Cross-Modal Environment Self-Adaptation During Object Recognition in Artificial Cognitive Systems

David Miralles (david.miralles@salle.url.edu)
La Salle - Universitat Ramon Llull
Guillem Garrofé
La Salle - Universitat Ramon Llull
Calota Parés
La Salle - Universitat Ramon Llull
Alejandro González
La Salle - Universitat Ramon Llull
Gerard Serra
La Salle - Universitat Ramon Llull
Alberto Soto
La Salle - Universitat Ramon Llull
Hans Op de Beeck
KU Leuven
Haemy Lee Masson
Johns Hopkins University
Xavier Sevillano
La Salle - Universitat Ramon Llull

Research Article

Keywords: Cross-modal, environment, ACS, Molyneux

Posted Date: July 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-754574/v1

Cross-modal environment self-adaptation during object recognition in artificial cognitive systems

David Miralles[1*], Guillem Garrofé[1], Carlota Parés[1], Alejandro González[1], Gerard Serra[1], Alberto Soto[1], Xavier Sevillano[1], Hand Op de Beeck[2], Haemy Lee Masson[3]

(1) Grup de recerca en Tecnologies Media, La Salle-Universitat Ramon Llull, Barcelona, Catalonia, Spain.
(2) Department of Brain and Cognition, Leuven Brain Institute, KU Leuven, Leuven, Belgium.
(3) Department of Cognitive Science, Johns Hopkins University, USA

Abstract

The cognitive connection between the senses of touch and vision is probably the best-known case of cross-modality. Recent discoveries suggest that the mapping between both senses is learned rather than innate. These evidences open the door to a dynamic cross-modality that allows individuals to adaptively develop within their environment. Mimicking this aspect of human learning, we propose a new cross-modal mechanism that allows artificial cognitive systems (ACS) to adapt quickly to unforeseen perceptual anomalies generated by the environment or by the system itself. In this context, visual recognition systems have advanced remarkably in recent years thanks to the creation of large-scale datasets together with the advent of deep learning algorithms. However, such advances have not occurred on the haptic mode, mainly due to the lack of two-handed dexterous datasets that allow learning systems to process the tactile information of human object exploration. This data imbalance limits the creation of synchronized multimodal datasets that would enable the development of cross-modality in ACS during object exploration. In this work, we use a multimodal dataset recently generated from tactile sensors placed on a collection of objects that capture haptic data from human manipulation, together with the corresponding visual counterpart. Using this data, we create a cross-modal learning transfer mechanism capable of detecting both sudden and permanent anomalies in the visual channel and still maintain visual object recognition performance by retraining the visual mode for a few minutes using haptic information. Here we show the importance of cross-modality in perceptual awareness and its ecological capabilities to self-adapt to different environments.

Introduction

Humans perceive the environment through multiple senses. A set of sensory information, acquired through each modality, is integrated and transformed into a supra-modal representation. This process requires cross-modality (also referred to as cross-modal transfer or cross-modal matching) – the cognitive ability to associate the sensory features acquired independently through multiple senses. Human manipulation of objects, a natural example of cross-modality, connects the senses of sight and touch from an early age, and this sensory connection is strengthened over the course of child development (1), and stays throughout the lifespan (2). In particular, vision and haptics are complementary to each other, improving the credibility of mental representation of object properties and recognition performance (3, 4, 5, 6). In this article, we present an Artificial Cognitive System (ACS) that builds on a cross-modality ability using human manipulation data, achieving perceptual awareness and a dynamic capacity to adapt to changing environments.

The question of how humans achieve cross-modality was sparked off by 17th-century natural philosopher William Molyneux. In his letter to John Locke, he questioned whether a congenitally blind person, who recently gained vision, would be able to visually recognize an object, previously known only by touch, or she/he would need to learn to make the intermodal transfer from touch to vision (7). To this day, the debate over whether this transfer ability is innate or acquired has led scientists to investigate cross-modality in newborns, animals, and congenitally blind individuals (8, 9, 10). Of these studies, a recent cross-modal matching experiment, conducted on congenitally blind individuals who later gained sight as adults, suggests that cross-modality is acquired, dynamic, and moldable rather than innate and predetermined (9). This dynamic nature of cross-modality allows human observers to modulate the strength of intermodal connection based on the reliability of the information derived from each modality

(5), an ability that has also been observed in various animal species, including capuchin monkeys (11), rodents (12), and even bumblebees (13).

In the past few years, there has been a growing interest in the development of cross-modality in artificial agents (14), especially robots, as it may facilitate the creation of systems that autonomously adapt to different environments. In particular, with the advancement of visuo-tactile sensors for haptic capture systems, researchers have started investigating the cross-modal connection between vision and touch in robotics (15, 16, 17). The tactile patterns that result from these sensors can be related to the images obtained using the visual mode, thus creating a framework that enables the establishment of cross-modal relationships. By gathering haptic data in the form of images, the dimensional gap between touch and vision features can be successfully overcome (18). Furthermore, the maturity of image recognition methods (e.g. deep learning algorithms) has allowed some progress in the study of the relationship between these two modes. However, existing haptic data is still a long way from what would be a haptic dexterous robot exploration similar to that of humans. To solve this problem, researchers have recently designed a glove that through a mechanoreceptor network can provide tactile patterns to the system, as well as information related to the dexterity of the human grasp (19). However, that work does not relate haptic data to visual mode. In this paper, we use a haptic capture system (20) that also leverages information from the human manipulation of objects, but with the ultimate goal of delving deeper into the design of cross-modality in ACS.

To achieve this objective, we designed and printed novel 3D objects that collect human exploration data with multiple capacitive touch sensors on the object surface. With this dataset, our ACS achieves cross-modality via transfer learning from touch to vision. Unlike other approaches (17), where corrupted inputs are incorporated during training time, we present a new mechanism that allows our ACS to use cross-modality to continuously monitor whether the information received from visual and haptic modes matches using cross-modality, hence being able to detect anomalies (e.g. blurred vision). Given that the two sensory modalities are independent but collaborative, like those of a human, we examine how our system dynamically changes the strength of the intermodal connection to better solve object recognition problems when we degrade the quality of visual information at test time. If mismatches between vision and touch channels persist over time, the ACS can autonomously retrain the faulty modality through transfer knowledge within few minutes. Our findings suggest that with the implantation of biologically inspired cross-modality, the ACS becomes perceptually aware of a faulty sensory modality and autonomously adapts to changing environments without losing performance.

Results

Haptic data and object recognition

We designed a system that captures haptic information generated by humans during the object manipulation process. This collected data is enough to create a haptic recognition system that outperforms humans in a classification task with similar 3D shapes (20), both in accuracy and response time. As illustrated in Fig. 1, the objects are six similar shapes we have digitally created and 3D printed (20, 21). The external surface of each object is totally covered with 24 copper pads equally distributed and connected to an electronic board placed inside, which also includes a gyroscope. During the object manipulation, this system samples data from all the sensors at 40Hz and sends it to a computer through wireless communication. Every sample is stored as a 24-bit array, h_j, called haptic state of the j sample, one bit for each copper pad (Fig. 1B), hence the system has no information about the relation between the location of the 24 sensors and the positions of their statuses in the array. Besides, since each sensor has its position inside the array, the resulting state would vary if sensors were placed differently. For this reason, and to make sure the presented algorithm does not use the sensors' order to recognize the objects, sensors

are placed in a way that every position in the haptic state corresponds to a sensor located in the same spatial location for every object. In the same way that touch receptors on the human hand would also receive input from corresponding locations on different objects when the relative orientation of the grasping hand and the objects would be held constant.

Our haptic dataset is based on these haptic states and their time evolution. The geometry of the objects affects their handling, and this is reflected in our data. To gather our haptic dataset, one participant is invited to manipulate each object with both hands and perform a random exploration task. Four series, each lasting for five minutes, have been recorded per object.

To perform automatic haptic object recognition, the dataset is divided into two parts: three series for training (15 minutes) and one for testing (5 minutes). To determine the probability of a set of n consecutive haptic states $(h_1, \dots, h_n)$ belonging to a specific object S_i ($i \in [1, \dots, 6]$), i.e., $P(S_i \mid h_1, \dots, h_n)$, we adopt a naïve Bayes approach as follows:

$$\widehat{S_\iota} = \underset{S_i,\ i \in [1,\dots,6]}{\operatorname{argmax}} \left[P(S_i) \prod_{j=1}^{n} P(h_j \mid S_i) \right]$$

Here the n-product $P(h_j \mid S_i)$ is the naïve condition, $P(S_i) = \frac{1}{6}$ is the probability of each object and $\widehat{S_\iota}$ the resulting prediction.

Our haptic object recognition system achieves an average accuracy of 89.63% after just 8 seconds of manipulation (average time for best accuracy in humans (20)), as shown in Fig. 1.

Multimodal dataset generation and visual object recognition

As stated earlier, the 3D printed objects have been produced from a 3D digital render. Moreover, from the data collected by the gyroscope located inside the objects and using the 3D renders, we synthesize a video with the movements of the objects caused by human manipulation (Fig. 2). We create one video frame for each haptic state; see Methods for details. We select the 5-minutes test data series mentioned above for this purpose. This opens the visual channel to our system. Now, the ACS can receive data from haptic and visual senses simultaneously (Fig. 2).

Since the ACS has previously had a haptic experience, it can autonomously tag the visual dataset through the results of the already trained haptic object recognition system, creating a cross-modal relationship. Obviously, after this transfer of knowledge, we can train a visual object recognition system from the new labelled visual dataset. We divide the 5-minute visual dataset into three parts: 60% for training, 20% for testing, and reserve the remaining 20% for a later experiment. Using this data, we train a set of convolutional neural networks (CNN) for visual object recognition. The visual object recognition system is based on a one-vs-all strategy combining six CNN classifiers, thus yielding a six-dimensional output, $v \in \mathbb{R}^6$. Each component of the output, v_i, is associated to the probability that the visual input corresponds to one of the 6 stimuli S_i (objects). To decide which is the corresponding object, the following criterion is adopted: a sample belongs to a certain class if only one of the values v_i exceeds a threshold, τ, ($\exists!\, v_i$ s.t. $v_i > \tau$). The value of τ is obtained from the relationship between the accuracy of the visual classifier and the different threshold values as shown in Fig. 3. Two other results can be obtained: a) Confusion (CF): there is more than one v_i value above the threshold ($\exists v_i, v_j$ s.t. $v_i > \tau$ and $v_j > \tau$ with $i \neq j$), that is, the visual object recognition system assigns the sample to more than one class. b) Ignorance (IG): there is no value of v_i above the threshold ($v_i < \tau, \forall i$), i.e., there is a lack of knowledge to decide which class it belongs to. Although there are several efficient methods for dealing with confusion, especially based on heuristics, at this stage we have preferred to include it as a false negative, even at the expense of visual object recognition performance. We will resolve the confusion with the help of the haptic mode, thus allowing true cross-modality, as we will see in the next section.

Visual object recognition results shown in Fig. 3A demonstrate that the visual channel of our ACS achieves 75% accuracy including CF and IG as false negatives.

Stressing the visual channel and the Molyneux mechanism

The Molyneux problem addresses the following question: would a person born blind that later regains sight as an adult, be able to visually recognize the shapes of objects previously experienced by touch? Recent empirical studies have pointed out that upon recovery of sight, subjects are initially unable to recognize these objects visually. However, after they experience the world with both senses, in a few days a cross-modal link is created allowing them to pass the Molyneux test (9). In the present study, this connection between the two senses equips the ACS with a cognitive mechanism that allows it to autonomously detect a faulty channel.

The aforementioned mechanism, which we have called the Molyneux mechanism, allows the ACS to continuously check if what it is seeing agrees with what it is touching. Hence, the ACS can detect sudden anomalies by comparing the classification labels of the two independent channels, haptic and visual (Fig. 3B).

More specifically, the ACS can encounter four different situations while comparing the labels given by the haptic and visual object recognition systems: i) both recognition systems agree, i.e., there is a match, ii) do not agree, i.e., there is a mismatch, iii) the visual object recognition system does not have enough knowledge to classify that sample (IG), which results in a mismatch, and iv) the visual channel object recognition system assigns the sample to more than one class (CF) and the ACS checks if one of this classes agrees with the haptic recognition (match) or, on the contrary, it does not (mismatch). It is worth noting that when both recognition systems are working properly, some short duration mismatches can appear, but the common situation is a continuous agreement between the visual and the haptic recognition systems. As these mismatches are short in time, they can be easily removed with a low pass filter; see Methods for details.

In order to test this mechanism and study the effectiveness of artificial cross-modality for perceptual awareness, we stressed our visual channel by applying a blur filter. Once applied, the accuracy of the visual object recognition system went down to approximately 20% for the 6 classes (see Fig. 3A). The ACS detected this anomaly using the Molyneux mechanism, and did it quickly, with an average delay of only 2.13 seconds (85.33 samples) after applying the blur filter to the visual input. In Fig. 4 we show the results for each object.

Self-adaptation to a new environment

The Molyneux mechanism enables ACS to realize, in a fully automatic way, that the haptic and visual object recognition systems lose coherence when a blur filter is applied to the visual channel. This situation does not affect the haptic classifier, which remains stable and consistently gives trustworthy information to the visual channel. Thus, the blurred images are tagged with the output of the haptic classifier, and the ACS can retrain the CNNs of the visual object recognition system in order to classify them correctly. Using the 20% (60 seconds) of the visual dataset that we had previously reserved after going to a blur filter, we retrained the visual object recognition system by grouping the retraining samples into 8-second batches (total of 7 batches) and iteratively feeding the CNNs with one batch at a time. For each retraining batch iteration, we calculate the current accuracy of the visual object recognition system in the blurred vision scenario. As shown in Fig. 5, the accuracy in the blurred vision scenario increases as the ACS receives more blurred visual information in its retraining process.

Discussion

Visual object recognition and cross-modality

As described in the Results section, the visual mode benefits from the previous experience of the haptic mode. The haptic classifier recognizes the object being manipulated and the ACS tags the information of the visual channel in real-time. This would be equivalent to the process where for the first time a human becomes aware of the connection that exists between a known manipulated object and its visualization. It is important to highlight the autonomy that this cross-modal knowledge transfer gives to the ACS since it does not need an external agent to tag the visual channel (Fig. 2B).

On the other hand, heuristic arguments are commonly used to solve the confusion in the outputs of a one-vs-all visual classifiers. However, in our approach, cross-modality allows us to solve confusion through the haptic mode, i.e. through the interaction with the environment (object). In this study, if we use a highest value heuristic, our visual model reaches an accuracy of 79%, whereas if we use cross-modality to solve confusion, the accuracy goes up to 81% (see Fig. 5). Although in this case accuracies are similar, if the haptic classifier is stable, the cross-modality will always equal or improve the results of heuristic methods. This result resembles the exchange of multi-modal information to discriminate a stimulus in humans, which is a very common process (3).

Perceptual awareness

The Molyneux mechanism defined in this article allows the ACS to check the coherence between two synchronized samples (haptic-visual) (Fig. 3B). In other words, this mechanism allows the ACS to answer the question: is what I touch and what I see the same object? As shown in Fig. 4, in the Results and also in Methods section, the study of these anomalies has been detailed to differentiate them from those that lengthen in time. Filtering visual-haptic classification pairs in real-time allows the ACS to realize that the visual object recognition system is not working properly when the blur filter is applied. High and stable accuracy over time of the haptic classifier is assumed for this study as shown in Fig. 1. The goal of the filtering process is to identify changes in the visual channel. It can be observed (see Fig. 4) that for objects lat00 and lon05 the change to blurred vision is not detected. This is because the visual object recognition system continues to classify these two objects correctly despite the blur, and therefore there is no incoherence of any kind other than a decrease in visual accuracies for these two objects, i.e., the filter does not detect any change in the visual channel.

Resilience from cross-modality

Realizing that the environment has changed is the first step in the process of self-adapting to it. The proposed design with two independent object recognition systems and the cross-modality approach allows the ACS to autonomously adapt to changes in the environment that affect one of its sensory modes using the information of the mode that remains stable. The results from Fig. 5 show how after just one minute of retraining the visual classifier (CNN), the ACS adapts to the blurred scenario with an accuracy of 69.1%

It is important to highlight the ecological character of the transfer of knowledge between modes in the sense proposed in (9), since ACSs could adapt to transformations or changes in their perception systems during their lifespan.

Limitations

With the approach proposed in this article, we are aware that we are simplifying the problem by using synthetic images and avoiding the occlusions caused by human hands during object manipulation. Nonetheless, even though this work is not focused in solving this issue, this apparent problem could be part of the solution, since these occlusions are strongly correlated with haptic data.

Another limitation of this approach is that, in this very first experiment where we have shown the benefits of cross-modality for self-adaptation in a changing environment where suddenly the vision channel gets blurred, we have not accounted for other errors that may cause the vision channel to stop working properly. Moreover, we have used single frames instead of a group of states in the visual model. Our goal with the presented experiment was to test the Molyneux mechanism and show the benefits of cross-modality for environment self-adaptation for object recognition tasks. We believe this is the very first step towards the development of perceptual awareness in ACSs for adaptation to changes in their perception systems, and further studies are required.

Opportunities for future research

Although the current trend is to place touch sensors on robotic hands' end-effectors, the use of sensors on objects is an equally important field of research, especially in obtaining data for ACS. In fact, it seems reasonable to assume there would be a correlation between what was obtained by the introduced haptic capture system and the point cloud that a robotic hand could generate if it could interact with that system. We hypothesize that placing sensors directly on objects is equivalent to obtaining data from a human-like robotic dexterous hands. This would allow us to integrate our ACS to a robotic hand such as Shadow Hand (22). Other researchers have showed that this hand (23) could exhibit high levels of dexterity in object manipulation tasks.

It would also be interesting to study other ways to stress the input channels. Two situations of special interest are: a) the inducement of errors in the haptic channel to study cross-modality in the opposite direction, b) the desynchronization between the haptic and visual channels.

Finally, to extend this work to unfamiliar objects, a deeper study of the Molyneux mechanism is necessary. To this end, we propose an analysis of the haptic and visual perceptual spaces used in neuroscience (24), which would allow us to understand how we can relate unknown objects with a previous experience of the ACS.

Methods

In this section, we provide the methods and procedures used in this research article.

Visual dataset generation

To generate the 3D renders, the six similar shapes meshes are placed independently on a Processing 3D scene with a 299x299 window to ensure generating square images during the rendering process. To this end, we use a dataset consisting of four files per object (20 minutes in total). Then, an independent rendering process has been performed for each of these files matching with our haptic dataset. In each scene, the object of study is initially positioned in the world origin (0,0,0), which is the central point of

view of a camera that remains static during the whole process. This camera position is the same for all the experiments.

Once the environment is set up, each object is texturized uniformly generating a UV Map with the same colors on each side of the shape. Using this approach, each side can be easily identified despite the symmetries present in all the shapes. Then, the data from the gyroscope collected during objects' human manipulation is used to perform rotations on the object matching the ones from the human manipulation experiment. Each haptic state has associated four values corresponding to each component of a quaternion that are used to perform each rotation, considering the center of the figure the origin of the rotation. This approach recreates the objects´ original movements since the gyroscope in the 3D printed shapes is placed inside the object on its center.

This rotation is performed for each haptic state and rendered from the world camera generating images that capture each current object rotation forming the visual dataset. This synchronization between the visual and the haptic dataset is what makes the experiments presented in this article possible.

Blurred images generation

In order to simulate a sudden loss of visual channel, we generate blurred images for all samples. These blurred images are generated by applying over the original images an average filter, convoluting the image with a 20x20 normalized box filter from OpenCV libraries.

Visual object recognition system and training process

In order to generate the visual recognition system, we follow a one-vs-all strategy. This strategy generates an independent visual classifier for each one of the six classes ($Model_{[c]}$ where c defines the class). $Model_{[c]}$ classifies a single image as belonging to class c or not. We adapt a pre-trained CNN model based on InceptionV3 architecture (25), and we change the last layer for a fully connected dense layer with 2 outputs, using a softmax activation function. Each of the $Model_{[c]}$classifiers is trained using categorical cross-entropy as loss function and a dropout of 0.4. The training is performed for 5 epochs using 60/20 split of our own dataset for training/test the models (the remaining 20 percent is reserved for the blur filter test). Samples of our dataset are split in two classes for each of the Model[c]: i) positive class: samples belonging to class c, and ii) negative class: samples that not belong to class c. Pre-trained initial models are initialized using the weights of using ImageNet dataset (26).

Visual retraining process

By using the remaining 20% of samples that were not used in the model training/testing process (which amount to around 60 seconds), each of the $Model_{[c]}$ classifiers is retrained using exactly the same parameters used for training the original $Model_{[c]}$. The difference now is that the initial CNN weights are the ones obtained after the previous training process. This retraining process is performed using one batch (8 seconds of samples) at a time, and is repeated sequentially up to 7 times, as the samples are grouped in 7 batches. This retraining process simulates the gradual adaptation of the visual model to new conditions, that in this work we simulate by a sudden loss of vision resulting in an input of blurred images. By using the presented gradual retraining process, the model adapts to these new visual conditions for the blurred vision scenario.

Molyneux mechanism and filtering

The haptic and visual classifiers constantly classify the haptic states and visual frames of the figures that are acquired at a 40 Hz frequency. This means that every 25 ms, by applying the Molyneux mechanism, the haptic classification is compared with the visual classification, checking if i) both channels are in agreement, and ii) there is a failure in one of the channels or not.

It is normal that some short duration mismatches between both channels appear although both channels are working properly. In order to provide a stable decision regardless of whether a channel is failing or not, a low pass filter is applied to the output of the Molyneux mechanism.

The low pass filter consists in a 6th order Butterworth filter offering a flat output for the passband frequencies and avoiding ripples. The first 1000 samples of each figure test file are used to study the duration of the mismatches when there is no failure in neither of the two channels, obtaining a mean duration of $\mu = 3.3$ samples and a standard deviation of $\sigma = 9.1$. Since it is desired to achieve a huge attenuation for the mismatches frequencies, the cutoff frequency is set a decade before the frequency corresponding to the $\mu + \sigma$ duration. Taking into account the sample rate of 40 Hz, the cutoff frequency can be calculated as:

$$f_{cutoff} = \frac{f_s}{10 \cdot (\mu + \sigma)} = 0.32\ Hz$$

After applying the filter, the output of the Molyneux mechanism fluctuates between 0 and 1 (see the orange plot in Fig. 4), where 0 corresponds to channels not matching (failure in one channel) and 1 to channels matching. In order to offer a binary response, as the one show in Fig. 4, the filter output goes through an hysteresis cycle, where the output goes from 0 to 1 if the input is higher than 0.8 and from 1 to 0 if the input is lower than 0.2.

References and Notes

1. Purpura, G., Cioni, G., & Tinelli, F. Development of visuo-haptic transfer for object recognition in typical preschool and school-aged children, *Child Neuropsychology* **24**, 657 (2018).
2. Norman, J. F., et al. Aging and the visual, haptic, and cross-modal perception of natural object shape, *Perception*, **35**(10), 1383-1395 (2006)
3. Gibson, J. J. *The ecological approach to visual perception* (Taylor and Francis, 1979).
4. Gepshtein, S., & Banks, M. S. Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, **13**(6), 483-488 (2003)
5. Ernst, M. O., & Bülthoff, H. H. Merging the senses into a robust percept. *Trends in cognitive sciences*, **8**(4), 162-169 (2004)
6. Lederman, S. J., & Klatzky, R. L. Haptic perception: A tutorial, *Attention, Perception, & Psychophysics* **71**, 1439 (2009).
7. Locke, J. *An essay concerning human understanding,* (Hackett Publising Company, 1996, original 1689).
8. Cheselden, W. VII. An account of some observations made by a young gentleman, who was born blind, or lost his sight so early, that he had no remembrance of ever having seen, and was couch'd between 13 and 14 Years of age. *Philosophical Transactions of the Royal Society of London*, *35*(402), 447-450 (1728)
9. Held, R. et al. The newly sighted fail to match seen with felt, *Nature neuroscience* **14**, 551 (2011).
10. Lewkowicz, D. J., & Lickliter, R. (Eds.). *The development of intersensory perception: Comparative perspectives*. Psychology Press. (2013)
11. Carducci, P., Squillace, V., Manzi, G., & Truppa, V. Touch improves visual discrimination of object features in capuchin monkeys (Sapajus spp.). *Behavioural processes*, **172**, 104044 (2020)
12. Hu, X., Urhie, O., Chang, K., Hostetler, R., & Agmon, A. A novel method for training mice in visuo-tactile 3-D object discrimination and recognition. *Frontiers in behavioral neuroscience*, **12**, 274 (2018)
13. Solvi, C., Al-Khudhairy, S. G., & Chittka, L. Bumble bees display cross-modal object recognition between visual and tactile senses. *Science*, **367**(6480), 910-912 (2020)
14. Billard, A. & Kragic, D. Trends and challenges in robot manipulation, *Science* **364** (2019).
15. Lin, J., Calandra, R., & Levine, S. Learning to identify object instances by touch: Tactile recognition via multimodal matching. *IEEE International Conference on Robotics and Automation,* 3644-3650 (2019)
16. Falco, P., Lu, S., Natale, C., Pirozzi, S., & Lee, D. Connecting touch and vision via cross-modal prediction, *IEEE Transactions on Robotics* **35**, 987 (2019).

17. Lee, M. A., Tan, M., Zhu, Y., & Bohg, J. Detect, Reject, Correct: Crossmodal Compensation of Corrupted Sensors. Preprint at https://arxiv.org/abs/2012.00201 (2020)
18. Li, Y., Zhu, J. Y., Tedrake, R. & Torralba, A. Connecting touch and vision via cross-modal prediction, *IEEE Conference on Computer Vision and Pattern Recognition* 10609-10618 (2019).
19. Sundaram, S. et al. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, **569**(7758), 698-702 (2019)
20. Miralles, D. et al. Artificial haptic recognition through human manipulation of objects, *Conference on Cognitive Computational Neuroscience*, (2019).
21. Gielis, J. A generic geometric transformation that unifies a wide range of natural and abstract shapes, *American journal of botany,* **90**, 333 (2003).
22. Walker R., Shadow dexterous hand technical specification, *Shadow Robot Company* (2005).
23. Andrychowicz, O. M., et al., Learning dexterous in-hand manipulation, *The International Journal of Robotics Research* **39**, 3 (2020).
24. Masson, H. L., Bulthé, J., De Beeck, H. P. O., & Wallraven C., Visual and haptic shape processing in the human brain: unisensory processing, multisensory convergence, and top-down influences, *Cerebral Cortex* **26**, 3402 (2016).
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision, *IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826 (2016)
26. Deng, W., et al. ImageNet: A Large-Scale Hierarchical Image Database, *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009)

Figures and Tables

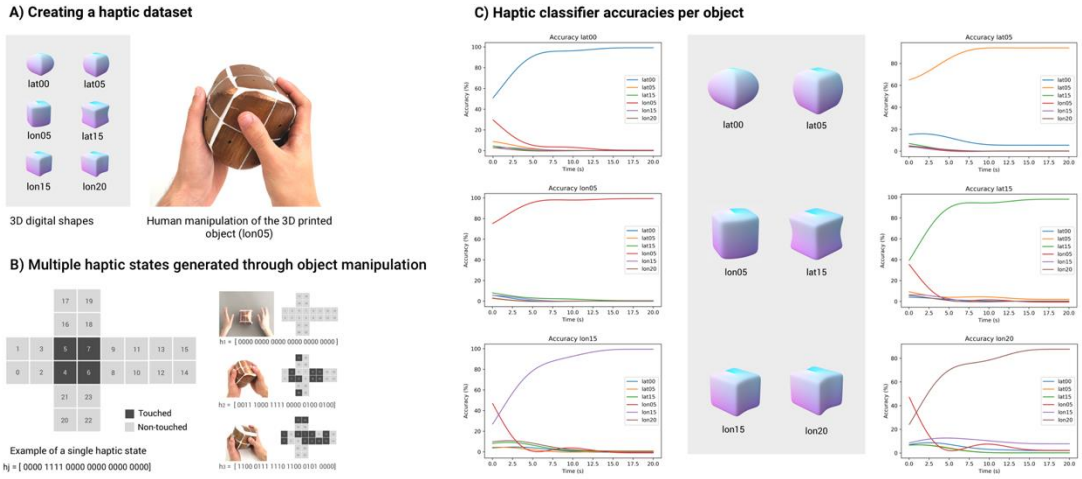

Fig. 1. From object human manipulation to automatic haptic classification. (**A**) The objects of this study are six 3D printed shapes labelled as lat00, lat05, lat15, lon05, lon15, and lon20. The participant sits in front of a computer and follows instructions on how to manipulate the objects randomly. (**B**) The human manipulation data of the objects are collected by the system and are stored as haptic states. A haptic state is represented by a 24-bit array and indicates the status of each sensor (touch/not touched) forty times per second. (**C**) ACS recognizes each object through a simple Bayes algorithm based on haptic states. As show in this subfigure, accuracies higher than 80% are consistently reached after a few seconds.

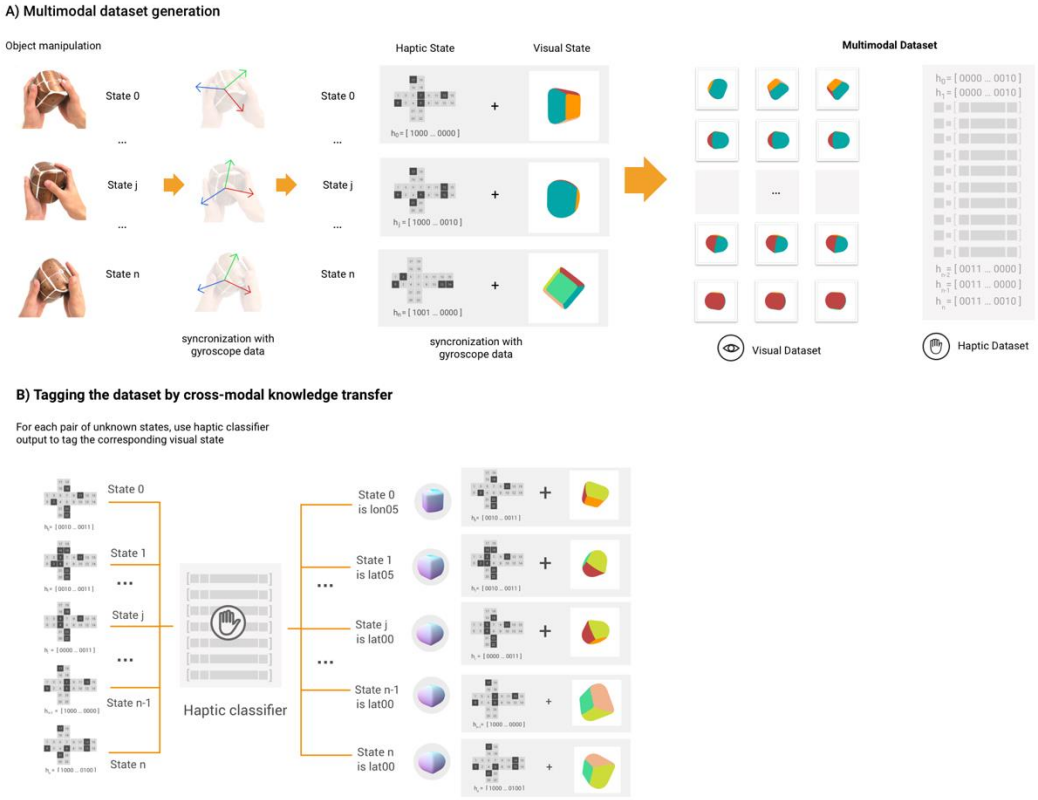

Fig. 2. Multimodal dataset generation through cross-modality. (**A**) Using the gyroscope, we can obtain the orientation of the object associated to each haptic state. From this information, we can draw (through 3D renders) a visual state of the object and associate it with its corresponding haptic state. From this, given any haptic dataset we can generate its corresponding visual dataset. With this method a new multi-modal dataset is created as a result of the combination of haptic and visual datasets. (**B**) Given that no manual annotation process has been carried out in the creation of the previous multi-modal dataset, the visual mode is not labelled. Here, we propose to tag the visual dataset from the results of the haptic object recognition system. This transfer of information generates a link between the two modes, visual and haptic, called cross-modality.

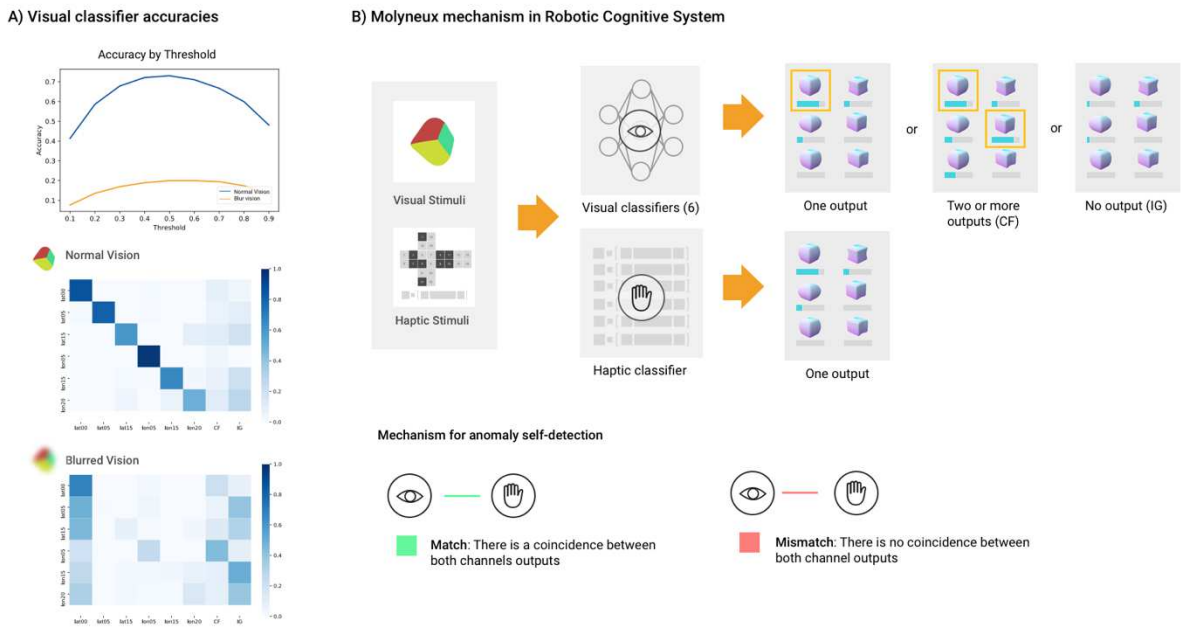

Fig. 3. Visual classifier threshold and the Molyneux mechanism. (**A**) With the synthesized video generated from human object manipulation, we trained a one-vs-all based CNN model, generating an individual classifier for each class. In order to determine the output label of a classifier, we have studied which accuracy threshold results in the best performance of the visual model. By setting the threshold to 0.5, we obtain the confusion matrices shown at the bottom after testing the model with normal vision images (like the ones used for training) and blurred vision images. (**B**) Every video frame and its corresponding haptic state are classified by the visual and haptic object recognition systems, respectively. The output of the visual recognition system can be i) a single class, ii) multiple classes (CF), or iii) none (IG). On the other hand, the label from the haptic recognition system is always univocal. By applying the Molyneux mechanism, we check if visual and haptic recognition agree, which we call a match. Otherwise, we call this disagreement a mismatch. In the case the visual recognition result is CF (ii), it will be a match if one of the possible classes agrees with the haptic classification.

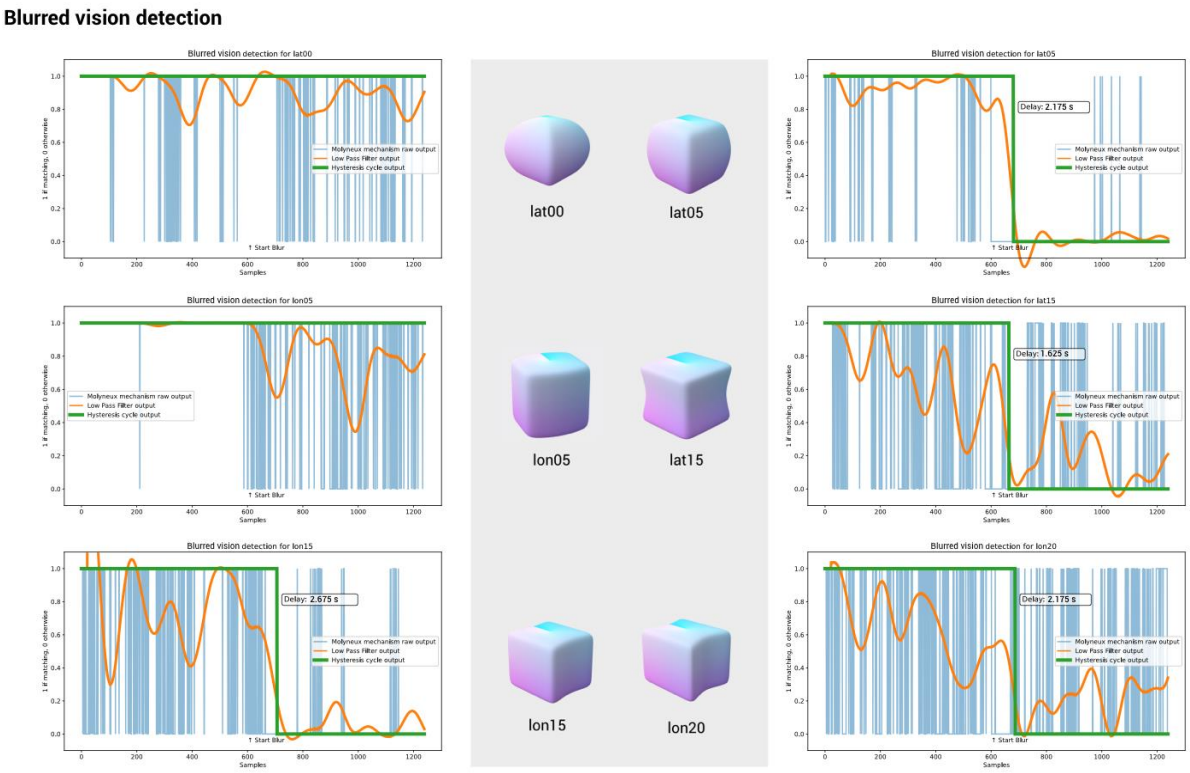

Fig. 4. Blurred vision detection through Molyneux mechanism and filtering. By applying the Molyneux mechanism to every visual-haptic pair, the ACS can determine if the two channels are matching or not, hence it can detect failures in one channel. Although there is a clear trend towards the two channels matching when both are working properly, short mismatches can appear (blue plot). In order to obtain an accurate decision whether the channels are matching or not, a low pass filter is applied to attenuate these mismatches (orange plot). Finally, as a way to offer a binary response that decides if the vision is blurred or not, the filter output goes through a hysteresis cycle (green plot). With this entire process the detection of the blurred vision is not immediate and has an average delay of 2.13 seconds (85.33 samples). Note that there is no mismatch in lat00 and lon05, that is, for these two objects the blur filter does not affect the recognition system.

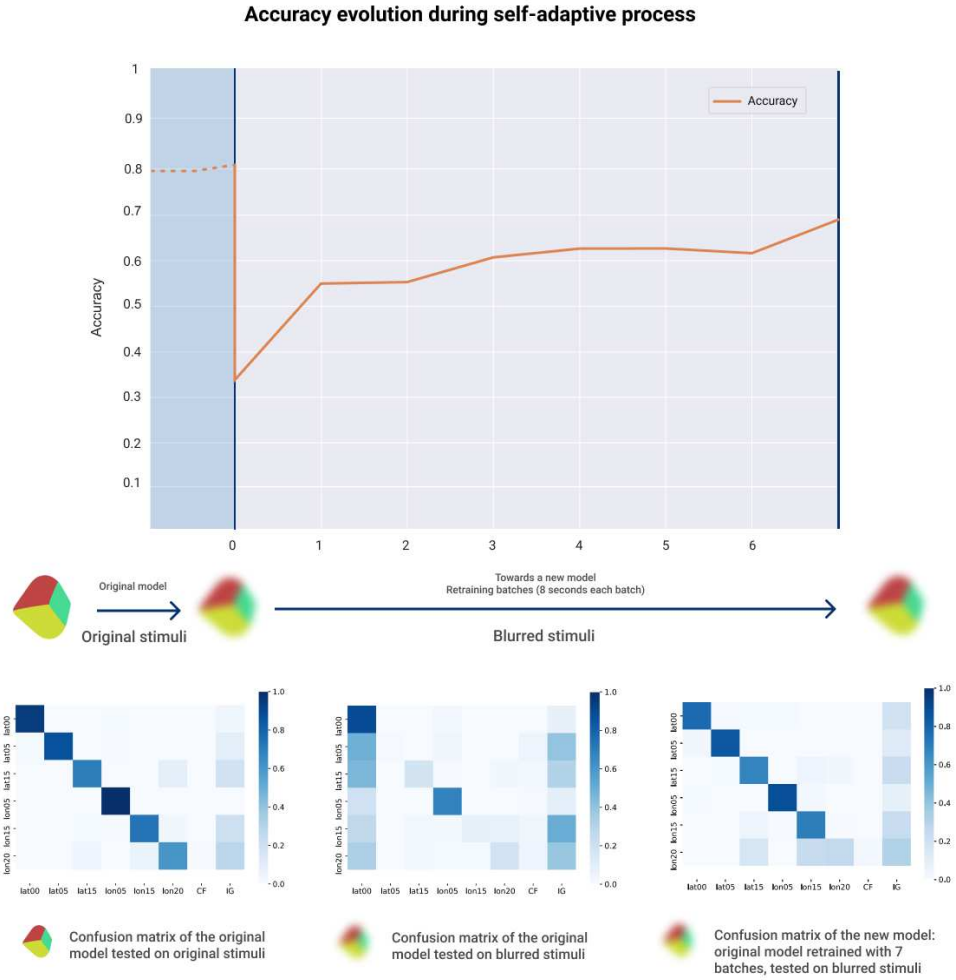

Fig. 5. Visual model adaptation to new stimuli acquisition conditions. Top plot shows the decrease in accuracy at the moment (t = 0) when blurred stimuli are first introduced. As blurred batches are incorporated to retrain the system, previous accuracy with original stimuli is now achieved with blurred stimuli (t=7). Bottom plots show the confusion matrices at the most relevant moments during this adaptation process: (Left) Original model tested with original stimuli, (Center) Original model tested with blurred stimuli (t=0), (Right) New model after the adaptation process with 7 batches of blurred images (t=7) tested with blurred stimuli.