Rigid tool affordance matching points of regard

Marta Sanzari and Fabrizio Natola and Federico Nardi and Valsamis Ntouskos and Mahmoud Qodseya and Fiora Pirri
ALCOR, Vision, Perception and Cognitive Robotics Laboratory
DIAG, Sapienza University of Rome, Italy
{sanzari, natola, ntouskos, qodseya, nardi, pirri}@diag.uniroma1.it

***Abstract*— In this abstract we briefly introduce the analysis of simple rigid object affordance by experimentally establishing the relation between the point of regard of subjects before grasping an object and the finger tip points of contact once the object is grasped. The analysis show that there is a strong relation between these data, in so justifying the hypothesis that people figures out how objects are afforded according to their functionality.**

I. Introduction

We consider the following problem: how the way an object is observed influences a subject understanding of what the objects afford. For example several authors (see [1], [2]) have highlighted that an object is grasped according to its functionality, and that a subject can figure out what an object affords in terms of its functional properties: *s/he may pick up the object in a way that reflects his or her understanding of its purpose as well as its physical composition* [1].

Several studies have faced in the last thirty years the relation between perception and affordance (see [3], [4], [5], just to mention few of them), though, as far as we know, few authors (e.g. [6], [7]) have explored the relation with attention, but none has considered the relation with the point of regard.

In our studies on egocentric grasping with subjects wearing the Gaze Machine (GM) [8] we have noticed a significant relation between the point of regard (POR) on the object to be grasped and the finger tips position on the object, once grasped. This relation is quite relevant because of two significant outcomes: 1) point of regard data can be collected in a simple way, and then a simple transformation can be applied to recover the grasping points; 2) the relation between perception and object functionalities can be established. These two items are, crucially, part of the experimental paradigm studying the point of regard in human performing tasks.

In this abstract we explain the method we used to perform this analysis, which can be easily reproduced also with an eye tracker, since we have not used the main facility of the gaze machine, which is to reconstruct the 3D scene, but just used the video and the point of regard projected in the video (see the accompanying video).

The main contribution of our work is to reproduce both the point of regard and the grasping points in a 3D model of the object that is grasped, for the purpose of studying the object affordance. In the next section we briefly describe the method and in Section 3 we explain the experiments we have made to support the data collection and the analysis.

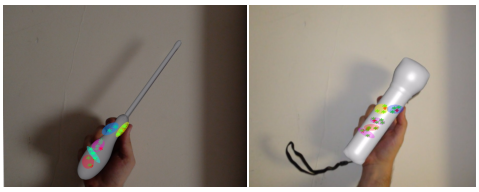

Fig. 1: Final 3D models of a screwdriver and a torch with collected PORs and finger tips clustered together.

II. Method, Modeling and Transformations

In this section we first briefly summarize the main steps of the method, and then we briefly illustrate the modeling and transformation parts of the method.

We consider a video of egocentric grasping, of a subject wearing the gaze machine (or alternatively a common eye tracker). We track the observed tool, so far making no assumption about clutter in the scene, which is occupied only by the observed tool. From the segmented mask of the tool, taken in the initial frame, we obtain the 3D model, say m. The 3D model m is fixed and referenced with respect to a global 3D frame.

Using the silhouette of the tracked object in the video, the model is transformed by a transformation $T_{m,I}$ so as to fit in the contour of the object in each frame I. Then, using the line of sight along which the model is aligned with the silhouette, the PORs, within the object contour, are back-projected on the 3D model m. In this way we obtain a full representation of the PORs collected in a time lapse ΔT on the 3D model. As discussed in the experimental section, the time lapse lasts at most 6 sec.

The sequence showing the grasping steps is treated in a similar way. While the subject is grabbing the object the hand is tracked and, finally, the finger tips position on the object are determined. In this case, we need to care only of the final image of this sequence showing the subject holding the object, say I_{hold}. The model m is transformed into the visible contour of the object. The finger tips are extracted and, again using the inverse transformation along the line of sight to fix the depth, the finger tips are back-projected on the model m. The two obtained sets are finally compared, the ratio between the number of PORs close to the finger

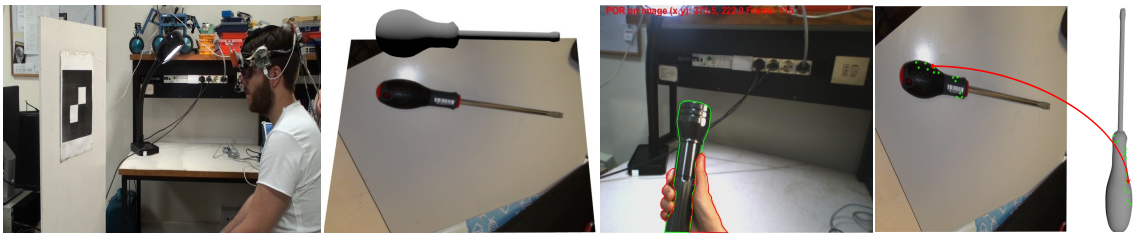

Fig. 2: **First**: GM calibration phase. **Second**: 3D model obtained from the tool silhouette. **Third**: segmented tool and hand. **Forth**: collected PORs back-projected on the model.

tips and the total number of points serve to evaluate the affordance matching hypothesis, the perceptual basis of the object functionality assessment.

As it can be appreciated in the accompanying video, the subject wearing the eye tracker moves her/his head and leans over the object, therefore to collect the point of regard it is necessary to track the object. Similarly, in order to obtain the final pose of the hand over the tool it is required to track the hand. For these two tasks we have been using the Chan-Vese method applied to each frame of the sequence, and exploiting the initialization at the initial frame together with simple tracking [9], [10].

On the other hand, we plan to explore tracking of both the hand and the tool. Note that the problem cannot be simply reduced to multiple tracking since the trajectory of the object - due to the motion of the subject wearing the GM - and the hand are different up to when the grasp is obtained (see for example [11]).

In the following we describe how the model is obtained from the segmentation mask of the initial frame, and how the bijective transformation is obtained. Let $I_0, I_1, \ldots, I_n$ be the frames of the video sequence, with I_n the frame in which the hand enters the scene, and $S_0, S_1, \ldots, S_n$ the silhouettes obtained via the Chan-Vese method. Let (I_0, S_0) be the initial pair. Then the model m is obtained by applying bending and stretching forces to the surface defined by the segmented mask, and solving the resulting energy functional by the finite element method [12]. A synthesis of the method is shown in the following algorithm:

Algorithm 1: Silhouette modeling

Input: (I_0, S_0), Parameters $\mathbf{q}_1$, $\mathbf{q}_2$, load $\mathbf{L}$
Output: model m
Generate a triangulation for S_0;
Choose the set of shape functions (at least quadratic) and the quadrature nodes;
Interpolate the shape functions at the quadrature nodes;
Assemble the stiffness matrix $\mathbf{K}$ and the load $\mathbf{L}$ using the quadrature rule;
Find the weights $\mathbf{X}$ of the shape functions solving the equation $\mathbf{KX} = \mathbf{L}$;
Compute mesh for m based on the triangulation.

As gathered above the model needs to be fitted in the silhouette of each frame (note that because of ego motion of the subject the silhouette changes continuously). The transformation is illustrated in the following algorithm:

Algorithm 2: Transformation between model and silhouette

Input: Model m, image silhouette pairs (I_i, S_i) with object silhouette S_i, $i = 0, \ldots, n$
Output: Bijective transformation T_{m,I_i} between reference m and image I_i
for $i = 1 : n$ **do**
  Detect a set of feature points F_0 in the silhouette S_0, (by keypoints, SURF features or similar) ;
  Detect a set of feature points F_i in the segment S_i,
  Project F_0 on m to obtain the 3D feature points X_0;
  Find feature matches $F_0 \leftrightarrow F_i$;
  Estimate 3D transformation $T^{(0)}_{m,I_i}$ based on $X_0 \leftrightarrow X_{I_i}$ up to an affine transformation;
  Apply T_{m,I_i} on m;
  Back-project each POR in S_i, along line of sight, to m;

Finally, in the last frame, using the finger tip segmentation and the transformation, the finger tips are back-projected on the model surface together with the collected PORs, see the last image of Figure 2.

III. Experiments

Experiments were performed with 10 volunteers. Each subject was asked to wear and calibrate the GM, and then to turn the head toward the table where a tool was placed, without notice, and with the subject ignoring the tools set. Then s/he is asked to pick the object up *very carefully*, and look at it while holding it. The experiment lasts about 10 sec. Scene data are acquired at 30 fps. Fixation of the tool cannot last more than 6 sec. while no time restriction is posed on the grasping and looking at the held object, since no more PORs are collected in this phase.

The results, illustrated in Figure 2, show the volunteers points of regard (PORs) and the contact points finger tips-tool. PORs are finally clustered with fingertip, see Figure 1. Accordingly, ratios between PORs close to finger tips and total number of PORs, while fixating the tools, are evaluated. These values are among 0.70 and 0.92: more deep analysis on affordance matching object functionalities based on PORs can start from here.

REFERENCES

[1] D. A. Rosenbaum, K. M. Chapman, M. Weigelt, D. J. Weiss, and R. van der Wel, "Cognition, action, and object manipulation." *Psychological bulletin*, vol. 138, no. 5, p. 924, 2012.

[2] P. Haggard, "Planning of action sequences," *Acta Psychologica*, vol. 99, no. 2, pp. 201–215, 1998.

[3] K. E. Adolph, M. A. Eppler, and E. J. Gibson, "Development of perception of affordances." *Advances in infancy research*, 1993.

[4] M. A. Eppler, "Development of manipulatory skills and the deployment of attention," *Infant Behavior and Development*, vol. 18, no. 4, pp. 391 – 405, 1995.

[5] E. J. Gibson and A. S. Walker, "Development of knowledge of visual-tactual affordances of substance," *Child Development*, vol. 55, no. 2, pp. pp. 453–460, 1984.

[6] L. Sun, U. Klank, and M. Beetz, "Eyewatchme;3d hand and object tracking for inside out activity analysis," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, June 2009, pp. 9–16.

[7] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The karlsruhe humanoid head," in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*. IEEE, 2008, pp. 447–453.

[8] F. Pirri, M. Pizzoli, and A. Rudi, "A general method for the point of regard estimation in 3d space," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 921–928.

[9] E. S. Brown, T. F. Chan, and X. Bresson, "Completely convex formulation of the chan-vese image segmentation model," *International journal of computer vision*, vol. 98, no. 1, pp. 103–121, 2012.

[10] T. F. Chan, L. Vese, *et al.*, "Active contours without edges," *Image processing, IEEE transactions on*, vol. 10, no. 2, pp. 266–277, 2001.

[11] T. Yang, Q. Pan, J. Li, and S. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 970–975 vol. 1.

[12] G. Celniker and D. Gossard, "Deformable curve and surface finite-elements for free-form shape design," in *SIGGRAPH*, 1991, pp. 257–266.