A Humanoid Robot Learning Audiovisual Classification By Active Exploration

Glareh Mir, Matthias Kerzel, Erik Strahl, Stefan Wermter
Knowledge Technology Group, Department of Informatics, University of Hamburg, Germany
{8gmir,kerzel,strahl,wermter}@informatik.uni-hamburg.de

2021 IEEE International Conference on Development and Learning (ICDL) | 978-1-7281-6242-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDL49984.2021.9515598

***Abstract*—We present a novel neurorobotic setup and dataset for active object exploration and audiovisual classification based on their material properties. In the robotic setup, a humanoid drops an item on a sloped surface and records the video image frames and raw audio of the collision of the surface and object. The novel dataset includes 32800 images and 1600 s of audio recording from 800 samples for 16 objects and will be made publicly available. We propose a novel neural architecture for the classification of the objects. A detailed analysis of results shows that different materials are easier classified either in the audio or the visual modality. As a main contribution, we can show that combining modalities can achieve an even higher classification accuracy of 90%.**

***Index Terms*—Crossmodal object recognition, Humanoid robots, Supervised learning, Robot learning, Multi-layer neural network**

I. INTRODUCTION

Infants interact with and identify objects in an interesting manner. At different ages, infants use combinations of different object properties to identify them [1]. There are many lessons from such interactions that can be carried over to robotic learning of objects and object properties. We present a novel neurorobotic setup and model for active audiovisual exploration and recognition of objects based on their material properties[1].

To learn about the material properties of an object, robots can employ different active exploration methods. Using tactile sensors, they can slide over an object's surface [2]; using proprioception, they can compress the object [3]. They can shake [4] the object or strike it [5] to elicit audio information. A robot can also slide an object over a surface or shove it [6]. We propose a novel setup in which a robot drops an object onto a sloped surface. The sound the object makes and the trajectory of movement reveal a great deal of information. Some objects create a loud noise, some bounce multiple times, some tumble, and others slide downward, making little sounds.

Our approach is inspired by humans' ability to exploit multiple modalities for robust and fast recognition, specifically audiovisual object recognition [7] [8]. We propose a supervised neural network architecture for classifying objects based on their audio, their motion (perceived visually), and their combined audiovisual properties. We show that combining the audio and visual modalities outperforms monomodal object classification in the proposed scenario.

Our main contributions are: *1)* A novel dataset[2] of 16 different (material, shape, weight) objects being individually dropped by the humanoid NICO (Neuro-Inspired COmpanion) with audiovisual recordings of the event from the humanoid's perspective. Each object is actively explored 50 times, resulting in a dataset of 800 audiovisual samples. *2)* Novel supervised neural models for learning to classify objects based on the recorded data. *3)* An in-depth analysis of monomodal versus crossmodal analysis abilities of the approach and an evaluation of the use of pretrained monomodal models.

II. RELATED WORK

Studies that have looked into human object classification capabilities suggest that humans benefit from effects of *Super-additivity in multisensory integration* by which multisensory

[1]Video available at https://www2.informatik.uni-hamburg.de/WTM/videos/Neurorobotic_Exploration.mp4

[2]The dataset will be available at https://www.inf.uni-hamburg.de/en/inst/ab/wtm/research/corpora.html

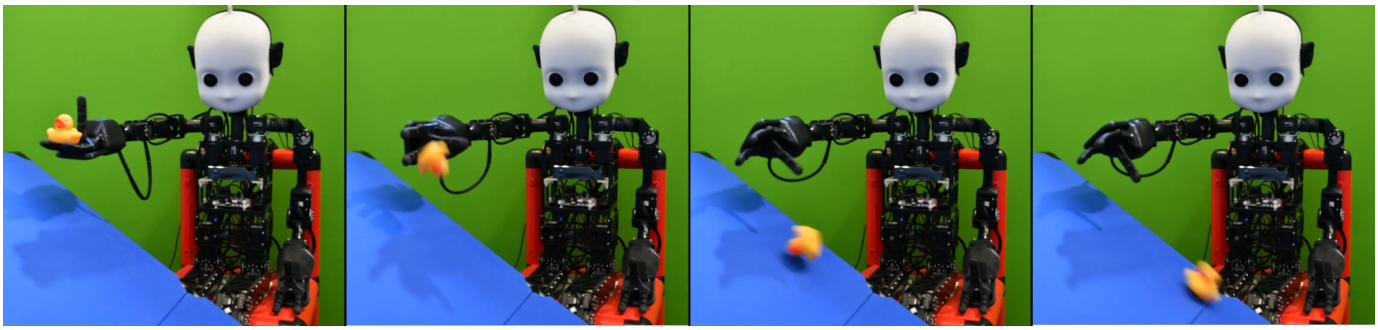

Fig. 1. "NICO", the humanoid, is collecting audiovisual samples by dropping object onto a sloped surface.

response exceeds the sum of modality-specific response [9]. This enhancement is natural for humans and helps human object recognition capabilities, specifically audiovisual object recognition [7]. It has been shown that humans could recognize animals faster and more accurate based on matched image and sound cues compared to unisensory stimuli [8]. There are many other studies supporting the notion of multisensory enhancement of response in humans compared to a unisensory one [10] [11] [12] [13].

Several studies have adapted this principle to neurorobotic setups. Sterling et al. [5] strike different objects to elicit sound and use a crossmodal neural network to estimate the object's shape and medium. Their neural network architecture has informed this study. Gao et al. [14] take a unique approach to learn audiovisual object models and then use visual context for audio source separation. Several active object exploration datasets have been realized with the NICO platform: The EMIL project [6] is used in studies on "Language Grounding" [15]. Active haptic [16] and auditory [17] exploration were successfully realized. Especially the aforementioned studies informed the design of the presented neurorobotic study.

III. APPROACH

In this section, we will first discuss the object set, the robotic scenario and sample collection method, and finally, we will present different neural learning approaches.

A. Object set

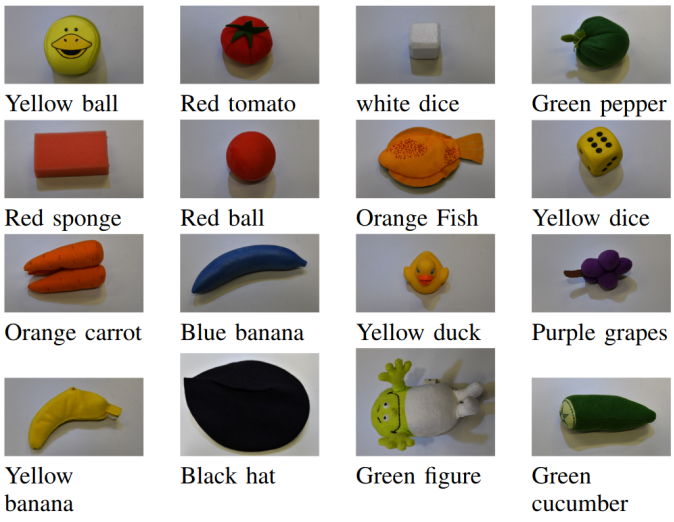

Fig. 2. Set of 16 objects for active audiovisual exploration. Objects are made from different materials ranging from hard plastic, soft plastic, rubber, styrofoam, sponges, and PVC foam to fabric and plush toys with various fillings.

Figure 2 shows all objects used to record the dataset. Most of the objects are toys made of various materials ranging from hard plastic, soft plastic, rubber, styrofoam, sponges, and PVC foam to fabric and plush toys with various fillings. We use the difference-of-image technique to extract only the motion and shape characteristics of the object during the active exploration and abstract away the color and texture properties of the objects (which would allow for easy classification). During the dataset recording, a wide range of audio and motion characteristics could be observed; some object landed with a loud noise, other bounced multiple times or tumbled while still, others slid down the slope almost inaudible.

B. NICO Humanoid and Experimental Setup

For the experiments we used NICO, the "Neuro-Inspired COmpanion" [18], seen in Figure 1. NICO is a humanoid with a wide range of capabilities, including sensing (vision, audio, tactile), acting (motor joints following human anatomy), and facial expression. For our purpose, we use its audio and visual recording capabilities, as well as grasping and arm movement. Experiments are realized using the NICO API[3] and *Robot Operating System (ROS)* for simultaneous action execution and synchronized data collection.

NICO uses two high-resolution RGB See3CAM_CU135 cameras located in the eye sockets, with overlapping fields of view for video recording and two Soundman OKM II binaural microphones located in 3D-printed ears. NICO has 30 degrees of freedom, distributed over two arms, two legs, and the neck. Our NICO has three-fingered SR-DH4D hands[4], with the two index fingers being controlled jointly as a single degree of freedom, as is the opposable thumb. As we aimed for audio data collection, the robot ego-noise was kept at a minimum. Therefore during recording NICO only used one degree of freedom for the yaw of the right wrist. However, during the operation overall, NICO also used multiple degrees of freedom from the shoulder, elbow, and head.

NICO's humanoid build makes it possible to place the robot in a child-sized environment to realize the experimental setup. The surface of impact is a slope made from a Quadro Construction Kit for child-sized furniture and playground elements. The slope is made of 3 planes of size 40×40 cm, stabilizing legs, and one plane of the same size for stopping object at the end of the slope. The objects are dropped from 25 cm above the slope. NICO is seated on a child-sized chair. The setup is depicted in Fig. 1.

C. Collection Method

To collect samples, the following scenario was designed, Figure 1 shows the NICO humanoid in action:

1) NICO bends its head forward at a 30 degrees angle
2) NICO puts its arm in the "ready state" to hold the object (hand is held straight, reaching forward, palm upwards)
3) The experimenter puts the object in NICO's hand
4) NICO waits for the start signal (keypress)
5) NICO starts recording stereo audiovisual information
6) NICO drops the object on the surface
7) NICO continues the stream for 2 seconds
8) NICO stops recording and goes back to the "ready state"

Using NICO's capabilities, the sample collection was orchestrated using a state machine through ROS. The recordings

[3]https://github.com/knowledgetechnologyuhh/NICO-software
[4]https://www.seedrobotics.com/

consist of 41 frames of stream images and 2 seconds of audio recording in the *.wav* format for any of the 800 samples, resulting in 32800 images and 1600 s of recordings. Relevant sensory information in both audio and video lasts about half a second; therefore, audio and visual tensors are automatically cut to focus on the window of interest during pre-processing.

D. Preprocessing of Audio and Visual Data

The recorded audio and visual data is pre-processed in the following way:

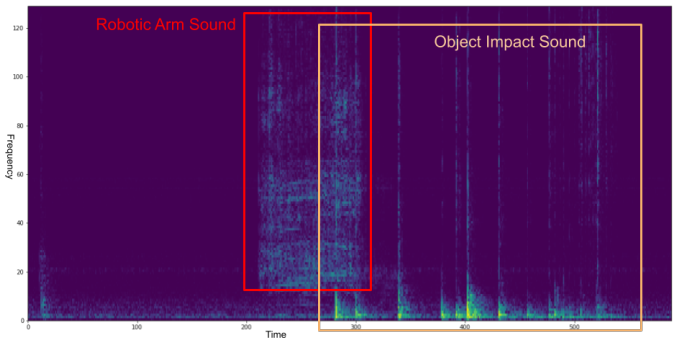

Fig. 3. The spectrogram displaying clear patterns.

1) Extraction of Audio Features: The raw audio sample is converted to log-scaled spectrograms of the size $(130, 600)$ pixels. As shown in Figure 3, ego-noise and object impact sound become clearly distinguishable. The relevant part of the spectrogram, or the region of interest, is then extracted.

2) Extraction of Visual-Motion Features: We extract the motion information from the visual stream by creating *difference of images (DOI)* frames, compare Tsironi et al. [19]. This has two advantages: First, we lose any information based on the objects' color and texture. Though this information would help classify the objects, it would not be a classification based on the material and resulting motion properties of the object, which is the aim of this study. By omitting color and texture information, the approach becomes invariant towards using the same object, e.g., in different colors: a block of Styrofoam will tumble in the same characteristic way, no matter in which color it is painted. Second, we suppress input that is not related to the moving object, enhancing the visual learning method's robustness against changes in the visual background and lighting conditions.

Using DOI will focus the learning approach on the motion of the object, e.g., bouncing, sliding, or tumbling. From the stream of input images, we create a sequence of images in which each image represents the difference between two consecutive image frames. Figure 4 shows an excerpt from a DOI sequence.

E. Deep Learning Networks

We present a novel neural network architecture for supervised learning of unimodal (audio or motion data only) and crossmodal classification of objects. We first design, optimize and train two unimodal networks for classification based on audio and visual data. The resulting models are the basis for the crossmodal classification and create an expectation for the contribution of each modality. In the second step, we utilize the pretrained unimodal networks as branches of a crossmodal network. We evaluate different learning procedures to evaluate the effect of using crossmodal information. Figure 5 displays the full crossmodal network.

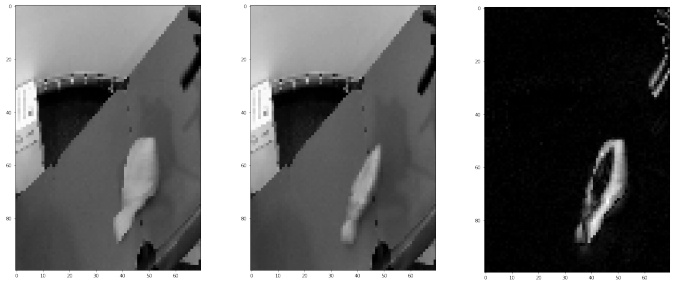

Fig. 4. Difference of images for one sample and one timestep is shown. By using differences of images as input to the neural classification approach, color and texture properties of the object are abstracted away.

IV. Experiment and Results

First, we report the results of uni-modal object classification using only the audio or visual features from the dataset. We show that different objects are easier to classify in either the visual and audio modality. We then report the results of the three different training approaches for the crossmodal network and show that using both modalities further enhances the classification accuracy.

To evaluate our models, we use 10-fold cross-validation. We use each fold to train the different unimodal and crossmodal network models. This provides us with the ability to train the crossmodal model on the same data that the imported models were trained on, ensuring that the imported unimodalities were not trained on the test sets.

We use automated Bayesian hyperparameter optimization [20] to optimize individual model architectures and learning rates. The optimized hyperparameters, their ranges and optimization results are shown in Table I.

1) Unimodal Audio Network: The unimodal audio network consists of a block of 3 convolution and pooling layers followed by a fully connected (dense) layer and an output layer with softmax activation. The input are the 130×150 spectrogram (image). The architecture is described in detail in Table I.

We achieve an audio classification accuracy of 0.6025 (± 0.0361). With the resulting confusion matrix shown in Figure 6, we can show that certain objects are classified very well using only audio data, e.g., the heavy blue banana, which consists of 3d printed hard plastic, while other objects are nearly indistinguishable.

2) Unimodal Visual Network: The architecture of this network follows conventional designs of visual networks; it consists of a block of 4 convolution and pooling layers

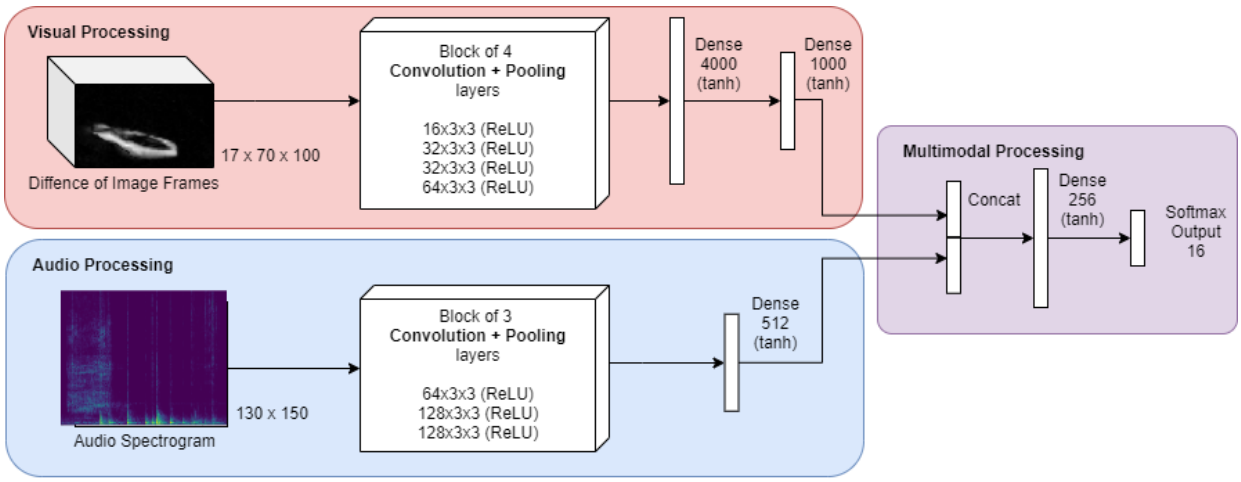

Fig. 5. The crossmodal learning network.

Layer Name	units	Activation	Optimized item
Audio Neural Network Model			
Conv2D	$64 \times 3 \times 3$	relu	filters
MaxPooling2D	4×4	-	-
Conv2D	$128 \times 3 \times 3$	relu	filters
MaxPooling2D	4×4	-	-
Conv2D	$128 \times 3 \times 3$	relu	filters
MaxPooling2D	4×4	-	-
Dense	512	tanh	units
Dense	16	softmax	-
Visual Neural Network Model			
TD(Conv2D)	$16 \times 3 \times 3$	relu	-
TD(AveragePooling2D)	2×2	-	-
TD(Conv2D)	$32 \times 3 \times 3$	relu	-
TD(AveragePooling2D)	2×2	-	-
TD(Conv2D)	$32 \times 3 \times 3$	relu	-
TD(AveragePooling2D)	2×2	-	-
TD(Conv2D)	$64 \times 3 \times 3$	relu	-
TD(AveragePooling2D)	2×2	-	-
Dense	4000	tanh	-
Dense	1000	tanh	-
Dense	16	softmax	-
Crossmodal Neural Network Model			
Concatenation	16×2	-	-
Dense	256	tanh	units
Dense	16	softmax	-

TABLE I
NETWORK SPECIFICATIONS

followed by two fully connected layers and an output layer with softmax activation. The input is a stack of 17 frames of 70×100 images. The architecture is described in Table I.

We achieve a visual classification accuracy of 0.8412 (± 0.0336) using only motion and shape but not texture or color information. The resulting confusion matrix (Figure 7) shows that a different subset of objects is classified well due to their characteristic motion pattern. Therefore, we hypothesize that combining both modalities will enhance the classification accuracy further.

3) Crossmodal Networks: The crossmodal network is created by taking the two unimodal networks and removing their output layers. Instead, the output of the last fully connected layers is concatenated and processed through one more dense layer before an output layer with softmax activation. The architecture is described in Table I. We explore three different training methods for the network: *a) Trained from scratch*: All network parameters are initialized randomly at the start of the training. *b) Retrained*: Parameters for the unimodal branches of the crossmodal network are imported from the unimodal experiment and are trained again during crossmodal training. *c) Pretrained*: As in the previous condition, the parameters for the audio and visual branches of the network are imported but fixed, only the parameters of the final fully connected and output layer are trained.

Name	Loss mean	Acc mean	Loss SD	Acc SD
Audio modality	1.5160	0.6025	0.2298	0.0361
Visual modality	0.7251	0.8412	0.1876	0.0336
Trained crossmodal	0.6493	0.8200	0.2956	0.1238
Retrained crossmodal	0.4692	0.8737	0.1802	0.0482
Pretrained crossmodal	0.3915	0.8975	0.1111	0.0388

TABLE II
CROSS VALIDATION RESULTS

We achieve classification accuracies of *a)* 0.8200 (± 0.1238), *b)* 0.8737 (± 0.0482) and *c)* 0.8975 (± 0.0366). Figures 8, 9 and 10 show the confusion matrices for the different crossmodal training variants. The results suggest that any form of pretraining reduces deviation. The highest accuracy is achieved by freezing the weights of the pretrained networks and only tuning the crossmodal layers of the network. This can be interpreted as training on multiple modalities causing destructive interferences [21], i.e. the back-propagated error from the concatenated modalities is not suitable for training the separate uni-modal network arms. Table II summarizes the results of 10-fold cross validation.

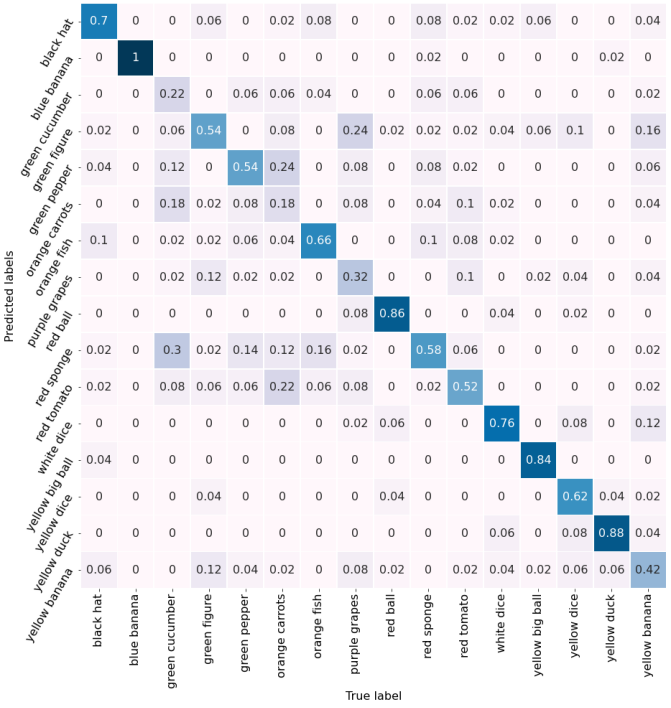

Fig. 6. The confusion matrix of audio model.

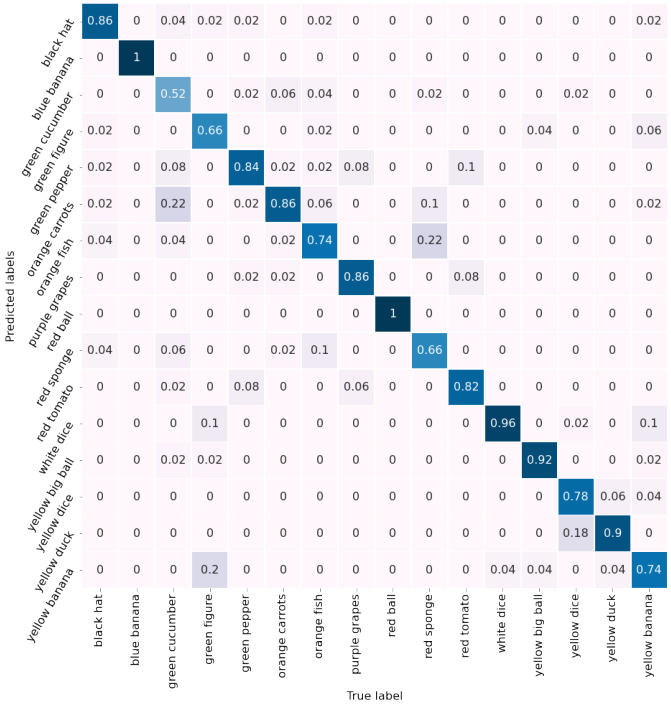

Fig. 8. The confusion matrix of trained crossmodal model.

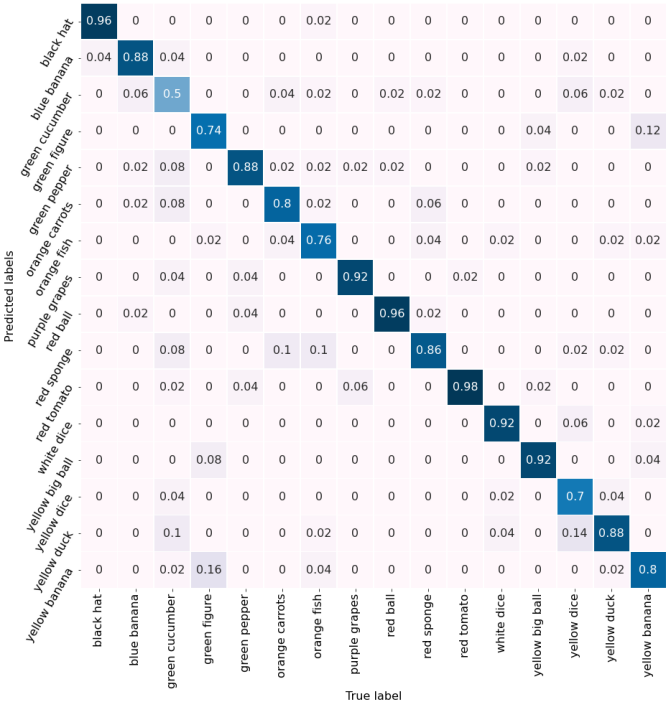

Fig. 7. The confusion matrix of visual model.

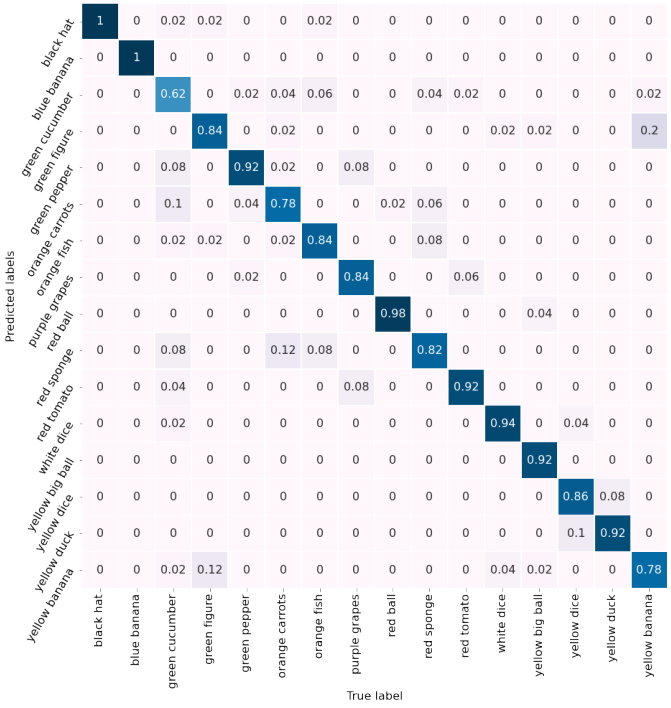

Fig. 9. The confusion matrix of retrained crossmodal model.

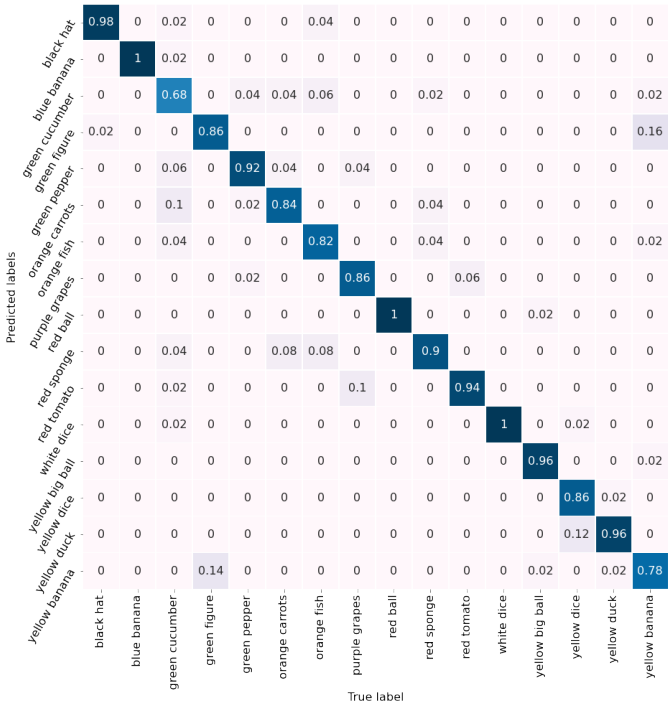

Fig. 10. The confusion matrix of pretrained crossmodal model.

V. CONCLUSIONS

We present a novel neurorobotic audiovisual learning setup in which a humanoid robot drops objects made from different materials onto a sloped surface. The setup is inspired by child-like ways of exploring object properties. Our main contributions are a publicly available dataset of 32800 images and 1600 s of audio recording from 800 samples for 16 objects and a novel neural architecture for classifying objects based on their characteristic sound and motion pattern that result from the drop. Using confusion matrices, we show that different objects are easier to classify in either the audio or visual modality. Our results show that the combination of the audio and visual modality outperforms unimodal classification. We also show that using pretrained unimodal networks enhances the classification accuracy as destructive interferences during learning are avoided. In future work, we will extend the experimental design to a larger object set and include a variety of active exploration procedures.

ACKNOWLEDGMENT

The authors gratefully acknowledge partial support from the German Research Foundation (DFG) under project CML (TRR 169).

REFERENCES

[1] T. Wilcox, R. Woods, C. Chapa, and S. McCurry, "Multisensory exploration and object individuation in infancy," *Developmental psychology*, vol. 43, no. 2, p. 479, 2007.

[2] M. Kerzel, M. Ali, H. G. Ng, and S. Wermter, "Haptic material classification with a multi-channel neural network," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 439–446, 2017.

[3] M. Kerzel, E. Strahl, C. Gaede, E. Gasanov, and S. Wermter, "Neuro-robotic haptic object classification by active exploration on a novel dataset," *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2019)*, 2019.

[4] M. Eppe, M. Kerzel, E. Strahl, and S. Wermter, "Deep neural object analysis by interactive auditory exploration with a humanoid robot," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 284–289, 2018.

[5] A. Sterling, J. Wilson, S. Lowe, and M. Lin, "Isnn: Impact sound neural network for audio-visual object classification," *ECCV*, 2018.

[6] S. Heinrich, M. Kerzel, E. Strahl, and S. Wermter, "Embodied multimodal interaction in language learning: the emil data collection," *Proceedings of the ICDL-EpiRob Workshop on Active Vision, Attention, and Learning (ICDL-Epirob 2018 AVAL)*, 2018.

[7] S. Werner and U. Noppeney, "Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization," *Cerebral Cortex*, vol. 20, no. 8, pp. 1829–1842, 2009.

[8] S. Molholm, W. Ritter, D. C. Javitt, and J. J. Foxe, "Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study," *Cerebral Cortex*, vol. 14, no. 4, pp. 452–465, 2004.

[9] T. Stanford and B. Stein, "Superadditivity in multisensory integration: Putting the computation in context," *Neuroreport*, vol. 18, pp. 787–92, 2007.

[10] A. Barutchu, C. Spence, and G. W. Humphreys, "Multisensory enhancement elicited by unconscious visual stimuli," *Experimental brain research*, vol. 236, no. 2, pp. 409–417, 2018.

[11] B. A. Rowland and B. E. Stein, "Temporal profiles of response enhancement in multisensory integration," *Frontiers in Neuroscience*, vol. 2, p. 33, 2008.

[12] D. Hecht, M. Reiner, and A. Karni, "Multisensory enhancement: Gains in choice and in simple response times," *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, vol. 189, pp. 133–43, 2008.

[13] F. L. Higgen, C. Heine, L. Krawinkel, F. Göschl, A. K. Engel, F. C. Hummel, G. Xue, and C. Gerloff, "Crossmodal congruency enhances performance of healthy older adults in visual-tactile pattern matching," *Frontiers in Aging Neuroscience*, vol. 12, p. 74, 2020.

[14] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–53, 2018.

[15] S. Heinrich, Y. Yao, T. Hinz, Z. Liu, T. Hummel, M. Kerzel, C. Weber, and S. Wermter, "Crossmodal language grounding in an embodied neurocognitive model," *Frontiers in Neurorobotics*, vol. 14, p. 52, 2020.

[16] M. Kerzel, E. Strahl, C. Gaede, E. Gasanov, and S. Wermter, "Neuro-robotic haptic object classification by active exploration on a novel dataset," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2019)*, Jul 2019.

[17] M. Eppe, M. Kerzel, E. Strahl, and S. Wermter, "Deep neural object analysis by interactive auditory exploration with a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2018.

[18] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "Nico—neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 113–120, 2017.

[19] E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," *Neurocomputing*, vol. 268, pp. 76–86, 2017.

[20] J. Mockus, *Bayesian Approach to Global Optimization: Theory and Applications*, ser. Mathematics and its Applications. Springer Netherlands, 1989.

[21] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, "A modulation module for multi-task learning with applications in image retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–416.