Computer Vision and Image Understanding 190 (2020) 102844

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Deep sensorimotor learning for RGB-D object recognition ☆

Spyridon Thermos [a,b,*], Georgios Th. Papadopoulos [a], Petros Daras [a], Gerasimos Potamianos [b]

[a] Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece
[b] Department of Electrical and Computer Engineering, University of Thessaly, 38221 Volos, Greece

ARTICLE INFO

Communicated by: Yasutaka Furukawa

Keywords:
Object recognition
Sensorimotor learning
Object affordance
Convolutional neural networks
Recurrent neural networks
3D convolutions

ABSTRACT

Research findings in cognitive neuroscience establish that humans, early on, develop their understanding of real-world objects by observing others interact with them or by performing active exploration and physical interactions with them. This fact has motivated the so-called "sensorimotor" learning approach, where the object appearance information (sensory) is combined with the object affordances (motor), *i.e.* the types of actions a human can perform with the object. In this work, the aforementioned paradigm is adopted, and a neuro-biologically inspired two-stream model for RGB-D object recognition is investigated. Both streams are realized as state-of-the-art deep neural networks that process and fuse appearance and affordance information in multiple ways. In particular, three model variants are developed to efficiently encode the spatio-temporal nature of the hand–object interaction, while an attention mechanism that relies on the appearance stream confidence is also investigated. Additionally, a suitable auxiliary loss is proposed for model training, utilized to further optimize both information streams. Experiments on the challenging SOR3D dataset, which consists of 14 object types and 13 object affordances, demonstrate the efficacy of the proposed model in RGB-D object recognition. Overall, the best performing developed model achieves 90.70% classification accuracy, which is further increased to 91.98% when trained using the auxiliary loss. The latter corresponds to 46% relative error reduction compared to the appearance-only classifier performance. Finally, a cross-view analysis on the SOR3D dataset provides valuable feedback for the viewpoint impact on the affordance information.

1. Introduction

Object recognition is defined as the ability of humans or machines to perceive 2D and 3D objects based on their visual attributes. Objects constitute key elements, crucial in scene understanding, action identification, and interaction prediction. Thus, the process of recognizing them in the context of an image or video has been an important research topic over the last decades. Robust object recognition remains an open challenge though, mainly due to the sole use of appearance-related information (Liang and Hu, 2015; Hong et al., 2015; Lee et al., 2018; Kanezaki et al., 2018). In fact, using shape, color, and texture cannot fully address the shape variation, deformations, occlusions, and illumination changes that occur in real-world scenarios.

Besides appearance, an object can also be described based on its supported set of "affordances", *i.e.* its functionalities or more specifically the set of actions that humans can perform while interacting with it. Thus, for example, recognizing a sponge based not only on its shape and texture, but also on its "graspable" and "squeezable" affordances, is plausible. The "affordance" term was first defined by Gibson (1977) in a different context, describing what the environment offers or provides to animals living in it. Based on this definition, affordance implies how the environment and animals complement each other. On the other hand, Minsky (1991) focused on a more specific definition of the term, arguing for the significance of classifying objects according to what they can be used for, namely what they can afford. Since then, several approaches have elaborated the affordance theory, targeting object recognition by utilizing their functionalities (Rivlin et al., 1995). The so-called function-based reasoning can be viewed as an approach applicable to environments in which objects are designed and used for specific purposes. Sutton et al. (1998) presented three possible ways of extracting the functional information of an object: (a) "function from shape", where the shape of the object provides some indication of it, (b) "function from motion", where an observer attempts to understand it by perceiving a task being performed with the object, and (c) "function from manipulation", where such is extracted by manipulating the object.

There is accumulated evidence that humans, at an early stage of their lives, perceive objects by combining their visual attributes with the feedback from interacting with them. This process is known as

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.102844.

* Corresponding author at: Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece.
E-mail address: spthermo@iti.gr (S. Thermos).

https://doi.org/10.1016/j.cviu.2019.102844
Received 23 November 2018; Received in revised form 24 May 2019; Accepted 2 October 2019
Available online 22 October 2019

"sensorimotor learning" (Piaget and Brown, 1985; Flavell, 1992; DiCarlo et al., 2012), due to the parallel processing of the "sensory" and "motor" information in the human brain. Indeed, it is well established by cognitive scientists that there are two main streams that process the aforementioned information (Ungerleider and Haxby, 1994; van Polanen and Davare, 2015): the ventral stream that runs in the inferotemporal cortex is involved in the recognition of objects, while the dorsal one that projects to the posterior parietal cortex is involved in the understanding of 3D space and action planning. Research findings indicate that the two streams process information both independently and in parallel, utilizing feedback loops and sharing information through neural connections that exist in multiple stages (Cloutman, 2013; Brandi et al., 2014). These identified interconnections enable the human brain to fuse sensory and motor information, so as to achieve robust cognition.

In this paper, motivated by the above facts, we investigate sensorimotor learning for RGB-D object recognition in the context of "function from motion", as defined by Sutton et al. (1998). Further, inspired by the complex neural network of the human brain, we adopt the Deep Learning (DL) paradigm (LeCun et al., 2015) to form two parallel information streams that process object appearance (sensory) and affordance (motor) information. These streams exploit DL architectures, primarily convolutional and recurrent neural networks, and are fused in multiple ways, in order to mimic the complex information exchange between the brain processing pathways. A schematic of our approach is depicted in Fig. 1.

The main contribution of this paper is therefore the introduction of DL-based sensorimotor learning in RGB-D object recognition. Specifically, three variants of the proposed two-stream sensorimotor modeling approach are considered that utilize different deep neural networks to encode the spatial-only or spatio-temporal correlations of suitable appearance and affordance input representations.

Additionally, inspired by the aforementioned neuro-scientific findings for the human brain complex information exchange at different levels of granularity, fusion at one or multiple layers of each model variant is extensively investigated. Regarding spatio-temporal information processing, the incorporation of an attention mechanism is also proposed, which forces the model to selectively attend to the affordance information, when the appearance one is not discriminative enough, as indicated by appropriate stream confidence measures.

Further, an auxiliary loss function is introduced, based solely on affordance predictions. The new loss is combined with the object prediction one, and the result is used to optimize both streams during training. In order to compute the auxiliary loss, a classifier is added after the last affordance stream layer, but later removed during inference.

Finally, an extensive quantitative evaluation of the proposed models is presented, using the challenging SOR3D corpus (Thermos et al., 2017)[1] that includes a significantly increased number of affordances compared to existing works in the literature (see Section 2.2 and Table 1). Besides comparison of the two-stream models with the appearance-only baseline, the best performing one is further benchmarked against traditional probabilistic fusion approaches. The evaluation is concluded with a cross-view analysis, providing valuable insights about how view-dependent is the affordance information and how each viewpoint affects model performance.

The remainder of the paper is organized as follows: Section 2 overviews related work on affordance-based recognition and sensorimotor learning; Section 3 presents the visual front-end and the single-stream models; Section 4 details the proposed two-stream sensorimotor modeling framework and the investigated fusion schemes; Section 5 presents the experimental results; and finally, Section 6 summarizes the paper.

[1] Publicly available at: http://sor3d.vcl.iti.gr/.

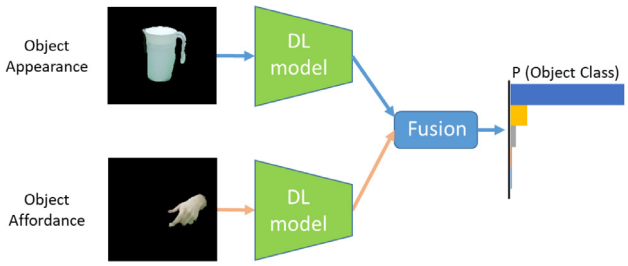


Fig. 1. Schematic of the deep learning (DL) based architecture of the proposed sensorimotor 3D object recognition framework. Following fusion of the object appearance and affordance processing streams, the object class is predicted.

2. Related work

Object recognition is a fundamental problem in computer vision. Focusing solely on object appearance attributes, relevant approaches can be divided into two main categories: methods that represent the object with hand-crafted features and ones that learn deep object representations exploiting the DL paradigm. Characteristic works of the first category are reported in the survey of Andreopoulos and Tsotsos (2013). Regarding DL-based methods, numerous works appear in the literature. For example, among others, Liang and Hu (2015) propose a recurrent Convolutional Neural Network (CNN) framework to classify objects based on their appearance and the context of the scene; Su et al. (2015) present a multi-view CNN with a view-pooling layer to categorize 3D objects; Qi et al. (2016) propose a volumetric CNN for object point-cloud processing and classification; Yu et al. (2018) utilize polynomial kernels and bilinear pooling in a CNN to aggregate local convolutional features in a 3D object representation; and Feng et al. (2018) propose a group-view CNN that models the hierarchical correlations among multiple 2D views of a 3D object, leading to a powerful 3D descriptor.

Besides appearance-based learning, there are extensive studies related to functional object recognition exploiting object affordances. In particular, affordance-oriented object recognition is investigated in the literature from two viewpoints: (a) embodiment and (b) observation. The former indicates the scenario where there is direct interaction of the perceiver with the object, while the latter denotes the scenario where the perceiver observes others interacting with the object.

2.1. Inferring object affordances from embodiment

Regarding the embodiment scenario, object affordances that are inferred from agent–object interaction have been recently leveraged in object recognition. In particular, Saxena et al. (2008) concentrate on robotic grasping of novel objects using a set of 2D object views labeled with grasping points. Additionally, Högman et al. (2016) propose a Gaussian process to model object-related sensorimotor "contigencies" (O'Regan and Noë, 2001) and categorize objects by "pushing" them and observing their displacement. Lyubova et al. (2016) employ the iCub and Meka robots to categorize objects by combining visual and proprioceptive knowledge with motion behavior observed during interaction. Focusing on more composite actions, Fitzpatrick et al. (2003) utilize robotic "push", "pull", and "poke" actions to further explore object representations, while Montesano et al. (2008) present a scenario where a robot with basic motor skills categorizes objects by observing human–object interactions and subsequently selecting its own action that will have the same effects on the object. Additionally, Jayaraman and Grauman (2018) propose a system for active visual recognition through agent–object interaction, where given an initial view of the object, the system predicts how the choice of motion alters the environment, and integrates the result of the object manipulation at

Table 1
Datasets that consist of "tool-objects" (indoor scenes) and have affordance information available.

Dataset	Interaction	Format	Objects	Affordances	Subjects	Samples	Public availability
Kjellström et al. (2011)	yes	RGB	6	3	4	28	no
Castellini et al. (2011)	yes	RGB	7	5	20	130	no
Kluth et al. (2014)	no	RGB	8	1	n/a	n/a	no
TTU (Zhu et al., 2015)	yes	RGB-D	10	3	1	452	no
ADE-Affordance (Song et al., 2016)	no	RGB	8	4	n/a	10,360	yes
IIT (Nguyen et al., 2017)	no	RGB-D	10	9	n/a	8,835	yes
COQE (Mottaghi et al., 2017)	no	RGB	10	1	n/a	5000	yes
SOR3D (Thermos et al., 2017)	yes	RGB-D	14	13	105	20,800	yes

Fig. 2. Examples of human–object interactions in the SOR3D dataset (Thermos et al., 2017), captured by three Kinect sensors.

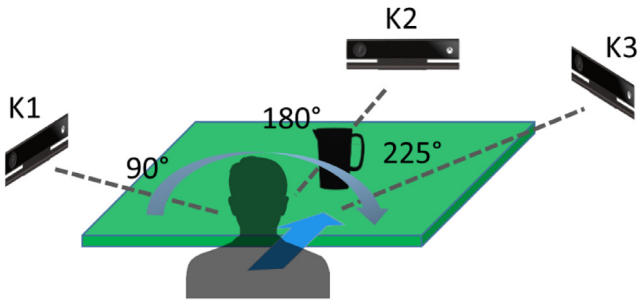


Fig. 3. Schematic representation of the SOR3D capturing setup. The three Kinect sensors (K1–K3) are placed from left to right at 90°, 180°, and 225° with respect to the subject orientation.

each time-step to classify the object. The system is trained end-to-end using reinforcement learning.

Besides recognition, object affordances provide valuable feedback for numerous tasks in the field of cognitive vision and developmental robotics, such as gaze control, semantic grasping, and action prediction (Ghadirzadeh et al., 2016; Giagkos et al., 2017; Jang et al., 2017; Oberlin and Tellex, 2018; Zambelli and Demiris, 2017). However, further elaboration on this aspect of sensorimotor learning lies outside the scope of this paper.

2.2. Observation-based sensorimotor learning

Learning to recognize objects by observing others interacting with them is a challenging machine perception task. However, recent works on observation-based sensorimotor object recognition mostly rely on simple fusion schemes (*e.g.* using simple Bayesian models or the product rule), hard assumptions (*e.g.* naive Gaussian prior distributions), and simplified experimental settings (*e.g.* few object types and simple affordances).

For example, Kjellström et al. (2011) utilize histograms of oriented gradients to model object appearance, while the global velocity, orientation, and joint angles of the hand are used to encode the affordance information. A binary SVM is trained for each stream, while the predicted object-hand pairs of 3 consecutive frames are utilized by factorial conditional random fields for the final object class prediction. This method is evaluated using a dataset of 6 objects and 3 affordances. Kluth et al. (2014) propose a framework where GIST-features of object appearance and affordance are used to form sensorimotor representations. Then, probabilistic reasoning comprised of a Bayesian network with information gain strategy is used for object classification, exploiting these representations. The method is evaluated on a dataset that consists of 8 object classes and a single affordance. Additionally, Castellini et al. (2011) encode the object appearance as frequency histograms of 200 bins, while 22 motor features provided by a motion-capture glove sensor are used as affordance representation. The appearance and affordance features are fused using positively weighted linear combination of Mercer kernels and are used to train a one-versus-all SVM for object classification. The algorithm is evaluated on a dataset of 7 objects and 5 affordances. Zhu et al. (2015) propose a framework aiming at understanding the affordance and the functional basis (*e.g.* the part of the hammer that touches a surface when hammering) of tool objects through observing a human, using them for task-oriented object recognition. They model the object appearance, action sequence, and physical quantities produced by the interaction using graphs and train a ranking-SVM classifier to recognize the objects. The framework is evaluated on a dataset consisting of 10 objects and 3 affordances.

Moving beyond experimental frameworks that rely on hand-crafted features, simple affordances, and hard assumptions, early work by Thermos et al. (2017) introduces the DL paradigm to sensorimotor object recognition and presents a large-scale dataset of 14 object types and 13 object affordances. There, two models are proposed: The first utilizes two CNNs that encode spatial-only information, while the second is based on a combination of CNNs with a recurrent neural network (RNN) for spatio-temporal information encoding. The latter is further optimized in Thermos et al. (2018) with the incorporation of an attention mechanism, which relies on the appearance stream confidence. Here, we extend these works by introducing new affordance input representations, a novel model based on 3D convolutions, and an auxiliary loss for improved model training. Further, various fusion strategies, originally introduced in Thermos et al. (2017), are investigated in conjunction with the newly introduced 3D CNN model.

The proposed sensorimotor models are evaluated on the SOR3D dataset (Thermos et al., 2017), which is a publicly available corpus that enables development and evaluation of sensorimotor object recognition methods. Table 1 reports 8 datasets that consist of tool-objects captured indoors and include affordance information. From these datasets, ADE-Affordance (Song et al., 2016), IIT (Nguyen et al., 2017), COQE (Mottaghi et al., 2017), and the one from Kluth et al. (2014) include only static objects with no interaction, while the affordance information is represented as pixel-wise annotation of the object part that enables a specific affordance (*e.g.* the handle of a cup is annotated as "graspable") followed by the corresponding bounding box. These datasets are mostly used for affordance-part detection and segmentation tasks. On the other hand, the datasets that include

Table 2
Supported object and affordance types in the SOR3D corpus (Table from (Thermos et al., 2017)). Considered object-affordance combinations are marked with √.

Object types	Affordances												
	Grasp	Lift	Push	Rotate	Open	Hammer	Cut	Pour	Squeeze	Unlock	Paint	Write	Type
Ball	√	√	√										
Book	√	√	√		√	√							
Bottle	√	√	√					√					
Box	√	√	√	√	√								
Brush	√	√									√		
Can	√	√	√										
Cup	√	√	√	√									
Hammer	√	√				√							
Key	√	√					√			√			
Knife	√	√					√						
Pen	√	√										√	
Pitcher	√	√	√	√				√					
Smartphone	√	√	√										√
Sponge	√	√	√	√					√				

hand–object interaction sessions, apart from SOR3D, consist of small numbers of samples and are not publicly available (note that the TTU dataset (Zhu et al., 2015) provides only the colored point clouds of the objects and not the hand–object sequences). Part of the information in Table 1 is also reported in the recent survey on visual affordances by Hassanin et al. (2018).

3. Visual front-end and single-stream models of appearance and affordance

In this section, the preprocessing framework, as well as the appearance and affordance input representations are detailed. Additionally, three single-stream models capable of encoding either spatial-only or spatio-temporal information are presented.

3.1. Classes, input streams, and preprocessing

As discussed in the previous section, the SOR3D dataset is used in this paper for sensorimotor object recognition. Briefly, it contains 6,943 sessions (samples are depicted in Fig. 2) of 14 object types and 13 affordance types, combined into 54 possible hand–object interactions (see Table 2), since not all object-affordance combinations are feasible. For each session, RGB and depthmap streams are provided, captured by three Kinect sensors (see also Figs. 2 and 3, as well as Section 5.4), thus yielding 20.8k interaction videos that contain the captured scene (*e.g.* the subject, object, desk interaction area, and surrounding background).

Based on the Kinect intrinsic parameters, the RGB frames are mapped to the corresponding depthmaps, the region that includes the hand–object interaction is defined, and a centered rectangular region (300 × 300 pixels) is cropped. Subsequently, using a simple thresholding method in the HSV color space (Vezhnevets et al., 2003), the background is removed, and the skin color pixels (*i.e.* those corresponding to the hand region) are separated from the object ones. As a result, the hand and object RGB and depthmap frames are provided separately, as depicted in Fig. 4. Note that the hand–object separation leads to the two types of information utilized in this work, namely the object appearance that is related to the object shape, color, and texture, and the object affordance that is related to the hand movement. This process aims to remove information that is not relevant to the interaction (*e.g.* background, tablecloth, etc.), in order to investigate the true added value of the affordance information. Note also that as this database is collected in a controlled lab environment (*i.e.* illumination, green tablecloth, no long sleeves), traditional segmentation approaches are very accurate, thus a more sophisticated semantic segmentation algorithm would offer no substantial gains.

Due to the low object intra-class variance, we choose to ignore the RGB information and instead encode the depth information using two different approaches. In the first one, we adopt the depth encoding

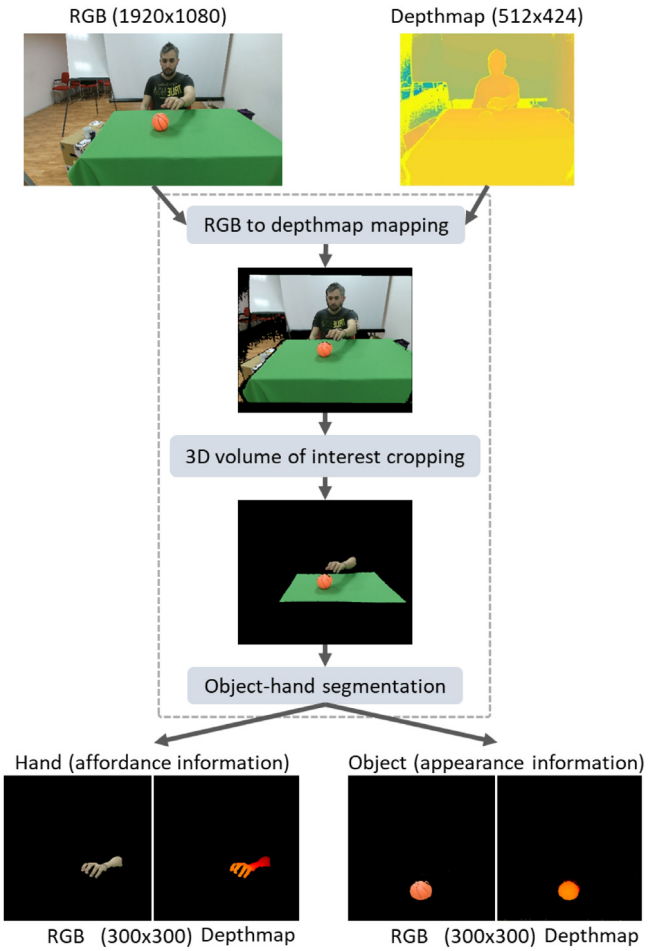


Fig. 4. Preprocessing overview. The captured RGB and depth raw data (top) are initially aligned, the 3D volume of interest is cropped (middle), and the hand and object RGB and depth representations are separated (bottom). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

algorithm introduced by Gupta et al. (2014), which relies on computing three depth-based features, namely the Horizontal disparity, the Height above the ground, and the Angle between the surface normals and the gravity direction of the captured scene (HHA). The three computed features are stacked to form a 3-channel representation that has the same width and height as the original depthmap. The second depth encoding approach is depthmap "colorization". Motivated by Eitel et al. (2015),

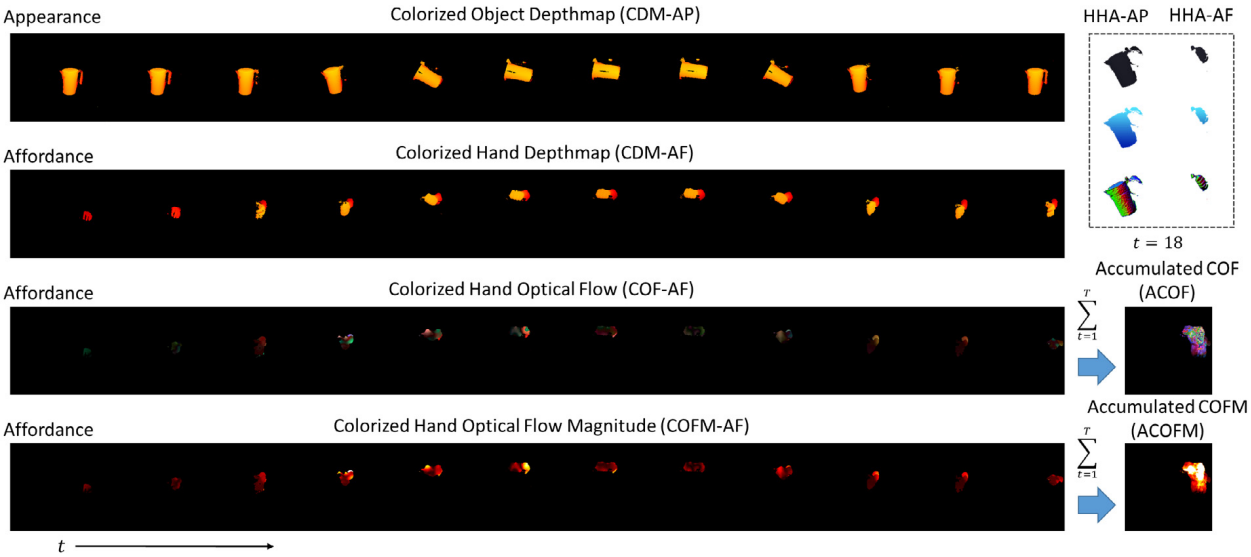


Fig. 5. Example video session "pour from pitcher" from the SOR3D corpus, sampled every 4 frames. The object appearance is depicted as colorized depthmaps and HHA encoding (only for an example frame, top-down: disparity, height, normals channels are shown), while the affordance information is depicted as colorized depthmaps, HHA encoding (same example frame and top-down presentation as HHA-AP), 3D optical flow, and 3D optical flow magnitude, as well as the accumulation of the latter two over the sequence of T frames. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

depth colorization is performed by normalizing all depth values in the interval $[0, 255]$ and then mapping each pixel distance to color values ranging from red (near) to yellow (far), transforming the one-channel depthmap to a three-channel color image. Note that the aforementioned approaches also enable the exploitation of transfer learning by using DL models pre-trained on large-scale image datasets (Pan and Yang, 2010; Tommasi et al., 2010; Yosinski et al., 2014).

Besides depth encoding, we further process the original depthmaps and RGB frames in order to compute the 3D optical flow of the interaction (relating to object affordance). Due to the development of affordable RGB-D sensors, several 3D flow computation methods have been proposed in the literature (Hadfield and Bowden, 2014; Hornácek et al., 2014; Quiroga et al., 2014). In this work we utilize the primal–dual algorithm proposed in Jaimez et al. (2015) due to its efficiency. In detail, the 3D motion vectors between two pairs of RGB-D images, as well as their magnitude are first computed. The 3D flow and its magnitude are then colorized by normalizing each axis values in the interval $[0, 255]$, transforming the 3D motion vectors into a three-channel image. We further choose to encode the 3D flow sequence into a single motion map, by accumulating the flow over the entire sequence, as such representations can be very informative (Wang et al., 2017).

To summarize, the information streams that are utilized as input to the proposed models are: (a) HHA encoding, (b) colorized depthmaps (CDM), (c) colorized 3D optical flow (COF) along with the accumulated colorized 3D optical flow (ACOF), and (d) colorized 3D optical

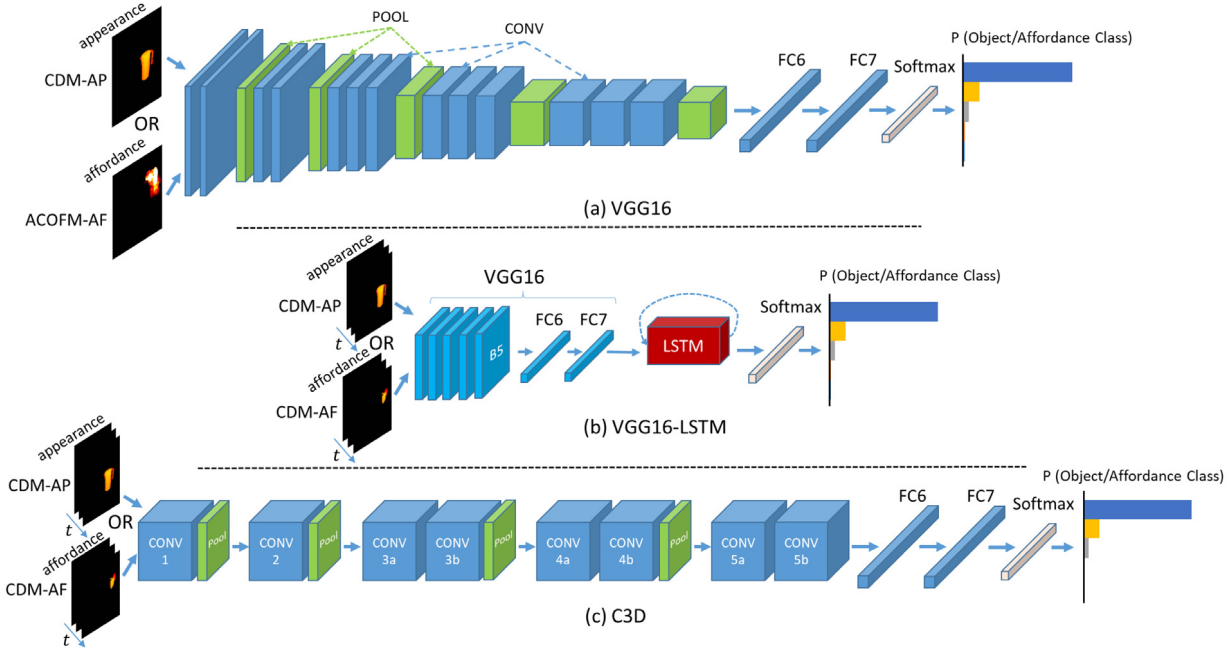


Fig. 6. Detailed architecture of the adopted single-stream models: (a) VGG16 that is capable of encoding spatial information; (b) VGG16-LSTM that utilizes a VGG16 and an LSTM to encode spatio-temporal information; and c) C3D that exploits 3D convolutions to encode spatio-temporal information. The CDM-AP or HHA-AP can be used as input appearance representations, while various affordance input representations of Fig. 5 can be used, as evaluated in Table 3 of Section 5.1.

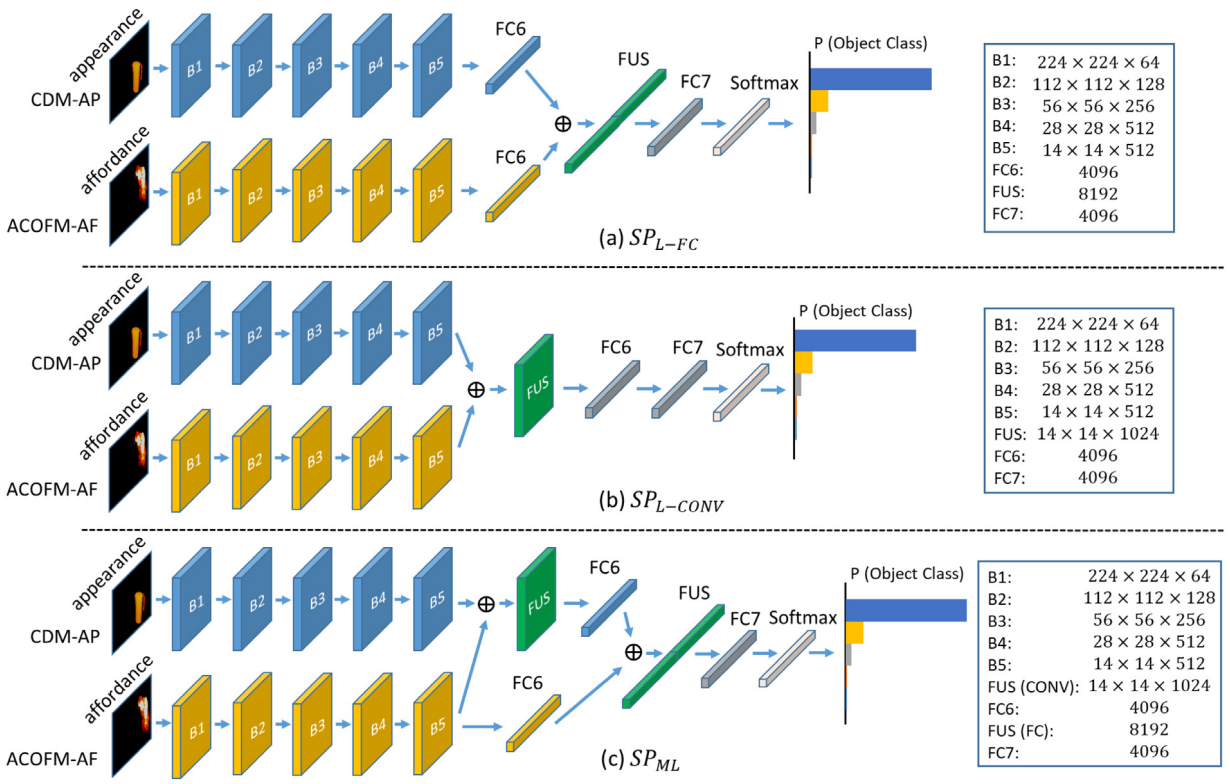


Fig. 7. Detailed architecture of the SP model for: (a) late fusion at the FC layer; (b) late fusion at the CONV layer; and (c) multi-layer fusion. Each block (B1–B5) corresponds to a CONV-RL-POOL sequence of VGG16, while FUS indicates feature fusion (concatenation). At the right side of each fusion scheme, the dimensionality of the activation matrix for each CONV block is reported as "$height \times width \times channels$" and for each FC layer as the number of neurons. The CDM-AP or HHA-AP representations can be used as input to the appearance stream, whereas the ACOF-AF and ACOFM-AF as input to the affordance stream.

flow magnitude (COFM), coupled with the corresponding accumulated one (ACOFM). Fig. 5 depicts an example of two appearance and six affordance input representations of a "pour from pitcher" session. For the sake of clarity, the information stream that processes the object appearance is denoted as the "appearance stream", while the one that processes the hand–object interaction is denoted as the "affordance stream". Additionally, the notation {AP, AF} is used to state that a specific input is received from the appearance or the affordance stream (*e.g.* CDM-AP denotes that the appearance stream receives colorized depthmaps as input).

3.2. Single-stream models

For the appearance and affordance information processing, three single-stream models are proposed, as detailed next.

The first model, depicted in Fig. 6(a), is the VGG16 (Simonyan and Zisserman, 2015) network, which encodes the spatial-only information of an input image. It consists of 5 blocks (B1–B5) of convolutional (CONV) layers and 2 Fully Connected (FC) ones, while each CONV group is followed by a pooling (POOL) layer. A Rectified Linear unit (RL) is used as activation function after each CONV and FC layer. Note that VGG16 can efficiently learn complex spatial feature representations and has been widely used for visual recognition purposes.

The second model is capable of encoding both spatial and temporal information, by processing sequences of 2D frames. As shown in Fig. 6(b), the model consists of a VGG16 network followed by a Long–Short Term Memory (LSTM) one (Hochreiter and Schmidhuber, 1997). (Finally, the C3D network Tran et al., 2015), which is also capable of encoding spatio-temporal information, is used as the third model. The C3D, depicted in Fig. 6(c), consists of 8 3D convolutional (3DCONV), 5 POOL, and 2 FC layers. Note that this model processes groups of video frames, stacked along the RGB-channel axis to form 3D representations.

The aforementioned models are separately trained for object and affordance recognition, using the 14 object and 13 affordance classes as ground truth, respectively. Note that all models use a Softmax layer for class prediction. Additionally, HHA encoding and CDM are used as input representations of object appearance, whereas all six affordance input representations reported in Section 3.1 are utilized to investigate their impact on RGB-D object recognition.

Regarding the VGG16 model, which predicts classes for individual images, the video-level object prediction is obtained by averaging all frame-level predictions. However, this process is not effective for affordance recognition. Intuitively, object affordances are explicitly described by hand motion, which is time-evolving. Thus, using the VGG16 model to predict the affordance class of a sequence would be inconsistent. This intuition is confirmed in Section 5.1 (see also Table 3), hence for most of the paper we utilize only the ACOF and ACOFM representations as input to the affordance VGG16, as they summarize the entire motion of the sequence accumulated within a single frame.

4. Two-stream models fusing appearance and affordance

Motivated by the two-stream hypothesis of the human brain sensorimotor learning process, the aforementioned single-stream models are fused in multiple ways in order to achieve robust object recognition. Three sensorimotor models are presented, where the appearance and affordance information exchange between the two streams is extensively investigated.

4.1. Two-stream spatial-only model (SP)

The two-stream spatial-only model (SP), depicted in Fig. 7, utilizes two VGG16 networks, one for appearance and the other for affordance information processing. The appearance VGG16 receives HHA-AP or

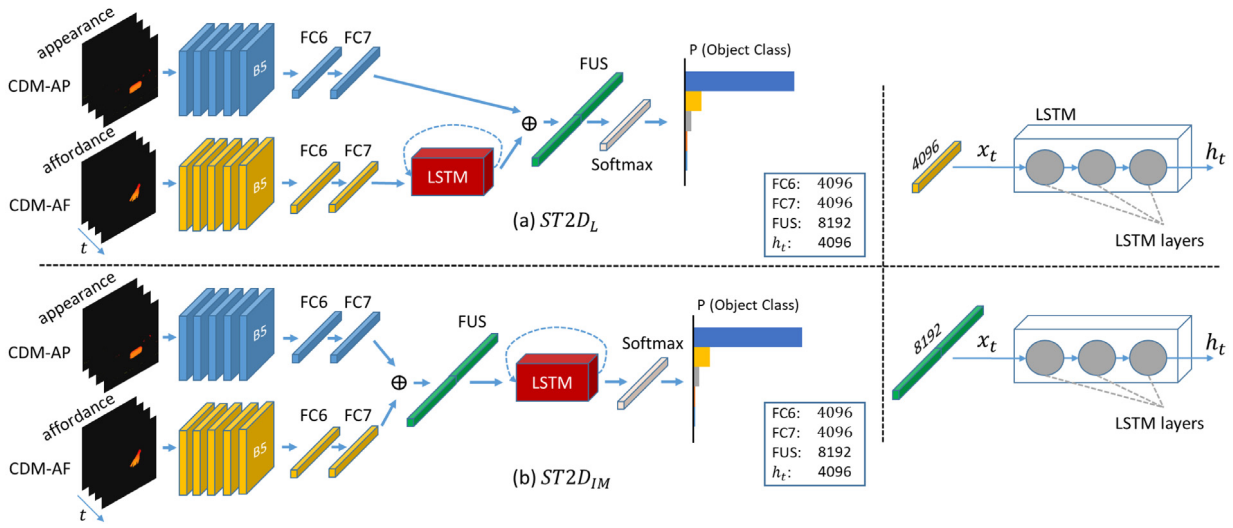


Fig. 8. Detailed architecture of the ST2D model for: (a) late fusion and (b) intermediate fusion. Blocks B1–B5, FC6, and FC7 correspond to the VGG16 network, while FUS indicates feature fusion. At the right side, the internal structure of the LSTM, as well as the input x and output h vector dimensionality for each fusion scheme are depicted. The appearance stream processes HHA-AP or CDM-AP inputs, while the HHA-AF, CDM-AF, COF-AF, and COFM-AF representations can be used as input to the affordance stream.

CDM-AP input, while the affordance one processes either ACOF-AF or ACOFM-AF representations, similarly to the single-stream affordance VGG16 described in Section 3.2. Three fusion schemes of the two streams are investigated: (a) late fusion at the FC layer level (SP_{L-FC}); (b) late fusion at the CONV layer level (SP_{L-CONV}); and (c) multi-layer fusion (SP_{ML}) that combines the aforementioned approaches.

Late fusion at the FC layer level (Fig. 7(a)) is realized by concatenating the activations of FC6 (*i.e.* the sixth VGG16 layer, which is a FC one) of each stream, after the RL non-linearity. After fusion, a single stream of a 4096-dimensional (dim) FC layer and a Softmax layer is formed.

Regarding late fusion at the CONV layer level (Fig. 7(b)), the activation maps after RL5 (non-linearity of CONV5) are concatenated along the channel dimension. In more detail, if $X_s^{h\times w\times d}$ represents each activation matrix, where $s \in \{\text{AP}, \text{AF}, \text{FUS}\}$ and h, w, d correspond to the height, width, and number of channels, then $X_{AP}^{14\times14\times512}$, $X_{AF}^{14\times14\times512}$ are the inputs and $X_{FUS}^{14\times14\times1024}$ is the output of the fusion. The latter is further convolved with 512 filters of 1×1 size and downsampled using a max-pooling layer (2×2 size), thus resulting in a $X_{FUS}^{7\times7\times512}$ activation matrix. Similarly to the FC layer late fusion, a single processing stream is formed that consists of 2 FC layers (4096-dim) and a Softmax layer.

Finally, in order to allow more complex information exchange at different levels of granularity between the two streams, a multi-layer fusion scheme is also investigated (Fig. 7(c)). In particular, the two streams are initially fused after the last CONV layer (RL5) and then fused again after FC6 (RL6). The appearance FC6 layer receives as input the fused activations, while the affordance FC6 receives the activations from POOL5 (POOL layer after CONV5) of the affordance stream only. Subsequently, the activations after RL6 of both streams are concatenated forming a 8192-dim feature, followed by a 4096-dim FC layer and a Softmax layer. Note that, in the multi-layer fusion case only, the weights of the affordance B1–B5 layers are not updated from the gradients computed at the CONV fusion level during back-propagation. In that way, the affordance stream contributes to the appearance one in multiple levels, without being particularly affected by the appearance information.

For the video-level object class prediction, the object probabilities for each frame of the sequence are averaged. Note that the affordance input representation remains unaltered, as it includes the aggregated information of the entire sequence.

4.2. Two-stream spatio-temporal 2D model (ST2D)

Another approach for modeling the dynamic nature of the affordance information is realized with the proposed two-stream Spatio-Temporal 2D model (ST2D). As shown in Fig. 8, we adopt the VGG16-LSTM structure to model the spatio-temporal nature of the hand–object interaction. Two fusion approaches are considered, namely intermediate ($ST2D_{IM}$) and late fusion ($ST2D_L$).

Regarding the $ST2D_{IM}$ model, the 4096-dim spatial feature vectors extracted by each VGG16 model (*i.e.* the activations after the RL7 layer) are concatenated and then processed by the LSTM at every time instant, namely at every frame of the input sequence. The LSTM encodes the temporal correlations of the interaction, while its internal state vector $[h(t)]$ (4096-dim) is further processed by a Softmax layer for the object class prediction.

On the other hand, $ST2D_L$ adopts the VGG16-LSTM structure only for the affordance information processing, since the appearance VGG16-LSTM performs significantly worse than VGG16 as a single-stream object classifier (see Section 5.1 and Table 3). Thus, for this model, the RL7 activations of the appearance stream (4096-dim) are concatenated with the internal state vector $[h(t)]$ (4096-dim) of the affordance VGG16-LSTM at every time instant. The outcome of the concatenation is further processed by a Softmax layer.

Regarding the final prediction, two approaches are investigated for both fusion schemes. These approaches aggregate the frame-level prediction to yield a video-level classification decision. Given a series of frame-level posteriors $p_{t,c}$, where $t = 1,\dots,T$ is the frame number and $c = 1,\dots,C$ the object class, the video-level classification decision $\hat{c}$ is given either by:

$$\hat{c}_{avg} = \arg\max_c \frac{1}{T}\sum_{t=1}^{T} p_{t,c} \quad , \tag{1}$$

employing the averaging (AVG) approach, or by:

$$\hat{c}_{w} = \arg\max_c \frac{1}{T}\sum_{t=1}^{T} t\, p_{t,c} \quad , \tag{2}$$

using the weighting (W) approach, respectively. Clearly, (1) indicates that all frame-level predictions contribute equally to the video-level one. However, the LSTM weights are updated after every processed frame, thus the affordance features prior to fusion should be more discriminative at the end of the sequence. Thus, we utilize (2) to force the model to focus more on the frame-level predictions over the last video frames.

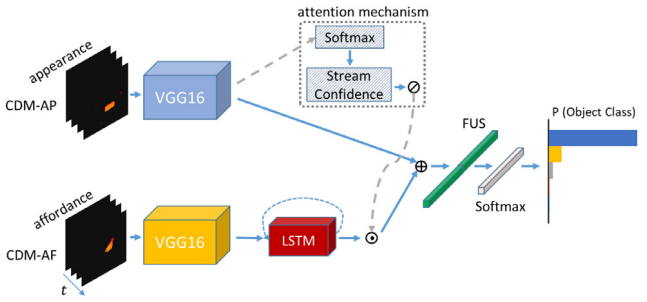


Fig. 9. Detailed architecture of the ST2D$_L$ model with an attention mechanism incorporated. The attention mechanism follows the final FC layer of the appearance VGG16 (top), which selectively attends to the affordance stream output (bottom). Symbols ⊘ and ⊙ represent normalization and frame-level multiplication, respectively.

4.3. Two-stream ST2D model with attention

Since our experiments in Section 5.2.2 indicate that neither of the ST2D fusion approaches exhibit satisfactory results in object recognition (see Table 5), we argue that the affordance information does not equally contribute to each frame-level prediction. For example, as shown in Fig. 5(top), the object can be easily identified at the first and the last frames of the video based only on its appearance. On the contrary, at the middle of the sequence, the object cannot be confidently identified due to the "handle" occlusion from the hand. Motivated by this observation, we utilize the attention mechanism proposed in Thermos et al. (2018), in order to incorporate the affordance information when it is truly needed, *i.e.* when the appearance features are not discriminative enough. Additionally, since fusing the two information sources prior to the LSTM leads to significantly inferior performance compared to the appearance VGG16 (see Tables 3 and 5), we choose to apply the attention only to the ST2D$_L$ model.

As depicted in Fig. 9, the attention mechanism consists of a Softmax classifier added after the last FC layer of the appearance CNN and a module that measures the appearance-based classifier confidence for each frame. The latter is used to selectively attend to the affordance features extracted from the LSTM, prior to fusion.

For this purpose, three confidence metrics are investigated. Denoting by $c_{t,n}, n = 1, \ldots, N$ the ranked $N-$best object class predictions of the appearance CNN classifier and by $p_{t,n}$ the corresponding posteriors at frame t, the first metric is the entropy $\mathcal{I}_{t,E}$, computed as:

$$\mathcal{I}_{t,E} = -\sum_{n=1}^{C} p_{t,n} \log\left(p_{t,n}\right) \ . \tag{3}$$

Clearly, $\mathcal{I}_{t,E}$ values that are close to zero indicate strong confidence, while larger values indicate difficulty in discrimination. The second investigated metric is the average $N-$best log-likelihood difference, computed as:

$$\mathcal{I}_{t,A} = \frac{1}{N-1} \sum_{n=2}^{N} (\log\left(p_{t,1}\right) - \log\left(p_{t,n}\right)) \ , \tag{4}$$

where $N \geq 2$. In contrast to entropy, larger values of $\mathcal{I}_{t,A}$ indicate high-confidence predictions. Finally, the last metric measures the log-likelihood dispersion among the $N-$best class predictions, and is given by:

$$\mathcal{I}_{t,D} = \frac{2}{N(N-1)} \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} (\log\left(p_{t,n}\right) - \log\left(p_{t,m}\right)) \ , \tag{5}$$

where $N \geq 2$. Similarly to (4), larger $\mathcal{I}_{t,D}$ values indicate high classification confidence. The aforementioned metrics have also been used in the context of audio–visual speech recognition (Potamianos and Neti, 2000; Adjoudani and Benoît, 1996).

Following the appearance classifier confidence measurement, the $\mathcal{I}_t$ values of all frames are normalized to $[0, 1]$ by:

$$w_t = \frac{\mathcal{I}_t - \mathcal{I}_{min}}{\mathcal{I}_{max} - \mathcal{I}_{min}} \ , \tag{6}$$

where $\mathcal{I}_{min}$, $\mathcal{I}_{max}$ are calculated over the entire video, and $w \in [0, 1]$ is the resulting confidence vector over all T video sequence frames. The last step of the attention mechanism is given by:

$$\hat{H} = \begin{cases} w \odot H & \text{if (3)} \\ (1 - w) \odot H & \text{if (4) or (5)} \end{cases}$$

where ⊙ indicates the frame-level multiplication of confidence values with the LSTM output matrix $H^{T \times M}$ ($M = 4096$ in our case, see also Fig. 8). Note that by multiplying w_t with h_t, the impact of the affordance information on the final prediction changes, since $\hat{H}^{T \times M}$ is concatenated with the appearance feature vector, as depicted in Fig. 9. The outcome of the concatenation is then followed by a Softmax layer, used to compute the object class posteriors.

4.4. Two-stream spatio-temporal 3D model (ST3D)

An alternative approach for modeling the spatio-temporal nature of time-evolving interactions is by incorporating the 3D CNN structure in a two-stream model. The two-stream Spatio-Temporal 3D model (ST3D) consists of two C3D ones, one for appearance and the other for affordance information processing. Note that we choose to process the appearance information using a C3D instead of a VGG16 model, as we observed that despite its slightly inferior performance as single-stream classifier for object recognition (see Section 5.1 and Table 3), it performs better when it is combined with the affordance C3D. Additionally, since its structure is very similar to VGG16, we investigate the same three fusion schemes as for the SP model (*i.e.* ST3D$_{L-FC}$, ST3D$_{L-CONV}$, ST3D$_{ML}$). The aforementioned fusion schemes are depicted in Fig. 10.

Note that, unlike ST2D, the C3D models used as appearance and affordance streams can selectively attend to both appearance and motion information. To support this hypothesis, Tran et al. (2015) use the deconvolution method proposed in Zeiler and Fergus (2014) to visualize the patterns learned by the C3D weights over video samples. They report that based on observations, the C3D starts by focusing on appearance in the first few frames and tracks the salient motion in the subsequent ones. Thus, unlike ST2D, no extra attention mechanism is incorporated to the model.

4.5. Auxiliary loss function

The training objective for the proposed two-stream models is to minimize the cross-entropy loss between the predicted object class and the ground truth. This loss is used to compute the gradients and update the weights of both streams. However, besides incorporating affordance information to improve object class prediction, further optimization of the models weights using an auxiliary loss function based solely on the affordance stream performance can be beneficial. In order to compute the auxiliary loss, the affordance features prior to fusion are used to train a Softmax classifier. The training objective of the new classifier is to minimize the cross-entropy loss between the predicted affordance class and the affordance ground truth. The two loss functions can be combined and used to optimize both streams. This aggregated loss is computed as:

$$\mathcal{L}_{agg} = -\frac{1}{K} \sum_{k=1}^{K} \left(y_{o,k} \log(p_{o,k}) + y_{a,k} \log(p_{a,k}) \right) \ , \tag{7}$$

where K is the total number of training samples, $y_{o,k}$ and $p_{o,k}$ are the object ground truth and predicted probability, and $y_{a,k}$ and $p_{a,k}$ are the affordance ground truth and predicted probability of sample k. The auxiliary loss can be applied to the affordance stream of any two-stream model, except for the ST2D$_{IM}$ where the two streams are fused before the LSTM. Fig. 11 depicts an example of the auxiliary loss applied to the SP$_{ML}$ model.

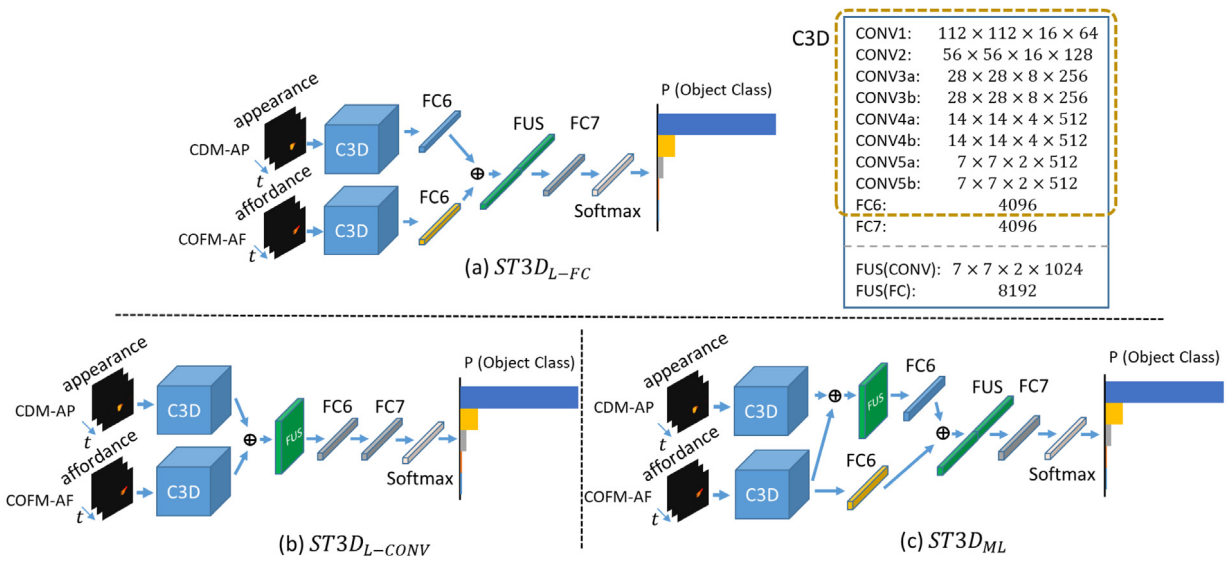


Fig. 10. Detailed architecture of the ST3D model for: (a) late fusion at the FC layer; (b) late fusion at the CONV layer; and (c) multi-layer fusion. At the upper right side the dimensionality of the activation matrix for each CONV block is reported as "$height \times width \times channels$" and for each FC layer as the number of neurons. The appearance stream processes HHA-AP or CDM-AP inputs, while the HHA-AF, CDM-AF, COF-AF, and COFM-AF representations can be used as input to the affordance stream.

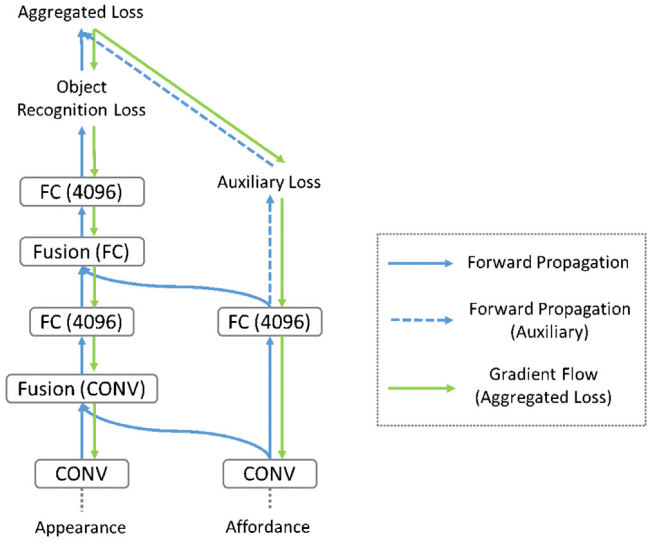


Fig. 11. Example of the auxiliary loss application on the SP_{ML} model. The auxiliary loss is computed based on the affordance classifier and then combined with the object recognition loss.

5. Experimental results

The presented single-stream and fusion models were evaluated using the SOR3D dataset for the task of object recognition. The data captured from the three viewpoints (see also Fig. 3) were accumulated into a unified (*i.e.* all-viewpoint) dataset, which was then split into training, validation, and test sets (25%, 25% and 50%) that correspond to approximately 5k, 5k and 10k hand–object interaction videos, respectively. For all 300 × 300 pixel extracted video frames, a 224 × 224 patch was randomly cropped and used as input to the models. All models were trained with the negative log-likelihood criterion, whereas for back-propagation, Stochastic Gradient Descent (SGD) with 0.9 momentum was used. The standalone VGG16 network was pre-trained on ImageNet (Deng et al., 2009), while the VGG16-LSTM and the C3D were pre-trained on Sports-1M (Karpathy et al., 2014). Subsequently, all models were fine-tuned on the SOR3D dataset with learning rate set to 5×10^{-3}, decreased by a factor of 5×10^{-1} when the validation accuracy curve plateaued. For fusion models training, $L2$ regularization (Ng, 2004) was incorporated in order to prevent over-fitting. All models were implemented using the Torch7 framework[2] on an Nvidia Titan X GPU.

5.1. Single-stream model evaluation

The first set of experiments deals with the evaluation of the single-stream models presented in Section 3.2. The results are reported in Table 3 in terms of overall object and affordance recognition accuracy. For each video sequence, a set of 20 uniformly selected video frames was provided to each single-stream model, while due to computational and memory restrictions the input sequence length for the C3D model was set to 8 frames. For the latter, a sliding window of 8 frames was applied to each sequence and the window-level predictions were averaged to provide a video-level one. The aforementioned setup was used during both training and testing. The frame-level predictions of the VGG16 were also averaged to provide a single prediction for each video.

Regarding object recognition, VGG16 yielded the best overall accuracy compared to the VGG16-LSTM and C3D models for both CDM-AP and HHA-AP input representations. From the aforementioned representations, CDM-AP performed slightly better than HHA-AP (*i.e.* 85.12% over 84.98%), mainly due to the nature of the captured data, *i.e.* height and disparity are more informative in outdoor scenes, or indoor ones that consist of large objects (*e.g.* furniture). Based on the reported results, the VGG16 model that processes CDM-AP input representation was considered as the appearance-only baseline for the rest of the experiments. Further, due to the CDM-AP superiority over HHA-AP, the former was considered as appearance input representation for all two-stream models evaluation.

In order to truly understand the impact of the affordance information on object recognition, we firstly evaluated the affordance encoding efficiency of each single-stream model. For this experiment only the affordance information was utilized providing the target labels, with the experimental framework remaining unaltered. However, the last layer of each network was restructured to predict probabilities based on the 13 affordance classes. From the results reported in Table 3,

[2] http://torch.ch.

Table 3
Recognition accuracy of the three single-stream models of Section 3.2 on the test set of the SOR3D database for various appearance and affordance input stream representations. Object recognition accuracy (%) is reported in the upper part of the table (appearance stream) and affordance recognition accuracy (%) in the lower part (affordance stream).

Input Stream	VGG16	VGG16-LSTM	C3D
HHA-AP	84.98	73.96	84.45
CDM-AP	85.12	74.33	84.67
HHA-AF	56.89	67.44	79.12
CDM-AF	57.28	69.27	81.44
COF-AF	58.32	68.02	82.68
COFM-AF	58.49	68.85	83.19
ACOF-AF	80.84	n/a	n/a
ACOFM-AF	81.92	n/a	n/a

Table 4
Object recognition results (in accuracy, %) on the SOR3D test set, using different SP-based fusion and training schemes and affordance inputs (in conjunction with CDM-AP input).

Input Stream (Regularization)	SP_{L-FC}	SP_{L-CONV}	SP_{ML}
ACOF-AF	87.03	87.93	89.10
ACOFM-AF	87.40	88.24	89.43
ACOFM-AF (aux. loss)	88.37	89.63	90.79
ACOFM-AF ($L2$)	87.92	88.55	89.95
ACOFM-AF (aux. loss, $L2$)	88.54	89.81	91.12

we can conclude that when processing individual frames (*i.e.* HHA-AF, CDM-AF, COF-AF, and COFM-AF) the VGG16 model cannot implicitly capture the temporal information of the affordance. Additionally, we observe that the VGG16-LSTM model cannot efficiently encode the temporal correlations of the sequence, mainly due to the short and fine-grained interaction. On the other hand, the C3D model yields satisfactory results for all affordance input representations, while the VGG16 one performs considerably well when using accumulated 3D flow as input.

5.2. Two-stream model evaluation

In this section, the fusion models evaluation is detailed. It should be noted that for all fusion model experiments the appearance stream receives CDM input (CDM-AP), as discussed in Section 3.1. Thus, the appearance input is not reported in Tables 4–8 for simplicity.

5.2.1. SP model evaluation

Table 4 shows the performance of the SP model, in terms of object recognition accuracy. From the presented results, it can be seen that using the ACOFM input representation is advantageous compared to the ACOF one. Thus for the rest of the SP model evaluation, the former representation is utilized. Further, the late fusion of CONV features (*i.e.* fusion after RL5) appears to perform better compared to the late fusion at the FC layer level. Note that at the FC layers the spatial information is lost, thus fusing CONV layer activations leads to more discriminative post-fusion features. Interestingly, SP_{ML} outperforms the aforementioned late fusion schemes. Using this fusion approach, the model learns both mid-level and high-level feature representations, without losing the spatial correspondence due to the feature-flattening at the FC layers.

Additionally, significant performance improvement can be observed when the auxiliary loss (see Section 4.5) is incorporated. This result reflects the importance of the affordance modeling optimization in parallel with the overall object recognition task. Further, regularization with the $L2$ norm leads to higher accuracy. In fact, the SP_{ML} model trained using the auxiliary loss and the $L2$ norm outperforms the appearance-only VGG16 by an absolute 6%.

Fig. 12(b) visualizes the confusion matrix of the best performing SP_{ML} model (ACOFM-AF, aux. loss, $L2$) on the SOR3D test set. It can be observed that this fusion scheme boosts recognition performance of all supported objects over the appearance-only VGG16 with CDM-AP input (see Fig. 12(a)), demonstrating the additional discriminative power of affordance information.

Table 5
Object recognition results using the $ST2D_{IM}$ and $ST2D_{L}$ models in conjunction with the averaging (AVG) and weighting (W) video-level prediction approaches for various affordance input representations and CDM-AP input.

Input Stream	$ST2D_{IM}$-AVG	$ST2D_{IM}$-W	$ST2D_{L}$-AVG	$ST2D_{L}$-W
HHA-AF	79.33	80.17	86.12	86.53
CDM-AF	79.65	80.43	86.50	86.87
COF-AF	78.98	79.94	86.30	86.64
COFM-AF	79.08	80.04	86.38	86.72

Table 6
Object recognition results using different ST2D-based fusion and training schemes and affordance inputs (in conjunction with CDM-AP input).

Input Stream (Regularization).	$ST2D_{IM}$-W	$ST2D_{L}$-W	$ST2D_{L}$-W (attention)
HHA-AF	80.17	86.53	89.14
CDM-AF	80.43	86.87	89.84
COF-AF	79.94	86.64	89.91
COFM-AF	80.04	86.72	90.02
COFM-AF (aux. loss)	n/a	86.86	90.18
COFM-AF ($L2$)	80.42	86.78	90.09
COFM-AF (aux. loss, $L2$)	n/a	86.95	90.31

5.2.2. ST2D model evaluation

Experimental results of the ST2D-based fusion evaluation are reported in Tables 5–7. In all cases, as in Section 5.1, a set of 20 uniformly selected frames was provided as input to the respective networks.

Table 5 reports the comparative evaluation of the averaging and weighting video-level prediction for the $ST2D_{L}$ and $ST2D_{IM}$ models. It can be observed that, for both fusion schemes and all affordance input representations, the weighting approach leads to better overall accuracy than the averaging one. Thus, for the rest of the ST2D experiments reported in Tables 6 and 7, the weighting video-level prediction is adopted.

Table 6 shows that the $ST2D_{IM}$ performs worse than the appearance-only VGG16 (*i.e.* 80.43% over 85.12%). Thus, we conclude that the LSTM cannot efficiently encode the time-evolving object manipulation, using a sequence of fused representations as input. Note that the latter is the result of fusing the two information streams at the FC layer-level, where the spatial correspondence is lost; thus, the LSTM has difficulty learning temporal correlations for both appearance and affordance. In contrast, the $ST2D_{L}$ fusion scheme outperforms the appearance-only VGG16 for all affordance input representations. In detail, the $ST2D_{L}$ scheme with HHA-AF input yields an absolute improvement of 1.41% compared to the appearance-only VGG16, which is further improved by CDM-AF, COF-AF, and COFM-AF to 1.75%, 1.52%, and 1.6% boosts, respectively.

The performance of $ST2D_{L}$ is further improved when the attention mechanism is incorporated. Based on Table 7, the $N-$best log-likelihood dispersion metric ($N = 3$) is selected as it yields the best overall accuracy. The inclusion of the attention mechanism leads to a performance boost for all affordance input representations (see rightmost column of Table 6). Note also that the attention-based model using COFM-AF slightly outperforms the ones that use HHA-AF and CDM-AF as input representations. One plausible reason is that the 3D optical flow of the hand movement, prior to and after the interaction, may not contain significant affordance information, thus its impact to the final prediction should be small for the corresponding frames. The application of the auxiliary loss to the attention-based $ST2D_{L}$ model with COFM-AF input, in combination with $L2$ regularization, yields a 90.31% object recognition accuracy.

Table 7
Object recognition results of the $ST2D_L - W$ model with CDM-AP and CDM-AF inputs using attention in conjunction with the following confidence estimation metrics: (a) the entropy, (b) the average $N-$best log-likelihood difference ($N = 3$), and (c) the $N-$best log-likelihood dispersion ($N = 3$).

Confidence Metric	Test Acc. (%)
Entropy	88.91
N-best difference	89.27
N-best dispersion	89.84

Table 8
Object recognition results using different ST3D-based fusion schemes and affordance inputs. CDM-AP is used as appearance input representation.

Input Stream (Regularization)	$ST3D_{L-FC}$	$ST3D_{L-CONV}$	$ST3D_{ML}$
HHA-AF	87.12	87.92	88.76
CDM-AF	87.97	88.32	89.23
COF-AF	88.06	88.65	89.79
COFM-AF	88.49	89.14	90.47
COFM-AF (aux. loss)	89.12	90.02	91.44
COFM-AF ($L2$)	88.86	89.58	90.70
COFM-AF (aux. loss, $L2$)	89.67	90.88	**91.98**

The confusion matrix of the attention-based $ST2D_L$ (COFM-AF, aux. loss, $L2$) model on the SOR3D test set is given in Fig. 12(c). It can be seen that it confuses objects that are very small or thin and their manipulation is very similar (*e.g.* small-size ones, like "Key", "Pen", etc.).

5.2.3. ST3D model evaluation

For the ST3D model evaluation, we used sequences of 20 uniformly selected frames as input for each stream combined with an 8-frame sliding window, similarly to the single-stream C3D experiment (see Section 5.1). Note that in contrast to the LSTM learning process (see Section 4.2), the window-level predictions of the C3D model are independent from each other, thus for the final prediction the averaging approach was used.

Table 8 reports the overall accuracy of the ST3D models. Similarly to the SP evaluation, $ST3D_{ML}$ outperforms $ST3D_L$ for all affordance inputs, due to the information sharing at different levels of granularity. Additionally, training both schemes with the auxiliary loss and $L2$ regularization leads to additional performance improvement. From the reported results, it can be observed that using 3D flow information instead of colorized depthmaps is advantageous. The latter is in accordance with the results presented in Table 3 for affordance recognition. Furthermore, it must be noted that $ST3D_{ML}$ with COFM-AF input, which is the best performing approach in this paper, outperforms the appearance-only VGG16 by 6.86% (*i.e.* 91.98% over 85.12%), which corresponds to an approximately 46% relative error reduction.

From a practical perspective, the ST3D model handles both the lack of temporal information modeling of the SP model and the difficulty of the ST2D one to learn the spatio-temporal correlations of fine-grained interactions. Additionally, it can better exploit 3D optical flow information, which explicitly describes the motion between sequential frames, thus making the recognition easier as the network does not need to estimate motion implicitly.

Finally, the confusion matrix of the $ST3D_{ML}$ model (COFM-AF, aux. loss, $L2$) on the SOR3D test set is depicted in Fig. 12(d). Notice that this model boosts recognition performance of all objects, while further improving it for the most challenging ones (*e.g.* "Key", "Knife", and "Pen"), by modeling the affordance information more efficiently.

5.3. Comparison with probabilistic fusion

The best performing fusion model, namely $ST3D_{ML}$ that utilizes COFM-AF as input and is trained with auxiliary loss and $L2$ regularization, is also comparatively evaluated against typical probabilistic fusion approaches of the literature. To perform a fair comparison, two C3D models are trained following the process presented in Section 5.1, using CDM-AP and COFM-AF as input representations. The product rule for fusing the appearance and the affordance C3D output probabilities is adopted as the first probabilistic approach. Additionally, after removing both Softmax classifiers, the concatenated FC7 activations of the appearance and affordance C3D models are used to train a one-versus-all SVM classifier with RBF kernel (Castellini et al., 2011; Kjellström et al., 2011), as well as a naive Bayes classifier (Högman et al., 2016). From the results presented in Table 9, it can be observed that the evaluated probabilistic fusion approaches fail to increase object recognition accuracy compared to the appearance-only VGG16 baseline. On the contrary, the proposed $ST3D_{ML}$ model exhibits a significant performance increase.

Table 9
Comparative evaluation of the $ST3D_{ML}$ (CDM-AP, COFM-AF, aux. loss, $L2$) model, three probabilistic fusion methods, and the appearance-only VGG16 baseline. In all cases, object recognition accuracy (%) is reported.

Model	Fusion Layer	Test Acc. (%)
Appearance-only VGG16 baseline	no fusion	85.12
Product Rule	Softmax	77.91
SVM (Kjellström et al., 2011)	RL7	84.77
Bayes (Högman et al., 2016)	RL7	80.63
$ST3D_{ML}$	RL5, RL6	**91.98**

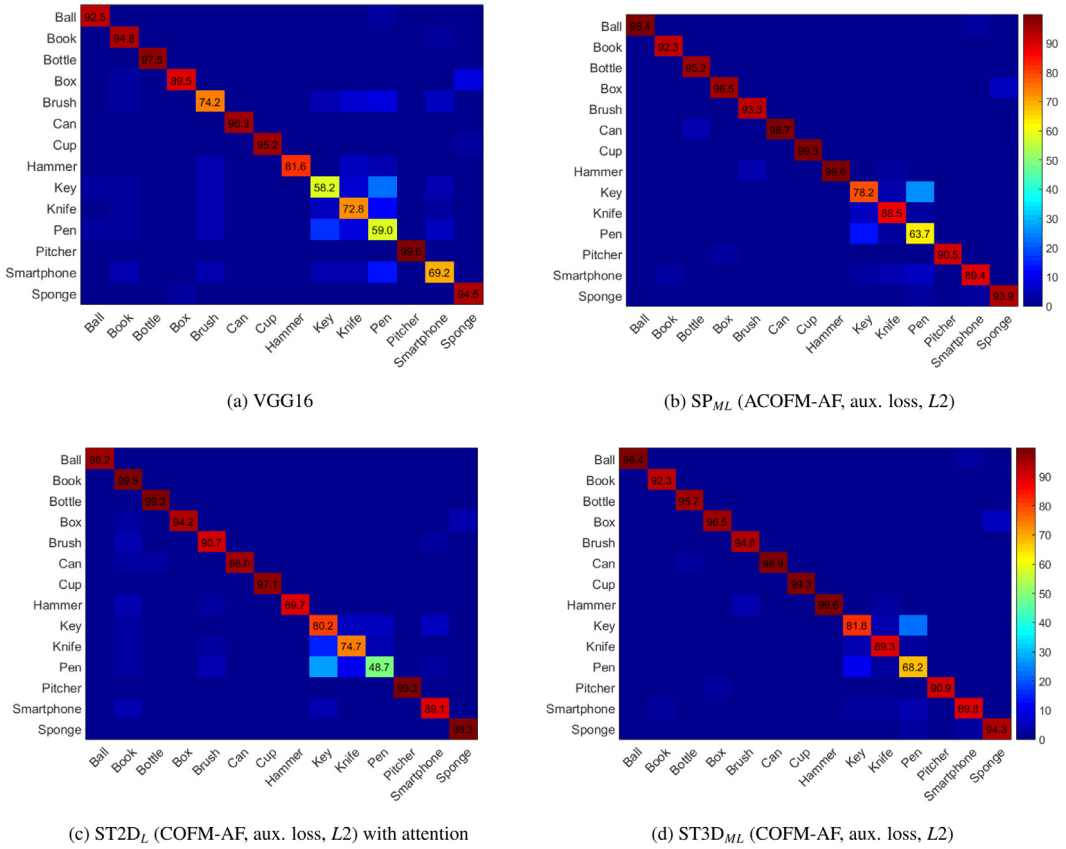
Table 10
Cross-view object recognition results using the appearance-only VGG16 and $ST3D_{ML}$ (COFM-AF, aux. loss, $L2$) models. The last row reports the results using the original SOR3D training and test sets.

Training Set	Test Set	VGG16	$ST3D_{ML}$
K_1	K_2, K_3	51.74	55.13
K_2	K_1, K_3	53.28	57.80
K_3	K_1, K_2	49.42	53.96
K_2, K_3	K_1	62.43	69.74
K_1, K_3	K_2	66.14	72.86
K_1, K_2	K_3	78.65	85.33
K_1, K_2, K_3	K_1, K_2, K_3	85.12	91.98

5.4. Cross-view analysis

In this section, we perform a cross-view analysis on the SOR3D data, in order to evaluate the contribution of each viewpoint to the performance of the appearance-only VGG16 and the $ST3D_{ML}$ model. For this analysis, the three viewpoints of the SOR3D capturing setup, depicted in Fig. 3, are denoted as K_1, K_2, and K_3. For each $K_i, i \in V = \{1, 2, 3\}$, the evaluated model is initially trained using the K_i data and tested on the $K_{V-\{i\}}$ set, and then trained with $K_{V-\{i\}}$ data and tested on the K_i one. It must be noted that no viewpoint fusion is considered for any of the experiments. The appearance-only VGG16 employs the CDM-AP input representation, while the $ST3D_{ML}$ utilizes CDM-AP and COFM-AF inputs, auxiliary loss, and $L2$ regularization.

Intuitively, the affordance information should be significantly more viewpoint-dependent, since the starting point of the hand movement is different from each viewpoint and the actual interaction may not always be visible (*e.g.* the handle of the cup might be from the opposite side of the RGB-D sensor). From the results presented in Table 10, it can be observed that both models perform worse when trained on one or two viewpoints and tested on the rest. Additionally, it can be seen that, contrary to the aforementioned intuition, the starting point of the hand does not significantly affect the models performance, and the affordance information is discriminant even if some parts of the interaction are not entirely visible. We can further conclude that K_1 and K_2 are the most critical viewpoints for both appearance and affordance exploitation, as their absence from the training set leads to inferior overall classification accuracy.



(a) VGG16

(b) SP_{ML} (ACOFM-AF, aux. loss, $L2$)

(c) $ST2D_L$ (COFM-AF, aux. loss, $L2$) with attention

(d) $ST3D_{ML}$ (COFM-AF, aux. loss, $L2$)

Fig. 12. Object recognition confusion matrices of the appearance-only VGG16 and the best performing fusion scheme of each two-stream model. Training parameters, such as affordance input and regularization of each model, are reported inside the parentheses. In all cases, CDM-AP is used as the appearance stream representation.

6. Conclusion

In this paper, the application of sensorimotor learning in RGB-D object recognition was investigated, following the observation learning scenario. Three DL-based models that fuse appearance and affordance information by adopting multiple fusion schemes were presented. Further, six alternative representations were used as input to the affordance stream in order to maximize the information gain by incorporating affordance information. An attention mechanism based on appearance stream confidence was developed, and an auxiliary loss for fusion model optimization based on the affordance stream performance was also introduced. The 3D convolution based two-stream model with multi-layer fusion was experimentally shown to significantly improve the appearance-only baseline and outperform the rest of the proposed models, as well as alternative probabilistic fusion methods of the literature. A cross-view analysis concluded the study, providing intuition concerning viewpoint contribution to model performance and viewpoint-dependency of the affordance information.

Our future research goals regarding sensorimotor learning are to evaluate the proposed algorithms on more challenging data and to experiment with more lightweight models. Regarding the former, the SOR3D dataset will be enhanced by adding more object types coupled with higher intra-class variance, as well as new interaction sessions combining multiple affordances and objects in a more cluttered and close to real-world setup (*e.g.* using multiple objects to cook in a kitchen). The enriched dataset will enable the development of sensorimotor algorithms applicable in real-world scenarios such as smart homes (Poland et al., 2009; Alam et al., 2012). A state-of-the-art semantic segmentation method will be added to address the hand–object separation in these scenarios. Further, the exploitation of shallower models with less parameters will be investigated, aiming at lightweight two-stream architectures with close to state-of-the-art performance that will be embedded in artificial agents for real-world applications.

Acknowledgments

The work presented in this paper was partially supported by the European Commission under contract H2020-762111 VRTogether.

References

Adjoudani, A., Benoît, C., 1996. On the integration of auditory and visual parameters in an HMM-based ASR. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading By Humans and Machines: Models, Systems, and Applications. Springer, Berlin Heidelberg, pp. 461–471.

Alam, M.R., Reaz, M.B.I., Mohd Ali, M., 2012. A review of smart homes – Past, present, and future. IEEE Trans. Syst. Man Cybern. C 42 (6), 1190–1203.

Andreopoulos, A., Tsotsos, J.K., 2013. 50 years of object recognition: Directions forward. Comput. Vis. Image Underst. 117 (8), 827–891.

Brandi, M.L., Wohlschläger, A., Sorg, C., Hermsdörfer, J., 2014. The neural correlates of planning and executing actual tool use. J. Neurosci. 34 (39), 13183—13194.

Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B., 2011. Using object affordances to improve object recognition. IEEE Trans. Auton. Mental Dev. 3 (3), 207–215.

Cloutman, L.L., 2013. Interaction between dorsal and ventral processing streams: Where, when and how? Brain Language 127 (2), 251–263.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 248–255.

DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73 (3), 415—434.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., 2015. Multimodal deep learning for robust RGB-D object recognition. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 681–687.

Feng, Y., Zhang, Z., Zhao, X., Ji, R., Gao, Y., 2018. GVCNN: Group-view convolutional neural networks for 3D shape recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 264–272.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., Sandini, G., 2003. Learning about objects through action - initial steps towards artificial cognition. In: Proc. IEEE International Conference on Robotics and Automation, ICRA, vol. 3, pp. 3140–3145.

Flavell, J.H., 1992. Cognitive development: Past, present, and future. Dev. Psychol. 28 (6), 998–1005.

Ghadirzadeh, A., Bütepage, J., Kragic, D., Björkman, M., 2016. Self-learning and adaptation in a sensorimotor framework. In: Proc. IEEE International Conference on Robotics and Automation, ICRA, pp. 551–558.
Giagkos, A., Lewkowicz, D., Shaw, P., Kumar, S., Lee, M., Shen, Q., 2017. Perception of localized features during robotic sensorimotor development. IEEE Trans. Cognitive Dev. Syst. 9 (2), 127–140.
Gibson, J.J., 1977. The theory of affordances. In: Shaw, R., Bransford, J. (Eds.), Perceiving, Acting, and Knowing: Toward an Ecological Psychology. Lawrence Erlbaum, Hillsdale NJ, pp. 67–82.
Gupta, S., Girshick, R.B., Arbeláez, P.A., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation. In: Proc. European Conference on Computer Vision, ECCV, pp. 345–360.
Hadfield, S., Bowden, R., 2014. Scene particles: Unregularized particle-based scene flow estimation. IEEE Trans. Pattern Anal. Mach. Intell. 36 (3), 564–576.
Hassanin, M., Khan, S., Tahtali, M., 2018. Visual affordance and function understanding: A survey. CoRR abs/1807.06775.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Högman, V., Björkman, M., Maki, A., Kragic, D., 2016. A sensorimotor learning framework for object categorization. IEEE Trans. Cognitive Dev. Syst. 8 (1), 15–25.
Hong, C., Yu, J., You, J., Chen, X., Tao, D., 2015. Multi-view ensemble manifold regularization for 3D object recognition. Inform. Sci. 320, 395–405.
Hornácek, M., Fitzgibbon, A., Rother, C., 2014. SphereFlow: 6 DoF scene flow from RGB-D pairs. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 3526–3533.
Jaimez, M., Souiai, M., Gonzalez-Jimenez, J., Cremers, D., 2015. A primal-dual framework for real-time dense RGB-D scene flow. In: Proc. IEEE International Conference on Robotics and Automation, ICRA, pp. 98–104.
Jang, E., Vijayanarasimhan, S., Pastor, P., Ibarz, J., Levine, S., 2017. End-to-end learning of semantic grasping. In: Levine, S., Vanhoucke, V., Goldberg, K. (Eds.), Proc. Conference on Robot Learning. In: Proceedings of Machine Learning Research, PMLR, vol. 78, pp. 119–132.
Jayaraman, D., Grauman, K., 2018. End-to-end policy learning for active visual categorization. IEEE Trans. Pattern Anal. Mach. Intell. (early access).
Kanezaki, A., Matsushita, Y., Nishida, Y., 2018. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 5010–5019.
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1725–1732.
Kjellström, H., Romero, J., Kragic, D., 2011. Visual object-action recognition: Inferring object affordances from human demonstration. Comput. Vis. Image Underst. 115 (1), 81–90.
Kluth, T., Nakath, D., Reineking, T., Zetzsche, C., Schill, K., 2014. Affordance-based object recognition using interactions obtained from a utility maximization principle. In: European Conference on Computer Vision Workshops, ECCVW, pp. 406–412.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lee, K., Lee, K., Min, K., Zhang, Y., Shin, J., Lee, H., 2018. Hierarchical novelty detection for visual object recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1034–1042.
Liang, M., Hu, X., 2015. Recurrent convolutional neural network for object recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 3367–3375.
Lyubova, N., Ivaldi, S., Filliat, D., 2016. From passive to interactive object learning and recognition through self-identification on a humanoid robot. Auton. Robots 40 (1), 33–57.
Minsky, M., 1991. Society of mind: A response to four reviews. Artificial Intelligence 48 (3), 371–396.
Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J., 2008. Learning object affordances: from sensory motor coordination to imitation. IEEE Trans. Robot. 24 (1), 15–26.
Mottaghi, R., Schenck, C., Fox, D., Farhadi, A., 2017. See the glass half full: reasoning about liquid containers, their volume and content. In: Proc. IEEE International Conference on Computer Vision, ICCV, pp. 1889–1898.
Ng, Andrew Y., 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Proc. International Conference on Machine Learning (ICML),
Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G., 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp. 5908–5915.
Oberlin, J., Tellex, S., 2018. Autonomously acquiring instance-based object models from experience. In: Bicchi, A., Burgard, W. (Eds.), Robotics Research, vol. 2. Springer International Publishing, pp. 73–90.
O'Regan, J.K., Noë, A., 2001. A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. 24 (5), 939–1031.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Piaget, J., Brown, T., 1985. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. University of Chicago Press.
Poland, M.P., Nugent, C.D., Wang, H., Chen, L., 2009. Smart home research: Projects and issues. Int. J. Ambient Comput. Intell. 1 (4), 32–45.
Potamianos, G., Neti, C., 2000. Stream confidence estimation for audio-visual speech recognition. In: Proc. International Conference on Spoken Language Processing, ICSLP, vol. 3, pp. 746–749.
Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J., 2016. Volumetric and multi-view CNNs for object classification on 3D data. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 5648–5656.
Quiroga, J., Brox, T., Devernay, F., Crowley, J.L., 2014. Dense semi-rigid scene flow estimation from RGBD images. In: Proc. European Conference on Computer Vision, ECCV, pp. 567–582.
Rivlin, E., Dickinson, S.J., Rosenfeld, A., 1995. Recognition by functional parts. Comput. Vis. Image Underst. 62 (2), 164–176.
Saxena, A., Driemeyer, J., Ng, A.Y., 2008. Robotic grasping of novel objects using vision. Int. J. Robot. Res. 27 (2), 157–173.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations, ICLR.
Song, H.O., Fritz, M., Goehring, D., Darrell, T., 2016. Learning to detect visual grasp affordance. IEEE Trans. Autom. Sci. Eng. 13 (2), 798–809.
Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition. In: Proc. IEEE International Conference on Computer Vision, ICCV, pp. 945–953.
Sutton, M., Stark, L., Bowyer, K., 1998. Function from visual analysis and physical interaction: a methodology for recognition of generic classes of objects. Image Vis. Comput. 16 (11), 745–763.
Thermos, S., Papadopoulos, G. Th., Daras, P., Potamianos, G., 2017. Deep affordance-grounded sensorimotor object recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 49–57.
Thermos, S., Papadopoulos, G. Th., Daras, P., Potamianos, G., 2018. Attention-enhanced sensorimotor object recognition. In: IEEE International Conference on Image Processing, ICIP, pp. 336–340.
Tommasi, T., Orabona, F., Caputo, B., 2010. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, pp. 3081–3088.
Tran, D., R., L. Bourdev, Fergus, Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks. In: Proc. IEEE International Conference on Computer Vision, ICCV, pp. 4489–4497.
Ungerleider, L.G., Haxby, J.V., 1994. 'What' and 'where' in the human brain. Curr. Opin. Neurobiol. 4, 157–165.
van Polanen, V., Davare, M., 2015. Interactions between dorsal and ventral streams for controlling skilled grasp. Neuropsychologia 79, 186–191.
Vezhnevets, V., Sazonov, V., Andreeva, A., 2003. A survey on pixel-based skin color detection techniques. In: Proc. GRAPHICON, pp. 85–92.
Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P., 2017. Scene flow to action map: a new representation for RGB-D based action recognition with convolutional neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 416–425.
Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: Proc. International Conference on Neural Information Processing Systems, NIPS'14, vol. 2. MIT Press, Cambridge, MA, pp. 3320–3328.
Yu, T., Meng, J., Yuan, J., 2018. Multi-view harmonized bilinear network for 3D object recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 186–194.
Zambelli, M., Demiris, Y., 2017. Online multimodal ensemble learning using self-learned sensorimotor representations. IEEE Trans. Cognitive Dev. Syst. 9 (2), 113–126.
Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Proc. European Conference on Computer Vision, ECCV, pp. 818–833.
Zhu, Y., Zhao, Y., Zhu, S.C., 2015. Understanding tools: Task-oriented object modeling, learning and recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2855–2864.