Visual saliency-based babbling of unknown dynamic environments
Leni Le Goff, Carlos Maestre, Stéphane Doncieux

To cite this version:

HAL Id: hal-02987424
https://hal.science/hal-02987424
Submitted on 3 Nov 2020

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

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Visual saliency-based babbling of unknown dynamic environments

Leni Legoff *,†, Carlos Maestre *,†, and Stephane Doncieux *,†
*UMR 7222, ISIR, Sorbonne Universités, UPMC Univ Paris 06, Paris, France
†UMR 7222, CNRS, ISIR, Paris, France
Email: {legoff, maestre, doncieux}@isir.upmc.fr

Our everyday environment contains many different objects and we are frequently confronted to new objects, may it be known objects with a new shape or color, or completely new objects (smartphones or tablet computers, for instance, did not exist at all in our environment a few years ago). A robot working in our environment should then be able to deal with such modifications. It should in particular be able to identify these objects and what to do with them, i.e. their affordances. Human infants learn these affordances through an interaction with the environment called body babbling [1]. Developmental robotics [2] encourages applying the same exploration step in robots. But, how to define an environment exploration strategy that would work on any kind of environment and object that the robot may encounter before knowing them and their features?

A widely used hypothesis consists in restricting the scenario composition, providing this a priori information to help segmenting the visual scene and then to help the babbling to focus on the objects thus identified. Typical hypotheses are that objects lay on a flat surface [3], or can be discriminated by an easy to detect color [4]. Although these approaches can perform properly in isolated controlled environments they would fail in open-ended scenarios, where it is not possible to envision all possible situations [5].

Other hypotheses are based on how humans attention is attracted by specific regions of the scene based on their visual saliency [6]. It is based on the variation of some properties in a visual scene (as color, intensity, shape, or orientation). Previous works ([7], [8], [9]) create a saliency map based on one or more properties of the scene. Interaction with the scene might yield new details to improve the saliency map [10]. However, this interaction should be guided without a priori information of the environment.

In this work we propose an autonomous babbling of unknown environments, named Visual Saliency Babbling, driven by the salient regions of raw images of the scene obtained from a fixed RGB-D camera. First, Visual Saliency Babbling identifies the salient regions of the environment, without any previous scene assumption, and then randomly interacts with one of them using an available inverse kinematics model. Once the region has been reached, the robot’s arm comes back to an initial position, and the modifications that eventually resulted from this interaction are recorded for a future object identification process thanks to a motion detector. These operations are repeated until no more salient regions are detected, or a maximum number of iterations is reached.

The salient regions of the environment are called Surfaces Of Interest (SOI). Visual Saliency Babbling generates the set of SOI from the point cloud generated by a RGB-D camera. Therefore, a SOI is a salient region in the point cloud. The SOI are defined on the basis of two visual features: SIFT keypoints¹, salient regions of a 2D image with an associated descriptor; and supervoxels, clusters in the point cloud, called voxels, that result from a segmentation of the environment respecting the limits of the objects (bottom left image of Figure 1). A supervoxel provides 3D information about a region, e.g.

Fig. 1. Steps of the generation of the set of SOI associated to the initial set-up used during the experiment. At the top left, raw image captured by the camera. At the top right, SIFT keypoints computed using the raw image. At the bottom left, supervoxels identified using the point cloud of the initial set-up. At the bottom right, blue regions represent the set of SOI computed using the SIFT keypoints and the supervoxels, projected over the previous point cloud.

¹http://www.vlfeat.org/api/sift.html
Cylinders, a car toy and a cube. The exploration is performed of a plain table and six objects to interact with: two balls, two supervoxels were computed, and finally 60 SOI were created. 1, 202 SIFT keypoints were identified around the objects, 233 the bottom right image of Figure 1. In the process of Figure SIFT keypoint to a supervoxel to extract SOI (blue regions in of SIFT keypoints into the point cloud; (5) matching of each projection of Figure 1); (3) supervoxels computation on the point cloud (top right image of Figure 1); (4) projection of SIFT keypoints into the point cloud; (5) matching of each SIFT keypoint to a supervoxel to extract SOI (blue regions in the bottom right image of Figure 1). In the process of Figure 1, 202 SIFT keypoints were identified around the objects, 233 supervoxels were computed, and finally 60 SOI were created.

EVALUATED SCENARIO AND RESULTS

The scenario to test Visual Saliency Babbling is composed of a plain table and six objects to interact with: two balls, two cylinders, a car toy and a cube. The exploration is performed by a Crustcrawler Pro-Series robotic arm\(^2\) (Figure 1). An Asus Xtion PRO LIVE camera\(^3\) provides the scene information.

Figure 2 depicts the explored SOI during a babbling of 25 iterations producing 11 contacts. The exploration is mainly focused on the regions where the objects are located, and it is adapted to the new position of the objects after each contact. An online video showing several executions of the experiment is available at http://youtu.be/YykKhgA6TY7E.

Several experiments with different number of iterations have been executed (Figure 3). The results obtained show a correlation between the number of iterations and the number of contacts produced. Therefore, Visual Saliency Babbling seems a suitable exploration mechanism for a robot to interact dynamic unknown environments.

ACKNOWLEDGEMENTS

This research is sponsored in part by the DREAM project\(^4\). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 640891. This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d’Avenir programme under reference ANR-11-IDEX-0004-02.

REFERENCES


\(^2\)http://www.crustcrawler.com/products/ProRoboticArm/
\(^3\)https://www.asus.com/
\(^4\)http://www.robotsthatdream.eu/