Making Sense of Audio Vibration for Liquid Height Estimation in Robotic Pouring

Hongzhuo Liang[1], Shuang Li[1], Xiaojian Ma[2,3],
Norman Hendrich[1], Timo Gerkmann[4], Jianwei Zhang[1]

***Abstract*— In this paper, we focus on the challenging perception problem in robotic pouring. We make use of audio vibration sensing and design a deep neural network PouringNet to predict the liquid height from the audio fragment during the robotic pouring task. PouringNet is trained on our collected real-world pouring dataset with multimodal sensing data, which contains more than 3000 recordings of audio, force feedback, video and trajectory data of the human hand that performs the pouring task. Evaluations on the dataset and robotic hardware demonstrate that our model generalizes well across different pouring settings. The related code, dataset, and video are available at https://lianghongzhuo.github.io/AudioPouring.**

I. Introduction

Robotic pouring [1] is a crucial robotic task in both domestic and industrial environments. Recent approaches to solving robotic pouring problem mostly rely on visual sensing [2]. By leveraging a camera situated in front of the target container, the current liquid height can be regressed from the visual features of the captured image. However, these approaches cannot generalize to opaque containers since the liquid height cannot be seen or could suffer from poor estimation error. On the other hand, haptic sensing is another important modality for the perception of robotic pouring. For example, when the force and torque feedback on the manipulator is available, we can either estimate the volume of liquid being poured or directly learn a pouring policy in an end-to-end manner [3]. However, the correlation between haptic information and the pouring liquid can be rather complicated and are varied among different end effectors and containers. These drawbacks in existing perception methods suggest that the robust and accurate perception in robotic pouring remains an open problem.

To this end, we propose to tackle these issues by leveraging the modality of acoustics [4]. Inspired by how human judge the liquid height during pouring with their hearing, we try to design a model that can estimate the position of liquid height with audio vibration. This is based on the observation that the vibrational frequency of the air in the container will change as the level of liquid rises during the pouring procedure. Moreover, estimating liquid height using audio vibration is immediate. Thus there is no need to explicitly perform an integration, which further reduces the prediction bias and achieves more accurate results.

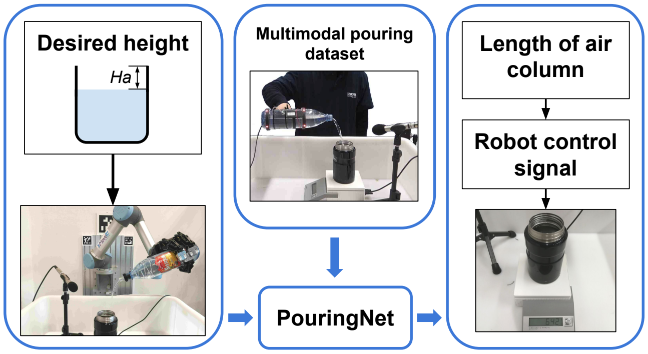


Fig. 1. Our robotic pouring system. (Left) Given the robot a target liquid level, audio vibrations during the pouring manipulation by a robot are recorded by a microphone then fed into PouringNet. (Center) PouringNet is trained offline to predict the length of the air column of the target container from our multimodal pouring dataset. (Right) The length of air column H_a predicted by PouringNet is used to guide the robotic pouring.

II. Data Preparation

Our dataset setup as shown in Fig. 2. It includes a source container, three different target containers (referred to as glass, thermos, and mug, shown in the first three items of Fig 4(b)), a Behringer B-5 microphone (44.1 kHz), an ATI Mini40 Force / Torque sensor (500 Hz), a Maul Logic digital scale (1 Hz), a Logitech web camera (30 Hz), and a PhaseSpace Impulse X2E motion tracking system (240 Hz). The height of the glass, thermos, and mug respectively are 127 mm, 150 mm and 99 mm. We placed the source containers relative to the bottom center of the microphone at a horizontal distance of 250 mm and a vertical distance of 750 mm. For each pouring trial, the subject held the handle of the source container and started pouring task at an angle varying $8^\circ \pm 15^\circ$ and at a random position which is relative to the mouth of the target container ranging from 450 mm to 500 mm. Pouring during the training only involved water.

III. Pouring Network

We design a recurrent deep network (PouringNet) P_θ to predict the length of the air column H_a. θ defines the parameters of our proposed PouringNet. The network architecture is shown in Fig. 3.

The height predictor (a 2-layer MLP) takes the recurrent vector as input and performs a regression of the temporary length of the air column. The height predictor is supervised with a mean squared error (MSE) loss $\mathcal{L}_{height}$

[1] TAMS (Technical Aspects of Multimodal Systems), Department of Informatics, Universität Hamburg

[2] Center for Vision, Cognition, Learning, and Autonomy, Department of Statistics, University of California, Los Angeles (UCLA)

[3] Beijing National Research Center for Information Science and Technology (BNRist), State Key Lab on Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University

[4] SP (Signal Processing), Department of Informatics, Universität Hamburg

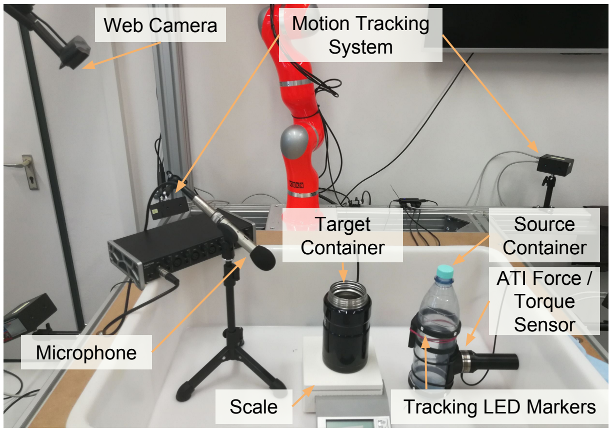


Fig. 2. Pouring setup used to collect our multi-modal dataset.

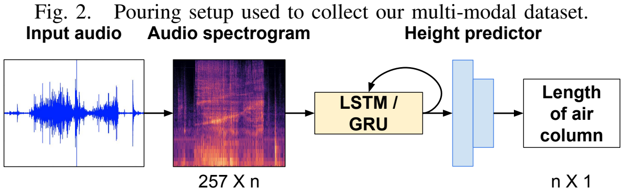


Fig. 3. PouringNet architecture. The blue rectangular denotes a fully-connected layer following with a batch normalization layer and a rectified linear unit.

$$\mathcal{L}_{height} = \|\hat{H}_a - H_a\|^2. \tag{1}$$

In addition, leveraging the principle that the liquid height in the target container is monotonically increased, we introduce an auxiliary $\mathcal{L}_{mono}$ to enforce the estimated length of the air column being decreasing along the time t

$$\mathcal{L}_{mono} = \sum_t [\max(0, (\hat{H}_{a_{t+1}} - \hat{H}_{a_t}))]. \tag{2}$$

Overall loss. Combining with $\mathcal{L}_{height}$ and $\mathcal{L}_{mono}$, the complete training objective for PouringNet is defined by $\mathcal{L}_{audio}$

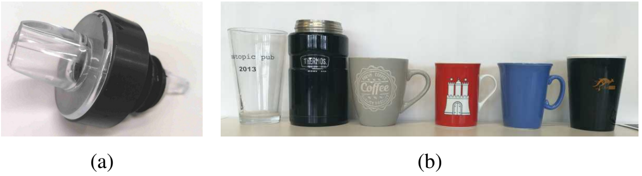
$$\mathcal{L}_{audio}(\theta) = \mathcal{L}_{height} + \alpha \cdot \mathcal{L}_{mono}, \tag{3}$$

where α is a hyperparameter for balancing these two loss functions. In our implementation, we set it to 0.01 for the best performances via some preliminary experiments.

IV. EXPERIMENT

We evaluate the adaptability and robustness of our audio-based perception method in pouring experiments with a UR5 robot. We design four groups of robotic experiments: evaluation on different target containers, different microphone positions, different initial liquid heights and different types of liquid. Due to the space limit, here we only present the results on different target containers.

In this experiment, we kept the distance between the target containers and the microphone the same as in our original dataset. During the robotic pouring, we varied the target length of the air column between [40 mm, 50 mm, 60 mm, 70 mm, 80 mm] for three target containers in our dataset and three unseen target containers in Fig. 4(b). The water was poured for five times to each considered height of each target container.



(a) (b)

Fig. 4. (a) The spout equipped on the source container in robotic experiments. (b) From left to right, the first three target containers are the target containers in our datasets: a glass, a stainless steel cup, and a mug; the latter three target containers are the unseen containers used in robotic experiments: a red mug, a blue mug and a plastic cup.

For numerical results, we converted the height error of each cup to weight error in this experiment shown in Table I. In previous work on robotic pouring, Schenck et al., [2] reported a mean error of 38 ml and Do et al., [5] achieved a mean volume error 22.53 ml over three different target containers. Compared to their results, the robotic pouring with our audio-based perception system can achieve higher precision.

TABLE I

Absolute Mean Amount Errors and Standard Deviations

Glass	Thermos	Mug
9.54 ± 7.81mm	9.91 ± 8.48mm	13.79 ± 11.04mm

Red Mug	Blue Mug	Plastic Cup
7.92 ± 7.14mm	6.42 ± 6.31mm	10.72 ± 8.70mm

V. Conclusion and Future Work

This paper presents a real-time perception system used for estimating the liquid height in robotic pouring. We offer a multimodal pouring dataset including audio-frequency recordings, liquid real-time weight, force and torque feedback, video and motion trajectories. With this dataset, we develop a robust audio-based perception model named PouringNet. Making use of the force, motion trajectories and visual data from our multimodal dataset in robotic pouring would be an exciting direction of future research.

ACKNOWLEDGMENT

This research was funded by the German Research Foundation (DFG) and the National Science Foundation of China (NSFC) in project Crossmodal Learning, DFG TRR-169/NSFC 61621136008. And it was partially supported by project STEP2DYNA (691154).

References

[1] M. Tamosiunaite, B. Nemec, A. Ude, and F. Wörgötter, "Learning to pour with a robot arm combining goal and shape learning for dynamic movement primitives," *Robotics and Autonomous Systems*, vol. 59, no. 11, pp. 910–922, 2011.

[2] C. Schenck and D. Fox, "Visual closed-loop control for pouring liquids," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[3] L. Rozo, P. Jiménez, and C. Torras, "Force-based robot learning of pouring skills using parametric hidden markov models," in *IEEE International Workshop on Robot Motion and Control*, 2013.

[4] J. A. Cacciola and R. J. Cacciola, "Liquid level detector using audio frequencies," 1997, uS Patent 5,623,252.

[5] C. Do and W. Burgard, "Accurate pouring with an autonomous robot using an rgb-d camera," in *Intelligent Autonomous Systems(IAS)*, 2019.