USDOT Number Localization and Recognition From Vehicle Side-View NIR Images

Orhan Bulan, Safwan Wshah, Ramesh Palghat, Vladimir Kozitsky and Aaron Burry
Palo Alto Research Center (PARC)
800 Phillips Rd. Webster NY 14580
orhan.bulan,safwan.wshah,ramesh.palghat,vladimir.kozitsky,aaron.burry@parc.com

Abstract

Commercial motor vehicles are mandated to display a valid U.S. Department of Transportation (USDOT) identification number on the side of the vehicle. Automatic recognition of USDOT numbers is of interest to government agencies for the efficient enforcement and management of the commercial trucks. Near infrared (NIR) cameras installed on the side of the road, to capture an image of an incoming truck, can capture USDOT images without distracting the drivers. In this paper, we propose a computer vision based method for recognizing USDOT numbers using an NIR camera system directed at the side of the commercial vehicles. The developed method consists of two stages; first, we localize the USDOT tag in the captured image using the deformable part model (DPM). Next, we train a convolutional neural network (CNN) using street-view house number (SVHN) dataset[1] and sweep the trained classifier across the localized region. Based on the calculated scores, we infer the digits and their locations using a probabilistic inference method based on Hidden Markov Models (HMM). The most likely digit sequence is determined by applying the Viterbi algorithm. A data set of 1549 *images was collected on a public roadway and is used to perform the experiments.*

1. Introduction

The Federal Motor Carrier Safety Administration (FMCSA) requires that an active and valid USDOT identification number must be properly displayed on both sides of the vehicle where the identification number is preceded by the letters "USDOT" [1]. Unlike license plates, USDOT numbers are assigned to vehicle-owners rather than vehicles so that commercial vehicles that belong to the same person/company have the same identification number to enable collecting and monitoring company's safety information acquired during audits and inspections [1].

[1]The SVHN data set is restricted for non-commercial use only and is no way used in a commercial product.

Several transportation management entities are interested in automated recognition of USDOT numbers using NIR cameras mounted at the road-side at weigh/inspection stations. The motivation for transportation management entities to read USDOT numbers is two-fold: a) to ensure rapid validation of vehicle credentials and, b) to pre-populate vehicle and driver information at a inspection/weigh station. The reason for USDOT recognition in addition to license plate recognition (LPR) is to increase the recognition accuracy by fusing the results from LPR and DOT recognition systems. Reading USDOT number reduces the pool of eligible LPR numbers which, in turn, increases the overall accuracy and reduces congestion at the inspection/weigh stations.

A common practice to locate and recognize USDOT numbers in captured images is by using Optical Character Recognition (OCR) engines to first identify the text regions in the image and then reading the "USDOT" tag in the text regions using an OCR engine. OCR-based localization, however, yields low detection performance due to the fact that the USDOT NIR images are captured under a variety of illumination conditions (e.g., day and night time, different weather conditions etc.) and can be noisy and have low contrast. Using pre-trained OCR engines (e.g., Tesseract) for recognizing the USDOT number after localization is also challenging given the poor image quality.

In this paper, we propose an end-to-end computer vision based method for recognizing USDOT numbers from vehicle side-view images captured by a NIR camera, which is commonly used in transportation imaging systems to ensure that the illumination does not distract drivers. Our algorithm for USDOT number recognition consists of two stages: first, we utilize an elastic deformation model [17, 3] for localizing the USDOT tag in the captured image. Next, we train a CNN using SVHN dataset [9] and sweep the trained classifier across the localized USDOT number region. Based on the calculated scores, we infer the digits and their locations using a probabilistic inference method based on Hidden Markov Models (HMM). The most likely digit

sequence is determined by applying the Viterbi algorithm. We evaluated the performance of the proposed method on a dataset of 1549 images collected on a public roadway.

Recognizing text from natural scene images has been considered in several studies in the literature and is still an active research area [5, 15, 16, 10, 11, 2, 12, 4]. Some of the methods in the literature leverages the language model to constrain word recognition problem [15, 16], which is not feasible in the USDOT number recognition. Several others first perform a segmentation of the characters and performs classification to recognize the text [2, 12]. The character segmentation is, however, highly challenging with poor image quality (e.g., noisy, low contrast etc.). In order to address recognition in poor quality images, sequential character recognition using, for example, a sliding window search has been exploited [15, 16]. In a very recent work, a hybrid approach is proposed for segmentation and recognition using CNN and HMM model to recognize the street-view house numbers [5].

Our first contribution in this paper is posing a new problem of recognizing USDOT numbers from vehicle side-view images captured by NIR cameras and proposing an end-to-end system for the USDOT number recognition problem. Our second contribution is localizing the USDOT number in the captured NIR image by detecting the USDOT tag using a deformable part model. Unlike the classical localization approaches based on OCRs, which is challenging in low quality images, the proposed localization algorithm can accurately detect the USDOT tag even in noisy and low constrast images. Our third contribution is applying a "transfer learning" where CNN classifier is trained using SVHN dataset and applied in a different domain to eliminate data gathering and manual annotation required for training the classifier.

The organization of this article as follows. Section 2 briefly summarizes the image acquisition for capturing USDOT images. In Sec. 3, we describe the details of our methodology for localizing USDOT number from captured NIR images using the deformable parts model. Sec. 4 presents our method for recognizing USDOT number from the localized regions. Evaluation of the methods using real world road images are presented in Sec. 5. Sec. 6 discusses the key findings of the present study.

2. Image Acquistion

A road-side NIR camera is installed and used to collect a dataset of 1549 USDOT images for both day and night time. Figure 1 shows samples of images of the side of a truck with high and low contrast. The particular images not only contain the US DOT number at the bottom of the door, but the vehicle identification number in the center and the name of the trucking company above that. Other trucks may contain additional information such as the weight of the truck and the location of the trucking company. A full image of the truck may also contain text in the form of marketing signage, and information that identifies the function of the truck

Figure 1. USDOT image samples with high contrast (a) and low contrast (b) acquired by an NIR camera.

The USDOT number is written on the side of the truck following the USDOT tag. The USDOT tag and number can be written with a variety fonts and sizes on the side of the truck. This variation, in addition to other source of noises, low contrast and illumination variation, poses a challenge on the recognition accuracy. Figure 2 illustrates several examples of cropped USDOT number images with variety of fonts and sizes. Some of these images are very noisy since captures of the side of the truck are not always performed under optimal conditions. Imaging at night in an ongoing traffic requires a NIR camera with low exposure time causing images to be noisy and low contrast. There are also other variations that mitigate the recognition performance such as dark on light or light on dark, embossing, glare, extra characters, blur etc. All these variations in turn have a negative impact on the recognition accuracy. So the localization and recognition system has to account for this variation in the operational phase.

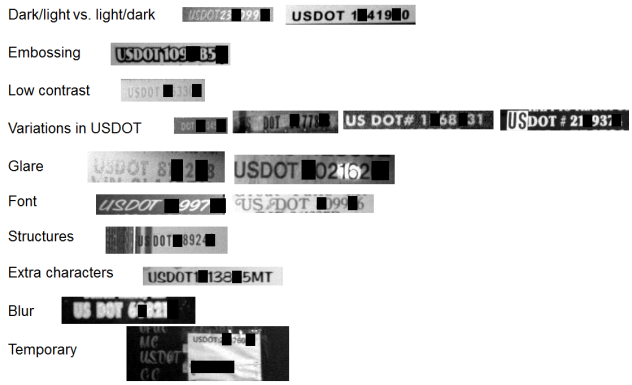

Figure 2. USDOT number images with a variety of fonts and quality.

3. USDOT Number Localization

Figure 3 shows an overview of our end-to-end methodology for USDOT number recognition. For localizing the USDOT number, we first detected the USDOT tag in the captured image as the number is typically preceded by the tag. Specifically, we train an appearance-based model to

detect and locate USDOT tag in a given vehicle side-view image using a deformable part model where the score for a particular configuration of landmarks L in a given image I is defined as

$$S(I, L) = App(I, L) + Shape(L) \quad (1)$$

In this function, $App(I, L)$ leverages the histogram of oriented gradients (HoG) features extracted at each pixel location. The appearance evidence of each of the landmarks is included in the App and the evidences for the spatial locations of the landmarks with respect to each other is included in the $Shape$ term. In [3], this model was viewed as a linear classifier with unknown parameters, which is learned during training using a latent SVM. The training constructs a model by learning the appearance at each landmark point and the relationship between points.

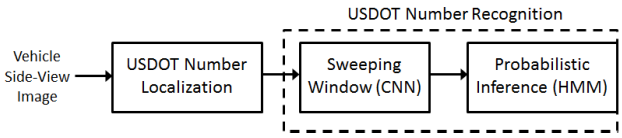

Figure 3. An overview of the methodology for recognizing USDOT numbers from side-view NIR images.

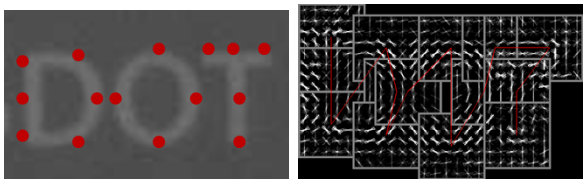

Figure 4. Landmarks located on a USDOT tag (a) and the trained model using the located marks (b).

We note that USDOT number is sometimes preceded by DOT instead of USDOT tag. In order to capture this variation, we trained our model for DOT, which is included in both tags. For this purpose, we located 15 landmarks in our positive samples to model the appearance of DOT/USDOT tag. The red dots in Fig. 4 show the located landmarks and the resulting DPM model. The number of landmark points can be adjusted based on the amount of unique features to be included. For example, for detecting faces in different poses in an image, more than 30 landmark points are needed in order to include unique facial features such as nose and eyes. In DOT/USDOT tag localization, the edges are mostly straight lines or curves. Hence, there is less benefit to include many landmark points as increasing number of landmark points can significantly increase the amount of manual work in the training phase and computational cost in the online application. Since the relationships between landmark points are processed through dynamic programing, the end points of the sequence of points cannot be connected. The choice of where to position the end points can potentially affect the performance and thus must be done with care. Also note that in our specific implementation, we used only one mixture to develop the model but more than one mixture can also be used to capture different appearances of the USDOT tag.

For an incoming image I, we identify a list of candidate "USDOT" tag regions by maximizing the score function Eq. (1) over L using dynamic programming to find the best configuration of parts [17, 3].

$$S^*(I) = \max_L S(I, L) \quad (2)$$

Once the location of the "USDOT" tag is detected, the location of the USDOT number can be determined with respect to the location, size, and aspect ratio of the detected tag. In vast majority of the cases, the USDOT number is located on the right side of the DOT/USDOT tag and has the same size and aspect ratio as the USDOT tag. After plotting the histograms of USDOT number locations/sizes with respect to the USDOT tag locations/sizes, we observe that histograms form approximate Gaussian distributions with respect to the USDOT tag sizes and locations. Based on the observed histograms, an image patch next to the detected USDOT/DOT tag whose height and width is determined by the size of the detected tag is determined as the location of the USDOT number. Figure 5 shows sample image patches automatically cropped from a region next to detected USDOT tags based on the observed histograms and after applying low level image processing for tight cropping.

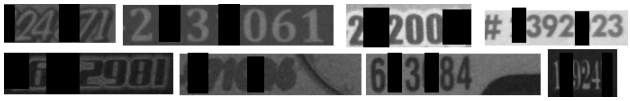

Figure 5. Localized USDOT numbers.

4. USDOT Number Recognition

As shown in Fig. 5, the localized USDOT numbers show a large variation in fonts, which requires that the training data should capture this variation. We used SVHN dataset in the training for each digit (i.e. 0 to 9) and trained a CNN classifier.

CNNs have shown outstanding image classification performance in many fields [8]. The success of CNNs is attributed to their ability to learn rich mid-level image representations as opposed to hand-designed low-level features used in other image classification methods [13]. Recently, image representations learned with CNNs have been efficiently transferred to other visual recognition tasks with limited amount of training data [7]. In the problem of recognizing USDOT numbers, the data labeling is very expensive. This paper showed that transferring the knowledge

of CNN models learned on similar dataset such as SVHN to recognize the USDOT numbers works very well. In our implementation, the CNN classifier is trained using the network described in [8]. We applied the method described by Jarret et al. [6] of using locally normalizing sets of internal features, at each stage of the model. And the use of smoother pooling functions, such as the L2 instead of max-pooling [14] showed better results. To minimize the false positive rate we trained four different CNN models at different orientations (0,90,180,270) where each model was trained separately and final score was calculated by averaging over the four classifiers.

After the classifier is trained, a fixed window is swept across a localized USDOT image and the classifier is applied to the image at each window location. The result is a matrix of character confidences, for each digit, at each window location. The vertically cropped USDOT image is resized to a height of 32 pixels and a 32×32 pixel window is swept across it. This window size appears to be well suited for localized USDOT images. Figure 6 illustrates the sweeping window process. The figure shows a window (in red) being swept across the localized USDOT number. We also plot the maximum score at each window location. The peaks in the plot represent the set of candidate digits.

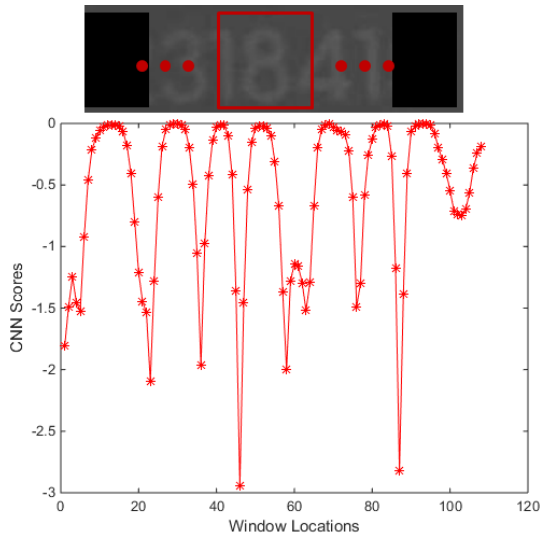

Figure 6. Sliding window OCR.

For decoding the USDOT number for a given sequence length N, we formulate the problem as

$$\mathbf{c}^* = \underset{\mathbf{c}}{\operatorname{argmax}}\, p(c_1, c_2, ..., c_N, x_1, x_2, ..., x_N) \qquad (3)$$

where $c_i \in 0, 1, ..., 9$'s represent possible digits from the digit set and x_i's represent the corresponding digit locations. Modeling the multidimensional density function, especially in the absence of physically inspired model, constitutes a hard task. We therefore, use HMM to model the problem of finding the highest probability sequence and simplify the joint density function in Eq. 3. Figure 7 shows a schematic of the HMM where A represents the transition matrix and O the emission matrix. $A_{j,i}$ represents the transition probability to go from character c_j at x_j to character c_i at x_i. O_i is the OCR probability for character c_i at location x_i. In the HMM formulation, the optimization problem in Eq. 3 reduces to

$$\mathbf{c}^* = \underset{\mathbf{c}}{\operatorname{argmax}} \prod_{i=1}^{N} p(c_i|x_i)p(x_i|x_{i-1}) \qquad (4)$$

where $p(c_i|x_i)$ and $p(x_i|x_{i-1})$ represent the emission and transition probabilities, respectively, which need to be modeled to find the highest probability sequence. We modeled the transition probability as a function of the digit spacing $(x_i - x_j)$. Even though the digit to digit spacing within a localized USDOT image shows variation for some digits (e.g., the spacings between 1 and the other digits are usually larger) it typically has a pre-dominant frequency (i.e., corresponding to the average digit spacing) that can be estimated by performing a fast Fourier transform (FFT) analysis on the calculated CNN scores. Following the calculation of average digit spacing T, we modeled the transition probability as a step function where the probability is set 0 if the digit spacing deviates more than $T/2$ from the average spacing (T) between the digits. Using a step function instead of a Gaussian distribution for the transition probability yielded better results as the digit spacing can largely vary in USDOT numbers between different digits.

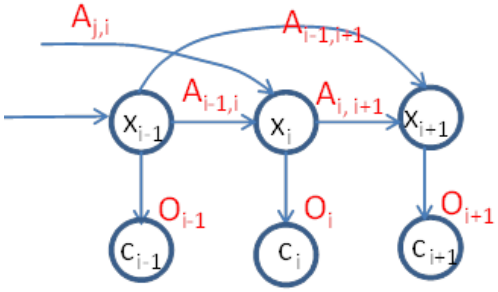

Figure 7. Graphical representation of HMM model for USDOT number decoding.

We next calculate the emission probabilities $(p(c_i|x_i))$ from the calculated CNN scores. We observe that the CNN scores can vary between 0 to -25 where higher CNN score means higher confidence for the estimated digit. We also note that CNN scores infer different confidences for different digits. In order to address this variation we calculated the CNN scores for the validation samples in the SVHN dataset and fit exponential distributions for each digit based on the calculated scores. Figure 8 shows the exponential distributions fitted for each digit given the CNN scores of

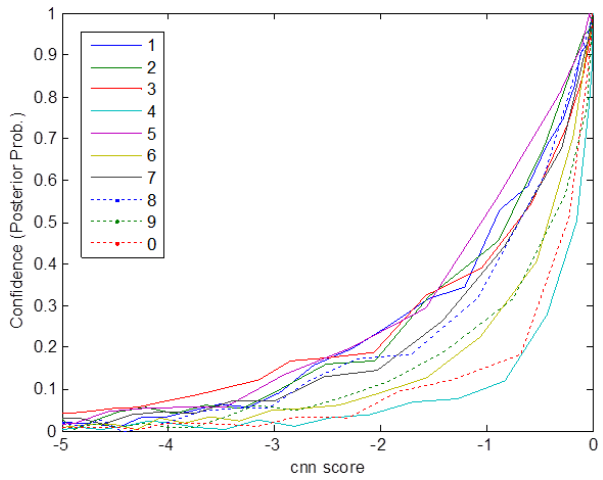

Figure 8. Posterior probabilities for different digits versus CNN scores.

the SVHN dataset. Based on these distributions, we normalize the CNN scores of the USDOT numbers and calculate the emission probabilities for each digit.

Given the emission and transition probabilities, we decoded the highest probability digit sequence using Viterbi algorithm, which leverages dynamic programming to solve the optimization problem in Eq. 4. Note that our problem formulation for finding the best sequence is for a given code length N. The typical length of USDOT numbers change from 5 to 8 with the majority of them having a length of 6 and 7 (e.g., $> 80\%$ in our USDOT dataset). The Viterbi algorithm can be repeated for different code lengths and the best code can be selected from the codes returned by different code lengths.

5. Experiments

In this section, we evaluated the performance of the proposed algorithm for USDOT number localization and recognition. The algorithm is implemented in Matlab and tested on a set of images acquired by the image acquisition system as described in Sec. 2.

We have conducted our experiments on a database of USDOT images acquired from a real-world application setting. In our set, we had 1549 USDOT images captured during both day-time and night time. The resolution of the captured images was 2448×1248 and the size of the height of the USDOT tag was changing between 20 to 100 pixels. Figure 1 shows two sample USDOT images in our database captured during daytime, where the size, font and contrast of the USDOT tags/numbers vary.

5.1. Localization

We first trained an appearance model for DOT tag using 15 landmarks located around the tag. In our training, we used 100 positive and 100 negative samples. Figure 9 shows the USDOT tags used for training the appearance model. For each positive sample, we manually located the landmarks and trained our model using a linear SVM. The model was then tested on 1549 test images and we observed detection scores. In 90% of the images, our model correctly located the USDOT tag as the highest score window. In 5% of the images, the proposed method identified the USDOT tag as the second highest window and 2% of the cases the USDOT tag was identified as the third highest score window. This provided significant improvement over the OCR-based localization approach.

Figure 9. USDOT tags used for training the appearance model.

5.2. Recognition

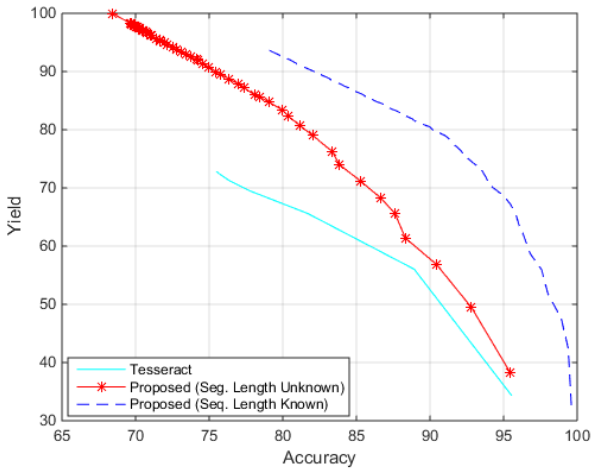

Figure 10. Accuracy-yield curves for the proposed probabilistic inference approach for USDOT number decoding using CNN features and its comparison with Tesseract.

We next evaluated the performance of the proposed probabilistic inference approach on the localized USDOT images. In Fig. 10, the proposed approach for USDOT number decoding is compared with Tesseract on the localized images. In order to make a fair comparison, we limited the dictionary to only digits for Tesseract. For the proposed ap-

proach, we plotted two curves in the figure. The dotted red curve corresponds to the case when the length of the digit sequence is unknown. In this case, the Viterbi decoding is repeated for each code length and the best code is selected among the codes returned for each code length using heuristics based on the code probability. The dashed blue curve represents the case assuming the code length is known and Viterbi decoding is performed only for the given code length to find the highest probability code sequence. This case forms an upper bound on the Viterbi decoding when the code length is unknown. As is evidenced by Fig. 10, the proposed approach vastly outperforms Tesseract, especially in the high yield region. We expect that the performance of the unknown code length case can be further improved by (a) rejecting false characters that show up as peaks in the CNN scores by introducing a null character in the OCR training set to represent background and half characters in a manner similar to [5], and (b) training a classifier to pick the best code length.

Figure 11. Performance of the proposed USDOT number recognition framework. The images in the first row presents the wrong recognitions due to the characters circled. The images in the second row shows accurately recognized USDOT numbers. The contrast of the images is enhanced for visualization purposes.

Figure 11 presents the USDOT number recognition results for several sample images. The proposed probabilistic inference approach has correctly recognized the USDOT numbers in the second row. The images in the first row shows the wrong recognitions due to the characters circled. Note that the errors in the first, second and fourth image in the first row is due the confusion between the digits 1 and 7. The error in the third image is because of the extra hash tag character at the beginning of the digit sequence.

6. Conclusion

In this paper, we pose a new problem for recognizing USDOT numbers from vehicle side-view images captured by NIR cameras. The image regions where the USDOT number is located, can be extracted by generating an appearance model for USDOT tags. The appearance based localization can provide improvement over an OCR-based localization approach especially for the poor quality images in the existence of noise and low contrast. A CNN classifier trained using samples from SVHN dataset eliminates data gathering and manual annotation required for training the classifier for USDOT numbers. A probabilistic inference based on HMM and Viterbi decoding for finding the highest probability digit sequence outperforms Tesseract for USDOT number recognition.

References

[1] http://www.wisconsinlimo.org/public/news/wlanews2005october.pdf. U.S. DOT Number Regulation. 1

[2] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. Photoocr: Reading text in uncontrolled conditions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 785–792. IEEE, 2013. 2

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 1, 3

[4] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013. 2

[5] Q. Guo, D. Tu, J. Lei, and G. Li. Hybrid cnn-hmm model for street view house number recognition. In *Asian Conference on Computer Vision (ACCV) workshop on deep learning on visual data*. Singapore, 2014. 2, 6

[6] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2146–2153. IEEE, 2009. 4

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 3

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. 3, 4

[9] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5. Granada, Spain, 2011. 1

[10] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In *Computer Vision–ACCV 2010*, pages 770–783. Springer, 2011. 2

[11] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3538–3545. IEEE, 2012. 2

[12] L. Neumann and J. Matas. Scene text localization and recognition with oriented stroke detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 97–104. IEEE, 2013. 2

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, 2014. 3

[14] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3288–3291. IEEE, 2012. 4

[15] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1457–1464. IEEE, 2011. 2

[16] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012. 2

[17] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 1, 3