chist-era

CORSMAL
Collaborative object recognition, shared manipulation and learning

The CORSMAL Challenge

Audio-visual object classification for human-robot collaboration

Performance scores

SORBONNE UNIVERSITÉ

Queen Mary University of London

EPFL

Performance scores

The CORSMAL challenge evaluates and ranks the teams by assigning a 100 point-based score that accounts for a set of objective performance scores and an assessment of the submitted source code for reproducibility. The integrity of the submitted results is ensured by having a public test set with no annotations available to the teams and a private test set with both the data and the annotations not available to the teams. For the public test set, teams run their models on their own and submit their results in a constrained amount of time. For the private test set, organisers will install, run and evaluate the teams' models.

To provide a sufficient granularity into the behaviour of the various components of the pipeline, we use 10 performance scores for the challenge tasks across public and private sets. The first 7 scores quantify the accuracy of the estimations for the 5 main tasks. The last 3 scores are an indirect evaluation of the impact of the estimations on the quality of human-to-robot handover and delivery of the container by the robot. Performance scores will be also computed individually for the public CCM test set, the private CCM test set, and their combination. The scores cover filling level, filling type, container capacity, container width at the top, width at the bottom, and height, container mass, filling mass, object mass (container + filling), and the delivery of the container upright and at a pre-defined target location (see Tab. 1).

T1 and T2

For the tasks of filling level and filling type classification, we compute precision, recall, and F1-score for each class k across all the configurations belonging to class k, J_k. *Precision* is the number of true positives over the total number of true positives and false positives for each class k (P_k). *Recall* is the number of true positives over the total number of true positives and false negatives for each class k (R_k). The *F1-score* is the harmonic mean of precision and recall, defined as

$$F_k = 2\frac{P_k R_k}{P_k + R_k}. \tag{1}$$

We then compute the weighted average F1-score, F_1, across the K classes,

$$F_1 = \sum_{k=1}^{K} \frac{J_k F_k}{J}, \tag{2}$$

where J is the total number of configurations (e.g., for either the public test set, the private test set, or their combination) and J_k is the number of configurations associated to class k. Note that $K = 3$ for filling level classification and $K = 4$ for filling type classification.

T3, T4 and T5

For the tasks of container capacity estimation and container mass estimation, we compute the relative absolute error between the estimated measure $a \in \{\tilde{}^j; \tilde{m}_c^j\}$, and the true measure, $b \in \{^j; m_c^j\}$, for each configuration j,

$$\varepsilon(a; b) = \frac{|a - b|}{b}. \tag{3}$$

For the task of container dimensions estimation, where $a \in \{\tilde{w}_t^j; \tilde{w}_b^j; \tilde{h}^j\}$ and b is the corresponding annotation, we use the normalisation function $\rho_1(\cdot\,;\cdot)$ [2]:

$$\rho_1(a; b) = \begin{cases} 1 & \text{if } b = a = 0 \\ 1 - \frac{|a - b|}{b} & \text{if } |a - b| < b \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that the first condition is never used in our case, as $b > 0$.

For the indirect high-level filling mass estimation[1], we compute the relative absolute error between the estimated,

[1] Note that the score for filling mass is not a linear combination of the scores outputted for filling level classification, filling type classification, and capacity estimation, but it takes into consideration the formula for computing the filling mass based on the estimations of each task for each configuration. This means that a method with lower T1, T2 and T3 scores can obtain a higher filling mass score compared to other methods because the performance on each run is more accurate in general.

$\hat{m}_f^j$, and the true filling mass, m_f^j, for each configuration j, unless the annotated mass is zero (empty filling level),

$$(\hat{m}_f^j, m_f^j) = \begin{cases} 0, & \text{if } m_f^j = 0 \wedge \hat{m}_f^j = 0; \\ \hat{m}_f^j, & \text{if } m_f^j = 0 \wedge \hat{m}_c^j \neq 0; \\ \frac{|\hat{m}_f^j - m_f^j|}{m_f^j}, & \text{otherwise.} \end{cases} \tag{5}$$

With reference to Tab. 1, we compute the score, $s_i \in [0, 1]$ (where 1 is best), across all the configurations J for each measure as follows:

$$s_i = \begin{cases} \frac{1}{J} \sum_{j=1}^{J} \mathbb{1} e^{-\varepsilon(a,b)} & \text{if } a \in \{\hat{m}_c^j, \hat{w}_t^j, \hat{w}_b^j, \hat{h}^j\} \\ \frac{1}{J} \sum_{j=1}^{J} \mathbb{1} \, \mathbb{1}(a,b) & \text{if } a \in \{\hat{w}^j, \hat{w}_b^j, \hat{h}^j\} \\ \frac{1}{J} \sum_{j=1}^{J} \mathbb{1} e^{-\varepsilon(a,b)} & \text{if } a = \hat{m}_f^j, \end{cases} \tag{6}$$

with $i = \{3, \ldots, 8\}$. The indicator function $\mathbb{1} \in \{0, 1\}$ is 0 only when $a \in \{\hat{m}_c^j, \hat{w}_t^j, \hat{w}_b^j, \hat{h}^j, \hat{m}_f^j\}$ in recording j is not estimated. Note that estimated and annotated measures are strictly positive, $a > 0$ and $b > 0$, except for filling mass that can be 0 when the container is empty.

Object safety and accuracy of delivery

An innovative feature of the CORSMAL challenge is the use of the estimation results by the teams in the context of a specific use-case. For this reason, we evaluate the object safety and the delivery accuracy using a human-to-robot handover simulator [1].

Object safety is the probability that the predicted force applied by the robot (based on the object mass given by the sum of container and filling masses) will enable a gripper to hold the container without dropping it or breaking it. We compute object safety as an exponential function that accounts for the difference between the predicted normal force F_j and the required normal force at the real object mass (as available to the simulator), $\hat{F}_j$, for each configuration j:

$$\psi_j = e^{\frac{|\hat{F}_j - F_j|}{\hat{F}_j} \ln(1-c)}; \tag{7}$$

where c controls the sensitivity of ψ_j. A negative difference represents an increase in the probability of dropping the container and a positive difference represents an increase in the probability of breaking the container. We

T1	T2	T3	T4	T5	Description	Unit	Measure	Score	Weight	Type	R2S
●	○	○	○	○	Filling level		λ^j	$s_1 = F_1(\lambda^1, \ldots, \lambda^J, \hat{\lambda}^1, \ldots, \hat{\lambda}^J)$	$\pi_1 = 1/8$	D	○
○	●	○	○	○	Filling type		τ^j	$s_2 = F_1(\tau^1, \ldots, \tau^J, \hat{\tau}^1, \ldots, \hat{\tau}^J)$	$\pi_2 = 1/8$	D	○
○	○	●	○	○	Capacity	mL	γ^j	$s_3 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j e^{-\varepsilon^j(\gamma^j, \hat{\gamma}^j)}$	$\pi_3 = 1/8$	D	○
○	○	○	●	○	Container mass	g	m_c^j	$s_4 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j e^{-\varepsilon^j(m_c^j, \hat{m}_c^j)}$	$\pi_4 = 1/8$	D	○
○	○	○	○	●	Width at top	mm	w_t^j	$s_5 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j \sigma_1(w_t^j, \hat{w}_t^j)$	$\pi_5 = 1/24$	D	○
○	○	○	○	●	Width at bottom	mm	w_b^j	$s_6 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j \sigma_1(w_b^j, \hat{w}_b^j)$	$\pi_6 = 1/24$	D	○
○	○	○	○	●	Height	mm	h^j	$s_7 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j \sigma_1(h^j, \hat{h}^j)$	$\pi_7 = 1/24$	D	○
●	●	●	○	○	Filling mass	g	m_f^j	$s_8 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j e^{-\varepsilon^j(m_f^j, \hat{m}_f^j)}$	$\pi_8 = 1/8$*	I	○
●	●	●	●	●	Object mass	g	m^j	$s_9 = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}_j \psi^j(m^j, \hat{F}^j)$	$\pi_9 = 1/8$*	I	●
●	●	●	●	●	Pose at delivery	(mm, °)	(α^j, β^j)	$s_{10} = \frac{1}{J} \sum_{j=1}^{J} \psi_j(\alpha^j, \beta^j, \eta, \phi)$	$\pi_{10} = 1/8$*	I	●
●	●	○	○	○	Joint filling type and level			$s_{11} = F_1(\lambda^1, \tau^1, \ldots, \hat{\lambda}^1, \hat{\tau}^1, \ldots)$	{	D	{
○	○	●	○	●	Container capacity and dimensions			$s_{12} = s_3/2 + (s_4 + s_5 + s_6)/6$	{	D	{
●	●	●	●	●	Overall score			$S = \sum_{l=1}^{10} \pi_l s_l$	{	I	{

KEY { T: task, D: direct score, I: indirect score, R2S: measured in the real-to-simulation framework [1].

* weighted by the number of performed tasks.

Table 1: Performance scores for a given test set (public, private or their combination). For a measure a, its corresponding ground-truth value is $\hat{a}$. All scores are normalised and presented in percentages (the higher, the better). $F_1(\,)$ is the weighted average F1-score. Filling amount and type are sets of classes (no unit).