numerous educational scenarios with different learning subjects have been implemented, such as learning handwriting [13,14], a second language [15], and social emotional learning [16]. Additionally, many CRI scenarios have been designed for exploring special research goals, for example, understanding child engagement [17], defining parameters in a long-term interaction [12], and finding appropriate ways to adjust the curriculum on CRI [18]. This extensive range of applications and research

results indicate the great potentials of robotic use in children's edification. Consequently, providing non-technical child professionals with robust robotic systems with powerful perception systems could bring forward momentum in the exploitation of CRI [19].

Regarding robot perception, during Human–Robot Interaction (HRI), the analysis of the human emotional state is crucial in developing empathetic robots [20]. Empathy for robots is essential to decode human movements and expressions which carry emotional information. Consequently, the robotic agents can adapt their behavior and actions towards the user appropriately, aiming to establish a healthy long-term interaction relationship [21]. Additionally, a robot with the ability to perceive different body movements and actions could create rich interaction scenarios [22].

In this work, we propose a novel visual robotic perception system for CRI scenarios. Following the above directions, while developing our system, we are dealing with several research questions. How can the tremendous opportunities that deep architectures offer to the computer vision field be employed in order to create a visual perception system for robotic applications specifically for children? Is it possible to leverage these deep architectures to create a CRI system with the joint capability of both action and emotion recognition and an appropriate trade-off between computational efficiency and performance? Furthermore, can IL be used in order to allow the perception system to recognize new actions without forgetting older ones, and which class of IL methods performs better?

To the best of our knowledge, this is the first work considering IL for action recognition in CRI to tackle the fact that new classes have to be recognized by the system in separate edutainment scenarios. Besides this core novelty of our work, we also examine several parameters of the system, in order to balance performance and computational efficiency, and propose a combined visual perception system for action and emotion recognition in the context of CRI. In short, the main contributions of this work are threefold:

- We propose a novel system based on powerful lightweight deep neural network-based architectures for action and affect recognition.
- We evaluate the emotion and action recognition architectures thoroughly, achieving state-of-the-art results on two children's databases. We also perform ablation studies for both modules regarding the effect of the pretraining scheme and the number of sampled segments in the resulting performance and computational efficiency.
- We wrap the action recognition architecture around an IL system that allows novel actions to be easily added. Specifically, the proposed perception system gives the opportunity to a non-technical expert to extend and adjust the action classes contextually. This is achieved by extending the existing and well-known IL technique called iCaRL (i.e., Incremental Classifier and Representation Learning) [23], so that it can be applied on videos instead of frames, within the premises of the Temporal Segment Networks (TSN) framework. We perform extensive evaluations of the IL system under various parameters and conditions, and compare it against other IL methods, proving its robustness, efficacy, and lightweightness.

The remainder of the paper is organized as follows: Section 2 presents previous works in action and emotion recognition along with the IL applications in HRI scenarios. In Section 3, we present our visual perception system and its modules extensively. Section 4 includes our experimental results and ablation studies on the EmoReact and BabyRobot databases. Finally, in Section 5, we combine and summarize our important findings and future directions, and Section 6 concludes the work.

2. Related Work

Recently, the research for designing intelligent perception systems for robotic agents, especially social robots, has gained interest due to the advances in recognition techniques like action, speech, and emotion recognition. However, as mentioned above, the human recognition models trained on adults' data show significantly decreased accuracy when applied to children's data [22]. Thus, the presented related work focuses on designed or evaluated systems on children data, when available.

Regarding child affect recognition, in [24], Castellano et al. present a system that perceives affective expressions of children while playing chess with an iCat robot and modifies the robot behavior to result in a more engaging and friendly interaction. An adaptive robot behavior based on the perceived emotional responses was also developed for a NAO robot in [25]. Goulart et al. propose in [26] an equivalent computational system using visual information captured from RGB and infrared thermal cameras. Filippini et al. [17], classify children's emotional state during interactions with Mio Amico Robot using thermal signal analysis and managed to understand their engagement level. In [27], Lopez-Rincon proposes a Convolutional Neural Network (CNN) to identify children's facial expressions. Lastly, in [28], we proposed a two-branch architecture leveraging both body posture and facial expressions for identifying children's emotions during CRI scenarios.

Contrary to the popularity of the human action recognition problem, child action recognition is not among the famous computer vision problems. A notable work [29] by Marinoiu et al. proposed a CNN architecture for action and continuous emotion recognition, deploying 3D skeleton data of the participants focusing on children with Autism Spectrum Disorder (ASD). To the best of our knowledge, this is the only work combining jointly children action and emotion recognition along with our proposed perception system. Our previous work [6] focused on exploring different feature extraction approaches, encoding methods, and fusion techniques for proposing a multi-view system for action recognition during CRI. Recently, Zhang et al. [30] dealt with the action recognition problem for ASD children and proposed a Long Short-Term Memory based network fed with the extracted children's skeleton after a denoising filter.

Several algorithms have been proposed for IL in the computer vision literature since it constitutes a crucial attribute for real-world deployment of any machine learning system. Since there are not related works concerning IL for CRI, we will give a general overview of the IL field and its most popular methods, several of which are used and evaluated in this work. An interesting categorization of the developed neural network methods for continual lifelong learning is presented in [31] according to how they mitigate catastrophic forgetting. Conceptually, these approaches can be divided into (i) the regularization methods imposing constraints on the update of the neural weights (i.e., [32–34]), (ii) the dynamic architectures concerning those that change their architectural properties such as the number of the used neurons (i.e., [35]), and (iii) the complementary learning systems and memory replayed methods (i.e., [23,36–38]).

In End-to-End Incremental Learning (EEIL) [36], a combination of a memory dataset (also called experience replay) and knowledge distillation—which was initially proposed for transfer learning between different networks—was employed to incrementally add new images and classes to an image classification network. iCaRL [23] proposed a similar system for IL over long periods by decoupling the data representation and the classifier. On the other hand, [37] proposed using generative adversarial networks to mimic data the model has seen in the past, while [39] proposed a brain-inspired replay of the internal representations of the model. In [40], the authors proposed the IL2M network (Incremental Learning with Dual Memory), which rectifies the predictions using a dual memory and is based on the saved certainty statistics of predictions of classes from previous tasks. The Memory Aware Synapses (MAS) [34] method is based on the online computation of the importance of neural network parameters. Finally, Learning without Forgetting (LwF) [33] used knowledge distillation and the output of old tasks in new data to avoid forgetting old tasks. To study the class-incremental learning on the image classification task in-depth, we refer the reader to Masana et al. [41].

Concerning continual learning in HRI, Churamani et al. [42] discuss its importance for creating fully adaptive affective robots and how to utilize it for perception and behavior learning with adaptation. In [43], a CNN classifier for object detection was enriched with incremental learning capabilities to add new object classes for classification, while in [44], adaptive incremental learning through interaction of social robots with humans was proposed. The online incremental classification resonance network in [45] imbued the

face identification system of the Mybot robot increasing the number of faces it can identify. Tuyen et al. [46] also used an incremental learning model, which identified the cultural traits of humans it interacted with. Finally, Lesort et al. [47] summarize real use cases of continual learning for robotic applications, reasons for deploying incremental learning, and the challenges faced in these tasks.

3. Materials and Methods

An overview of the perception system is presented in Figure 1. Considering the constant need for introducing new action classes during new edutainment scenarios, the action recognition module is wrapped in an IL system. On the other hand, the emotion recognition module does not need to be constantly modified and is trained only once. While research has shown that personalizing emotion recognition in the context of continual learning increases performance [48,49], the same can be argued for action recognition (personalizing) [50,51]. In this work, we focus on IL in the context of allowing the addition of new classes to the system—personalized adaptation is out of our scope. We will first present the action and emotion recognition modules and then the IL action system.

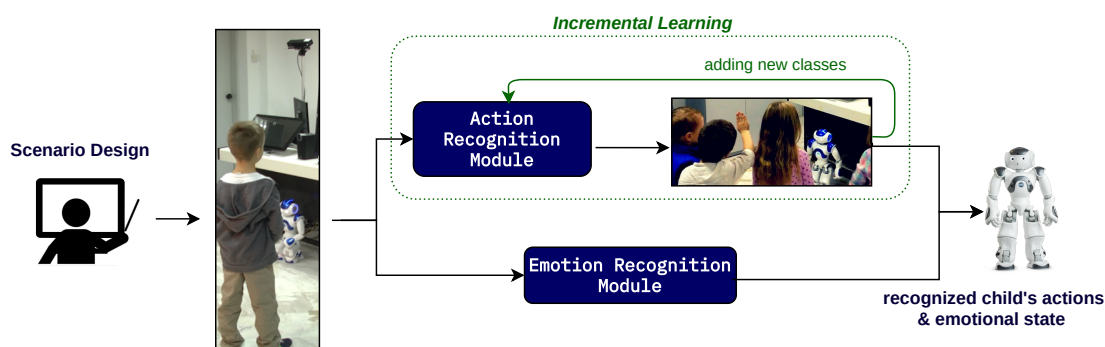


Figure 1. The proposed visual perception system for child–robot interaction scenarios.

3.1. Action Recognition

The action recognition module (Figure 2) is based on the TSN system [52]. The TSN system and practices have been initially used for large-scale action recognition. Under this system, the modality of interest (RGB or optical flow) in the input video V is split into K different segments $\{S_1, S_2, \dots, S_K\}$ of equal duration, and then a snippet T_k of N consecutive frames is sampled from each segment. Subsequently, a backbone CNN is applied to each snippet, represented as $F(T_k; W_{cnn})$, producing a feature vector $L_k(V)$ for each snippet, and then a fully connected layer $H(L_k; W_{fc})$ is applied on $L_k(V)$ to produce the class scores $S_k(V)$. Finally, the segmental consensus function G is used to produce the final class scores from those of each snippet. Common choices of the segmental consensus function include averaging, weighted averaging, or the maximum; here, we use simple averaging.

The whole process is described by the following equation:

$$S = G(S_k) = G(H(F(T_k; W_{cnn}); W_{fc})|_{k \in K}). \quad (1)$$

Traditionally, in the TSN system, both optical flow and RGB images are used to train two different networks separately, which are then fused to produce the final output.

The random sampling of the TSN system allows long-term temporal modeling without introducing redundant information that exists in the sequential frames of a video. In addition, it reduces overfitting and allows for quicker training and inference, which are both crucial in CRI, where usually a small amount of data are available, and real-time recognition is imperative. In order to force the networks to focus on the child and its actions, we first perform pose tracking on the input video, and then we crop the image

around the child during temporal sampling using the detected skeleton. The same process is applied to the flow stream as well.

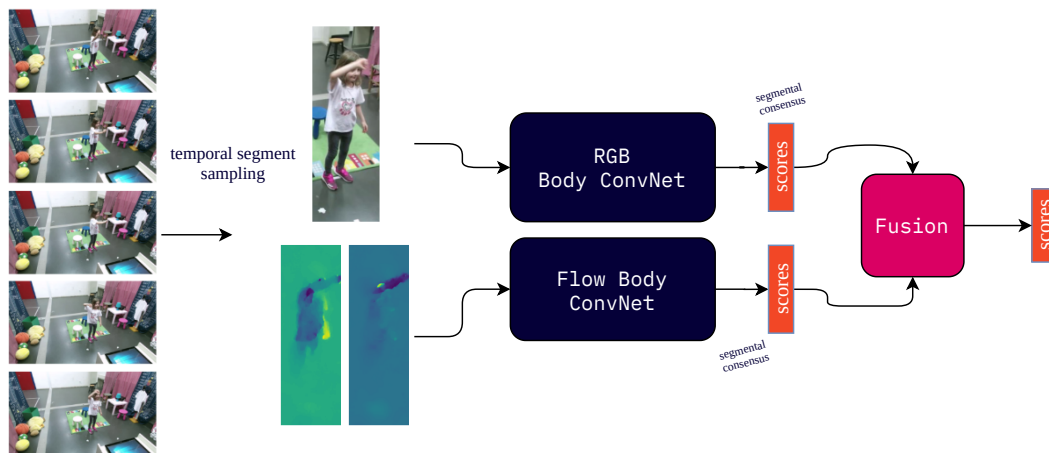


Figure 2. The TSN system used for action and emotion recognition in the robotic edutainment system.

3.2. Emotion Recognition

The emotion recognition module follows the same principles as the action recognition one, both for convenience and due to the proven efficacy of TSNs for emotion recognition [53,54]. In comparison, however, to action recognition, the input video (both RGB and optical flow) is cropped around the child’s face by using facial landmarks detected using OpenFace [55]. In addition, the emotion recognition module is trained for multi-label classification by applying a sigmoid function on the output scores instead of softmax, and training with binary cross-entropy loss (BCE). Furthermore, apart from the categorical emotions output, we add another fully connected layer that enables predictions in the Valence axis of the dimensional emotion model. The dimensional emotion model was initially proposed in [56], and its Valence axis measures how “positive” an emotion is. As a result, the network is trained concurrently with both the BCE loss of the categorical predictions and the mean-squared error loss for the dimensional emotion predictions, as a form of multi-task learning. The choice of training a model with both models of affect is motivated by the fact that many recent emotion recognition databases (such as BoLD [54], EMOTIC [57], and EmoReact [58]) include both multi-label categorical emotion annotations and dimensional emotion annotations.

3.3. Incremental Learning for Action Recognition

When a deep neural network is retrained with new classes, it suffers from *catastrophic forgetting*: the knowledge for the previous classes tends to be forgotten, and its performance decreases dramatically. A naive approach to tackle this would consider retraining the network with an entire dataset of all new and old classes. However, it is easy to see that this becomes computationally prohibitive, especially when new classes need to be added continuously to the network. In order to allow the action recognition module to have incremental learning capabilities, we wrap it around an IL system. To do this, we extend the iCaRL method [23], allowing it to be applied to videos within the premises of the TSN framework and CRI.

Under this system, several samples (called exemplars) are retained for each class that the system has seen, called the *memory budget* B . In the literature [41], there are two methods for defining B . The usual method involves a fixed budget size as the number of classes varies, while the second approach involves a fixed number of exemplars-per-class, resulting in a linear increase in the budget as the number of classes in the system increases. In this second approach, if we defined as E the number of exemplars-per-class and C the total number of seen classes, we have $B = C \cdot E$. In our IL setup, we follow the second approach

since in our specific use-case we do not expect the number of classes to become large-scale; this linear increase in the budget becomes a bottleneck only in large-scale applications.

When a new *phase* of IL takes place for the action recognition module (i.e., when one or more new classes need to be introduced to it), the samples of the new classes are combined with the exemplars in the memory to form the *combined* dataset, which is then used to retrain the action recognition module. During training, the network learns to minimize both the cross-entropy loss, as well as a distillation loss for the old classes [23]. Simple random sampling is used to select the exemplars that are held out from each class. Various works have shown [36,41] that random sampling achieves competitive results to other approaches such as herding [59] (which uses the distance of class samples to the mean exemplar).

In contrast to other approaches of IL, classification of new samples in iCaRL (and our extended method) uses a representation-based classifier. For each class the system holds in the memory a prototype vector M_i , which is the average feature vector of all exemplars of the class, $M_i = \frac{1}{E} \sum_{j=1}^E L_j(V)$. Within the premises of the TSN framework, the feature vector of each exemplar $L(V)$ is calculated as:

$$L(V) = G(L_k(V)) = G(F(T_k; \mathbf{W})|_{k \in K}). \quad (2)$$

Then, a new sample video V_n is assigned the class label with the nearest prototype vector $\arg \min_{i=1 \dots C} \|L(V) - M_i\|$.

3.4. Edutainment Scenario Example

Next, we present an example of an edutainment scenario that can be implemented within the proposed robotic system, underlining the sequence of the child and robot actions along with the incremental learning phase. In this scenario, the concept of angles in mathematics is the learning subject.

ROBOT: Today, we will learn together the obtuse angles. Do you remember what an angle is?

CHILD: [Points to a corner of the room]

ROBOT: [Recognizes the pointing gesture. Then responds] You are pointing at something. Could you make an angle with your hands to show it clearly?

CHILD: Yes! [says while expressing happiness].

ROBOT: [Recognizes the emotion and the action, and says] Great, this is an acute angle! I've read about the obtuse angles, but it was difficult for me. Could you show me with your hands?

TEACHER: They don't know yet. We are going to learn it and then they'll show to you.

[While the lesson is ending. . .]

ROBOT: I would like some of you to perform actions to depict the obtuse angles. Then, you can examine if I learned them and if I remember the acute and the right angles that you've taught me once before.

[While the children are performing their actions, the robot is collecting exemplars for the new classes. After the incremental learning phase, the robot asks children to perform actions, both new and old, and recognize them along with their emotional states.]

3.5. Databases and Training Methods

We will now describe the databases used for evaluating the previously described methods and include details for the training process.

3.5.1. Action Recognition

We evaluate the action recognition module on the BabyRobot action database [6]. The BabyRobot database contains 25 children, aged six to ten years old, performing various actions collected while playing a game with multiple robotic agents. The dataset features a total of 13 actions: painting a wall, cleaning a window, driving a bus, swimming, dancing, working out, playing the guitar, digging a hole, wiping the floor, ironing a shirt, hammering

a nail, reading a book, and background movement. The BabyRobot database is multi-view, featuring different Kinect cameras placed around the room. For our single-view action evaluation, we use only camera Kinect #1, located at the top right corner regarding the child, as shown in Figure 2, providing a full-body view of the child.

For the backbone CNN of the action recognition module, we use a Batch Normalization Inception (BNInception) [60] architecture for the RGB and Flow streams and consider two different pretraining schemes: pretraining on the Kinetics [61] action recognition dataset, or alternatively pretraining on the ImageNet database. The BNInception architecture was selected because the weights of the pretrained model in both of these databases are publicly available. To have a direct comparison with the best-published result in [6], we perform leave-one-child-out cross-validation. The training and scheduling parameters are empirically selected as follows: we train each network for 60 epochs with stochastic gradient descent (SGD), starting with a learning rate of 1×10^{-4} and decreasing it by a factor of 10 at 20 and 40 epochs. We use the default TSN values of one frame length for the RGB segments and five frames for the Flow segments.

3.5.2. Emotion Recognition

The emotion recognition module is evaluated on the EmoReact [58] dataset. This contains 1102 videos of 63 children, aged between 4 and 14, expressing emotions while discussing different topics. The dataset is collected from the YouTube channel React and features multi-label annotations on eight different categorical emotions: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, Frustration. Furthermore, in each video the valence is annotated on a scale of 1 to 7, with 1 corresponding to a completely negative emotion and 7 to a completely positive emotion.

The CNN architecture we use for both modalities is a standard residual network with 50 layers (ResNet50) [62]. The networks are trained using binary cross-entropy loss for the categorical emotions and mean squared error for the valence. We perform the same ablation studies as with the action recognition module. We report results on the test set of EmoReact (the dataset includes a standard train-validation-test split) using the balanced (per-class) and unbalanced area under the receiver operating characteristic (ROC AUC) for the categorical predictions, and mean-squared error for the dimensional emotion prediction.

3.5.3. Incremental Learning

To evaluate the IL method, we create an augmented and more challenging setup. We merge the BabyRobot action dataset with the additional BabyRobot gesture dataset [22], which includes seven gestures (make a circle with hands, tell someone to come closer, greet someone, nod, point at something, tell someone to sit down, and make a stop sign) performed by the same children included in the action dataset and captured by the same camera. After merging, the augmented dataset now comprises a total of 20 classes. To speed up the training and evaluation process, we create a training/testing split with 20 children in the training set and 5 children in the test set (instead of leave-one-child-out cross-validation which would geometrically increase the number of models needed to train). We use the same training scheme as before (i.e., adding novel classes to the action recognition module) during each IL phase (60 epochs with learning rate reduction at 20 and 40 epochs). We repeat all experiments 10 times to account for different seeds (averaging to get the final results). The order of the classes in each run is selected randomly.

4. Results

We will now present the results of the action and emotion recognition modules individually and of the IL system under various conditions. The source code for all models and experiments can be found at <https://github.com/filby89/incremental-learning-CRI> (accessed on 1 November 2021).

4.1. Action Recognition

4.1.1. Number of Segments

In our first ablation study, we explore the effect of the number of sampled segments for the RGB and optical flow (Table 1), as a function of both the final performance of each modality and the computational complexity of the system (in terms of elapsed seconds per train and inference epoch). We can see that the system's performance increases as we increase the number of segments used in the TSN framework for both modalities. However, after a threshold, the increased computational burden does not reflect equal performance gains. Accordingly, we select five segments for both the RGB and the optical flow for the rest of the experiments on the action recognition module.

Table 1. Performance and elapsed time per training and validation epoch of the *action recognition module* for varying numbers of sampled segments.

Segments	Accuracy (%)	Time/Training Epoch (s)	Time/Validation Epoch (s)
RGB			
1	36.74	5.2	0.4
3	40.95	6.0	0.8
5	47.43	8.8	1.0
10	49.56	14.6	1.4
Flow			
1	58.75	5.4	0.6
3	71.77	10.3	1.2
5	75.96	16.3	1.8
10	76.82	31.3	3.2

4.1.2. Pretraining

For our second ablation study in Table 2, we present varying pretraining schemes. We can see the intuitive fact that pretraining on the Kinetics action databases results in much higher performance than ImageNet pretraining for both modalities. In the same table, we also present the final fusion result obtained with an empirical weighted average scheme (assigning 0.8 weight to optical flow and 0.2 to RGB), as well as the previous state-of-the-art method of Dense Trajectory Ensemble features and the C3D convolutional network of [6]. We see that the optical flow modality outperforms the previous results, and combination with the RGB modality results in an additional, albeit small, performance improvement.

Table 2. Results of the *action recognition module* on the BabyRobot action dataset using leave-one-child-out cross-validation.

Model	Accuracy (%)
RGB-Kinetics	47.43
RGB-ImageNet	42.46
Flow-Kinetics	75.96
Flow-ImageNet	65.26
RGB-Kinetics + Flow-Kinetics	76.55
RGB-ImageNet + Flow-ImageNet	64.37
Dense Traj. Ensemble [6]	74.15
C3D [6]	59.38

4.2. Emotion Recognition

4.2.1. Number of Segments

Like in the action recognition module, while generally increasing the number of segments increases the performance (the only exception being five RGB segments where the

balanced ROC AUC is lower than in three segments), we have diminishing gains (Table 3). We find that an appropriate trade-off between performance and the computational burden is to select in the rest of our experiments three segments for the RGB modality and five segments for the Optical Flow. Note that results in the unbalanced case are higher, because the network learns to classify better more frequent emotions in the dataset (an example being Happiness).

Table 3. Performance and elapsed time per training and validation epoch of the *emotion recognition module* for varying numbers of sampled segments.

Segments	ROC AUC		MSE	Time/Training Epoch (s)	Time/Validation Epoch (s)
	Balanced	Unbalanced			
RGB					
1	0.683	0.773	0.032	8	6
3	0.713	0.786	0.030	26	16
5	0.703	0.785	0.030	39	27
10	0.716	0.789	0.029	73	53
Flow					
1	0.580	0.739	0.038	36	22
3	0.583	0.737	0.036	102	73
5	0.615	0.756	0.036	170	123
10	0.636	0.754	0.036	351	256

4.2.2. Pretraining

We compare pretraining our networks on the ImageNet dataset against pretraining them on the largest facial expression dataset, AffectNet [63]. We have trained a ResNet50 on AffectNet, achieving 59.49% accuracy on the validation set (test set is not available). The effect of pretraining can be seen in Table 4. We observe that the AffectNet RGB pretrained model achieves higher performance than the ImageNet one. Compared to the RGB ones, the Flow network achieves a lower ROC AUC, and pretraining on AffectNet results in higher unbalanced ROC AUC and lower mean-squared error. Average fusion increases the final ROC AUC, but not the mean squared error (MSE) of the predicted valence. In the same Table, we also list the previous state-of-the-art result on the EmoReact dataset [58], which used features from the OpenFace [55] framework along with a support vector machine (SVM), showing that our method achieves significantly better emotion recognition performance.

Table 4. Results of the *emotion recognition module* on the categorical and continuous emotions of the EmoReact dataset.

Model	ROC AUC		MSE
	Balanced	Unbalanced	
RGB-AffectNet	0.713	0.786	0.030
RGB-ImageNet	0.657	0.735	0.044
Flow-AffectNet	0.615	0.756	0.036
Flow-ImageNet	0.643	0.752	0.039
RGB-ImageNet + Flow-ImageNet	0.682	0.765	0.039
RGB-AffectNet + Flow-AffectNet	0.725	0.789	0.031
RGB-AffectNet + Flow-ImageNet	0.724	0.791	0.032
OpenFace with SVM [58]	0.62	-	-

4.3. Incremental Action Learning

As we also saw during the evaluation of the Action Recognition module in Table 2, the RGB modality offers a minuscule increase in accuracy but takes a considerable amount of

time to train. As a result, we opt to remove the RGB modality in the final action recognition system and use only the optical flow for the IL experiments.

4.3.1. Ablation Study—Number of Exemplars per Class

Our first evaluation study for IL compares the iCaRL method for TSNs with other methods that use experience replay (modified for the TSN framework as well): EEIL [36], IL2M [40], and simple fine tuning (i.e., using only training with the combined dataset). Results for various numbers of exemplars-per-class ($E = 2, 5, \text{ and } 10$) and two different numbers of total phases ($T = 5, 10$) can be seen in Figure 3. Note that depending on the total number of phases, different number of classes are added per phase (the total number of classes 20 divided by the number of phases). For five total phases, four new classes are added per phase, and the accuracy x-axis starts at $x = 4$ while the forgetting x-axis one phase later at $x = 8$. Similarly, for $T = 10$ the number new classes added per phase are 2, the accuracy x-axis starts at $x = 2$ and the forgetting one at $x = 4$.

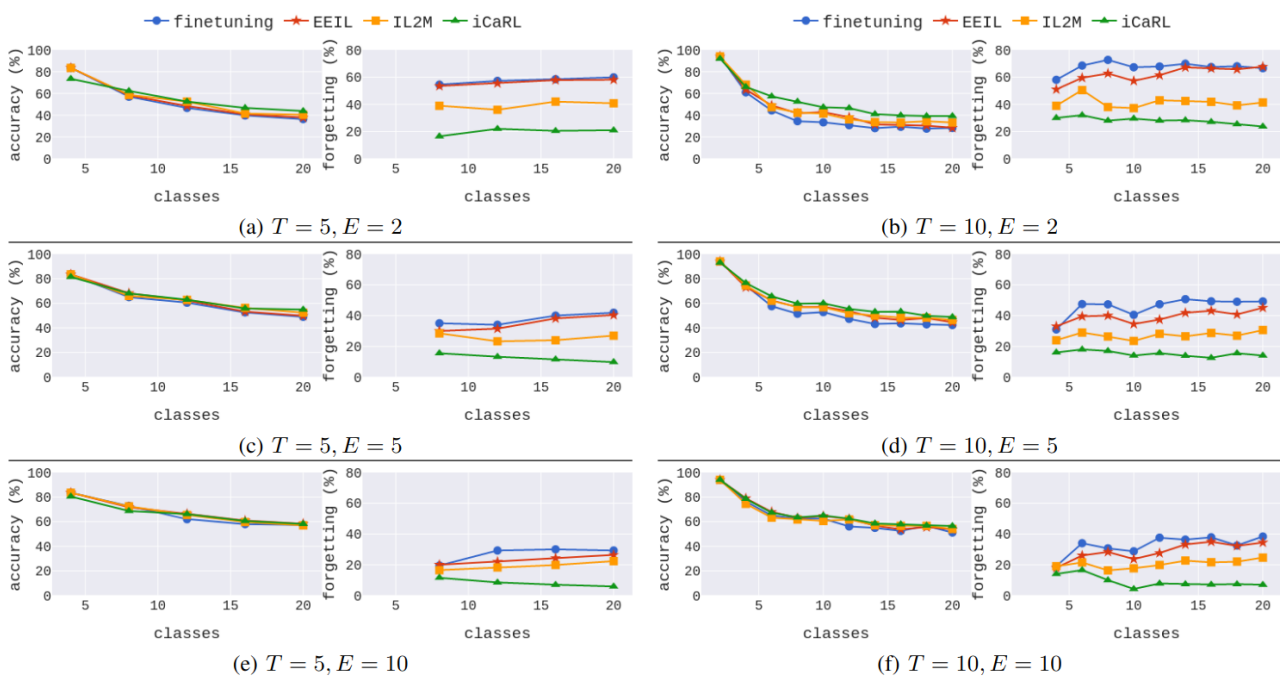


Figure 3. Comparison of the proposed extended iCaRL for videos against alternatives that use experience replay with varying number of exemplars-per-class (E). The left column shows results for a total of $T = 5$ IL phases and the right column for $T = 10$ IL phases. Note that for $T = 5$ the new classes added per phase are $20/5 = 4$ and the accuracy x-axis starts at $x = 4$ while the forgetting x-axis one phase later at $x = 8$. Similarly, for $T = 10$ the number of new classes added per phase are 2, the accuracy x-axis starts at $x = 2$ and the forgetting one at $x = 4$.

We can observe that the extended iCaRL for TSNs exhibits the least percentage of forgetting across all different setups and the higher accuracy in most cases. Interestingly, as we increase the number of exemplars, all IL methods present competitive results on accuracy; however, EEIL and fine tuning suffer significantly from catastrophic forgetting. This implies that iCaRL (and IL2M to an extent) in these cases have an inherent trade-off, trying to balance performance in the newly introduced classes against the old classes. On the other hand, EEIL and fine tuning highly disregard old knowledge and only achieve high performance on new classes. We believe that the iCaRL for TSNs, although it employs random video segments to build the video representation, draws its power from the representation-learning based classifier.

4.3.2. Ablation Study–Evaluation against Regularization Methods

In Figure 4, we also compare the extended iCaRL method for TSNs against regularization-based methods which do not use experience replay: LwF [33], elastic weight consolidation (EWC) [32], and MAS [34]. In the same figure, we also show the result of training every phase with the full dataset (i.e., not performing IL—referred as “Joint”). Note that iCaRL with $E = 5$ and 10 exemplars presents comparative forgetting percentage to “Joint”, which further strengthens our conjecture about the superiority of the representation-learning based classifier. The other methods suffer significantly from catastrophic forgetting, and as the number of classes increases, they exhibit poor performance.

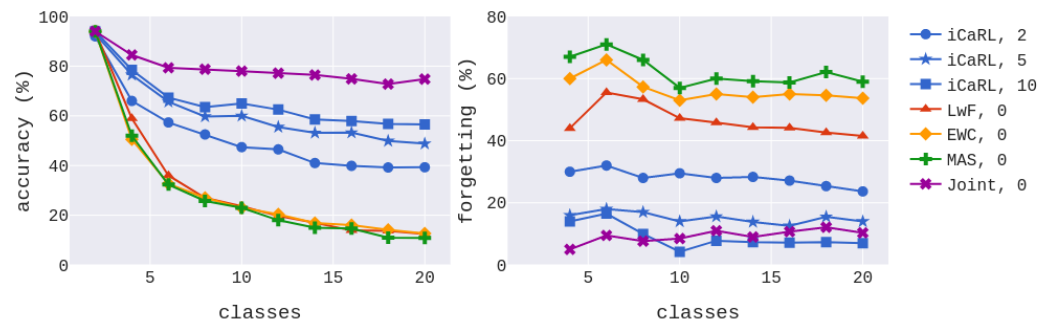


Figure 4. Evaluation of iCaRL for videos against regularization methods ($T = 10$).

4.3.3. Ablation Study–Training Time and Total Accuracy

Finally, Table 5 presents the aggregated results of the considered methods, including the average accuracy and forgetting across all phases, as well as the average time taken to train one phase. We can see that iCaRL achieves the highest average accuracy and least forgetting while having similar computational burden compared to the other methods that use exemplars. On the other hand, while the regularization methods are computationally efficient, they achieve poor performance. Finally, when comparing the “Joint” method (i.e., no IL) against iCaRL with 10 exemplars, we observe a 39% relative reduction in time per phase, but only a 16% relative reduction in accuracy and similar forgetting scores, validating our reasoning for creating an IL setup for action recognition.

Table 5. Average accuracy, forgetting, and time taken for one IL phase for the extended iCaRL for TSNs and other IL methods ($T = 10$).

Method	# Exemplars	Accuracy (%)	Forgetting (%)	Time (s)
EEIL [36]	2	45.06	62.09	443
	5	58.44	39.47	600
	10	65.05	28.73	845
Fine tuning	2	41.16	67.34	289
	5	54.98	45.77	365
	10	63.18	32.78	503
iCaRL [23]	2	52.11	28.00	305
	5	61.55	15.17	387
	10	66.06	9.04	533
IL2M [40]	2	46.37	41.41	292
	5	58.92	27.09	370
	10	64.03	20.59	507
EWC [32]	0	30.72	56.51	252
LwF [33]	0	31.62	46.50	253
MAS [34]	0	29.66	62.22	252
Joint	–	79.06	9.32	873

5. Discussion

This section combines and summarizes our most important findings on all three aspects of the proposed visual perception system and pinpoints factors influencing its performance.

- The choice of the database on which the visual perception models are pretrained greatly affects the recognition accuracy. The system's effectiveness is boosted significantly by employing pretrained models on datasets that are directly related to our desired recognition tasks. More specifically, for action recognition, pretraining on the Kinetics dataset (which includes human actions) results in better action recognition performance compared to pretraining on the ImageNet dataset (for object recognition). Similarly, for the emotion recognition task, pretraining on the AffectNet dataset of facial expressions significantly boosts the system's performance, compared to ImageNet pretraining.
- The number of the sampled segments for the TSN is highly correlated to the recognition task. For action recognition, we note that there is a considerable (yet diminishing) improvement in accuracy by increasing the number of the segments. At the same time, while increasing of the number of segments for the emotion task does translate in an improved performance trend, the results do fluctuate a lot. The mean duration of the videos in the action dataset is 4.23 s and the emotion dataset is 5.06 s. Thus, as both databases have videos with comparable duration, we should look for the cause of this difference in the coded information. Indeed, more sampled segments of an action video imply a more comprehensive understanding of the presented action, since typically an action consists of many different movements. On the other hand, there are cases where emotion is depicted only with a single movement in a short time, for example, smiling for expressing happiness.
- Concerning the information streams, we experimented with a spatial stream that takes as input RGB video frames and a temporal one that takes the optical flow derived from the video as input. The experimental results demonstrate that the primary stream of information for action recognition is the temporal one, with the spatial stream offering a small only performance boost. The opposite is observed in the emotion recognition task, where the spatial stream achieves the best performance, and the temporal one has a small impact.
- Regarding incremental learning, we compare various methods that use experience replay (such as the proposed extended iCaRL for TSNs) and others that impose constraints on the update of the networks. We note that catastrophic forgetting is greater on regularization methods, while memory-replay methods tend to do better. The proposed extended iCaRL for TSNs achieved the best forgetting score across all setups and the higher accuracy in most cases. Furthermore, we also conducted ablation studies on the size of the memory and its impact on accuracy, forgetting, and time to train, proving the efficiency of the proposed system. Methods using dynamic architectures have not been explored yet, as they were considered more computationally demanding, and we aim to explore them in the future.

In the future, regarding the desired application for the proposed CRI system, we aim to evaluate our system in real-world data where both the action and the emotion recognition system will be deployed simultaneously. Due to the ongoing COVID-19 pandemic, we could not conduct experiments with children, which would be more challenging than the separate datasets. Additionally, we aim to integrate the proposed advanced perception system with the robotic edutainment system proposed in [64] (where we partially presented some of the above results). Finally, having feedback on the use by non-technical experts will help highlight and overcome other difficulties that they may face.

6. Conclusions

The proposed lightweight visual perception system for Child–Robot Interaction scenarios employs deep architectures consisting of two perception modules for action and

emotion recognition. According to our extensive experiments and ablation studies, the best trade-off between efficiency and performance for the action recognition module is to use only the Optical Flow modality and five sampled segments yielding about 76% accuracy. For the emotion recognition module, again based on extensive experimentation, we opt to use only the RGB modality in the final system with three sampled temporal segments, which results in 0.79 unbalanced ROC AUC score. Thus, the selection of parameters, such as the information stream and the number of the sampled segments, depends on the recognition task, that is, developing a robust recognition system for children should use both modalities, RGB for emotion recognition and Optical Flow for action recognition. We also note that both perception modules have achieved state-of-the-art results on the EmoReact and BabyRobot action datasets while considering the computational costs. Finally, the proposed extended iCaRL for videos with $E = 5$ exemplars per class achieved an appropriate trade-off between accuracy/forgetting and time to train a new phase. In general, we see that the memory-replay methods perform better than regularization methods on both accuracy and forgetting.

In conclusion, to evaluate the whole system, one has to consider accuracy, computational costs, and inference time. Keeping the time between the child's action and the robot's reaction short is essential in a robotic application that targets children. Otherwise, the child's interest could diminish, and the interaction could stop. Our proposed visual perception system considers all of the above to carry out its purpose and efficiently accommodate CRI scenarios.

Author Contributions: Conceptualization, all authors; methodology, all authors; software, P.P.F. and N.E.; validation, N.E., P.P.F.; investigation, N.E., P.P.F.; writing—original draft preparation, N.E., P.P.F.; writing—review and editing, G.P., P.M.; visualization, N.E., P.P.F.; supervision, G.P., P.M.; project administration, G.P., P.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is carried out/funded in the context of the project “Intelligent Child–Robot Interaction System for designing and implementing edutainment scenarios with emphasis on visual information” (MIS 5049533) under the call for proposals “Researchers’ support with an emphasis on young researchers- 2nd Cycle”. The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014–2020.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to the fact that this research includes experiments with pre-recorded datasets.

Informed Consent Statement: This research includes experiments with pre-recorded datasets and we did not conduct any live experiment involving humans.

Data Availability Statement: In this research, we used two databases to train and evaluate the proposed visual robotic system. The first one is the BabyRobot dataset (<http://babyrobot.eu>, accessed on 1 November 2021) [22] and the second is the EmoReact database (<https://www.behnaznojavan.com/emoreact>, accessed on 1 November 2021) [58].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Scoglio, A.; Reilly, E.; Gorman, J.; Drebing, C. Use of social robots in mental health and well-being research: systematic review. *J. Med. Internet Res.* **2019**, *21*, e13322. [[CrossRef](#)]
2. Góngora Alonso, S.; Hamrioui, S.; de la Torre Díez, I.; Motta Cruz, E.; López-Coronado, M.; Franco, M. Social robots for people with aging and dementia: A systematic review of literature. *Telemed. e-Health* **2019**, *25*, 533–540. [[CrossRef](#)] [[PubMed](#)]
3. Lambert, A.; Norouzi, N.; Bruder, G.; Welch, G. A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 1804–1817. [[CrossRef](#)]
4. Belpaeme, T.; Kennedy, J.; Ramachandran, A.; Scassellati, B.; Tanaka, F. Social robots for education: A review. *Sci. Robot.* **2018**, *3*, eaat5954. [[CrossRef](#)] [[PubMed](#)]

5. Tsiami, A.; Filntisis, P.P.; Efthymiou, N.; Koutras, P.; Potamianos, G.; Maragos, P. Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
6. Efthymiou, N.; Koutras, P.; Filntisis, P.P.; Potamianos, G.; Maragos, P. Multi-View Fusion for Action Recognition in Child–Robot Interaction. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
7. Kennedy, J.; Lemaignan, S.; Montassier, C.; Lavalade, P.; Irfan, B.; Papadopoulos, F.; Senft, E.; Belpaeme, T. Child speech recognition in human-robot interaction: Evaluations and recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction, Vienna, Austria, 6–9 March 2017.
8. Wood, L.; Dautenhahn, K.; Robins, B.; Zaraki, A. Developing child–robot interaction scenarios with a humanoid robot to assist children with autism in developing visual perspective taking skills. In Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal, 28–31 August 2017.
9. Pulido, J. C. and González, J.C.; Suárez-Mejías, C.; Bandera, A.; Bustos, P.; Fernández, F. Evaluating the child–robot interaction of the NAOTherapist platform in pediatric rehabilitation. *Int. J. Soc. Robot.* **2017**, *9*, 343–358. [[CrossRef](#)]
10. Robinson, N.L.; Connolly, J.; Hides, L.; Kavanagh, D.J. A Social Robot to Deliver an 8-Week Intervention for Diabetes Management: Initial Test of Feasibility in a Hospital Clinic. In Proceedings of the International Conference on Social Robotics, Golden, CO, USA, 14–16 November 2020.
11. Boccanfuso, L.; Barney, E.; Foster, C.; Ahn, Y.A.; Chawarska, K.; Scassellati, B.; Shic, F. Emotional robot to examine different play patterns and affective responses of children with and without ASD. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016.
12. Davison, D.P.; Wijnen, F.M.; Charisi, V.; van der Meij, J.; Evers, V.; Reidsma, D. Working with a social robot in school: A long-term real-world unsupervised deployment. In Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction, Cambridge, UK, 23–26 March 2020.
13. Chandra, S.; Dillenbourg, P.; Paiva, A. Children teach handwriting to a social robot with different learning competencies. *Int. J. Soc. Robot.* **2019**, *2*, 721–748. [[CrossRef](#)]
14. Gargot, T.; Asselborn, T.; Zammouri, I.; Brunelle, J.; Johal, W.; Dillenbourg, P.; Archambault, D.; Chetouani, M.; Cohen, D.; Anzalone, S.M. “It Is Not the Robot Who Learns, It Is Me.” Treating Severe Dysgraphia Using Child–Robot Interaction. *Front. Psychiatry* **2021**, *12*, 596055. [[CrossRef](#)]
15. Kennedy, J.; Baxter, P.; Senft, E.; Belpaeme, T. Social robot tutoring for child second language learning. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016.
16. Wolfe, E.; Weinberg, J.; Hupp, S. Deploying a social robot to co-teach social emotional learning in the early childhood classroom. In Proceedings of the 13th Annual ACM/IEEE International Conference on Human–Robot Interaction, Chicago, IL, USA, 5–8 March 2018.
17. Filippini, C.; Spadolini, E.; Cardone, D.; Bianchi, D.; Preziuso, M.; Sciarretta, C.; del Cimmuto, V.; Lisciani, D.; Merla, A. Facilitating the Child–Robot Interaction by Endowing the Robot with the Capability of Understanding the Child Engagement: The Case of Mio Amico Robot. *Int. J. Soc. Robot.* **2020**, *13*, 677–689. [[CrossRef](#)]
18. Senft, E.; Lemaignan, S.; Bartlett, M.; Baxter, P.; Belpaeme, T. Robots in the classroom: Learning to be a Good Tutor. In Proceedings of the 4th Workshop on Robots for Learning (R4L) at HRI2018, Chicago, IL, USA, 5 March 2018.
19. Druin, A.; Hendler, J.A.; Hendler, J. *Robots for Kids: Exploring New Technologies for Learning*; Academic Press: Cambridge, MA, USA, 2000.
20. Hone, K. Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interact. Comp.* **2006**, *18*, 227–245. [[CrossRef](#)]
21. Bickmore, T.W.; Picard, R.W. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.* **2005**, *12*, 293–327. [[CrossRef](#)]
22. Efthymiou, N.; Filntisis, P.P.; Koutras, P.; Tsiami, A.; Hadfield, J.; Potamianos, G.; Maragos, P. ChildBot: Multi-Robot Perception and Interaction with Children. *arXiv* **2020**, arXiv:2008.12818.
23. Rebuffi, S.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental classifier and representation learning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017.
24. Castellano, G.; Leite, I.; Pereira, A.; Martinho, C.; Paiva, A.; Mcowan, P.W. Multimodal affect modeling and recognition for empathic robot companions. *Int. J. Hum. Robot.* **2013**, *10*, 1350010. [[CrossRef](#)]
25. Tielman, M.; Neerinx, M.; Meyer, J.; Looije, R. Adaptive emotional expression in robot-child interaction. In Proceedings of the 2014 9th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Bielefeld, Germany, 3–6 March 2014.
26. Goulart, C.; Valadão, C.; Delisle-Rodriguez, D.; Funayama, D.; Favarato, A.; Baldo, G.; Binotte, V.; Caldeira, E.; Bastos-Filho, T. Visual and Thermal Image Processing for Facial Specific Landmark Detection to Infer Emotions in a Child–Robot Interaction. *Sensors* **2019**, *19*, 2844. [[CrossRef](#)]
27. Lopez-Rincon, A. Emotion recognition using facial expressions in children using the NAO Robot. In Proceedings of the International Conference on Electronics, Communications and Computers, Puebla, Mexico, 27 February–1 March 2019.
28. Filntisis, P.P.; Efthymiou, N.; Koutras, P.; Potamianos, G.; Maragos, P. Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Robot. Autom. Lett.* **2019**, *4*, 4011–4018. [[CrossRef](#)]

29. Marinoiu, E.; Zafir, M.; Olaru, V.; Sminchisescu, C. 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
30. Zhang, Y.; Tian, Y.; Wu, P.; Chen, D. Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder. *Sensors* **2021**, *21*, 411. [[CrossRef](#)]
31. Parisi, G.I.; Kemker, R.; Part, J.L.; Kanan, C.; Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Netw.* **2019**, *113*, 54–71. [[CrossRef](#)] [[PubMed](#)]
32. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [[CrossRef](#)] [[PubMed](#)]
33. Li, Z.; Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2935–2947. [[CrossRef](#)] [[PubMed](#)]
34. Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
35. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsell, R. Progressive neural networks. *arXiv* **2016**, arXiv:1606.04671.
36. Castro, F.M.; Marin-Jiménez, M.J.; Guil, N.; Schmid, C.; Alahari, K. End-to-end incremental learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
37. Shin, H.; Lee, J.; Kim, J.; Kim, J. Continual learning with deep generative replay. *arXiv* **2017**, arXiv:1705.08690.
38. Maracani, A.; Michieli, U.; Toldo, M.; Zanuttigh, P. RECALL: Replay-based Continual Learning in Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021.
39. Van de Ven, G.M.; Siegelmann, H.T.; Tolias, A.S. Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **2020**, *11*, 1–14. [[CrossRef](#)]
40. Belouadah, E.; Popescu, A. I2m: Class incremental learning with dual memory. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
41. Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A.D.; van de Weijer, J. Class-incremental learning: Survey and performance evaluation. *arXiv* **2020**, arXiv:2010.15277.
42. Churamani, N.; Kalkan, S.; Gunes, H. Continual Learning for Affective Robotics: Why, What and How? In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 31 August–4 September 2020.
43. Dehghan, M.; Zhang, Z.; Siam, M.; Jin, J.; Petrich, L.; Jagersand, M. Online Object and Task Learning via Human Robot Interaction. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019.
44. Zhang, H.; Wu, P.; Beck, A.; Zhang, Z.; Gao, X. Adaptive incremental learning of image semantics with application to social robot. *Neurocomputing* **2016**, *173*, 93–101. [[CrossRef](#)]
45. Park, J.Y.; Kim, J.H. Online Incremental Classification Resonance Network and Its Application to Human–Robot Interaction. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1426–1436. [[CrossRef](#)] [[PubMed](#)]
46. Tuyen, N.T.V.; Jeong, S.; Chong, N.Y. Emotional Bodily Expressions for Culturally Competent Robots through Long Term Human–Robot Interaction. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018.
47. Lesort, T.; Lomonaco, V.; Stoian, A.; Maltoni, D.; Filliat, D.; Díaz-Rodríguez, N. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Inf. Fusion* **2020**, *58*, 52–68. [[CrossRef](#)]
48. Barros, P.; Parisi, G.; Wermter, S. A personalized affective memory model for improving emotion recognition. In Proceedings of the Intl. Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
49. Churamani, N.; Gunes, H. CLIFER: Continual Learning with Imagination for Facial Expression Recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020.
50. Costante, G.; Galieni, V.; Yan, Y.; Fravolini, M.L.; Ricci, E.; Valigi, P. Exploiting transfer learning for personalized view invariant gesture recognition. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014.
51. Fourie, C.K.; Lasota, P.A.; Shah, J.A. Motivating Incremental, Personalized Models of Human Behavior for Structured Environments. In Proceedings of the Behavioral Patterns and Interaction Modelling for Personalized Human–Robot Interaction, Cambridge, UK, 23–26 March 2020.
52. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision ECCV, Amsterdam, The Netherlands, 11–14 October 2016.
53. Filntisis, P.P.; Efthymiou, N.; Potamianos, G.; Maragos, P. Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss. In Proceedings of the European Conference on Computer Vision ECCV, Glasgow, UK, 23–28 August 2020.
54. Luo, Y.; Ye, J.; Adams, R.B.; Li, J.; Newman, M.G.; Wang, J.Z. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *Int. J. Comput. Vis.* **2020**, *128*, 1–25. [[CrossRef](#)]

55. Baltrušaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), Xi'an, China, 15–19 May 2018.
56. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; MIT Press: Cambridge, MA, USA, 1974.
57. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Emotion recognition in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 22–25 July 2017.
58. Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C.E.; Morency, L.P. EmoReact: A multimodal approach and dataset for recognizing emotional responses in children. In Proceedings of the 18th Acm International Conference on Multimodal Interaction ICMI, Tokyo, Japan, 12–16 November 2016.
59. Welling, M. Herding dynamical weights to learn. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009.
60. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, Las Vegas, NV, USA, 26 June–1 July 2016.
61. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, Hawaii, 22–25 July 2017.
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, Las Vegas, Nevada, 26 June–1 July 2016.
63. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [[CrossRef](#)]
64. Efthymiou, N.; Filintisis, P.; Potamianos, G.; Maragos, P. A robotic edutainment framework for designing child–robot interaction scenarios. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference PETRA, Corfu, Greece, 29 June–2 July 2021.