



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Latent object characteristics recognition with visual to haptic-audio cross-modal transfer learning

Citation for published version:

Saito, N, Moura, J, Uchida, H & Vijayakumar, S 2024, Latent object characteristics recognition with visual to haptic-audio cross-modal transfer learning. in *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Institute of Electrical and Electronics Engineers, 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, Abu Dhabi, United Arab Emirates, 14/10/24. <https://doi.org/10.48550/arXiv.2403.10689>

Digital Object Identifier (DOI):

[10.48550/arXiv.2403.10689](https://doi.org/10.48550/arXiv.2403.10689)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Latent Object Characteristics Recognition with Visual to Haptic-Audio Cross-modal Transfer Learning

Namiko Saito¹, João Moura¹, Hiroki Uchida², and Sethu Vijayakumar¹

Abstract—Recognising the characteristics of objects while a robot handles them is crucial for adjusting motions that ensure stable and efficient interactions with containers. Ahead of realising stable and efficient robot motions for handling/transferring the containers, this work aims to recognise the unobservable latent object characteristics. While vision is commonly used for object recognition by robots, it is ineffective for detecting hidden objects. However, recognising objects indirectly using other sensors is a challenging task. To address this challenge, we propose a cross-modal transfer learning approach from vision to haptic-audio. We initially train the model with vision, directly observing the target object. Subsequently, we transfer the latent space learned from vision to a second module, trained only with haptic-audio and motor data. This transfer learning framework facilitates the representation of object characteristics using indirect sensor data, thereby improving recognition accuracy. For evaluating the recognition accuracy of our proposed learning framework we selected shape, position, and orientation as the object characteristics. Finally, we demonstrate online recognition of both trained and untrained objects using the humanoid robot Nextage Open. See our accompanying video here: <https://www.youtube.com/watch?v=sOHqPC1uusg>

I. INTRODUCTION

Interacting with containers is a fundamental task in robotics, applicable to warehouse, delivery, and home use. These scenarios require robots to handle containers efficiently while keeping their content safe and stable, i.e. without having the objects topple, mix, and get damaged through impact. Therefore, robots need to adjust their motion (speed, acceleration, and trajectory) to the specific characteristics of the contained objects. However, often these objects are hidden by the container’s cover; therefore, ahead of realizing stable and quick robot motions for handling/transferring the containers, this work aims to recognise the unobservable latent object characteristics.

Recognising latent object characteristics inside a closed container presents a significant challenge. One solution involves leveraging indirect sensing modalities such as haptic and/or audio cues to infer these latent characteristics as [1], [2]. However, we will show that simply using haptic and audio as indirect sensing fails to accomplish good recognition. To enable using haptic-audio as indirect sensing to estimate the latent characteristics, in this work we propose a cross-modal transfer learning framework, where we first

¹Authors are with the School of Informatics, The University of Edinburgh, Edinburgh, U.K., and with The Alan Turing Institute, London, U.K.

²Author is with the Department of Intermedia Art and Science, Waseda University, Tokyo, Japan

This work is supported by the EU H2020 Project Harmony (Grant No. 101017008) and Japan Science and Technology Agency (JST) Moonshot R&D (Grant No. JPMJMS2031).

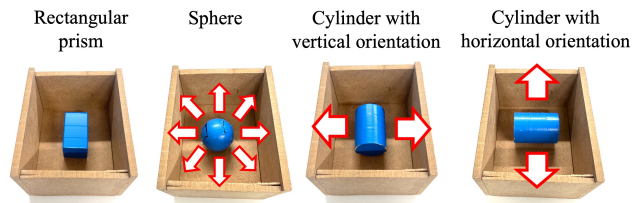


Fig. 1: A rectangular prism typically offers greater stability compared to a cylinder or a sphere. The ease with which a cylinder rolls is contingent upon its orientation, while a sphere possesses the capability to roll in any direction.

learn a prediction model based on vision reconstruction, which can directly observe the target characteristics, and transfer its latent space to a second prediction model learnt only with haptic-audio sensing information. This approach aims to bridge the gap between direct and indirect sensing modalities, facilitating more accurate recognition of hidden latent object characteristics using haptic and audio cues.

There are various characteristics that influence object behaviour as addressed by Gao *et al.* [3], who focus on factors such as mass, fragility, deformability, and certain properties that humans can estimate from objects’ appearance. Funabashi *et al.* [4] examine attributes like heaviness, softness, and slipperiness, which are related to the sense of touch. Additionally Gonçalves *et al.* [5] categorize objects based on characteristics such as area, squareness, circularity, and other geometric properties. As an exemplar, in this work, we focus on recognising the shape, orientation, and position of the objects within the container to assess our transfer learning framework, as they provide measurable ground truth that allows for numerical evaluation of accuracy.

Shape, orientation, and position are fundamental attributes that influence object behavior and are often concealed by the container’s cover. For example, as shown in Fig. 1, shape and orientation affect the direction of motion. While rectangular-shaped objects are stable, spheres can easily move in every direction. For the cylinder, in addition to the shape, one has to consider the orientation when predicting the direction of its rolling motion. Moreover, the position of the object within the container also influences its motion. Objects positioned at the centre of the container have more freedom of movement, while those placed near the corners are constrained by the container’s walls, limiting their motion options.

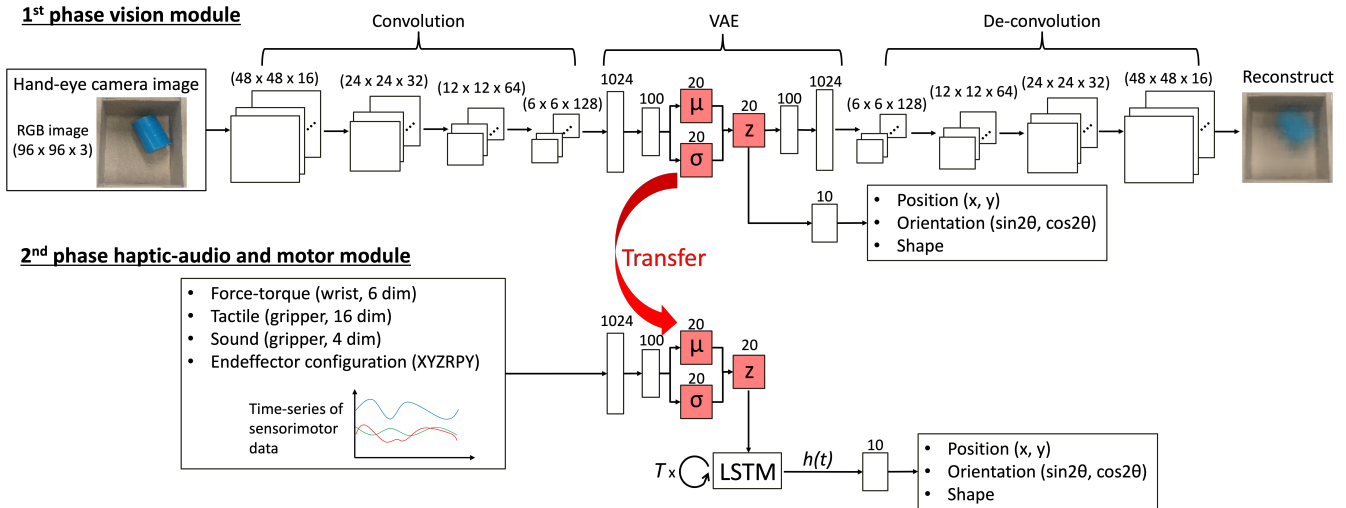


Fig. 2: Overall learning model. We train the first phase vision module, and then transfer the latent space to the second phase haptic-audio and motor module to use it for initiating the training process of the latter module. Our objective is to discern object position, orientation, and shape utilising indirect information derived from haptic-audio cues.

II. RELATED WORK

A. Object characteristics recognition with vision, haptics and audio

There are several methods for object detection [6]–[8], and recognition [9]–[11] with vision. In contrast, Gao *et al.* [3] addressed the understanding of physical concepts such as mass, fragility, and deformability for robot manipulation employing LLM and relied on human annotations based on vision to achieve this understanding.

Using GelSight [12] vision-tactile sensors, Anzai *et al.* [13] achieved perception of object pose within the hand, Lin *et al.* [14] achieved object identification from data samples through object picking, while Jiang *et al.* [15] realized the perception of material, shape, and pose of objects. Falco *et al.* [16] and Li *et al.* [17] employed cross-modal learning, integrating tactile and vision, for recognising object shapes.

Sterling *et al.* [18] utilised impact sound data generated during hitting an object to estimate the object’s geometry and material properties, and also reconstructed image data. Chen *et al.* [1] introduced a learning model to predict object shape by analysing acoustic vibrations captured when dropping an object into a container. Additionally, Niwa *et al.* [2] employed acoustic diffraction of audible sound to estimate the distance of an object, even when occluded by another object.

The aforementioned research successfully recognised static object characteristics. However, challenges remain in recognising characteristics dynamically as robots perform tasks. Our objective is to address this gap by developing a system capable of identifying object characteristics on the fly, while the robot is executing a task. This online recognition is crucial for generating targeted motions that depend on the object characteristics in future stages of the task.

B. Dynamic object characteristics recognition during motions

Marturi *et al.* [19] demonstrates a robot arm tracking and grasping moving objects using a vision system. Additionally, Chen *et al.* [20] achieved 6D pose dynamic estimation through optical flow analysis while manipulating objects. Funabashi *et al.* [4] used uSkin [21] skin-type sensors which can detect force in 3-axis, to recognise heaviness, softness, and slipperiness of an object while a multi-fingered hand grasps it.

Saito *et al.* [22] and Gemici *et al.* [23] represent the characteristics of objects using the neural networks’ latent space values. These networks utilise multimodal data which includes vision, tactile, and force sensing collected while the robot arm is in motion.

Previous work relies on sensing modalities such as vision, touch, or vibration that can directly observe the target object characteristics. In contrast, this work focuses on the scenario of latent target characteristics, for instance the case of recognizing visual information, such as position and shape of an object inside of a container, where the robot can only indirectly sense these characteristics. This indirect sensing presents additional challenges in prediction and recognition.

III. APPROACH

The goal of this work is to achieve recognition of object characteristics while in motion, focusing particularly on the case of indirect sensing of those characteristics. To address this challenge, we propose a cross-modal transfer learning method. We choose a box transfer task as an example to evaluate our method. In this task, a humanoid robot, Nextage Open from Kawada Robotics, grasps a box containing a small object and moves the box to different poses, randomly generated. As the robot moves, our learning model predicts

the characteristics of the small object, such as its shape, position, and orientation, on the fly.

Previous research on cross-modal learning [14], [16], [17], which introduced learning models trained with both vision and tactile data to predict object class or shape, addressed occlusion and variable lighting conditions. Inspired by these studies, we propose a two-phase cross-modal transfer learning approach. Initially, we train a vision module by opening the lid of the container to enable direct visual observation of the object. Subsequently, we transfer the learned latent space to the second phase, where we train the model using haptic and audio data, allowing the robot to sense the object indirectly assuming the container is closed. We hypothesise that warm starting the training of the second phase, which uses haptic and audio data, with the latent space of the visual module will improve the recognition of the object characteristics.

In summary, our contributions are as follows:

- Proposing a two-phase learning framework for recognising latent object characteristics.
- Demonstrating that warm starting the initial haptic-audio latent state with the learnt visual latent state significantly improves the recognition of visual characteristics when using only indirect haptic-audio and motor sensing.
- Validating the proposed learning framework on online prediction of object shape, position, and orientation on a physical humanoid robot setup.

IV. METHOD

Fig. 2 depicts the overall architecture of the learning modules. We train the vision module in the first phase and transfer its latent space to the second phase haptic-audio and motor module. This transfer serves as the initial state value, improving the training convergence of the second phase module. Subsequently, as a baseline, we train the second phase to recognise object position, orientation, and shape, without using the warm start from the visual module latent space.

A. 1st phase vision module

This module aims to predict object characteristics from RGB images in situations where objects are directly observable within the container. To minimize the cost of collecting training data, we captured snapshot images of objects randomly placed inside the box, without moving the robot.

The structure of the module consists of convolution layers, variational autoencoder (VAE) [24] and de-convolution layers, with the number of neurons detailed in Fig. 2. In the convolutional layers, we utilise a kernel size of 4×4 , a stride of 2, and zero padding of 1. These layers convolute the image data and reconstruct it (y_{image}). Additionally, we incorporate a fully connected layer from the latent space of the VAE to output continuous values representing position and orientation (y_{pos} , y_{ori}) and shape labels (y_{shape}).

The architecture uses a self-supervised approach for image reconstruction and supervised learning with ground truth data

for characteristics prediction. The loss function L is

$$L = \alpha D_{\text{KL}}(Q||P) + \beta E_{\text{reconst}} + E_{\text{predict}}, \quad (1)$$

where

$$D_{\text{KL}}(Q||P) = -\frac{1}{2} \sum_{i \in N} (1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2), \quad (2)$$

and

$$E_{\text{reconst}} = \frac{1}{N} \sum_{i \in N} (y_{\text{image},i} - y_{\text{image},i}^{\text{G}})^2, \quad (3)$$

and

$$E_{\text{predict}} = \frac{1}{N} \sum_{i \in N} ((y_{\text{pos},i} - y_{\text{pos},i}^{\text{G}})^2 + (y_{\text{ori},i} - y_{\text{ori},i}^{\text{G}})^2 + (y_{\text{shape},i} - y_{\text{shape},i}^{\text{G}})^2), \quad (4)$$

where D_{KL} means KL divergence and σ and μ are parameters of the Gaussian distribution. Q and P represent stochastic model of encoder and decoder, i.e., $Q = q_{\varphi}(z|x)$ and $P = p_{\theta}(z)$, where φ is the parameters of encoder, θ is the parameters of decoder, x is input, and z is the latent variable. And y^{G} is the ground truth output, N is the number of snapshot images. The hyperparameters α and β are used to adjust the loss, with values set to 0.25 and 0.5, respectively, to prioritize the prediction error in Eq. (4). Due to overlaps between the information necessary for image reconstruction and for predicting object characteristics, we adopt this architecture to efficiently represent it in the latent space. Combining multiple errors as shown in Eq. (1) prevents the over-fitting resulting from only using the prediction error, i.e. the reconstruction error serves as a regularization term which improves convergence properties, leading to smaller number of epochs. We train this module for 5,000 epochs until the training error converges, using the Adam optimizer [25] to update the neural network weights. The middle layer of the VAE uses a Sigmoid activation function while the remaining layers use a ReLU activation function.

B. 2nd phase haptic-audio and motor module

The objective of the second phase learning is to predict object characteristics using time-series data from haptic, audio, and motor data, useful for when direct observation of objects is unavailable. We record sensorimotor data while commanding the robot to random poses within its workspace, resulting in swinging motions of the container.

The structure of this module comprises an encoder part of VAE and a long short-term memory (LSTM) with a fully connected layer, with the number of neurons detailed in Fig. 2. After training the first phase module, we transfer the latent space value, highlighted in red in Fig. 2, to the second module to use it as the initial value. The latent space should encapsulate the information necessary to predict object position, orientation, and shape, which is directly learned from vision. Our aim is to leverage this representation to facilitate learning with haptic, audio, and motor data. Since the second module handles time-sequential data, we employ LSTM, using sigmoid activation functions and with a sequence of T

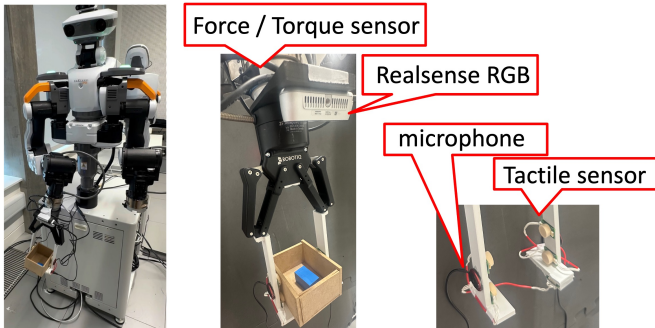


Fig. 3: Nextage robot and sensor settings.

time steps, where we iteratively input each single time-step data $x(t)$ (where t denotes the current time step). LSTM recurrently learns these inputs and encodes the sequential information in its hidden space $h(t)$ as

$$h(t) = \text{LSTM}(z(t), z(t-1), \dots, z(t-T+1)), \quad (5)$$

with

$$z(t) = \text{Encode}(x(t)). \quad (6)$$

Then, we use a full-connect layer to compute the position, orientation, and shape from the LSTM hidden latent space as

$$y_{\text{pos}}(t), y_{\text{ori}}(t), y_{\text{shape}}(t) = \text{FullConnect}(h(t)) \quad (7)$$

The loss function is

$$E = \frac{1}{N} \sum_{i \in N} ((y_{\text{pos},i}(t) - y_{\text{pos},i}^G(t))^2 + (y_{\text{ori},i}(t) - y_{\text{ori},i}^G(t))^2 + (y_{\text{shape},i}(t) - y_{\text{shape},i}^G(t))^2), \quad (8)$$

where here N represents the number of sequential datasets. We train this module for 20,000 epochs, using the Adam optimizer, until the error converges.

C. Evaluation

We conducted experiments to assess whether the modules can accurately detect position, orientation, and shape using offline untrained sequential data. To evaluate the contribution of cross-model transfer learning, we compared our approach with a baseline where learning starts with haptic, audio, and motor data from scratch, without support from the first phase module. Furthermore, we demonstrated object recognition while the robot manipulates the container using both trained and “untrained” objects.

V. EXPERIMENTAL SETUP

A. Hardware design and Objects

In the robot experiment, we utilise the right arm of Nextage Open, which has 6 degrees of freedom (DOF) as depicted in Fig. 3. We equip the arm with a Realsense camera D435i, an ATI gamma force and torque sensor, and a Robotiq 2f-140 gripper. To enhance gripping stability, we 3D printed gripper fingers to replace the original Robotiq gripper’s fingers, are shown in Fig. 3. Additionally, we

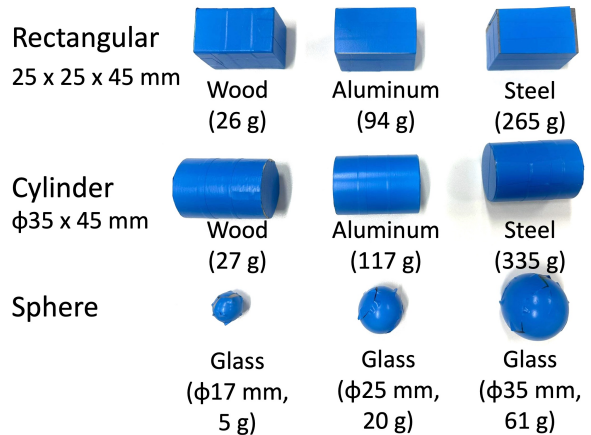


Fig. 4: 9 different shapes and sizes / weights objects for training.

integrate four Shokkaku Pot tactile sensors, developed by Touchence Inc., within the gripper fingers and install two piezo contact microphones on the outer sides.

We used a box measuring $100 \times 120 \times 70$ mm (width \times length \times height), built with 6.3 mm thick wooden boards. In training the modules, we utilised 9 different objects of various shapes, sizes, and weights, as summarized in Fig. 4. To remove any colour-based distinctions in the first vision module, we covered all the objects with the same blue colour. During the evaluation, we tested 8 untrained objects with characteristics distinct from those of the trained objects.

B. Dataset

We recorded the following data during the experiment:

- XYZ position and yaw-pitch-roll of the right arm end-effector (6 dimensions).
- RGB images (96 pixels width \times 96 pixels height \times 3 channels).
- Force and torque data (6 dimensions).
- Tactile sensor data (4 points \times 4 units).
- Piezo microphone data (2 units).

Prior to inputting the data into the model, we normalize the values of image data to the range $[0, 255]$ and other sensorimotor data to the range $[0.2, 0.8]$.

For the first vision module, we gathered snapshot images with one of the training objects randomly placed inside the box. Specifically, we collected 270 images for training the first module, with 30 images captured for each of the 9 training objects. Additionally, we acquired 36 images for testing, with 4 images taken for each of the 9 objects.

For the second haptic-audio and motor module, we commanded the right arm of the robot to random locations and orientations within its workspace. As the robot swung the box randomly, we recorded sequential data from tactile sensors, force-torque sensors, microphones, and end-effector configurations at a frequency of 50 Hz. We collected 72 sequence datasets, each comprising 1000 time steps, with 8 sequences captured for each of the 9 objects. During

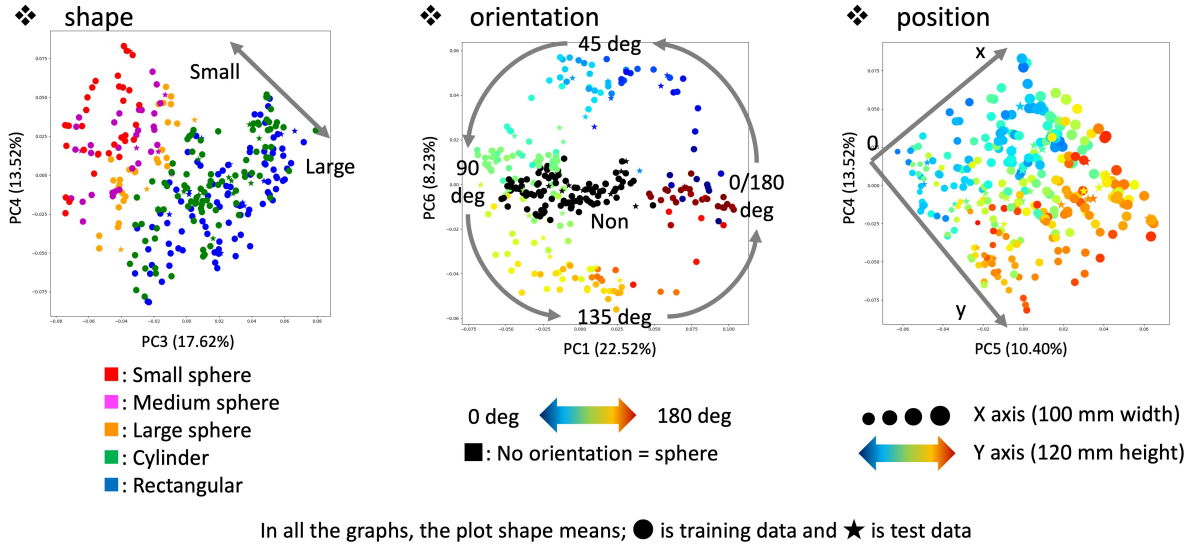


Fig. 5: PCA on latent space after training the 1st phase vision module, which is transferred to the 2nd phase haptic-audio module. The latent space represents the object shape, orientation and position.

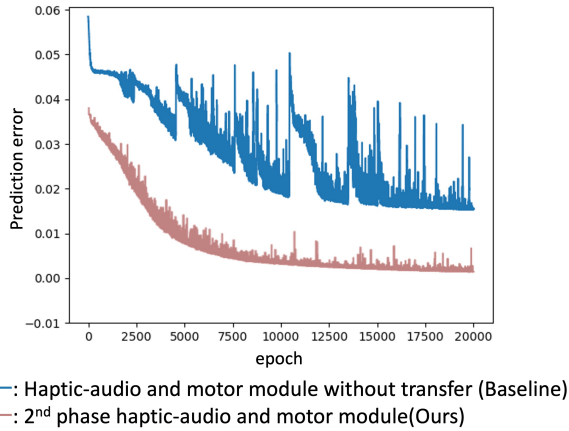


Fig. 6: Assessment of prediction error throughout 20,000 epochs of training, contrasting the performance between the proposed method and the baseline approach.

the training of the second module, we configured the time window to 250 steps, equivalent to 5 seconds, and slid it with a stride of 50 time steps. In other words, we trained the model with a total of 1,152 sequential datasets (16 windows \times 72 sequences). For testing the second module, we utilised 135 different sequential datasets, with 15 datasets recorded for each of the 9 objects.

C. Specification of Ground Truth for Recognition

For object shape, we assigned labels to each dataset using a single dimension as 0.2 for the rectangular prism, 0.35 for the cylinder, 0.5 for the large sphere, 0.65 for the medium sphere, and 0.8 for the small sphere. We trained the modules to predict these labelled values. We use interpolated regression values in place of a softmax activation function in the last layer of our network. This enables the representation

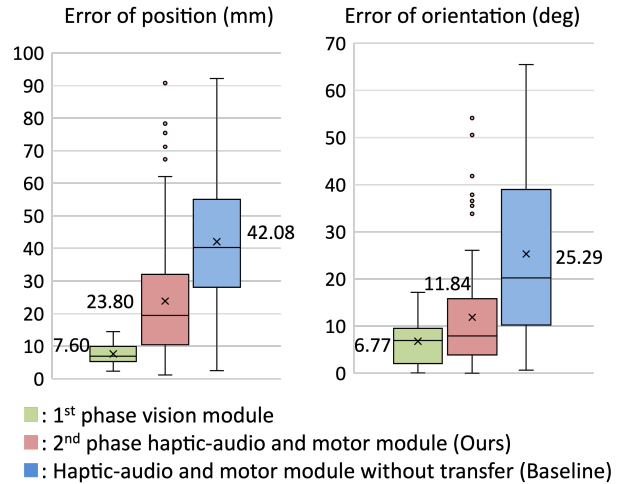


Fig. 7: Position and orientation recognition result, showing the prediction error. In a box plot, the cross shows the average, the lower edge of the box represents the first quartile, the line drawn inside the box represents the second quartile, and the upper edge of the box represents the third quartile. The vertical lines represent the maximum and minimum values and the dots are outliers.

of shapes as continuous geometric patterns instead of just classification of shapes. By doing so, we expect that our model can effectively handle the variability and randomness of daily shaped objects, providing a more nuanced and flexible understanding of shape representations. For classifying the object shapes in evaluation, we set the classification threshold at the equidistant value between the numerical labels. Then, we evaluated the classification success based on the number of matches between the predictions and the

❖ 1st phase vision module

Recognition \ Ground truth	Rectangular	Cylinder	Sphere		
			Large	Mid	Small
Rectangular	6	5	0	0	0
Cylinder	6	7	2	0	0
Sphere	Large	0	0	2	0
	Middle	0	0	0	4
	Small	0	0	0	0

❖ 2nd phase haptic-audio and motor module (Ours)

Recognition \ Ground truth	Rectangular	Cylinder	Sphere		
			Large	Mid	Small
Rectangular	29	1	0	0	0
Cylinder	14	34	3	0	0
Sphere	Large	2	10	10	2
	Middle	0	0	2	15
	Small	0	0	0	0

❖ Haptic-audio and motor module without transfer (baseline)

Recognition \ Ground truth	Rectangular	Cylinder	Sphere		
			Large	Mid	Small
Rectangular	19	3	0	0	0
Cylinder	24	40	0	0	0
Sphere	Large	2	2	2	12
	Middle	0	0	13	3
	Small	0	0	0	0

Fig. 8: Object shape recognition results which show the number of correct and incorrect predictions.

ground truth shapes.

To obtain the ground truth object positions and orientations, we utilised OpenCV, an open-source computer vision library, to detect blue square or circular shapes in camera images. The position contains 2 dimensions: the x and y coordinates of the object’s centre. For setting the ground truth orientation, we initially detected the degree of inclination (θ) of the long edge in the images. However, expressing the orientation with the original degree data posed a problem: the actual object orientation is similar between 0 and 179 degrees, yet the values are the most distant. To address this issue, we defined the ground truth orientation with 2 dimensions: $(\sin(2\theta), \cos(2\theta))$. This representation enables us to express how close the orientation is to vertical, horizontal, and 45-degree directions clockwise and counterclockwise. For spheres, which have no inherent orientation, we set both orientation dimensions to 0.

VI. RESULTS AND DISCUSSION

A. Latent space representation

To analyse the latent space of the VAE, transferred from the first module to the second module, and assess its representation of object characteristics, we applied Principal Component Analysis (PCA) to its latent space z , which is 20 dimensions. Fig. 5 shows a clear distribution of the objects

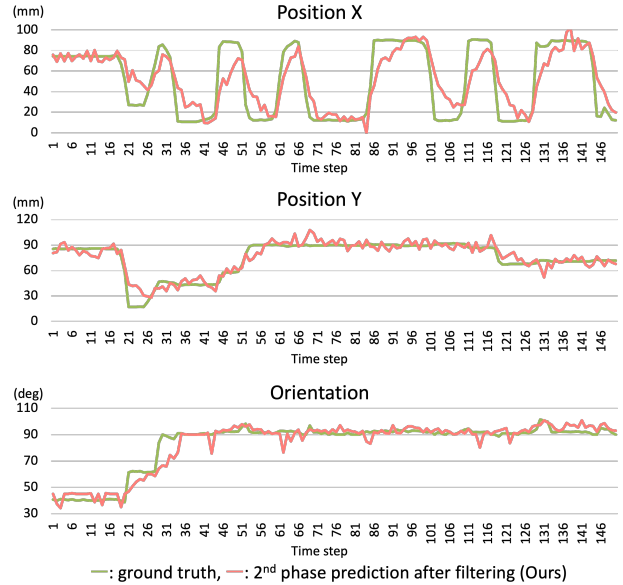


Fig. 9: Tracking results observed over a 15-second duration (consisting of 150 time steps at a rate of 10 Hz) during the real robot experiment with the wood cylinder and employing the proposed model

characteristics along the principal components. Observing the PC3 and PC4 space, we observe that object shapes are ordered according to the size of the objects. Additionally, on PC1 and PC6, the representation of orientation demonstrates degrees in a circular manner, with spheres centred at (0, 0) value. Furthermore, PC1 primarily indicates whether the orientation is vertical or horizontal, while PC6 expresses clockwise or counterclockwise inclination. Finally, PC4 and PC5 express the x and y positions in the box. In conclusion, the latent space effectively distributes and represents geometric and spatial information of the object, which can potentially provide valuable support for the second phase learning.

B. Prediction error convergence during training

To evaluate whether our proposed transfer method improves the second phase learning, we compared the prediction error calculated with Eq. (8) with a baseline approach that uses the same haptic-audio and motor learning, without the warm start from the vision module. Fig. 6 shows the convergence of the error during training over 20,000 epochs, where it is evident that the proposed method starts with a lower error than the baseline from the beginning. Furthermore, while the prediction error of the baseline exhibits significant fluctuations, the proposed method converges smoothly to a lower level of error. Therefore, we conclude that the transfer approach can effectively support training by significantly decreasing the prediction error of object characteristics.

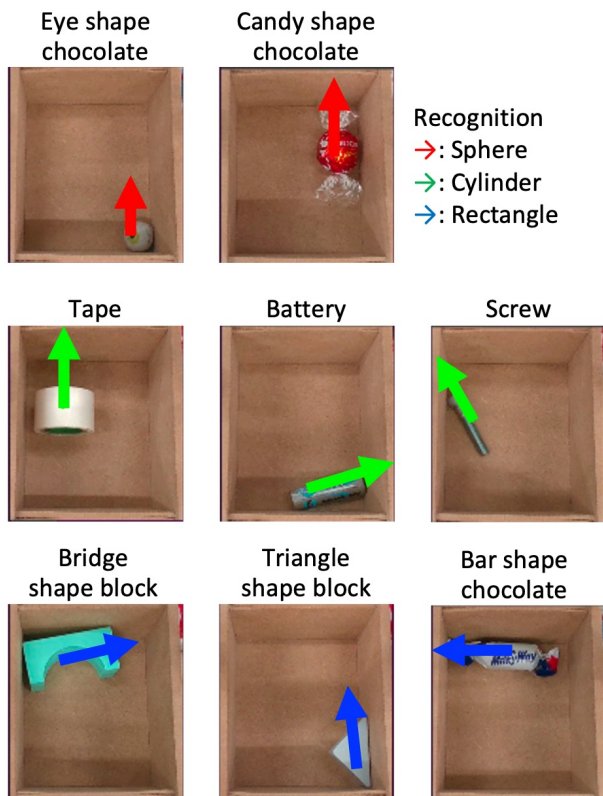


Fig. 10: Snapshots of prediction results in the real robot experiment using untrained objects with the proposed model. The arrows are showing the predicted shape with their colour, orientation with their direction, and position with the position they start.

C. Prediction result with test data

Fig. 7 shows the prediction error of position and orientation, comparing the first phase vision module, our proposal, and the baseline, using the offline dataset. With our proposal, the average error was 23.80 mm for position and 11.84 degrees for orientation. Although these errors are significantly larger when compared with using vision for prediction, that is a much simpler problem setting because the position and orientation are directly observable in that case. In the baseline, without using vision, the average error was 42.08 mm for position and 25.29 degrees for orientation, with much larger variance. This indicates that our proposal contributes to the accuracy of predicting position and orientation, when only using indirect observations.

The tables in Fig. 8 display the results of shape recognition, where the rows represent the actual shape and the columns represent predictions. For the first phase vision module, shape recognition is based on direct observation with camera images. The success rate of recognition was 63.9% (23/36 samples), which is relatively low due to the similarity in size between rectangular prisms and cylinders, making them difficult to distinguish from image snapshots. Conversely, both our proposed model and the baseline use sequential sensorimotor data to predict shape, without relying

on vision. The success rate of our proposed model was 71.1% (96/135 samples), while the baseline achieved a success rate of 47.4% (64/135 samples). As a result, our proposal exhibited the highest accuracy in shape recognition. This suggests that the second module could leverage the first phase and further improve shape recognition.

D. Robot experiment with trained and untrained objects

We conducted experiments by commanding the robot to swing the box and recognise the object characteristics online, outputting data at a rate of 10 Hz. With our proposal, predictions ($y_{\text{pos}}(t)$, $y_{\text{ori}}(t)$, $y_{\text{shape}}(t)$) calculated using Eq. (7) are independent of previous predictions ($y_{\text{pos}}(t-n)$, $y_{\text{ori}}(t-n)$, $y_{\text{shape}}(t-n)$). Consequently, the prediction can be noisy including outliers and exhibit abrupt changes, making it unsuitable for tracking. To address this issue, we applied a low-pass filter to the output of the model as:

$$\tilde{y}(t) = 0.4 \cdot y(t) + 0.3 \cdot \tilde{y}(t-1) + 0.2 \cdot \tilde{y}(t-2) + 0.1 \cdot \tilde{y}(t-3), \quad (9)$$

where $\tilde{y}(t)$ is the filtered output.

Fig. 9 illustrates the tracking results for 150 time steps, which is a 15-second period, with predictions after passing through the smoothing filter. Additionally, we provide the attached video demonstrating the tracking results after applying the smoothing filter to the model's predictions.

We further evaluated our model using eight untrained objects, as shown in Fig. 10. Despite their characteristics differing from the trained objects, the proposed model demonstrated the ability to recognise their shape, position, and orientation. The model exhibited generalization capabilities, indicating its potential applicability to a broader range of objects. The attached video showcases the tracking results after applying the smoothing filter to the predictions for these untrained objects, further demonstrating the effectiveness and generalization of our proposed model.

VII. CONCLUSION

In conclusion, this study presents a novel approach to latent object characteristics recognition in robotic manipulation tasks. By employing a two-phase cross-modal transfer learning method, we demonstrate the effectiveness of leveraging visual information from the first phase to improve recognition accuracy with haptic-audio and motor data in the second phase. Our experiments show that the proposed method outperforms the baseline approach, achieving better accuracy in predicting object position, orientation, and shape. Furthermore, our model exhibits generalization capabilities, successfully recognising untrained objects with characteristics similar to the trained categories. These results highlight the potential of our approach in enhancing robotic perception and manipulation in real-world scenarios.

Recognizing characteristics using indirect observations is challenging. In this work we selected characteristics such as shape, position and orientation because first we could generate a ground truth references for validation and second because the first vision module can directly observe and encode them in its latent space representation. However, there

are other object types of characteristics that also influence the object motion behaviour, such as stiffness, friction, and even deformability, that might have less reliable direct observation from the vision module and make it difficult to generate ground truth references for validation. Our framework can, in principle, handle these other types of characteristics but that remains untested.

Despite the promising results achieved in this study, there remain areas for improvement. As noted in Section VI-D, our current model lacks the incorporation of previous prediction information, which deviates from physical principles. Consequently, our model struggles to handle rigid objects remaining motionless and requires five-second time windows and relatively large swinging-box movements to hit the object against a wall at least once for localization. To address this limitation, future work will focus on implementing a Markov model to assess the reliability of output from the learning model based on previous predictions and output the modified recognition result. By incorporating this information, we aim to enhance the precision of predicting the next position and orientation of objects.

In future work, our focus will shift towards integrating the object characteristics information recognised by our proposed model into robot control systems. Specifically, we plan to develop a control system capable of adjusting speed, acceleration, and trajectory based on object characteristics from our recognition model. By doing so, we aim to enable robots to perform stable and efficient transportation tasks, ensuring that they can interact with objects in a quick and reliable manner in the context of various applications, including logistics, manufacturing, and the service industries.

REFERENCES

- [1] B. Chen, M. Chiquier, H. Lipson, and C. Vondrick, "The boombox: Visual reconstruction from acoustic vibrations," in *5th Annual Conference on Robot Learning*, 2021.
- [2] H. Niwa, T. Ogata, K. Komatani, and O. G. Hiroshi, "Distance estimation of hidden objects based on acoustical holography by applying acoustic diffraction of audible sound," in *IEEE International Conference on Robotics and Automation*, 2007, pp. 423–428. DOI: [10.1109/ROBOT.2007.363823](https://doi.org/10.1109/ROBOT.2007.363823).
- [3] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *arXiv*, 2023. DOI: [10.48550/arXiv.2309.02561](https://doi.org/10.48550/arXiv.2309.02561).
- [4] S. Funabashi, G. Yan, F. Hongyi, A. Schmitz, L. Jamone, T. Ogata, and S. Sugano, "Tactile transfer learning and object recognition with a multifingered hand using morphology specific convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022. DOI: [10.1109/TNNLS.2022.3215723](https://doi.org/10.1109/TNNLS.2022.3215723).
- [5] A. Gonçalves, J. Abrantes, G. Saponaro, L. Jamone, and A. Bernardino, "Learning intermediate object affordances: Towards the development of a tool concept," in *4th International Conference on Development and Learning and on Epigenetic Robotics*, 2014, pp. 482–488. DOI: [10.1109/DEVLRN.2014.6983027](https://doi.org/10.1109/DEVLRN.2014.6983027).
- [6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, Springer International Publishing, 2016, pp. 21–37.
- [8] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242. DOI: [10.1109/IWSSIP48289.2020.9145130](https://doi.org/10.1109/IWSSIP48289.2020.9145130).
- [9] P. Loncomilla, J. Ruiz-del-Solar, and L. Martínez, "Object recognition using local invariant features for robotic applications: A survey," *Pattern Recognition*, vol. 60, pp. 499–514, 2016, ISSN: 0031-3203. DOI: [10.1016/j.patcog.2016.05.021](https://doi.org/10.1016/j.patcog.2016.05.021).
- [10] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 681–687. DOI: [10.1109/IROS.2015.7353446](https://doi.org/10.1109/IROS.2015.7353446).
- [11] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928. DOI: [10.1109/IROS.2015.7353481](https://doi.org/10.1109/IROS.2015.7353481).
- [12] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017. DOI: [10.3390/s17122762](https://doi.org/10.3390/s17122762).
- [13] T. Anzai and K. Takahashi, "Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9361–9368. DOI: [10.1109/IROS45743.2020.9341799](https://doi.org/10.1109/IROS45743.2020.9341799).
- [14] J. Lin, R. Calandra, and S. Levine, "Learning to identify object instances by touch: Tactile recognition via multimodal matching," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3644–3650. DOI: [10.1109/ICRA.2019.8793885](https://doi.org/10.1109/ICRA.2019.8793885).
- [15] J. Jiang and S. Luo, "Chapter 2 - robotic perception of object properties using tactile sensing," in *Tactile Sensing, Skill Learning, and Robotic Dexterous Manipulation*, Academic Press, 2022, pp. 23–44.
- [16] P. Falco, S. Lu, A. Cirillo, C. Natale, S. Pirozzi, and D. Lee, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5273–5280. DOI: [10.1109/ICRA.2017.7989619](https://doi.org/10.1109/ICRA.2017.7989619).
- [17] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [18] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [19] N. Marturi, M. Kopicki, A. Rastegarpanah, V. Rajasekaran, M. Adjigble, R. Stolkin, A. Leonardis, and Y. Bekiroglu, "Dynamic grasp and trajectory planning for moving objects," *Auton Robot*, vol. 43, pp. 1241–1256, 2019. DOI: [10.1007/s10514-018-9799-1](https://doi.org/10.1007/s10514-018-9799-1).
- [20] T. Chen and D. Gu, "6d object pose tracking with optical flow network for robotic manipulation," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 8048–8053, 2023, 22nd IFAC World Congress, ISSN: 2405-8963. DOI: [10.1016/j.ifacol.2023.10.930](https://doi.org/10.1016/j.ifacol.2023.10.930).
- [21] T. P. Tomo, A. Schmitz, W. K. Wong, H. Kristanto, S. Somlor, J. Hwang, L. Jamone, and S. Sugano, "Covering a robot fingertip with uskin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 124–131, 2018. DOI: [10.1109/LRA.2017.2734965](https://doi.org/10.1109/LRA.2017.2734965).
- [22] N. Saito, T. Ogata, S. Funabashi, H. Mori, and S. Sugano, "How to select and use tools? : Active perception of target objects using multimodal deep learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2517–2524, 2021. DOI: [10.1109/LRA.2021.3062004](https://doi.org/10.1109/LRA.2021.3062004).
- [23] M. C. Gemic and A. Saxena, "Learning haptic representation for manipulating deformable food objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 638–645. DOI: [10.1109/IROS.2014.6942626](https://doi.org/10.1109/IROS.2014.6942626).
- [24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv*, 2013. DOI: [10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114).
- [25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, 2014. DOI: [10.48550/arXiv.1412.6980](https://doi.org/10.48550/arXiv.1412.6980).