Robotics and Autonomous Systems 60 (2012) 396–410

Contents lists available at SciVerse ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Extracting data from human manipulation of objects towards improving autonomous robotic grasping

Diego R. Faria, Ricardo Martins, Jorge Lobo *, Jorge Dias

Institute of Systems and Robotics, DEEC - University of Coimbra - Polo II, 3030-290 Coimbra, Portugal

ARTICLE INFO

Article history:
Available online 23 August 2011

Keywords:
Human demonstration
Manipulation task representation
Motion pattern
Probabilistic object representation
Contact points
Stable grasp

ABSTRACT

Humans excel in manipulation tasks, a basic skill for our survival and a key feature in our manmade world of artefacts and devices. In this work, we study how humans manipulate simple daily objects, and construct a probabilistic representation model for the tasks and objects useful for autonomous grasping and manipulation by robotic hands. Human demonstrations of predefined object manipulation tasks are recorded from both the human hand and object points of view. The multimodal data acquisition system records human gaze, hand and fingers 6D pose, finger flexure, tactile forces distributed on the inside of the hand, colour images and stereo depth map, and also object 6D pose and object tactile forces using instrumented objects. From the acquired data, relevant features are detected concerning motion patterns, tactile forces and hand-object states. This will enable modelling a class of tasks from sets of repeated demonstrations of the same task, so that a generalised probabilistic representation is derived to be used for task planning in artificial systems. An object centred probabilistic volumetric model is proposed to fuse the multimodal data and map contact regions, gaze, and tactile forces during stable grasps. This model is refined by segmenting the volume into components approximated by superquadrics, and overlaying the contact points used taking into account the task context. Results show that the features extracted are sufficient to distinguish key patterns that characterise each stage of the manipulation tasks, ranging from simple object displacement, where the same grasp is employed during manipulation (homogeneous manipulation) to more complex interactions such as object reorientation, fine positioning, and sequential in-hand rotation (dexterous manipulation). The framework presented retains the relevant data from human demonstrations, concerning both the manipulation and object characteristics, to be used by future grasp planning in artificial systems performing autonomous grasping.

1. Introduction

One of the key elements of the performance of robotic platforms is the ability to perform autonomous grasping, manipulation, exploration and characterisation of not completely known objects. Autonomous grasping and learning by imitation are topics that have been the focus of interest of many research groups in robotics for decades. The research aims to learn and model the human dexterity to endow a robot with such skills. The main objectives inside grasping strategy are to ensure stability and the ability of grasping unknown objects.

In robotics, the analysis of human movements has been applied in research areas related with task learning by imitation of human demonstrations. This approach is based on the principles described by several studies from human developmental sciences that, humans can acquire skills by watching and analysing others performing tasks. The challenge of using human demonstrations to model the manipulation strategies that will be performed by robotic platforms consists of building the bridge between the observation of human demonstrations and the reproduction of movements in the robotic platform that produce the same effect.

In this work, we propose a framework to extract relevant information from human demonstration using multimodal data overlaid with object information, having both the perspective of the object state during the manipulation task and the perspective of the human performing the manipulation. Identifying different stages of a manipulation task and characterising each phase of the task is important so as to retain the context in which different grasps and forces were used. One of the main elements in a manipulation task is the object being manipulated, and the effect of the human hand actions on the transformation of the object status from the starting conditions to the task goal. In this work the object is modelled as an object-centred probabilistic volumetric model, which is used to represent the contact regions and forces that enabled successful grasps during the human demonstrations. The object centred framework will facilitate future matching for an

* Corresponding author.
E-mail address: jlobo@isr.uc.pt (J. Lobo).

0921-8890/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2011.07.020

artificial system observing objects and searching for cues on how to grasp it taking into account the task context.

The next subsection describes some relevant previous work on this research topic. Section 2 presents the feature extraction implemented for the multimodal data from human demonstrations. The following section builds upon these features to segment and identify manipulation stages and derive a generalised task representation. Results are presented for two test cases: homogeneous (fixed grasp) and dexterous manipulation. Section 4 presents the object probabilistic volumetric map, a proposed representation of contact and tactile data for successful stable grasp. Section 5 presents how the manipulation knowledge acquired from human demonstrations can be unified with object information into a framework to be used in grasping planning strategies and autonomous grasping, with results, conclusion and future work presented at the end.

1.1. Related work

In this work, we address a subtopic of human movement analysis concerning object manipulation. Typically, in the literature, the techniques for movement analysis have two main approaches. The first group represents the movements at the trajectory level and generalise the representation of the movements through the extraction of statistical regularities from several human demonstrations of the movements, and the second group of approaches that propose a symbolic learning and encoding of movements based on the supervised labelling and segmentation of the primitives during the learning stage.

An example for the first class of approaches is provided by Calinon et al. [1], that proposes to extract continuous constraints from a set of demonstrations using different initial configurations of the manipulated object. The Cartesian trajectories are projected using the Principal Components Analysis (PCA). The spatio-temporal constraints are then represented through Gaussian Mixture Models (GMM). The approach has been successful in a robotic platform that reproduces a generalised version (obtained using Gaussian mixture regression) of a demonstrated task. Another example from the same class was suggested by Ogawara et al. [2], that presents a method to detect repeated motion patterns in a long motion sequence. The approach assumes that repeated motion patterns are structured information that can be obtained without the knowledge of the context of motions. The method was evaluated and compared to other previous works, by detecting repeated interactions between humans and objects in everyday manipulation tasks. The method has shown a greater performance in terms of detectability and computational time. Pastor et al. [3] proposed an approach to learn motor skills from human demonstrations modelled using a set of differential equations – Dynamic Movement Primitive (DMP) framework – and developed a library of movements by labelling each recorded movement according to task and context.

In this article, we present an approach that combines the previous group of methods (trajectory level) with the second group of methods, which associate symbolic learning with encoding of manipulation movements. The typical approach of these other methods is to initiate primitive sequence detection in the human demonstrations stream of data, followed by pattern recognition methods which provide the most probable temporal sequence of primitives. An example of this technique is used by Kondo et al. [4], that proposes a method to describe in-hand manipulation demonstration movements by recognising a sequence of contact state transitions between the human hand and the manipulated object. The recognition algorithm is based on a dynamic programming approach by comparing the similarity of the contact state transition between an input sequence and template manipulation primitives. The work by Kruger et al. [5] presents the automatic extraction of action primitives and the corresponding grammar from continuous movements of several human demonstrations of grasping tasks. The approach considers that all the actions can be described by a set of elementary building blocks and there is a set of rules (grammar) that define how the action primitives can be combined. The action primitives are represented by parametric hidden Markov models. One of the key elements of those platforms is their ability to handle and explore objects as shown by Klatzky and Lederman [6], Biederman et al. [7] or Sahbani et al. [8].

In this article, we also explore a multi-sensing approach to estimate the regions of the object that are going to be grasped by analysing the visual gazing performed by the subject during the preliminary moments of the grasp execution, as proposed by Flanagan et al. [9]. Other related works, such as Bohg et al. [10], show the analysis of human-grasping movements as the combination of a descriptor based on visual shape context with a non-linear classification algorithm that leads to the detection of stable grasping points for a variety of objects. According to recognition by components theory (RBC) [7], humans are able to recognise objects by separating them into geometric icons. Assuming that an object with a similar graspable part can be grasped in the same manner, for example, handbags, mugs etc. that are composed of curved parts like cylinders, it is possible to segment the objects in primitives for grasp planning, considering some shapes of single parts. The constituting parts of an object shape influence the choice of an object graspable part, independent of their orientations. The relative size of the object component is very important to select the graspable part [8]. In our work, a probabilistic description is used for the representation of 3D objects, which is then segmented in parts by approximating each object part using superquadrics primitives as proposed by El-Khoury and Sahbani [11]. In our representation, we associate data on object graspable parts such as contact points and tactile force obtained from demonstrations of in-hand manipulation with successful grasps.

2. Feature detection on multimodal data from human demonstrations

In this work, the human demonstrations play an important role in the ability of learning and identification of manipulation tasks, as well as to learn stable grasps. The data acquired will be used to model and extract the relevant aspects of the human demonstration, as well as for providing input for the methods presented in this work to represent the manipulation tasks and estimate the contact regions and candidate grasps for stable grasping on homogeneous [12] and dexterous manipulation tasks. The experimental activities with humans executing manipulation tasks are performed in our experimental area presented in Fig. 1. The experimental area is equipped with multiple data acquisition devices in order to capture the different types of data used by humans to perform successful manipulation tasks. The system records human gaze, 6D pose of hand and fingers, finger flexure, tactile forces distributed on the inside of the hand, colour images and stereo depth map. Using objects instrumented with inertial and force sensors, 6D pose and tactile forces on the object are also captured. An online database, the Handle Project — Data Collection Database [13], is publicly available with the datasets collected.

An essential step to achieve the learning and classification is the feature detection process, so that only the relevant information is used to represent a set of data. In this section, an overview of feature detection for trajectory identification and grasp transitions during tasks of in-hand manipulation is presented. A discrete representation of hand trajectory by curvatures and hand orientation along the trajectory is used. Grasps and grasp transitions are detected based on the spatial distribution and intensity of tactile forces on the hand.

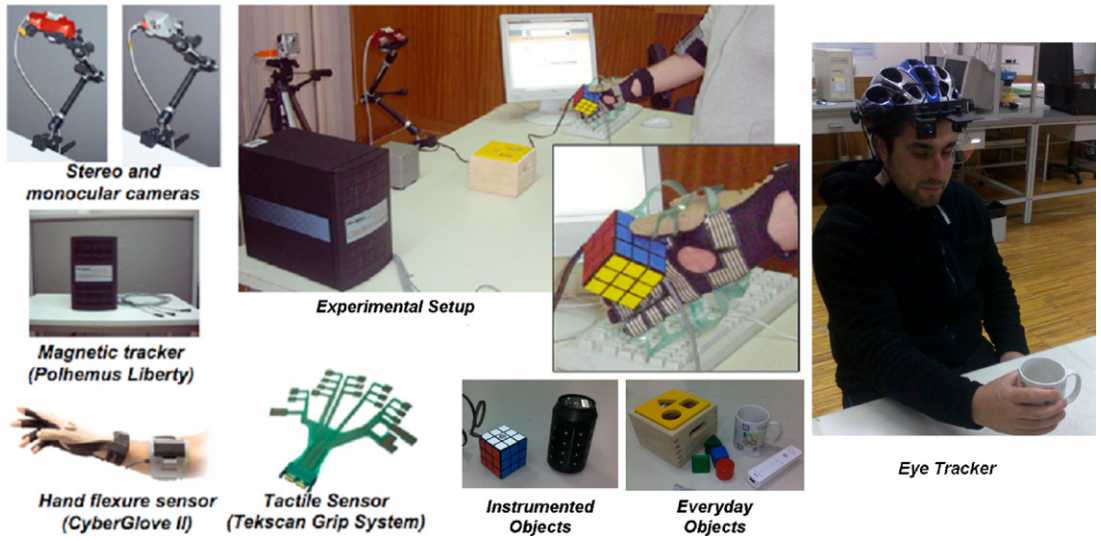

Fig. 1. Global overview of the experimental area, data acquisition devices and objects available.

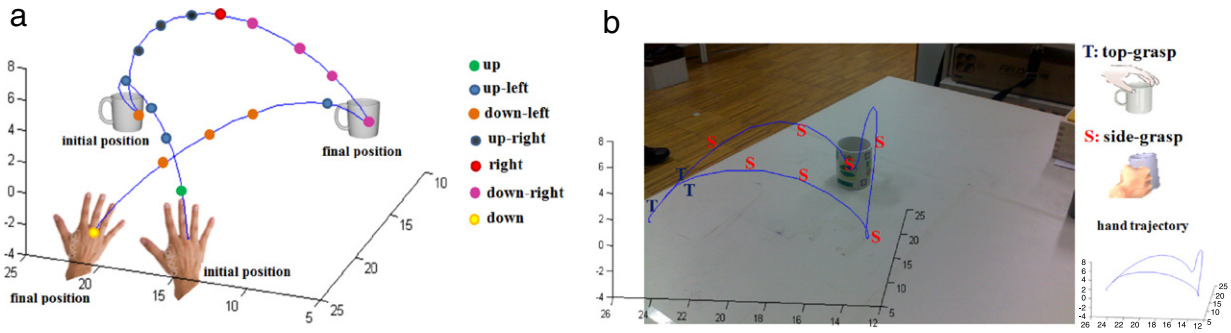

Fig. 2. (a) Example of a 3D trajectory of pick-up and place and possible curvatures along the trajectory. (b) Example of hand orientation along the trajectory.

2.1. Curvatures and hand orientation as trajectory features

In order to learn and characterise the hand trajectories, we are discretising them into significant changes in direction along the trajectories, hereafter named curvatures, and also detecting some pre-defined hand orientations with respect to a vertical reference. In this work, we are working with hand trajectories in 3D space. With our experimental setup, we have 6D pose data at 30 Hz given by the tracker device that is attached to the fingertips and on the back of the hand. A smoothing mean filter with a centred window of 9 samples is used, followed by a 0 to 1 scale normalisation using the initial and final points as reference. The trajectory is then segmented into action phases. The hand trajectory curvatures and hand orientation along each phase will be used to characterise each segment, so as to identify all the phases of the human manipulation demonstrations.

Along the hand trajectory, we are considering the discretisation of curvature along 8 key directions, i.e., $C \in \{up, down, left, right, up\text{-}left, up\text{-}right, down\text{-}left, down\text{-}right\}$. These are derived from the trajectory with a threshold on the level of significant change that triggers a new feature [14]. The hand orientation is represented as $O \in \{top, side, hand\text{-}out\}$, and derived from the plane formed by three fingers (index, middle and ring finger) [14].

Fig. 2 shows examples of features extraction along a hand trajectory: (a) illustrates the curvatures along a 3D trajectory (pick-up and place); (b) shows the hand orientation along the same trajectory.

2.2. Primitive detection for grasping transitions

Depending on the action phase (e.g. reach, load, lift, hold, replace, unload, release) of a typical manipulation movement, different types of signals (position/orientation of the fingers, distal phalanges and wrist, joints flexure level, tactile sensing) dynamically change their role and importance on the control of the object manipulation strategies. During the manipulation of objects, the contact signatures between the object and the different regions of the hand surface, as well as the configuration of the human hand joint flexure level, are important factors on the definition and characterisation of those strategies.

As demonstrated by the seminal work [15], it is possible to categorise the typical prehensile patterns of the hand that the humans use to hold daily life objects. Kamakura et al. [15] propose 14 patterns under 4 categories. The categorisation of the prehensile postures of the hand is made, taking in consideration the different regions of the hand that contact with the object during the interaction with it. Each contact tactile signature is the consequence of the mechanical configuration of the hand and implicitly the characteristics of the object being held. Kamakura et al. [15] just consider isolated prehensile patterns which are defined as the state of the hand in which the object is held without losing contact. The work also does not analyse the intensity of the contact.

In this work, although a manipulation task is segmented and modelled as a sequence of static prehensile primitives, those prehensile primitives are not analysed just for isolated statistical purposes. Besides the spatial configuration of the active contact areas,

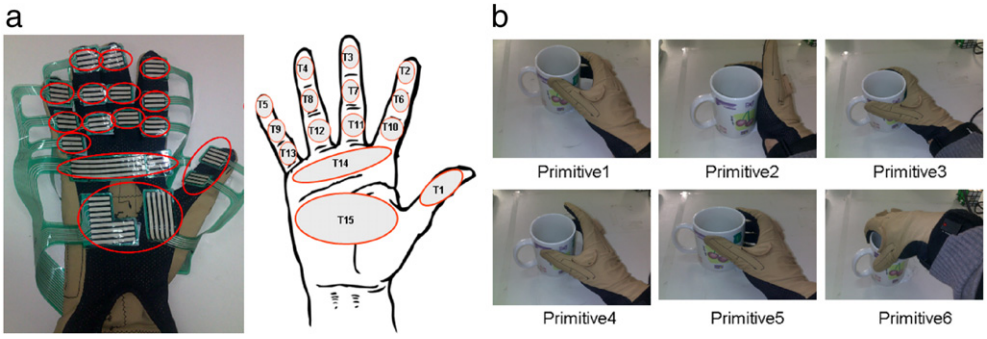

Fig. 3. (a) Representation of the 15 tactile regions defined in the human hand. (b) Six of the pre-defined grasp configurations used to estimate the static contact templates.

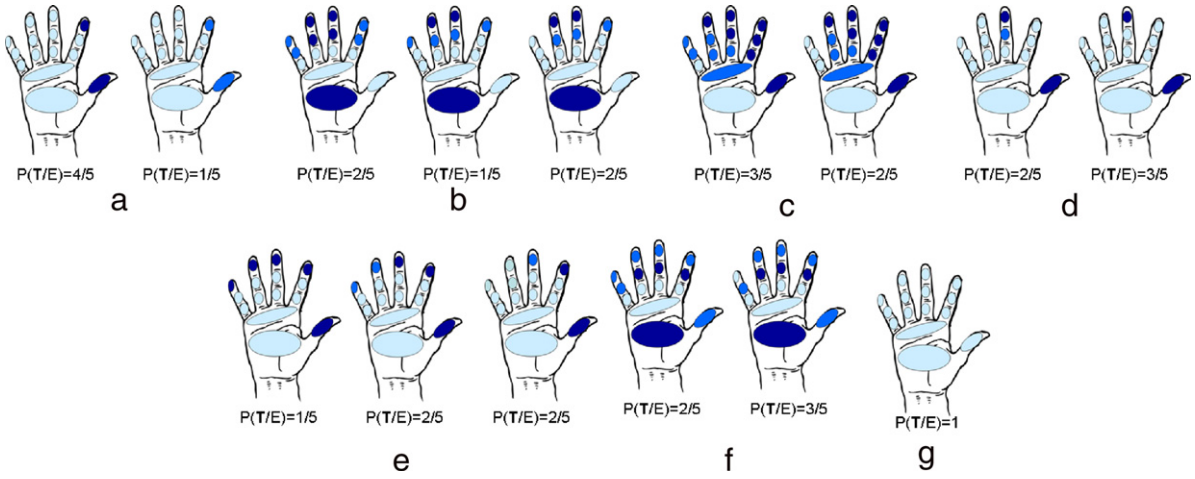

Fig. 4. Conditional probability density distribution of the primitives. (a) primitive 1; (b) primitive 2; (c) primitive 3; (d) primitive 4; (e) primitive 5; (f) primitive 6 and (g) primitive 7.

the prehensile primitives are also described by the intensity of the contacts. As analysed by Johansson and Flanagan [16], the tactile intensity is one of the fundamental variables to distinguish the different stages of a manipulation task. The same prehensile pattern class can proportionate different contact intensity configurations, depending on the stage of the manipulation task.

Using tactile sensor information, we can achieve a symbolic level generalisation of manipulation tasks by human demonstrations. In this section, tactile signatures of some manipulation primitives are addressed for future classification and identification of manipulation tasks. The tactile sensing device consists of 360 sensing elements (*Tekscan Grip System* sensor) which are distributed along the hand palm and fingers surface. The sensing elements are grouped in 15 regions as presented in Fig. 3, corresponding to different areas of the hand.

A variable T_i is assigned to each of these regions: $\mathbf{T} = \{T_1, T_2, \ldots, T_{15}\}$. The domain for each variable can be defined as $T_i \in \{NotActive, LowActive, HighActive\}$. The *NotActive*, *LowActive* and *HighActive* define the level of activation of that region during the in-hand manipulation task. Considering the raw sensor output 8-bit integer value, the *NotActive* state of a variable T_i corresponds to an average output of the sensing elements that is between 0 and 10, the *LowActive* is between 26 and 190, and the *HighActive* between 190 and 255. The general model of the framework used to describe the different templates of primitives can be defined by the set of variables **T**.

The three levels of discretisation of the tactile contact intensity of the hand (*NotActive, LowActive, HighActive*) was considered appropriate to characterise and distinguish the basic functional levels of activation of the hand for potential similar prehensile hand configurations, but in different stages of interaction with the object (e.g. *NotActive*: pre-grasp, transition between successive re-grasps; *LowActive*: initial contact with the object, hand region partially involved in this stage of the task; *HighActive*: hand region highly involved in this stage of the task). The seven tactile primitives were selected as a subset of tactile primitives that can be used to demonstrate the proposed concept in the type of manipulation tasks presented in this work.

During the experiments, the tactile sensing device was attached permanently to the glove, so all the subjects wear the tactile sensors attached to the glove in the same positions. A full calibration of the sensing elements is not required for the functional analysis of the hand regions involved in each grasp type. The activation level is discretised in three levels, and initially a uniform pressure is applied to normalise the scale factor across the sensing elements.

The set of pre-defined templates comprises a total of seven templates. The contact state templates primitives are estimated from seven different grasp configurations. Six of those seven grasp configurations are demonstrated in Fig. 3(b). The remaining one corresponds to when there is no contact between the hand and the object (*Primitive7*). The variable E indicates a primitive where $E \in \{Primitive1, Primitive2, \ldots, Primitive7\}$. In order to estimate the parameters of the template parameters **T** of each of the seven pre-defined primitives, several human demonstrations of the different static contact configurations of the human hand and the object were performed. The templates are extracted on each demonstration and the probability distribution $P(\mathbf{T}|E)$ is built.

In order to estimate the parameters of **T** for each of the seven pre-defined contact state template primitives, five demonstrations of each grasp configuration presented previously were performed by a subject. The conditional probability density distribution functions of **T** for each contact state template is shown in Fig. 4.

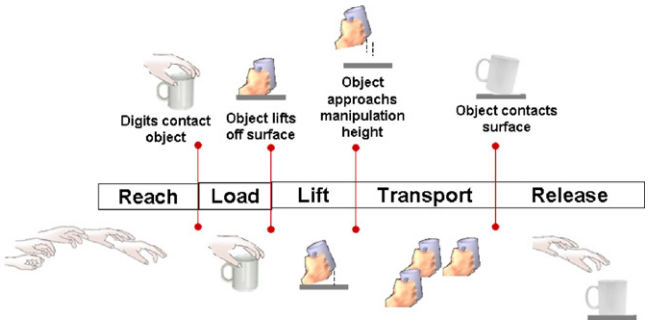

Fig. 5. Example of action phases in a simple homogeneous manipulation task, where the same grasp is employed during the manipulation.

3. Segmenting and identifying manipulation stages

Segmenting a task in action phases can help us to characterise each movement of a task as well as to understand the behaviour of the hand in each phase. By knowing the action phases of a task, we can discriminate easily a fixed grasp task (homogeneous manipulation) from a dexterous task due to the action transitions. Simple tasks (e.g. object displacement) can be composed of the following action phases: reach, load, lift, hold/transport, unload and release. A dexterous task is characterised by having the in-hand manipulation phase, where fine movements are performed with the intention of re-configuring the object state while it is being held by the hand. Dexterous tasks are composed of the following action phases: reach, load, lift, in-hand manipulation, unload and release. Fig. 5 illustrates an example of action phases in a simple homogeneous manipulation task, where the in-hand manipulation is replaced by a fixed grasp transport phase.

By observing the multimodal data, some assumptions can be made to find those phases during a task. For example, in the reaching phase, there is no object movement, the load phase is active when there is tactile information, and the transport phase when the object is moving. Since we have a synchronised data acquisition, by using the timestamps, we can analyse the multimodal data to know the state of each sensor in a specific time. Another option is segmenting by a probabilistic classification. Since we can extract features from the sensor signals, we can learn from multiple observations and then characterising each phase in a probabilistic way. Dealing with the uncertainty of sensor noise due to real world is a reason for adopting a probabilistic approach to automatically classify the action phases.

The next subsections present the probabilistic classification of manipulative tasks at the trajectory level and the symbolic representation dealing with tactile information to identify grasp transitions in dexterous tasks.

3.1. Hand trajectory classification

Following the segmentation of the hand trajectory of a given manipulation task, we characterise each action phase by looking at the hand pose along the trajectory. A classification step can be applied when a subject is performing a manipulation task, allowing the online identification of the trajectory class. A trajectory can be learned and later identified by using the Bayesian classification to reach degrees of belief. By continuous classification based on multiplicative updates of beliefs, hand trajectory phases may be classified by learning the curvatures and hand orientation along the trajectory to recognise a specific class of trajectories.

Bayesian models have already proven their usability for robotic perception and action [17]. Here we are addressing an example applied to trajectory classification. The probability of each type of grasp is updated for each hand displacement, i.e., we know what kind of manipulation task is more probable to happen by analysing the highest probability variable that indicates the trajectory types. Assuming that the trajectory was segmented by action phases, the probability distribution of the features that possibly identify the trajectory is computed at each hand displacement. To understand the general classification model, some definitions are required: g is a known manipulation task goal from all possible G (manipulation task classes, e.g. in case of simple tasks: object displacement or lift the object, reposing it in the same initial position.); c is a certain value of feature C (curvature types); i is a given index from all possible hand displacements composed of a distance D (corresponding to a segmented action phase). The learning phase provides the probability distribution of each class of features for all trajectories of a dataset. A learned table is computed (histogram), which is composed of the types of features and their probability distribution in each action phase of the trajectory. More details of the learning phase can be found in our previous work [14,18]. In this work, G refers to simple tasks, but it is possible to learn dexterous tasks (e.g. pick-up and write, toy sorting, pick-up the object, rotate and place it in other pose) for identification, and also use more relevant information during learning, such as grasp transitions as shown in 3.2.

The probability $P(c|gi)$ that a feature C has a certain value c can be defined by learning the probability distribution $P(C|Gi)$. The probability $P(o|gi)$ that a feature O has a certain value o can be defined by learning the probability distribution $P(O|Gi)$. Knowing $P(c|Gi)$, $P(o|Gi)$ and their priors $P(G)$, we are able to apply the Bayes rule and compute the probability distribution for G given a hand displacement i concerning the hand displacement of the learned table and the features c and o. Initially, the manipulation task variables (priors) $p(G)$ have a uniform distribution, and during the classification, their values are updated by applying the Bayes rule. The features O and C are independent. Each class of features are found in the same trajectory for classification. Having two feature classes for any given trajectory allows us to better characterise the type of trajectory. The trajectory classification step is shown in (1) and (2), for example, to identify pick and place (pp) and pick and lift (pl):

$$P(g_{pp}|c_{k+1}, o_{k+1}, i) = \frac{P(c_{k+1}, i|g_{pp})P(o_{k+1}, i|g_{pp})P(g_{pp})}{\sum_j P(c_{k+1}, i|g_j)P(o_{k+1}, i|g_j)P(g_j)} \quad (1)$$

$$P(g_{pl}|c_{k+1}, o_{k+1}, i) = \frac{P(c_{k+1}, i|g_{pl})P(o_{k+1}, i|g_{pl})P(g_{pl})}{\sum_j P(c_{k+1}, i|g_j)P(o_{k+1}, i|g_j)P(g_j)}. \quad (2)$$

In (1) and (2), the variable j is an index that represents all possible manipulation tasks. We formulate the equation in a recursive way. Assuming that for each hand displacement we can find new curvatures and new hand orientation, we can express the on-the-fly behaviour by using the index k that represents a certain displacement performed by the person during the hand movement. The classification is based on the highest probability, defined by a threshold (e.g. 0.7). We expect that a manipulation task movement that is being performed by a subject will produce a trajectory hypothesis with a significant probability. Other features could also eventually be used to identify a manipulation task, such as finger flexure, and tactile information (force intensity), but in our approach we need this task context and map the other features to an object centred model that contains geometric and grasp configuration and sequence data.

3.2. Grasp transition classification in in-hand manipulation tasks

After extracting the primitives regarding tactile signatures in in-hand manipulation tasks, a classification method is used to

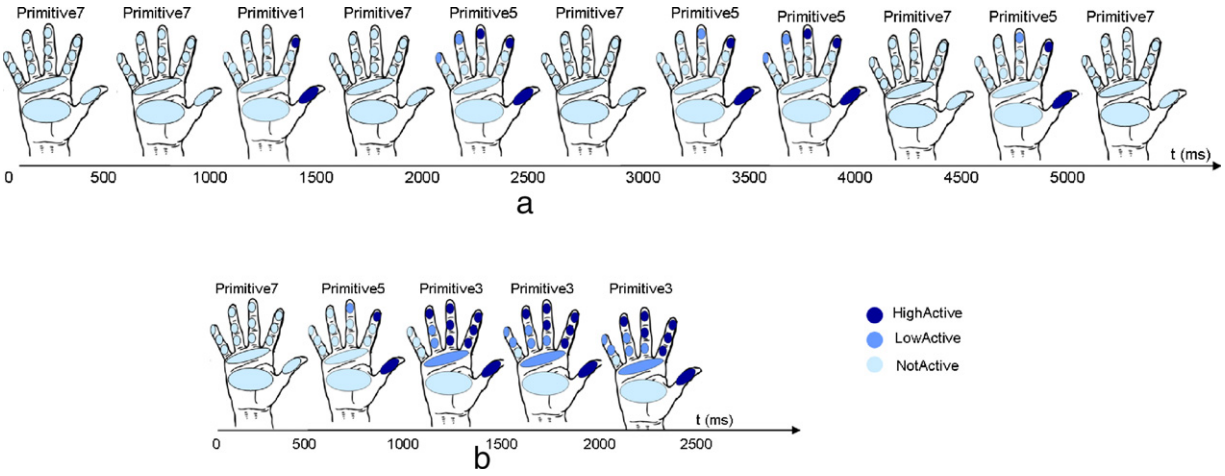

Fig. 6. Contact primitives detection on raw input data. (a) Task I – Mug reorientation; (b) Task II – Mug elevation.

Fig. 7. Task demonstration. (a) Task I – Mug reorientation (one cycle grasp-release cycle is shown); (b) Task II – Mug elevation.

identify the grasp transition in each action phase of a manipulation task.

In order to proceed to the detection of the pre-defined primitives, the tactile inputs produced during the in-hand manipulation demonstration are integrated during equal time intervals. The integrated data during one time slot Δt is classified according to the following expression:

$$P_{\Delta t}(E|\mathbf{T}) = \frac{P_{\Delta t}(\mathbf{T}|E)P(E)}{P_{\Delta t}(\mathbf{T})}, \tag{3}$$

where $P_{\Delta t}(\mathbf{T}|E)$ is achieved from the primitive demonstration training session. $P(E)$ is the probability of a template, and $P_{\Delta t}(\mathbf{T})$ is the probability of a model measurement. The template E with maximum likelihood is the template assigned to that time slot. The previous expression can be rewritten as follows:

$$P_{\Delta t}(E = Primitive_i|\mathbf{T} = (T_1, \ldots, T_{15})) = \frac{P_{\Delta t}(\mathbf{T} = (T_1, \ldots, T_{15})|E = Primitive_i)P(E = Primitive_i)}{\sum_{j=1}^{7} P_{\Delta t}(\mathbf{T} = (T_1, \ldots, T_{15})|E = Primitive_j)P(E = Primitive_j)}. \tag{4}$$

The output of the primitives detection stage is a raw temporal sequence of the templates corresponding to the pre-defined primitives. In order to test the primitives detection approach of the pre-defined contact state templates, two tasks were defined. For both tasks, the manipulated object is a mug and the starting configuration (position and relative orientation to the subject) is the same. *Task I* consists of the reorientation of the mug in order to position the grasp of the mug in a configuration suitable to be grasped by the handle by the subject right hand. *Task II* consists of grasping the mug without reorientation and elevating it. The detection of the primitives is made by using the average value over each of the **T** tactile inputs. The results for the primitives detection on raw data inputs for *Task I* and *Task II* are shown in Fig. 6, segmented in 0.5 s blocks. Fig. 7 shows some frames of the *TaskI* and *Task II* demonstration.

The estimated primitive corresponding to the input data of each segment is made by calculating, for each primitive in the redefined set, $P_{\Delta t}(E = Primitive_i|\mathbf{T} = (T_1, \ldots, T_{15}))$, given the input tactile data **T** and selecting the primitive that maximises the previous expression.

The input data (i.e. human demonstration) are sequentially fragmented by the primitives detection algorithm. Typically, the first segments correspond to the template of *Primitive7*, when there is no contact between the hand and the object. This period corresponds to the movement of the hand towards the object that is going to be manipulated.

Task I manipulation movements were segmented in a repetitive sequence of grasping and release of the object in order to reorientate the mug placed on the top of the table to be grasped correctly. This sequence of grasp-release allows the subject performing the experiment to reposition the hand on the object, adapting the grasp configuration to the new pose of the object, to maximise the effectiveness of the subsequent hand actuation on the object. The fingers involved on the reorientation of the mug are predominantly the thumb, index and middle fingers. The ring and little fingers have a less intensive participation on those movements, although the assigned primitive is the same.

The second manipulation task (*Task* II) was decomposed on a series of primitives that involved the participation of high

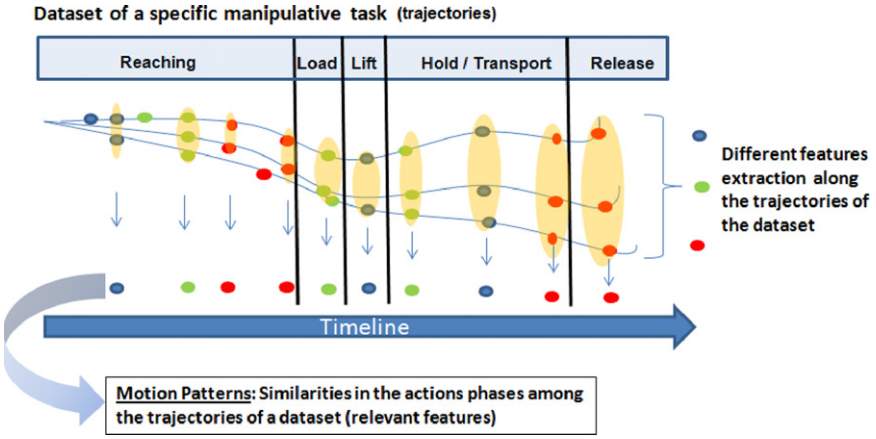

Fig. 8. Motion patterns: similarities detection in the action phases of the trajectories in a dataset of a manipulation task.

extensions of the finger's surface. The task does not require the execution of grasp-release sequences of movements.

3.3. Detecting motion patterns and segmenting action phases

An important step to model the human actions and behaviours is the motion pattern detection during an activity. In different daily tasks, the motion assumes an important key point to describe a specific action. The variety of human activity in everyday environment is very diverse; the same way that repeated performances of the same activity by the same subject can vary, similar activities performed by different individuals are also slightly different. If a particular motion pattern appears many times in long-term observations, this pattern must be meaningful to a user or to a task. In this work, we are focusing on manipulation tasks at the trajectory level to find similarities (significant patterns) given by multiple observations. The intention is to learn and to generalise a specific task by the hand movement, including finger motion as well as object trajectory during the task. This is useful for task recognition in robot imitation learning, and it can be applied in the future in such a way that the generalised movements can be used in other contexts by a robot.

To model a manipulation task, it is necessary to extract relevant information (patterns) from multiple observations. By looking for similarities among the features of a dataset of trajectories, it is possible to represent the dataset by its relevant features as illustrated in Fig. 8. The relevant information are repeated motion patterns that are used to generate a generalised trajectory.

In this work, we use some grasp classes to estimate the grasp type along the task to estimate the grasp transitions when a human is manipulating the object. In each task, it is necessary to identify the types of the grasping/gesture defined and then compute the probability distribution $P(Grasp|Observation)$ of each one along the action phases of the task for each trajectory by analysing the grasping occurrences.

The dataset of trajectories are aligned temporally, applying as a pre-processing step Dynamic Time Warping (DTW), a pattern-based method that allows the sequential information description of the data by the temporal distortion between different examples [19]. The next step is to detect the features and compute the probability distribution of the feature occurrences. Then similarities in all trajectories of a dataset are found, i.e. features with high probability (high occurrence in all the trajectories). A threshold is set on this probability to obtain a set of relevant features. The representation of a dataset of a specific task at trajectory level is given by the general form of the data. It is obtained after selecting the relevant features and then applying a regression on the spatial information of the relevant features.

3.4. Trajectory generalisation for task representation

There are some possibilities to achieve the general form (a smoothed trajectory) of a dataset of trajectories. The first one is an interpolation applied after the features selection (similarities between trajectories) as a function of arc length along a space curve using parametric splines. The second way is by using the spatio-temporal information of all features extracted from all trajectories of a dataset, where a polynomial regression is applied to fit the data to have a smoothed trajectory. The polynomial regression can be a good choice due to the curvilinear response during the fit and it can be adjusted because it is a special case of the multiple linear regressions model. In case of applying regression, to have a correct fit, the regression need to be done locally, at subregions of the trajectory due to the shape of trajectories. In general, for our data, a cubic order polynomial regression is enough for the fitting. In this type of curvilinear regression, the choice of degree and the evaluation of the quality of the fitting depend on an empirical analysis. Although polynomial regression fits a non-linear model to the data, as a statistical estimation problem, it is linear, in the sense that the regression function is linear in the unknown parameters that are estimated from the data. It is based on least square fitting.

The hand motion generalisation is useful to represent a task. For each dataset, we intend to have a generalised data to be used in the future to endow a robot to perform the generalised movement. Each task will be represented by the generalised hand trajectory combined with the learned force intensities, grasp transition and contact points for a stable grasp in each action phase of the task.

4. Object probabilistic volumetric map

Some grasping strategies for robotic systems are based on analysing object geometric properties and fitting suitable grasps, others on learning from human demonstrations for specific objects. In our approach, we try to encompass both concepts, using a volumetric map of the object, overlaid with data from human demonstrations in a probabilistic framework. Although this mapping is for specific objects, the representation can be used to match with partially observed new unknown objects that have similar geometric distributions.

In our work, a probabilistic representation of the object is used to have prior information of the object used in a manipulation task. In our previous work [20], we presented our in-hand exploration method of obtaining this object volumetric data, where contour following was used. This "exploratory procedure" is what humans also use for determining the geometry of an object [6]. For better representation of the object, in [20], the visual information

complements the 3D model with texture information. The model is based on probabilistic occupancy grid methods [21–23], used in robotics for 2D mapping and extended to 3D. In this work, we extend the previous proposed approach to learn the contact regions location on the object for a stable grasp as presented in Section 4.5. Since we already have the 3D representation of the object in a volumetric map, we can overlay the relevant data about human visual gaze, and contact points on the object surface.

When performing in-hand exploration of objects, the key idea is to use the hand to extract object geometrical information. During the in-hand exploration, the object might be moved or even released and re-grasped, for example, when one uses the other hand to assist the hand performing the in-hand exploration. This task becomes more complex than exploring objects fixed in a specific position. To deal with moving objects during the in-hand exploration, the object rotation and translation need to be taken into consideration. Knowing the object initial position and object displacements, we can compute the transformations to have all points in the same frame of reference. In our case, we have a *6DoF* sensor attached to the object so that we can map the hand contact points to an object centred frame of reference and properly register the point clouds to build the object model.

4.1. Probabilistic volumetric map

The occupancy of each individual voxel in the map is assumed to be independent from the other voxels' occupancy and thus O_C is a set of independent random variables, where $_C$ is the cell index and O_C the value indicating the level of occupancy of the cell. The in-hand exploration measure Z_{grasp} updates the occupancy knowledge. Initially, we have a uniform distribution for $P(O_C)$. Since no preliminary knowledge is available. $P(Z_{grasp}|O_C)$ indicates the probability of having a measurement Z_{grasp}, given the occupancy.

We model the measurement error due to sensor noise with a Gaussian distribution along with the sensor measurements. Due to the size of each cell relative to the standard deviation of the magnetic tracking sensor's measurements (up to 4 mm), inside each cell we consider a 3D isotropic Gaussian probability distribution, $P(Z_{grasp}|O_C)$, centred at the cell central point with standard deviation 0.4 cm and mean value equal to the cell central point coordinates of the cell. The probability distribution given the sensor's measurements is given by:

$$P(Z_{grasp}|[O_C = 1]) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}, \quad (5)$$

where $P(Z_{grasp}|O_C)$ represents the probability distribution of the sensor measurement given O_C and $|\Sigma|$ represents the determinant of Σ, the sensor noise variance.

The probability distribution on the occupancy $P(O_C|Z_{grasp})$ for each voxel is given by:

$$P(O_C|Z_{grasp}) = \frac{P(Z_{grasp}|[O_C = 1])P([O_C = 1])}{P(Z_{grasp}|[O_C = 0])P([O_C = 0]) + P(Z_{grasp}|[O_C = 1])P([O_C = 1])}, \quad (6)$$

where $P([O_C = 0]) = 1 - P([O_C = 1])$; $P(Z_{grasp}|[O_C = 1])$ is given by (5) and $P(Z_{grasp}|[O_C = 0])$ is a uniform distribution.

When using vision, the sensor model $P(Z_{vision}|O_C)$ needs to be defined. Stereoscopic systems are usually implemented as deterministic algorithms returning visual properties like range values. Adopting the solution proposed by Rocha et al. [24], we can have the voxel occupancy belief as the Gaussian distribution, where the range distance between the sensor and the detected obstacle, and the distance between the sensor and the voxel centre are used to compute the occupancy. This solution relies on sensor calibration to estimate global values for sensor model parameters. The probability distribution on the occupation probability $P(O_C|Z_{vision})$ for each voxel is similar to the one used for in-hand exploration [20].

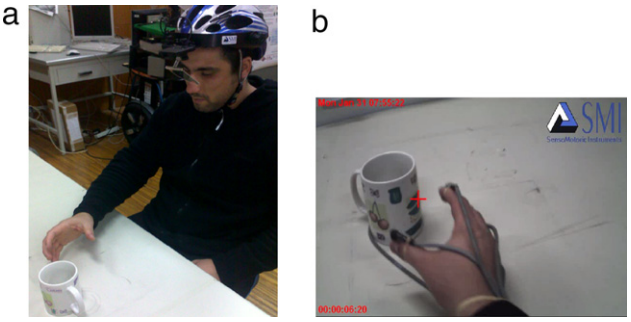

Fig. 9. Eye tracker: (a) subject performing a manipulation using the eye tracker; (b) typical output of the eye tracker. Red cross indicates the estimated gaze direction.

4.2. Multi-modality and fusion

In this work, the analysis of the stable grasps executed by humans during manipulation tasks is performed using a multimodal approach, in order to capture the multiple signals and strategies. An object centred probabilistic volumetric model is used to represent the multimodal data and map contact regions, gaze and tactile forces during stable grasps. One aspect that characterises the manipulation task is the trajectory described by the hand (fingers, palm, wrist) to reach and contact the object, in order to perform the initial stable grasp. The location of the contact points of the fingertips in the object surface are acquired using *Polhemus Liberty* motion tracking system.

The biological signals related to tactile inputs are also relevant to do the fine control of the manipulation tasks. The information about the level of activity of each region of the hand, during the contact with the object while the initial stable grasp, is acquired using the tactile sensing array *Tekscan Grip System* and the methods presented in Sections 2 and 3.

The gaze has been used as an analysis tool of physiological responses to stimuli as an indication of cognition. The gaze in response to visual, auditory or cognitive stimulus is measured, during the manipulation task, using an *SMI iView* eye tracker device. The gaze provides important cues about the strategies used to find and anticipate the appropriated region of the object to be grasped. The eye tracking system uses infra-red illumination and computer-based image processing. The pupil is detected and after calibration, the pupil centre location is translated into gaze data. The gaze direction is mapped by the system in the scene images by a red cross as presented in Fig. 9.

4.3. Object modelling using entropy as confidence level

To combine more than one sensing modality to obtain the object model, we need to compute the posterior of the occupation probability given the observations, $P(Oc|Z_{vis}Z_{grasp})$, for each voxel.

We are adopting entropy as a confidence level of the sensor models, by computing weights to perform late fusion as mixture models. The weights are computed using each entropy value computed from (7) for each local map, for example, vision and in-hand exploration. In a Bayesian framework, each model contributes to the result of the inference in proportion to its probability. Mixture models are usually presented directly as weighted sums of distributions. Through the mixture model, we can achieve the combination of different models into one. Examples of weighing mixture

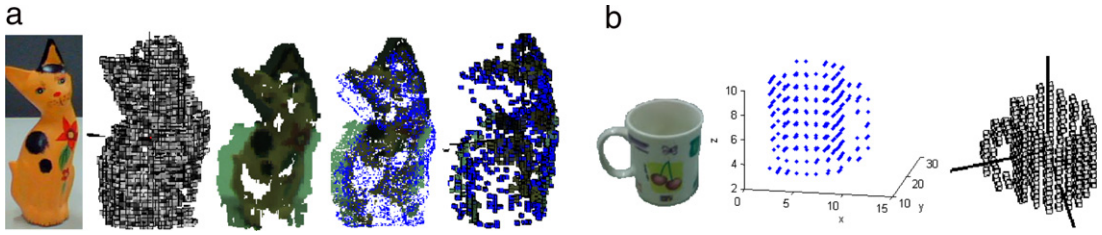

Fig. 10. Probabilistic representation of objects derived from in-hand exploration. (a) A wooden cat; computed probabilistic map; visual information (textured points cloud); probabilistic map with texture information (only voxels with probability higher than 0.8). (b) Image of a mug following the view of the occupied voxel's central points, and the last image shows the computed map derived from in-hand exploration (object-centred representation).

models using the Bayesian framework can be found in [25]. Here, the intention is to update a global map by looking at the different sensor data to know the confidence of each sensor.

Through Bayesian techniques, we can implement the sensor fusion and use entropy H as a confidence level. A confidence variable w will be used as the weight for each sensor. The weight w can be expressed as a prior $P(w)$ in the Bayesian rule. For each sensor, we can compute the entropy of the posterior probabilities as follows:

$$H(P(Oc|Z)) = -\sum_{c} P(O_c|Z)\log(P(Oc|Z)), \tag{7}$$

where $P(O_C|Z)$ represents the posterior probability of the occupancy of each cell in the map achieved by a specific sensor. The variable Z represents the sensor measurements and c is the index of each grid cell. Through the entropy H, we can achieve the probability distribution of the weights of each sensor. The weights are computed as follows:

$$w = 1 - \left(\frac{h}{\sum_{i=0}^{n} H_i}\right), \tag{8}$$

where w is the weight result; h is the current value of entropy that is being transformed in a weight; i is the index for each entropy value computed by (7).

Given the confidence of the occupied cells achieved by each sensor, we can fuse the sensor's belief by multiplying each local map to the correspondent sensor's weight reached by the entropy. For each cell of the volumetric map we can compute the mixture model belief for local maps fusion:

$$P(O_c|Z_1, \ldots, Z_S) = \sum_{i=1}^{S} P(w_i)P(Oc|Z_i), \tag{9}$$

where S represents the number of sensors.

Using (9), we update a global map with the probability distribution of each cell achieved by different sensors for data fusion. By employing entropy as the confidence level, we will be sure of the confidence of each sensor, that is, which is more reliable and then we build the global map from local maps (vision and in-hand exploration) with more certainty of the measures of the sensors. The only concern that needs to be taken into consideration on using the methodology proposed is the computational cost due to the necessity of calculating (7) and (9) for each cell.

A calibration between the sensors is needed to work with the local maps and the global one in the same frame of reference. In the previous work [18], we have presented an approach for the sensor's calibration (magnetic tracker and vision system), following the approach presented in [26].

4.4. Frame of reference for object-centred representation

We are adopting an object-centred representation by estimating the frame of reference of each object by its geometrical properties. For that, we compute the 3D moment invariants to find the centroid of the point cloud which depends on the distribution of the points of the object surface. The centroid will be located at the densest part of the point cloud. The 3D moment invariants are a measure of the spatial distribution of the mass of a shape. Let $p(x, y, z)$ be a local continuous density function which is represented by the probability of a voxel to be occupied (e.g. occupied, $p(x, y, z) \geq 0.7$; empty, otherwise). To estimate the location of the centroid of the point cloud, we first compute the *zero*th moment (sum of the voxels' probabilities) followed by the first moments for each axis x, y and z (sum of the product of all x by the probability of the respective voxel being occupied; the same for y and z). Then the centroid (c_x, c_y, c_z) is computed by the normalisation of each c by the *zero*th moment. The centroid is useful not only to define the frame of reference of the object, but it is used later with the contact points location to estimate how the object was grasped.

Fig. 10 shows examples of in-hand exploration of some objects and corresponding probabilistic volumetric maps.

In the next subsection, the contact points acquired from human demonstrations are overlaid on the object surface to store data on how the object is grasped.

4.5. Candidate contact points on object surface

The contact points for stable grasp is acquired by human demonstrations during a manipulation task. The fingers' locations given by the tracker device when in contact with the object surface (i.e. when the tactile sensors are activated) are overlaid to the object model to represent the contact points in the object point of view.

The object models used in the manipulation tasks are previously achieved using the approach presented of object probabilistic volumetric map. Since we know the object initial position (given by the tracker sensor attached to the object), the fingers' positions are easily overlaid to the object model to have the candidate contact points of stable grasp along the manipulation task. Using the tactile information we know exactly the instant that the fingertips touch the object and the force intensity applied on the object. This information is possible to acquire due to the timestamps used in our synchronised data acquisition.

Fig. 11 shows the contact points acquired during a human demonstration of stable grasps. The contact points were acquired during the load phase of a fixed grasp task (pick-up and place) overlaid in the object model. We can see different demonstrations of contact points for top-grasp (precision grip) and side-grasp.

The contact points between the human hand and the object surface are represented in the object-centred volumetric map during the load phase. The estimation of the location of those contact regions is made by combining two approaches. In one of

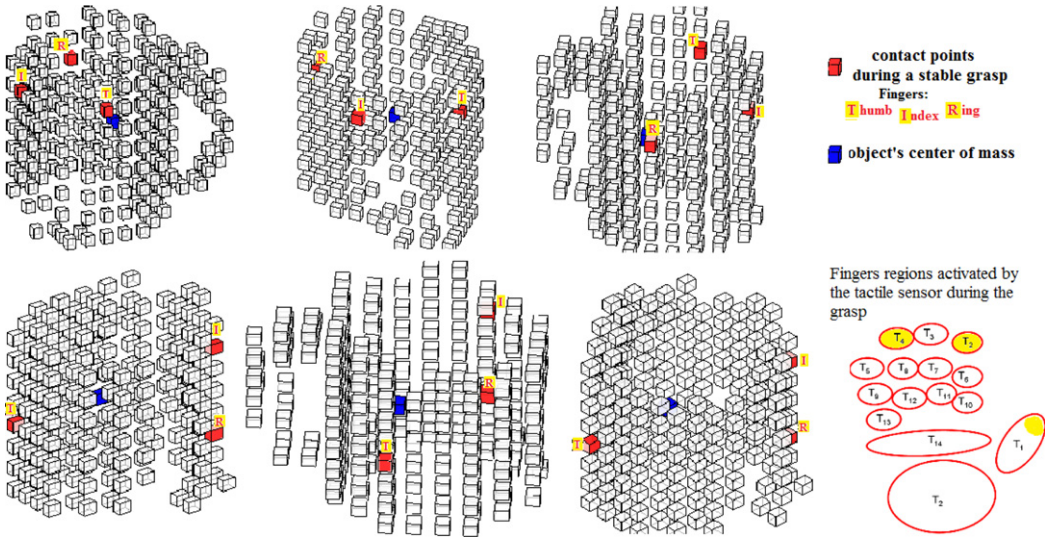

Fig. 11. Contact points on the surface of the object. Example of top-grasp (precision grip) — the images on the top are different views of the same contact points; Only three fingers are shown representing the thumb (T); index (I); and ring finger (R). By using the regions defined of the tactile sensing, we can see the regions of the hand that touched the object during the stable grasp. Bottom images: Example of side-grasp — different views of the same contact points; In these examples, only three fingers are shown representing the thumb (T); index (I); and ring finger (R).

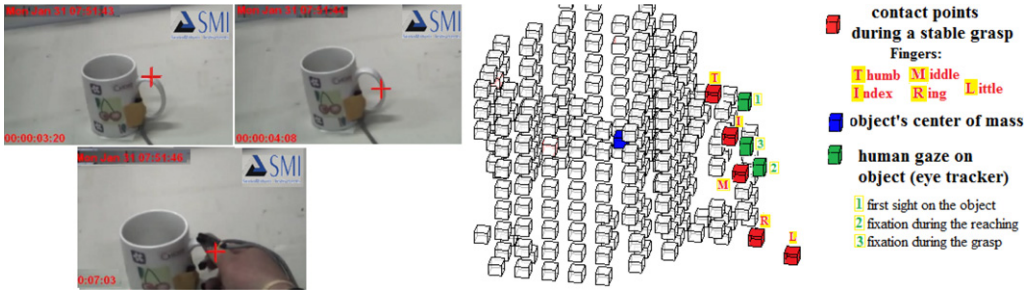

Fig. 12. Human gaze during grasping and the contact points on the object surface. Task: Reaching and Grasping by the Object Handle. The visual gaze during the grasping shows that the human usually looks to the region of the object where will be performed the grasp.

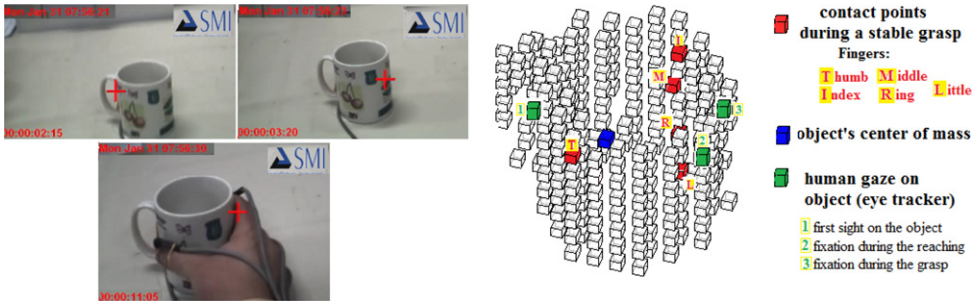

Fig. 13. Human gaze during grasping and the contact points on the object surface. Task: Reaching and Grasping the mug by side-grasp. This type of grasping was chosen due to the orientation of the object — it influences the type of grasping.

the approaches, the location of the fingertips is mapped in the volumetric map. The contact point positions are known when there is tactile intensities information. The second approach is by using an eye tracker to estimate the regions of the object that are observed by the subject while doing the reach motion planning and during the load phase. Fig. 12 shows some snapshots of the estimated observed regions during the manipulation of a mug, placed in a configuration where the handle of the mug is completely visible by the subject. The volumetric map represents both the observed regions of the mug and the regions which were effectively grasped. Fig. 13 represents the results achieved in a situation where the handle of the mug was not completely visible to the subject. Although, during initial instants, the attention of the subject is captured by the partially visible handle of the mug, due to its inaccessibility, the subject chooses to grasp the mug using a side grasp applied to the lateral regions of the mug.

Given the location of the contact points, we can infer the associated human hand grasp type. The average of the sum of squared distances between the contact points, for example, thumb to index, index to middle and so on, D_f, is computed. Then, the average of the sum of the squared distance of each isolated contact point to the centre of mass of the object, D_c, is computed. These two resulting values characterise a grasp type. For each grasp type $g \in \{\textit{top-grasp}, \textit{side-grasp}, \textit{hand-out}, \textit{grasp-by-handle}\}$, we have thresholds obtained from statistics of D_f and D_c from labelled datasets of grasp types and contact points that allow us

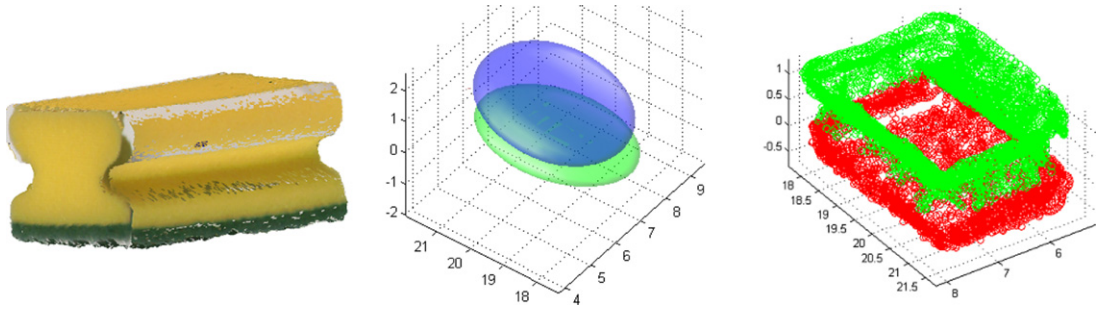

Fig. 14. Object segmentation of a sponge with the method proposed, using $K_{max} = 2$ and 3D points acquired by in-hand exploration of the object surface. On the left is an image of the object; in the middle the 2 clusters (Gaussian density functions) obtained from the segmentation process; and on the right the colour labelled 3D points of the segmented clusters.

to classify the grasps observed. This approach works well for the set of grasps defined, since the measures used encompass the object–hand relationship as well as the finger to finger spread. For small sets of predefined grasp types, it is enough to discriminate.

4.6. Segmentation of key object components

Since we have a probabilistic representation of the object model, it is possible to find geometrical primitives from the object. We are clustering the output of the probabilistic volumetric map to find the possible object components. By clustering, we can achieve outlier removal and we can also keep the position and size information of the object. According to the points cloud disposition, using the known method of Gaussian mixture models (GMM), we can find the most suitable clustering that will represent a component of the object. The intention is to simplify the global object shape in components, approximating it in basic primitives.

Gaussian mixture models have proven to be good models for points clustering where each cluster corresponds to a Gaussian function. Therefore, given a set of points, it is possible to find the GMM using the well-known method, Expectation Maximisation (EM).

A big issue that is raised in the scientific community when dealing with GMM is how to select suitable components given a point cloud, that is, the number of clusters K. This is an issue that the researchers are still working with. In the literature, we can find methods that can select the best number of clusters given the data. An example is the Bayesian Information Criterion (BIC), first suggested by [27]. Another possibility for the number of clusters selection is the Minimum Description Length (MDL) penalty function [28]. For both methods, it is necessary to set a maximum number of clusters, K_{max}, and then the method finds the appropriate number of clusters. In this work we use $K_{max} = 3$ and the MDL penalty function [28] to select the number of clusters K.

The basic idea here is to compute the Gaussian density functions given the volumetric model, clustering the relevant points reaching the segmentation of the global object shape in components that will be geometrical primitives later.

For test cases, we are using daily objects, most of which having simple shapes, allowing different candidate grasps. In case of object perception, this approach takes in consideration the variability of the in-hand manipulation motion among different subjects, as well as, the noise and uncertainty of the sensor measurements, by creating an initial probabilistic representation of the explored object based on a volumetric map. This initial step eliminates hypothetical erroneous data that could induce wrong representation of the object with consequent implications during the primitives segmentation phase. After some tests segmenting different daily objects, we could see that 3 components were enough to represent simple objects.

Fig. 14 shows a segmentation example of object components. The segmentation was tested using the points cloud given by occupied cells in the probabilistic volumetric map.

Fig. 15 shows more examples of object component segmentation by clustering the regions of the object. The clustering is done based on the distribution of the points, so that, for the same object, the number of points used to represent the model of the object can influence the segmentation process. For the same object, but acquired in different modality, e.g. in-hand exploration or visual information or laser scanner, can have similar or different segmentation results depending on the 3D points cloud structure. Usually, when the points cloud of the same object with the same sensor modality is acquired several times, the segmentation is very similar, and we can have similar results when the shape approximation by superquadrics is generated. In Fig. 15, we can see daily objects acquired with different sensors, some with more points than others. The coloured region in each object represents the points belonging to the same GMM cluster. This figure shows different segmentations for the same object due to the different number of points, e.g. by laser scanner, usually we have objects close to 150000 points; by in-hand exploration, 500 to 10000 points (it also depends on the exploration time).

4.7. Shape approximation using superquadrics

Since we have the object segmented in components, we now approximate each part by a geometrical primitive. For the extraction of these primitives, also known as geons, we are adopting superquadrics [29,30]. The advantage of using this methodology is the higher variety of shape options and also due to the facility of computing the parameters of the superquadrics that enclose important cues such as scale and orientation.

The 3D points' data need to be fitted to a superquadric model to represent a primitive. In our case, the points of each Gaussian density function will represent a shape. To estimate the parameters of the superquadric model, the gradient least-square minimisation based on Levenberg–Marquardt method [30] is used.

Fig. 16 shows an example of superquadrics generated after applying the object segmentation process.

After the object perception process, we can identify the object graspable part by using the information acquired during human demonstrations (contact points to know the object region where the grasping was performed). In the next subsection, we give details of how to find the object graspable part.

5. Object centred framework for manipulation knowledge

In the previous sections, we addressed how features extracted from hand trajectories and tactile data of in-hand manipulation could enable the segmentation and classifications of the action

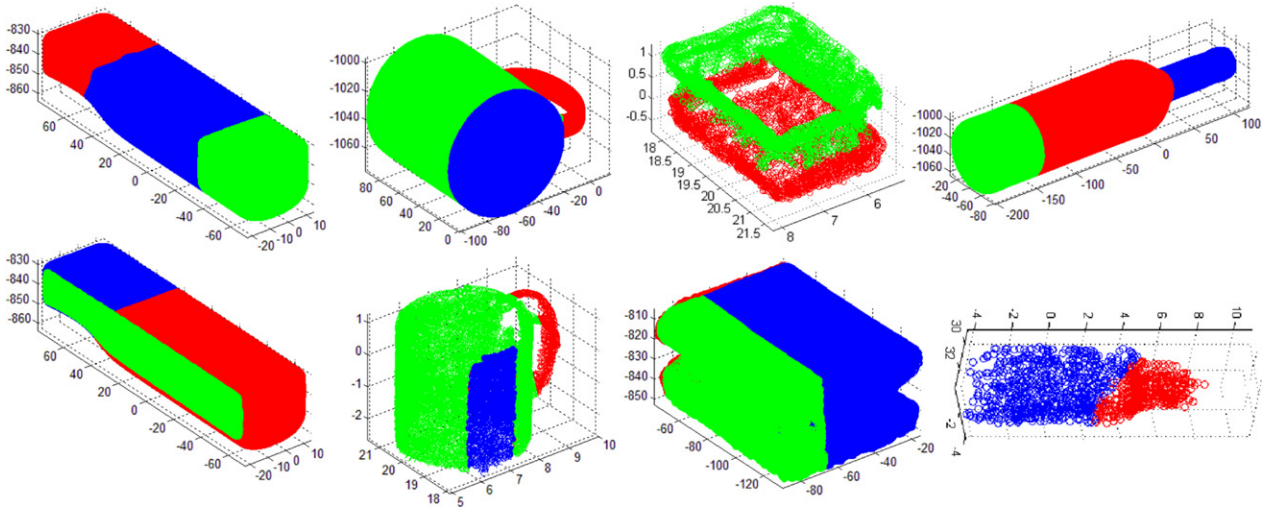

Fig. 15. Daily objects (*wii-mote*, mug, sponge and bottle) segmented using the method proposed. The same objects were acquired by different modalities with diverse point densities that resulted in different segmentations of components.

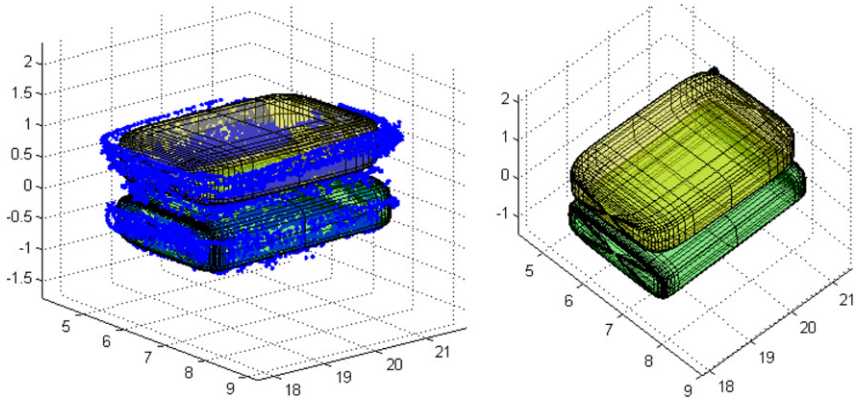

Fig. 16. Object (sponge) shape approximation using superquadrics after the segmentation process shown in Fig. 14. Left image shows the superquadrics on the 3D points cloud of the computed probabilistic map and (b) are the superquadrics (boxes) generated for this object.

phases and corresponding grasp transitions, and also how an object probabilistic volumetric map could be constructed. This will now be unified in a framework that associates to the object probabilistic model the hand approach vectors, initial grasps, and sequencing of grasps during in-hand dexterous manipulation. The rational behind this framework is that artificial systems when confronted with objects can first perform a match of the partial shape observed with the volumetric map, and use the data in the framework to key the possible approach trajectories and grasps span for manipulating the object.

When searching in the framework for object graspable parts, we also need to take into account the task context. Humans not only make some type of segmentation and identification of object parts in order to choose the best place to grasp, but are also task oriented in this choice.

From the multimodal data, we are able to extract relevant information of the manipulation tasks performed, such as different phases of manipulation during the hand trajectory. From the hand trajectories and tactile data, we identify the grasp types and transitions. These are mapped onto the object probabilistic volumetric model, so as to retain the relevant data from human demonstrations, concerning both the manipulation and object characteristics. The object centred framework will facilitate future matching for an artificial system observing objects and searching for cues on how to grasp it, and also taking into account the task context.

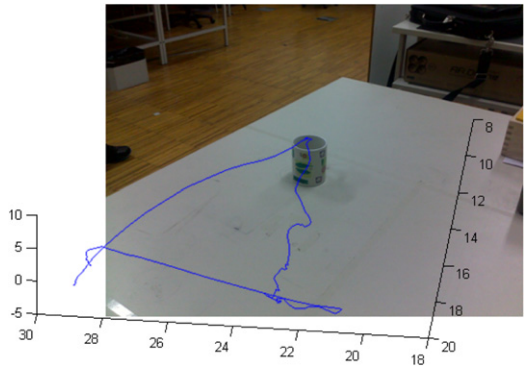

Fig. 17. Hand trajectory (sensor attached to the back of the hand) during the in-hand manipulation task (pick-up the mug, rotate and release it).

Figs. 17–19 present some of the data collected in this framework. Fig. 17 shows the hand trajectory during the in-hand manipulation task. The task is pick-up the object, rotate and repose it in another location. Fig. 18(a) shows the object trajectory where it is possible to visualise the object rotation during its trajectory for the same task; and (b) shows some transitions of the hand shape during the same task. Fig. 19 shows the object point of view, that is, the sequence of some contact points location during the in-hand manipulation phase.

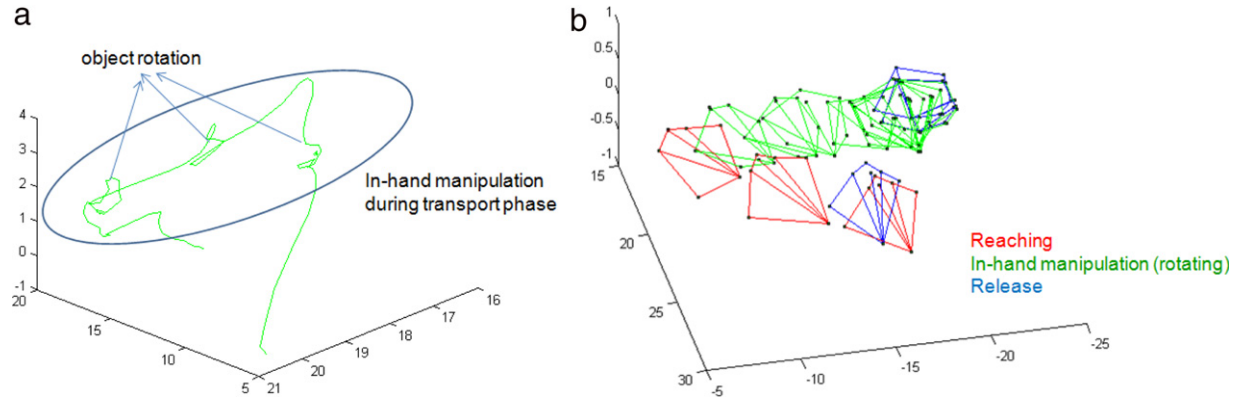

Fig. 18. (a) Object trajectory during the in-hand manipulation task (green); the blue circle shows the in-hand manipulation phase (object rotation along the trajectory.) (b) the graphs represent some transitions of hand shapes during the in-hand manipulation task. The nodes represent the fingertips and top of the hand locations.

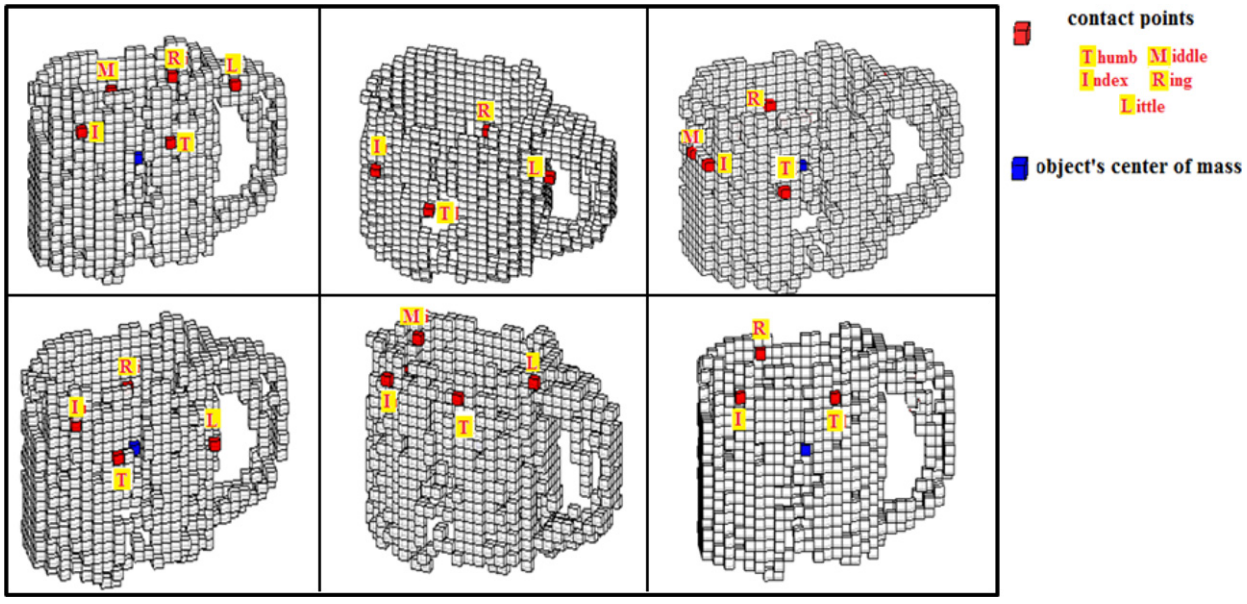

Fig. 19. Sequence of some contact points overlaid in the static representation of the object (volumetric map) during the manipulation task. The contact points are given by the fingertips locations during the in-hand manipulation phase.

From the human demonstrations, we obtain the task context for which distinct grasp types and object graspable parts where used. Given a set of observations to represent a type of task $\mathcal{T}$, with $\mathcal{T} \in$ {pick-up and place; pick-up and lift; pick-up and pour/tilt}, we have the probability of each type of grasp G represented as $P(G|\mathcal{T})$. For the object graspable part, we can identify the object component from the locations of the contact points on the object surface where the grasp was performed. Given a set of observations to represent a type of task $\mathcal{T}$, we have the probability of each object component geometrical primitive $C \in \{prim_1, prim_2, prim_3\}$ being the graspable part $P(C|\mathcal{T})$. The probability distributions are obtained from the occurrence statistics in datasets of the given task. Given a task-context, we can estimate the object graspable part B as follows:

$$P(B = prim_i|\mathcal{T}) = \frac{P(\mathcal{T}|C = prim_i)P(C = prim_i)}{\sum_j P(\mathcal{T}|C = prim_j)P(C = prim_j)}, \quad (10)$$

where the posterior information $P(B = prim_i|\mathcal{T}\ C)$ is computed for each object primitive C of a specific task $\mathcal{T}$; the likelihood $P(\mathcal{T}\ C|B = prim_i)$ is the learned probability for each primitive of the object given a specific task, and the normalisation factor is the sum of the probability of each object primitive being the graspable part.

In the following examples, all objects are limited to having three components, since the daily objects used are, in general, composed of few simple primitives. Fig. 20 shows statistics of the chosen object graspable part from human demonstrations. Three objects are used, a mug, a bottle and a wii-mote. The chosen primitive is the component with a high probability of being grasped. Using our framework, we can also detect which primitive was chosen for each type of task context and object. In Fig. 20, the top part of the bottle and *wii-mote* are represented as *prim*1, *prim*2 is the middle part and *prim*3 is the bottom part. For the mug, *prim*1 is the top part, *prim*2 is the bottom part and *prim*3 is the handle part. For the bottle and mug, all primitives are cylinders, and for the *wii-mote*, they are all boxes.

After learning a set of objects and the task context, when the object is observed again in the same context, the system is able to detect the graspable part as shown in Fig. 21.

In case of unknown objects, we have adopted a generalisation process, i.e., to try to reuse the prior knowledge for other contexts or using similarities. For example, if the unknown object has at

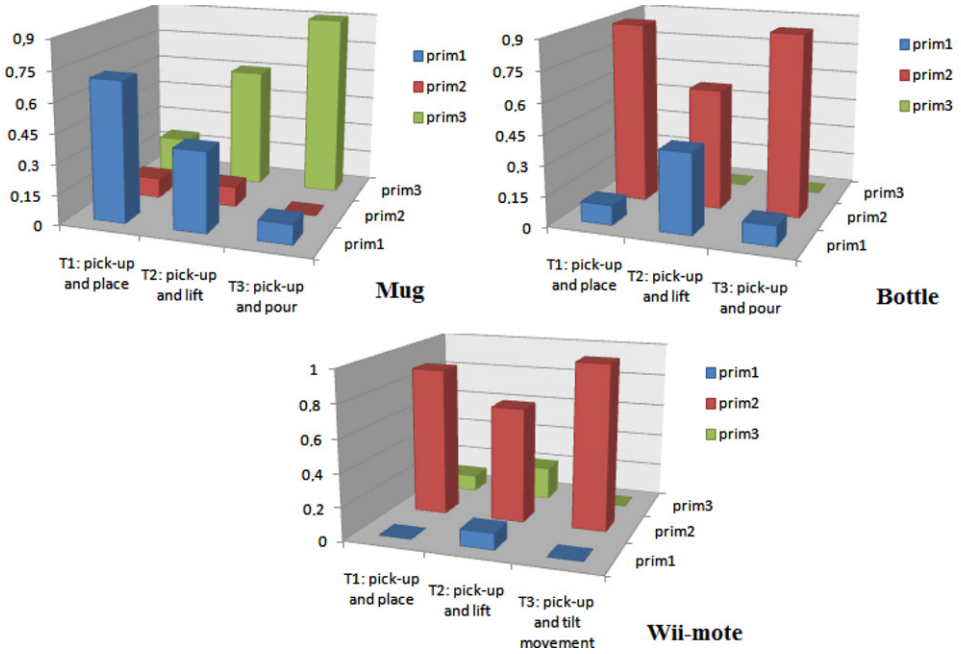

Fig. 20. Statistics for object graspable part after human demonstrations. Three different tasks performed many times by five different individuals. By analysing the probability distribution of the chosen primitives to perform the grasp, we can estimate the object graspable part given the task context.

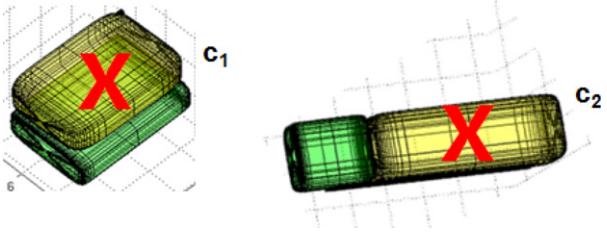

Fig. 21. Identified object graspable parts for the sponge and the *wii-mote*.

least one primitive in common with a known object, a similar grasp can be attempted. This will be addressed in a future work.

6. Conclusions and future work

Using multimodal data to learn from human demonstrations of manipulation tasks, we can learn and derive suitable models of manipulation tasks and of the manipulated objects, and later apply them for autonomous grasping by robotic systems. The outputs of this work can be used in different robotic applications as demonstrated in [31], by integrating the generalised representation of the manipulation movements to reach and grasp the object and the regions of the objects suitable to provide successful grasps. The constraints introduced by the models presented can be integrated by those applications during the estimation and synthesis of movements in new scenarios. We presented our proposed feature extraction of the multimodal data collected from human demonstrations of manipulation tasks. Based on these features, segmentation of the action phases and trajectory classification was accomplished. From the motion patterns, a generalised probabilistic representation for each type of task was derived. Results show the successful break down of action phases along a trajectory, as well as the suitability of the selected features as descriptors for the probabilistic approach used in task identification. Using the contact regions and tactile force intensities, a classification of grasp transitions was implemented, based on a set of grasp primitives. The implemented probabilistic approach for grasp primitive identification was able to correctly classify the grasp sequences in different tasks. An object probabilistic volumetric map was proposed to overlay the partially observed volume of the object with data about human visual gaze when initiating a grasp task, hand-object contact points and tactile forces. Results of this representation were presented that suggest its suitability for grasp planning since a unified model has the relevant observed information on how to grasp the object. The segmentation of the object into components will facilitate future matching for an artificial system observing objects and searching for data on how to perform successful grasping taking into account the task context.

Acknowledgements

The research leading to these results has been partially supported by the HANDLE project, which has received funding from the European Community's 7th Framework Programme under grant agreement ICT 231640; by the Portuguese Foundation for Science and Technology (FCT), with scholarships for Ricardo Martins (SFRH/BD/65990/2009) and Diego Faria (SFRH/BD/30655/2006); and by the Institute of Systems and Robotics at Coimbra University.

References

[1] S. Calinon, F. Guenter, A. Billard, On learning, representing, and generalizing a task in a humanoid robot, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37 (2) (2007) 286–298.
[2] K. Ogawara, Y. Tanabe, R. Kurazume, T. Hasegawa, Detecting repeated motion patterns via dynamic programming using motion density, in: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, 2009, pp. 1743–1749.
[3] P. Pastor, H. Hoffmann, T. Asfour, S. Schaal, Learning and generalization of motor skills by learning from demonstration, in: Proceedings of the 2009 IEEE International Conference on Robotics and Automation, ICRA'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 1293–1298.
[4] M. Kondo, J. Ueda, T. Ogasawara, Recognition of in-hand manipulation using contact state transition for multifingered robot hand control, Robotics and Autonomous Systems 56 (2008) 66–81.
[5] V. Kruger, D. Herzog, S. Baby, A. Ude, D. Kragic, Learning actions from observations, IEEE Robotics and Automation Magazine 17 (2) (2010) 30–43.

[6] R.L. Klatzky, S. Lederman, Intelligent Exploration by The Human Hand, Springer-Verlag, NY, USA, 1990, pp. 66–81.
[7] I. Biederman, Recognition by components: a theory of human image understanding, 94 (2) 1987, pp. 115–147.
[8] A. Sahbani, S. El-Khoury, A hybrid approach for grasping 3D objects, in: Intelligent Robots and Systems, 2009, IROS 2009. IEEE/RSJ International Conference on, 2009, pp. 1272 –1277.
[9] J.R. Flanagan, M.C. Bowman, R.S. Johansson, Control strategies in object manipulation tasks, Current Opinion in Neurobiology 16 (6) (2006) 650–659. motor systems/neurobiology of behaviour.
[10] J. Bohg, D. Kragic, Learning grasping points with shape context, Robotics and Autonomous Systems 58 (4) 2010, pp. 362–377.
[11] S. El-Khoury, A. Sahbani, A new strategy combining empirical and analytical approaches for grasping unknown 3D objects, Robotics and Autonomous Systems 58 (5) (2010) 497–507.
[12] S.B. Kang, K. Ikeuchi, Toward automatic robot instruction from perception-temporal segmentation of tasks from human hand motion, IEEE Transactions on Robotics and Automation 11 (5) (1995) 670–681. doi:10.1109/70.466599.
[13] Handle Project - Data Collection Database, 2011, URL: http://paloma.isr.uc.pt/DataCollectionDB/handle.
[14] D.R. Faria, J. Dias, 3D hand trajectory segmentation by curvatures and hand orientation for classification through a probabilistic approach, in: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'09, St. Louis, MO, USA, 2009.
[15] N. Kamakura, M. Matsuo, H. Ishii, F. Mitsuboshi, Y. Miura, Patterns of static prehension in normal hands, The American Journal of Occupational Therapy: Official publication of the American Occupational Therapy Association 34 (7) (1980) 437–445. http://view.ncbi.nlm.nih.gov/pubmed/6446851.
[16] R.S. Johansson, J.R. Flanagan, Coding and use of tactile signals from the fingertips in object manipulation tasks, Nature Reviews Neuroscience 10 (5) (2009) 345–359. http://www.ncbi.nlm.nih.gov/pubmed/19352402.
[17] P. Bessire, C. Laugier, R. Siegwart, Probabilistic Reasoning and Decision Making in Sensory-Motor Systems, 1st ed. Springer Publishing Company, Incorporated, 2008.
[18] D.R. Faria, H. Aliakbarpour, J. Dias, Grasping movements recognition in 3D space using a bayesian approach, in: Proceedings of ICAR'2009 — The 14th International Conference on Advanced Robotics, Munich, Germany, June 22–26, 2009.
[19] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 1 (1978) 43–49.
[20] D.R. Faria, R. Martins, J. Lobo, J. Dias, Probabilistic representation of 3D object shape by in-hand exploration, in: Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'10, Taipei, Taiwan, 2010.
[21] H.P. Moravec, Sensor fusion in certainty grids for mobile robots, AI Magazine 9 (2) (1988) 61–74.
[22] A. Elfes, Using occupancy grids for mobile robot perception and navigation, IEEE Computer 22 (1989) 46–57.
[23] S. Thrun, Robotic mapping: a survey, in: Exploring Artificial Intelligence in the New Millennium, Morgan Kaufmann, San Mateo, CA, 2002.
[24] R. Rocha, J. Dias, A. Carvalho, Exploring information theory for vision-based volumetric mapping, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, pp. 1023–1028.
[25] F. Colas, J. Diard, P. Bessière, Common bayesian models for common cognitive issues, Acta Biotheoretica 58 (1–2) (2010) 191–216.
[26] J. Lobo, J. Dias, Relative pose calibration between visual and inertial sensors, International Journal of Robotics Research 26 (6) (2007) 561–575.
[27] G.E. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (2) (1978) 461–464.
[28] G. Rissanen, Modeling the shortest data description, Automatica 14 (1978) 465–471.
[29] A.H. Barr, Superquadrics and angle preserving transformations, IEEE Computer Graphics and Applications 1 (1) (1981) 11–23.
[30] L. Chevalier, F. Jaillet, A. Baskurt, Segmentation and superquadric modeling of 3D objects, WSCG 11 (2003) 232–239.
[31] A. Billard, S. Calinon, R. Dillmann, S. Schaal, Robot programming by demonstration, in: Springer Handbook of Robotics, 2008, pp. 1371–1394.

Diego R. Faria is a Ph.D. student and researcher at the Institute of Systems and Robotics — Department of Electrical Engineering and Computers — University of Coimbra, Portugal. He is under the supervision of Prof. Jorge Dias (advisor) and Prof. Jorge Lobo (co-advisor). He is sponsored by a Ph.D. scholarship from the Portuguese Foundation for Technology and Sciences. He has graduated in Information Systems Technology in 2000 and has finished a Computer Science Specialisation Course in 2002 at the State University of Londrina, Brazil. He holds an M.Sc. degree in Computer Science from the Federal University of Parana, Brazil, since 2005. Currently, Diego Faria is collaborating as researcher on the European Project HANDLE within the 7° framework FP7. His research interest is Robotic Grasping, Multimodal Perception, Imitation Learning, Computer Vision and Pattern Recognition.

Ricardo Martins was born on the 15th of October 1984 in Proença-a-Nova, Portugal. Ricardo has an M.Sc. degree in Biomedical Engineering from the University of Coimbra, Portugal obtained in 2008. Currently, he is a Ph.D. student and researcher at the Institute of Systems and Robotics, Department of Electrical Engineering and Computers, University of Coimbra, Portugal. He is sponsored by a Ph.D. scholarship from the Portuguese Foundation for Technology and Sciences. He is collaborating as researcher on the European Project HANDLE within the 7° framework FP7. His research interests are Robotic Grasping/Haptics, Multimodal Perception and Imitation Learning.

Jorge Lobo (Jorge Nuno de Almeida e Sousa Almada Lobo) was born on the 23rd of September 1971, in Cambridge, UK. In 1995, he completed his five year course in Electrical Engineering at Coimbra University. In April 2002, he received the M.Sc degree, and in June 2007 he received the Ph.D degree from the University of Coimbra. He was a junior teacher in the Computer Science Department of the Coimbra Polytechnic School, and later joined the Electrical and Computer Engineering Department of the Faculty of Science and Technology at the University of Coimbra, where he currently works as Assistant Professor. He is responsible for courses on Digital Design, Microprocessors and Computer Architecture. His current research is carried out at the Institute of Systems and Robotics, University of Coimbra, working in the field of computer vision, sensor fusion, and mobile robotics. Current research interests focus on inertial sensor data integration in computer vision systems, Bayesian models for multimodal perception of 3D structure and motion, and real-time performance using GPUs and reconfigurable hardware. He has participated in several national and European projects, most recently in BACS, Bayesian Approach to Cognitive Systems, and HANDLE.

Jorge Dias born on March 7, 1960, in Coimbra, Portugal and has a Ph.D. degree on Electrical Engineering at University of Coimbra, specialisation in Control and Instrumentation, November 1994. Jorge Dias conducts his research activities at the Institute of Systems and Robotics (ISR-Instituto de Sistemas e Robótica) at University of Coimbra. Jorge Dias' research area is Computer Vision and Robotics, with activities and contributions on the field since 1984. He has several publications on Scientific Reports, Conferences, Journals and Book Chapters. Jorge Dias teaches several engineering courses at the Electrical Engineering and Computer Science Department, Faculty of Science and Technology, University of Coimbra. He is responsible for courses on Computer Vision, Robotics, Industrial Automation, Microprocessors and Digital Systems. He is also responsible for the supervision of Master and Ph.D. students on the field of Computer Vision and Robotics.