HAL
open science

Visuo-Tactile Recognition of Daily-Life Objects Never Seen or Touched Before

Zineb Abderrahmane, Gowrishankar Ganesh, André Crosnier, Andrea Cherubini

To cite this version:

Zineb Abderrahmane, Gowrishankar Ganesh, André Crosnier, Andrea Cherubini. Visuo-Tactile Recognition of Daily-Life Objects Never Seen or Touched Before. ICARCV 2018 - 15th International Conference on Control, Automation, Robotics and Vision, Nov 2018, Singapore, Singapore. pp.1765-1770, 10.1109/ICARCV.2018.8581230 . hal-01869015

HAL Id: hal-01869015

https://hal.science/hal-01869015v1

Submitted on 6 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Visuo-Tactile Recognition of Daily-Life Objects Never Seen or Touched Before

Zineb Abderrahmane, Gowrishankar Ganesh, André Crosnier and Andrea Cherubini

***Abstract*— This study proposes a visuo-tactile Zero-Shot object recognition framework. The proposed framework recognizes a set of novel objects for which no tactile or visual training data are available. It uses visuo-tactile training data collected from known objects to recognize the novel ones, given their attributes. This framework extends the haptic Zero-Shot Learning framework that we proposed in [1] with vision, which enables a multimodal recognition system. In our test with the PHAC-2 dataset, the system was able to get a recognition accuracy of 72% among 6 objects that were never touched or seen during the training phase.**

I. INTRODUCTION

Object recognition is an important ability, and a fundamental pre-requisite for many of the cognitive and social abilities of robots. Most state of the art recognition approaches are based on multi-class classification. They classify any encountered object as one of the previously experienced objects during the training phase, and never as a novel object, the robot has not been previously trained on. However, there are many objects that the robot can encounter in real life, and training the robot on all of them is infeasible. This is due to the effort and time required for collecting sensory training data from each one of them. Thus, the robot is usually trained on a limited set of objects and encountering novel objects is very often. This makes important to recognize these latter without collecting training data from any of them.

Information from multiple senses can be used for object recognition; the most prominent ones for this task are vision and touch. Artificial object recognition systems use either vision [2] or touch [3], but less frequently both together [4]. In this work, we provide probably the first visuo-tactile recognition system that can handle novel objects, i.e. objects that have neither been seen or touched during the training phase.

We cope with novel objects by so called Zero-Shot Learning (ZSL). A Zero-Shot Learning system generalizes the models learned from objects seen and/or touched during the training phase, to recognize novel ones. This can be done by describing training and novel objects using attributes, which are semantic properties a human can use to describe each object (e.g. round, soft and bumpy). Then, attributes are used to recognize novel objects based on visuo-tactile data collected from training ones. While tactile data are relevant to perceive the object's material, texture and compliance properties, the addition of vision can improve the performance by perceiving properties such as shape and color. In this paper, we suggest an attribute-based framework that enables visuo-tactile ZSL. To the best of our knowledge, several studies have been carried out on visual ZSL [5], only one on haptic (tactile and shape) ZSL [1], and there is no studies on visuo-tactile ZSL. A recent study suggested a hybrid system by combining visual and tactile data, but they performed tactile ZSL with trained visual features [6]. Here, we exploited the PHAC-2 dataset [7], that provides both haptic and visual data for 60 daily-life objects (see examples in Fig. 1).

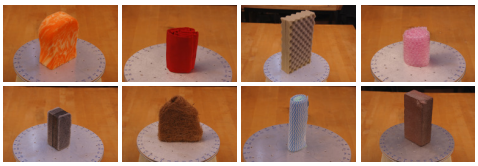

Fig. 1. Example images taken for PHAC-2 objects [7].

We improved our haptic ZSL framework proposed in [1] as follows: First, we improved attributes learning by replacing hand-crafted feature extractors with deep Convolutional Neural Networks (CNNs). CNNs were used to classify attributes based on both tactile and visual data. Second, we adapted the Direct Attributes Prediction (DAP) model used in [1] to take into account both visual and tactile modalities. For recognizing a novel object, three scenarios were investigated and compared: (1) the use of vision only, by assuming that the object cannot be touched, e.g. it is far from the robot, (2), the use of touch only, e.g. in case the robot operates in the dark, and (3), the use of both vision and touch. Finally, we improved the visuo-tactile ZSL by extending the PHAC-2 attribute set by adding visual ones capturing more object properties that cannot be felt using touch.

This paper is organized as follows. Sect. II presents related work on visual and tactile ZSL. In Sect. III, we present the theoretical framework of attribute-based ZSL. Next, we propose our solution for integrating visual and tactile modalities in Sect. IV. We then present our experimental setup and experimental evaluation in sections V and VI respectively. Finally, in Sect. VII, conclusions are provided.

II. RELATED WORK

A. Visual Zero-Shot Learning

Although many broad image datasets are available for object recognition (such as ImageNet [8]), image labeling for all possible classes is still intractable. This justifies the great attention gained by ZSL in visual recognition. Lampert et al. [9] designed the first attribute-based ZSL system for recognizing novel animal classes. By describing

Zineb Abderrahmane is supported by the Ministry of Higher Education and Scientific Research of Algeria through the Excellence Fellowship

All authors are with LIRMM, Universit de Montpellier, CNRS, Montpellier, France firstname.lastname@lirmm.fr

Gowrishankar Ganesh is also with CNRS-AIST JRL UMI3218/RL 1-1-1 Umezono, 305-8560 Tsukuba g.ganesh@aist.go.jp

animals using attributes (e.g. furry, small and tail), they used available images to train a classifier per attribute. Then, these classifiers were used to classify novel classes, solely given their attribute-based description. Further improvements of this framework have been proposed, by adapting it to large scale datasets [10], generalizing to real-valued attributes [11], [12], developing an online incremental approach [13], reducing human effort by automatically designing attributes [14], handling attributes unreliability [15], designing a hierarchical transfer model [16], and recovering missing class-attribute associations [17]. Another approach [18], [19] performs ZSL based on classes textual descriptions available on linguistic databases (e.g. Wikipedia). A third approach [20], [21], [22] classifies novel classes based on their direct or hierarchical relationships with training classes. Several approaches have been reviewed and compared in [5].

B. Tactile Zero-Shot Learning

Tactile recognition systems suffer not only from the difficulty of labeling data, but also from the difficulty of collecting them. Tactile data collection requires robot-object interaction, which is time consuming, especially because some sensors need a stable contact with the object's surface to obtain good-quality measures, e.g. [23] maintained the contact with the object for 20 seconds. Nevertheless, tactile ZSL has gained much less research attention than visual ZSL, which motivated us to propose a Zero-Shot haptic (tactile and shape) recognition system in [1].

In [1], we used the state of the art PHAC-2 dataset [7]. This dataset was used in multiple studies to improve robot haptic perception using the provided haptic attributes . Gao et al. [24] used deep learning for recognizing PHAC-2 attributes from haptic and visual data. The authors of [25] improved haptic perception by making use of attributes correlations and learned attributes in a multi-label setting instead of separately as in [24].

C. Visuo-Tactile Fusion

Many studies showed the efficiency of combining visual and tactile data for improving different robotic tasks. For instance, Gao et al [24] incorporated both modalities for understanding objects' properties, Ghanbari et al. [26] for assisting humans in cell injection task and Yamashiro et al. [27] for estimating friction properties. In a recent study, authors of [6] proposed a visuo-tactile dictionary learning for ZSL of the eight material categories that group PHAC-2 objects. They showed that incorporating both visual and tactile modalities is effective for performing ZSL. This encouraged us to extend our previous haptic ZSL framework [1] by adding vision. Since we use PHAC-2 objects which have simple and similar shapes, we do not consider kinesthetic (shape) data and we only take tactile data combined with visual data.

Our contribution w.r.t the state of the art is the design of a visuo-tactile ZSL framework which improves the haptic ZSL framework that we proposed in [1]. This work is different from [6] that uses visuo-tactile data for ZSL. They perform a material-based ZSL by considering the eight material classes, the PHAC-2 objects belong to, whereas, here, we perform an instance-based object recognition. In addition, they use visual data available for novel classes to perform tactile ZSL, whereas, here, we recognize novel objects having neither visual nor tactile data.

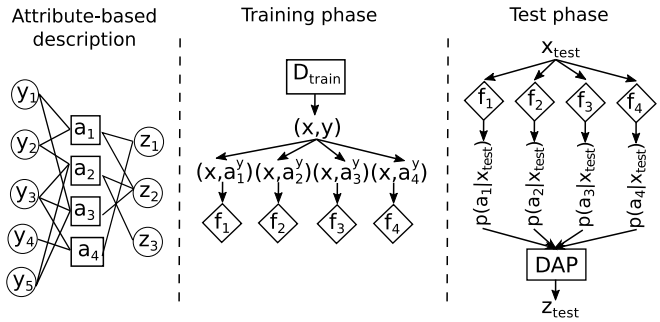

Fig. 2. Attribute-based ZSL (solution overview for N=5, M=4 and L=3): First, both Y and Z objects are described using A attributes. Then, a classifier f_m is learned for each attribute. Last, attributes classifiers are used by the DAP model to infer the object class.

III. Attribute-Based Zero-Shot Learning

Zero-Shot Learning (ZSL) is the problem of training and testing a recognition system on two disjoint sets. Let $Y = \{y_1, \ldots, y_N\}$ be the set of objects the robot has been trained on, giving training set $D_{train} \subset X \times Y$, where X is the feature space where collected sensory raw data are represented. During the test phase, the robot collects $x_{test} \in X$ by exploring an unknown object, which should be classified as one of the novel objects $Z = \{z_1, \ldots, z_L\}$. Since $Y \cap Z = \emptyset$, the robot has no training data for Z objects and needs an auxiliary information about objects to classify x_{test} as one of Z objects based on sensory training data collected from Y objects.

The solution applied in [1], illustrated in Fig. 2, is largely used by many studies. It consists in defining a set of attributes $A = \{a_1, \ldots, a_M\}$ (e.g. round, plastic, concave, etc. in [1]). Then, each object $o \in Y \cup Z$ is described using A. This associates o with a deterministic vector $\boldsymbol{a^o} = [a_1^o, \ldots, a_M^o]$, where for $m = 1, \ldots, M$: $a_m^o = 1$ if attribute a_m is a property present in object o (e.g. $a_m = concave$ for $o = cup$) and $a_m^o = 0$ otherwise. Authors of [9] proposed two models for using the attributes layer to make use of training data collected from Y to recognize Z objects. In [1], we chose the Direct attributes Prediction (DAP) model which is more popular and showed better performance in [9].

The DAP model uses D_{train} to learn a classifier per attribute. During the training phase, for each attribute $a_m \in A$, a probabilistic binary classifier $f_m : X \longrightarrow [0,1]$ is trained on $D_{train}^m = \{(x_i, a_m^{y_n}),\ s.t.\ (x_i, y_n) \in D_{train}\}$. During the test phase, the test sample x_{test} is input to each trained f_m which returns the posterior $f_m(x_{test}) = p(a_m \mid x_{test})$. It predicts the presence of attribute a_m in the object from which x_{test} was collected. Then, the posterior of each novel object $z_l \in Z$ is computed given $\boldsymbol{a^{z_l}}$ and all attributes posteriors as follows:

$$p(z_l \mid x_{test}) = \frac{p(z_l)}{p(\boldsymbol{a^{z_l}})} \prod_{m=1}^{M} p(a_m^{z_l} \mid x_{test}), \qquad (1)$$

By replacing object and attribute priors with a uniform distribution, the test sample x_{test} is classified as the object

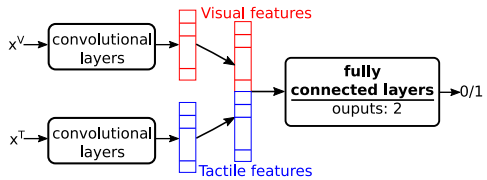

Fig. 3. Visuo-tactile CNN classifying an attribute as absent (0) or present (1) based on tactile and visual data.

having the highest posterior:

$$z_{test} = \underset{z_l \in Z}{\operatorname{argmax}}\, p(z_l \mid x_{test}). \tag{2}$$

IV. Visuo-Tactile Zero-Shot Learning

In this section, we adapt the DAP model to perform tactile, visual or visuo-tactile ZSL. A robot equipped with visual sensors can provide visual images in X^V about the object when it is in its field of view. In addition, by physically interacting with the object, the robot tactile sensors provide tactile data samples in X^T. X^V and X^T are feature spaces in which visual and tactile data are represented respectively. Thus, our aim is to adapt the framework presented in Sect. III for data samples $x = [x^V, x^T]$ in $X = X^V \times X^T$.

A. Attributes Learning

The first step is to learn from training data how to predict the presence of each attribute in an object, i.e. to compute $\{p(a_1 \mid x), \ldots, p(a_M \mid x)\}$ given $x = [x^V, x^T]$. We propose three solutions. The first solution learns a binary classifier per attribute that uses tactile data x^T only. The second one learns a binary classifier per attribute using visual data x^V only. The advantage of separating tactile and visual modalities for attributes prediction is that the system is operational even when only one sensor modality is available. The third solution learns one classifier per attribute using $x = [x^V, x^T]$. It makes use of both visual and tactile data to learn the attribute.

First, to learn a tactile classifier per attribute, we replace the hand-crafted feature extractor and the SVM classifier used in [1] by a CNN, which requires representing tactile signals in the form of a tactile image. By deriving a tactile image from x^T, the CNN automatically extracts discriminative features and predicts the presence of the attribute in the object at the same time. Likewise, we train a binary CNN per attribute that predicts its presence from visual images x^V .

The third solution classifies both tactile and visual data at the same time using one CNN. As illustrated in Fig. 3, we extract features from each modality separately using an independent convolutional part per modality. Then, we concatenate the tactile and visual CNN features to form one visuo-tactile feature vector that we classify using a fully connected neural network.

B. Visuo-Tactile DAP

The second step is to use outputs of attribute classifiers to compute posteriors of Z objects according to (1). This requires the prediction of all attributes posteriors $p(a_m \mid x_{test})$. However, choosing CNNs for attributes classification provides us with a classification score $s_m(x_{test}) \in \mathbb{R}$. To transform this score into a posterior probability, we use a sigmoid function as follows:

$$p(a_m \mid x) = 1/(1 + e^{-s_m(x)}). \tag{3}$$

Thus, for each attribute, we have three posteriors given the test sample $x_{test} = [x^V, x^T]$: $p(a_m \mid x^T)$ returned by the tactile CNN, $p(a_m \mid x^V)$ returned by the visual CNN and $p(a_m \mid x^V, x^T)$ returned by the visuo-tactile CNN. Thus, inferring $p(a_m \mid x_{test})$ used in (1) can be performed based on:

1) **Tactile data only:** by replacing $p(a_m \mid x_{test})$ in (1) with $p(a_m \mid x^T)$. We refer to this method as Tactile-ZSL (denoted T-ZSL);
2) **Visual data only:** by replacing $p(a_m \mid x_{test})$ in (1) with $p(a_m \mid x^V)$. We refer to this method as Visual-ZSL (denoted V-ZSL);
3) **Both tactile and visual data:** we proceed in two ways:
   a) by replacing $p(a_m \mid x_{test})$ in (1) with $p(a_m \mid x^V, x^T)$. We refer to this method as "Visuo-Tactile Features Concatenation ZSL" (denoted VT-FC-ZSL);
   b) by combining the two independent visual and tactile attributes posteriors to compute a visuo-tactile attribute posterior:

$$\begin{aligned} p(a_m \mid x_{test}) = {} & tact(a_m)\ p(a_m \mid x^T) \\ & + vis(a_m)\ p(a_m \mid x^V), \end{aligned} \tag{4}$$

   where $tact(a_m)$ and $vis(a_m)$ are user-tuned scores given to the importance of tactile and visual modalities for classifying attribute a_m respectively, s.t. $tact(a_m) + vis(a_m) = 1$. We refer to this method as "Visuo-Tactile Scores Merging ZSL" (denoted VT-SM-ZSL).

V. Database Description

A. PHAC-2 Dataset

In this work, we use the state of the art PHAC-2 dataset [7]. This dataset describes a set of 60 objects, having various texture, material and compliance properties, using 25 haptic attributes. We reduce them to 19 binary attributes (as in [1]): $A = \{absorbent, bumpy, compressible, cool, fuzzy, hard, hairy, metallic, porous, rough, scratchy, slippery, smooth, soft, solid, springy, squishy, textured, thick\}$.

Each object has been explored 10 times using the gripper of a PR2 robot equipped with 2 BioTac sensors. Each exploration trial consists of a sequence of 4 exploration steps: squeeze, hold, slow slide and fast slide. In addition, 8 images have been taken from different viewpoints for each of 53 objects.

B. Data Augmentation and Pre-processing

The available data are very few; only 10 tactile samples and 8 visual samples per object. This requires both tactile and visual data augmentation. As in [24], we augmented tactile data by combining data from both BioTacs and by sub-sampling the signals measured by each BioTac using five different starting points, resulting in 100 samples per

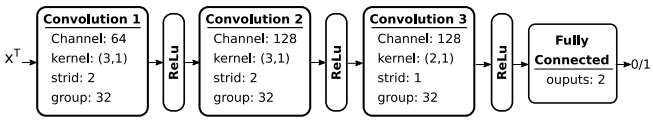

Fig. 4. Architecture of CNN for predicting attribute presence (1) or absence (0) based on tactile data.

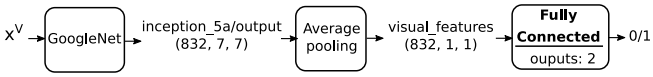

Fig. 5. Architecture of CNN for predicting attribute presence (1) or absence (0) based on visual data.

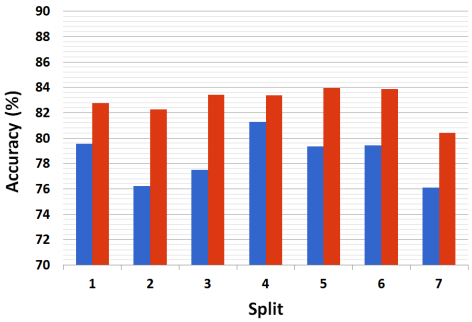

Fig. 6. Attribute binary classification accuracies for all object splits averaged over all attributes (method of [1] in blue and tactile CNN in red).

object instead of 10 (10 trials $\times$ 2 BioTacs $\times$ 5 starting points). Each sample includes 23 BioTac channels: static pressure, vibrations, temperature, heat flow and 19 electrode voltages. Principal component analysis (PCA) was applied to electrode voltages and the first four principal components were kept, giving 8 signals for each BioTac. Then, each of the 8 signals has been sub-sampled to 30 time samples. Next, by separating between the four exploration steps, the 8 signals of each exploration step were concatenated to obtain a 32-dimensional signal (4 exploration steps $\times$ 8 signals). Therefore, the space in which tactile data are represented is $X^T = \mathbb{R}^{32\times 30}$. On the other hand, 8 RGB images of resolution (224×224) yield a visual features space $X^V = \mathbb{R}^{3\times 224\times 224}$. This space was augmented by rotating each image multiple times and zooming in object's surface, resulting in 80 images per object instead of 8.

C. Objects Splits

By definition, performing ZSL requires the splitting of the object set into two disjoint sets: Y and Z. Since we aim at developing a visuo-tactile recognition system and visual data are not available for all of the 60 objects, we keep only the 53 objects for which visual data are available. We randomly select 6 objects ($\approx 10\%$) having different attributes vectors as Z objects, and the remaining objects as Y objects. We repeat the process 7 times in order to generate 7 random (Z,Y) splits to ensure the independence of the results from the choice of objects.

VI. Evaluation and Results

A. Implementation Choices

We have implemented our framework using Python based on [1], [28], [29]. CNNs were implemented using caffe [30]. The architecture of tactile CNN is the same as in [24] and is illustrated in Fig. 4. Fig. 5 illustrates the visual CNN architecture, we used a pre-trained model of GoogleNet [31] as a feature extractor for the visual data. We compared the BVLC [32] and MINC [33] GoogleNet pre-trained models and we found relatively similar results. Thus we chose the MINC model to have results comparable with [24]. Then, the extracted visual features are averaged and classified using a fully connected neural network that predicts the presence of the attribute. Finally, the convolutional parts of each of Fig. 4 and Fig. 5 are used to extract tactile and visual features for the visuo-tactile CNN illustrated in Fig. 3.

B. Attributes Learning

First of all, we focus on attributes learning using tactile data only. We compare our previous hand-crafted features extractor and SVM method [1] with the current deep classification method. In Fig. 6, we illustrate for each of the seven object splits defined in Sect. V-C the average classification accuracy of all attributes. We note that for all splits, CNN classification performs better than SVM with an average improvement of 4.37%. This shows the efficiency of deep learning in automatically extracting features and classifying them at the same time, compared to hand-crafted features and separate classification.

In Sect IV-A, we proposed three methods for learning attributes: based on tactile data only, based on visual data only and based on both visual and tactile CNN features. In Fig. 7, we compare the three classifiers of each attribute trained and tested based on split 1. We note that the performance changes from an attribute to another. Some attributes such as *bumpy, metallic* and *squishy* are better classified using tactile data. Some attributes such as *rough, springy* and *textured* are better classified using visual data. Others such as *absorbent, compressible* and *hard* are better classified using both tactile and visual data. Overall, 8 attributes are better classified using tactile data, 3 using visual data and 8 using visuo-tactile data. The fact that few attributes are better classified using vision only is obvious, since the attributes were defined in [7] to describe the haptic sensation of the objects. Besides, for 8 attributes, merging visual and tactile data improved learning compared to learning from each modality separately, which is promising for combining both modalities using VT-SM-ZSL.

C. Visuo-Tactile DAP

Here we present results of classifying a test sample $x_{test} = [x^V, x^T]$ as one of the 6 objects in Z. Knowing that we have zero training data for each of the 6 objects, classifying them with traditional classifiers gives an average classification accuracy of 16.67% which is equal to chance.

In table I, we compare DAP classification accuracies for classifying Z objects based on x^T only (T-ZSL), x^V only (V-ZSL) and $[x^V, x^T]$ (VT-FC-ZSL). Results show that most of splits are better classified with visuo-tactile data, some of them with tactile data only, and none of splits with visual data only. This was expected from results of attributes learning.

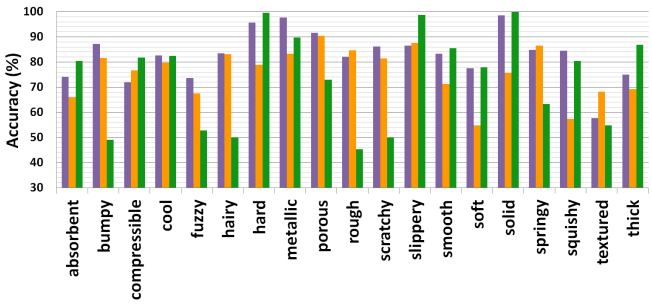

Fig. 7. Attributes classification accuracies for split 1: purple with tactile data alone, yellow with visual data alone, and green with both visual and tactile data.

TABLE I

COMPARISON OF TACTILE, VISUAL AND VISUO-TACTILE ZSL RECOGNITION ACCURACIES (%).

set	T-ZSL	V-ZSL	VT-FC-ZSL
1	60.5	31.46	**71**
2	40.5	46.25	**53.33**
3	43.67	54.17	**61.95**
4	**62.83**	37.71	56.28
5	41.83	28.13	**42.97**
6	**64.33**	33.96	57.73
7	33.33	35.21	**54.88**
average	49,57	38,13	**57,31**

However, we note that even though visual data alone are not very efficient for classifying objects, they efficiently improve tactile recognition in 5 out of 7 cases.

Motivated by the good results obtained with the concatenation of visual and tactile features using VT-FC-ZSL, we continue investigating another method for performing visuo-tactile DAP which is VT-SM-ZSL. According to (4), the importance of both modalities for classifying each attribute $tact(a_m)$ and $vis(a_m)$ should be estimated. For this, we compare in table II three methods of computing $tact(a_m)$ and $vis(a_m)$. The first *binary* method gives a binary importance $tact(a_m) = 1$ and $vis(a_m) = 0$ if the attribute classification accuracy using tactile data is better than using visual data, and $tact(a_m) = 0$, $vis(a_m) = 1$ otherwise. The second *weighted* method gives a real valued importance: $tact(a_m) = acc^t_m/(acc^t_m + acc^v_m)$ and $vis(a_m) = acc^v_m/(acc^t_m + acc^v_m)$ where acc^t_m, acc^v_m are respectively the classification accuracies of tactile CNN and of visual CNN trained on classifying attribute a_m and tested on validation data. The third, and simple *uniform* method assumes the same importance for both vision and tactile, i.e $tact(a_m) = vis(a_m) = 0.5$. Results show that the average accuracy of the binary method is greater than the other two methods. The elimination of the least performing modality for each attribute helped to perform visuo-tactile DAP by taking the best of each modality.

D. Adding Visual Attributes

In the experiments above, we used the haptic attributes provided with the PHAC-2 dataset to perform the ZSL. We obtained an improvement of 7,74% when adding visual features to tactile ones (see tables I and II). This motivated us to try adding more visual attributes to further improve the visuo-tactile ZSL.

TABLE II

VT-SM-ZSL RECOGNITION ACCURACIES (%).

set	binary	weighted	uniform
1	**55.27**	52.87	53.65
2	41.10	51.39	**51.79**
3	48.25	59.74	**61.13**
4	**57.08**	49.63	48.67
5	33.86	42.23	**42.34**
6	**49.98**	44.47	43.08
7	**51.53**	34.84	33.84
average	**48.15**	47.88	47,79

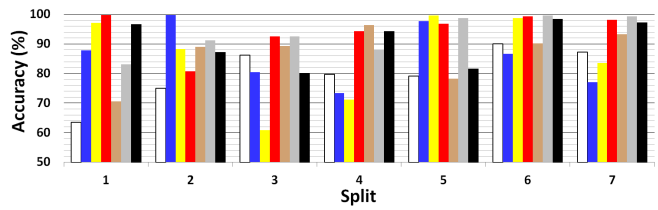

Fig. 8. Color attributes classification using visual images (bars are colored by the colors they represent).

While haptic attributes describe the texture, compliance and material properties, visual attributes can describe better the shape and color properties. Given that PHAC-2 objects have simple shapes with flat parallel sides, we assumed that adding visual attributes describing objects shapes would not be very effective. We therefore extended the attributes set with a set of color attributes. By observing objects' images, we defined a set of 7 colors shared between all objects, which are $A_c = \{white, blue, yellow, red, beige, silver, black\}$. Each object-color pair is associated with a binary value given by a human operator.

First, for each color attribute, we trained a visual CNN having the same architecture as for the haptic attributes (see Fig. 5). Fig. 8 illustrates the classification accuracies of color attributes CNNs, obtained for each split. We note that classification performance varies from an attribute to another and from one set to another. Overall, all colors have been classified with more than 60% accuracy.

Next, we improved the DAP classification results by extending the haptic attributes with color attributes. We first improved V-ZSL by classifying all the 26 attributes (haptic and color) using visual data only. Results (reported in the first column of table III) show the significant improvement of recognition accuracy for almost all object splits, compared to V-ZSL in table I. This highlights the effectiveness of adding visual attributes along with haptic ones. Only split 3 shows a degradation in terms of accuracy, but this is coherent with the fact that it has the lowest average attributes classification accuracy of 83.15% and the lowest accuracy of 60.83% for classifying attribute *yellow*. Furthermore, in table III, we added color attributes for VT-FC-ZSL and VT-SM-ZSL by giving $vis(a_c) = 1$ and $tact(a_c) = 0$ for all color attributes. Compared to tables I and II, this addition improved object classification for almost all splits, with an accuracy of 86% for split 6.

TABLE III

RECOGNITION ACCURACIES (%) WHEN ADDING COLOR ATTRIBUTES TO VISUAL AND VISUO-TACTILE ZSL.

set	V-ZSL+C	VT-FC-ZSL+C	VT-SM-ZSL+C
1	47.29	**77.48**	62.88
2	54.58	**66.67**	54.29
3	48.13	**59.58**	57.21
4	66.25	**75.1**	73.19
5	49.38	77.82	**80.38**
6	62.5	**77.82**	69.52
7	46.46	68.4	**86.27**
average	53,51	**71.74**	66.53

VII. CONCLUSION

In this paper, we proposed a visuo-tactile recognition framework, capable of recognizing novel daily-life objects based on their attribute-based description and without collecting any visuo-tactile data about them. We showed how replacing hand-crafted feature extraction with CNNs improved attributes learning (see Fig. 6). In addition, integrating visual data to tactile data (see tables I and II) significantly improved the Zero-Shot recognition accuracy. Finally, extending haptic attributes with visual ones improved the recognition performance (see table III). The obtained improvement consolidates previous studies results which highlighted the importance of visuo-tactile collaboration for improving robotic tasks.

REFERENCES

[1] Z. Abderrahmane, G. Ganesh, A. Crosnier, and A. Cherubini, "Haptic Zero-Shot Learning: Recognition of objects never touched before," *Robotics and Autonomous Systems*, vol. 105, pp. 11–25, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889017307492

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[3] A. Schmitz, Y. Bansho, K. Noda, H. Iwata, T. Ogata, and S. Sugano, "Tactile object recognition using deep learning and dropout," in *IEEE-RAS Int. Conf. on Humanoid Robots*, 2014, pp. 1044–1050.

[4] 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 Int. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 5273–5280.

[5] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning - the good, the bad and the ugly," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] H. Liu, F. Sun, B. Fang, and D. Guo, "Cross-modal zero-shot-learning for tactile object recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–9, 2018.

[7] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, "Robotic learning of haptic adjectives through physical interaction," *Robotics and Autonomous Systems (RAS)*, vol. 63, pp. 279–292, 2015.

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

[9] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 951–958.

[10] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," in *European Conf. on computer vision (ECCV)*. Springer, 2010, pp. 1–14.

[11] D. Parikh and K. Grauman, "Relative attributes," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2011, pp. 503–510.

[12] Y. Cheng, X. Qiao, X. Wang, and Q. Yu, "Random forest classifier for zero-shot learning based on relative attribute," *IEEE Trans. on neural networks and learning systems*, vol. 29, no. 5, pp. 1662–1674, 2018.

[13] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa, "Online incremental attribute-based zero-shot learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3657–3664.

[14] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 771–778.

[15] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Advances in Neural Information Processing Systems*, 2014, pp. 3464–3472.

[16] Z. Al-Halah and R. Stiefelhagen, "How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2015, pp. 837–843.

[17] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen, "Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5975–5984.

[18] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2013, pp. 2584–2591.

[19] J. Lei Ba, K. Swersky, S. Fidler *et al.*, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2015, pp. 4247–4255.

[20] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where–and why? semantic relatedness for knowledge transfer," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 910–917.

[21] T. Mensink, E. Gavves, and C. G. Snoek, "COSTA: Co-occurrence statistics for zero-shot classification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2441–2448.

[22] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot object recognition by semantic manifold distance," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2635–2644.

[23] E. Kerr, T. M. McGinnity, and S. Coleman, "Material classification based on thermal and surface texture properties evaluated against human performance," in *IEEE Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 444–449.

[24] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016, pp. 536–543.

[25] H. Liu, F. Sun, D. Guo, B. Fang, and Z. Peng, "Structured output-associated dictionary learning for haptic understanding," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1564–1574, 2017.

[26] A. Ghanbari, X. Chen, W. Wang, B. Horan, H. Abdi, and S. Nahavandi, "Haptic microrobotic intracellular injection assistance using virtual fixtures," in *IEEE Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, 2010, pp. 781–786.

[27] D. Yamashiro, S. Tanaka, and H. T. Tanaka, "Active estimation of friction properties with haptic vision," in *IEEE Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, 2008, pp. 1329–1332.

[28] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker. Penn-haptics-bolt. [Online]. Available: https://github.com/IanTheEngineer/Penn-haptics-bolt

[29] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell. "deep learning for tactile understanding from visual and haptic data". [Online]. Available: https://people.eecs.berkeley.edu/∼yg/icra2016/

[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. on Multimedia*. ACM, 2014, pp. 675–678.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[32] Caffe. [Online]. Available: https://github.com/BVLC/caffe

[33] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database (supplemental material)," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.