Vehicle-Rear: A New Dataset to Explore Feature Fusion for Vehicle Identification Using Convolutional Neural Networks

Icaro O. de Oliveira[1], Rayson Laroca[2], David Menotti[2], Keiko V. O. Fonseca[1], Rodrigo Minetto[1]

[1] Federal University of Technology - Paraná, Curitiba, Brazil
[2] Federal University of Paraná, Curitiba, Brazil

ABSTRACT

This work addresses the problem of vehicle identification through non-overlapping cameras. As our main contribution, we introduce a novel dataset for vehicle identification, called Vehicle-Rear, that contains more than three hours of high-resolution videos, with accurate information about the make, model, color and year of nearly 3,000 vehicles, in addition to the position and identification of their license plates. To explore our dataset we design a two-stream Convolutional Neural Network (CNN) *that simultaneously uses two of the most distinctive and persistent features available: the vehicle's appearance and its license plate. This is an attempt to tackle a major problem: false alarms caused by vehicles with similar designs or by very close license plate identifiers. In the first network stream, shape similarities are identified by a Siamese CNN that uses a pair of low-resolution vehicle patches recorded by two different cameras. In the second stream, we use a CNN for Optical Character Recognition* (OCR) *to extract textual information, confidence scores, and string similarities from a pair of high-resolution license plate patches. Then, features from both streams are merged by a sequence of fully connected layers for decision. In our experiments, we compared the two-stream network against several well-known CNN architectures using single or multiple vehicle features. The architectures, trained models, and dataset are publicly available at* `https://github.com/icarofua/vehicle-rear`.

1. INTRODUCTION

Identifying vehicles through non-overlapping cameras is an important task to assist surveillance activities such as travel time estimation, enforcement of speed limits, criminal investigations, and traffic flow. The vehicle identification problem can be formally defined as the process of assigning the same label to distinct instances of the same object as it moves over time in a network of non-overlapping cameras [1]. The remarkable progress of emerging technologies in producing low-cost cameras, capable of acquiring high-definition images, has made the infrastructure to tackle this problem become pervasive in many cities.

Although extensively investigated [2–7], this research problem is far from being solved since several challenges come from the high inter-class similarity, caused by vehicles of the same make, model and/or color that often look exactly the same, see Figure 1(a), vehicles with similar license plate identifiers, see Figure 1(b), and from the high intra-class dissimilarity, caused by abrupt illumination changes or camera viewpoints, that makes two instances of the same vehicle have differences, see Figure 1(c). In the remainder of this section, we detail our research problem and the main contributions of this work.

1.1. Research problem

The main issue of existing datasets for vehicle identification [8–10] is the fact that the authors intentionally redacted the license plate identifier in all images to respect privacy restrictions, and, as explained later, the knowledge extracted from this unique identifier is essential for solving certain difficult matching problems, e.g., the correct identification of distinct but visually similar vehicles, as shown in Figure 1(a). However, in some regions/countries, the license plates are linked/related to the vehicle and not to the respective drivers/owners; in other words, in such cases it is not possible to obtain any public information about the vehicle owner based on the license plate. One of the countries where this occurs is Brazil [11], where we collected images to create a novel dataset for vehicle identification that contains labeled license plate information.

In this work, we consider a road network topology structured as shown in Figure 2, where the rear license plate is legible in most cases – it is worth noting that in some countries/regions, e.g. several states in the United States, the li-

This is an author-prepared version. The published version is available at the *IEEE Xplore Digital Library* (DOI: 10.1109/ACCESS.2021.3097964).

Fig. 1. Examples of challenging scenarios for vehicle identification: (a) similar vehicles with different license plates; (b) similar license plate strings and distinct vehicles; and (c) same vehicle under different lighting conditions. The combination of attributes, e.g. vehicle appearance and textual information from the license plate region, can help to improve the recognition since two similar vehicles may have considerably different license plates and vice versa.

cense plate is attached only to the vehicle's rear. The images are taken from an elevated surveillance camera that records simultaneously multiple road lanes. Each vehicle of interest typically enters the field of view through the bottom part of the frame and leaves through the top side. As can be noted, not every vehicle seen in one camera appears in the other.

1.2. Contributions

This work has two main contributions for the vehicle identification problem:

- We introduce a novel dataset, called Vehicle-Rear, composed of high-resolution videos, that, to the best of our knowledge, is the first to consider the same camera view of most city systems used to enforce speed limits – i.e., rear view of the vehicles with their license plates legible in most cases; Vehicle-Rear is associated with accurate information about each vehicle: make, model, color and year, as well as the image coordinates of each license plate region and its corresponding ASCII string;
- We propose a novel two-stream CNN architecture that uses the most distinctive and persistent features for vehicle identification: coarse-resolution image patches, containing the *vehicle shape*, feed one stream, while high-resolution *license plate patches*, with string identifiers easily readable by humans (as present in the Vehicle-Rear dataset), feed the other stream. Such multi-resolution strategy helps to minimize the computational effort while it makes possible to capture the essential details for vehicle identification;

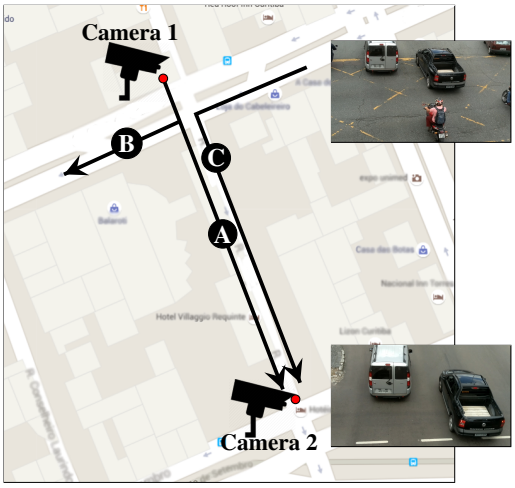


Fig. 2. Illustration of the experimental environment setup: a pair of low-cost Full-HD cameras, depicted by red dots, properly calibrated and time synchronized are monitoring two distinct traffic lights on the same street, 546 ft away. The road network is structured in such a way that some vehicles are monitored only by Camera 1, see route B; only by Camera 2, see route C; or by both cameras, see route A.

We believe that the creation of a publicly available dataset containing images captured in real-world scenarios and labeled information about both the vehicle and its license plate represents a step forward in designing different approaches to vehicle identification, since state-of-the-art algorithms for vehicle identification take advantage of only one of these attributes [2, 12, 13]. We hope that our dataset and deep architecture can also be useful for other machine learning problems such as vehicle model identification, time-travel estimation, among others.

The remainder of this paper is organized as follows. In Section 2, we review the literature on vehicle identification. The proposed Vehicle-Rear dataset is described in Section 3. The two-stream architecture is described in Section 4, and the experimental evaluation is reported in Section 5. In Section 6 we discuss some alternative architectures, and in Section 7 we state the conclusions.

2. RELATED WORK

Vehicle identification is an active field of research with many algorithms and an extensive bibliography. As observed by Tian *et al.* [14], this problem is still an open issue for future developments of networked video surveillance systems, in which the road camera infrastructure is used to extract vehicle trajectories for behavior analysis and pattern discovery. Traditionally, algorithms proposed for this task were based on the comparison of electromagnetic signatures captured from a pair of inductive or magnetic sensors [15, 16]. This class of systems can benefit from the existing infrastructure to capture vehicle signature profiles from inductive-loop detectors [3], weight-in-motion devices [17], and microloop sensors [4]. However, as stated by Ndoye *et al.* [4], such signature-based algorithms are complex and depend on complicated data models or extensive calibrations.

Video-based algorithms have been proven essential for vehicle identification. As describe in the surveys of Deng *et al.* [13], Wang *et al.* [18], and Khan & Ullah [19], handcrafted image descriptors [20–23] were the first attempt to solve this problem, e.g. Zapletal and Herout [24] and Chen *et al.* [25] used HOG descriptors, Cabrera *et al.* [20] used HAAR descriptors, while Cormier *et al.* [26] used Local Binary Patterns (LBP) [22] – all these works also combined other handcrafted descriptors. Zhang *et al.* [27] used Scale-Invariant Feature Transform (SIFT) [21] to distinguish between subordinate categories with similar visual appearance, caused by a huge number of car design and models with similar appearance. In particular, SIFT was widely explored to extract distinctive key points from the vehicle for feature correspondence [28].

The use of Siamese-based architectures for the specific problem of vehicle identification is common. Tang *et al.* [7] proposed to fuse deep and handcrafted features using a *Triplet Siamese Network* [29] – a network that attempts to minimize the distance between an anchor and a positive sample and to maximize the distance between the same anchor and a negative sample. Yan *et al.* [5] proposed a novel Triplet Loss Function, which uses both the intra-class variance and the inter-class similarity in vehicle models, but using only vehicle shape features. Liu *et al.* [6] developed a coarse-to-fine algorithm for vehicle identification that filters out potential matchings with handcrafted and deep features based on color and shape, and then used a Siamese network for the license plate regions.

The idea of multi-stream Convolutional Neural Networks (CNNs) has also been considered by many authors to tackle different identification problems. Ye *et al.* [30] proposed a two-stream architecture that uses static video frames and optical flow features for video classification. Similarly, Chung *et al.* [31] proposed a two-stream Siamese architecture that is also based on spatial and temporal information extracted from RGB frames and optical flow features but for person re-identification. Zagoruyko *et al.* [32] described distinct Siamese architectures to compare image patches. In particular, they developed a two-stream architecture that explores multi-resolution information by using the central part of an image patch and the surrounding part of the same patch. Specifically for vehicle identification, Oliveira *et al.* [33] proposed a two-stream network fed by small patches from the vehicle shape and the license plate region, and Guo *et al.* [2] proposed a three-stream network where one stream extracts global features from the vehicle shape and the other two streams learn to locate vehicle features, such as windscreen and car-head parts.

Architectures designed to recognize patterns in temporal sequences, such as Long Short-Term Memory (LSTM) [34], ensembles [35], and spatio-temporal (3D) convolutions [36], may also have a major impact on vehicle identification [37, 38]. As an example, Shen *et al.* [37] noted that if a vehicle is seen by cameras 1 and 3 then it should also appear in camera 2; thus, if no candidate is observed by camera 2, any subsequent match should have very low confidence. The authors employed a Siamese network fed with the vehicle's shape and temporal metadata to model this scenario, and an LSTM to evaluate the visual and spatio-temporal differences of neighboring states along with path proposals. The dataset used in their experiments, VeRi-776 [39], was acquired by 20 cameras. Zhou *et al.* [38] exploited an adversarial bi-directional LSTM network to create a vehicle representation from one camera view that would allow modeling transformations across continuous view variations. Generative Adversarial Networks (GANs) were also explored to generate samples to facilitate the vehicle identification task [40].

License plate recognition, as we used in this work, is one of the key attributes for successful vehicle identification and deep networks have achieved many advances in this field. Li *et al.* [41] first extracted sequential features from the license plate patch using a CNN in a sliding window manner. Then, Bidirectional Recurrent Neural Networks (BRNNs) with LSTM were applied to label the sequential features, while Connectionist Temporal Classification (CTC) was employed for sequence decoding. The results showed that their method attained better recognition rates than the baselines. Nevertheless, Dong *et al.* [42] claimed that such a method is very fragile to distortions caused by viewpoint change and therefore is not suitable for license plate recognition in the wild. Thus, a license plate rectification step is employed first in their approach, which leverages parallel Spatial Transform Networks (STNs) with shared-weight classifiers. Recently, Selmi et al. [43] trained a Mask-RCNN [44] to predict 37 positive classes (0-9, A-Z, and one Arabic word). Despite the fact that promising results were reported in their experiments, the chosen model (with an input size of 530×300 pixels) is much more computationally expensive than those used in other works (e.g., [45–47]) for license plate recogni-

tion, which makes it difficult (or even impossible) for it to be employed in some real-world applications – especially those where multiple vehicles can coexist on the scene.

Silva & Jung [48] proposed a YOLO-based model to simultaneously detect and recognize all characters within a cropped license plate. While impressive frames per second (FPS) rates were reported in their experiments, less than 65% of the license plates on the test set were correctly recognized since the character classes in the training set used by them were highly unbalanced. Accordingly, Laroca *et al.* [47, 49] and Silva & Jung [46, 50] retrained that model, called CR-NET, with enlarged training sets composed of real images and many other artificially generated. In all these works, the retrained networks became much more robust for the detection and classification of real characters.

As final remarks, although some previous studies have shown the importance of feature fusion for vehicle identification (e.g., [2, 6, 7]), none of them explored a camera infrastructure specifically designed for traffic law enforcement as those available in many cities, where the vehicle's rear license plate is legible in most cases. Considering such camera views, it is possible to develop a novel and robust two-stream architecture that combines two decisive features for vehicle identification: (i) shape features from the vehicle rear-end and (ii) textual features from the license plate region.

3. THE VEHICLE-REAR DATASET

As detailed in Table 1, the Vehicle-Rear dataset consists of 10 videos – five from Camera 1 and five from Camera 2 (20 minutes long each video) – captured by a low-cost 5-megapixel CMOS image sensor, time-synchronized, with a resolution of $1{,}920 \times 1{,}080$ pixels at 30.15 frames per second.

Table 1. Vehicle-Rear dataset: detailed information about the number of vehicles, with and without a legible license plate, recorded by Cameras 1 and 2; and the number of true matchings between Camera 1 and 2.

	Camera 1		Camera 2		
Set	# Vehicles	# Plates	# Vehicles	# Plates	# Matchings
01	385	342	277	245	199
02	350	301	244	225	179
03	340	312	273	252	203
04	280	258	230	196	147
05	345	299	242	205	165
Total	1,700	1,512	1,266	1,123	893

We chose a busy avenue of the city, with traffic of different types of vehicles, and different periods of the day to record the videos so that each set has very specific lighting conditions (see Figure 3). Note that temporal information can also be explored in the Vehicle-Rear dataset since for each vehicle we have between [5-25] frame occurrences per camera (depending on the vehicle speed); thus, redundant information could be used to further improve the vehicle identification.

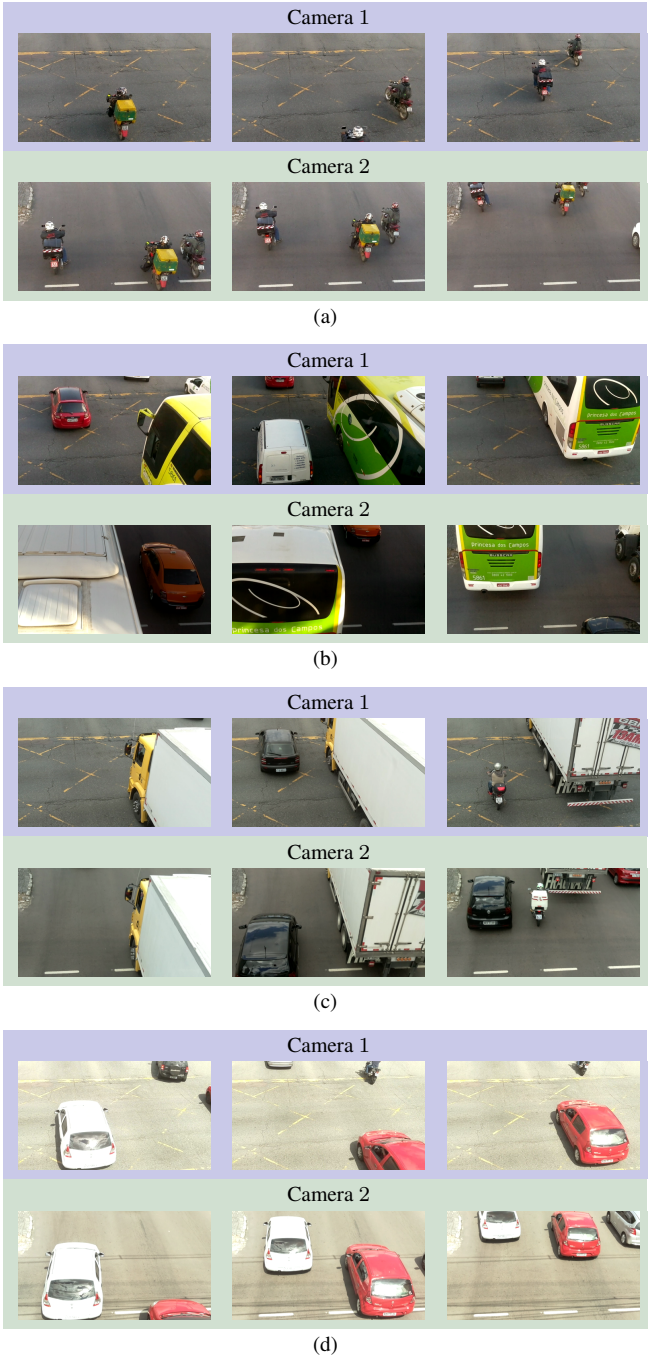


Fig. 3. Image sequences from the proposed Vehicle-Rear dataset. The temporal sequences show examples of (a) motorcycles; (b) cars and buses; (c) trucks; (a) and (c) in normal weather conditions; (b) dark frames caused by the motion of large vehicles; and (d) severe lighting conditions.

For each video, we provide a ground truth XML file in

which each entry, corresponding to a distinct vehicle, has an axis-aligned rectangular box of the first license plate occurrence, the corresponding identifier in ASCII code, the frame position, as well as the vehicle's make, model, color and year, which were recovered from the database of the National Traffic Department of Brazil (DENATRAN). We remark that the DENATRAN database is publicly available, that is, there is no restriction on access to such information. As far as we are aware, the proposed dataset is the first public dataset for vehicle identification to provide information on the appearance of the vehicles and also on their license plates.

Figure 4 and Figure 5 show the diversity of our dataset in relation to vehicle automakers and colors, respectively. As can be seen, there is a considerable imbalance – as is likely the case for every dataset – since vehicles of certain brands and colors sell more than others. Nevertheless, according to our experiments, such imbalance did not significantly affect the results obtained by the evaluated models.

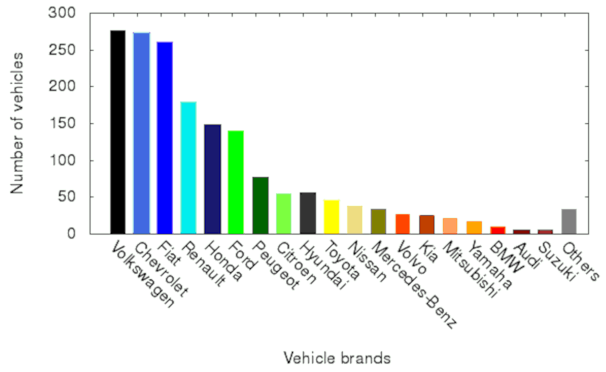


Fig. 4. Vehicle histogram by brand in the Vehicle-Rear dataset.

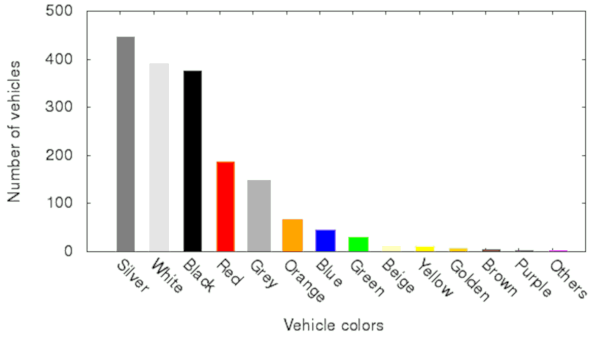


Fig. 5. Vehicle histogram by color in the Vehicle-Rear dataset.

Finally, it is worth noting that the licenses plates of vehicles in Brazil, where the images were collected, are linked/related to the vehicle and no public information is available about the vehicle drivers/owners; hence, a license plate remains the same after a change in vehicle ownership [51]. Considering the height and distance of the cameras, as well as the fact that they record the rear view of vehicles, identifying the driver/owner from the captured frames in our dataset is not possible, to the best of our knowledge. Finally, as detailed in Section 6, this study was officially authorized to collect and explore open data such as the Vehicle-Rear dataset.

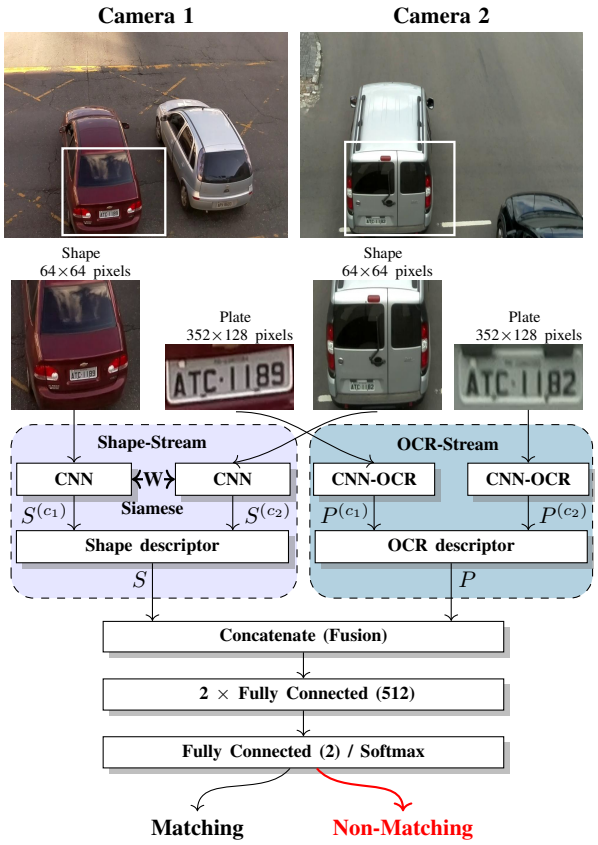


Fig. 6. Inference scheme of the proposed two-stream Siamese neural network for vehicle matching.

4. VEHICLE IDENTIFICATION ARCHITECTURE

In order to explore the attributes of the proposed dataset, we design a two-stream neural network, as shown in Figure 6, that uses the most distinctive and persistent features available for vehicle identification: coarse-resolution image patches, containing the vehicle shape, feed one stream, while high-resolution license plate patches, easily readable by humans, feed the other stream. Such a multi-resolution strategy helps to minimize the computational effort while making it possible to capture the necessary details for the recognition. We developed a text descriptor, i.e., Optical Character Recognition (OCR), which is combined with the shape descriptor through a sequence of fully connected layers for decision. Further details on these key steps are presented in the remainder of this section.

4.1. Preliminaries

For our problem, let $S^{(c_1)} = \langle s_1^{(c_1)}, s_2^{(c_1)}, \ldots, s_m^{(c_1)} \rangle$ and $S^{(c_2)} = \langle s_1^{(c_2)}, s_2^{(c_2)}, \ldots, s_m^{(c_2)} \rangle$ be two m-dimensional vectors representing the deep features extracted with a Siamese network from shape patches recorded by cameras 1 and 2, respectively. Also, let $\mathcal{C}_n = \{c_0, c_1, \ldots, c_{n-1}\}$ be a non-empty alphabet consisting of n unique elements. Then, let $f : \mathcal{C} \rightarrow \mathcal{N}$ be a one-to-one function (bijection) that maps elements of the alphabet $\mathcal{C}$ to unique real numbers $\mathcal{N}$ according to Equation (1)

$$f(c_i) = \frac{i}{n-1} \tag{1}$$

where i is the element position in the alphabet, such that $0 \leq i < n$, and n denotes the set size. The alphabet used to build the license plate identifiers is composed by 26 letters and 10 digits, thus, $\mathcal{C}_{36} = \{A, \ldots, Z, 0, \ldots, 9\}$. This mapping is shown in Figure 7. Note that the lexicography order is used to establish the mapping function f. As a consequence, no special arrangement among similar characters, such as D, O, Q and 0, was done.

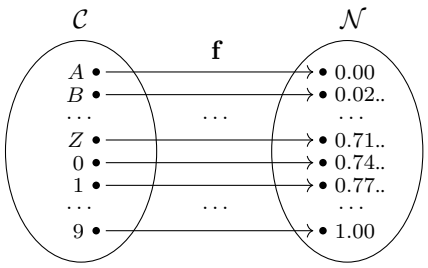


Fig. 7. A bijective function (f) to map license plate characters (domain $\mathcal{C}$) to real numbers (range $\mathcal{N}$).

4.2. Shape similarities

The shape similarities are identified by a Siamese network, which hereinafter is referred to as Shape-Stream. This particular class of neural architecture was introduced by Bromley *et al.* [52] and consists of two identical networks that share the same weights. We choose a Siamese network to compare shape similarities because it is an effective and simple architecture to solve image matching problems.

The shape descriptor is defined as a new vector according to Equation (2)

$$S = S^{(c_1)} - S^{(c_2)} = (s_1, s_2, \ldots, s_m) \tag{2}$$

where each component s_i is given by an L_1 (Manhattan) distance, that is, $s_i = |s_i^{(c_1)} - s_i^{(c_2)}|$ for cameras c_1 and c_2. The twin networks guarantee that two similar image patches will not be mapped to very different locations in the feature space since they compute the same function and their weights are tied; therefore, it is expected that the vector components are small for two instances of the same vehicle and large otherwise. The deep features were extracted with a low complex VGG-based CNN [33], called Small-VGG, formed by a reduced number of convolutional layers in order to save computational effort, as shown in Table 2.

Table 2. The CNN architecture used by the Siamese network in the Shape-Stream.

#	Layer	Filters	Size	Input	Output
0	conv	64	$3 \times 3/1$	$64 \times 64 \times 3$	$64 \times 64 \times 64$
1	max		$2 \times 2/2$	$64 \times 64 \times 64$	$32 \times 32 \times 64$
2	conv	128	$3 \times 3/1$	$32 \times 32 \times 64$	$32 \times 32 \times 128$
3	max		$2 \times 2/2$	$32 \times 32 \times 128$	$16 \times 16 \times 128$
4	conv	128	$3 \times 3/1$	$16 \times 16 \times 128$	$16 \times 16 \times 128$
5	max		$2 \times 2/2$	$16 \times 16 \times 128$	$8 \times 8 \times 128$
6	conv	256	$3 \times 3/1$	$8 \times 8 \times 128$	$8 \times 8 \times 256$
7	max		$2 \times 2/2$	$8 \times 8 \times 256$	$4 \times 4 \times 256$
8	conv	512	$3 \times 3/1$	$4 \times 4 \times 256$	$4 \times 4 \times 512$
9	max		$2 \times 2/2$	$4 \times 4 \times 512$	$2 \times 2 \times 512$

4.3. License plate similarities

The plate similarities are then identified by using textual information extracted from fine-resolution license plate image patches (OCR-Stream). We observed through a series of experiments, as detailed in Section 6, that the same approach we used for shape was not very accurate to distinguish between very similar license plate regions. The textual content, on the other hand, makes it possible to explore the syntax that defines the license plate layouts and, thus, to improve the recognition. Inspired by the tremendous advances in machine learning achieved by CNNs, we used a state-of-the-art architecture (CR-NET) [48] for OCR that has proven to be robust to recognize license plates from various countries [46, 47], but here it was fine-tuned for the Brazilian license plate layout (i.e., three letters followed by four digits).

The OCR architecture, as described by Silva & Jung [48] and later improved by Laroca *et al.* [47], consists of the first eleven layers of YOLO [53] and four other convolutional layers added to improve non-linearity, as shown in Table 3. The network was trained to predict 35 character classes (0-9, A-Z, where the letter 'O' is detected/recognized jointly with the digit '0') – however, for the sake of simplicity of definitions, we will assume a complete alphabet with 36 characters in the remainder of this section. Furthermore, some swaps of digits and letters, which are often misidentified, were used to improve the recognition: [1 ⇒ I; 2 ⇒ Z; 4 ⇒ A; 5 ⇒ S; 6 ⇒ G; 7 ⇒ Z; 8 ⇒ B] and [A ⇒ 4; B ⇒ 8; D ⇒ 0; G ⇒ 6; I ⇒ 1; J ⇒ 1; Q ⇒ 0; S ⇒ 5; Z ⇒ 7].

We created an OCR descriptor by combining the textual content extracted from both license plates. For that purpose, we propose a scheme to map characters to real numbers as follows.

Table 3. The CNN-OCR architecture for license plate recognition as proposed by Silva & Jung [48] and improved by Laroca *et al.* [47].

#	Layer	Filters	Size	Input	Output
0	conv	32	$3 \times 3/1$	$352 \times 128 \times 3$	$352 \times 128 \times 32$
1	max		$2 \times 2/2$	$352 \times 128 \times 32$	$176 \times 64 \times 32$
2	conv	64	$3 \times 3/1$	$176 \times 64 \times 32$	$176 \times 64 \times 64$
3	max		$2 \times 2/2$	$176 \times 64 \times 64$	$88 \times 32 \times 64$
4	conv	128	$3 \times 3/1$	$88 \times 32 \times 64$	$88 \times 32 \times 128$
5	conv	64	$1 \times 1/1$	$88 \times 32 \times 128$	$88 \times 32 \times 64$
6	conv	128	$3 \times 3/1$	$88 \times 32 \times 64$	$88 \times 32 \times 128$
7	max		$2 \times 2/2$	$88 \times 32 \times 128$	$44 \times 16 \times 128$
8	conv	256	$3 \times 3/1$	$44 \times 16 \times 128$	$44 \times 16 \times 256$
9	conv	128	$1 \times 1/1$	$44 \times 16 \times 256$	$44 \times 16 \times 128$
10	conv	256	$3 \times 3/1$	$44 \times 16 \times 128$	$44 \times 16 \times 256$
11	conv	512	$3 \times 3/1$	$44 \times 16 \times 256$	$44 \times 16 \times 512$
12	conv	256	$1 \times 1/1$	$44 \times 16 \times 512$	$44 \times 16 \times 256$
13	conv	512	$3 \times 3/1$	$44 \times 16 \times 256$	$44 \times 16 \times 512$
14	conv	200	$1 \times 1/1$	$44 \times 16 \times 512$	$44 \times 16 \times 200$
15	detection				

The OCR descriptor is composed by the mapped characters, alternated with its classification scores so as to aggregate knowledge about the confidence of each prediction. Moreover, the descriptor also contains the similarities between both license plate identifiers. Namely, for two aligned strings, we compute a character-by-character distance using a step function, as shown in Equation (3)

$$d(c_i, c_j) = \begin{cases} 0 & \textbf{if } f(c_i) - f(c_j) = 0 \\ 1 & \textbf{otherwise} \end{cases} \quad (3)$$

where c_i and c_j are two characters that belong to set $\mathcal{C}_{36}$, detailed in Section 4.1, and f is the mapping function of Equation (1). Observe that two characters are equal or distinct for the step function, i.e., the notion of proximity does not exist. For example, although letter A is mapped to value 0.00, B to 0.02... and Z to 0.71..., the distance between A and B is the same distance between A and Z (1 for both cases). However, the confidence scores, associated with each character, may help the network to decide the weight of such distances.

The OCR descriptor is illustrated in Figure 8.

5. EXPERIMENTS

In this section, we describe an extensive set of experiments comparing several CNN/OCR architectures.

For training, evaluation and testing it is necessary to pairwise image patches. If we have n_1 vehicles passing through Camera 1 and n_2 vehicles passing through Camera 2, then we can create $n_1 \times n_2$ image pairs, where n_1 is the maximum number of matching pairs and $(n_1 \times n_2) - n_1$ is the approximate number of non-matching pairs. Note that we have highly imbalanced sets from non-matching pairs ($(n_1 \times n_2) - n_1$) compared to matching pairs. Therefore, in order to have more matching pairs, we used the MOSSE algorithm [54] to track

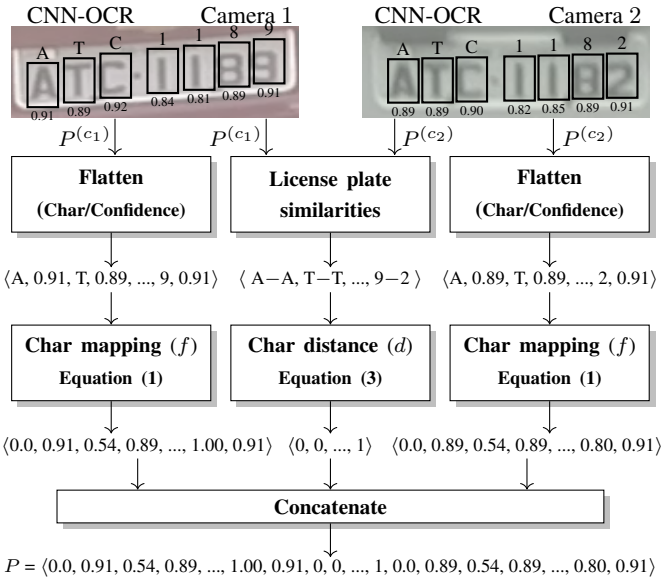


Fig. 8. The OCR-descriptor scheme: the ASCII characters and the corresponding classification confidences are extracted from both license plate regions with the CNN-OCR architecture; then, they are combined to create a text descriptor.

a vehicle for m consecutive frames, and only for the matching pairs we used all its m frame occurrences to create new matching pairs. An advantage of using such a technique is that the object appearance in a sequence of consecutive frames usually has small image variations – due to the vehicle motion, scene illumination changes, image noise, etc. – that produces distinct pairs. This process is depicted in Figure 9. Using the strategy described above, we generated 5 sets of matching/non-matching pairs, as listed in Table 4.

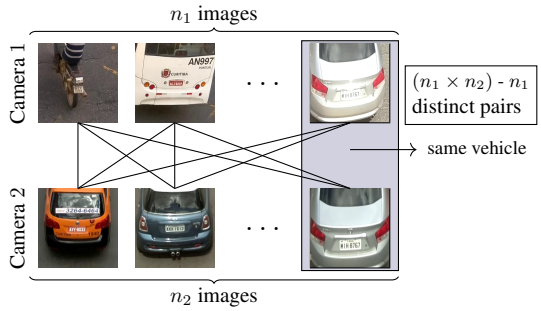
Table 4. Number of matching/non-matching image pairs generated within each set.

Set	# Non-matching pairs	# Matching pairs
01	83,250	19,560
02	66,722	17,370
03	76,681	19,520
04	49,650	14,177
05	60,313	16,030
Total	336,616	86,657

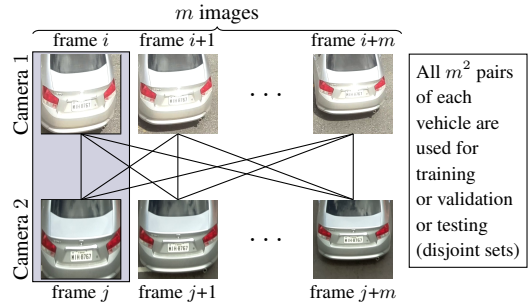
5.1. Experimental Setup

The CNN-OCR model was trained using the Darknet framework[1], while the other models were trained using Keras[2].

[1] https://github.com/AlexeyAB/darknet/
[2] https://keras.io/



(a) Non-matching pairs



(b) Matching pairs

Fig. 9. Generation of image pairs for training, validation and testing. The same procedure is used for the license plates.

We performed our experiments on an Intel i7-8700K 3.7GHz CPU, 64GB RAM, with an NVIDIA Titan Xp GPU.

Our experiments were performed using Ubuntu 14.04, Python 3.7, OpenCV 3.4.1, Keras 2.3.1 and TensorFlow 1.15.2. All networks were trained using the Adam optimizer with a learning rate of 10^{-4}, batch size = 128, and epochs = 10. The architectures and trained models are publicly available at `https://github.com/icarofua/vehicle-rear`.

We remark that we evaluated different input sizes, as well as number of filters in the convolutional layers, for both vehicle and license plate images, but better results were not achieved. In this sense, it is also worth noting that both models chosen by us (Small-VGG and CNN-OCR) are relatively lightweight compared to others commonly used in the literature, despite the fact that they have reached impressive results [33, 55, 56]. More specifically, Small-VGG has 1.7M parameters and requires 0.317 GFLOPs, while CNN-OCR has 3.3M parameters and requires 5.899 GFLOPs.

5.2. Evaluation Metrics

The quantitative criteria we used to assess the performance of each model are precision P and recall R, as defined in Equation (4)

$$P = \frac{tp}{tp + fp} \qquad R = \frac{tp}{tp + fn} \tag{4}$$

where tp denotes the number of true matchings between Cameras 1 and 2, fp is the number of false matchings, and fn the number of true matchings missed by the respective model. For ranking purposes, we also consider the F-score, which is the harmonic mean of precision and recall, as shown in Equation (5)

$$F = \frac{2}{1/P + 1/R} \tag{5}$$

We chose F-score over accuracy since the number of non-matching pairs is much larger than matching pairs and, thus, for highly imbalanced data, we can have a very low true matching rate but a very high accuracy.

5.3. Data Augmentation

For data augmentation in vehicle shape images, we used random crops between 0 and 8 pixels, scale between 0.8 and 1.2, and shear between -8 and 8. In license plate images, we used scale between 0.8 and 1.2, translation between -10% and 10%, rotation between -5 and 5, and shear between -16 and 16 (note that these parameter values were defined based on experiments performed in the validation set). We used Albumentations [57], which is a well-known Python library for image augmentation, to apply these transformations.

5.4. Ablation Study

As shown in Table 5, we evaluated the use of several CNNs architectures for the identification task. In all experiments, we used 5 rounds of cross-validation using the 5 sets listed in Table 4. For each round, we used 2 sets for training, 1 for validation, and 2 for testing. We started with sets 01 and 02 for training, 03 for validation, and 04 and 05 for testing; then we used 02 and 03 for training, 04 for validation, and 05 and 01 for testing; then 03 and 04 for training, and so on. Therefore, $\bar{P}, \bar{R}$ and $\bar{F}$ are the average values of precision, recall and F-score for these 5 rounds.

Table 5. Vehicle identification performance based on shape: for these experiments, we evaluated several CNN architectures, exclusively based on shape features, in the Siamese Shape-Stream using different image sizes.

One-Stream (Shape-only)	$\bar{P}$	$\bar{R}$	$\bar{F}$
Resnet8 (128 × 128 px)	54.01%	89.89%	66.86%
Lenet5 (128 × 128 px)	89.74%	71.09%	78.61%
Resnet6 (128 × 128 px)	73.70%	86.59%	78.74%
MC-CNN (64 × 64 px)	83.00%	82.42%	82.63%
GoogleNet (112 × 112 px)	79.51%	91.30%	84.38%
Matchnet (128 × 128 px)	89.05%	92.86%	90.75%
Small-VGG (64 × 64 px)	**90.43%**	**92.54%**	**91.35%**

For license plate recognition, we compared the performance of the CNN-OCR architecture against two commer-

cial systems: *Sighthound* [58] and *OpenALPR*[3] [59]. These systems were chosen since they are commonly used as baselines in the license plate recognition literature [46, 47, 60] and also because they are robust for the detection and recognition of various license plate layouts [58, 59]. It should be noted that, due to commercial reasons, little information is given about the network models used in such systems. As can be seen in Table 6, the CNN-OCR architecture achieved an F-score of 94.1% if we consider a perfect match (correct matching of all characters), however, if we consider partial OCR readings, then we can have an F-score of 97.7% by allowing one misreading and 98.6% for two misreadings. In any scenario, CNN-OCR considerably outperformed the Sighthound and OpenALPR commercial systems.

Table 6. Vehicle identification performance based on OCR: comparison of the results achieved by the CNN-OCR architecture with those obtained by two well-known commercial systems. For this evaluation, we consider as true matchings the cases where exactly the same license plate characters were predicted in cameras 1 and 2.

	Partial matching 2 errors			Partial matching 1 error			Perfect matching		
OCR	$\bar{P}$	$\bar{R}$	$\bar{F}$	$\bar{P}$	$\bar{R}$	$\bar{F}$	$\bar{P}$	$\bar{R}$	$\bar{F}$
Sighthound	99.9%	84.5%	91.5%	100%	81.5%	90.0%	100%	66.0%	79.3%
OpenALPR	99.9%	83.2%	90.7%	100%	80.4%	89.1%	100%	70.0%	82.2%
CNN-OCR*	99.8%	92.0%	95.7%	100%	86.7%	92.8%	100%	74.1%	84.9%
CNN-OCR	**99.9%**	**97.3%**	**98.6%**	**100%**	**95.5%**	**97.7%**	**100%**	**88.8%**	**94.1%**

* CNN-OCR trained without using any images belonging to our scenario.

It is important to highlight that we employed datasets proposed by several research groups from different countries (the same ones used by Laroca *et al.* [47]), with only 445 more images belonging to our scenario, to train the CNN-OCR architecture so that it is robust for various license plate layouts. In this way, as shown in Figure 10, CNN-OCR is able to correctly recognize license plates from various countries.

As the commercial systems were not tuned specifically for our dataset/scenario, we also report in Table 6 the results achieved by CNN-OCR when it was trained without using any images belonging to our scenario. It is remarkable that CNN-OCR still outperformed both commercial systems despite the fact that they are trained in much larger private datasets, which is a great advantage, especially in deep learning-based approaches [46, 47]. This experiment also highlights the importance of fine-tuning the CNN-OCR model to our scenario in order to achieve outstanding results.

Figure 11 shows some examples in which CNN-OCR failed to correctly recognize all license plate characters. As can be seen, errors occur mostly due to partial occlusions, extreme light conditions, and degraded license plates. Note that such conditions may cause one character to look very similar to another, and thus even humans can misread these

[3]Although OpenALPR has an open source version, the commercial version (the one used in our experiments) employs different algorithms for OCR trained with larger datasets to improve accuracy [49, 59].

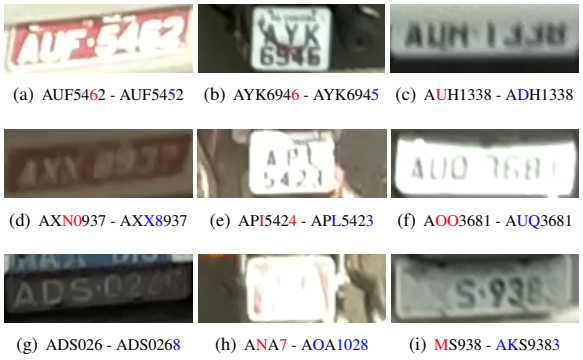
(a) BAG0160 (b) APY7367 (c) AXM6295

(d) 3J66282 (e) 997JDG (f) KJW5804

(g) RK069AV (h) ZG4100AC (i) VW4X4WP

Fig. 10. Examples of license plates that were correctly recognized by the CNN-OCR architecture. The images in the first row belong to our dataset while the others belong to public datasets acquired in other countries.

license plates (we even had to explore multiple frames and vehicle make/model information to check if the labeled string was correct in such challenging cases).



(a) AUF5462 - AUF5452 (b) AYK6946 - AYK6945 (c) AUH1338 - ADH1338

(d) AXN0937 - AXX8937 (e) API5424 - APL5423 (f) AOO3681 - AUQ3681

(g) ADS026 - ADS0268 (h) ANA7 - AOA1028 (i) MS938 - AKS9383

Fig. 11. Examples of license plates that were partially or not recognized by the CNN-OCR architecture. For each license plate, we show the predicted and ground truth strings, where the red and blue characters denote the CNN-OCR misreadings and the ground truth, respectively.

Finally, as can be seen in Table 7, the fusion of appearance information (vehicle shape features obtained by the best network found in our experiments shown in Table 5) with textual information (OCR) using the proposed two-stream neural network, as described in Section 4, increased the F-score by nearly 5% over each feature separately.

We believe that both features have a significant level of complementarity, that is, even if CNN-OCR does not recognize all license plate characters correctly, it is still possible to correctly match the image pairs in most of the cases by using the textual and confidence information available, as well as the characters and shape similarity features. Figure 12 shows some classification results obtained by our two-stream neural network.

Table 7. Vehicle identification performance based on shape and textual features: performance of the proposed two-stream network by using the best CNN for shape (Small-VGG) and the best OCR model (CNN-OCR). For comparison, we included the performance of each stream when used alone.

Architecture	$\bar{P}$	$\bar{R}$	$\bar{F}$
One-Stream (Shape)	90.43%	92.54%	91.35%
One-Stream (CNN-OCR)	100.00%	88.80%	94.10%
Two-Stream (Shape + CNN-OCR)	**99.35%**	**98.50%**	**98.92%**

As an additional contribution, we shared in GitHub[4] three alternative architectures that explore the same features but use additional streams and temporal information.

6. DISCUSSION

In this section, we compare the proposed Vehicle-Rear dataset with four other well-known datasets described in the literature, namely, VeRi-776 [61], VERI-Wild [8], VehicleID [10], and CityFlowV2 [9]. An overview of these datasets is presented in Table 8. As can be seen, our dataset is the only one with visible/legible/labeled license plate identifiers and with all videos recorded in Full-HD resolution. Furthermore, Vehicle-Rear and CityFlowV2 are the only datasets that provide uncropped frames, enabling the design of vehicle identification approaches that explore the entire scene. Another point worth noting is that none of the public datasets for vehicle identification – except ours – have motorcycle images, despite the fact that motorcycles are one of the most popular transportation means in metropolitan areas, especially in developing countries [47, 62]. On the other hand, the images in the Vehicle-Rear dataset were not collected by as many cameras as those from CityFlowV2 and VeRi (776 and Wild), nor in multiple views.

In summary, the main advantage of the proposed dataset, compared to existing ones, is that it enables the development of novel approaches/architectures for vehicle identification (both cars and motorcycles) based on the license plate identifiers in conjunction with vehicle shape features.

As can be seen in Figure 13, even if we consider only images from the vehicle's rear, in most of the cases the license plate identifier is illegible for the VeRi-776 dataset, and the authors did not provide the bounding boxes and strings of the license plates in cases where they are legible, and it would be impractical (i.e., a very laborious task) to scan/label them to train/evaluate our networks. Moreover, two state-of-the-art commercial systems that are widely employed to locate and recognize license plates from various regions/countries, Sighthound [58] and OpenALPR [59], rejected or failed in 79% and 96% of all images available in the VeRi-776 dataset, respectively. We emphasize that even though in [6, 39] the authors claimed that they extended the VeRi-776 dataset with license plate annotations, these annotations were not made

[4] https://github.com/icarofua/vehicle-rear

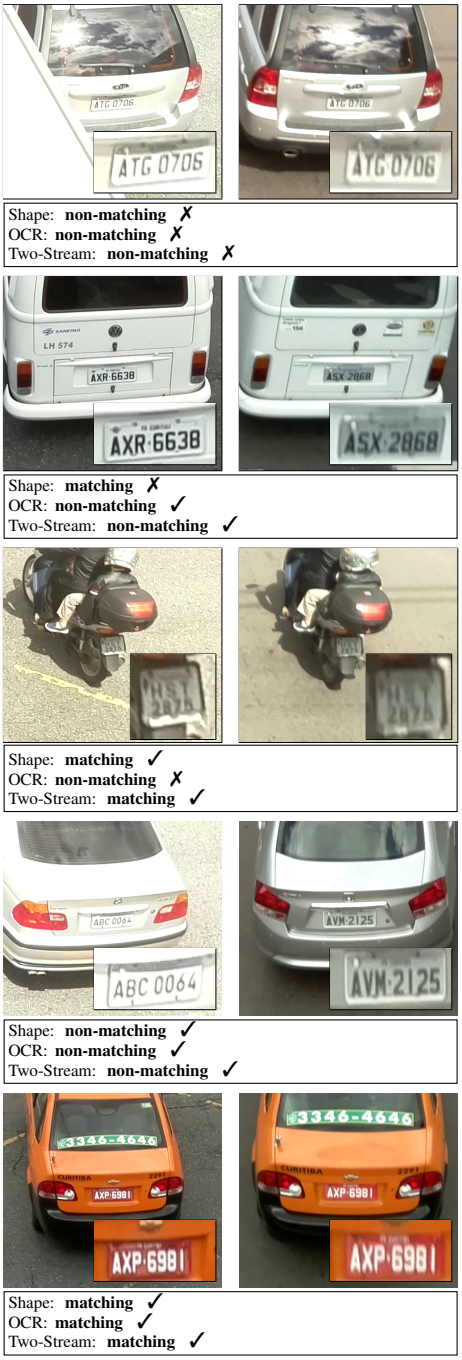


Fig. 12. Inference results: the first three rows show examples where the three architectures failed: partial occlusion; CNN-Shape failed (similar shape); CNN-OCR failed (HBI-20 for the left plate, HLG-297 for the right one, while the ground truth is HST-2875). In the last two examples, all architectures found a true non-matching and a matching, respectively.

Table 8. A comparison of publicly available datasets for vehicle identification with the proposed dataset called Vehicle-Rear. The entries marked with * refer to cases where only cropped patches (i.e., vehicle bounding boxes and not the entire scene) are provided by the authors.

Dataset	Image Resolution	# Vehicles	# Cameras	# Boxes Labeled	Multi-view	Video-based	Motorcycles	License Plate Information
VeRi-776 [61]	✗*	776	20	49,357	✓	✗	✗	✗
VERI-Wild [8]	✗*	40,671	174	416,314	✓	✗	✗	✗
VehicleID [10]	✗*	10,319	2	90,000	✗	✗	✗	✗
CityFlowV2 [9]	$[1280 \times 720]$ to $[1920 \times 1080]$	880	46	333,931	✓	✓	✗	✗
Vehicle-Rear (ours)	$[1920 \times 1080]$	2,093	2	26,161	✗	✓	✓	✓

available due to privacy restrictions (according to the first author of [6, 39, 61]). In the VERI-Wild [8], CityFlowV2 [9] and VehicleID [10] datasets, on the other hand, it is not even possible to exploit information from the license plate regions for vehicle identification, as they were purposely redacted in all images (with a black bounding box) by the respective authors because of privacy restrictions. For the record, CityFlowV2 is an updated version – with refined annotations – of CityFlow [63].

In this sense, we remark that the above datasets – as well as others available in the literature – have a different purpose from the one introduced in this work, as they have images from urban surveillance cameras in different resolutions and viewpoints. As stated in [13], these datasets have high inter-similarity (similar visual appearance for two different makes, model and type of vehicles) and high intra-variability.

Lastly, it is important to highlight that the Vehicle-Rear dataset is part of a cooperation agreement[5] between the universities involved in this project and the city where the videos were recorded. This agreement involves **free and open access** to the data mentioned in this article.

7. CONCLUSIONS

In this paper, we introduced a novel dataset for vehicle identification that, to the best of our knowledge, is the first to consider the same camera view of most city systems used to enforce traffic laws; thus, it enables to extract features with quality and also to retrieve accurate information about each vehicle, reducing ambiguity in recognition.

To explore the Vehicle-Rear dataset, we designed a two-stream CNN architecture that combines the discriminatory power of two key attributes: the vehicle appearance and license plate recognition. For this purpose, we proposed a novel approach to compute textual similarities from a pair of license plate regions which were then combined with shape similarities extracted from a Siamese architecture.

The proposed architecture achieved precision, recall and F-score values of 99.35%, 98.5%, 98.92%, respectively. The combination of both features (vehicle shape and OCR) brought an F-score boost of nearly 5%, solving very challenging instances of this problem such as distinct vehicles with very similar shapes or license plate identifiers.

Finally, although we achieved an F-score of 98.92% there is still room for improvement. Some open research problems are (i) designing novel networks that could extract vehicle information with the same quality from even smaller image patches; (ii) designing a one-stream architecture that has performance comparable to multi-stream architectures; and (iii) exploring other fine-grained attributes or temporal sequences for vehicle identification.

[5] A copy of the cooperation agreement can be obtained upon request.

Acknowledgments

This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grants 428333/2016-8, 313423/2017-2, 309292/2020-4 and 308879/2020-1, and in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) under Grant 88887.516264/2020-00. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. Additionally, we thank all the support given by Curitiba's City Hall, Aditya Choudhary for his help with the code, and Diogo C. Luvizon for all his support in recording the videos used in our experiments.

References

[1] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[2] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4328–4338, Sep. 2019.

[3] W. H. Lin and D. Tong, "Vehicle re-identification with dynamic time windows for vehicle passage time estimation," *IEEE Trans. on Intelligent Transportation Systems (ITS)*, vol. 12, no. 4, pp. 1057–1063, 2011.

[4] M. Ndoye, V. F. Totten, J. V. Krogmeier, and D. M. Bullock, "Sensing and signal processing for vehicle re-identification and travel time estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 119–131, 2011.

[5] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group sensitive triplet embedding for vehicle re-identification," *IEEE Transactions on Multimedia*, pp. 2385–2399, 2018.

[6] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2018.

[7] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li, "Multi-modal metric learning for vehicle re-identification in traffic surveillance environment," in *IEEE International Conference on Image Processing*, 2017, pp. 2254–2258.

(a) VeRi-776 [61]

(b) VERI-Wild [8]

(c) VehicleID [10]

(d) CityFlowV2 [9]

Fig. 13. Vehicle rear images of four public datasets: in the VeRi-776 dataset (a), most license plates are not legible and the authors did not provide any annotations for the plates; in the VERI-Wild, VehicleID and CityFlowV2 datasets (b-d), the license plates were redacted due to privacy restrictions.

[8] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

[9] M. Naphade *et al.*, "The 5th AI city challenge," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2021.

[10] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[11] P. da República, "LEI Nº 9.503, DE 23 DE SETEMBRO DE 1997 - Código de Trânsito Brasileiro." http://www.planalto.gov.br/ccivil_03/leis/l9503compilado.htm, 1997, accessed: 2021-06-14.

[12] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3275–3287, July 2018.

[13] J. Deng, M. S. Khokhar, M. U. Aftab, J. Cai, R. Kumar, J. Kumar *et al.*, "Trends in vehicle re-identification past, present, and future: A comprehensive review," *arXiv preprint arXiv:2102.09744*, 2021.

[14] B. Tian, B. T. Morris, M. Tang, Y. Liu, Y. Yao, C. Gou, D. Shen, and S. Tang, "Hierarchical and networked vehicle surveillance in ITS: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 557–580, 2015.

[15] R. O. Sanchez, C. Flores, R. Horowitz, R. Rajagopal, and P. Varaiya, "Vehicle re-identification using wireless magnetic sensors: Algorithm revision, modifications and performance analysis," in *IEEE International Conference on Vehicular Electronics and Safety*, 2011, pp. 226–231.

[16] S. Charbonnier, A.-C. Pitton, and A. Vassilev, "Vehicle re-identification with a single magnetic sensor," in *IEEE International Instrumentation and Measurement Technology*, 2012, pp. 380–385.

[17] I. Christiansen and E. L. Hauer, "Probing for travel time: Norway applies avi and wim technologies for section probe data," *UC Berkeley Transportation Library*, pp. 41–44, 1996.

[18] H. Wang, J. Hou, and N. Chen, "A survey of vehicle re-identification based on deep learning," *IEEE Access*, vol. 7, pp. 172 443–172 469, 2019.

[19] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Computer Vision and Image Understanding*, vol. 182, pp. 50–63, 2019.

[20] R. R. Cabrera, T. Tuytelaars, and L. Van Gool, "Efficient multi-camera detection, tracking, and identification using a shared set of haar-features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 65–71.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[23] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single line text regions," *Pattern Recognition*, vol. 46, no. 3, pp. 1078–1090, 2013.

[24] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–31.

[25] H. C. Chen, J.-W. Hsieh, and S.-P. Huang, "Real-time vehicle re-identification system using symmelets and homs," in *IEEE Intl. Conference on Advanced Video and Signal Based Surveillance*, 2018, pp. 1–6.

[26] M. Cormier, L. W. Sommer, and M. Teutsch, "Low resolution vehicle re-identification based on appearance features for wide area motion imagery," in *IEEE Winter Applications of Computer Vision Workshops*, 2016, pp. 1–7.

[27] C. Zhang, X. Wang, J. Feng, Y. Cheng, and C. Guo, "A car-face region-based image retrieval method with attention of sift features," *Multimedia Tools and Applications (MTA), Springer*, pp. 1–20, 2016.

[28] D. C. Luvizon, B. T. Nassu, and R. Minetto, "A video-based system for vehicle speed measurement in urban roadways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1393–1404, 2017.

[29] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[30] H. Ye, Z. Wu, R.-W. Zhao, X. Wang, Y.-G. Jiang, and X. Xue, "Evaluating two-stream CNN for video classification," in *ACM International Conferecen on Multimedia Retrieval*, 2015, pp. 435–442.

[31] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *IEEE international conference on computer vision*, 2017, pp. 1983–1991.

[32] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[33] I. de Oliveira, K. V. O. Fonseca, and R. Minetto, "A Two-Stream Siamese neural network for vehicle re-identification by using non-overlapping cameras," in *IEEE Intl. Conference on Image Processing*, 2019, pp. 1–4.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] R. Minetto, M. Pamplona Segundo, and S. Sarkar, "Hydra: An ensemble of con-

volutional neural networks for geospatial land classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6530–6541, 2019.

[36] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[37] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *IEEE International Conference on Computer Vision*, 2017, pp. 1900–1909.

[38] Y. Zhou and L. Shao, "Vehicle re-identification by adversarial bi-directional lstm network," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, vol. 00, Mar 2018, pp. 653–662.

[39] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 869–884.

[40] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, Aug 2019.

[41] H. Li, P. Wang, M. You, and C. Shen, "Reading car license plates using deep neural networks," *Image and Vision Computing*, vol. 72, pp. 14–23, 2018.

[42] M. Dong, D. He, C. Luo, D. Liu, and W. Zeng, "A CNN-based approach for automatic license plate recognition in the wild," in *British Machine Vision Conference (BMVC)*, 2017, pp. 1–12.

[43] Z. Selmi, M. B. Halima, U. Pal, and M. A. Alimi, "DELP-DAR system for license plate detection and recognition," *Pattern Recognition Letters*, vol. 129, pp. 213–223, 2020.

[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision*, Oct 2017, pp. 2961–2969.

[45] G. R. Gonçalves, M. A. Diniz, R. Laroca, D. Menotti, and W. R. Schwartz, "Real-time automatic license plate recognition through deep multi-task networks," in *Conference on Graphics, Patterns and Images*, Oct 2018, pp. 110–117.

[46] S. M. Silva and C. R. Jung, "Real-time license plate detection and recognition using deep convolutional neural networks," *Journal of Visual Communication and Image Representation*, p. 102773, 2020.

[47] R. Laroca, L. A. Zanlorensi, G. R. Gonçalves, E. Todt, W. R. Schwartz, and D. Menotti, "An efficient and layout-independent automatic license plate recognition system based on the YOLO detector," *IET Intelligent Transport Systems*, vol. 15, no. 4, pp. 483–503, 2021.

[48] S. M. Silva and C. R. Jung, "Real-time brazilian license plate detection and recognition using deep convolutional neural networks," in *Conference on Graphics, Patterns and Images*, Oct 2017, pp. 55–62.

[49] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust real-time automatic license plate recognition based on the YOLO detector," in *International Joint Conference on Neural Networks*, 2018, pp. 1–10.

[50] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *European Conference on Computer Vision (ECCV)*, Sept 2018, pp. 593–609.

[51] Wikipedia, "Vehicle registration plates of Brazil," https://en.m.wikipedia.org/wiki/Vehicle_registration_plates_of_Brazil, 2021, accessed: 2021-06-14.

[52] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *International Conf. on Neural Information Processing Systems*, 1993, pp. 737–744.

[53] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[54] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2544–2550.

[55] R. Laroca, V. Barroso, M. A. Diniz, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "Convolutional neural networks for automatic meter reading," *Journal of Electronic Imaging*, vol. 28, no. 1, p. 013023, 2019.

[56] R. Laroca, A. B. Araujo, L. A. Zanlorensi, E. C. de Almeida, and D. Menotti, "Towards image-based automatic meter reading in unconstrained scenarios: A robust and efficient approach," *IEEE Access*, vol. 9, pp. 67 569–67 584, 2021.

[57] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.

[58] S. Z. Masood, G. Shu, A. Dehghan, and E. G. Ortiz, "License plate detection and recognition using deeply learned convolutional neural networks," *arXiv preprint arXiv:1703.07330*, 2017.

[59] OpenALPR Software Solutions, "OpenALPR Library & Cloud API," http://www.openalpr.com/, 2019.

[60] G. R. Gonçalves, M. A. Diniz, R. Laroca, D. Menotti, and W. R. Schwartz, "Multi-task learning for low-resolution license plate recognition," in *Iberoamerican Congress on Pattern Recognition*, Oct 2019, pp. 251–261.

[61] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *IEEE International Conference on Multimedia and Expo*, July 2016, pp. 1–6.

[62] G.-S. J. Hsu and C.-W. Chiu, "A comparison study on real-time tracking motorcycle license plates," in *IEEE Image, Video, and Multidimensional Signal Processing Workshop*, 2016, pp. 1–5.

[63] Z. Tang *et al.*, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019, pp. 8789–8798.