# Analysis on viewing angles for filling level estimation

Vishal Kashyap
210684838
ex21021@qmul.ac.uk

*Abstract*

**This paper investigates the effect of viewing angles for estimating filling level in a container. The objective is to find the best viewing angle and the discuss the add-value of combining multiple angles. This is investigated by generating results using different combinations of views on the same model. It is shown that for the defined model, there is no significant add-value in using more than one view. An alternate architecture is also discussed to combine information for these views better.**

## I. Introduction

Recent advances in computer vision and Artificial Intelligence have significantly improved robot manipulation and control tasks. Having been limited to industries, researchers are now focusing on creating robots for home chores. Chores like pouring liquid into containers or human-robot handovers involve estimating their filling level. There has been research on estimating the filling level using single RGB images (Modas *et al.,* 2021; Mottaghi *et al*., 2017). Though applicable, robots deployed with a camera can use a sequence of frames, adding to the information already present in the single frame. Vision researchers have focused on this approach but have relied on a pre-selected set of containers (Do and Burgard, 2018; Schenck and Fox, 2017). The algorithm must generalise to unseen containers and their contents for this application to succeed.

Iashin *et al.* (2020) use a sequence of frames from four different angles to determine the filling level. They focus on the generalisation of the model by testing it on containers not seen during training. However, they do not analyse the best camera angle or the add-value in using multiple cameras from different angles. This paper will build on the research done by Iashin *et al.* (2020) and answer these questions. Section II will set up the problem and state the objectives. Section III will discuss a key work in this area. Datasets used and evaluation measures will be described in Section IV. Section V and VI will discuss the results obtained and their conclusions.

## II. Problem Definition

This paper will evaluate the usefulness of different camera angles for estimating the filling level of container using computer vision. The vision architecture and features (camera feed from four different angles) developed by Iashin *et al.* (2020) will be used. This paper has three primary objectives. First, evaluate the performance when only one of the four viewing angles are used to train the model. Second, determine if using multiple angles add value to using just one angle. Third, limit the training dataset to transparent and translucent containers and repeat the tests. The third objective is essential as using RGB feed from the camera is not effective on opaque objects and may lead to noisy results (Mottaghi *et al.,* 2017).

For all tests, the features, model architecture, training procedure and evaluating criteria have been kept identical to Iashin *et al.* (2020). Changes have only been made to the amount of training data used, depending on the viewing angles chosen for the test. Thus, the results obtained by them will be used as a benchmark.

These tests will help us choose the best camera angle for filling tasks. Understanding the add-value of using multiple angles will help make decisions on the optimal number of cameras to use under economic constraints. This is also essential to understand and make changes to the model architecture to use all angles in a better way.

## III. Key Work

Iashin *et al.* (2020) have formulated filling level estimation as a classification problem with three possible values, 0%, 50% and 90%. They convert about 0.5 seconds of video streams to RGB (R(2+1)d features which are 512-d vectors. These features achieve state of the art performance in action recognition tasks and provide significant benefits over 3D convolutions (Tran *et al*., 2018). Streams from each camera are encoded separately using GRU units. Softmax is applied to the sum of logits from these GRU units to estimate the filling level (Fig. 1, left).

## IV. Evaluation Criteria

The model has been trained on the Corsmal containers manipulation dataset (Xompero *et al.,* 2020). This dataset has recordings of 15 containers being filled by 12 different subjects under two different backgrounds and lighting conditions. The 15 containers are split into 5 boxes, 5 drinking glasses and 5 cups. The containers are either transparent, translucent or opaque. The dataset provides recording from multiple sensors, of which we will focus on the RGB streams from 4 cameras.

The dataset has been split into public train, public test and private test. Training is done using a 3-fold-cross-validation on the public train set with 30 epochs per fold. An average of the best F1 scores of the three validation sets is used to track the performance of different classifiers.

## V. Discussion

### *A. Tests on all containers*

For the first set of tests, all 9 containers were considered. Table I below shows that view 3 gave the best results (0.734) when the model was trained on features from a single view. This performance is comparable to the benchmark in which all views were used (0.747). This raises concerns on the add-value of views 1, 2 and 4.

Results from using a combination of two and three views build on this concern. When view 3 is combined with other views (in either three-view or two view tests), we get better F1 scores. For example, view 3 alone has a score of 0.734; combining it with view 2 gives 0.742, using views 2, 3, 4 give a score of 0.758. However, combining views other than view 3 gives almost the same or lower F1 score. Nonetheless, many scores in single, two view and three view tests are very close

TABLE I. Results from tests with all 9 containers

| | **Single View tests** | | | | |
|---|---|---|---|---|---|
| View | ***All*** | ***1*** | ***2*** | ***3*** | ***4*** |
| F1 score | 0.747 | 0.728 | 0.718 | 0.734 | 0.724 |
| | **Three View tests** | | | | |
| View | ***All*** | ***1, 2, 3*** | ***1, 2, 4*** | ***1, 3, 4*** | ***2, 3, 4*** |
| F1 score | 0.747 | 0.752 | 0.726 | 0.745 | 0.758 |
| | **stack-feature model test** | | | | |
| ***Model*** | ***Benchmark*** | | | ***stack-feature model*** | |
| F1 score | 0.747 | | | 0.751 | |

to the benchmark which uses all four views. This raises concerns about the model's ability to combine information from multiple views.

### B. *Tests on transparent and translucent containers*

Four out of nine containers tested in section V.A are opaque. Using just vision information to estimate the filling level in an opaque container is not reasonable. To verify if the results from section V.A hold, the same set of tests were run limiting to five transparent or translucent containers.

As expected, the F1 scores for tests with 5 containers are significantly better. Like the 9 container tests, view 3 is the best while view 2 gives the worst score in single view tests. Again, the scores are comparable to the benchmark for single view, two view and three view tests.

TABLE II. Results from tests with all 5 containers

| | **Single view tests** | | | | |
|---|---|---|---|---|---|
| Camera | ***All*** | ***1*** | ***2*** | ***3*** | ***4*** |
| F1 score | 0.790 | 0.783 | 0.772 | 0.800 | 0.787 |
| | **Three view tests** | | | | |
| Camera | ***All*** | ***1, 2, 3*** | ***1, 2, 4*** | ***1, 3, 4*** | ***2, 3, 4*** |
| F1 score | 0.790 | 0.800 | 0.805 | 0.812 | 0.809 |
| | **stack-feature model test** | | | | |
| ***Model*** | ***Benchmark*** | | | ***stack-feature model*** | |
| F1 score | 0.790 | | | 0.853 | |

### C. *Stack-feature test*

In the benchmark model (Fig. 1, left), each GRU unit is trained separately using data from a single view. The information from different views is combined later while summing up the logits from all four models. An alternative model (Fig. 1, right) could stack R(2+1)d features from the four views and feed these to a single GRU unit. The stacked feature would now be a vector of dimension 2048-d. The stacked-feature model performs on par with the benchmark when tested on all 9 containers (see table I). When tested on 5 containers, the F1 score of the stacked-feature model is significantly better (see table II). The results from 5 container test should be more reliable due to the elimination of noisy prediction from opaque containers, as discussed in section II.

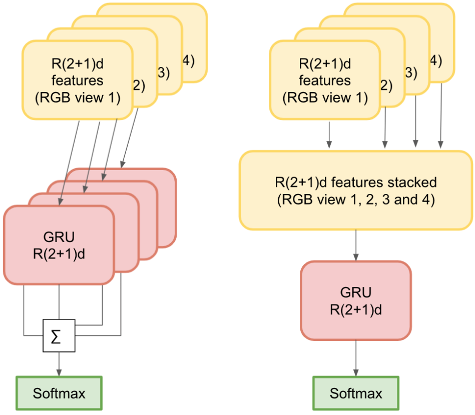
R(2+1)d features (RGB view 1)
2)
3)
4)
GRU R(2+1)d
Σ
Softmax
R(2+1)d features (RGB view 1)
2)
3)
4)
R(2+1)d features stacked (RGB view 1, 2, 3 and 4)
GRU R(2+1)d
Softmax

Fig. 1. Left: Model used by Iashin *et al.* (2020). Right: Modified architecture to use stacked features. (Modified from source: Iashin *et al.* (2020))

## VI. Conclusion

This article borrows the vision model from Iashin *et al.* (2020) and analyses the effect of multiple viewpoints on the estimation of filling levels in containers. It was found that, when using feed from a single view, view 3 is the best while view 2 gave the worst result. Using multiple views showed no significant add-value as the scores from single, two, three and four views tests were comparable. An alternate model (stacked-feature) was suggested to combine the information from different views better. This model gave significantly better results compared to the benchmark.

The next steps would be to test these models on public and private datasets to see if the results hold. Future research can focus on finding better ways to combine streams from multiple views. Another direction could be finding the best angle or combination of angles for the task (not limited to the given four).

## References

Iashin, V., Palermo, F., Solak, G. and Coppola, C. (2021). Top-1 CORSMAL Challenge 2020 Submission: Filling Mass Estimation Using Multi-modal Observations of Human-Robot Handovers. *Pattern Recognition. ICPR International Workshops and Challenges*, pp.423–436.

Xompero, A., Sanchez-Matilla, R., Mazzon, R., Cavallaro, A.: Corsmal containers manipulation (2020). https://doi.org/10.17636/101CORSMAL1, http://corsmal. eecs.qmul.ac.uk/containers manip.html

Modas, A., Xompero, A., Sanchez-Matilla, R., Frossard, P. and Cavallaro, A. (2021). Improving Filling Level Classification with Adversarial Training. *2021 IEEE International Conference on Image Processing (ICIP).*

Mottaghi, R., Schenck, C., Fox, D. and Farhadi, A. (2017). See the Glass Half Full: Reasoning About Liquid Containers, Their Volume and Content. *2017 IEEE International Conference on Computer Vision (ICCV).*

Do, C. and Burgard, W. (2018). Accurate Pouring with an Autonomous Robot Using an RGB-D Camera. *Intelligent Autonomous Systems 15*, pp.210–221.

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

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M. (2018). *A Closer Look at Spatiotemporal Convolutions for Action Recognition.* [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8578773 [Accessed 6 Nov. 2020]