ATTENTION-ENHANCED 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

Sensorimotor learning, namely the process of understanding the physical world by combining visual and motor information, has been recently investigated, achieving promising results for the task of 2D/3D object recognition. Following the recent trend in computer vision, powerful deep neural networks (NNs) have been used to model the "sensory" and "motor" information, namely the object appearance and affordance. However, the existing implementations cannot efficiently address the spatio-temporal nature of the human-object interaction. Inspired by recent work on attention-based learning, this paper introduces an attention-enhanced NN-based model that learns to selectively focus on parts of the physical interaction where the object appearance is corrupted by occlusions and deformations. The model's attention mechanism relies on the confidence of classifying an object based solely on its appearance. Three metrics are used to measure the latter, namely the prediction entropy, the average N-best likelihood difference, and the N-best likelihood dispersion. Evaluation of the attention-enhanced model on the SOR3D dataset reports 33% and 26% relative improvement over the appearance-only and the spatio-temporal fusion baseline models, respectively.

***Index Terms*—** Sensorimotor object recognition, attention mechanism, stream fusion, deep neural networks

1. INTRODUCTION

During the last decades significant research effort has focused on the field of 2D/3D object recognition. Though this task is crucial in a variety of fields, such as automation, security, and robotics, it remains an open challenge in real-world scenarios. In particular, the existing illumination variation, occlusions, and deformations are problems that cannot be addressed by the sole use of static object appearance features, such as shape, texture, and color [1, 2, 3].

On the other hand, recent studies in computer vision have adopted a more human-inspired approach, the so-called "sensorimotor learning" [4, 5, 6, 7]. Cognitive neuroscience argues that human object perception is based on fusing the object appearance with its affordance [8, 9], namely its functionalities or more specifically the set of actions that a human can perform with the object. Thus, recognizing a mug not only based on its appearance attributes such as the cylindrical shape and the hole on the top, but also based on its "graspable" and "pourable" affordances appears to be more reasonable [10]. In fact, humans in the early stages of their life do not rely on semantic labels for object understanding, but perform active exploration and physical interactions with the real-world objects or learn from observing others interacting with them [11, 12].

Aided by advances in learning feature representations with convolutional (CNNs) and recurrent neural networks (RNNs) [13, 14], sensorimotor object recognition has significantly evolved over the past decade. In fact, using NN-based models, the human-inspired two-stream information processing approach [15, 16] can be modeled more efficiently. In particular, in our recent work [6], we investigated various two-stream fusion architectures to improve object recognition and achieved promising results. Besides sensorimotor learning, multi-stream fusion has been widely investigated in a variety of tasks, such as action recognition [17, 18], audio-visual speech recognition [19, 20], and voice activity detection [21]. Additionally, spatial and temporal attention mechanisms have been integrated to multi-stream architectures, aiming at learning the most discriminative features of each input modality, leading to significant performance improvement over plain fusion. For example, in [22] the proposed temporal attention mechanism fuses visual with motion features by attending to the most informative frames of the video in order to generate a description, while in [23] cross-link attentional layers are developed between the temporal and spatial streams in order to enhance the human foreground area for video-based action recognition.

In this paper, we investigate the integration of an attention mechanism within the sensorimotor object recognition framework. The spatio-temporal late fusion model of [6] is adopted, and the attention mechanism is integrated in the object appearance stream, prior to fusion. We argue that the object appearance features are the most informative for object recognition, while the affordance information should be attended when the appearance one is failing. The prediction entropy, the average N-best likelihood difference, and the N-best likelihood dispersion metrics are evaluated for the appearance classifier's confidence measurement. The proposed attention-enhanced model is experimentally shown to perform better than both appearance-only and the spatio-temporal baselines.

The remainder of the paper is organized as follows: Section 2 presents the sensorimotor object recognition task addressed in this paper and reviews our baseline models introduced in [6], Section 3 details the proposed attention mechanism and its integration within the recognition framework, while Section 4 presents our experiments. Finally, Section 5 summarizes the paper.

2. TASK DESCRIPTION AND BASELINE APPROACH

In this section, the object recognition problem based on hand-object interaction input is discussed. Notice that all models presented in the study utilize RGB-D sequences of the SOR3D dataset [6] captured by Kinect sensors (more details are provided in Section 4). A

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

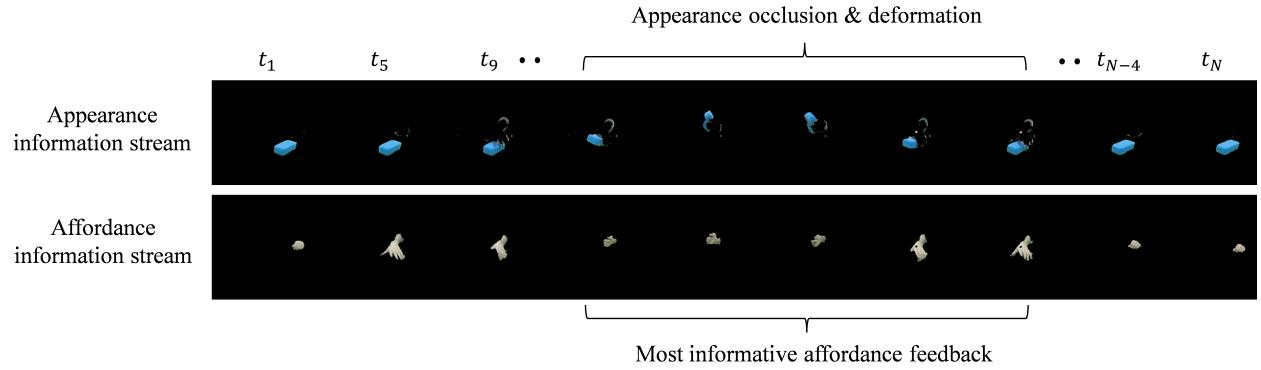

Fig. 1: Example video session "squeeze sponge" from SOR3D [6], sampled every 4 frames. The object appearance and the corresponding affordance information are presented as RGB frames.

sub-sampled sequence of the available data is depicted in Fig. 1. There are two available streams of information: a) the appearance stream, which encodes the appearance features of the object, and b) the affordance stream, which encodes the motion features during an interaction.

2.1. Single-stream Modeling

As single-stream baseline approach for the object recognition task, a CNN is used to train a classifier that recognizes objects based solely on their appearance features. Additionally, a baseline model for affordance representation learning is defined, which consists of a CNN for feature extraction, followed by a Long-Short Term Memory (LSTM) [24]. This model exploits the capabilities of the CNN to model the spatial correlations of the input, while subsequently takes advantage of the LSTM for efficiently encoding the temporal dynamics of the interaction. Further details for the CNN structure and the information flow for each model are presented in Section 4.

2.2. Two-stream Fusion

As already noted, fusing multiple streams of information leads to more discriminative feature representations, therefore to more confident predictions. In this context and given the demonstrated effectiveness of the sensorimotor approach, a model that fuses the appearance with the affordance information is investigated. There are several approaches to fuse the aforementioned streams, however in this work the best performing spatio-temporal model from [6] is adopted. In detail, the features of the last fully connected (FC) layer $x_t^{1\times F}$ of the appearance CNN and the hidden state vector $h_t^{1\times M}$ (M LSTM hidden units) of the last layer of the affordance CNN-LSTM are concatenated for each frame $t = 1, \ldots, T$, and are processed by a Multilayer Perceptron (MLP). The MLP consists of 2 FC layers and is followed by a Softmax layer. The adopted late fusion model is depicted in Fig. 2 (excluding the green box) and serves as the spatio-temporal (ST) baseline.

Though the latter model performs better than the appearance-only CNN in the object recognition task, we argue that the affordance information is truly informative when the actual interaction takes place. As shown in Fig. 1, the object can be easily identified at the first and the last frames of the video based solely on its appearance features. On the contrary, at the middle of the video, it is hard to predict the object label with confidence, mainly due to occlusions and deformation. Thus, the affordance features are mostly needed as additional information during the interaction, namely when the prediction based on the appearance features is not so confident. To support this hypothesis, an attention mechanism that operates before the stream fusion is proposed.

3. ATTENTION MECHANISM

The proposed attention mechanism is based on the appearance stream confidence. As depicted in Fig. 2 (green box), a Softmax layer is added after the last FC layer of the appearance CNN, which predicts the label of the object for each frame. The new layer is followed by a module that measures the appearance-based classifier confidence for the entire frame sequence. The output of the latter is used to selectively attend to the affordance features extracted by the affordance CNN-LSTM stream, prior to the fusion MLP.

In order to measure the appearance classifier confidence, we investigate three different metrics. Let $c_{t,n}, n = 1, \ldots, N$ be the ranked $N-$best object class predictions of the appearance CNN classifier, $\mathcal{C}$ the number of the object classes, and $p_{t,n} = Pr(c_{t,n}|x_t)$ the probability distribution after the Softmax given the appearance feature vector x_t at frame t. As the first metric, the entropy $\mathcal{I}_{t,E}$ is computed for the probability distribution as:

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

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:

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

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

$$\mathcal{I}_{t,D} = \frac{2}{N(N-1)}\sum_{n=1}^{N-1}\sum_{m=n+1}^{N}(\log\left(p_{t,n}\right) - \log\left(p_{t,m}\right)), \tag{3}$$

where $N \geq 2$. Similarly to (2), larger $\mathcal{I}_{t,D}$ values indicate high classification confidence. It must be noted that the presented metrics

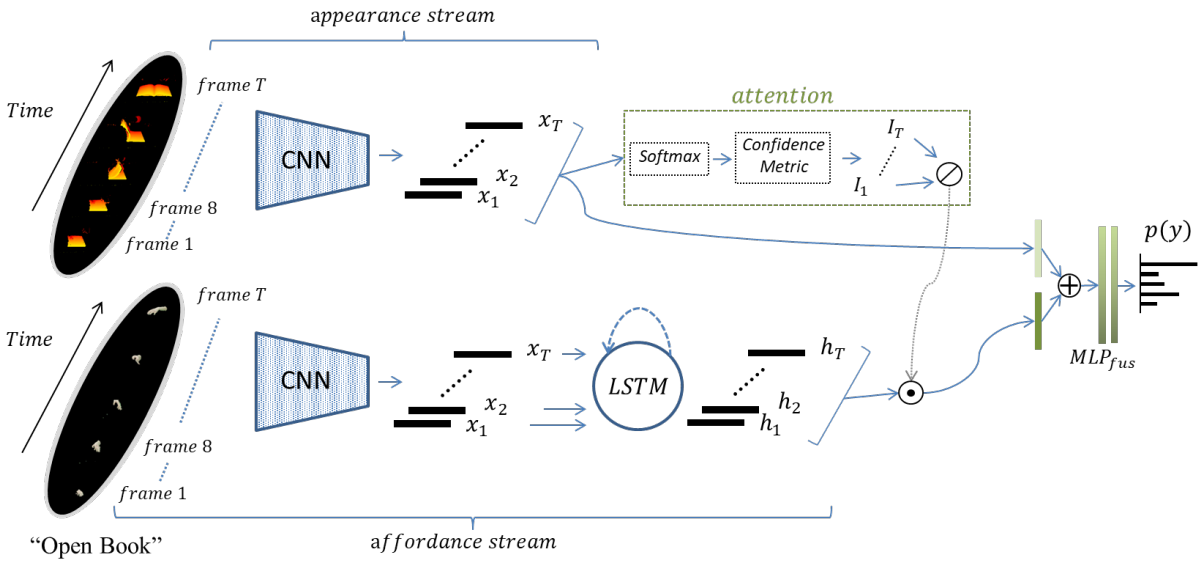

Fig. 2: Detailed architecture of the proposed spatio-temporal late fusion model. The green box includes the attention mechanism modules attached to the final FC layer of the appearance CNN (top), which selectively attends to the affordance CNN-LSTM output (bottom). The feature fusion MLP follows (right-most), while $\oslash$, $\odot$, and $\oplus$ represent normalization, frame-level multiplication, and concatenation.

have been also used in the context of audio-visual speech recognition [25, 26]. Following the appearance classifier confidence measurement, the $\mathcal{I}_t$ values of all frames are normalized to $[0, 1]$ by:

$$w_t = \frac{\mathcal{I}_t - \mathcal{I}_{min}}{\mathcal{I}_{max} - \mathcal{I}_{min}}, \tag{4}$$

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

$$\hat{H} = \begin{cases} w \odot H & \text{if (1)} \\ (1 - w) \odot H & \text{if (2) or (3)} \end{cases}$$

where $\odot$ indicates the frame-level multiplication of confidence values with the LSTM output matrix $H^{T \times M}$. Notice that by multiplying w_t with the corresponding h_t, the mechanism alters the impact of the affordance information on the final prediction, since $\hat{H}^{T \times M}$ is fused with the appearance features as:

$$\hat{p}_t = \text{softmax}(\phi(\text{concat}(x_t, \hat{h}_t))), \tag{5}$$

where $x_t \in X^{T \times F}$ denotes the appearance feature vector, ϕ is the fusion MLP followed by a Softmax function, and $\hat{p}_t$ is the probability distribution of the attention-based ST (AST) model (Fig. 2) for the $t-$th frame.

Regarding the final prediction, two approaches are investigated. Both aggregate the frame-level prediction of the AST model to yield a video-level decision for the object label. Given a series of frame-level predictions $\hat{p}_{1,c}, \ldots, \hat{p}_{t,c}, \ldots, \hat{p}_{T,c}$ from (5), the video-level classification decision y is given either by:

$$y_{avg} = \arg\max_c \frac{1}{T} \sum_{t=1}^{T} \hat{p}_{t,c}, \tag{6}$$

as the averaging (AVG) approach, or by:

$$y_w = \arg\max_c \frac{1}{T} \sum_{t=1}^{T} t\, \hat{p}_{t,c}, \tag{7}$$

as the weighting (W) approach, respectively. Clearly, the latter forces the model to focus more on the frame-level predictions over the last frames of the video, while the former treats all frame-level predictions equally.

4. DATASET AND EXPERIMENTS

As noted in Section 2, the presented models are trained and evaluated using the SOR3D dataset [6]. SOR3D is the broadest and most challenging public dataset in the sensorimotor object recognition literature. It consists of 20,830 RGB-D instances of various length, that include 14 object types and 13 affordance ones. The instances are provided as sequences of RGB and depthmap frames with 300×300 pixel resolution, depicting the segmented object and the corresponding hand, as shown in Fig. 1. We use the same training, validation and test set as in [6].

For the presented experiments, the colorized depth maps of the object appearance (CDM-AP) were used as input to the appearance CNN, while for the affordance CNN-LSTM both colorized depth maps of the hand (CDM-AF) and the corresponding colorized magnitude of the computed 3D optical flow [27] (3DFM-AF) were used, respectively. Notice that the depth map colorization enables the fine-tuning of pre-trained deep neural networks, and it is considered a common practice [28].

Regarding the CNNs included in the models, the widely-known VGG-16 [29] was used. For the ST and the AST models, the appearance features were extracted from the last FC layer of the appearance VGG, while the spatial affordance features that were propagated to the LSTM were extracted from the last FC layer of the affordance VGG. All VGGs were pre-trained on ImageNet [30], and the utilized LSTM consisted of 4096 hidden units.

For the experiments, each input frame was randomly cropped to a 224×224 resolution. The negative log-likelihood criterion was selected during training, whereas for back-propagation, Stochastic Gradient Descent (SGD) with momentum was used. Regarding the

Model	Input Stream(s)	Test Acc. (%)
Appearance-only [6]	CDM-AP	85.12
ST_{AVG} [6]	CDM-AP, CDM-AF	86.50
ST_{AVG}	CDM-AP, 3DFM-AF	86.38
ST_W	CDM-AP, CDM-AF	86.87
ST_W	CDM-AP, 3DFM-AF	86.72
AST_{AVG}	CDM-AP, CDM-AF	89.27
AST_{AVG}	CDM-AP, 3DFM-AF	89.41
AST_W	CDM-AP, CDM-AF	89.84
AST_W	CDM-AP, 3DFM-AF	**90.02**

Table 1: Comparative evaluation of the appearance-only, ST, and AST ($N-$best log-likelihood dispersion metric, $N = 3$) models. The second column presents the input streams used for each experiment. The first two lines correspond to the baseline models of this study (from [6]).

Confidence Metric	Test Acc. (%)
Entropy	88.75
$N-$best difference	89.12
$N-$best dispersion	89.27

Table 2: Comparative evaluation of the AST_{AVG} model using: a) the entropy, b) the average $N-$best log-likelihood difference ($N = 3$), and c) the $N-$best log-likelihood dispersion ($N = 3$). The model is evaluated using CDM-AP and CDM-AF as input streams.

baseline approach, the appearance VGG was fine-tuned with a learning rate (LR) set to 5×10^3 for 30 epochs. For the ST and the AST approaches, the latter is used for feature extraction and confidence measurement. Subsequently, the pre-trained affordance VGG, the LSTM, and the fusion MLP were jointly fine-tuned on SOR3D, with LR set to 1×10^2 for 90 epochs. The LR was decreased by a factor of 2×10^2 when the validation accuracy curve plateaued. The models were implemented using the Torch7[1] framework and a Nvidia Titan X GPU.

Table 1 reports the performance of the baseline appearance-only CNN and the baseline ST model using CDM-AP and CDM-AF inputs as presented in [6]. In order to compare the AST performance with the aforementioned baselines, the best metric for the attention mechanism must be determined. The performance of the AST model for the metrics presented in Section 3 is given in Table 2. For a fair comparison, CDM-AP and CDM-AF are used as inputs to the appearance and affordance streams, while the final prediction is based on frame-level prediction averaging. It can be observed that the attention mechanism based on the $N-$best log-likelihood dispersion metric ($N = 3$) yields the best overall accuracy (89.27%). In fact, the AST model outperforms both the appearance-only CNN and the ST model using any of the presented confidence metrics. Thus, in the remaining experiments, the AST model with the $N-$best log-likelihood dispersion metric is adopted.

Table 1 also reports the overall accuracy achieved by the ST and AST models for different input modalities and frame-level prediction aggregation. From the presented results, it can be observed that the weighting frame-level prediction approach leads to superior performance over the averaging one, for all ST and AST evaluated models. Additionally, the AST models outperform the ST ones for any input combination, supporting our hypothesis that the affordance features are most informative during the hand-object interaction, hence when the object appearance is cluttered. Notice that even though 3D optical flow is more informative than depth when representing motion, the ST model fails to exploit its capabilities. On the contrary, the AST model that uses 3DFM-AF instead of CDM-AF as affordance input, achieves marginally better overall accuracy. One plausible reason is that the 3DFM 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 AST_W model with CDM-AP and 3DFM inputs constitutes the best performing scheme, achieving 90.02% accuracy. In fact, it outperforms the appearance-only and the ST baseline models by 4.9% and 3.52% absolute, while also achieving an absolute increase of 3.15% in the overall accuracy, compared to the best performing ST one. This corresponds to 33%, 26%, and 25% relative classification error reduction, respectively. The object recognition confusion matrices of the best performing ST and AST models are depicted in Fig. 3. It can be seen that the proposed attention mechanism boosts the performance of all evaluated object types, favoring the most deformable ones (*e.g.* "box", "brush", and "sponge").

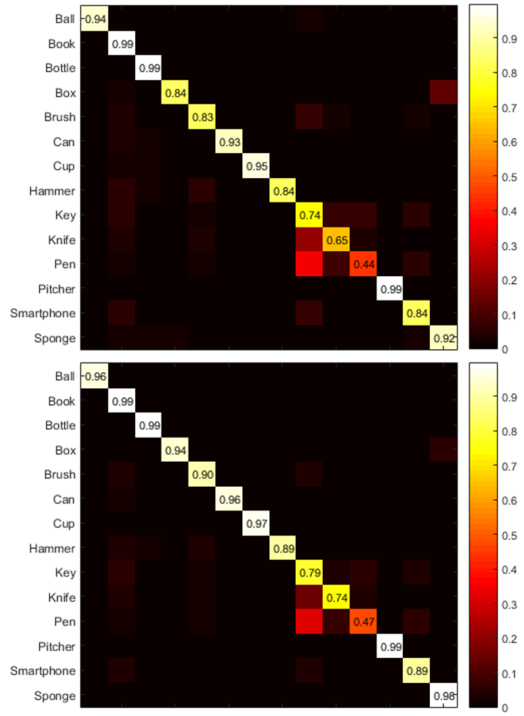

Fig. 3: Object recognition confusion matrices of the best performing: a) ST_W (top), and b) AST_W (bottom) model, respectively.

5. CONCLUSION

In this paper, the problem of attention-enhanced sensorimotor object recognition was investigated. An attention mechanism that selectively attends to the affordance information, when the appearance features are not discriminative enough, was introduced. The latter was integrated to the spatio-temporal fusion model presented in [6] and was experimentally shown to outperform both the no-attention and the appearance-only models by a significant margin. Future work will investigate the integration of intra-stream attention mechanisms to force better single-stream representation learning, prior to the fusion module.

[1] http://torch.ch/

6. REFERENCES

[1] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3367–3375.

[2] 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," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2014, pp. 647–655.

[3] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5648–5656.

[4] T. Kluth, D. Nakath, T. Reineking, C. Zetzsche, and K. Schill, "Affordance-based object recognition using interactions obtained from a utility maximization principle," in *Proc. European Conference on Computer Vision Workshops (ECCVW)*, 2014, pp. 406–412.

[5] V. Högman, M. Björkman, A. Maki, and D. Kragic, "A sensorimotor learning framework for object categorization," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 1, pp. 15–25, 2015.

[6] S. Thermos, G. Th. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 49–57.

[7] J.-B. Alayrac, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Joint discovery of object states and manipulation actions," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2146–2155.

[8] 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.

[9] E. Rivlin, S. J. Dickinson, and A. Rosenfeld, "Recognition by functional parts," *Computer Vision and Image Understanding*, vol. 62, no. 2, pp. 164–176, 1995.

[10] Y. Zhu, A. Fathi, and L. Fei-Fei, "Reasoning about object affordances in a knowledge base representation," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 408–424.

[11] M.-L. Brandi, A. Wohlschläger, C. Sorg, and J. Hermsdörfer, "The neural correlates of planning and executing actual tool use," *The Journal of Neuroscience*, vol. 34, no. 39, pp. 13183–13194, 2014.

[12] D. L. K. Yamins and J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, pp. 356–365, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

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

[16] L. L. Cloutman, "Interaction between dorsal and ventral processing streams: Where, when and how?," *Brain and Language*, vol. 127, no. 2, pp. 251–263, 2013.

[17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1933–1941.

[18] G. Papadopoulos and P. Daras, "Human action recognition using 3D reconstruction data," *IEEE Transactions on Circuits and Systems for Video Technology (To Appear)*, 2018.

[19] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[20] A. Garg, G. Potamianos, C. Neti, and T.S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2003, vol. I, pp. 24–27.

[21] S. Thermos and G. Potamianos, "Audio-visual speech activity detection in a two-speaker scenario incorporating depth information from a profile or frontal view," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 579–584.

[22] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 4203–4212.

[23] A. Tran and L.-F. Cheong, "Two-stream flow-guided convolutional attention networks for action recognition," in *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, 2017, pp. 3110–3119.

[24] J. Schmidhuber, D. Wierstra, M. Gagliolo, and F. Gomez, "Training recurrent networks by Evolino," *Neural Computation*, vol. 19, no. 3, pp. 757–779, 2007.

[25] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000, vol. 3, pp. 746–749.

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

[27] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers, "A primal-dual framework for real-time dense RGB-D scene flow," in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 98–104.

[28] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 681–687.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.