Object Recognition and Pose Estimation by means of Cross-Correlation of Mixture of Projected Gaussian

Mauro Antonello[1], Sukhan Lee[2], Naguib Ahmed[2], and Emanuele Menegatti[1]

[1] University of Padova, department of Information Engineering,
via Gradenigo 6/B, 35131 Padova, Italy
[2] Intelligent Systems Research Institute, Sungkyunkwan University,
Suwon, Rep. of Korea

Abstract. In this work we present a novel approach for object recognition and pose estimation suited for a personal robot environment. The described technique addresses the peculiar requirements of this field, such as the system responsiveness and the capability of learning new object models online. Our method makes use of a novel variant of the Mixture of Gaussian in order to learn the spatial distribution of the object features. Such distribution is trained online while the robot moves in the environment. The resulting models are then compared through a fast cross-correlation based method, thus obtaining at the same time the optimal registration point.

Keywords: object recognition, pose estimation, cross correlation, mixture of projected Gaussian

1 Introduction

One of the primary capabilities required by personal robots is recognizing the surrounding environment with high responsiveness, often combined with object recognition and grasping tasks. Moreover, in a real-world scenario a robot should be able to deal with new objects when it is placed in a new environment. Although actual state of art object recognition and pose estimation methods can achieve high recognition performance [1][2][3], the computational load is often too heavy for real-time applications, especially if used in an online training context. In this work we propose a novel technique for object recognition and pose estimation, focused on the key needs of the personal robotic environment. We exploit the capability of a personal robot to explore and self-localize in the environment to improve both object and scene models online. As the robot moves in the environment, it collect a sparse set of features and keypoints to create a statistical model of their spatial distribution, continuously refined as soon as new frames are collected. Online statistical approaches are widely exploited in robotic-oriented algorithms, in particular particle filtering [4][5] or Kalman filters are key representatives. A variant of the proposed algorithm is presented as case study in section 3, showing a concrete application in a personal robot object recognition module.

2 Proposed Method

Assuming the robot is provided with an RGB-D sensor, the proposed algorithm detects the local features (i.e. SIFT) in the environment analyzing the RGB-D stream. In order to obtain full 6DOF keypoints we make use of 2D SIFT keypoints and we reproject them to the 3D surface in the point cloud. As analized by Munaro et. al [6], this approach gives better results in comparison to other 3D descriptors. Exploiting the robot self-localization the feature locations are referred to the world coordinates and the algorithm creates a model containing explicit spatial information of the salient points. More precisely, the information regarding the feature location and orientation is expressed as a probability distribution, namely Mixture of Projected Gaussian (MoPG). This particular variant of the Mixture of Gaussian is used to correctly handle 6DOF data points, which lie on the $SE(3)$ manifold [7]. Hence, object and scene models consist in a set of MoPG describing the probability to find a given feature at a given pose. In our implementation all MoPG are trained online following the procedure described by Engel et al. [8].

In order to both compare and register two models we apply the Cross-Correlation (CC) operator (see Fig. 2). A finite sum of Gaussian PDFs (i.e. MoPG components) is a function closed under convolution operator, this closure applies even to the CC since by definition the CC of two real continuous functions, $(f_1 \star f_2)(\mathbf{t})$ can be expressed in terms of their convolution:

$$(f_1 \star f_2)(\mathbf{t}) := f_1(-t) * f_2(t)\,. \tag{1}$$

The closure of the MoPG over the CC operation ensures that the result is still a MoPG. Moreover, the CC between Gaussian PDFs is computationally fast since given two Multivariate Gaussian distributions X_1, X_2 such that:

$$\begin{aligned} \mathbf{x}, \boldsymbol{\mu}_i \in \mathbb{R}^n, \quad \boldsymbol{\Sigma}_i \in \mathbb{R}^{n\mathrm{x}n}, \quad i \in \{1,2\} \\ X_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad X_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \\ f_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)}\,. \end{aligned}$$

Such convolution is given by:

$$\begin{aligned} (f_1 * f_2)(\mathbf{x}) &= \int^{\mathbb{R}^n} f_1(\boldsymbol{\tau})\, f_2(\mathbf{x}-\boldsymbol{\tau})\, d\boldsymbol{\tau} \\ &= \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}_c|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)' \boldsymbol{\Sigma}_c^{-1} (\mathbf{x}-\boldsymbol{\mu}_c)}\,. \end{aligned} \tag{2}$$

The convolution result is again a Multivariate Normal distributed PDF, with $\boldsymbol{\mu}_c = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$. From (1) and (2) we obtained the closed form for the CC between two Multivariate Gaussian and, consequently, for MoPG distributions. Once obtained the cross-correlation MoPG of two models we have a function to indicate how good is the registration between the two models at

a given pose, see Fig. 2. In order to find peaks in this MoPG we applied an efficient mode finding algorithm, which is similar to the algorithm for MoGs proposed in [9]. Comparing the scene model with an object model by means of the described technique does not require prior segmentation of the scene and can find multiple occurrences of the object at the same time, thus making our algorithm extremely efficient and well suited for real-time applications.

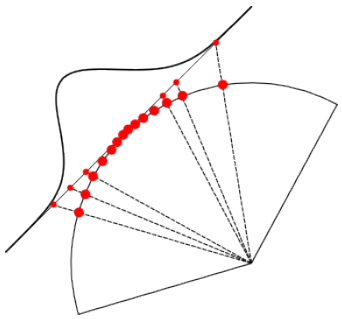

Fig. 1. Components of an MoPG lies in a space tangent to the manifold, thus avoiding issues implied by some statistical concepts (i.e. averaging) on the $SE(3)$ manifold.

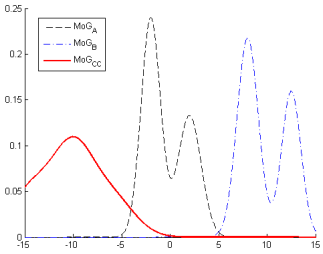

Fig. 2. The cross-correlation function $\mathrm{MoG_{CC}}$ is used as an index of the registration quality of $xtMoG_{\mathrm{A}}$ and $\mathrm{MoG_B}$ when $\mathrm{MoG_B}$ is translated by a given value.

3 Case Study

The algorithm presented in the previous section exploits the variation of the object appearance with respect to the pose to learn the model. This variation is caught as a MoPG representing spatial distribution of local feature in the object texture but the algorithm is generally applicable to a wider choice of features other than photometric local features (i.e. SIFT). For instance, the proposed MoPG can be applicable as well to geometrical features such as 3D shape descriptors. Similarly to what proposed in [10], in our case study these global features consist in the object height, width and aspect ratio. Differently from local features, in case of global features is the feature descriptor that may take different values with respect to the orientation as well as of the distance, due to the complexity of the 3D shape of the object. This is illustrated in the left side of Fig. 3, where the measured height of the milk box varies according to its orientations as the thin panel at the top may not be detectable at the particular orientations of milk box. This leads to a singularity in measurements for the given sensor and sensing algorithm. The proposed MoPG is expected to overcome this representation problem by providing a multi-modal likelihood distribution of the descriptor over the pose space.

Framework Setup The proposed case study is now being carried out as part of the collaboration between the Intelligent Autonomous Systems Laboratory (IAS-Lab) in Padua University and the Intelligent Systems Research Institute (ISRI) in Sungkyunkwan University. The result will be extensively analyzed on a 3rd generation of domestic service robot capable of errand services, HomeMate[3]. HomeMate is an outcome of the Korus-tech project sponsored by the Ministry of Science, ICT and Planning (MSIP) of Korea, for which ISRI is serving as the Principal Investigator with "YUJIN Robot Co." and "Bonavision Co." as domestic partners, as well as "Georgia Tech" and "Penn State University" as US partners. Three generations of "Homemates" had been developed and tested in Korean elderly centers, as well as in US elderly homes for market survey. 3rd Generation of HomeMate (right side of Fig. 3) is equipped with MS Kinect RGB-D and Bumblebee 2 Stereo cameras mounted on a Pan/Tilt Module, StarGazer Indoor Localization System, Laser Range Sensor, 5DOF manipulator, HP EliteBook 2760p Tablet PC, and a Linux-based integrated PC. HomeMate is a cognitive consumer robot developed for elderly care with such services as errand, video chatting, game playing, as well as medicine alarm. HomeMate is designed to operate in unstructured indoor environment. It can navigate in a cluttered dynamic environment, identify objects and users with a highly adaptive vision system, and interact with users fluidly. Main challenges HomaMate has to cope with are poor illumination conditions, high clutter, dynamic changes in environment, and low attention span of elderly users.

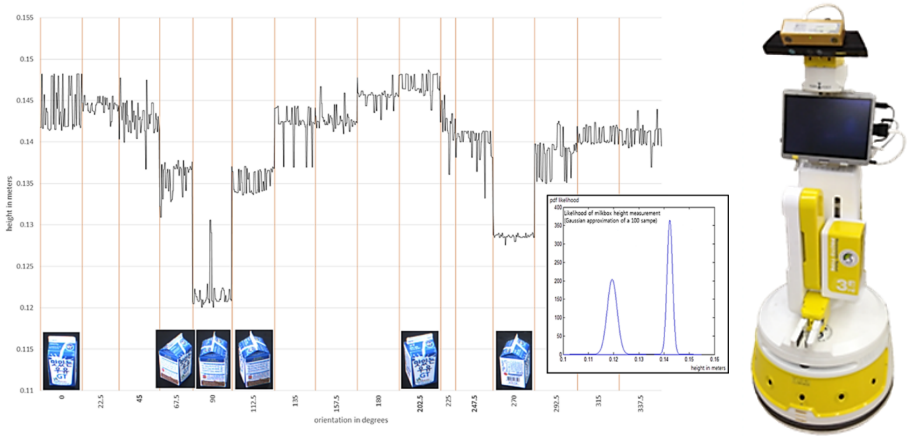

Fig. 3. Left: Example of singularity in height measurement of milk box in certain orientations causing feature likelihood to be multimodal. Right: 3rd generation of HomeMate cognitive consumer robot

[3] Korus-Tech Project: cognitive consumer robot.
HomeMate: http://isrc.skku.ac.kr/project/HomeMate.php

References

1. Tombari, F., Di Stefano, L.: Object Recognition in 3D Scenes with Occlusions and Clutter by Hough Voting. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology, IEEE (November 2010) 349–355
2. Xie, Z., Singh, A., Uang, J., Narayan, K., Abbeel, P.: Multimodal Blending for High-Accuracy Instance Recognition. IROS (2013)
3. Aldoma, A., Tombari, F., Prankl, J., Richtsfeld, A., Stefano, L.D., Vincze, M.: Multimodal Cue Integration through Hypotheses Verification for RGB-D Object Recognition and 6DOF Pose Estimation. ICRA (2013) 2096–2103
4. Browatzki, B., Tikhanoff, V., Metta, G., Bulthoff, H.H., Wallraven, C.: Active object recognition on a humanoid robot. In: 2012 IEEE International Conference on Robotics and Automation, IEEE (May 2012) 2021–2028
5. Lee, S., Lu, Z.: Dependable 3D object recognition with two-layered particle filter. In: Proceedings of the 5th International Confernece on Ubiquitous Information Management and Communication - ICUIMC '11, New York, New York, USA, ACM Press (February 2011) 1
6. Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A Feature-based Approach to People Re-Identification using Skeleton Keypoints. ICRA (2014)
7. Feiten, W., Lang, M.: MPG-a framework for reasoning on 6 dof pose uncertainty. Workshop on Manipulation Under Uncertainty (2011)
8. Engel, P., Heinen, M.: Incremental learning of multivariate Gaussian mixture models. Advances in Artificial IntelligenceSBIA 2010 (2011) 82–91
9. Carreira-Perpinan, M.: Mode-finding for mixtures of Gaussian distributions. Pattern Analysis and Machine Learning (2000) 1–23
10. Lee, S., Ilyas, M., Jaewoong, K., Naguib, A.: Evidence filtering in a sequence of images for recognition. In: 2012 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE (October 2012) 1–8