VA2Mass: Towards the Fluid Filling Mass Estimation via Integration of Vision & Audio Learning*

Qi Liu[1], Fan Feng[1], Chuanlin Lan[1], and Rosa H.M. Chan[1][0000−0003−4808−2490]

Department of Electrical Engineering,
City University of Hong Kong, Hong Kong, China
rosachan@cityu.edu.hk

Abstract. Robotic perception of filling mass estimation via multiple sensors and deep learning approaches is still an open problem due to the diverse pouring durations, small pixel ratio for target objects and complex pouring scenarios. In this paper, we propose a practical solution to tackle this challenging task via estimating filling level, filling type and container capacity simultaneously. The proposed method is inspired by how humans observe and understand the pouring process via the cooperation among multiple modalities, i.e., vision and audio. In a nutshell, our proposed method is divided into three folds to help the agent shape a rich understanding of the pouring procedure. First, the agent obtains the prior of container categories (i.e., cup, glass or box) through the object detection framework. Second, we integrate the audio features with the prior to make the agent learn a multi-modal feature space. Finally, the agent infers the distribution of both the container capacity and fluid properties. The experimental results show the effectiveness of the proposed method, which ranked as 2^{nd} runner-up in the CORSMAL Challenge of Multi-modal Fusion and Learning For Robotics in ICPR 2020.

Keywords: Multi-modal perception · Robotic learning · Deep learning.

1 Introduction

Deep learning, especially Deep Neural Networks (DNNs) have achieved state-of-the-art performances in various tasks such as speech recognition, visual object recognition, and image classification [16] [8] [29] [2]. There is increasing interest in using DNNs and learning techniques (i.e., meta learning, transfer learning, continual learning, reinforcement learning and their intersections) for robotic learning in different robotic tasks, including manipulation, SLAM, motion control, and image processing [17] [36] [27] [22] [32] [12]. However, the environments

* The work described in this paper was partially supported by grant from Guangdong-Hong Kong-Macau Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence Fund (No. 20019009).

and tasks are dynamic and complex for robotics. Thus how to construct a stable, efficient and robust robotic learning system is still an open challenge [3] [34] [33].

In this work, we focus on one specific challenging task in robotic learning: estimating the properties of containers with different fillings poured, which is crucial for human-robot cooperation. In fact, pick-and-place and pouring are the top 2 frequently executed motions in household chores [25] [26]. The elders or disabilities can perform better on such daily activities like objects pick-up, place and handovers with the assistance of the smart robots. In this paper, we focus on the estimation of the mass of the filling. The task is divided into 3 folds:

- **Task (1)**: filling level classification;
- **Task (2)**: filling type classification;
- **Task (3)**: container capacity estimation.

The only prior information known to the robot is the container categories and the containers vary in their physical properties, i.e., shape, material, texture, transparency, and deformability. Task (1) is to classify how full are the containers. The containers can be either empty or filled with an unknown content at 50% or 90% of the whole capacity of the container. There are three classes: 0 (empty), 50 (half full), and 90 (full). Task (2) is to classify which content is filled in the containers. The containers can be either empty or filled with an unknown content. There are four filling type classes: 0 (empty), 1 (pasta), 2 (rice), 3 (water). Task (3) is to estimate the container capacity. The containers vary in shape and size.

There are three main challenges for filling mass estimation with CORSMAL Containers Manipulation dataset. Firstly, the duration of each recording in the dataset is different from each other (Figure 1), which makes it extract inconsistent dimensions of the modality features. Secondly, the target objects (the container and the content) occupied small pixel ratio compared with the subject or desk in the recording (Figure 2), so it is impossible to track the objects by common classification models (e.g., VGG-16 pre-trained with ImageNet) [35]. Thirdly, the scenarios in the dataset vary in target occlusion and subject motion (e.g., the container is held by the subject or not) (Figure 5). These issue would compromise the estimation performance.

To this end, we propose a solution to reduce the above issues and improve the performance by leveraging the modality of RGB images and acoustics. Inspired by how humans judge the physical properties of the filling content and the containers with their vision and hearing, our solution is to obtain the container prior first, followed by classifying the filling level and filling type with audio vibration, and sampling the container capacity from the Gaussian process regression. Empirically, given the container type (e.g., cup, glass, food box), humans deal with filling level (i.e., empty, half full and full) prediction by hearing the changes of the vibrational frequency of the air in the container right after the pouring procedure, especially for the liquid or empty filling. Besides, under the same prerequisite, humans can distinguish the filling type (i.e., empty, pasta, rice and water) via the nature frequency of the filling content.

The main contributions of this work are summarized as follows:

1. We propose to tackle the estimation task for physical properties of the filling and its corresponding container (i.e., filling level, filling type and container capacity) in the perception of human pouring. We adopted YOLOv4 [5] to classify the container as a prior, followed by audio feature extraction and fitting of container capacity distribution.
2. We have done a lot of experiments on the CORSMAL Containers Manipulation dataset, which proves the effectiveness and competitiveness of our method.

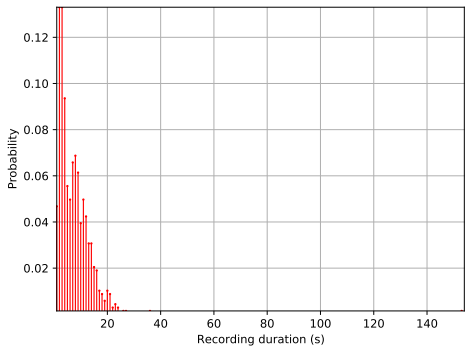

Fig. 1. Discrete probability mass function over the recording durations.

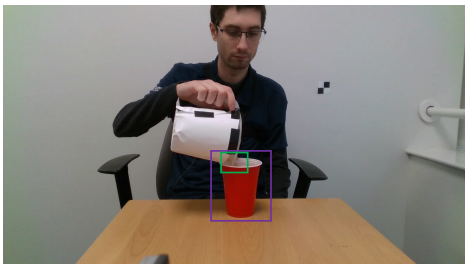

Fig. 2. The target object pixel ratios in the visual recording. The purple bounding box indicates the container, while the green bounding box indicates the filling content.

2 Related Work

In this section, we focus on reviewing pouring tasks via different manipulation learning approaches.

2.1 Motion trajectory based approaches

P. Pastor et al. learned dynamic motor primitives (DMPs) from human demonstrations for pouring tasks, and used these to generalize to different container (i.e., cup) placements [23]. Similarly, M. Muehlig et al. encoded demonstrated bimanual pouring trajectories using GMM [21]. In contract, S. Brandi et al. proposed learning in a feature space defined by the warped parameters, in order to automatically generalize between objects [6].

2.2 Modality sensing based approaches

Vision, as the one of commonly used modality in our daily lives, has been explored to robotic learning, including perception of the pouring process. K.J. Pithadiya et al. [28] looked at different edge detection algorithms for detecting whether or not water bottles are over or under filled. However, instead of determining the actual liquid height, the detected edges are compared to a reference line to determine this. For the restaurant industry, R. Bhattacharyya et al. [4] use RFID tags for liquid level detection in beverage glasses and liquor bottles. Besides, RGB-D cameras are also adopted, as the extension of vision sensing, to tackle the robotic pouring [10] [9]. The auditory information contributes to the pouring task when the filling interacts with the air column of the container [14] [7] [11] [18]. For example, S. Griffith bootstraped classification learning about how objects interact with water with auditory and proprioceptive feature [11]. Recently, inspired by the phenomenon that the resonance frequencies implicitly relate to the length of the air column of the container, H. Liang et al. [18] designed a perception-based deep neural network to estimate the liquid filling level of the target containers. In addition to the vision and acoustic sensing, touch modality is also widely used in robotic pouring tasks, especially force and torque sensing [30] [13] [31].

3 Method

The overall pipeline of our method is shown in Figure 3. All tasks is based on container prior, i.e., container category, by YOLOv4 [5]. After that, the process is divided into two parts. One part is for Task (1) and Task (2) by a multilayer perceptron (MLP) input with audio features, the other part is for Task (3) by using Gaussian process regression.

3.1 Container detection

To infer the container category, we first extract frames from recorded videos of four-view cameras (i.e., C_1, C_2, C_3 and C_4), then we conduct container detection via YOLOv4 pretrained on MS COCO dataset [19] for each frame, from which we obtain the bounding boxes and categories of the containers. Finally, we follow the majority rule, that votes for cup, glass and box among extracted frames, to

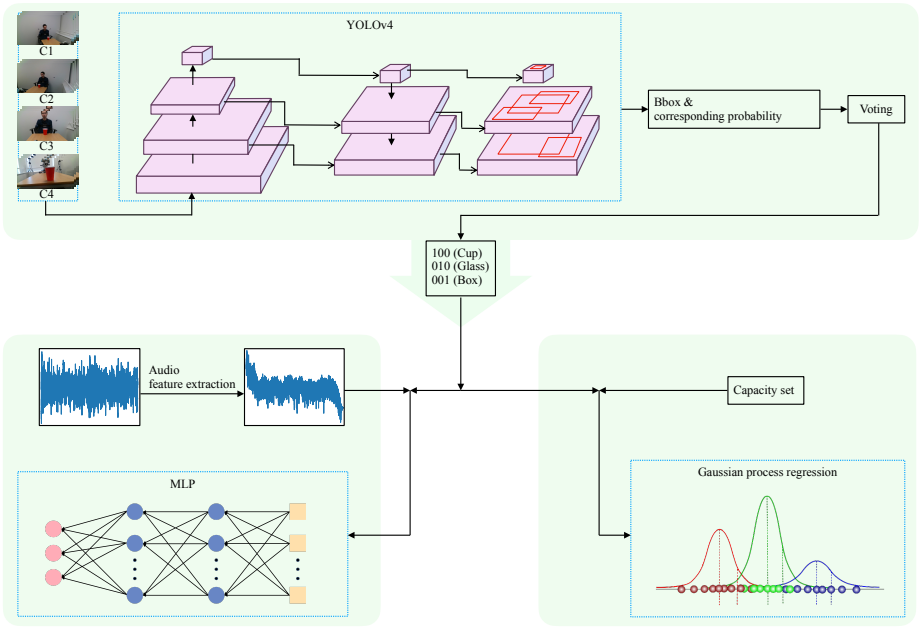

Fig. 3. The overall pipeline for filling mass estimation. C_i indicates the views from the i^{th} camera (Intel RealSense D435i), C_1 is the view from the left side of the manipulator, C_2 is the view from the right side of the manipulator, C_3 is the front view with the camera mounted on the manipulator, C_4 is the view from the moving camera worn by the demonstrator (human).

decide the container category for each video. As shown in Figure 2, the target container occupied small pixel ratio in the recording, YOLOv4 is capable to capture the small objects and predict reasonable category for the container.

3.2 Filling level and filling type classification (Task (1) and Task (2))

Audio feature extraction As shown in Figure 4, we use spectrogram as the audio feature by Discrete Fourier Transform (DFT):

$$F(u) = \sum_{x=0}^{N-1} f(x) e^{-j\frac{2\pi ux}{N}} \tag{1}$$

Where x and u are the input audio signal and sampling frequency respectively, with N denotes the length of input signal.

Specifically, we re-sample the raw audio with the frequency of $16,600Hz$, and select the last $32,000$ data points as input audio signal. The dimension of

output frequency feature is also 32,000, and we select the $u >= 0$ part, whose dimension is 16,001.

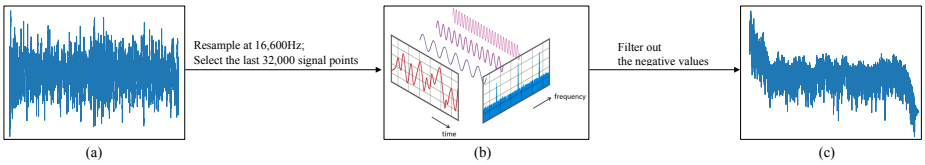

Fig. 4. The procedure of audio feature extraction. (a) Sampled audio sequence; (b-c) Spectrogram of the sampled audio using FFT.

Backbone models The utilized classification model is Multi-Layer Perceptron (MLP) with 2 hidden layers. Given the container category, we trained 6 MLP models for filling level and filling type classification. Specifically, we use 3 different MLPs for filling level classification, each of which belongs to each container category. Likewise 3 distinct MLPs are utilized for filling type classification. The loss function and optimizer are cross-entropy function and Adam [15]. Detailed experimental settings (i.e., learning rate, number of neurons, etc) are listed in Section 4.2.

3.3 Container capacity estimation (Task (3))

In this task, the objective is to estimate the container capacity from the data. We observe that for each category, the capacity distribution can be fit by Gaussian distribution well.

Thus we utilize the Gaussian process regression to learn the capacity distribution of each container. The pipeline can be divided into 3 steps:

- Step 1: Infer the category label x_i from the object detection model.
- Step 2: Construct the training set: $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$, where each sample is (x_i, y_i), where x_i and y_i indicate the label of object category and its corresponding capacity. N is the number of samples.
- Step 3: Conduct the Gaussian process regression. For a new input X^* in test-set, we have $\hat{\mathbf{y}}^* = K\left(X^*, X\right) K(X, X)^{-1}\mathbf{y}$, where $\mathbf{y}^*$ is the predicted value and K is the covariance function defined by $K(A, B)_{ij} = \exp\left(-\frac{1}{2}\left|A_i - B_j\right|^2\right)$.

4 Experiments

4.1 CORSMAL Containers Manipulation dataset

CORSMAL distributes an audio-visual-inertial dataset [1] of people interacting with containers, for example while pouring a filling into a glass or shaking

a food box. The CORSMAL Container Manipulation dataset is collected with four multi-sensor devices (one on a robotic arm, one on the human chest and two third-person views) and a circular microphone array [1]. Each device is equipped with an RGB camera, a stereo infrared camera and an inertial measurement unit. In addition to RGB and infrared images, each device provides synchronised depth images that are spatially aligned with the RGB images. All signals are synchronised, and the calibration information for all devices, as well as the inertial measurements of the body-worn device, is also provided. Besides, the dataset is collected under three different scenarios (Figure 5) with an increasing level of difficulty, caused by occlusions of the target objects or subject motion. In the first scenario, the subject sits in front of the robot, while a container is on a table. The subject pours the filling into the container, while trying not to touch the container (cup or glass), or shakes an already filled food box, and then initiates the handover of the container to the robot. In the second scenario, the subject sits in front of the robot, while holding a container. The subject pours the filling from a jar into a glass/cup or shakes an already filled food box, and then initiates the handover of the container to the robot. In the third scenario, a container is held by the subject while standing to the side of the robot, potentially visible from one third-person view camera only. The subject pours the filling from a jar into a glass/cup or shakes an already filled food box, takes a few steps to reach the front of the robot and then initiates the handover of the container to the robot. In our proposed solution, we used two modalities to tackle the filling mass estimation, which are RGB images from all views to classify the containers, and audio from the microphones to tackle Task (1) and Task (2).

4.2 Exeprimental details

Here we list all the experimental details. Table 1 illustrates the hyper-parameters of MLP models used in Task (1) and Task (2). All the MLPs in the experiments are trained from scrach. For YOLOv4, we utilize the pre-trained model from MS COCO dataset [19]. Table 2 shows the model and library acquisitions in the experiments. All the methods are implemented using PyTorch [24] toolbox with an Intel Core i9 CPU and 8 Nvidia RTX 2080 Ti GPUs.

Table 1. Architecture & hyper-parameters of MLP used in Task (1) & Task (2).

Architecture & Hyper-parameters	Chosen Values
# Hidden layers	2
# Neurons in the 1^{st} hidden layer	$3,096$
# Neurons in the 1^{st} hidden layer	512
# Epochs	200
learning rate	0.05

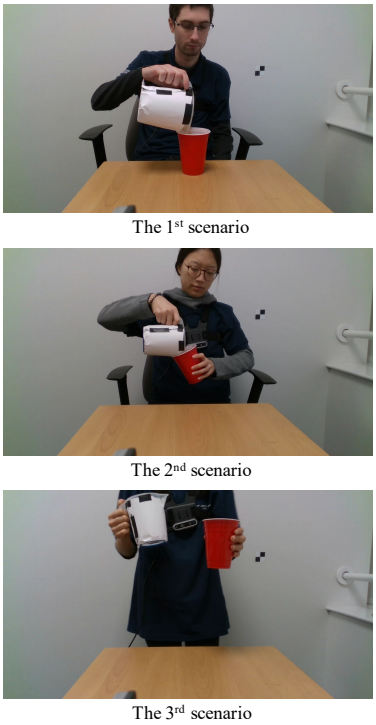

Fig. 5. The example visualization of three different scenarios in the CORSMAL Containers Manipulation dataset.

Table 2. Model and library acquisitions in the experiments.

Pre-trained models and libraries	Acquisition
YOLOv4 pre-trained model	`https://github.com/kiyoshiiriemon/yolov4_darknet`
Librosa lib [20]	`https://librosa.org/doc/latest/index.html`

4.3 Evaluation Metric

For classification tasks (i.e., Task (1) and Task (2)), we used the Weighted Average F1-score (WAFS) as the evaluation metric. For the capacity estimation task (i.e., Task (3)), the Average Capacity Score (ACS) under the relative absolute error is employed in the experiments. For evaluating the filling mass estimation, the Average filling Mass Score (AMS) under the relative absolute error is used to evaluate the performance of our proposed method.

4.4 Experimental results

Table 3, Table 4, Table 5 and Table 6 show the performance results for the filling mass estimation, Task (1), Task (2) and Task (3) respectively. Though the Gaus-

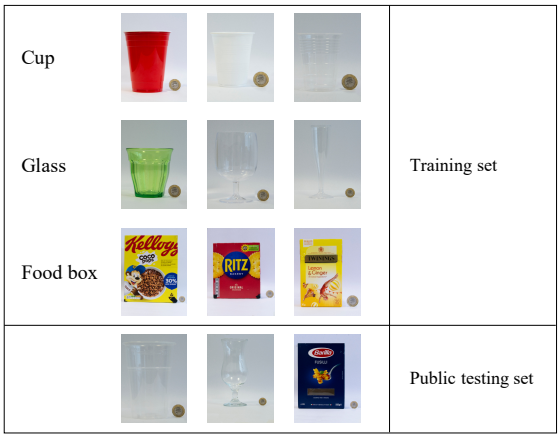

Fig. 6. Containers from training set and public testing set of CORSMAL Containers Manipulation dataset.

sian process regression showed effectiveness on container capacity estimation, we find that our method is relatively weak on filling level and filling type classification. The reason would lie in the audio feature extraction, which the number of signal points we select would be the background noise in the recordings of complex scenarios (i.e., the container is held by the subject, or the subject is potentially visible from one third-person view camera only). Therefore, the trained models would probably classify the filling level or filling type as the empty. In the future, we plan to extract the spectrogram based on the regular time windows.

Table 3. Performance on filling mass estimation.

Team/Baseline	Performance		
	Public testset	Private testset	overall
Because It's Tactile	64.98	65.15	65.06
HVRL	63.32	61.01	62.16
Ours	52.80	54.14	53.47
NTNU-ERC	38.56	39.80	39.18
Baseline (Random)	38.47	31.65	35.06
Challengers	29.25	23.21	26.23
Baseline (SCC-Net)	28.02	22.92	25.47
Baseline (Mask R-CNN + RN18)	19.46	25.47	14.53

5 Conclusion

In this paper, we present the solution for filling mass estimation. In our method, the container detection by YOLOv4 is served as the prior, then we extract the au-

Table 4. Performance on filling level classification.

Team/Baseline	Performance		
	Public testset	Private testset	overall
Baseline (SCC-Net)	84.21	80.98	82.66
Because It's Tactile	78.14	78.14	79.65
HVRL	82.63	74.43	78.56
Challengers	50.73	47.08	48.71
Baseline (Mask R-CNN + RN18)	58.51	32.93	47.00
Ours	44.31	42.70	43.53
Baseline (Random)	38.47	31.65	35.06
NTNU-ERC	-	-	-

Table 5. Performance on filling type classification.

Team/Baseline	Performance		
	Public testset	Private testset	overall
HVRL	97.83	96.08	96.95
Because It's Tactile	93.83	94.70	94.26
Baseline (SCC-Net)	93.34	92.85	93.09
NTNU-ERC	81.97	91.67	86.89
Challengers	78.58	71.75	75.24
Ours	41.77	41.90	41.83
Baseline (Random)	21.24	27.52	27.52
Baseline (Mask R-CNN + RN18)	30.85	13.04	23.05

Table 6. Performance on container capacity estimation.

Team/Baseline	Performance		
	Public testset	Private testset	overall
NTNU-ERC	66.92	67.67	67.30
Ours	63.00	62.14	62.57
Because It's Tactile	60.56	60.58	60.57
HVRL	57.19	52.38	54.79
Baseline (Random)	31.63	17.53	24.58
NTNU-ERC	-	-	-
Challengers	-	-	-

dio feature into MLPs for filling level and filling type classification, and conduct Gaussian process regression for container capacity estimation. The experimental results indicate the effectiveness of our proposed method for filling mass estimation task. The rank of our proposed method in the CORSMAL Challenge of Multi-modal Fusion and Learning For Robotics also indicates the effectiveness.

References

1. A. Xompero, R. Sanchez-Matilla, R.M., Cavallaro, A.: CORSMAL Containers Manipulation (1.0) [Data set], `https://doi.org/10.17636/101CORSMAL1`
2. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing **22**(10), 1533–1545 (2014)
3. Bae, H., Brophy, E., Chan, R.H., Chen, B., Feng, F., Graffieti, G., Goel, V., Hao, X., Han, H., Kanagarajah, S., et al.: Iros 2019 lifelong robotic vision: Object recognition challenge [competitions]. IEEE Robotics & Automation Magazine **27**(2), 11–16 (2020)
4. Bhattacharyya, R., Floerkemeier, C., Sarma, S.: Rfid tag antenna based sensing: Does your beverage glass need a refill? In: 2010 IEEE International Conference on RFID (IEEE RFID 2010). pp. 126–133. IEEE (2010)
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
6. Brandi, S., Kroemer, O., Peters, J.: Generalizing pouring actions between objects using warped parameters. In: 2014 IEEE-RAS International Conference on Humanoid Robots. pp. 616–621. IEEE (2014)
7. Clarke, S., Rhodes, T., Atkeson, C.G., Kroemer, O.: Learning audio feedback for estimating amount and flow of granular material. Proceedings of Machine Learning Research **87** (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Do, C., Burgard, W.: Accurate pouring with an autonomous robot using an rgb-d camera. In: International Conference on Intelligent Autonomous Systems. pp. 210–221. Springer (2018)
10. Do, C., Schubert, T., Burgard, W.: A probabilistic approach to liquid level detection in cups using an rgb-d camera. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2075–2080. IEEE (2016)
11. Griffith, S., Sukhoy, V., Wegter, T., Stoytchev, A.: Object categorization in the sink: Learning behavior–grounded object categories with water. In: Proceedings of the 2012 ICRA Workshop on Semantic Perception, Mapping and Exploration. Citeseer (2012)
12. Gu, S., Holly, E., Lillicrap, T., Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 3389–3396. IEEE (2017)
13. Huang, Y., Sun, Y.: Learning to pour. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7005–7010. IEEE (2017)
14. Ikeno, S., Watanabe, R., Okazaki, R., Hachisu, T., Sato, M., Kajimoto, H.: Change in the amount poured as a result of vibration when pouring a liquid. In: Haptic Interaction, pp. 7–11. Springer (2015)

15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
17. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. The International Journal of Robotics Research **34**(4-5), 705–724 (2015)
18. Liang, H., Li, S., Ma, X., Hendrich, N., Gerkmann, T., Sun, F., Zhang, J.: Making sense of audio vibration for liquid height estimation in robotic pouring. arXiv preprint arXiv:1903.00650 (2019)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
20. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. vol. 8, pp. 18–25 (2015)
21. Muhlig, M., Gienger, M., Hellbach, S., Steil, J.J., Goerick, C.: Task-level imitation learning using variance-based movement optimization. In: 2009 IEEE International Conference on Robotics and Automation. pp. 1177–1184. IEEE (2009)
22. Nair, A., Bahl, S., Khazatsky, A., Pong, V., Berseth, G., Levine, S.: Contextual imagined goals for self-supervised robotic learning. In: Conference on Robot Learning. pp. 530–539. PMLR (2020)
23. Pastor, P., Hoffmann, H., Asfour, T., Schaal, S.: Learning and generalization of motor skills by learning from demonstration. In: 2009 IEEE International Conference on Robotics and Automation. pp. 763–768. IEEE (2009)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Neural Information Processing Systems (NeurIPS). pp. 8024–8035 (2019)
25. Paulius, D., Huang, Y., Milton, R., Buchanan, W.D., Sam, J., Sun, Y.: Functional object-oriented network for manipulation learning. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2655–2662. IEEE (2016)
26. Paulius, D., Jelodar, A.B., Sun, Y.: Functional object-oriented network: Construction & expansion. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–7. IEEE (2018)
27. Pierson, H.A., Gashler, M.S.: Deep learning in robotics: a review of recent research. Advanced Robotics **31**(16), 821–835 (2017)
28. Pithadiya, K.J., Modi, C.K., Chauhan, J.D.: Selecting the most favourable edge detection technique for liquid level inspection in bottles. International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) ISSN pp. 2150–7988 (2011)
29. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
30. Rozo, L., Jiménez, P., Torras, C.: Force-based robot learning of pouring skills using parametric hidden markov models. In: 9th International Workshop on Robot Motion and Control. pp. 227–232. IEEE (2013)
31. Saal, H.P., Ting, J.A., Vijayakumar, S.: Active estimation of object dynamics parameters with tactile sensors. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 916–921. IEEE (2010)

32. Sanchez-Matilla, R., Chatzilygeroudis, K., Modas, A., Duarte, N.F., Xompero, A., Frossard, P., Billard, A., Cavallaro, A.: Benchmark for human-to-robot handovers of unseen containers with unknown filling. IEEE Robotics and Automation Letters **5**(2), 1642–1649 (2020)
33. She, Q., Feng, F., Hao, X., Yang, Q., Lan, C., Lomonaco, V., Shi, X., Wang, Z., Guo, Y., Zhang, Y., et al.: Openloris-object: A robotic vision dataset and benchmark for lifelong deep learning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 4767–4773. IEEE (2020)
34. Shi, X., Li, D., Zhao, P., Tian, Q., Tian, Y., Long, Q., Zhu, C., Song, J., Qiao, F., Song, L., et al.: Are we ready for service robots? the openloris-scene datasets for lifelong slam. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3139–3145. IEEE (2020)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR 2015 (2015)
36. Yang, P.C., Sasaki, K., Suzuki, K., Kase, K., Sugano, S., Ogata, T.: Repeatable folding task by humanoid robot worker using deep learning. IEEE Robotics and Automation Letters **2**(2), 397–403 (2016)