Do We Train on Test Data? The Impact of Near-Duplicates on License Plate Recognition

Rayson Laroca[1], Valter Estevam[1,2],
Alceu S. Britto Jr.[3], Rodrigo Minetto[4], David Menotti[1]

[1]Federal University of Paraná, Curitiba, Brazil
[2]Federal Institute of Paraná, Irati, Brazil
[3]Pontifical Catholic University of Paraná, Curitiba, Brazil
[4]Federal University of Technology-Paraná, Curitiba, Brazil

June 2023

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Visão
Robótica
Imagem

Automatic License Plate Recognition (ALPR)

A usual *Automatic License Plate Recognition (ALPR)* system.

Automatic License Plate Recognition (ALPR)

A usual *Automatic License Plate Recognition (ALPR)* system.

ALPR has many practical applications:

- Toll collection;
- Vehicle access control in restricted areas;
- Traffic law enforcement.

Automatic License Plate Recognition (ALPR)

A usual *Automatic License Plate Recognition (ALPR)* system.

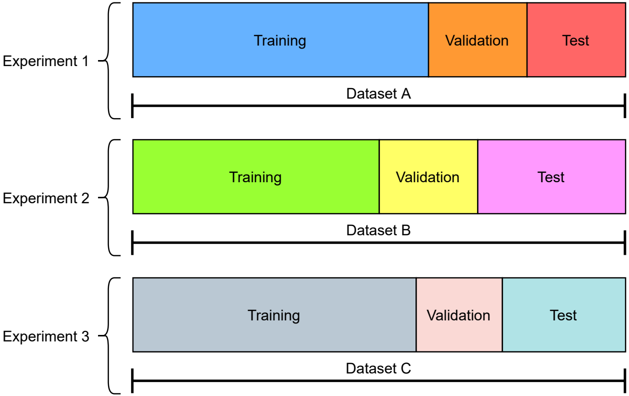
ALPR has many practical applications:

- Toll collection;
- Vehicle access control in restricted areas;
- Traffic law enforcement.

Current research has mostly focused on the **License Plate Recognition (LPR)** stage.

Problem Statement [1/2]

LPR methods are typically evaluated using images from public datasets, which are divided into ***disjoint*** *training and test sets using standard splits or following previous works (when there is no standard split).*



Problem Statement [2/2]

Although the images for training and testing belong to disjoint sets, the splits traditionally adopted in the literature were defined without the authors considering that ***the same license plate may appear in multiple images.***

Problem Statement [2/2]

*Although the images for training and testing belong to disjoint sets, the splits traditionally adopted in the literature were defined without the authors considering that **the same license plate may appear in multiple images.***

As a result, we found that there are many ***near-duplicates*** (i.e., different images of the same license plate) in the training and test sets of datasets widely explored in ALPR research.

Near-Duplicates – AOLP dataset

(a) Subset AC

(b) Subset LE

(c) Subset RP

(d) Subset AC

(e) Subset AC

(f) Subset RP

In the split protocols traditionally adopted in the literature,
<u>some of these images are in the training set and others are in the test set</u>.

Near-Duplicates – CCPD dataset

Subset Base | Subset Base | Subset Base | Subset Base

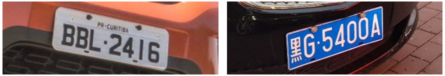
(a) Training set

Subset Challenge | Subset Challenge | Subset Weather | Subset Weather

(b) Test set

Many vehicles/license plates appear in both training and test images in the CCPD dataset.

- State-of-the-art ALPR approaches **rectify (unwarp)** the detected license plates before feeding them to the recognition model:



(a) detected license plates

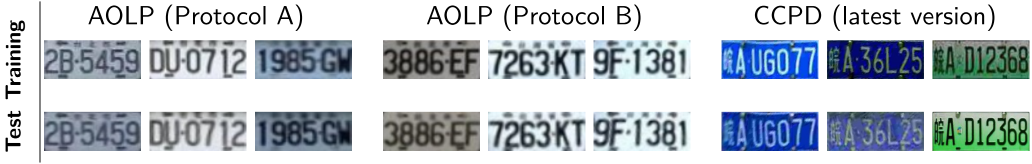
(b) rectified license plates

Near-Duplicates – LP Rectification

- State-of-the-art ALPR approaches **rectify (unwarp)** the detected license plates before feeding them to the recognition model:

(a) detected license plates (b) rectified license plates

Hence, the presence of duplicates in the training and test sets means that LPR models are, in many cases, **being trained and tested on essentially the same images**:



Examples of *near-duplicates* in the training and test sets of the AOLP and CCPD datasets.

Research Question

Research Question

To what extent have such near-duplicates impacted the evaluation of deep learning-based models applied to LPR?

Experimental Setup

We explored the two most popular datasets in the field:

- AOLP (https://github.com/avlab-cv/aolp);
- CCPD (https://github.com/detectrecog/ccpd).

Experimental Setup

We explored the two most popular datasets in the field:

- AOLP (`https://github.com/avlab-cv/aolp`);
- CCPD (`https://github.com/detectrecog/ccpd`).

We created ***fair splits*** for each dataset, where:

- There are no duplicates in the training and test sets;
- The key characteristics of the original partitions are preserved as much as possible.

Experimental Setup

We explored the two most popular datasets in the field:

- AOLP (https://github.com/avlab-cv/aolp);
- CCPD (https://github.com/detectrecog/ccpd).

We created ***fair splits*** for each dataset, where:

- There are no duplicates in the training and test sets;
- The key characteristics of the original partitions are preserved as much as possible.

We compared the performance of six well-known Optical Character Recognition (OCR) models applied to LPR under the traditional (adopted in previous works) and fair protocols:

OCR Model	Original Application
CNNG	License Plate Recognition
Holistic-CNN	License Plate Recognition
Multi-Task	License Plate Recognition

OCR Model	Original Application
STAR-Net	Scene Text Recognition
TRBA	Scene Text Recognition
ViTSTR-Base	Scene Text Recognition

Results – AOLP [1/2]

Results achieved under the AOLP-A[1,2] (adopted in previous works) and AOLP-Fair-A (ours) protocols.

Model	AOLP-A ↑	AOLP-A-Fair ↑	Gap ↓	Rel. Gap ↓
CNNG	98.88%	95.63%	3.25%	290.2%
Holistic-CNN	96.75%	93.11%	3.64%	**112.0%**
Multi-Task	97.33%	93.79%	3.54%	132.6%
STAR-Net	98.69%	95.83%	2.86%	218.3%
TRBA	**99.18%**	**96.94%**	2.24%	273.2%
ViTSTR-Base	98.74%	**96.94%**	**1.80%**	142.9%

The error rates were **more than twice as high** in the experiments conducted under the fair protocol, which has no duplicates.

[1] Protocol A: images divided into training and test sets with a 2:1 ratio.
[2] AOLP-A: 46.9% of the test images have duplicates in the training set.

Results – AOLP [2/2]

Results achieved under the AOLP-B[3,4] (adopted in previous works) and AOLP-Fair-B (ours) protocols.

Model	AOLP-B ↑	AOLP-B-Fair ↑	Gap ↓	Rel. Gap ↓
CNNG	**98.91%**	96.80%	2.11%	193.6%
Holistic-CNN	98.42%	96.30%	2.12%	134.2%
Multi-Task	98.42%	95.29%	3.13%	198.1%
STAR-Net	98.47%	96.46%	2.01%	131.4%
TRBA	98.75%	**97.47%**	**1.28%**	**102.4%**
ViTSTR-Base	98.75%	97.31%	1.44%	115.2%

The error rates were **more than twice as high** in the experiments conducted under the fair protocol, which has no duplicates.

[3] Protocol B: the AC and LE subsets are used for training, while the RP subset is used for testing.
[4] AOLP-B: 67.6% of the test images have duplicates in the training set.

Results – AOLP [2/2]

Results achieved under the AOLP-B[3,4] (adopted in previous works) and AOLP-Fair-B (ours) protocols.

Model	AOLP-B ↑	AOLP-B-Fair ↑	Gap ↓	Rel. Gap ↓
CNNG	**98.91%**	96.80%	2.11%	193.6%
Holistic-CNN	98.42%	96.30%	2.12%	134.2%
Multi-Task	98.42%	95.29%	3.13%	198.1%
STAR-Net	98.47%	96.46%	2.01%	131.4%
TRBA	98.75%	**97.47%**	**1.28%**	**102.4%**
ViTSTR-Base	98.75%	97.31%	1.44%	115.2%

The error rates were more than twice as high in the experiments conducted under the fair protocol, which has no duplicates.

The ranking of OCR models **changed** when they were trained and tested under fair splits.
Best model: **CNNG** → **TRBA**

[3]Protocol B: the AC and LE subsets are used for training, while the RP subset is used for testing.
[4]AOLP-B: 67.6% of the test images have duplicates in the training set.

Results – CCPD

Results achieved on the CCPD dataset under the standard[5] and CCPD-Fair protocols.

Model	CCPD ↑	CCPD-Fair ↑	Gap ↓	Rel. Gap ↓
CNNG	**88.24%**	**86.93%**	1.31%	11.1%
Holistic-CNN	77.01%	75.41%	1.60%	7.0%
Multi-Task	83.01%	81.84%	**1.17%**	**6.9%**
STAR-Net	78.53%	73.33%	5.20%	24.2%
TRBA	75.83%	71.48%	4.35%	18.0%
ViTSTR-Base	79.06%	76.37%	2.69%	12.9%

[5] CCPD's standard protocol: 19.1% of the test images have duplicates in the training set.

Results – CCPD

Results achieved on the CCPD dataset under the standard[5] and CCPD-Fair protocols.

Model	CCPD ↑	CCPD-Fair ↑	Gap ↓	Rel. Gap ↓
CNNG	**88.24%**	**86.93%**	1.31%	11.1%
Holistic-CNN	77.01%	75.41%	1.60%	7.0%
Multi-Task	83.01%	81.84%	**1.17%**	**6.9%**
STAR-Net	78.53%	73.33%	5.20%	24.2%
TRBA	75.83%	71.48%	4.35%	18.0%
ViTSTR-Base	79.06%	76.37%	2.69%	12.9%

The CCPD dataset has ≈ 157K test images:

- The lowest performance gap of **1.17%** translates to **1,800+** additional license plates being misrecognized under the fair split (vs. the standard one);
- The highest gap of **5.20%** represents a staggering number of **8,000+** more license plates being incorrectly recognized under the fair split.

[5] CCPD's standard protocol: 19.1% of the test images have duplicates in the training set.

Results – Overview

AOLP dataset

The high fraction of near-duplicates in the splits traditionally adopted in the literature ***may have hindered the development and acceptance of more efficient LPR models*** *that have strong generalization abilities but do not memorize duplicates as well as other models.*

Results – Overview

AOLP dataset

*The high fraction of near-duplicates in the splits traditionally adopted in the literature **may have hindered the development and acceptance of more efficient LPR models** that have strong generalization abilities but do not memorize duplicates as well as other models.*

CCPD dataset

*Our experiments provide a clearer picture of the true capabilities of LPR models compared to prior evaluations using the standard split, which has duplicates. Results revealed a decrease in the average recognition rate from **80.3%** to **77.6%** when the experiments were conducted under a fair split without duplicates.*

What about other datasets?

Other Datasets [1/3]

The <u>EnglishLP</u>, <u>Medialab LPR</u>, and <u>PKU</u> datasets lack an official split protocol.

These datasets are customarily divided into training and test sets **randomly** without the authors noticing that the same vehicle/license plate may appear in multiple images.

(a) <u>EnglishLP</u> (b) <u>Medialab LPR</u> (c) <u>PKU</u>

The presence of near-duplicates has also been overlooked in such setups.

Other Datasets [2/3]

The ReId dataset:

- 105,923 images in the training set;
- 76,412 images in the test set.

52,394 of the test images (**68.6%**) have near-duplicates in the training set.

(a) Training set

(b) Test set

Examples of near-duplicates in the ReId dataset.

Other Datasets [3/3]

There are duplicates **even across different datasets.**

(a) Images from the ChineseLP dataset

(b) Images from the CLPD dataset

Both datasets contain images scraped from the internet.

Conclusions

- Our experiments on the AOLP and CCPD datasets showed that near-duplicates have significantly biased the evaluation and development of deep learning-based models for LPR;

Conclusions

- Our experiments on the AOLP and CCPD datasets showed that near-duplicates have significantly biased the evaluation and development of deep learning-based models for LPR;
- As this problem has not yet received due attention from the community, the existence of near-duplicates has recurred in evaluations conducted on several other public datasets;

Conclusions

- Our experiments on the AOLP and CCPD datasets showed that near-duplicates have significantly biased the evaluation and development of deep learning-based models for LPR;
- As this problem has not yet received due attention from the community, the existence of near-duplicates has recurred in evaluations conducted on several other public datasets;
- We hope this work will encourage LPR researchers:
  - To train/assess their models using the fair splits[6] we created for the AOLP and CCPD datasets;
  - To beware of duplicates when performing experiments on other datasets.

[6] The fair splits as well as the list of near-duplicates we have found are publicly available for further research.

International Neural Network Society
IJCNN 2023
INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS

Thank you!

https://raysonlaroca.github.io/supp/lpr-train-on-test/

CAPES CNPq NVIDIA

PUCPR GRUPO MARISTA

UFPR UNIVERSIDADE FEDERAL DO PARANÁ