Geometric versus Learned Scale Localization for Volume Estimation of Handheld Syringes in Clinical Video

Jan Maris | Supervisors: Nergis Tömen, Xucong Zhang

j.w.maris@student.tudelft.nl

1 Context & Motivation

- Medication preparation in paediatric and neonatal care is a high-risk process.
- Despite digital prescribing systems and mandated double-checks, medication errors still occur.
- Reading a syringe by eye is error-prone, and routine safety checks often miss it.
- **Goal:** explore the feasibility of *vision-assisted verification* of medication preparation steps

2 Background & Related Work

- Existing camera-based syringe-volume systems rely on a fixed camera or controlled setup.
- Our two methods build on established ideas:
  - Derive volume from plunger position and known capacity. *[1]*
  - Read volume from the syringe's silhouette using a per-type model *[2]*
  - Localize the scale's marks directly as keypoints, as done on analog gauges *[3]*

RESEARCH QUESTION

How accurately can syringe volume be estimated from real clinical video by a geometric, calibration-based method versus a learned keypoint detector, given an annotated liquid line?

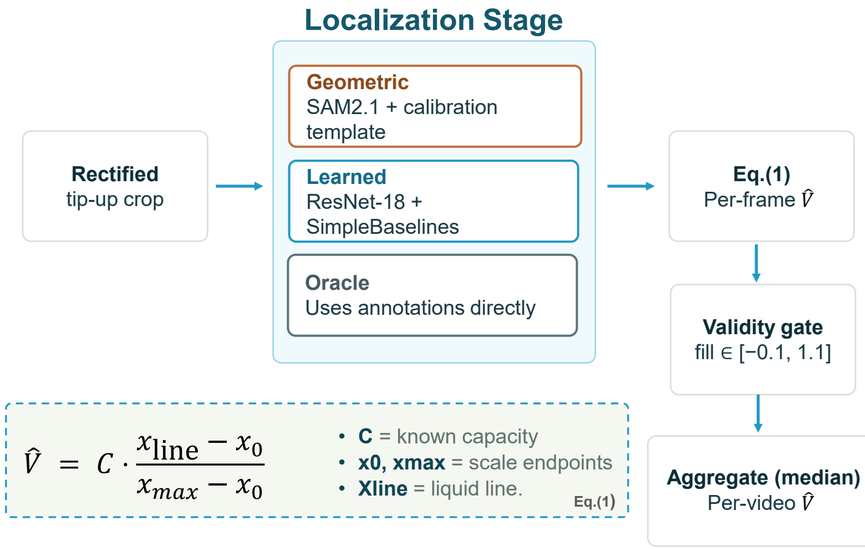
We isolate **one** sub-problem: localizing the two scale endpoints. The liquid line and bounding box are supplied as annotated inputs.

References

[1] Amit Nissan, Fadi Mahameed, Sapir Gershov, Aeyal Raz, and Shlomi Laufer. Efficient computer vision pipeline for automated anesthetic injection documentation. *Computer Assisted Surgery*, 30(1):2582020, 2025.
[2] Hem K. Regmi, Jerry Nesamony, Scott M. Pappada, Thomas J. Papadimos, and Vijay Devabhaktuni. A system for real-time syringe classification and volume measurement using a combination of image processing and artificial neural networks. *Journal of Pharmaceutical Innovation*, 14(4):341–358, 2019.
[3] Maurits Reitsma, Julian Keller, Kenneth Blomqvist, and Roland Siegwart. Under pressure: learning-based analog gauge reading in the wild. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

4 Methodology

SHARED VOLUME PIPELINE (PER FRAME → PER VIDEO)



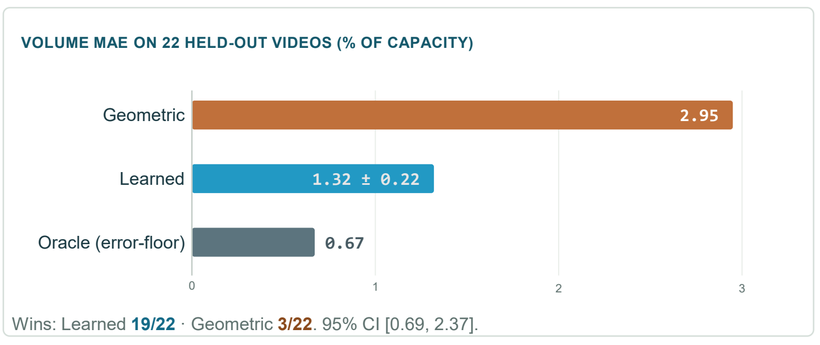
$$\hat{V} = C \cdot \frac{x_{\text{line}} - x_0}{x_{max} - x_0} \quad \text{Eq.(1)}$$

- **C** = known capacity
- **x0, xmax** = scale endpoints
- **Xline** = liquid line.

5 Experiments

- **H1: Accuracy.** Learned localizes better on free pose.
  Test: leave-one-video-out, 5 seeds; paired Wilcoxon / sign / permutation
- **H2: Generalization.** Learned error rises on an unseen type.
  Test: leave-one-type-out (hold each type out in turn).
- **H3: Error character.** If the geometric method's error is a systematic per-video offset and the learned method's is frame noise, aggregation should favor the learned method.
  Test: compare per-frame vs per-video MAE

6 Results



Wins: Learned **19/22** · Geometric **3/22**. 95% CI [0.69, 2.37].

Note: Oracle uses keypoint annotations directly, the gap to the oracle is the localization error

- **H1.** Across all frames: learned wins **19/22** compared to geometric, +1.63 pp, p = 0.001, robust across seeds.
  - Excluding the learned method's worst video (max mark off-frame): **0.86 ± 0.04 %cap**, near the 0.67 floor.
- **H2.** Type absent from training → **5.32 %cap** (~4×). Headline accuracy holds only for **known types.**
- **H3.** Aggregation gain (per-frame − per-video MAE) is **~1.6–2.6 %cap** for learned vs **1.0 %cap** for geometric, despite a similar single-frame error (per-frame MAE **3.97 vs 3.1–3.6**). Gain favors learned and geometric retains a larger per-video error

7 Limitations

- Holds for **trained syringe types only**;
- Ground truth is a **manual visual reading**; box, liquid line, per-frame orientation are annotated inputs
- Out-of-plane **tilt uncorrected:** part of the irreducible (oracle) error; it may also drive the geometric method's extra offset, **not measured directly.**
- Small single-institution set (26 videos, 4 types); fills below 0.41% capacity absent

8 Conclusions & Future Work

- Learned scale localization is feasible on free-hand syringes within known types: **1.32 ± 0.22 %cap** overall, improving to **0.86 ± 0.04 %cap** when the full scale is in frame.
- The two fail differently: geometric by **bias**, learned by **removable noise**

FUTURE WORK

- Synthetic multi-design training → close the unseen-type gap
- Read capacity from printed numerals (OCR) → drop known-type assumption
- Estimate pose → correct out-of-plane tilt.