Deep Affordance-grounded Sensorimotor Object Recognition

Spyridon Thermos[1,2]   Georgios Th. Papadopoulos[1]   Petros Daras[1]   Gerasimos Potamianos[2]

[1]Information Technologies Institute, Centre for Research and Technology Hellas, Greece
[2]Department of Electrical and Computer Engineering, University of Thessaly, Greece

{spthermo,papad,daras}@iti.gr   gpotam@ieee.org

Abstract

It is well-established by cognitive neuroscience that human perception of objects constitutes a complex process, where object appearance information is combined with evidence about the so-called object "affordances", namely the types of actions that humans typically perform when interacting with them. This fact has recently motivated the "sensorimotor" approach to the challenging task of automatic object recognition, where both information sources are fused to improve robustness. In this work, the aforementioned paradigm is adopted, surpassing current limitations of sensorimotor object recognition research. Specifically, the deep learning paradigm is introduced to the problem for the first time, developing a number of novel neuro-biologically and neuro-physiologically inspired architectures that utilize state-of-the-art neural networks for fusing the available information sources in multiple ways. The proposed methods are evaluated using a large RGB-D corpus, which is specifically collected for the task of sensorimotor object recognition and is made publicly available. Experimental results demonstrate the utility of affordance information to object recognition, achieving an up to 29% relative error reduction by its inclusion.

1. Introduction

Object recognition constitutes an open research challenge of broad interest in the field of computer vision. Due to its impact in application fields such as office automation, identification systems, security, robotics, and the industry, several research groups have devoted intense efforts in it (see for example the review in [2]). However, despite the significant advances achieved in the last decades, satisfactory performance in real-world scenarios remains a challenge. One plausible reason is the sole use of static object appearance features [8, 18, 19]. Such cannot sufficiently handle the object appearance variance, occlusions, deformations, and illumination variation.

Research findings in cognitive neuroscience establish that object recognition by humans exploits previous experiences of active interaction with the objects of interest. In particular, object perception is based on the fusion of sensory (object appearance) and motor (human-object interaction) information. Central role in this so-called "sensorimotor object recognition" theory has the notion of object affordances. According to Gibson [10], "the affordances of the environment are what it offers the animal", implying the complementarity between the animal and the environment. Based on this theory, Minsky [22] argues on the significance of classifying items according to what they can be used for, *i.e.* what they afford. These theoretical foundations have resulted to the so-called function-based reasoning in object recognition, which can be viewed as an approach applicable to environments in which objects are designed or used for specific purposes [27]. Moreover, the work in [30] describes three possible ways for extracting functional (affordance) information for an object: a) "Function from shape", where the object shape provides some indication of its function; b) "Function from motion", where an observer attempts to understand the object function by perceiving a task being performed with it; and c) "Function from manipulation", where function information is extracted by manipulating the object. The present work focuses on (b).

Concerning the neuro-physiological and the corresponding cognitive procedures that take place in the human brain during sensorimotor object recognition, it is well established that there are two main streams that process the visual information [32]: a) the dorsal, which projects to the posterior parietal cortex and is involved in the control of actions (motor), and b) the ventral, which runs to the inferotemporal cortex and is involved in the identification of the objects (sensory). There is accumulated evidence that the two streams interact at different information processing stages [5]: a) computations along the two pathways proceed both

independently and in parallel, reintegrating within shared target brain regions; b) processing along the separate pathways is modulated by the existence of recurrent feedback loops; and c) information is transferred directly between the two pathways at multiple stages and locations along their trajectories. These identified interconnections indicate how the human brain fuses sensory and motor information to achieve robust recognition. Motivated by the above, mimicking the human sensorimotor information processing module in object recognition by machines may hold the key to address the weaknesses of current systems.

Not surprisingly, affordance-based information has already been introduced into the object recognition problem. However, current systems have been designed based on rather simple classification, fusion, and experimental frameworks, failing to fully exploit the potential of the affordance stream. In particular, these works have not yet exploited the recent trend in computer vision of employing very deep Neural Network (NN) architectures, the so-called "Deep Learning" (DL) paradigm. DL methods outperform all previous hand-crafted approaches by a large margin [17, 31, 35].

In this paper, the problem of sensorimotor 3D object recognition is investigated using DL techniques. The main contributions lie in the:

- **Design of novel neuro-biologically and neuro-physiologically grounded neural network architectures** for sensorimotor object recognition, exploiting the state-of-the-art automatic feature learning capabilities of DL techniques; to the best of the authors' knowledge, this is the first work that introduces the DL paradigm to sensorimotor object recognition.

- **NN-based implementation of multiple recent neuro-scientific findings** for fusing the sensory (object appearance) and motor (object affordances) information streams in a unified machine perception computational model. Until now, such neuro-scientific findings have not been transferred to computer vision systems.

- **Large number of complex affordance types** supported by the proposed methodology. In particular: a) a significantly increased number of affordances compared to current works that only use few (up to 5); b) complex types of object affordances (*e.g.* squeezable, pourable) that may lead to complex object manipulations or even significant object deformations, compared to the relatively simple binary ones currently present in the literature (*e.g.* graspable, pushable); and c) continuous-nature affordance types, moving beyond plain binary analysis of presence/non-presence of a given affordance, while modeling the exact dynamics of the exhibited affordances.

- Introduction of a **large, public RGB-D object recognition dataset**, containing several types of interactions of human subjects with a set of supported objects. This is the first publicly available sensorimotor object recognition corpus, including 14 classes of objects, 13 categories of affordances, 105 subjects, and a total number of approximately 20,8k human-object interactions. This dataset yields sufficient data for training DL algorithms, and it is expected to serve as a benchmark for sensorimotor object recognition research.

- **Extensive quantitative evaluation** of the proposed fusion methods and comparison with traditional probabilistic fusion approaches.

The remainder of the paper is organized as follows: Section 2 presents related work in the field of sensorimotor object recognition. Section 3 discusses the introduced 3D object recognition dataset. Section 4 details the designed NN architectures. Section 5 presents the experimental results, and Section 6 concludes this paper.

2. Related Work

Most sensorimotor object recognition works have so far relied 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). In particular, Kjellström *et al.* [15] model the spatio-temporal information by training factorial conditional random fields with 3 consecutive object frames and extract action features using binary SVMs; a small dataset with 6 object types and 3 affordances is employed. Högman *et al.* [13] propose a framework for interactive classification and functional categorization of 4 object types, defining a Gaussian Process to model object-related Sensori-Motor Contigencies [25] and then integrating the "push" action to categorize new objects (using a naive Bayes classifier). Additionally, Kluth *et al.* [16] extract object GIST-features [24] and model the possible actions using a probabilistic reasoning scheme that consists of a Bayesian inference approach and an information gain strategy module. A visuo-motor classifier is implemented in [4] in order to learn 5 different types of grasping gestures on 7 object types, by training an SVM model with object feature clusters (using K-means clustering) and a second SVM with 22 motor features (provided by a CyberGlove); the predictions are fused with a weighted linear combination of Mercer kernels. Moreover, in the field of robotics, affordance-related object recognition has relied on predicting opportunities for interaction with an object by using visual clues [1, 11] or observing the effects of exploratory actions [20, 23].

Clearly, design and evaluation of complex data-driven machine perception systems for sensorimotor object recog-

Figure 1. Examples of human-object interactions captured by the 3 Kinect sensors employed in the corpus recording setup.

Dataset	Types	Affordances	Interactions	Subjects	Available
[16]	8	1	n/a	n/a	no
[13]	4	1	4	Robot arm	no
[15]	6	3	7	4	no
[4]	7	5	13	20	no
Introduced	**14**	**13**	**54**	**105**	**yes**

Table 1. Characteristics of sensorimotor object recognition corpora reported in the literature, compared to the presented one. Number of object types, affordances, human-object interactions, and subjects, as well as data public availability, are reported.

nition based on the state-of-the-art DL framework has not been considered in the literature. Such systems should not depend on over-simplified or hard assumptions and would target the automatic learning of the highly complex sensorimotor object recognition process in realistic scenarios.

3. RGB-D Sensorimotor Dataset

In order to boost research in the field of sensorimotor object recognition, a large-scale dataset of multiple object types and complex affordances has been collected and is publicly available at http://sor3d.vcl.iti.gr/. The corpus constitutes the broadest and most challenging one in the sensorimotor object recognition literature, as summarized in Table 1, and can serve as a challenging benchmark, facilitating the development and efficient evaluation of sensorimotor object recognition approaches.

The corpus recording setup involved three synchronized Microsoft Kinect II sensors [21] in order to acquire RGB (1920×1080 resolution) and depth (512×424 resolution) streams from three different viewpoints, all at 30 Hz frame rate and an approximate 1.5 meters "head-to-device" distance. A monitor was utilized for displaying the "prototype" instance before the execution of every human-object interaction. Additionally, all involved subjects were provided with a ring-shaped remote mouse, held by the other hand than that interacting with the objects. This allowed the participants to indicate by themselves the start and end of each session (*i.e.* performing real-time annotation). Before the execution of any interaction, all objects were placed at a specific position on a desk, indicated by a marker on the table cloth. The dataset was recorded under controlled environmental conditions, *i.e.* with negligible illumination variations (no external light source was present during the experiments) and a homogeneous static background (all human-object interactions were performed on top of a desk covered with a green tablecloth). Snapshots of the captured video streams from each viewpoint are depicted in Fig. 1.

Regarding the nature of the supported human-object interactions, a set of 14 object types was considered (each type having two individual instantiations, *e.g.* small and big ball). The appearance characteristics of the selected object types varied significantly, ranging from distinct shapes (like "Box" or "Ball") to more challenging ones (like "Knife"). Taking into account the selected objects, a respective set of 13 affordance types was defined, covering typical manipulations of the defined objects. Concerning the complexity of the supported affordances, relatively simple (*e.g.* "Grasp"), complex (*e.g.* leading to object deformations, like affordance "Squeeze"), and continuous-nature ones (*e.g.* affordance "Write") were included. In contrast, other experimental settings in the literature have mostly considered simpler and less time evolving affordances, like "Grasp" and "Push". In Table 2, all supported types of objects and affordances, as well as all combinations that have been considered in the dataset, are provided. As listed, a total of 54 object-affordance combinations (*i.e.* human-object interactions) are supported. All participants were asked to execute all object-affordance combinations indicated in Table 2 at least once. The experimental protocol resulted in a total of 20,830 instances, considering the data captured from each Kinect as a different human-object interaction instance. The length of every recording varied between 4 and 8 seconds.

4. Sensorimotor Object Recognition

We now proceed to describe the proposed sensorimotor object recognition system. Specifically, we first provide its overview, followed by details of video data processing and the DL modeling approaches considered.

4.1. System overview

The proposed system is depicted in Fig 2. Initially, the collected data are processed by the visual front-end module. This produces three visual feature streams, one of which corresponds to conventional object appearance, while the remaining two capture object affordance information. These streams are subsequently fed to appropriate DL architectures, implementing single-stream processing systems for the recognition of object types and affordances. Eventually, appearance and affordance information are combined to yield improved object recognition, following various fusion strategies.

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	√	√	√	√					√				

Table 2. Supported object and affordance types in the presented corpus. Considered object-affordance combinations are marked with √.

4.2. Visual front-end

The RGB data stream is initially mapped to the depth stream for each Kinect, making use of a typical calibration model [12]. Since the exact positioning of the Kinect sensors is known in the developed capturing framework, a "3D volume of interest" is defined for each Kinect, corresponding to the 3D space above the desk that includes all performed human-object interactions. Pixels that correspond to 3D points outside the defined volumes of interest are considered as background and the respective RGB/depth values are set to zero. Subsequently, a centered rectangular region (300×300 resolution), containing the observed object manipulations, is cropped from the aligned RGB and depth frames. Then, using a simple thresholding technique in the HSV color space [33], pixels corresponding to the desk plane (tablecloth) are removed, and subsequently skin color pixels (corresponding to the hand of the performing subject) are separated from the object ones.

For the extracted object and hand depth maps, a "depth colorization" approach is followed, similar to the one introduced by Eitel *et al.* [9]. The depth colorization enables the common practice of utilizing networks (transfer learning [26, 34]) pre-trained on ImageNet [6], and fine-tuning them on the collected data. In particular, the depth value at every pixel location is linearly normalized in the interval $[0, 255]$, taking into account the minimum and maximum depth values that have been measured in the whole dataset and for all pixel locations. Subsequently, a "hot colormap" is applied for transforming each scalar normalized depth value to a triplet RGB one. In parallel, the 3D flow magnitude of the extracted hand depth maps is also computed. The algorithm of [14] for real-time dense RGB-D scene flow estimation is used. Denoting by $F_t(x, y, z)$ the 3D flow-field of the depth video at frame t, the 3D magnitude field $\sum_{t=1}^{T} |F_t(x, y, z)|$ is considered, accumulated over the video duration (T frames). For the latter field, the same colorization approach with the RGB case is applied (Fig. 3). Thus, the visual front-end provides three information streams (Fig 2 middle): a) colorized object depth maps, b) colorized hand depth maps, and c) colorized hand 3D flow magnitude fields.

4.3. Single-stream modeling

For each information stream, separate NN architectures are designed, as depicted in Fig 4. Regarding the **appearance stream**, the well-known VGG-16 network [29], which consists of 16 layers in total, is used for analyzing the appearance of the observed objects. The VGG-16 model consists of 5 groups of Convolutional (CONV) layers and 3 Fully Connected (FC) ones. After each CONV or FC layer, a Rectified Linear Unit (RL) one follows. For the rest of the paper, the notation depicted in Fig 4 (top) is used (*e.g.* $CONV4_3$ is the 3rd CONV layer of the 4th group of convolutions).

Regarding the **affordance stream**, the colorized hand depth map and 3D flow magnitude are alternatively used for encoding the corresponding dynamics. In particular, two distinct NN architectures, aiming at modeling different aspects of the exhibited motor (hand) actions, are designed: a) a "Template-Matching" (TM), and b) a "Spatio-Temporal" (ST) one. The development of the TM architecture is based solely on the use of CNNs (Fig 4 top), aiming at estimating complex multi-level affordance-related patterns along the spatial dimensions. The different CONV layers of the employed CNN now model affordance-related patterns of increasing spatial complexity. With respect to the development of the ST architecture, a composite CNN (VGG-16) - Long-Short Term Memory (LSTM) [28] NN is considered, where the output of a CNN applied at every frame is subsequently provided as input to an LSTM. This architec-
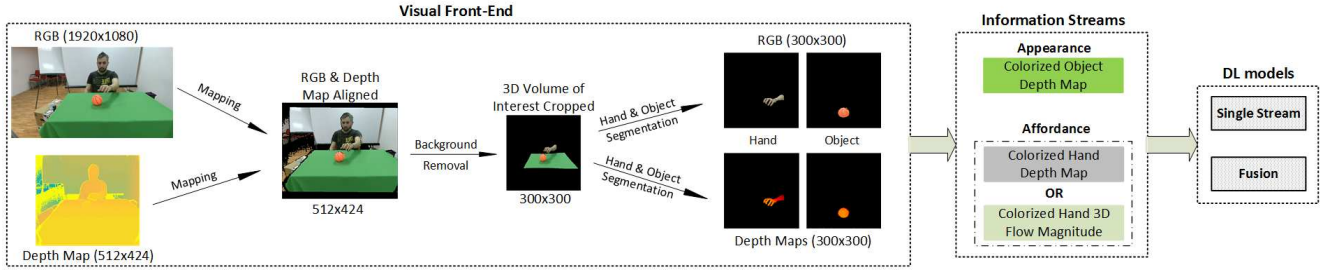
Figure 2. System overview. The visual front-end module (left) processes the captured data, providing three information streams (middle) that are then fed into a single-stream or fusion DL model (right).

ture is similar to the generalized "LRCN" model presented in [7]. The aim of the ST architecture (Fig 4 bottom) is to initially model correlations along the spatial dimensions and subsequently to take advantage of the LSTM sequential modeling efficiency for encoding the temporal dynamics of the observed actions. In preliminary experiments, the colorized hand depth map led to better results than using 3D flow information as input. A set of 20 frames, uniformly sampled over the whole duration of the observed action, are provided as input to the LSTM.

4.4. Fusion architectures

Prior to the detailed description of the evaluated sensori-motor information fusion principles, it needs to be noted that these are implemented within two general NN architectures, namely the "Generalized Template-Matching" (GTM) and the "Generalized Spatio-Temporal" (GST) one. The GTM (Fig. 5) and the GST (Fig. 6) architectures are derived from the corresponding TM and ST ones, respectively, and their fundamental difference concerns the nature of the affordance stream modeling; GTM focuses on model-

ing correlations along the spatial dimensions, while GST relies on encoding time-evolving procedures of the performed human actions. Anatomical studies [3] on the physiological interconnections between the ventral and the dorsal streams have resulted, among others, in the following dominating hypothesis: The ventral (appearance) stream might receive up-to-date action-related information from the dorsal (affordance) stream, in order to refine the object internal representation [32].

4.4.1 Late fusion

Late fusion refers to the combination of information at the end of the processing pipeline of each stream. For the GTM architecture, this is implemented as the combination of the features at: a) the same FC layer, or b) the last CONV layer. The FC layer fusion is performed by concatenating the FC features of both streams. It was experimentally shown that fusion after the RL6 layer was advantageous, compared to concatenating at the output of the FC6 layer
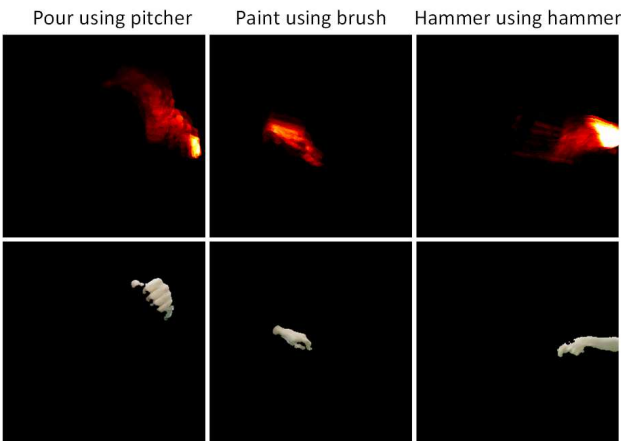

Figure 3. Examples of colorized flow magnitude field (top row) and RGB snapshots of the corresponding actions (bottom row).
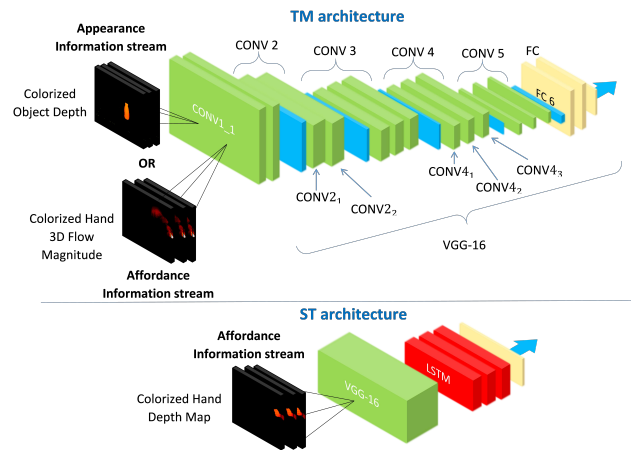

Figure 4. Single-stream models. Top: appearance CNN for object recognition, and affordance CNN (TM architecture). Bottom: affordance CNN-LSTM (ST architecture). The CNN layer notation used in this paper is depicted at the top figure.
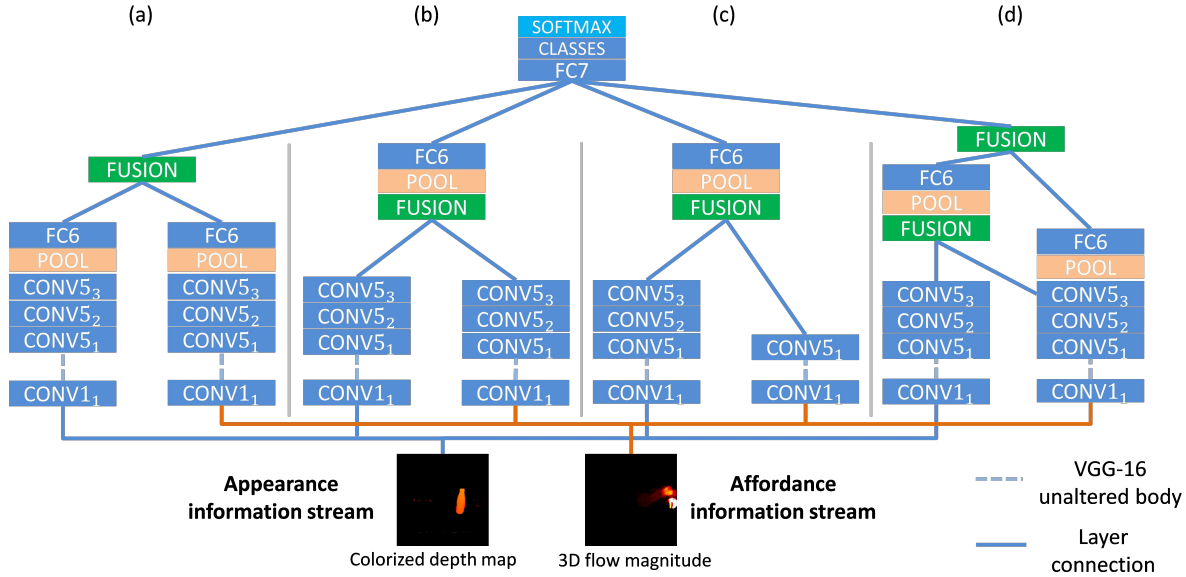
Figure 5. Detailed topology of the GTM architecture for: a) late fusion at FC layer, b) late fusion at last CONV layer, c) slow fusion, and d) multi-level slow fusion. In each case, the left stream represents the appearance and the right the affordance network, respectively.

(*i.e.* before the nonlinearity). After fusion, a single processing stream is formed (Fig. 5a). Regarding fusion at the last CONV layer, the $RL5_3$ activations of both appearance and affordance CNNs are stacked. After feature stacking, a single processing stream is again formed with four individual structural alternatives, using: i) 1 CONV (1×1 kernel size) and 1 FC, ii) 2 CONV (1×1 kernel size) and 1 FC, iii) 1 CONV (1×1 kernel size) and 2 FC (best performance, depicted in Fig. 5b), and iv) 2 CONV (1×1 kernel size) and 2 FC layers.

For the GST architecture, the late fusion scheme considers only the concatenation of the features of the last FC layers of the appearance CNN and the affordance LSTM model, as depicted in Fig. 6a. In particular, the features of the FC7 layer of the appearance CNN and the internal state vector $[h(t)]$ of the last LSTM layer of the affordance stream are concatenated at every time instant (*i.e.* at every video frame). Eventually, a single stream with 2 FC layers is formed. On the other hand, there is accumulated evidence that asynchronous communication and feedback loops occur during the sensorimotor object recognition process [5]. In this context, an asynchronous late fusion approach is also investigated for the GST architecture. Specifically, the GST late fusion scheme (Fig. 6a) is again applied. However, the information coming from the affordance stream [*i.e.* the internal state vector $h(t)$ of the last LSTM layer] is provided with a time-delay factor, denoted by $\tau > 0$, compared to the FC features of the appearance stream; in other words, the features of the affordance stream at time $t - \tau$ are combined with the appearance features at time t.

4.4.2 Slow fusion

Slow fusion for the GTM architecture corresponds to the case of combining the CONV feature maps of the appearance and affordance CNNs in an intermediate layer (*i.e.* not the last CONV layer) and subsequently forming a single processing stream, as depicted in Fig. 5c. For realizing this, two scenarios are considered, which correspond to the fusion of information from the two aforementioned CNNs at different levels of granularity: a) combining the feature maps of the appearance and the affordance CNN from the same layer level; and b) combining the feature maps of the appearance and the affordance CNN from different layer levels. The actual fusion operator is materialized by simple stacking of the two feature maps. It needs to be noted that only appearance and affordance feature maps of same dimensions are combined. For the GST architecture, the slow fusion scheme considers only the concatenation of the features of the $RL7$ layer of the appearance and the affordance CNNs models, followed by an LSTM model, as can be seen in Fig. 6b.

In order to simulate the complex information exchange routes at different levels of granularity between the two streams, a multi-level slow fusion scheme is also examined. In particular, the two streams are connected both at an intermediate/last CONV and at the FC layers. The particular NN topology that implements this multi-level slow fusion scheme for the GTM architecture is illustrated in Fig. 5d.

In the remainder of the paper, the following naming convention is used for describing the different proposed NN
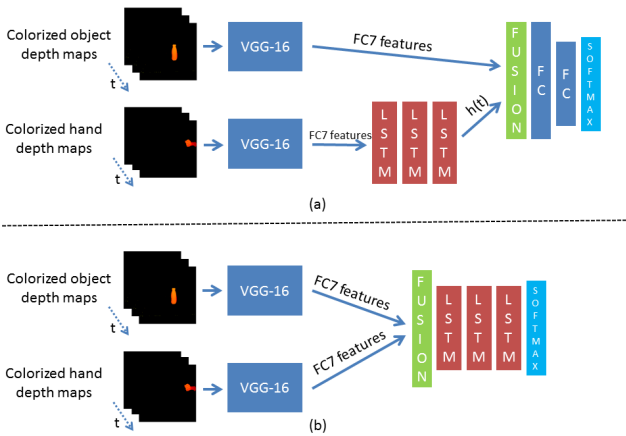
Figure 6. Detailed topology of the GST architecture for: a) late fusion and b) slow fusion.

architectures: $GAT_{FT}(param)$, where the Generalized Architecture Type, $GAT \in \{GTM, GST\}$, and the Fusion Type, $FT \in \{LS, LA, SSL, SML\} \equiv \{$Late Synchronous, Late Asynchronous, Slow Single Level, Slow Multi Level$\}$ and $param$ indicates the specific parameters for each particular fusion scheme (as detailed above). At this point, it needs to be highlighted that any further information processing performed in the affordance stream after the fusion step does not contribute to the object recognition process; hence, it is omitted from the descriptions in this work.

5. Experimental Results

The proposed NN architectures are evaluated using the introduced dataset. The involved human subjects were randomly divided into training, validation, and test sets (25%, 25%, and 50%). The utilized VGG-16 network was pre-trained on ImageNet. For all 300×300 formed frames, a 224×224 patch was randomly cropped and provided as input to the NNs. The negative log-likelihood criterion was selected during training, whereas for back-propagation, Stochastic Gradient Descent (SGD) with momentum set equal to 0.9 was used. The GTM- and GST-based NNs were trained with learning rate set to 5×10^{-3} (decreasing by a factor of 5×10^{-1} when the validation error curve plateaued) for 60 and 90 epochs, respectively. For the implementation, the *Torch*[1] framework and a Nvidia Tesla K-40 GPU were used.

5.1. Single-stream architecture evaluation

The first set of experiments concerns the evaluation of the single-stream models (Section 4.3). From the results presented in Table 3 (only overall classification accuracy is

---
[1]http://torch.ch/

given), it can be observed that object recognition using only appearance information yields satisfactory results (85.12% accuracy). Regarding affordance recognition, the TM architecture outperforms the ST one, indicating that the CNN model encodes the motion characteristics more efficiently than the composite CNN-LSTM one. For the ST model 3 LSTM layers with 4096 hidden units each were used, based on experimentation.

5.2. GTM and GST architectures evaluation

In Table 4, evaluation results from the application of different GTM-based fusion schemes (Section 4.4) are given. From the presented results, it can be seen that for the case of late fusion combination of CONV features (*i.e.* fusion at the $RL5_3$ layer) is generally advantageous, since the spatial correspondence between the appearance and the affordance stream is maintained. Concerning single-level slow fusion models, different models are evaluated. However, single-level slow fusion tends to exhibit lower recognition performance than late fusion. Building on the evaluation outcomes of the single-level slow and late fusion schemes, multi-level slow fusion architectures are also evaluated. Interestingly, the $GTM_{SML}(RL5_3^{app}, RL5_3^{aff}, RL6)$ outperforms all other GTM-based models. This is mainly due to the preservation of the spatial correspondence (initial fusion at the CONV level), coupled with the additional correlations learned by the fusion at the FC level.

Experimental results from the application of the GST-based fusion schemes (Section 4.4) are reported in Table 5. In all cases, a set of 20 uniformly selected frames were provided as input to the respective NN. Additionally, two evaluation scenarios were realized, namely when for the final object classification decision the prediction of only the last frame was considered ("last-frame") or when the predictions from all frames were averaged ("all-frames"). For the case of the synchronous late fusion, it can be seen that the averaging of the predictions from all frames is advantageous. Concerning the proposed asynchronous late fusion scheme, evaluation results for different values of the delay parameter are given. It can be observed that asynchronous fusion leads to decreased performance, compared to the synchronous case, while increasing values of the delay parameter τ lead to a drop in the recognition rate. Moreover, the slow fusion approach results to a significant decrease of the object recognition performance.

From the presented results, it can be observed that the $GTM_{SML}(RL5_3^{app}, RL5_3^{aff}, RL6)$ architecture constitutes the best performing scheme. The latter achieves an absolute increase of 4.31% in the overall recognition performance (which corresponds to an approximately 29% relative error reduction), compared to the appearance CNN model (baseline method). For providing a better insight, the object recognition confusion matrices obtained from the applica-

Method	Task	Accuracy (%)
Appearance CNN	object recognition	85.12
Affordance CNN	affordance recognition	81.92
Affordance CNN-LSTM	affordance recognition	69.27

Table 3. Single-stream results for object and affordance recognition.

GTM-based fusion architecture [after fusion]	Accuracy (%)
GTM_{LS}(FC6)	87.40
GTM_{LS}(RL5$_3$) [1 CONV, 1 FC]	87.65
GTM_{LS}(RL5$_3$) [1 CONV, 2 FC]	88.24
GTM_{LS}(RL5$_3$) [2 CONV, 1 FC]	87.64
GTM_{LS}(RL5$_3$) [2 CONV, 2 FC]	86.40
GTM_{SSL}(RL3$_3^{app}$, RL3$_3^{aff}$)	78.74
GTM_{SSL}(RL4$_3^{app}$, RL4$_3^{aff}$)	87.20
GTM_{SSL}(RL4$_3^{app}$, RL4$_1^{aff}$)	85.82
GTM_{SSL}(RL5$_3^{app}$, RL5$_1^{aff}$)	88.13
GTM_{SML}(RL5$_3^{app}$, RL5$_1^{aff}$, RL6)	88.23
GTM_{SML}(RL5$_3^{app}$, RL5$_3^{aff}$, RL6)	**89.43**

Table 4. Object recognition results using different GTM-based fusion schemes.

tion of the GTM_{SML}(RL5$_3^{app}$, RL5$_3^{aff}$, RL6) architecture and the appearance CNN are given in Fig. 7. From the presented results, it can be observed that the proposed fusion scheme boosts the performance of all supported object types. This demonstrates the discriminative power of affordance information. Additionally, it can be seen that objects whose shape cannot be efficiently captured (*e.g.* small-size ones like "Pen", "Knife", "Key", etc.) are favored by the proposed approach. Moreover, affordance information is also beneficial for objects that exhibit similar appearance (*e.g.* "Brush" with "Pen" and "Knife").

5.3. Comparison with probabilistic fusion

The GTM_{SML}(RL5$_3^{app}$, RL5$_3^{aff}$, RL6) architecture is also comparatively evaluated, apart from the appearance CNN model, with the following typical probabilistic fusion approaches of the literature: a) the product rule for fusing the appearance and the affordance CNN output probabilities, b) concatenation of appearance and affordance CNN features and usage of a SVM classifier (RBF kernel) [4, 15], and c) concatenation of appearance and affordance CNN features and usage of a naive Bayes classifier [13]. From the results in Table 6, it can be seen that the literature approaches for fusing the appearance and affordance information streams fail to introduce an increase in the object recognition performance in the introduced challenging dataset; the aforementioned methods were evaluated under significantly simpler experimental settings. On the contrary, the proposed approach exhibits a significant performance increase over the baseline method (appearance CNN).

GST-based fusion architecture [after fusion]	Accuracy (%)
GST_{LS}(last-frame)	86.28
GST_{LS}(all-frames) [1 CONV, 2 FC]	**86.50**
GST_{LA}(all-frames, $\tau = 2$)	86.42
GST_{LA}(all-frames, $\tau = 4$) [1 CONV, 2 FC]	86.17
GST_{LA}(all-frames, $\tau = 6$) [1 CONV, 2 FC]	85.28
GST_{SSL}(all-frames) [1 CONV, 2 FC]	79.65

Table 5. Object recognition results using different GST-based fusion schemes.

Fusion architecture	Fusion Layer	Accuracy (%)
Appearance CNN	no fusion	85.12
Product Rule	Softmax	73.45
SVM [15, 4]	RL7	83.43
Bayes [13]	RL7	75.86
GTM_{SML}	RL5$_3^{app}$, RL5$_3^{aff}$, RL6	**89.43**

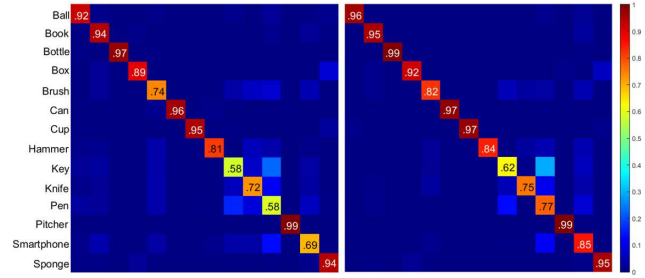Table 6. Comparative evaluation of GTM_{SML}(RL5$_3^{app}$, RL5$_3^{aff}$, RL6) architecture.

Figure 7. Object recognition confusion matrices of appearance CNN (left) and GTM_{SML}(RL5$_3^{app}$, RL5$_3^{aff}$, RL6) architecture (right).

6. Conclusions

In this paper, the problem of sensorimotor 3D object recognition following the deep learning paradigm was investigated. A large public 3D object recognition dataset was also introduced, including multiple object types and a significant number of complex affordances, for boosting the research activities in the field. Two generalized neuro-biologically and neuro-physiologically grounded neural network architectures, implementing multiple fusion schemes for sensorimotor object recognition were presented and evaluated. The proposed sensorimotor multi-level slow fusion approach was experimentally shown to outperform similar probabilistic fusion methods of the literature. Future work will investigate the use of NN auto-encoders for modeling the human-object interactions in more details and the application of the proposed methodology to more realistic, "in-the-wild" object recognition data.

References

[1] A. Aldoma, F. Tombari, and M. Vincze. Supervised learning of hidden and non-hidden 0-order affordances and detection in real scenes. *ICRA*, pp. 1732–1739, 2012.

[2] A. Andreopoulos and J. K. Tsotsos. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, 117(8):827–891, 2013.

[3] M. L. Brandi, A. Wohlschläger, C. Sorg, and J. Hermsdörfer. The neural correlates of planning and executing actual tool use. *The Journal of Neuroscience*, 34(39):13183–13194, 2014.

[4] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo. Using object affordances to improve object recognition. *IEEE Transactions on Autonomous Mental Development*, 3(3):207–215, 2011.

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

[6] J. Deng, W. Dong, R. Socher, L. .-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. *CVPR*, pp. 248–255, 2009.

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*, pp. 2625–2634, 2015.

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. *ICML*, pp. 647–655, 2014.

[9] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. *IROS*, pp. 681–687, 2015.

[10] J. J. Gibson. The theory of affordances. In R. Shaw and J. Bransford (eds.). *Perceiving, Acting, and Knowing: Toward an Ecological Psychology.*, pp. 67–82. Lawrence Erlbaum, Hillsdale NJ, 1977.

[11] T. Hermans, J. M. Rehg, and A. Bobick. Affordance prediction via learned object attributes. *ICRA Workshops*, 2011.

[12] D. Herrera, J. Kannala, and J. Heikkilä. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2058–2064, 2012.

[13] V. Högman, M. Björkman, A. Maki, and D. Kragic. A sensorimotor learning framework for object categorization. *IEEE Transactions on Cognitive and Developmental Systems*, 8(1):15–25, 2015.

[14] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers. A primal-dual framework for real-time dense RGB-D scene flow. *ICRA*, pp. 98–104, 2015.

[15] H. Kjellström, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011.

[16] T. Kluth, D. Nakath, T. Reineking, C. Zetzsche, and K. Schill. Affordance-based object recognition using interactions obtained from a utility maximization principle. *ECCV Workshops*, pp. 406–412, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *NIPS*, pp. 1097–1105, 2012.

[18] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. *CVPR*, pp. 3367–3375, 2015.

[19] Y. Liu, H. Zha, and H. Qin. Shape topics: A compact representation and new algorithms for 3D partial shape retrieval. *CVPR*, pp. 2025–2032, 2006.

[20] N. Lyubova, S. Ivaldi, and D. Filliat. From passive to interactive object learning and recognition through self-identification on a humanoid robot. *Autonomous Robots*, 40(1):33–57, 2016.

[21] Microsoft. Meet Kinect for Windows. `https://developer.microsoft.com/en-us/windows/kinect`. [Online].

[22] M. Minsky. Society of mind: A response to four reviews. *Artificial Intelligence*, 48(3):371–396, 1991.

[23] B. Moldovan, M. van Otterlo, P. Moreno, J. Santos-Victor, and L. D. Raedt. Statistical relational learning of object affordances for robotic manipulation. *ICILP*, 2012.

[24] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

[25] J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–1031, 2001.

[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CVPR Workshops*, pp. 512–519, 2014.

[27] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *Computer Vision and Image Understanding*, 62(2):164–176, 1995.

[28] J. Schmidhuber, D. Wierstra, M. Gagliolo, and F. Gomez. Training recurrent networks by Evolino. *Neural Computation*, 19(3):757–779, 2007.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[30] M. Sutton, L. Stark, and K. Bowyer. Function from visual analysis and physical interaction: a methodology for recognition of generic classes of objects. *Image and Vision Computing*, 16(11):745–763, 1998.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, pp. 1–9, 2015.

[32] V. van Polanen and M. Davare. Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, 79:186–191, 2015.

[33] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. *GraphiCon*, pp. 85–92, 2003.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *NIPS*, pp. 3320–3328, 2014.

[35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *ECCV*, pp. 85–92, 2014.