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

## Solution for CORSMAL Challenge of Multi-modal Fusion and Learning For Robotics in ICPR2020

By Concatenation (Qi Liu, Fan Feng, Chuanlin Lan & Rosa H.M. Chan)
With affiliation City University of Hong Kong
On 14 Jan. 2021

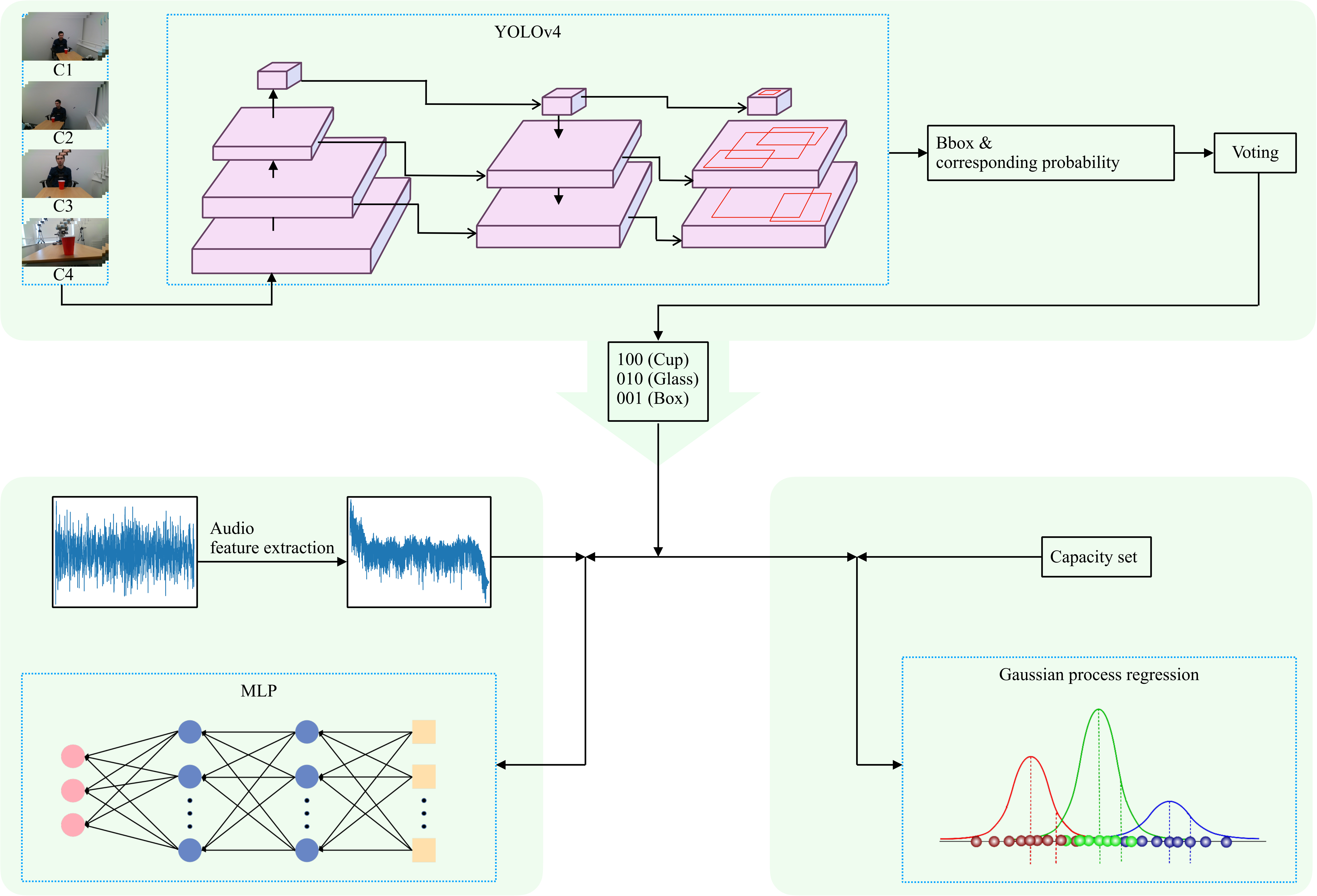
CityU

# Inspiration

- Containers have various nature frequencies due to their physical properties.
  - Physical properties: material, stiffness, texture.
- Given the specific container, the vibrational frequency varies if poured with different filling contents.
  - Filling contents: empty, water, pasta and rice.

# Method

- **Based on container prior.**
    - Modality used: RGB from all views;
    - YOLOv4[1] pre-trained on MS COCO[2].
- **Filling level and filling type classification.**
    - Modality used: Audio;
    - Multi-Layer Perceptron (MLP) with 2 hidden layers.
- **Container capacity estimation.**
    - Modality used: RGB from all views;
    - Gaussian process regression to fit category-based capacity distribution.





1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dolla ́r, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)

# Method (Con’t)

## Filling level and filling type classification

1. Audio feature extraction.
   - Re-sample at 16,600Hz;
   - Select the last 32,000 data points;
   - Discrete fourier transform (DFT).
2. Classification model.
   - Two hidden-layer (3,096-512) MLP;
   - Learning rate: 0.05;
   - #Epochs: 200.

# Method (Con't)

## Container capacity estimation

1. Infer the category label $x_i$ from the object detection model;
2. Construct the training set;
   - $\mathscr{D} = (\mathbf{X}, \mathbf{y}) = \left\{ \left( \mathbf{x}_i, y_i \right) \mid i = 1, \ldots, N \right\}.$
3. Conduct the Gaussian process regression.
   - For a new input $X^*$ in test-set, we have $\hat{\mathbf{y}}^* = K\left(X^*, X\right) K(X, X)^{-1} \mathbf{y}.$

   $\mathbf{y}^*$ - The predicted value.

   $K$ - the covariance function defined by $K(A, B)_{ij} = \exp\left( -\frac{1}{2} \left| A_i - B_j \right|^2 \right).$

# Experimental result

## 2nd runner-up

| Task | Performance | | |
|---|---|---|---|
| | Public test | Private test | Overall |
| **Task 1 - Filling level** | 44.31 | 42.70 | 43.53 |
| **Task 2 - Filling type** | 41.77 | 41.90 | 41.83 |
| **Task 3 - Container capacity** | 63.00 | 62.14 | 62.57 |
| **Overall Task - Filling mass** | 52.80 | 54.14 | 53.47 |

- Weak in Task 1 and Task 2.
  - # signal points we select would be the background noise in the recordings of complex scenarios;
  - Future work: extract the spectrogram based on the regular time windows.

# Conclusion

Method summary

- Container detection is served as the prior;
- Audio features is fed into MLPs for filling level and filling type classification;
- Gaussian process regression for container capacity estimation.

The proposed method

- can be tuned for better performance or computation efficiency.
    - e.g., be equipped with different backbone models.
- is useful for smart robots helping with daily activities like objects pick-up, place and handovers.

# Thank you for listening !

Report