ETH zürich CVL | BMIC USZ Universitäts Spital Zürich

MICCAI

Conformal forecasting for surgical instrument trajectory

Sara Sangalli[1,*], Gary Sarwin[1,*], Ertunc Erdil[1], Carlo Serra[2], Alessandro Ceretta[2,3], Victor Staartjes[2] and Ender Konukoglu[1]

[1] Computer Vision Lab, ETH Zurich, [2] University Hospital of Zurich, [3] University of Bologna

[*] Equal contributions

Introduction

- Machine learning techniques, including instrument trajectory forecasting, are transforming **real-time support** and **training in neurosurgery**.
- We apply conformal techniques to jointly quantify uncertainty in **instrument trajectory predictions**.
- We produce **uncertainty maps**, enabled by **conformal prediction**, as an essential step towards safe, interpretable, and trustworthy automated surgical systems.

Problem definition

- We build our forecasting network on [1]: processing of a surgical video frame **sequence** $s_t = x_{t-d:t} := \{x_\tau\}_{\tau=t-d}^{t}$ → **object detector** to identify anatomical structures and the surgical instrument → **latent space** mapping → NN predicts the **change in instrument center location** for the next h frames, modeled as a movement vector.

- **Goal:** build **uncertainty intervals** for the **phase** and **magnitude** of the forecast vector, which are guaranteed to contain the ground truth phase and magnitude with at least a user specified probability 1 – α.

Method

Split conformal prediction (CP) [2]

D_{cal} is a set of exchangeable video sequences $s_1, \ldots, s_n$, with corresponding ground truth motion vectors $v_1, \ldots, v_n$, and forecast vectors $\hat{v}_1, \ldots, \hat{v}_n$

s_{n+1} is drawn from the same ditribution, the conformity score for the phase (the same holds for magnitude) can be computed as absolute error residuals:

$$R_i^{CP} = |\angle v_i - \angle \hat{v}_i|, i \in D_{cal}$$

For the target coverage 1 – α, the quantile of the empirical distribution of the residuals is:

$$Q_{1-\alpha}(R^{CP}, D_{cal}) := (1-\alpha)\left(1+\frac{1}{|D_{cal}|}\right)\text{-th empirical quantile of } \{R_i^{CP} : i \in D_{cal}\}$$

This results in the following prediction interval for the phase for the test sample:

$$PI_\alpha(s_{n+1}) = [\angle\hat{v}_{n+1} - Q_{1-\alpha}(R^{CP}, D_{cal}), \angle\hat{v}_{n+1} + Q_{1-\alpha}(R^{CP}, D_{cal})]$$

This PI satisfies the **marginal coverage guarantee**:

$$\mathbb{P}\left(\angle v_{n+1} \in PI_\alpha(s_{n+1})\right) \geq 1 - \alpha$$

Conformalised quantile regression (CQR) [3]

CQR offers **adaptive** intervals for the desired coverage. Regression networks $\hat{Q}(s)$ are trained with a Pinball loss:

$$\mathcal{L}_\alpha(y, \hat{y}) := \begin{cases} \alpha(y - \hat{y}) & \text{if } y - \hat{y} > 0, \\ (1-\alpha)(\hat{y} - y) & \text{otherwise} \end{cases}$$

To produce lower and upper quantile predictions:

$$\{\hat{q}_{\frac{\alpha}{2}}(s), \hat{q}_{1-\frac{\alpha}{2}}(s)\}$$

Conformity scores quantify the error of the two regressed quantiles:

$$R_i^{CQR} = max\{\hat{q}_{\frac{\alpha}{2}}(s_i) - \angle v_i, \angle v_i - \hat{q}_{1-\frac{\alpha}{2}}(s_i)\}, i \in D_{cal}$$

The prediction interval with the same guarantees as CP, is constructed as:

$$PI_\alpha(s_{n+1}) = [\hat{q}_{\frac{\alpha}{2}}(s_{n+1}) - Q_{1-\alpha}(R^{CQR}, D_{cal}), \hat{q}_{1-\frac{\alpha}{2}}(s_{n+1}) + Q_{1-\alpha}(R^{CQR}, D_{cal})]$$

Multiple-testing corrections

The goal is to obtain a **joint predictive interval** that simultaneously provides guarantees for both phase and magnitude.

Multiple testing issues require **correction**: techniques from the literature on individual test level are used to restore valid coverage guarantees:

Bonferroni correction [4]:

$$\alpha_{corr} = \alpha / k$$

Sidak correction [5]:

$$\alpha_{corr} = 1 - (1-\alpha)^{1/k}$$

Max-Rank correction [6]: directly on the ranking of nonconformity scores across all variables.

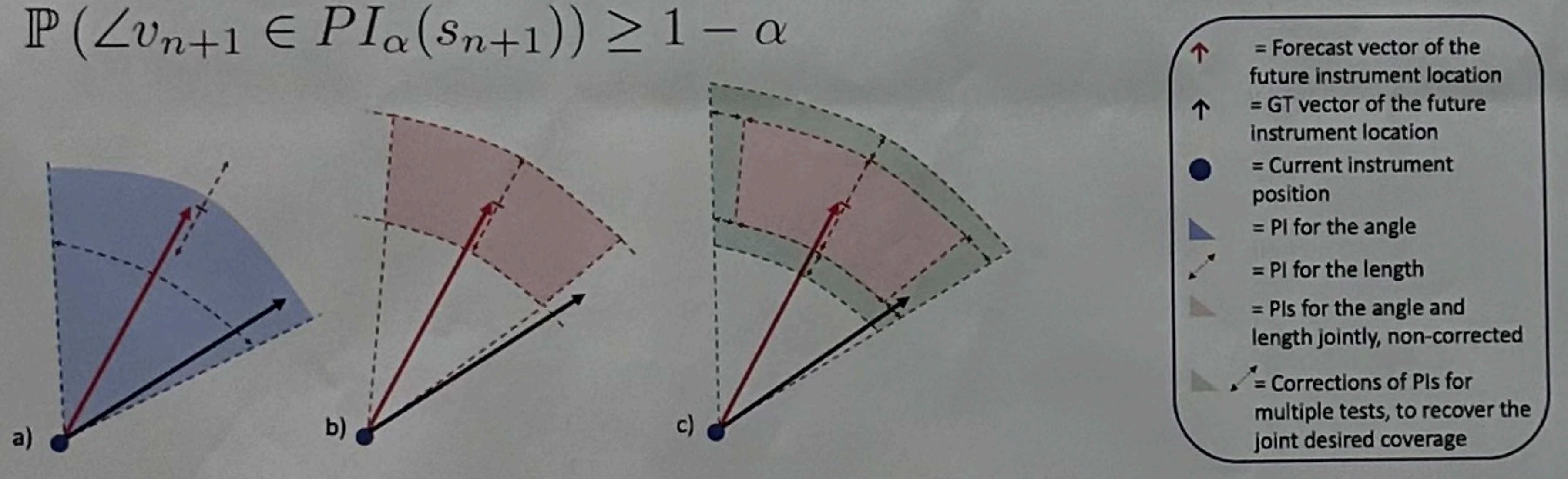

Fig. 1. Qualitative illustration of how Conformal Prediction (CP) for instrument trajectory forecasting.
a): CP applied independently to the angle and the length.
b): Joint intervals obtained by merging the independent ones without corrections, here failing to cover the angle.
c): Multiple-test corrections restore valid coverage for both quantities.

Experiments

Dataset: 144 pituitary surgery videos. 77 videos used to train a detector (~10000 labels, 15 anatomy classes, 1 instrument class). Detector used to create pseudolabels for frames of the remaining 67 videos. 57 used to train the forecasting network and CQR heads and 10 for testing. The test set has 6 patients for calibration and 4 for evaluation, randomly drawn 20 times; predictions are made independently, ensuring exchangeability for conformal prediction.

Forecasting network: transformer encoder with three fully connected layers outputs a 16D latent from 64 frames; predicts the next 8 frames with errors of 47° in angle and 0.2 in length, normalized to image size [1].

CQR network: 4 fully connected layers with ReLU activations, batch normalization, and dropout.

Method	Target Coverage = 60%		Target Coverage = 70%	
	Coverage (%)	PI Size (↓)	Coverage (%)	PI Size (↓)
CP angle	59.7 ± 3.7	78.5° ± 3.6°	69.8 ± 2.3	111.9° ± 3.4°
CQR angle	59.5 ± 3.5	**69.3° ± 3.1°**	69.4 ± 2.8	**103.5° ± 3.4°**
CP length	59.5 ± 12.1	0.25 ± 0.03	67.9 ± 12.9	0.31 ± 0.04
CQR length	60.4 ± 9.5	**0.19 ± 0.02**	69.2 ± 9.8	**0.24 ± 0.03**
CP joint, non corr.	31.0 ± 7.7	–	43.8 ± 9.0	–
CQR joint, non corr.	33.2 ± 5.4	–	45.8 ± 6.9	–
CP joint, Bonf. corr.	59.3 ± 8.4	169.8°, 0.41	67.8 ± 6.4	211.6°, 0.5
CQR joint, Bonf. corr.	61.5 ± 6.6	165.9°, 0.17	70.0 ± 6.0	212.7°, 0.21
CP joint, Sidak corr.	55.2 ± 8.7	151.8, 0.38	65.4 ± 7.7	200.0°, 0.46
CQR joint, Sidak corr.	57.2 ± 6.9	144.2°, 0.15	67.7 ± 6.2	200.5°, 0.19
CP joint, Max-Rank corr.	58.2 ± 7.9	105.7°, 0.83	68.3 ± 7.7	206.4°, 0.48
CQR joint, Max-Score corr.	59.0 ± 4.2	75.2°, 0.59	68.9 ± 4.0	204.8°, 0.20

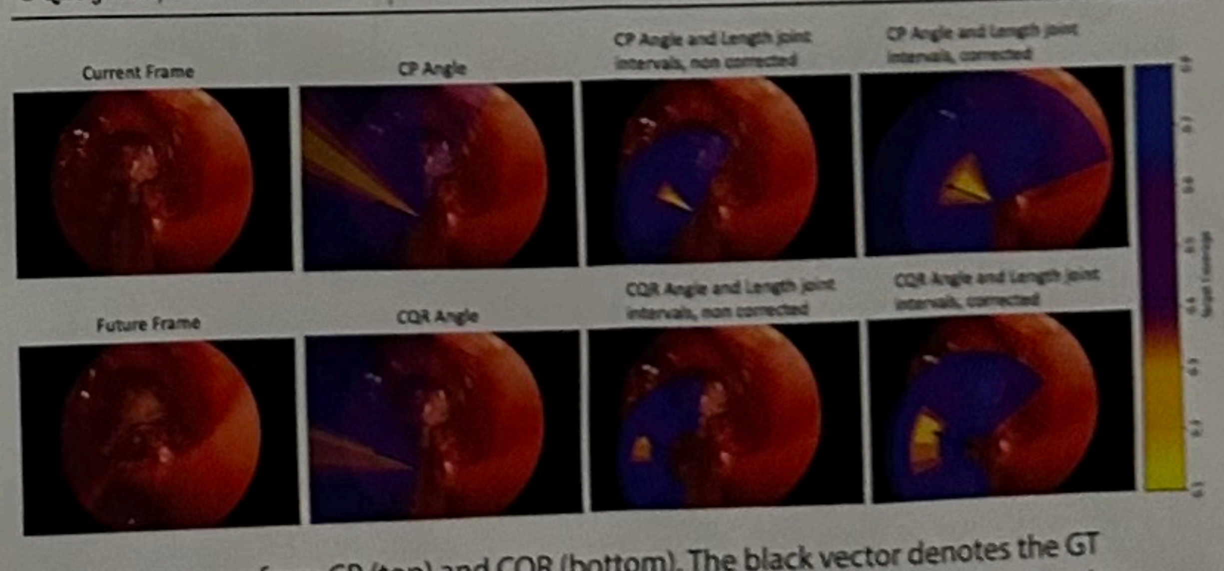

Fig. 2. Heatmaps from CP (top) and CQR (bottom). The black vector denotes the GT trajectory. Target coverage ranges from 10% (yellow) to 80% (blue). Left: Angle-only intervals—CQR yields sharper intervals and better coverage than CP. Center: Joint intervals without correction—coverage fails as expected. Right: Sidak-corrected joint intervals—recalibration restores validity, with CQR providing tighter bounds.

Results

- CQR generates **more precise PIs** compared to CP → CQR learns the data distribution for the specified intervals adaptively, while CP utilizes fixed thresholds, independent of the input.
- **Length** shows **higher variability**, especially for CP.
- For **joint intervals**, as expected, coverage drops significantly (by 25-30%) without multiple test corrections.
- **Applying corrections** successfully restores coverage, particularly for CQR, while naturally increasing PI sizes to ensure joint validity.

This study was financially supported by:
1. The LOOP Zürich – Medical Research Center, Zurich, Switzerland,
2. Personalized Health and Related Technologies (PHRT), project number 222, ETH domain and
3. Clinical Research Priority Program (CRPP) Grant on Artificial Intelligence in Oncological Imaging Network, University of Zürich,
4. The SNSF (Project IZKSZ3_218786).

[1] Sarwin, G., Carretta, A., Staartjes, V., Zoll, M., Mazzatenta, D., Regli, L., Serra, C., Konukoglu, E.: Anatomy might be all you need: Forecasting what to do during surgery (2025), https://arxiv.org/abs/2501.18011
[2] Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Found. Trends Mach. Learn. 16(4), 494–591 (Mar 2023)
[3] Romano, Y., Patterson, E., Candes, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
[4] Vovk, V., Wang, R.: Combining p-values via averaging. Political Methods: Quantitative Methods eJournal (2012), https://api.semanticscholar.org/CorpusID:88512881
[5] Šidák, Z.: Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62(318), 626–633 (1967). https://doi.org/10.1080/01621459.1967.10482935, https://doi.org/10.1080/01621459.1967.10482935
[6] Timans, A., Straehle, C.N., Sakmann, K., Naesseth, C.A., Nalisnick, E.: Max-rank: Efficient multiple testing for conformal prediction. arXiv e-prints arXiv:2311.10900 (Nov 2023). https://doi.org/10.48550/arXiv.2311.10900