Conformal Predictors for Efficient Video Text Spotting

Omar Ben Tanfous Sankha Subhra Mukherjee Neil M. Robertson

Terminal Industries Ltd.

BMVC 2025

The 36th British Machine Vision Conference 24th - 27th November 2025, Sheffield, UK

Introduction

Motivation

- Frame-level text recognition in videos is often inconsistent due to motion blur, occlusion, and illumination changes.
- Recognition reliability is strongly linked to model uncertainty.
- Conformal Prediction (CP) offers a lightweight, post-hoc, and statistically rigorous method to estimate uncertainty.

***Objective**: How can per-frame uncertainty estimates be leveraged to improve video text spotting performance?*

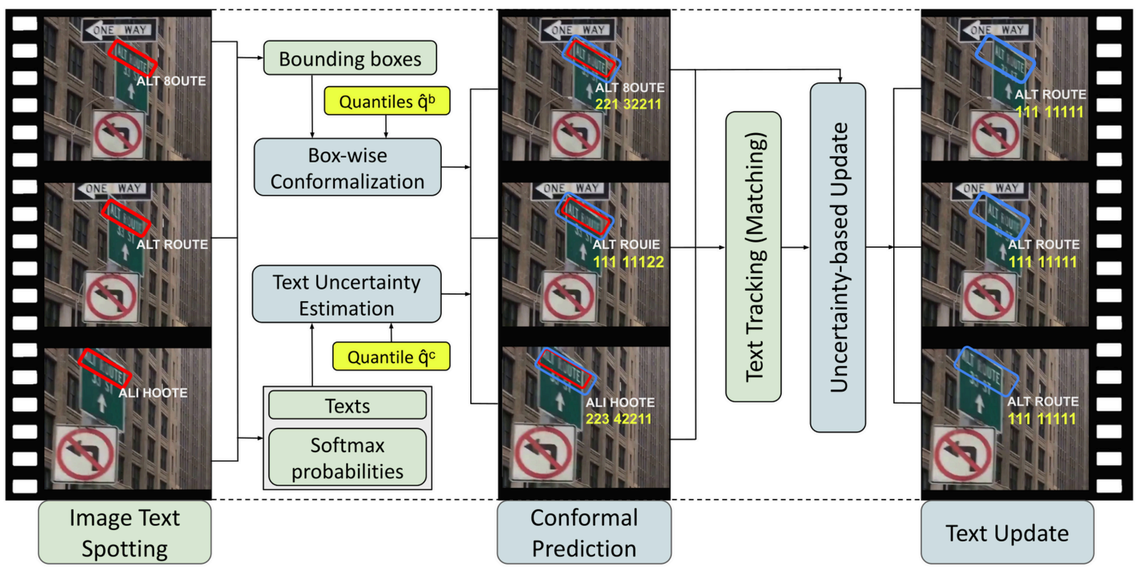
Contributions

- A novel uncertainty-aware method for selecting the most confident text recognition over a track of texts.
- A second CP module to refine the predicted text bounding boxes - improving text tracking performance.
- Improvements over SOTA on four benchmarks: ArtVideo, DSText, BOVText and ICDAR15.

Key Ideas

- **No retraining required:** CP calibrates any pretrained model using a small held-out set.
- **Statistically guaranteed:** CP Provides formal confidence intervals with predefined coverage.
- **Lightweight and efficient:** Adds minimal computational overhead at inference time.
- **Improves temporal stability:** Uses uncertainty estimates to correct unreliable frame-level predictions.

Proposed Approach



- A *pretrained* SOTA text spotter [1] is used to detect and recognize text in video frames.
- CVTS (Conformal Video Text Spotter) proceeds as follows:
  - A box-wise conformal predictor calibrates the detected bounding boxes.
  - The recognized texts and their softmax probabilities are passed to a text conformal predictor, which estimates per-character uncertainty scores (shown in yellow in the images).
  - After text matching via a *pretrained* tracker [1], an uncertainty-aware algorithm updates texts across tracks.

Text Conformal Prediction (TCP)

- **Concept**: Produces a *prediction set* that contains the true character with high probability.
- **Key Properties:**
  - **Coverage**: Guarantees the true character lies within the prediction set with prob ≥ $(1 - \alpha)$.
  - **Adaptiveness**: The clearer the text, the smaller the prediction set.
- **Calibration Step** (per model and dataset):
  1. Compute the score of the correct class for each calibration sample.
  2. Determine the $(1 - \alpha)$ % quantile, $\hat{q}$, of these scores.
- **Inference Step:**
  - For a new sample, include all characters whose scores exceed $\hat{q}$ to form the prediction set.
  - The uncertainty estimate is defined as the size of this prediction set.

Box-wise Conformal Prediction (BCP)

- **Calibration Step:**
  a. Let $B_k = \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$ and $\hat{B}_k = \{\hat{x}_{\min}, \hat{y}_{\min}, \hat{x}_{\max}, \hat{y}_{\max}\}$ be the ground-truth and predicted bounding boxes, respectively.
  b. Compute the nonconformity scores as $(\hat{x}_{\min} - x_{\min},\ \hat{y}_{\min} - y_{\min},\ \hat{x}_{\max} - x_{\max},\ \hat{y}_{\max} - y_{\max})$
  c. For each coordinate $c \in \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$ compute the conformal quantile using the Bonferroni correction.
- **Inference Step:** Define the conformalized bounding box for a test input as

$$\hat{C}_\alpha(X_{\text{test}}) = \{\hat{x}_{\min} - q^{x_{\min}},\ \hat{y}_{\min} - q^{y_{\min}},\ \hat{x}_{\max} + q^{x_{\max}},\ \hat{y}_{\max} + q^{y_{\max}}\}$$

Experimental Results

End-to-End text spotting

Method	ArTVideo			DSText		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
TransDETR [2]	49.3	2.8	69.7	26.43	-22.63	79.73
LOGO [3]	—	—	—	46.26	12.94	78.78
GoMatching [1]	82.3	75.7	**83.5**	46.27	23.24	80.41
CVTS (Ours)	**83.3**	**76.1**	**83.5**	**47.87**	**25.08**	**80.42**

Method	BOVText			ICDAR15		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
LOGO [3]	—	—	—	68.07	12.94	73.00
GloTSFormer [4]	—	—	—	74.20	62.50	78.20
GoMatching [1]	62.7	**52.9**	**87.2**	80.11	72.20	78.52
CVTS (Ours)	**65.0**	52.9	**87.2**	**80.92**	**72.54**	**78.54**

Text Tracking

Method	BOVText		ICDAR15		ArTVideo		DSText	
	IDF1	MOTA	IDF1	MOTA	IDF1	MOTA	IDF1	MOTA
GoMatching[1]	68.3	68.8	72.87	62.41	76.5	68.2	58.51	48.09
CVTS (BCP)	**68.4**	**69.0**	**73.42**	**62.69**	**76.7**	**68.3**	**58.69**	**48.15**

- CVTS achieves superior end-to-end text spotting and tracking performance over the GoMatching baseline across four datasets, **without model retraining.**

Ablation Study

Modules	ArTVideo				DSText		
	IDF1	MOTA	MOTP	FPS	IDF1	MOTA	MOTP
CVTS (overall)	**83.3**	**76.1**	83.5	0.63	**47.93**	**25.19**	80.39
CVTS (TSF)	83.2	75.9	83.5	0.65	46.95	24.55	**80.52**
CVTS (BCP)	82.3	75.8	83.5	0.66	46.31	22.57	80.43
CVTS (TCP)	82.4	76.0	83.5	0.66	46.32	23.35	80.41
GoMatching [1]	82.3	75.7	83.5	0.66	46.27	23.24	80.41

Modules	BOVText			ICDAR15		
	IDF1	MOTA	MOTP	IDF1	MOTA	MOTP
CVTS (overall)	**65.0**	52.9	87.2	78.01	70.34	78.73
CVTS (TSF)	64.8	52.9	87.2	77.06	69.67	**78.76**
CVTS (BCP)	63.0	52.9	87.2	**80.91**	**72.53**	78.54
CVTS (TCP)	62.8	52.9	87.2	80.23	72.13	78.54
GoMatching [1]	62.7	52.9	87.2	80.11	72.20	78.52

Conclusion

- CVTS provides a statistically grounded, lightweight, and modular enhancement for VTS systems through uncertainty quantification.
- Experiments show consistent improvements over SOTA, demonstrating the value of integrating conformal prediction into VTS and its potential for real-world applications.

[1] Haibin He, Maoyuan Ye, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Gomatching: A simple baseline for video text spotting via long and short term matching. In NeuRIPS, 2024
[2] Weijia Wu, Yuanqiang Cai, Chunhua Shen, Debing Zhang, Ying Fu, Hong Zhou, and Ping Luo. End-to-end video text spotting with transformer. In IJCV, 2024
[3] Hongen Liu, Xu Huang, Yan Sun, Kai Liu, Xiang Zhou, Yu Qiao, et al. Logo: Video text spotting with language collaboration and glyph perception model. In arXiv, 2024.
[4] Han Wang, Hui Wang, Shiliang Wang, Wenhao Yang, Shuai Liu, and Wenyu Liu. Glotsformer: Global video text spotting transformer. In arXiv, 2024.