Reduced Annotation Based on Deep Active Learning for Arabic Text Detection in Natural Scene Images

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY-NC-SA 4.0

SUBMISSION DATE / POSTED DATE

21-12-2021 / 23-12-2021

CITATION

Boukthir, Khalil; Qahtani, Abdulrahman M.; Almutiry, Omar; dhahri, habib; Alimi, Adel (2021): Reduced Annotation Based on Deep Active Learning for Arabic Text Detection in Natural Scene Images. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17327963.v1>

DOI

[10.36227/techrxiv.17327963.v1](https://doi.org/10.36227/techrxiv.17327963.v1)

Reduced Annotation Based on Deep Active Learning for Arabic Text Detection in Natural Scene Images

Khalil Boukthir^a, Abdulrahman M. Qahtani^c, Omar Almutiry^d, Habib Dhahri^d, Adel M. Alimi^{a,b}

^aREsearch Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia.

^bDepartment of Electrical and Electronic Engineering Science, Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa.

^cDepartment of Computer Science, College of Computers and Information Technology, Taif University, P.O.Box. 11099, Taif 21944, Saudi Arabia.

^dCollege of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia.

ABSTRACT

Providing labeled Arabic text images dataset for scene text detection is inherently difficult and costly at the same time. Consequently, only few small datasets are available for this task. Previous work has only focused on the data augmentation technique of small datasets; however, the images generated with these techniques cannot reproduce the complexity and variability of natural images.

In this paper, we propose a new Arabic text images dataset using the Google Street View service named TSVD. The dataset contains 7k images collected from different Tunisian cities. It is much more diverse and complex than current image datasets. Taking advantage of this dataset to train a CNN models, annotation is required for building high performance models. The annotation task consumes a lot of time and effort for researchers due to its repetitiveness. The development time of text detection systems in natural images is valuable with an effective use. We believe that we have developed a Deep Active Learning algorithm for the annotation phase. A Deep Active Learning algorithm for the annotation phase has been developed by approaching the annotation suggestion task using a deep learning text detector. Deep convolutional neural networks (CNN) are used to perform the text detection in natural scene images. Our deep active learning framework combines convolutional neural network (CNN) and active learning approach. This reduces annotation effort by making pertinent suggestions on the most effective annotation areas. We utilize uncertainty provided by CNN models to determine the maximum uncertain areas for annotation. Deep active learning is shown in order to reduce significantly the number of training samples required and also to minimize the annotation work of our dataset up to 1/5.

Keywords: Active learning, Deep learning, Annotation, Natural scene images, Text detection

1. Introduction

A large dataset of scene text images is an important and difficult factor in the extraction of textual information in natural images. It is the input to any computer vision system. Actually, Convolutional Neural Network (CNN) is the most popular network for text detection in natural scene image. Nonetheless, their training is usually based on large sets of fully supervised

data. In particular, complex scripts seem to require a lot of samples as Arabic. A solution to this problem has been proposed in [1] EASTR, Saudi Arabia streets text images dataset and in [2] for real-time Arabic scene text detection. However, unfortunately, they are not available publically due to the shortage of dataset for Arabic scene text detection. Hence, it is roughly inaccessible to train CNN models for annotating Arabic dataset during the annotation phase.

In this work, we propose improvements similar to [3], to the annotation problem of our dataset of street view images. We have made three key contributions:

First, we propose a new semi-automatic method for generating natural scene text images from the streets of Tunisian cities called TSVD. We can see the image generation workflow ar-

e-mail: khalil.boukthir@regim.usf.tn (Khalil Boukthir),
amqahtani@tu.edu.sa (Abdulrahman M. Qahtani),
oalmutiry@ksu.edu.sa (Omar Almutiry),
hdhahri@ksu.edu.sa (Habib Dhahri),
adel.alimi@regim.usf.tn (Adel M. Alimi)

chitecture in Fig. 1. We believe that a large-scale dataset of nature text images is an essential resource to train a deep learning model of detection of Arabic text images in the wild. Extracting panorama images from Google Street View requires us to do a preprocessing step to eliminate useless images. The large number of extracted images makes the manual preprocessing a very difficult task. In this respect, we use a standard CNN tools for this step, which detects scene text with both high accuracy and efficiency.

The second contribution is the result of our semi-automatic method to generate a new dataset of Arabic text images in the street. This dataset, called TSVD (Tunisia Street View Dataset), is suitable for training high performance Arabic scene text detectors. The key difference with the existing text datasets as those of [1] and [2] is basically that it is the only publicly available ones and bearing in mind that it is suitable for training Arabic Text Detectors. Some samples of our dataset are presented in Fig. 3.

Finally, the use of the text annotation deep active learning approach (see Fig. 2) is both an efficient and a fast way to annotate data. The method used is inspired by [3] with improvements similar to [4] to the training CNN models. We show that using a conjunction of these three methods: CNN, bootstrapping and uncertainty sampling, the amount of annotation work can be considerably minimized to 60% as an average. In addition, we demonstrate that deep active learning improves the rapidity of annotation. Furthermore, our experiments are performed on a portion of our dataset (700 images). We use data augmentation essentially to increase the sampling rate. It helps to show the best results with precision and dice index.

The paper is organized as follows: Section 2 discusses the related works. Section 3 describes the materials and methods used to generate our annotated dataset for text detection in natural scene images. This section depicts how to use a simple process for generating a street view dataset and the deep active learning method is introduced consecutively. The experiments and results are shown in section 4 and so is the conclusion in section 5.

Fig. 1: Flowchart of generating TSVD.

2. Related works

The proposed method is related to three main research topics, namely: Street View Datasets (see Table 1), Arabic Text Scene Detection Datasets (see Table 2) and Dataset Text Detection Annotation, whose literature is reviewed in the following.

2.1. Street View Datasets

In this section we will describe some datasets that use Google Street View as a data source (see Table 1).

Google Street View dataset [5] used for the geolocation of an image utilizing a novel multiple nearest neighbor feature matching method using Generalized Minimum Clique Graphs (GMCP). This dataset contains 102k high quality Google Street View images.

The Street View Text (SVT) dataset [6] was harvested from Google Street View. Images text in this dataset exhibits high variability and often has low resolution. In dealing with outdoor street level imagery, we note two characteristics: firstly, it consists of 350 images annotated with word-level axis-aligned bounding boxes, and secondly, it contains smaller and lower resolution text.

StreetLearn [7], is a limited dataset of Google Street View images, has been approved for use with the StreetLearn project and academic research. We are releasing these Google Street View panoramas (approximately 143k) and street connectivity graphs covering two cities, Manhattan and Pittsburgh.

GSV [8], is a dataset of street intersections of four cities in North America and data of 403 crowd workers. By evaluating it through using Tohme smart system, it occurs that it has similar performance in ramp detection, compared to a manual labeling approach only (F-measure: 84% vs 86% basic), but with a 13% reduction in effort time.

SVHN [9], is a real-world image dataset for developing machine learning and object recognition algorithms with minimal requirement on data preprocessing and formatting. It is possibly seen similar in flavor to MNIST (e.g., the images are of small cropped digits), but incorporates an order of magnitude of more labeled data (over 600k digit images) and comes from a significantly harder, unsolved, real world problem (recognizing digits and numbers in natural scene images). SVHN is obtained from house numbers in Google Street View images.

2.2. Arabic Text Scene Detection Datasets

ARASTEC [10], is a dataset of Arabic characters where the source is scene images from where individual characters were manually segmented. This dataset contains a set of 260 photographed images of sign boards, hoardings and advertisements.

ARASTI [11], is an Arabic scene text dataset containing 374 images taken under real life situations without caring about the environmental conditions. The Arabic word image dataset contains 1280 cropped word images from Scene Texts.

EASTR-42K [1], the proposed data captured bilingual (English and Arabic) scene text images and tried to cover every possible word permutation of Arabic language with its variant shapes. This dataset covers huge variety of English and

Fig. 2: Deep active learning annotation architecture.

Arabic scene text appeared in unconstrained environment. It contains 2107 Arabic and 983 English text lines with 784 as multi-lingual.

ASAYAR [12], is used for the localization of text in traffic panels. It consists of 3 sub-datasets: Arabic-Latin scene text localization, traffic sign detection, and directional symbol detection. The dataset contains 1763 images collected from different Moroccan highways, and annotated manually, using 16 object categories. The fully annotated ASAYAR images contain more than 20k bounding box objects with Arabic and French language. Containing 1375 text-based panels images with word and line level annotations.

In [2], the lack of publicly available datasets motivates Rajae et al. which creates their own dataset for Real-time Arabic scene text detection. It contains 575 images manually processed along with data augmentation to enrich the process of training a fully convolutional neural networks model.

2.3. Dataset Text Detection Annotation

Text image annotation aims at labeling the text boxes in each image. Recent images annotation phases are often done manually. Hence, the scene text image is manually annotated into different text lines [1], [2], [11]. For all types of scripts, manual annotation is costly in terms of both time and staff [13]. ASAYAR [12] was manually annotated by four researchers during nine months (more than 2000 hours of work). In addition, different solutions have been proposed to deal with the presence of this phase of text annotation in the videos. Oussama et al. [14], propose a semi-automatic method to annotate the AcTiv dataset. Generally, any standard dataset essentially undergo annotation for ground truth. A department of IBM Almaden Research Center employs a deep active learning algorithm in order to reduce the annotation of a dataset of cell images [3].

3. Materials and methods

3.1. TSVD Dataset

3.1.1. Properties

Google Street view images come from two sources, either Google or public contributors. TSVD is built on the basis of

Table 1: Datasets extracted from Google Street View.

Dataset	Size	Type of content
SVHN [9]	600k	Images of house numbers
StreetLearn [7]	143k	Google Street View panoramas
Google Street View [5]	102k	Google Street View images
GSV dataset [8]	1k	Street intersections
Street View Text [6]	350	Business signage

Google Street View images. This paper is written with a dataset consisting of 7000 images of 5 Tunisian cities. The main properties of our dataset are described as follows:

- **Scale:** TSVD aims to provide the most comprehensive and diverse coverage of the Arabic world in the street view images. The 7k images are collected for each city. Table 3 show the distributions of the number of images per city for the current TSVD. If we compare it with other datasets of the same nature, the scale of our dataset is competitive (see Table 2).

- **Accuracy / Positioning:** A team of Google researchers describes the technical challenges involved in capturing, processing, and serving street-level imagery on a global scale [15], [16]. Recorded photographs must be associated with accurate positioning. This is done via a Global Positioning System (GPS), wheel speed sensor, and inertial navigation sensor data.

- **Diversity:** One of the goals of building our dataset is mainly to have text images with variable appearances and positions. Data TSVD is collected from 5 Tunisian cities. The nameplate and the Shop Signs of the selected cities streets contain different Arabic texts with different quality, font, background and script resolution.

- **Resolution:** Images were acquired with a high resolution Ladybug2 camera (resolution 1024 x 768) by Point Grey Research. But after cropped panorama images, resolution go to low 512 x 512.

Fig. 3: Examples of images from our dataset TSVD.

Table 2: Arabic Text Detection Datasets.

Dataset	Size	Type of content	Availability
EASTR-42K [1]	42K	Textlines, Words, Characters	Private
ASAYAR [12]	1763	Textlines, Words	Public
Dataset for Real-time Arabic scene text detection [2]	575	Words	Private
ARASTI [11]	374	Words	Public
ARASTEC [10]	260	Characters	Public

Table 3: TSVD dataset images by city.

City	Number of images
Tunis	1083
Sfax	593
BenArous	1885
Bizert	2348
Djerba	1091

3.1.2. Constructing Dataset

- Data Collection: To easily collect street view images, we chose five Tunisia cities where the Google Street View showing much heavier car circulation than others likewise. The list of cities is given in Table 3. In order to identify each position in Google Street View, it is necessary to access via a link consisting of four following parts: GPS coordinates of the position of the street view camera, h: degree of horizontal rotation (between 0 and 360), t: degree of vertical rotation (between 0 and 180) and the ID panorama.

Next, we will get ID panorama list and GPS coordinates. We use an Html page that uses a JavaScript code and Google Maps Platform developed by Petr et al. [17]. In Fig. 5, we explain the Google Street View link that has a specific format by inserting the IDs and GPS coordinates in the specific positions in to order to generate a list of Google Street View links.

The free Street View Grabber¹ utility allows to extracting

Fig. 4: The cropped image panorama using the tool Grabber Street View (a) The initial panorama image (b) Cropped images to generate.

panoramic images from Google Street View, in a very simple way, just copy and past the Google Street View link in this tool. A quality setting adjusts the size of the extracted images, which makes it even possible to export multiple panoramas at the same time. For that purpose, we have generated street view link lists. Additionally, the quality is always set to maximum for resolution setting, enabling the Crop Panorama option to help us to generate 91 images from a single panorama image (see Fig. 4).

- Data preprocessing: Data preprocessing is an important step in the building dataset process. One of the major drawbacks in using these cropped images is gathering a great

¹Link: shorturl.at/dsFO1

https://www.google.com/maps/@34.7330599,10.7551391,3a,75y,227.35h,97.12t/data=!3m6!1e1!3m4!1s!1sa10KFNvfqSwNwnla70sGeA!2e0!7i13312!8i6656

(a) (b) (c) (d)

Fig. 5: Description of a Google Street View link (a) GPS coordinates of the position of the street view camera (b) Degree of horizontal rotation: between 0° and 360° (c) Degree of vertical rotation: between 0° and 180° and (d) The panorama ID.

number of textless images. We used a shell script to delete the images at the bottom and at the top, particularly those representing sky as well as the roads. At this step, the number of images has been demineralized to 20% of the initial number. Here is the shell script used to remove the 39 top images, which are constantly representing the sky and the 39 bottom ones showing the paved road:

```
patch_int = 1; ## index of the initial image
patch_f = 39; ## index of the final image
for i in (seq patch_int patch_f);
do rm -- *tile$i*;
done;
```

3.2. Deep active learning annotation architecture

Our architecture relies on a deep active learning framework to annotate Arabic text in natural scene images. The basic idea is inspired by a Suggestive Annotation Algorithm proposed in [4]. We propose some special designs for adapting this architecture to efficiently annotate arabic text in natural images. In addition, we propose to use uncertainty sampling proposed in [3]. Then we adapt the deep active learning framework to annotate the arabic text in natural images by CNN, which is a scene text detector.

To deal with the specificity of scene text annotation, we propose the deep active learning annotation for Arabic text detection in natural scene image, which is composed by three main components: CNN bootstrapping, uncertainty sampling, and annotation suggestion. The overall architecture is depicted in Fig. 2, and detailed in the following.

```
The data is divided into N = 10 batches
Data_labeled = batch(1)
For i from 1 to N-1
  Data_unlabeled = batch(i+1)
  Data_labeled_list = Bootstrap sampling (Data_labeled)
  For j from 1 to 3
    Train CRAFT (Data_labeled_list[j])
    Select better CRAFT
  Test CRAFT on Test(Data_unlabeled)
  Obtain region score on Val(Data_unlabeled)
  Data_labeled = Data_labeled + Samples in Val(Data_unlabeled) whose
  zero boxing predicted in the images (data with labels)
```

Fig. 6: Algorithm for deep active learning.

3.2.1. CNN bootstrapping

At the initial stage of annotation process, we combined two techniques: CNN text detector as it effectively detects text area

in natural images and bootstrap sampling as it is a random sample conducted with replacement.

Starting with very little training data, we iteratively train a set of CNN models using the bootstrapping technique to generate training data. At the end of each stage, we extract useful information from these three CNN models to decide what will be the next batch of images to annotate. After acquiring the new annotation samples, the next stage started by using all the available annotated images.

The CNN model is a fully convolutional network architecture based on VGG-16 [18], which is essentially the extracting architecture feature used to encode the network's input into a certain feature representation. The decoding segment of the CNN is similar to U-Net [19] which has skip connections that aggregating low-level features.

3.2.2. Uncertainty sampling and annotation suggestion

The strategy chosen to augment the learning performance is to find the uncertain samples with the deep active learning approach. We will place them at each iteration for annotation. However, the CNN models tend to extract region score of text boxes. Samples with a probability greater than a threshold (0.7) are not selected by CNN detector. Then our method aims to select images without text boxes. This technique enables us to filter and select only the images that the model is not sure to predict. Those samples will be suggested for annotation.

4. Experiments and results

4.1. Implementation details

Our deep active learning framework is implemented in Keras. The experiments were realized based on the training procedure explained in the following. In our experiments, we use nine training iterations and three CNN models. In so far as the background-foreground network is concerned, the image was resized to 640×640 , while maintaining the original aspect-ratio, 700 images (10% of dataset) were used during training. The environment used for the evaluation is Google Colab with GPU.

4.2. Deep active learning algorithm

An annotation based on deep active learning is performed to label the images in our dataset. The algorithm of our approach is composed of three main parts: we start by training the three CNN models with the three equal dataset sets generated using the bootstrapping samples technique. Afterwards, the suggested images to annotate for the following iteration, which is without boxes, is done by selecting the best model predicted result of the three CNN models. Finally, it is in the last part of the

Table 4: Comparison of results of both methods: deep active learning method and deep learning method.

Dataset	Training samples	Training time (s)	Precision (%)	Processing time(s)	Dice index
TSVD	17247	22506	89.86	195	0.81
TSVD_AL	3708	77224	82.77	266	0.84

Table 5: Results on Arabic Text Detection Datasets.

Dataset	Precision (%)		Dice index	
	Training on 1/5 of dataset	Training on all dataset	Training on 1/5 of Dataset	Training on all dataset
ASYAR [12]	83.78	93.57	0.88	0.94
MLT17 [20]	73.26	81.55	0.73	0.76
MLT19 [21]	74.09	81.56	0.73	0.75
TSVD	82.77	89.86	0.84	0.81

deep active learning algorithm in which we will manually annotate the images selected in the previous step. They are added to the dataset in the next iteration. The algorithm for performing the deep active learning experiment is described in Fig. 6.

4.3. Deep active learning evaluation

The experiments were realized in two phases as in [3], we start by doing control experiments, just by training the CNN model on 80% of the training data and in turn 20% of the data is used to test the Arabic text detection metrics in natural scenes. The second phase of the experiment is the implementation of the algorithm described in the previous section. Due to the inherent difficulties in training CNN models with large sets of Arabic Text Scene Datasets images, only a slice of our dataset and available for the first annotation iteration is used. In order to face this problem, data augmentation is employed.

Fig. 7: Undetected samples for each iteration.

4.4. Results

The results, measured using the precision and the dice index, are reported in Table 4. It is worth noting that training the network using our deep active learning approach, is more effective than using deep learning. Specifically, we can see that the deep active learning experiment (denoted by ‘*_AL’) requires much

less training data to achieve similar precision and better detection metrics (dice index). On average, the number of training samples in the deep active learning experiment are reduced to 1/5 of the number of samples in the control experiment. This can be explained by the fact that few samples not detected by the CNN model and annotated manually are added to the training models on each batch. We can see the reduction in undetected samples in the Fig. 7.

Furthermore, the deep active learning approach is better for some cases (see Fig. 8 (b) and (e)) with a little difficult font, with the same precision in other cases (see Fig. 8 (c) and (f)) and for texts with an arbitrary orientation (see Fig. 8 (a) and (d)), the deep learning approach is more efficient.

Our approach shows competitiveness (see Table 5) on most public Arabic datasets: ASYAR [12], MLT17 [20] and MLT19 [21] and demonstrates precision ability by showing these performances using the trained models with 1/5 of the data.

The training time for deep active learning experiments is more noticeable than with deep learning approach, as shown in Table 4. Actually, in the deep active learning experiment, the three CNN models are trained at each iteration. On the other side, in deep learning experiment, the CNN model is trained once for the entire training data. This explains why the deep active learning take more time for training.

5. Conclusion

This article presents the generation of a new dataset of word-level annotated using a semi-automatic collection and preprocessing workflow and proposes a deep active learning framework for the detection of scene Arabic text images. A dataset for Arabic scene text detection is a challenge in terms of number and image invariant, 7k Arabic text images with a lot of background, font and background diversity. It will be useful to validate the detection system of Arabic text in natural images. A deep active learning approach may be employed to reduce the annotation burden on researchers and professionals for annotation tasks. Very few training samples are required for achieving

Fig. 8: (a),(b),(c): Text boxes detected using the deep learning approach
(d),(e),(f): Text boxes detected using the deep active learning approach.

good quality detection. In the light of the present paper's findings, it is worth noting that the number of training samples is reduced to 1/5 of the original training size on average.

Acknowledgement

We deeply acknowledge Taif University for Supporting this study through Taif University Researchers Supporting Project number (TURSP-2020/327), Taif University, Taif, Saudi Arabia. The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48. Thanks are also due to Tarek M.Hamdani, Fatma Ben Said, Lobna Haddad and Wael Ouarda for her discussions on preparing the manuscript.

References

- [1] Saad Ahmed, Saeeda Naz, Muhammad Razzak, and Rubiyah Yusof. A novel dataset for english-arabic scene text recognition (eastr)-42k and its evaluation using invariant feature extraction on detected extremal regions. *IEEE Access*, 01 2019.
- [2] Rajae Moumen, Raddouane Chiheb, and Rdouan Faizi. Real-time arabic scene text detection using fully convolutional neural networks. *International Journal of Electrical and Computer Engineering (IJECE)*, 11:1634, 04 2021.
- [3] Aritra Chowdhury, Sujoy K. Biswas, and Simone Bianco. Active deep learning reduces annotation burden in automatic cell segmentation. In John E. Tomaszewski and Aaron D. Ward, editors, *Medical Imaging 2021: Digital Pathology*, volume 11603, pages 94 – 99. International Society for Optics and Photonics, SPIE, 2021.
- [4] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017.
- [5] Amir Roshan Zamir and Mubarak Shah. Image geo-localization based on multiplenearst neighbor feature matching usinggeneralized graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1546–1558, 2014.
- [6] Kai Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464, Nov 2011.
- [7] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.
- [8] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. Tohme: Detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 189–204, 2014.
- [9] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS DLW*, 01 2011.
- [10] Maroua Tounsi, Ikram Moalla, Adel M. Alimi, and Frank Lebouregois. Arabic characters recognition in natural scenes using sparse coding for feature representations. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1036–1040, 2015.
- [11] Maroua Tounsi, Ikram Moalla, and Adel M. Alimi. Arasti: A database for arabic scene text recognition. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 140–144, 2017.
- [12] Mohammed Akallouch, Kaoutar Sefrioui Boujemaa, Afaf Bouhoute, Khalid Fardousse, and Ismail Berrada. Asayar: A dataset for arabic-latin scene text localization in highwaraffic panels. *IEEE Transactions on Intelligent Transportation Systems*, PP:1–11, 10 2020.
- [13] Chongsheng Zhang, Weiping Ding, Guowen Peng, Feifei Fu, and Wei Wang. Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4727–4743, 2021.
- [14] Oussama Zayene, Jean Hennebert, Sameh Masmoudi Touj, Rolf Ingold, and Najoua Essoukri Ben Amara. A dataset for arabic text detection, tracking and recognition in news videos-activ. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 996–1000. IEEE, 2015.
- [15] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, June 2010.
- [16] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380, Sep. 2009.
- [17] Petr Gronát, Michal Havlena, Josef Sivic, and Tomas Pajdla. Building streetview datasets for place recognition and city reconstruction. *Research Reports of CMP, Czech Technical University in Prague*, 2011.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. in *International Conference on Learning Representations (ICLR)*, abs/1409.1556, 2015.
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *LNCS*, 9351:234–241, 10 2015.
- [20] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khelif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng-lin Liu, and Jean-Marc Ogier. Icdar 2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1454–1459, 2017.
- [21] Nibal Nayef, Cheng-Lin Liu, Jean-Marc Ogier, Yash Patel, Michal Busta, Pinaki Chowdhury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Umapada Pal, and Jean-Christophe Burie. Icdar 2019 robust reading challenge on multi-lingual scene text detection and recognition — rrc-mlt-2019. pages 1582–1587, 09 2019.