Cutting Sayre's Knot: Reading Scene Text without Segmentation. Application to Utility Meters.

Lluís Gómez, Marçal Rusiñol and Dimosthenis Karatzas
Computer Vision Center, Dept. Ciències de la Computació
Edifici O, Univ. Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain.

***Abstract*—In this paper we present a segmentation-free system for reading text in natural scenes. A CNN architecture is trained in an end-to-end manner, and is able to directly output readings without any explicit text localization step. In order to validate our proposal, we focus on the specific case of reading utility meters. We present our results in a large dataset of images acquired by different users and devices, so text appears in any location, with different sizes, fonts and lengths, and the images present several distortions such as dirt, illumination highlights or blur.**

***Keywords*-Robust Reading, End-to-end Systems, CNN, Utility Meters.**

I. Introduction

When attempting to design algorithms for endowing computers with the ability to read text from images, one often stumbles upon Sayre's paradox [1]. This dilemma is expressed as: "*text cannot be recognized without being segmented and cannot be segmented without being recognized*". Such paradox is mostly apparent when dealing with handwritten text, but is also patent when considering scene text. Individual words have to be separated among them when dealing with cursive text in the same way that textual elements have to be separated from cluttered environments in natural scene text before executing the reading process. However, in order to effectively perform such segmentations, one should ideally recognize which text is written. This contradiction is usually addressed through first engineering text / non-text classifiers for a later proper text recognition process [2].

Pipelines that first segment text that is later fed to a recognition process present certain drawbacks. On one hand, any segmentation errors will affect the subsequent text recognition step. In order to overcome this, usually over-segmentation strategies are adopted, e.g. [3], [4]. On the other hand, the ground-truth acquisition for training the final systems will be more expensive since one has to provide not only the text transcriptions but also the localization of such text within the images. Synthetic data is often used to train those systems as a means for reducing the cost of human labeling, e.g. [2], [5].

The main motivation of our work is to bypass any implicit segmentation step, and propose a reading system that given an image directly outputs its contained text. Our research hypothesis is that convolutional neural networks (CNNs) could be trained to automatically read in such an end-to-end manner. During the training phase, full images and the corresponding text transcription would be provided, without any indication on where on the image the text appears.

Figure 1. Examples of utility meters. Public domain images similar to the ones in our private dataset shown for illustrative purposes.

In order to validate such an hypothesis, in this paper we focus on the specific case of reading utility meters from camera acquired images. We can see some examples of the type of images in Figure 1. The scenario we take as a proof of concept might initially appear simple, since we are just dealing with a 10-digit alphabet. However, the problem still shows the same challenges that we usually encounter when reading text in natural images. Text may appear in any location, with different size and fonts, different lengths, while artifacts such as dirt, illumination highlights, blur, etc. are common. In addition, in such an application scenario, spotting approaches (e.g. [3]) guided by a predefined dictionary of words to read are not suitable, as the task is inherently an "open dictionary" one over the set of possible digits. So, in this case, we do not take advantage of any language model nor predefined set of words to read, although obviously the different possible readings are finite.

A. Related Work

Reading text in natural images is a hot research topic with an increasing interest by the community in recent years. End-to-end scene text recognition pipelines are commonly based in a two-stage approach, first applying a text localization algorithm to the input image and then recognizing the text present in the cropped bounding boxes provided by the detector. In the localization stage the dominant trend nowadays is on using CNN based detectors [6], [7], [8], that have replaced traditional methods based on connected components analysis [9], [10], [11]. Scene text recognition from pre-segmented text has been approached in two different conditions: using a small provided lexicon per image (also known as the word spotting task) [12], [13], or performing unconstrained

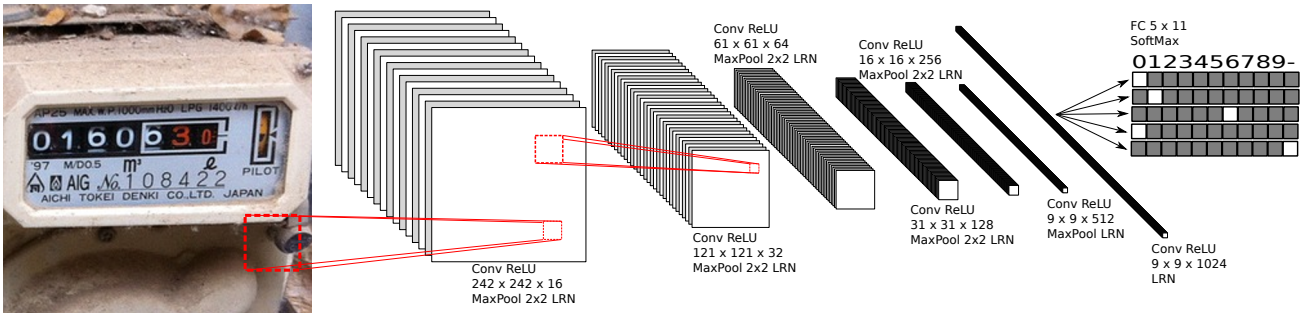


Figure 2. Our proposed architecture. Seven blocks composed of a 3x3 convolutional layer with ReLU activations, a 2x2 max pooling layer and a LRN layer. The number of convolutional filters is doubled at every block. A final step of five fully connected layers with softmax produce probability distributions over the 11 possible classes.

text recognition, i.e. allowing the recognition of out-of-dictionary words[14], [15]. True segmentation-free end-to-end approaches to date are limited to the interpretation of street signs [16],[17], where a canonical transcription of the street sign contents is sought. In [16] the problem is tackled through an LSTM based architecture, while in [17] attention models are employed in the same problem.

In the specific case of reading images of utility meters, different systems have been proposed in the literature, all of them following the paradigm of first detecting the meter reading zone, then segmenting individual digits and a later recognition problem. Reading zones are segmented by either using adaptive thresholding and mathematical morphology operations [18], [19] or by using supervised trained detectors such as multilayer perceptrons (MLP) [20], [21] or Haar cascades [22]. The later segmentation of digits has been performed by projection profiles [23], connected component analysis [18] or the MSER operator [24]. Finally, the digit classification step has been addressed by the use of Histogram of Gradients (HOG) features and Support Vector Machines (SVM) classifiers [24], CNNs [19], MLPs [25] or by directly using off the shelf OCR tools like Tesseract [19]. No public datasets have been made available in any of the previous works, respecting the privacy of the consumers, while no code is available.

In this paper we propose to avoid a two-stage pipeline of segmentation followed by recognition, and propose instead an end-to-end system that directly outputs the text in the scene in a segmentation-free manner. Moreover, we show that the resulting system is capable to detect the right length for the image string and to filter any non-significant digits, without any explicit training.

The rest of the paper is organized as follows. Section II presents the proposed methodology for the end-to-end convolutional neural network that is able to read text without any explicit segmentation step. Section III provides the implementation details of the proposed network and training procedures. Section IV presents the experimental results that were obtained while conclusions are drawn in Section V.

II. Architecture

A. Reading Utility Meters

Our model for reading utility meters is based on a single neural network that takes as input a meter image and is capable of producing the actual meter reading as output. It is important to notice that utility meters have different measurement's lengths, depending on the model. In our case, we assume that all meters have 4 or 5 significant digits. As shown in Figure 3, these digits are always followed by a set of non-significant digits, the decimal part of the reading that normally is not taken into account when billing the service and therefore is not relevant for the automatic meter reading. The separation between significant and non-significant digits has been traditionally tackled using color features [22], [21], [24], in our case the idea is that the network itself must learn both to ignore them and to find the correct measurement length.

Figure 3. Utility meters may have different reading lengths depending on the model. This particular model has 5 significant digits followed by three non-significant digits (in red) that correspond to the decimal part of the reading. The ground-truth data for this image would be the string "01971". Image source: Wikimedia Commons CC-BY-SA-3.0.

B. Proposed Method

We implement our model as a convolutional neural network that predicts each of the output digits simultaneously. The architecture of our network is shown in figure 2. It is observed that it is composed of a convolutional backbone

of seven blocks, each composed of a convolutional layer with Rectified Linear Units (ReLU) activations, a max pooling layer (except on *conv7*), and a local response normalization (LRN) layer. The number of convolutional filters is doubled at every block (starting from 16 at *conv1* layer up to 1024 at *conv7* layer) and the kernel size is 3×3 in all layers. All pooling layers use a kernel of 2×2 and a stride of 2. After the convolutional part we stack five independent fully connected layers, each one with a Softmax layer producing a probability distribution over the 11 possible classes (10 digits + 1 no-symbol class) for each significant digit of the final reading. The five outputs of the network are then treated as a typical classification output and trained using the Cross-Entropy loss function:

$$L = \frac{-1}{N} \sum_{n=1}^{N} \log(\hat{p}_{n,l_n}) \quad (1)$$

where N is the batch size, $\hat{p}_n$ is the prediction vector, and $l_n \in [0, 1, 2, ..., K-1]$ is the correct class label among the K classes for the n'th sample.

Intuitively for the network being able to read in an end-to-end manner it has to take into account global information extracted over the entire image to make the individual digit predictions. The initial convolutional layers extract visual features of the whole image, while the fully connected layers specialize in predicting the output probabilities for each digit.

This end-to-end model has several benefits over traditional methods of robust reading. First, the network can be trained with full images, without any explicit segmentation, and directly optimizes the end-to-end reading performance. Second, the particular design of the network allows for real time reading speeds while achieving high reading accuracy.

III. Implementation Details

We have implemented the end-to-end reading model using the Caffe [26] deep learning framework. We have trained the network from scratch using the RMSProp[1] optimizer for $500,000$ iterations with a batch size of 16 and an initial learning rate of 0.0001 that is decreased one order of magnitude every $50,000$ iterations. At training time we resize the input images to 512×512, subtract the train set mean, and do random crops of 483×483 as a data augmentation strategy. Figure 4 shows the evolution of the sum of the five losses over time. We appreciate how the network converges after $400,000$ iterations.

IV. Experiments

A. Dataset and Evaluation Protocol

In order to train and test the proposed reading system, we have used a private dataset of images of utility meters captured with mobile devices. Meter images are collected in real life by non-expert users, and therefore reflect real-life statistics. The dataset has a total of $222,198$ images of

[1] http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

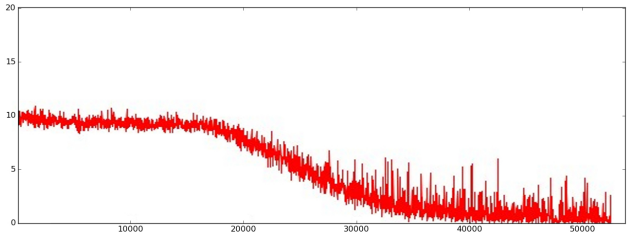


Figure 4. Training loss of our convolutional neural network over time.

utility meters from 13 different manufacturers comprising 47 different models of meters. Train and test splits were conducted in a stratified fashion looking at the meter model metadata, in order to ensure a correct balance. We ended with $177,758$ images being used during training while $44,440$ were kept for testing.

In order to evaluate the performance of the proposed system, we will compute its accuracy by counting in how many of the test images we obtain a perfect reading. That is, all the significant digits have been correctly read and the non-significant ones have been ignored, c.f. Fig. 7.

B. Baseline: Classic Segmentation and Recognition Pipeline

For the sake of completeness and in the light of the unavailability of any state of the art implementation or public dataset that would allow us a meaningful comparison to the state of the art, we have implemented a classic pipeline comprising separate segmentation and digit recognition stages. Apart from providing an indicative level of performance to compare against, this exercise has provided us with insight on the kind of limitations that such two-stage architectures present.

For the baseline pipeline and given the nature of the application, we have opted to implement an ad-hoc segmentation strategy rather than a generic text localizer. Following the published state of the art [22], [24], we have trained two Haar cascades [27]. The first Haar cascade deals with the detection of the reading zones of the meters and the second one is applied afterwards for the segmentation of the digits within the reading zone. We can see an example of the baseline segmentation results in Figure 5.

For the recognition stage, we have implemented two alternatives, on one hand an SVM digit classifier based on HOG features and on the other hand a LeNet based digit classifier CNN.

On our test dataset, if we just evaluate the individual digit recognizers by feeding the SVM or CNN the ground-truthed segmented digits, both recognizers yield individual digit recognition rates of around 95%. However, when evaluating the end-to-end task by using a two-stage approach, the performance is dramatically dropped to around 54%. Such a drop clearly indicates on one hand how segmentation errors are propagated and clearly affect the overall performance. On the other hand, it shows that recognition errors seem to be uniformly distributed in the meter images, i.e. a single digit error in a reading would

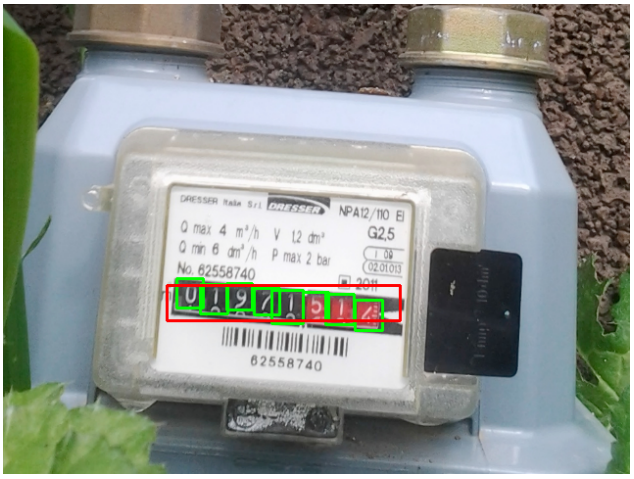


Figure 5. Example of the reading zone and digit segmentation obtained with Haar cascades.

translate to an end-to-end level error. These results are not meant to substitute state of the art, but are good indicators obtained of a pipeline reflecting the state of the art, in the same conditions as our system.

C. Results

The proposed system achieves an overall end-to-end recognition accuracy of 94.167% on the test set. This means that our CNN model was able to correctly read all significant digits, automatically filtering any non-significant ones, in $41,848$ of the $44,440$ test images. This represents a significant improvement of 40% compared to the baseline pipeline. At the level of digit recognition, the overall accuracy is 97.94%, while Table I shows the recognition accuracies of individual digits on the test set.

Table I
CLASSIFICATION ACCURACY FOR INDIVIDUAL DIGITS' PREDICTIONS.

	1st	2nd	3rd	4th	5th
Accuracy	99.043	98.393	97.596	97.313	97.353

It is worth to notice at this point how the leftmost digits are the ones with better accuracy. This is probably due to the fact that rightmost digits change with more frequency and thus are more prone to be captured in the position between two numbers, while it can also be partially an effect of the strong bias that leftmost digits exhibit. An interesting side effect of such behavior is that the majority of the errors being on the less significant digits means that the average reading error in the particular utility meter units (e.g. m^3 of water) is made smaller.

In the following we list the most common problems revealed by the error analysis done to our system's test outputs:

- **Strong glare or blur** are the most common sources of errors. As illustrated in Figure 6 in most cases the affected digits are hard to read even for humans.
- **Small scale**. Images captured from a long distance pose difficulties for two reasons: (1) the numbers become unreadable due to lack of resolution , while at the same time (2) the network has not seen many small-meter examples at training time.
- **Severe perspective distortion**. Although the network shows a robust performance for "moderate" distortions, it is not able to provide correct readings in extreme cases.
- **Capture errors**. The images are captured by non-expert users, who sometimes fail to include the full reading area in the image or provide upside down flipped images, causing the network to fail.
- **Annotation errors**. We have identified that approximately 10% of the errors correspond to images with wrong annotations.

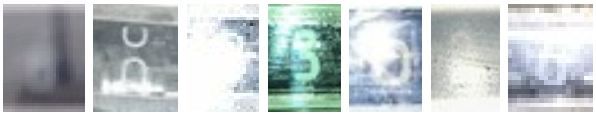


Figure 6. Examples of errors on individual digits affected by strong glare or blur. From left to right (Predicted/Truth): 0/6, 0/9, 1/8, 5/3, 6/0, 8/2, 0/6.

The above list of errors should not lead one to think that the rest of the images are good quality, controlled captures. As a matter of fact, the proposed method is very robust with reading meters in evidently complicated cases. Figure 7 shows some examples of correct readings where the performance of the algorithm is particularly robust.

The system's processing time for a single image (using a batch size of 1) is 31 ms. on a commodity GPU, thus providing a frame rate of 32 fps. In a $i7$ CPU the processing time rises to 1.19 seconds per image, still a pretty decent time.

D. Analysis of spatial context sensitivity

An exceptional outcome about the proposed method is not that the CNN is able to learn rich visual features for digit classification, but that it learns how to infer *which* is the correct digit to look at in order to make correct predictions in each of its independent outputs. Intuitively, to do so the network has to learn how to use contextual cues around the particular digits' locations.

In order to investigate from which parts of the image a certain classification prediction is coming from, we make use of the occlusion sensitivity visualization technique proposed by Zeiler and Fergus in [28]. For this we iteratively set a 32×32 square patch of the image to be all zero with a sliding window, and then look at how this occlusion affects the probability of the true class in each of the classifiers. Figure 8 shows a visualization of the five outputs' probabilities as a function of the occluder position for two given test images as 2-dimensional heat maps.

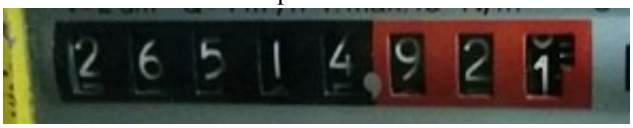
This example demonstrates that the model is not only localizing the digits within the scene, as the probability of the correct class drops significantly when the corresponding digit is occluded, but also that the surrounding area of the digit has a critical contribution to the network

Output: 00030

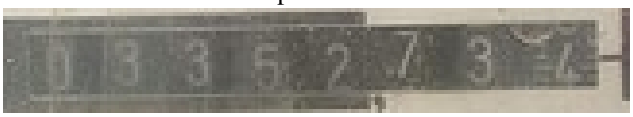
Output: 5918

Output: 00782

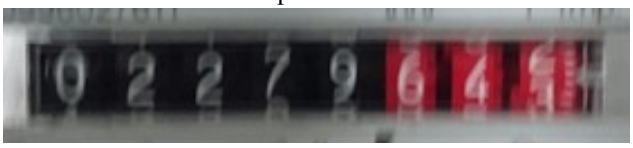
Output: 01412



Output: 26514



Output: 03352



Output: 02279

Output: 06602

Figure 7. Examples of correct readings in complicated cases. We show a cropped version of the reading area for visualization purposes and to conceal private meter information. The full input images are similar to the ones shown in Figures 1, 2, and 3.

predictions. In particular, we see how the four leftmost digits predictions are pretty sensitive to the left edge of the meter's reading area, while the predictions for the last two digits have a more sensible area at their right side, presumably because they have to figure out where is the significant/non-significant digits' boundary.

V. Conclusions

In this work we explored the possibility to perform true, segmentation-free, end-to-end reading in images, with a particular application to reading utility meters in natural scene images. The proposed CNN architecture can be trained in an end-to-end manner, and is able to directly output readings without any explicit text localization step, while it inherently learns to filter out non-significant digits present on the meter. The obtained results in a large dataset of images acquired by different users and devices demonstrate that the proposed system is able to correctly read the meter's measurements with high accuracy in real-time. Further analysis, reveal that such architectures are capable of not only implicitly localise areas of interest for each digit, but efficiently use the contextual information available.

We plan to extend this work in the future with more aggressive data augmentation strategies in order to cope with small meters, severe perspective distortion, and large translation variance. We also consider adding an objectness score output to the network so it can learn when a meter is present and well centered in the input image, providing thus a rejection option to the system.

Acknowledgments

This work was supported by the Spanish projects TIN2014-52072-P and TIN2017-89779-P and by the CERCA Programme / Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] K. Sayre, "Machine recognition of handwritten words: A project report," *Pattern Recognition*, vol. 5, no. 3, pp. 213–228, September 1973.

[2] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 1457–1464.

[3] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," in *arXiv preprint arXiv:1412.1842*, 2014.

[4] L. Gómez and D. Karatzas, "Textproposals: A text-specific selective search algorithm for word spotting in the wild," *Pattern Recognition*, vol. 70, pp. 60–74, October 2017.

[5] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Proceedings of the Workshop on Deep Learning, NIPS*, 2014.

[6] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4161–4167.

[7] Z. Tian, W. Huang, T. He, P. He, and Y.Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 56–72.

[8] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proceedings of the International Conference on Computer Vision*, 2017.

[9] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3538–3545.

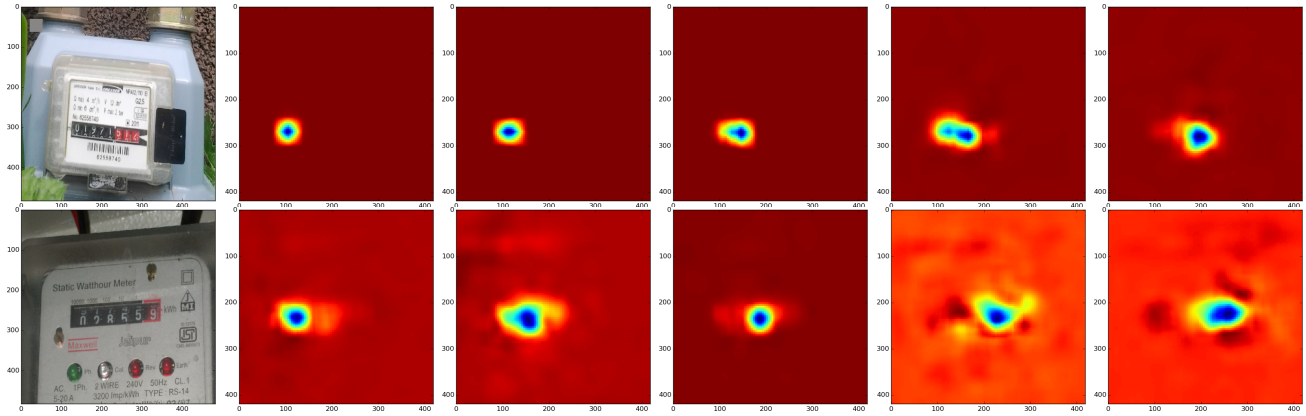


Figure 8. Visualization of the image parts responsible for the classification predictions. The five column-wise heat maps correspond to the correct class probabilities for each digit as a function of the position of a gray square occluder in the input images. Input images source: Wikimedia Commons CC-BY-SA-3.0.

[10] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2961–2968.

[11] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970–983, 2014.

[12] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[14] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[15] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," *arXiv preprint arXiv:1709.02054*, 2017.

[16] R. Smith, C. Gu, D.-S. Lee, H. Hu, R. Unnikrishnan, J. Ibarz, S. Arnoud, and S. Lin, "End-to-end interpretation of the french street name signs dataset," in *European Conference on Computer Vision*. Springer, 2016, pp. 411–426.

[17] Z. Wojna, A. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," *arXiv preprint arXiv:1704.03549*, 2017.

[18] D. Oliveira, R. Cruz, and K. Bensebaa, "Automatic numeric characters recognition of kilowatt-hour meter," in *Proceedings of the International Conference on Signal Image Technology and Internet Based Systems*, 2009, pp. 107–111.

[19] M. Cerman, G. Shalunts, and D. Albertini, "A mobile recognition system for analog energy meter scanning," in *Proceedings of the International Symposium on Visual Computing*, 2016, pp. 247–256.

[20] A. Nodari and I. Gallo, "A multi-neural network approach to image detection and segmentation of gas meter counter," in *Proceedings of the Conference on Machine Vision Applications*, 2011, pp. 239–242.

[21] M. Vanetti, I. Gallo, and A. Nodari, "GAS meter reading from real world images using a multi-net system," *Pattern Recognition Letters*, vol. 34, no. 5, pp. 519–526, April 2013.

[22] M. Chouiten and P. Schaeffer, "Vision based mobile gas-meter reading. machine learning method and application on real cases," in *Proceedings of the International Workshops on Electrical and Computer Engineering Subfields*, 2014, pp. 94–97.

[23] Z. Cai, C. Wei, and Y. Yuan, "An efficient method for electric meter readings automatic location and recognition," *Procedia Engineering*, vol. 23, pp. 565–571, 2011.

[24] I. Gallo, A. Zamberletti, and L. Noce, "Robust angle invariant GAS meter reading," in *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, 2015.

[25] H. Puttnies, V. Altmann, F. Golatowski, and D. Timmermann, "Cost-efficient universal approach for remote meter reading using web services and computer vision," in *Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks*, 2015.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[27] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.