

Effect of motivation on gaze behavior over time

1st Tanakan Pramot
Department of Computer Eng.
Faculty of Engineering
Kasetsart University
Bangkok, Thailand
tanakan.p@ku.th

2ndZeynep Yücel, 3rdAkito Monden
Department of Computer Science
Okayama University
Okayama, Japan
zeynep, monden@okayama-u.ac.jp

4th Pattara Leelaprute
Department of Computer Eng.
Faculty of Engineering
Kasetsart University
Bangkok, Thailand
pattara.l@ku.ac.th

Abstract—The goal of this study is to understand the effect of presence and lack of grasping motivation on the gaze behavior over a time window. To that end, we derive a heat map from the gaze data for each 1 sec time interval and consider its centroid to coincide with the centroid of the grasping polygon. By estimating the tendencies in assigning grasping regions, grasping polygons of unknown objects are estimated. The similarity of estimated polygons is computed using intersection rate and Jaccard index and it is found that (i) gaze behavior of motivated subjects present smaller correlation with grasping regions than unmotivated ones, and (ii) correlation increases as time elapses for both cases.

Index Terms—Grasping, gaze, attention

I. INTRODUCTION AND RELATED WORK

Using gaze data to discover grasping affordances is rather an unexplored field. Only recently, several studies are carried out to investigate the effect of familiarity of objects and the intention of the subjects (lifting or using) [4], readiness to act [5], as well as the effect of center bias on gazing and its evolution over time [1] and distribution of gaze over functional and manipulative parts of tool objects [2], [3]. The findings indicate that manipulative ends are gazed more often and this effect becomes stronger, provided that the subject is familiar with the object, is ready to act, and has the intention of “using” rather than “lifting”. Under the light of these findings, it seems that estimating grasping regions of objects from gaze information of subjects with intentions of using or lifting the objects, is not possible. In this study, we contrast such motivations with lack of motivation (i.e. free viewing) and analyze temporal variations in gazing.

II. DATASET AND EXPERIMENT PROCEDURE

We use the freely available “Learning to grasp” dataset of Cornell University, which contains 1034 images of a variety of graspable objects from various orientations, together with their annotations of grasping regions as quadrilaterals [6]. A random subset of 432 images are displayed as a slide show viewed by four subjects, where two subjects are instructed to image grasping the objects and the other two carry out a free viewing task, henceforth referred as motivated and unmotivated, respectively. As the subjects view the clips resting on a chin rest, their gaze information is collected by an infrared sensor oscillating at 70 Hz.

This research was supported by JSPS KAKENHI Grant Number 18K18168.

III. ESTIMATION OF GRASPING REGIONS

To reveal the relationships between gaze and grasping regions, we build a heat map for the set of gaze points within each 1 sec time interval. The centroids of these heat maps are considered as potential centroids of grasping regions.

For creating a grasping polygon at these locations, we need to specify their morphological properties, i.e. size, inertia ratio and orientation. These values are determined by identifying the tendencies in choice of grasping polygons in the ground truth data. Namely, several properties of the object are proposed to be constitutive elements in specifying these values and empirical relations, which are found to indicate a significant correlation, are modeled in a parametric way. These models are in turn used to specify the morphological properties of the grasping polygons of unknown objects. The agreement between the ground truth and generated polygons is quantified with intersection rate and Jaccard similarity index.

A. Determining centroids of grasping regions

Each image is displayed for 3 seconds, and a total of roughly 165 gaze points are gathered during the display period. Grouping gaze points for every 1 sec time interval, we obtain 3 subsets. A heat map is built for each subset y kernel density estimation. In order to get a continuous map out of these discrete coordinates, we apply a Gaussian kernel k_g with a spherical covariance matrix around each gazed point. We compute k_g with support of 101×101 px and a proper bandwidth in relation to the screen resolution, the distance of the subject to the screen and human field of view. The centroid of the map is considered as the potential centroid of the grasping region.

B. Determining morphological properties of grasping regions

We consider the object size S_o , orientation θ_o and inertia ratio R_o as potential determining factors on morphological properties of grasping regions. In order to derive S_o , θ_o and R_o , we obtain the binary foreground, F_B and compute S_o and C_o as the 0th and 1st moments of F_B .

Next, an elliptic model is built to derive θ_o and R_o by solving for the eigen values $\lambda_{p,s}$ and eigen vectors $v_{p,s}$ of F_B such that $\lambda_p > \lambda_s$, $F_B v_{p,s} = \lambda_{p,s} v_{p,s}$. The principal and secondary axes $\vec{r}_{p,s}$ of the elliptic model are aligned with the unit vectors $\hat{v}_{p,s}$. Thus, $\theta_o = \arctan 2(\hat{v}_p, \hat{v}_s)$.

For obtaining R_o , we rotate the image around C_o by $-\theta_o$ and compute its horizontal and vertical projections, which can be used to approximate $\|\vec{r}_{p,s}\|$ such that $R_o \in (0, 1]$ is computed as $\|\vec{r}_s\|/\|\vec{r}_p\|$. Thereby, three descriptors of the object $\{S_o, \theta_o, r_o\}$ are derived. Regarding grasping regions, similar descriptors as S_g , θ_g and R_g , are computed using directly the annotated vertices.

C. Empirical observations and models

To understand grasping intuitions, several descriptor pairs are proposed to have a potential correlation and their relative distributions are examined. If the empirical data is found to be correlated, a parametric model is built to reflect that relation.

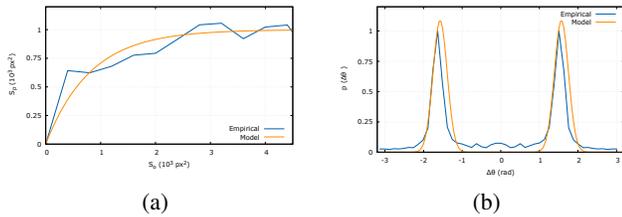


Fig. 1. Empirical observations relating morphological properties of polygons.

Firstly, we examine the relation between sizes, S_o and S_p , presented in Figure 1-(a). It is not surprising that grasping polygons of small objects are small and the size grows in a somewhat negative exponential relation with respect to the size of the object, settling around a stable value for very large objects. The model given in Eqn. 1 is proposed to reflect this relationship,

$$f(S_p|C_s, \lambda) = C_s(1 - \exp(-\lambda S_o)), \quad (1)$$

where C_s is the scaling factor and λ determines the rate of decay. After calibrating these parameters on the cumulative data, we obtain the approximation in Figure 1-(a).

When we examine the relation between inertia ratios, R_o and R_p , we see that R_p is roughly around 0.6, independent of R_o . Thus, R_p is assumed to come from the standard normal distribution, $R_p \sim \mathcal{N}(\mu_{rp}, \sigma_{rp})$.

Finally, we examine the relation between orientations, θ_o and θ_p and observe that the relative distribution suggests an offset. Thus, it is decided to study the relative orientation, $\Delta\theta_p = \theta_o - \theta_p$. The empirical distribution of $\Delta\theta_p$ is found not to be correlated with S_o or R_o . But the peaks around $\pm\pi/2$ in Figure 1-(b) suggest that the polygon is often positioned perpendicularly to the principal axis of the object. In order to model $\Delta\theta_p$, we propose using a von Mises distribution, which is the circular equivalent of the standard normal distribution,

$$f(\Delta\theta_p|\mu_{\Delta\theta_p}, \kappa_{\Delta\theta_p}) = \frac{\exp(\kappa_{\Delta\theta_p}(\Delta\theta_p - \mu_{\Delta\theta_p}))}{2\pi I_0(\kappa_{\Delta\theta_p})}, \quad (2)$$

where $\mu_{\Delta\theta_p}$ and $\sigma_{\Delta\theta_p}$ are analogous to the mean and standard deviations of normal distribution and I_0 is the modified Bessel function of order 0. The resulting model illustrated in Figure 1-(b) is considered to provide a satisfactory approximation.

IV. RESULTS AND CONCLUSION

We divide the set of 432 images into three subsets and calibrate the models defined in Section III-C using two of the subsets, where the third subset is used for estimation. Denoting the estimated grasping region with E and the set of ground truth polygons with G , we quantify estimation performance using two metrics, namely intersection rate $I(E, G)$ and Jaccard index $J(E, G)$. Intersection rate quantifies how often the estimated polygons have a nonzero intersection with any of the ground truth polygons, whereas Jaccard index measures how good this match is,

$$J(E, G) = \max_i \left(\frac{E \cap G_i}{E \cup G_i} \right). \quad (3)$$

Since in the ground truth several polygons are defined for each image, we consider the similarity with the best matching polygon in Eq. 3. In addition, we evaluate accuracy at every 1 sec interval (t_0 , t_m , t_f) and over entire display period (t_T).

TABLE I
INTERSECTION RATE AND JACCARD INDICES (%).

		t_0	t_m	t_f	t_T
$I(E, G)$	Motivated	63.19	85.06	84.49	89.69
	Unmotivated	77.43	85.99	85.30	90.27
$J(E, G)$	Motivated	14.86	22.91	23.69	23.41
	Unmotivated	22.72	30.08	32.25	32.68

Table I suggests that gaze of motivated subjects have roughly the same rate of intersection as the unmotivated subjects ($I(E, G)$). However, the Jaccard indices are much lower, indicating a lower degree of similarity. We consider two possibilities to explain this phenomena. Firstly, it could be due to the activate exploration of the environment of the motivated subjects, which is necessary to plan the subsequent actions. Secondly, for the *tools* with a well-defined grip and manipulating end, motivated subjects may examine the manipulating end more often than the unmotivated subjects ignoring the functional end. In addition, early saccades are found present less similarity, which could be due to the center bias as pointed out by [1].

REFERENCES

- [1] L. Van Der Linden, S. Mathôt, and F. Vitu, "The role of object affordances and center of gravity in eye movements toward isolated daily-life objects," *Journal of Vision*, vol. 15, no. 5, pp. 8–8, 2015.
- [2] N. Natraj, Y. Pella, A. Borghi, and L. Wheaton, "The visual encoding of tool-object affordances," *Neuroscience*, vol. 310, pp. 512–527, 2015.
- [3] R. M. Skiba and J. C. Snow, "Attentional capture for tool images is driven by the head end of the tool, not the handle," *Attention, Perception, & Psychophysics*, vol. 78, no. 8, pp. 2500–2514, 2016.
- [4] A. Belardinelli, M. Barabas, M. Himmelbach, and M. V. Butz, "Anticipatory eye fixations reveal tool knowledge for tool interaction," *Experimental brain research*, vol. 234, no. 8, pp. 2415–2431, 2016.
- [5] E. Ambrosini and M. Costantini, "Body posture differentially impacts on visual attention towards tool, graspable, and non-graspable objects.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 43, no. 2, p. 360, 2017.
- [6] D. Fischinger, M. Vincze, and Y. Jiang, "Learning grasps for unknown objects in cluttered scenes," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 609–616, IEEE, 2013.