Cross-modal Visuo-Tactile Object Recognition Using Robotic Active Exploration

Pietro Falco[1], Shuang Lu[1], Andrea Cirillo[2], Ciro Natale[2], Salvatore Pirozzi[2], and Dongheui Lee[1]

***Abstract*— In this work, we propose a framework to deal with cross-modal visuo-tactile object recognition. By cross-modal visuo-tactile object recognition, we mean that the object recognition algorithm is trained only with visual data and is able to recognize objects leveraging only tactile perception. The proposed cross-modal framework is constituted by three main elements. The first is a unified representation of visual and tactile data, which is suitable for cross-modal perception. The second is a set of features able to encode the chosen representation for classification applications. The third is a supervised learning algorithm, which takes advantage of the chosen descriptor. In order to show the results of our approach, we performed experiments with 15 objects common in domestic and industrial environments. Moreover, we compare the performance of the proposed framework with the performance of 10 humans in a simple cross-modal recognition task.**

I. INTRODUCTION

In order for robots to execute tasks in unstructured environments, multimodal perception plays a key role. Computer vision technologies have become essential for an effective analysis of the scene, for path planning, and observing the behavior of humans in the robot workspace. However, vision alone is often not enough to achieve sufficient perception capabilities of robotic systems in unstructured environments, due to variable light conditions, occlusions in cluttered scenes, and requirement on contact information between robot and environment. Tactile perception is of fundamental importance for robots that physically interact with the external environment. Wisely leveraging tactile information provides robots with enhanced perceptive capabilities. For these reasons, in the robotic community tactile perception and, in general, multimodal perception are becoming important research directions to support visual perception. Even though tactile and visuo-tactile multimodal perception have gained a great deal of interest, the field of cross-modal perception has not been profusely explored in robotics. A robotic system with cross-modal perception capability is able to leverage a-priori knowledge acquired with a sensing modality and to use it with a completely different sensorial modality at execution time. A practical example of the cross-modal perception is visuo-tactile cross-modal object recognition, which we deal with in this work. In neuroscience and psychology, cross-modal (or intermodal) object recognition is defined as *the name for the ability to recognize an object, previously inspected with one modality like vision, via a second modality like touch* [1], *without prior training in the second modality* [2]. In our robotic application, the robot observes only visual data in the training step. At execution time, exploiting the knowledge previously acquired by visual perception, the robot has the capability to recognize objects only with tactile perception, even if it did not touch any objects before. In a real world scenario, in fact, it can be very common that for the unstructured nature of the environment, a perception source can become unavailable at a given unexpected time. We investigate, then, if we can reuse the knowledge gained with the first perception source (vision) even if we have a different perception source active (tactile). As depicted in Fig. 1, we investigate in this work if it is possible to exploit tactile exploration to recognize the object with a classifier trained only with visual data.

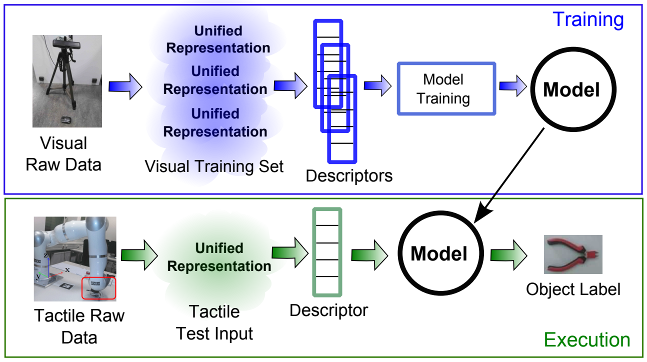


Fig. 1: Cross-modal recognition concept: training pipeline (top) and execution pipeline (bottom)

In this work, the following questions will be discussed:

1) What is a good representation for both visual and tactile data, which can be used in a compatible, interchangeable, and transparent fashion?
2) Given a visuo-tactile representation, can we use existing descriptors proposed by the computer vision community to deal with cross-modal recognition problems?
3) What classification algorithm should we adopt to recognize the objects?

We respond supporting our proposed choices with a large set of experiments. The rest of the paper is organized as follows. In Sec. II the related work is reported. Section III describes the proposed unified representation and the proposed descriptor. Both visual and tactile sensing setup are described in Sec. IV. Section V presents diverse experiments in order to show the performance of the proposed cross-modal classifier. Conclusions are reported in Sec. VI.

[1] Chair of Automatic Control Engineering, Technical University of Munich, Germany pietro.falco@tum.de

[2] Department of Industrial and Information Engineering, Università degli Studi della Campania "Luigi Vanvitelli", Aversa, Italy

II. RELATED WORK

In the robotic literature, diverse works exist related to monomodal and multimodal object recognition. Cross-modal object recognition has been studied especially in neuroscience and psychology.

Monomodal Recognition: in monomodal recognition, the classifier is trained with one sensing modality, usually visual or tactile, and it is used at execution time with the same modality to recognize objects. Visual object recognition has become a rather mature field and several works can be found in the literature. Due to the diffusion of low-cost RGB-D cameras, object recognition from 3D point clouds has gained a great deal of interest within vision and robotics communities. For this reason, several descriptors have been proposed in the last years, which exploit visual 3D point clouds. Example are Persistent Feature Histograms (PFH) [3], Fast PFH (FPFH) [4], Unique Signatures of Histograms (SHOT) [5], and Ensemble of Shape Functions (ESF) [6], and Spin Images (SI) [7].

In addition to computer vision, tactile perception has become of crucial importance for object recognition [8], material features detection [9], and slipping avoidance in grasping tasks [10]. In [11], a descriptor to recognize objects is presented with tactile data collected by a robotic hand. The work presented in [12] uses a bag-of-words approach, which recognizes object from low-resolution tactile images, obtained by grasping the object with a sensorized gripper. A bag-of-features framework in presented in [13]. It uses diverse tactile image descriptors to estimate a probability distribution over object identity as an object is explored, while [14] exploits texture properties to discriminate objects.

Multimodal Recognition: in the multimodal perception case, the training procedure is carried out by using both visual and tactile data in order to improve the accuracy. In [15], a deep learning method based on Convolutional Neural Networks (CNNs) is proposed, which achieves an enhanced accuracy in recognizing properties of materials through the fusion of tactile and visual data. The approach presented in [16] integrates visual and range data to recognize objects. In [17], visual features are combined with tactile glances to refine object models, obtaining more accurate information about surfaces. The work in [18] reconstructs 3-D models of unknown objects by fusion of visual and tactile information while objects are grasped.

Cross-modal Recognition: Cross-modal perception has been studied with great interest in the communities of psychology and neuroscience. Recent studies carried out on animals are reported in [19], [2]. In [20], a specific chapter is dedicated to cross-modal recognition. A classical study concerning intermodal matching on infants is reported in [21], while, in [22], visuo-tactile cross-modal perception in apes is investigated. To the best of our knowledge, robotic cross-modal visuo-tactical recognition has not been investigated.

III. CROSS-MODAL OBJECT RECOGNITION

In this section, the three elements of our cross-modal visuo-tactile framework are described, i.e., unified representation, features definition, and learning algorithm.

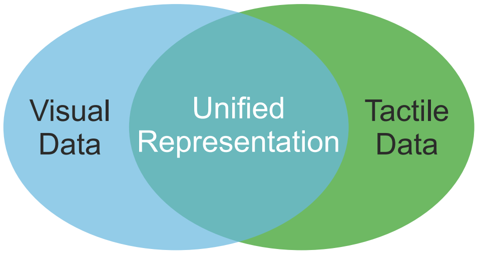


Fig. 2: A representation for cross-modal perception considers common information between the modalities

A. Representation and Preprocessing

The first point we address is how to represent visual and tactile data to allow an effective cross-modal perception. RGB-D cameras allow us to represent an object O as a set of points $\mathcal{P} = \{\boldsymbol{p}_0, \boldsymbol{p}_1, ..., \boldsymbol{p}_n\}$, defined hereafter as point cloud of O. Each vector $\boldsymbol{p} = (p_x, p_y, p_z)$ denotes the 3D position coordinates of the point $\boldsymbol{p}$. With the symbols $\mathcal{P}^v$ and $\boldsymbol{p}^v$ we indicate that the point cloud $\mathcal{P}$ and the point $\boldsymbol{p} \in \mathcal{P}$ is captured with visual perception. In order to derive a unified, compatible representation, *we represent tactile raw data as point clouds*, meaning by raw data the contact points between the object and the sensor. Even though representations based on tactile point clouds were used for shape reconstruction [23] and creation of object bounding boxes [24], this choice may appear naive for object recognition applications. In fact, modern tactile sensors can provide richer information than a point cloud, such as contact forces, textures, pressure maps, and friction coefficients. However, as graphically shown in Fig. 2, in order to achieve cross-modal capabilities, a representation is required that contains information common to both visual and tactile perception. The tactile point cloud representation of the object O is denoted as $\mathcal{P}^t = \{\boldsymbol{p}^t_0, \boldsymbol{p}^t_1, ..., \boldsymbol{p}^t_m\}$, where the symbol t denotes a point cloud acquired with a tactile perception system. In this work, we avoid the need of a registration step [25] between visual and tactile point clouds by choosing feature descriptors that are invariant to rotation and translation. Tactile and visual point clouds present significant differences in point density, in partiality of data, and in the characteristics of noise that affects the measurements. To derive a more effective unified representation, we equalize $\mathcal{P}^v$ and $\mathcal{P}^t$ in order to reduce the difference in point density and in partiality. Data partiality consists in missing points in visual and tactile clouds. Even when the position and orientation of the objects are the same in both tactile and visual exploration, the tactile and visual point clouds have different missing points. Besides partiality, visual and tactile point clouds present also different point densities. In order to alleviate these differences, we preprocess both tactile and visual point clouds through two main steps: *equalizing partiality of the data* and *uniforming point density*.

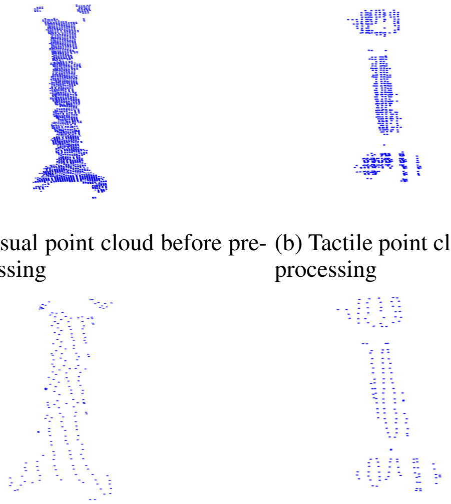
Equalizing Partiality: The method we adopt to handle data partiality is the Moving Least Squares (MLS) surface reconstruction [26]. This step allows us to filter the measurement noise and to recreate the missing parts of the surface. The core of the MLS approach is composed by three basic steps. First of all, we assume that points of each 3D tactile or visual point cloud $\mathcal{P}$ belong to a two-dimensional surface S. However, measurement noise corrupts the observed position of each point. As a consequence, the points of $\mathcal{P}$ will be ”near” S but do not belong to S. Given each point $\boldsymbol{p} \in \mathcal{P}$, the first step consists in finding a plane H that approximates locally the surface S in a region I of center $\boldsymbol{p}$ and radius r, called ”search radius”. The plane H is computed by using Principal Component Analysis (PCA). The points of the set I are projected onto H and upsampled with a step of $0.3\,\mathrm{mm}$. With these operation we transform the set I into the set $\tilde{I}$. The second step consists in fitting with a polynomial of order p_d the height of the points projected on H. We choose $p_d = 2$ and $r = 6\,\mathrm{cm}$. Setting $r = 6\,\mathrm{cm}$ confers rather strong filtering behavior and we can lose information in proximity of sharp edges. Typical values in monomodal visual perception are $r \in [1.5, 3]\,\mathrm{cm}$. However, in cross-modal perception, a strong filter is able to equalize cross-modal noise and in our case study allowed enhanced performance. A more detailed and formal description of the procedure can be found in [26]. In this work, the parameters are chosen with a grid search approach, maximizing the recognition accuracy.

Uniforming Density: The second step of the equalization procedure consists in applying a voxel grid filter [27] to downsample and ensure a more uniform point density. We apply the voxel filtering approach implemented in PCL. In this approach, the space is divided in 3D cubes (called voxels). All the points contained in each 3D box are substituted with their centroids. Following this procedure, the number of points will be equal to the number of voxels. Selecting appropriately the dimension of the voxels, the similarity of point density between tactile and visual data can be improved. In this work, we have empirically chosen cubic voxel with edge length $l = 5\,\mathrm{mm}$.

The procedure is summarized in Algorithm 1. An example of visual and tactile point clouds before and after preprocessing is shown in Fig. 3. The equalization step plays a key role in order to improve the performance (see Sec. V).

B. A Suitable Descriptor

After defining a unified representation based on point clouds, it is important to choose a suitable feature descriptor that allows cross-modal recognition. The choice of the feature descriptors strongly depends on the chosen representation of raw data. Since our unified representation is based on point clouds, we orient our research towards 3D point cloud descriptors. From the results reported in the computer vision literature [5] [6], we expect that SHOT and ESF are promising candidates for our problem. However, we obtained rather poor results with respect to the monomodal classification, since the data in the training and test set come from radically different sensing modalities. Even after the preprocessing, the differences in noise, resolution and partiality of the data found in the training and test set cannot be equalized perfectly. Therefore, the need of a new descriptor suitable for cross-modality. Following a strategy commonly adopted in communication engineering, we propose to increase the redundancy of the information associated to the descriptors. A straightforward way to increase the redundancy is finding a smart combination of different descriptors. In our case we expect benefit by combining SHOT and ESF, since they encode information with two different approaches. SHOT encodes point clouds with histograms of normal vectors [5], while ESF does not compute normals and encodes information based on shape functions [6]. The ESF descriptor is an ensemble of 10 concatenated histograms of shape functions consisting of angle, point distance, and area functions. Each histogram has 64 bin, for a total of 640 elements [6]. The SHOT computes a local reference frame Σ_c using the eigenvalue decomposition around an input point $\boldsymbol{c}$, in our case $\boldsymbol{c}$ is the centroid of the point cloud $\mathcal{P}$. Given the frame Σ_c, a sphere S_c of center $\boldsymbol{c}$ and radius r_c is defined. S_c is then split into 32 divisions and for each division a 11-bin histogram is computed. Each histogram contains angles that describe the directions of the normal vectors to each point $\boldsymbol{p} \in \mathcal{P}$ in the frame Σ_c. The descriptor concatenates the histograms into the final signature, obtaining a vector of 352 elements. We compute a single SHOT feature for each object and use it as a global feature.


(a) Visual point cloud before preprocessing (b) Tactile point cloud before preprocessing

(c) Visual point cloud after preprocessing (d) Tactile point cloud after preprocessing

Fig. 3: Visual and tactile point clouds

Algorithm 1 Equalization

```
function $\mathcal{P}$=equalize(PointCloud $\mathcal{P}^*$)
$\bar{\mathcal{P}}$ =MLS($\mathcal{P}^*, u_s = 0.3\,\mathrm{mm}$, $r = 6\,\mathrm{cm}$ ,$p_d = 2$)
$\mathcal{P}$ =voxelGridFilter($l = 5\,\mathrm{mm}$)
return($\mathcal{P}$)
```

Let $\boldsymbol{d}_{SHOT}$ be the SHOT descriptor associated to the point cloud $\mathcal{P}$ and $\boldsymbol{d}_{ESF}$ the ESF descriptor associated to the same point cloud. Both descriptors are column vectors. The first possible way to improve the performance is to simply concatenate $\boldsymbol{d}_{SHOT}$ and $\boldsymbol{d}_{ESF}$ so that:

$$\boldsymbol{d}_c = [\boldsymbol{d}_{SHOT}^T \quad \boldsymbol{d}_{ESF}^T]^T, \tag{1}$$

where $\boldsymbol{d}_c$ is the concatenated descriptor. Even if the concatenated descriptor $\boldsymbol{d}_c$ contains more information than $\boldsymbol{d}_{ESF}$, the improvement in accuracy was not significant. This can happen because the dimension of $\boldsymbol{d}_c$ is much higher than both SHOT and ESF. As a consequence, the classification problem is affected by the curse of dimensionality. Moreover, the high increase in dimension can be a limitation also in terms of training time and classification time, especially when scaling to very large databases. Therefore, we decided to exploit a data compression method. To compress the descriptor $\boldsymbol{d}_c$, we organize the vectors $\boldsymbol{d}_{SHOT}$ and $\boldsymbol{d}_{ESF}$ in the matrix:

$$\hat{\boldsymbol{D}} = [\boldsymbol{d}_{SHOT} \quad \tilde{\boldsymbol{d}}_{ESF}], \tag{2}$$

where $\tilde{\boldsymbol{d}}_{ESF} = [\boldsymbol{d}_{ESF} \quad \bar{\mathbf{0}}]$ and $\bar{\mathbf{0}}$ is a 0-vector of dimension $1 \times (640 - 352)$. We want to compress the information carried from the matrix $\hat{\boldsymbol{D}} \in \mathbb{R}^{640\times 2}$ into a vector $\boldsymbol{d}_r \in \mathbb{R}^{640}$. We leverage the data compression capability of Singular Value Decomposition (SVD) [28]. First, we center the matrix $\hat{\boldsymbol{D}}$, so that all columns are zero-mean. Let $\boldsymbol{D}$ be such a mean-centered matrix. The SVD of the matrix $\boldsymbol{D}$ is

$$\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathbf{T}}, \tag{3}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ...\boldsymbol{u}_{640}] \in \mathbb{R}^{640\times 640}$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2] \in \mathbb{R}^{2\times 2}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{\mathbf{640\times 2}}$ is the matrix that contains the singular values σ_1 and σ_2. We choose the compressed SVD descriptor $\boldsymbol{d}_r$ as

$$\boldsymbol{d}_r = \boldsymbol{u}_1\sigma_1. \tag{4}$$

This descriptor has dimension 640 and is a linear combination of the columns of the matrix $\boldsymbol{D}$. The best rank-1 approximation of the matrix $\boldsymbol{D}$ is given by the matrix $\boldsymbol{D}_1 = \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1$. As a consequence, $\boldsymbol{d}_r\boldsymbol{v}_1$ is the 1-rank matrix that minimizes the norm $\|\boldsymbol{D} - \boldsymbol{D}_1\|$.

Using $\boldsymbol{d}_r$ as a descriptor we obtain significantly better performance than using the $\boldsymbol{d}_c$. The descriptor $\boldsymbol{d}_r$, in fact, carries more information than both ESF and SHOT, but is less affected by the curse of dimensionality than the concatenated descriptor $\boldsymbol{d}_c$. We call the descriptor derived in Eq. (4) Cross-Modal ESF (CMESF). The CMESF descriptor consists in the basic ESF enriched with the information carried by SHOT.

C. Learning Algorithm

We compare k-Nearest Neighbor (k-NN), with different values of k and radial basis function kernel Support Vector Machines (SVM). Both k-NN and SVM are simple and widely-used algorithms for classification problems. We apply such learning algorithms to several state-of-the-art visual descriptors and with the one proposed in this work. We found that a suitable choice in the proposed framework is k-NN.

In more detail, to deal with the cross-modal recognition problem, we perform two steps. The first step consists in building a model $\mathcal{M}^v$, which embeds a-priori knowledge derived from visual perception. The second step is to exploit a-priori knowledge embedded in $\mathcal{M}^v$ with data from a different sensing modality.

Algorithm 2 Cross-modal Recognition

```
function l=recognize(PointCloud $\mathcal{P}_o^t$, Model $\mathcal{M}^{\mathrm{v}}$)
$\bar{\mathcal{P}}_o^t$ = equalize($\mathcal{P}_o^t$)
$\boldsymbol{d}_{SHOT}$ =computeSHOT($\bar{\mathcal{P}}_o^t$)
$\boldsymbol{d}_{ESF}$ =computeESF($\bar{\mathcal{P}}_o^t$)
$\hat{\boldsymbol{D}} = [\boldsymbol{d}_{SHOT} \quad \boldsymbol{d}_{ESF}]$
$\boldsymbol{D} = \mathrm{center}(\hat{\boldsymbol{D}})$
$[\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}] = \mathrm{svd}(\boldsymbol{D})$
$\boldsymbol{d}_{CMESF} = \boldsymbol{U}(:,1)\boldsymbol{\Sigma}(\mathbf{1},\mathbf{1})$
$l = k$-NN($\boldsymbol{d}_{CMESF}, \mathcal{M}^v$)
return($l$)
```

Building the model: We use visual point clouds of 15 objects and for each object we collect 40 examples. Each example i consists in a point cloud $\mathcal{P}_i$. For each point cloud $\mathcal{P}_i$, we compute the CMESF descriptor $\boldsymbol{d}_i$, which is the representation of the point cloud in the feature space. Let $\mathcal{V}$ be the set of the CMESF descriptors associated to all the examples, i.e., $\mathcal{V} = (\boldsymbol{d}_1, \boldsymbol{d}_2, ...\boldsymbol{d}_l)$, with $l = 600$ in our case. We derive the model $\mathcal{M}^v$ using the set $\mathcal{V}$. When using k-NN, the model $\mathcal{M}^v$ simply consists in the elements of $\mathcal{V}$. In case of SVM and other methods that require an explicit training step, we use $\mathcal{V}$ as training set and the model $\mathcal{M}^v$ is the trained classifier. The sensing system and the procedure to collect visual data is described in Sec. IV-A.

Exploiting the model for cross-modal recognition: We exploit the knowledge accumulated with visual perception in order to interpret tactile data at execution time. To test the performance of cross-modal recognition, we classify the outcome of 5 tactile explorations per object. The tactile sensing system and the exploration procedure are described in Sec. IV-B. After the acquisition of the tactile point cloud, we derive the descriptor $\boldsymbol{d}$ with the procedure described in Sec. III-B. To recognize the object through visual a-priori knowledge, we provide $\boldsymbol{d}$ as an input to the classifier which embeds the model $\mathcal{M}^v$. The output of such a classifier is the estimated class of the object.

The entire process of visuo-tactile recognition that adopts the CMESF descriptor is summarized in Algorithm 2. The inputs of the algorithm are the model $\mathcal{M}^v$, derived by visual data a-priori known and the point cloud observed by tactile sensors $\mathcal{P}_O^t$ at execution time. The output is the label l of the explored object O.

IV. SENSING SYSTEM

In order to implement the cross-modal object recognition and to show the validity of the design choices reported in Sec. III, we prepared an experimental setup, constituted by a visual perception system and a tactile perception system.

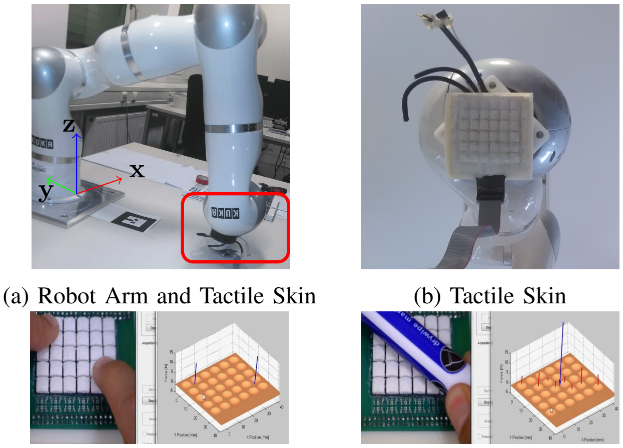
A. Visual Perception

The visual perception system is shown in Fig. 4. The visual point clouds are collected with a Kinect RGB-D camera. The objects and the camera are placed as depicted in Fig. 4. The collected point cloud of an object O is separated from the rest of the scene by an Euclidean cluster extraction (ECE) algorithm in the PCL libraries. The ECE algorithm removes the planes from the scene and clusters the remaining points with a kd-tree approach.

B. Tactile Perception

Tactile Sensing Setup: In the experiments, the SAPHARI tactile skin [29] is mounted on the end effector of a KUKA light weight robot, as shown in Fig 5. The tactile technology, originally presented in [30], exploits optoelectronic devices to detect the local deformation, generated by an external contact force applied to a deformable layer that covers the optoelectronic layer. The tactile skin consists in a grid of 6×6 sensor modules with a size of $5 \times 5\,\mathrm{cm}$ as a whole. Every sensing module, shown in Fig 5, has a unique spatial representation in the robotic base frame and provides the three component of the force applied on it. When the intensity of the contact force $\|\boldsymbol{F}_i\|$ for the module i is larger than a threshold θ, the contact point $\boldsymbol{p}_i$ is extracted. In this work, we have empirically chosen $\theta = 0.8\,\mathrm{N}$. The tactile readings for each object are then represented as three-dimensional point clouds, as described in Sec. III. The experimental setup for tactile exploration is illustrated in Fig. 5.

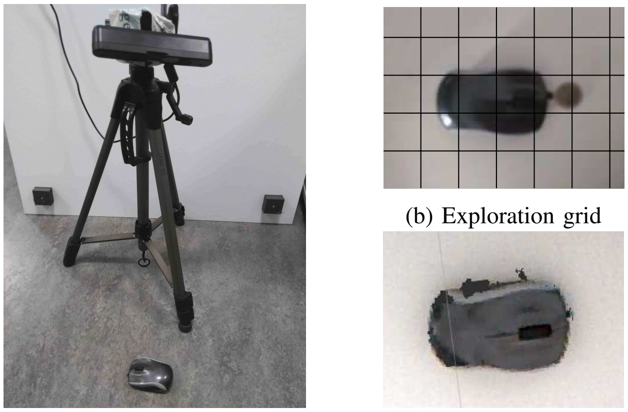
Exploration Strategy: Tactile exploration is a core part of tactile object representation and recognition. An appropriate exploration strategy ensures a good quality of the tactile point clouds. In this work, we explore the object by pressing along the $\boldsymbol{z}$ axis as depicted in Fig. 5. We take the robot base frame as the unified world reference frame, all tactile point clouds are represented in this frame. When a module of the tactile skin is in contact with the object, the point of contact between the tactile skin and the object is included in the point cloud $\mathcal{P}^t$ relative to the object O. The tactile readings are described as a six dimensional vector, encoding the pose and force information. During the exploration, Cartesian impedance controller is used, which allows the robot to interact with the object in a compliant manner. The robot is compliant along $\boldsymbol{z}$, so that it is able to explore without damaging any objects. As depicted in Fig. 4b, the end effector is controlled to move to each vertex of the grid and, after reaching a vertex, press the object. Each contact point between the skin and the object is represented as a point $p = (p_x, p_y, p_z)$ and it is expressed in the robot base frame. In this work, the tactile frame Σ_t is the robot base frame. However, the feature selection is frame-independent. All the objects are fixed on the table. The points of the table are removed with the planar filter algorithm implemented in PCL. We adopt a simple exploration strategy, which is particularly suitable for planar objects. The exploration procedure is described in Algorithm 3.



(a) Robot Arm and Tactile Skin (b) Tactile Skin

(c) Tactile Skin Readings (d) Tactile Skin Readings

Fig. 5: Experimental setup for tactile perception


(b) Exploration grid

(a) Visual Setup (c) Visual Point Cloud

Fig. 4: Visual Sensing System

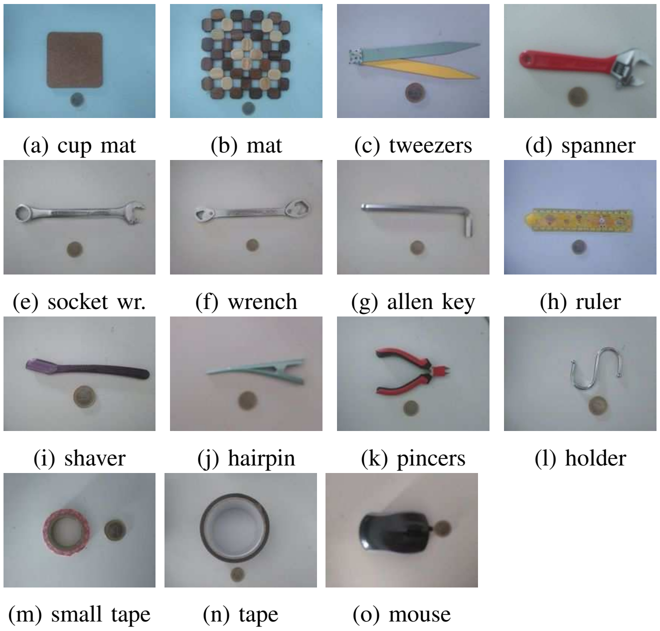
Algorithm 3 Exploration Strategy

```
$Traj_1 = (\boldsymbol{v}_1, \boldsymbol{v}_2, ...., \boldsymbol{v}_n)$    ▷ grid vertices in Fig. 4b
for $\boldsymbol{v}_j \in Traj_1$ do
    moveTo($\boldsymbol{v}_j$)  ▷ it brings robot from vertex to vertex
    press on vertex $\boldsymbol{v}_j$
    for each sensor module $i$ do
        read($\boldsymbol{F}_i$)
        if $\|\boldsymbol{F}_i\| \geq 0.8N$ then
            $\boldsymbol{p}_i \leftarrow (p_x, p_y, p_z)$
            $\mathcal{P} = \mathcal{P} \cup \{\boldsymbol{p}_i\}$
        end if
    end for
end for
```

V. Experiments

A. Description of the Dataset

We selected 15 objects, depicted in Fig. 6, which are typical of domestic and industrial environments. For each object, the tactile exploration procedure, described in Algorithm 3 has been repeated 5 times. Then, 40 samples from each object


(a) cup mat (b) mat (c) tweezers (d) spanner

(e) socket wr. (f) wrench (g) allen key (h) ruler

(i) shaver (j) hairpin (k) pincers (l) holder

(m) small tape (n) tape (o) mouse

Fig. 6: Objects used for experiments, each one close to a 1-euro coin

have been collected with the visual system in Fig. 4. After the data acquisition procedure, we have 40 visual and 5 tactile point clouds for each object. The visual point clouds are used to build the a-priori knowledge. In this case study, a-priori knowledge is constituted by the classifier trained with visual data, denoted in our work with the symbol $\mathcal{M}^v$. The tactile exploration data are then classified exploiting the a priori knowledge $\mathcal{M}^v$. It is important to emphasize that, with the proposed approach, the robot can classify objects using the sense of touch even the object has never been touched before, but it has been only seen by vision.

Feature descriptor	1-NN	3-NN	5-NN	SVM
PFH	15.20%	8.44%	8.44%	8.44%
FPFH	10.06%	10.06%	10.06%	13.34%
SI	21.33%	22.67%	21.33%	32.00%
SHOT	32.63%	32.63%	31.25%	35.79%
ESF	45.26%	45.26%	42.11%	32.63%
$\boldsymbol{d}_c$	45.26%	45.26%	42.11%	26.32%
CMESF	55.79%	**57.89**%	53.68%	33.68%

TABLE I: Recognition result without preprocessing

B. *Classification results*

In order to show the performance of the framework, we evaluate, in terms of accuracy, the proposed combination of (1) unified representation, (2) unified descriptor, and (3) suitable learning algorithm. We compare different state-of-the-art descriptors with the proposed CMESF, as shown in Table II. In order to show how strongly the preprocessing step impacts on the performance of cross-modal object recognition, we indicate in Table I and Table II the classification accuracy without and with preprocessing, respectively. In the light of the results mentioned above, the preprocessing allows an improvement up to 20% in the recognition accuracy and the importance of this step is crucial when dealing with visuo-tactile cross-modality. We compare k-NN and radial basis function kernel SVM, since they are simple and have recognized effectiveness in classification problems. The best performance is achieved by 5-NN combined with the CMESF descriptor. The accuracy is evaluated by classifying 5 tactile explorations per object using the visual knowledge embedded in the model $\mathcal{M}^v$, obtained through the different classification methods reported in Table II. We can observe that the descriptors using estimated normal vectors, i.e., PFH, FPFH, SI and SHOT, perform worse than ESF. CMESF is particularly suitable for cross-modal recognition, as the improvement with respect to basic ESF is, in our case study, almost 15% and the dimension remains the same. In Table III, the confusion matrix is reported for a more detailed analysis. We can notice the recognition performance is between 80% and 100% for all the objects except for 4: (j) the hairpin presents 20% accuracy, (n) the tape 40% accuracy, (e) the socket wrench 60% accuracy, (h) the ruler 60% accuracy.

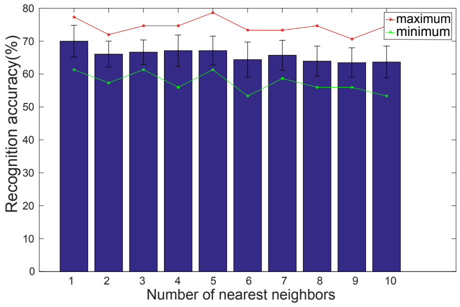


Fig. 7: Cross-modal recognition result on 10 observations

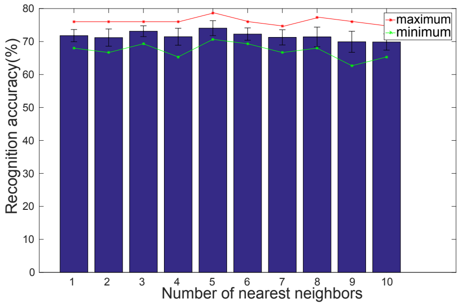


Fig. 8: Cross-modal recognition result on 30 examples

Feature Descriptor	1-NN	3-NN	5-NN	SVM
PFH	5.33%	12.00%	12.00%	10.67%
FPFH	9.33%	12.00%	14.67%	16.00%
SI	22.67%	28.00%	28.00%	40.00%
SHOT	37.33%	36.00%	34.67%	32.00%
ESF	60.00%	65.33%	65.33%	49.33%
$\boldsymbol{d}_c$	62.67%	68.00%	66.67%	30.67%
CMESF	72.00%	73.33%	**77.33**%	40.00%

TABLE II: Cross-modal Recognition Result

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	1.0	0	0	0	0	0	0	0.2	0	0	0	0.2	0	0	0
b	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	1.0	0	0.4	0	0	0	0	0	0	0	0	0	0
d	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0.8	0	0	0	0	0	0	0.2	0	0
g	0	0	0	0	0	0.2	1.0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0.6	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0.2	1.0	0	0	0	0	0.6	0
j	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0.8	1.0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

TABLE III: Confusion matrix of visual-tactile object recognition

.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0.05	0	0.75	0	0	0	0	0	0	0	0	0		0	0
d	0	0	0.05	0.75	0.1	0	0	0	0	0.1	0	0	0	0	0
e	0	0	0	0.15	0.9	0.1	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0.9	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0.95	0	0	0	0	0	0	0	0
h	0	0	0.1	0	0	0	0.05	0.8	0.2	0	0	0	0	0	0
i	0	0	0.1	0	0	0	0	0.2	0.8	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	0.75	0	0	0	0	0
k	0	0	0	0.1	0	0	0	0	0	0.15	1.0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0.95	0.05	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.95	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

TABLE IV: Confusion matrix of human object recognition.

C. Number of examples

In order to understand how the performance changes when the cardinality of the training set decreases, we define 60 reduced training sets divided in 3 classes. The first class of training sets is obtained by random sampling 30 observations from the complete training set. The random sampling procedure is carried out 20 times. We denote the first class as $\mathcal{C}^{30} = \{T_1^{30}, T_2^{30}, .., T_{20}^{30}\}$. The second class is obtained by random-sampling 20 examples from the complete training set. According to the notation adopted above, we denote the second class of training sets as $\mathcal{C}^{20} = \{T_1^{20}, T_2^{20}, .., T_{20}^{20}\}$. Also in this case the sampling procedure is carried out 20 times. The third class is obtained by random sampling 20 times 10 observations out of 40 and it is denoted as $\mathcal{C}^{10} = \{T_1^{10}, T_2^{10}, .., T_{20}^{10}\}$. Figures 8 and 7 show the accuracy of cross-modal recognition using the classes $\mathcal{C}^{30}$ and $\mathcal{C}^{10}$, respectively. In the figures we report the minimum (in green), the maximum (in red), and the average result emphasized by the histogram in blue. For each class, we define average, minimum and maximum values because we evaluate 20 randomly extracted reduced training sets per class. This part of our analysis is important to understand how the performance degrades when the number of examples decreases. The classifiers show an average recognition accuracy around 70% both with 30 observations (Fig. 8) and 20 observations.The result on 10 observations in the training set decreased to around 65% (Fig. 7). As expected, we can conclude that the recognition accuracy on larger training sets is higher. However, the accuracy does not decrease dramatically even in case of only 20 observations, i.e., half training set. With half the examples, the average accuracy decreases of 7%.

D. Comparison with human cross-modal object recognition

In order to have an ideal reference for assessing the performance of artificial cross-modal recognition, and because in the literature it is hard to find a cross-modal recognition algorithm, we compare the performance of our framework with a "golden standard", which is represented by the performance of humans. We arranged then a simple experiment, described in Algorithm 4, to have a first estimation of human performance in visuo-tactile cross-modal object recognition tasks. In this experiment, 10 participants, ranging in age from 20 to 30 years, were invited to look at the set of objects shown in Figure 6 for 2 minutes. Afterwards, each participant was blindfolded and invited to explore each object with one hand. Since human skin can sense also the temperature of the object, the participants wear a thin glove to maintain the tactile perception capability, but to reduce the perception of the temperature. The objects are placed on the table. The participants explore by touching each object for 10 s without seeing the objects. After that, the participants are invited to say which object was explored. Algorithm 4 describes the adopted protocol step by step. A picture of human tactile exploration is shown in Fig. 9. The average accuracy achieved is 89.7% and in Table IV the confusion matrix is reported. The accuracy of humans in this experiment is 12% better than the performance of the proposed method based on processed tactile point clouds and the CMESF descriptor. Comparing Tables III and IV we can notice that for most objects the performance of our framework are close to human performance in this case study.

Feature Descriptor	Visual	Tactile
FPFH	66.33%	91.67%
SI	98.00%	93.30%
SHOT	97.17%	92.00%
ESF	97.33%	**94.67**%
$\boldsymbol{d}_c$	98.50%	93.33%
CMESF	**98.67%**	**94.67**%

TABLE V: 1-NN Monomodal Recognition Result

E. Comparison with monomodal object recognition

The results of the cross-modal visuo-tactile object recognition framework are also compared with the monomodal visual recognition case. The results of the visual and tactile monomodal cases are reported in Table V. In this case study, the classifier is trained and tested with the same modality. The accuracy has been evaluated with a 10-fold cross-validation method. From Table V it is possible to see that both visual and tactile monomodal problems are, as expected, less challenging than the cross-modal case, since training set and test set are generated from the same perception mode. Most state-of-the-art descriptors achieve more than 90% accuracy in the monomodal case with 1-NN.

VI. CONCLUSION AND FUTURE WORK

In this work, we deal with cross-modal visuo-tactile object recognition. We train a classifier by using visual data from a Kinect camera and we recognize objects at execution

Fig. 9: Tactile exploration performed by a human

Algorithm 4 Protocol for the experiment with humans

```
1: The subject is invited to see all the objects for 2 minutes
2: The subject is invited to wear a thin glove that prevents
   from sensing the temperature of objects
3: The subject is blinded with a blind fold
4: The objects are put in a bag, one object is picked out
   and put on a table
5: The subject explores the object with the hand for 10 s
6: The subject tells the name of the explored object
7: Go to step 3 until every object has been picked out
```

time only with tactile data, without any a-priori tactile information. Through this case study, we aim at answering the three questions introduced in Sec. I. Answering to the first question, we propose in Sec. III-A the visuo-tactile point cloud representation combined with the procedure of equalization in resolution and partiality. As an answer to Question 2, we found that using descriptors from the computer vision community, as they are, is not the best option since we obtain only around 50% accuracy. However, combining two different descriptors, i.e., SHOT and ESF, with a SVD-based compression revealed a feasible strategy.Concerning Question 3, our experiments show that a suitable learning algorithm for the proposed descriptor is k-NN.

Future work will consist in implementing more complex exploration strategies, in investigating novel visuo-tactile descriptors, and in investigating deep learning methods to train a model from a huge amount of visual data.

ACKNOWLEDGMENT

This work has been partially funded by the Marie Curie Action LEACON, EU project 659265 and by the Technical University of Munich, International Graduate School of Science and Engineering (IGSSE).

REFERENCES

[1] "Psychology dictionary," http://psychologydictionary.org/.

[2] S. Schumacher, T. B. de Perera, J. Thenert, and G. von der Emde, "Cross-modal object recognition and dynamic weighting of sensory inputs in a fish," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7638–7643, 2016.

[3] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008*.

[4] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *IEEE International Conference on Robotics and Automation, 2009*.

[5] S. Salti, F. Tombari, and L. Di Stefano, "Shot: unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[6] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3d object classification," in *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, 2011, pp. 2987–2992.

[7] A. Johnson, "A representation for 3d surface matching," Ph.D. dissertation, Ph. D. thesis, Carnegie Mellon University, 1997.

[8] S. Luo, W. Mou, K. Althoefer, and H. Liu, "Novel tactile-sift descriptor for object shape recognition," *Sensors Journal, IEEE*, no. 9, 2015.

[9] H. Liu, X. Song, J. Bimbo, L. Seneviratne, and K. Althoefer, "Surface material recognition through haptic exploration using an intelligent contact sensing finger," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 52–57.

[10] G. De Maria, P. Falco, C. Natale, and S. Pirozzi, "Integrated force/tactile sensing: The enabling technology for slipping detection and avoidance," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 3883–3889.

[11] M. M. Zhang, M. D. Kennedy, M. A. Hsieh, and K. Daniilidis, "A triangle histogram for object classification by tactile sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4931–4938.

[12] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, "Object identification with tactile sensors using bag-of-features," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 2009, pp. 243–248.

[13] Z. Pezzementi, E. Plaku, C. Reyda, and G. D. Hager, "Tactile-object recognition from appearance information," *Robotics, IEEE Transactions on*, vol. 27, no. 3, pp. 473–487, 2011.

[14] M. Kaboli, R. Walker, and G. Cheng, "Re-using prior tactile experience by robotic hands to discriminate in-hand objects via texture properties," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2242–2247.

[15] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," *arXiv preprint arXiv:1511.06065*, 2015.

[16] S. Gould, P. Baumstarck, M. Quigley, A. Y. Ng, and D. Koller, "Integrating visual and range data for robotic object detection," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2, 2008.*

[17] M. Bjorkman, Y. Bekiroglu, V. Hogman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, 2013, pp. 3180–3186.

[18] J. Ilonen, J. Bohg, and V. Kyrki, "Fusing visual and tactile sensing for 3-d object reconstruction while grasping," in *IEEE Int. Conference on Robotics and Automation (ICRA)*, 2013, pp. 3547–3554.

[19] J. M. Cloke, D. L. Jacklin, and B. D. Winters, "The neural bases of crossmodal object recognition in non-human primates and rodents: a review," *Behavioural brain research*, vol. 285, pp. 118–130, 2015.

[20] G. Calvert, C. Spence, and B. E. Stein, *The handbook of multisensory processes*. MIT press, 2004.

[21] A. N. Meltzoff and R. W. Borton, "Intermodal matching by human neonates," *Nature*, 1979.

[22] R. K. Davenport, C. M. Rogers, and I. S. Russell, "Cross modal perception in apes," *Neuropsychologia*, vol. 11, no. 1, pp. 21–28, 1973.

[23] M. Meier, M. Schöpfer, R. Haschke, and H. Ritter, "A probabilistic approach to tactile shape reconstruction," *Robotics, IEEE Transactions on*, vol. 27, no. 3, pp. 630–635, 2011.

[24] K. Charusta, D. Dimitrov, A. J. Lilienthal, and B. Iliev, "Extraction of grasp-related features by human dual-hand object exploration," in *Advanced Robotics, 2009. ICAR 2009. International Conference on*.

[25] S. Li and D. Lee, "Fast visual odometry using intensity-assisted iterative closest point," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 992–999, 2016.

[26] D. Levin, "Mesh-independent surface interpolation," in *Geometric modeling for scientific visualization*. Springer, 2004, pp. 37–49.

[27] "Downsampling a pointcloud using a voxelgrid filter," http://pointclouds.org/documentation/tutorials/voxel_grid.php.

[28] D. Kalman, "A singularly valuable decomposition: the svd of a matrix," *The college mathematics journal*, vol. 27, 1996.

[29] A. Cirillo, F. Ficuciello, C. Natale, S. Pirozzi, and L. Villani, "A conformable force/tactile skin for physical human–robot interaction," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 41–48, 2016.

[30] A. Cirillo, P. Cirillo, G. De Maria, C. Natale, and S. Pirozzi, "An artificial skin based on optoelectronic technology," *Sensors and Actuators A: Physical*, vol. 212, pp. 110–122, 2014.