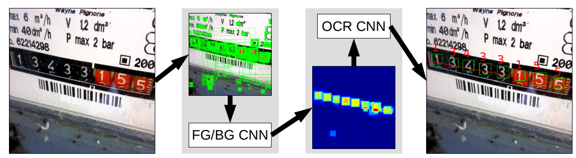
Augmented Text Character Proposals and Convolutional Neural Networks for Text Spotting from Scene Images

Alessandro Zamberletti, Ignazio Gallo, Lucia Noce
University of Insubria,
Department of Theoretical and Applied Science,
Via Mazzini, 5, 21100 Varese, Italy
{a.zamberletti, ignazio.gallo, lucia.noce}@uninsubria.it

Abstract

In this work we propose a novel method for text spotting from scene images based on augmented Multi-resolution Maximally Stable Extremal Regions and Convolutional Neural Networks. The goal of this work is augmenting text character proposals to maximize their coverage rate over text elements in scene images, to obtain satisfying text detection rates without the need of using very deep architectures nor large amount of training data. Using simple and fast geometric transformations on multi-resolution proposals our system achieves good results for several challenging datasets while also being computationally efficient to train and test on a desktop computer.



a) Test image b) Proposals c) Text localization d) Text reading

Figure 1: A visual overview of the proposed model. Given a test image (a), augmented proposals (b) are extracted and processed by a CNN to build a text localization map in which potential areas of interest are highlighted (c). High intensity regions from the text confidence map are further processed to recognize text elements of interest (d).

1. Introduction

Text localization and recognition (text spotting) from scene images and digital documents is an interesting task that finds applications in multiple commercial areas where automated systems can replace human workers in carrying out tedious repetitive data entry tasks.

In the last few years, researchers were able to obtain new state-of-the-art results for text spotting from scene images; however, recent state-of-the-art algorithms are often difficult to reproduce as they use very deep architectures [1] and/or large datasets which sometimes are not publicly available due to copyright restrictions [2].

In this manuscript, instead of focusing our attention on increasing either the deepness of the text localization and recognition classifiers, or the amount of labeled training data, we optimize the data that is fed to the proposed model by maximizing the detection recall of multi-resolution text character proposals extracted from scene images.

Initially, we tackle the problem of reading analogic flow meters from natural images, showing that two slight variants of LeNet [3], trained solely on augmented Multi-resolution Maximally Stable Extremal Regions (MR-MSER) [4], can reach nearly human detection accuracies and fast recognition times.

We then incrementally prove the generality of the proposed method by applying it to the task of license plate recognition [5, 6] and unconstrained text localization from scene images [7, 8, 9], obtaining state-of-the-art results for the first and competitive performances for the second.

For all the evaluated datasets, replacing augmented proposals with their respective non augmented versions leads to a dramatic reduction in terms of detection rates.

2. Related Works

2.1. Text Localization and Recognition

Algorithms for text localization and recognition can be classified as either region based [10] or connected component based [11, 12]. Region based methods exploit local features and sliding window classifiers to localize and read potential text components, while connected component based methods analyze the geometric properties of proposals extracted from the processed images and identify the ones corresponding to text characters.

Region based methods are prone to errors as some local windows in scene images are indistinguishable from text characters; while the performances of connected component based methods strongly depend on the ability of capturing text characters as individual proposals.

To overcome these previously mentioned limitations, multiple hybrid approaches have been proposed.

In [1], a single very large CNN is used for integrated text localization and recognition of Street View House Numbers and CAPTCHAs, thus removing the need for using local windows or proposals. While this approach seems promising, it can only be applied to text sequences whose length is known a priori, and a large amount of training data is required to obtain acceptable results.

In [9], multiple very large CNN, trained solely on synthetic data, are used to localize and read text word proposals from Edge Box and ACF detector. Even though this latest approach is similar to the one we propose, we work at text character level using augmented MR-MSER proposals in place of synthetically generated training data.

2.2. Text Character Proposals

Since text characters usually show uniform color characteristics, Maximally Stable Extremal Regions (MSER) [13] are widely used in literature as proposals for text characters in scene images [11, 12].

In order to maximize the coverage provided by MSER proposals over text regions, multiple variants of the original MSER algorithm have been proposed: M-CHN MSER [11], MR-MSER [4], and E-MSER [12], to name a few.

Generic object proposal/detection methods (Edge Box, Selective Search, ACF, *etc.*) have also been recently used in place of MSER variants for text proposal generation [9].

In this work, we compare most of the previously mentioned proposal approaches, showing that our proposal augmentation technique (Fig. 2) can significantly boost detection recall (number of ground-truth text character annotations covered by at least one generated proposal) for all the evaluated datasets.

3. Proposed Method

3.1. Text Localization

The proposed text localization pipeline is visually summarized in Fig. 3. Given a test image, MR-MSER are computed as in [4]: MSER proposals are extracted at each level of a scale pyramid, which has 1 octave per scale and a total of 3 scales. Ulike [4], no Gaussian smoothing is applied between octaves; Δ parameter is set to 3 to maximize the number of extracted proposals. On average, 8k MR-MSER proposals are extracted from a 640×480 rgb scene image.

The idea behind the use of MR-MSER is that unstable text regions in the original image may become stable at

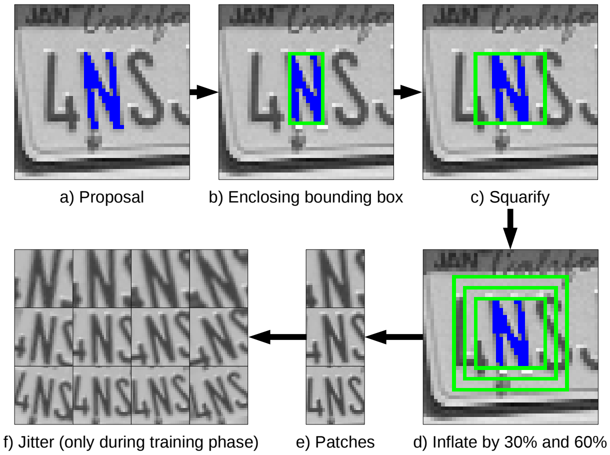


Figure 2: Proposal augmentation pipeline. Given a proposal and its bbox (a,b): the bbox is *squarified* without moving its center (c); two additionals bboxes are obtained by inflating the *squarified* bbox by 30% and 60% of its area (d); resulting patches are randomly rotated within $[-\frac{\pi}{4}, \frac{\pi}{4}]$ (e,f).

lower scales in the pyramid, where most image details are lost and colors are merged together [4].

To increase detection recall of MR-MSER proposals over text regions in the processed images, we adopt the augmentation pipeline described in Fig. 2 : (i) the original proposal is *squarified* without moving its center, (ii) neighboring text characters and background noise are captured by inflating the *squarified* proposal by 30% and 60% of its area in every dimension, (iii) the *squarified* original proposal, together with its inflated variants, are resized to 28×28 pixel.

Given a single MR-MSER proposal, a total of 3 augmented variants are generated; this provides us with roughly 24k image patches per image that need to be classified as either containing text characters of interest (foreground - FG) or noise (background - BG).

The task of classifying image patches as belonging to either FG or BG is approached using a slight variant of LeNet [3]. The proposed architecture has a total of three convolutional hidden layers with $[128, 256, 512]$ units each, and two fully connected layers containing 512 units. Max pooling with 2×2 window size is performed after each convolutional step. Kernel size and stride are fixed to respectively 3 and 1 for all the convolutional layers. The final classification is performed using Softmax.

The network is trained using augmented MR-MSER proposals extracted from images from the given training dataset. Positive samples are obtained by selecting augmented MR-MSER proposals having Intersection-over-union $IoU > 0.5$ with at least one ground-truth text character annotation. An equal amount of negative training patches ($IoU = 0$ for every ground-truth text character an-

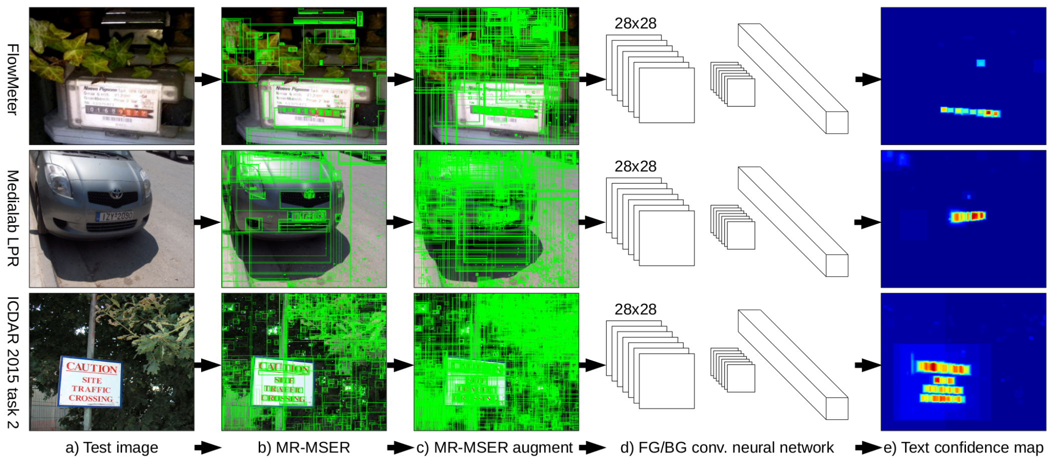


Figure 3: Text localization pipeline. MR-MSER proposals extracted from the given test images (a,b) are augmented (c) and processed by the CNN (d). FG/BG prediction values are stacked together to form text confidence maps (e).

Table 1: Implementation details. Times refer to a 640×480 rgb image processed on a Intel Xeon E5-1620 at 3.5 GHz, NVIDIA GTX 980, and C++ Caffe Deep Learning library.

Task	Time (ms)	Comp.	# of prop.
MR-MSER (loc.)	14.4	CPU	8k
Prop. augment (loc.)	2.80	CPU	24k
FG/BG CNN (loc.)	47.2	GPU	1k
OCR CNN (read)	13.8	GPU	< 100

notation) are randomly selected.

Since the network is relatively small (10MB SDRAM), we can apply on-line jitter to the training patches while maintaining acceptable training speeds (≈ 1000 sample/s). As shown in Fig. 2, each training patch is randomly rotated four different times within $[-\frac{\pi}{4}, \frac{\pi}{4}]$ radians, thus, starting from roughly 24k augmented patches we generate an average of 96k randomly rotated patches per training image.

As in [10], confidence values provided by the CNN for each augmented MR-MSER proposal are stacked together to build a text confidence map in which high intensity regions denote potential text components of interest.

As shown in Fig. 3 and Table 1, the proposed localization pipeline works well for heterogeneous images, and requires on average 64.4 ms to be completed on a GTX 980 GPU.

3.2. Text Reading

The proposed text reading pipeline is visually summarized in Fig. 4. Given the normalized text confidence map produced by the previously described text localization step, we gather augmented MR-MSER proposals representing potential regions of text as follows: (i) each proposal is assigned a score computed as the average intensity of its pixels in the text confidence map, and (ii) proposals with score higher than 0.9 are considered potential regions of text.

Since flow meter and license plate images contain single text lines of interest, to discard additional non-text proposals we compute the best fit line for the data using Weighted Linear Least-squares over proposal centers and scores, and remove proposals that do not overlap that line. This routine cannot be used for ICDAR images as they may contain multiple lines of text (we only use threshold on ICDAR).

Non discarded proposals (roughly 1k per image) are then processed by a CNN that performs OCR and assigns each of them a digit/letter and a confidence value. The network has the same architecture of the one used for text localization, and it is trained using the same data gathering and on-line jitter techniques. Non-maximum Suppression (NMS) is finally performed over proposal confidence values; NMS overlap threshold is set to 0.1 IoU to discard nested proposals generated by our augmentation technique.

Text reading requires 13.8 ms on a GTX 980 GPU.

4. Experiments

4.1. Datasets

The proposed method has been evaluated using the following three datasets: FlowMeter, Medialab LPR [5] and ICDAR 2015 task 2 [7].

FlowMeter contains 6050 train and 168910 test scene images of gas flow meters. All the images were acquired using smart phones, and typically contain non-horizontal flow meters as well as difficult light conditions, lack of focus, motion blur, reflections, gravel on the digits, *etc.*

Medialab LPR contains 680 scene images of car license

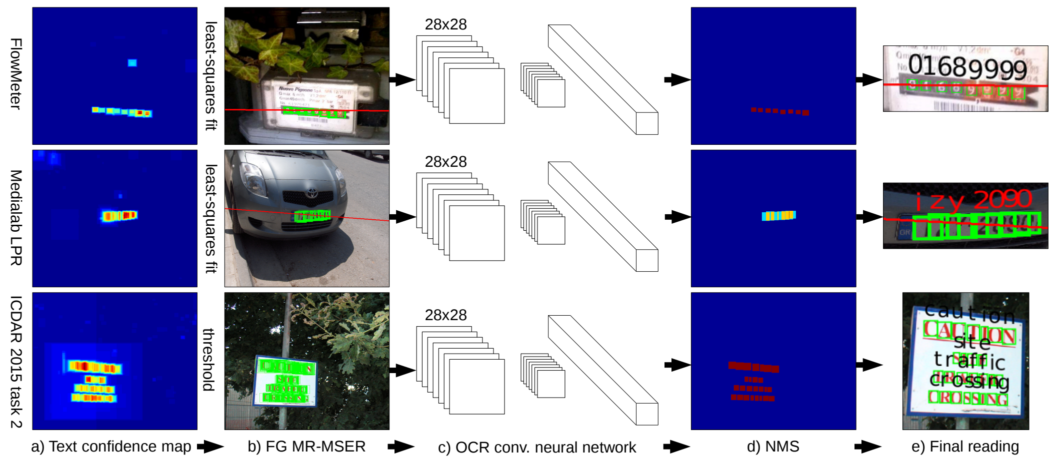


Figure 4: Text reading pipeline. Proposals overlapping potential regions of text in text confidence maps (a,b) are OCRed by the CNN (c). Non-max Suppression (NMS) is performed over CNN prediction scores (d) to obtain the final readings (e).

plates, obtained by merging all the collections from Medialab website (as in [5, 6]). [1] Similarly to competing methods, none of those images were used for training our model; instead, we used a total of 790 manually tagged training images from Zemris DB and UCSD Car LPR datasets. [2]

ICDAR 2015 Task 2 contains 229 train and 233 test scene images of focused text, it has been the reference dataset for text localization for the last decade due to its difficulty and large number of competing approaches. [3]

4.2. Results

Evaluation results for FlowMeter, Medialab LPR and ICDAR 2015 Task 2 datasets are listed in Tables 2, 3 and 4 respectively. Results for Tables 2 and 3 are measured using Sequence Transcription Accuracy metric [1], namely the rate of test images for which the predicted sequence of numbers/letters matches their respective ground-truth data.

Recall (R), Precision (P) and Hmean for Table 4 are measured using DetEval evaluation tool [7].

The proposed method achieves nearly human performances for FlowMeter dataset, state-of-the-art results for Medialab LPR, and competitive results for ICDAR 2015 Task 2. For this latest dataset, unlike most competing approaches (see website), our model has been trained solely on samples gathered from the original training set.

Unsurprisingly, accuracies drop when not using augmented proposals; in fact, as shown in Fig. 5, augmented MR-MSER achieves on average 20% more detection recall on FlowMeter dataset for all the evaluated IoU values, compared with MSER [13] and MR-MSER [4].

[1] http://www.medialab.ntua.gr/research/LPRdatabase.html
[2] http://vision.ucsd.edu/belongie-grp/research/carRec/car_data.html
[3] http://rrc.cvc.uab.es/?ch=2&com=evaluation

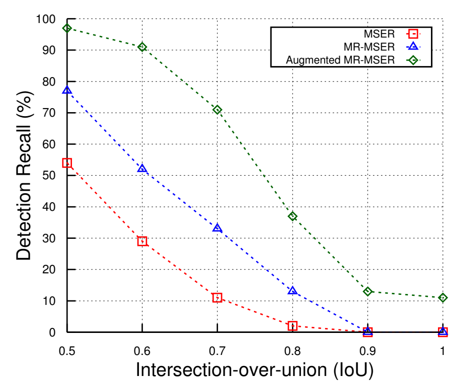


Figure 5: Text chatacter detection recall of MSER, MR-MSER and augmented MR-MSER proposals for FlowMeter dataset, while varying IoU coverage tolerance.

As also shown in Fig. 6 and Table 1, augmented MR-MSER provides the best compromise between detection recall and computational complexity among the evaluated augmented proposal algorithms [4, 11, 13].

Detection recall is measured as the percentage of ground-truth text character annotations *covered* by proposals; a text character is considered *covered* if there exists at least one proposal having $IoU > x$ with the ground-truth bounding box of that character; x varies on the horizontal axis in Fig. 5, and it is fixed to 0.5 in Fig. 6.

Using augmented proposals, CNN provides text localization/reading predictions for each text character based both on the character patch (the original proposal), and its surroundings (the inflated proposals). This is similar to processing the original image at a multi-resolution level and leads to accurate text confidence maps.

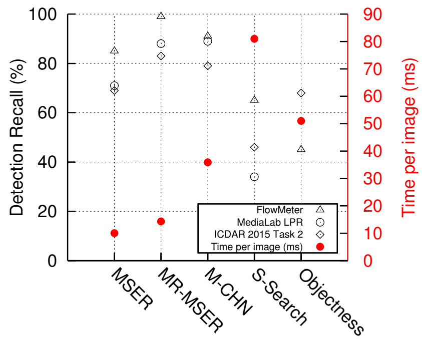


Figure 6: Text character detection recall evaluation and timings information for multiple augmented proposal algorithms [4, 11, 13]. Detection recall is computed at 0.5 IoU.

5. Conclusion

We have proposed a novel method for text spotting from scene images. Our goal was to achieve acceptable detection rates and low computational complexity; to this end, we introduced a fast geometric-based MR-MSER proposal augmentation technique which enhances detection recall of text characters in scene images. Using small LeNet variants and augmented proposals, our system localizes and recognizes text characters of interest from 640×480 rgb images in roughly 78.2 ms on a desktop machine, can be fully trained in few hours (2-8), and achieves competitive results for several challenging text spotting datasets.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce GTX 980 GPU used for this research.

Table 2: FlowMeter DB (168910 test images).

Method	Acc. (%)	Speed (img/s)
Human performance	95.1	0.08
SVM+HOG	67.4	2.10
Proposed	93.6	12.8
Proposed (no augment)	83.1	12.9

Table 3: Medialab LPR (680 test images).

Method	Acc. (%)	Speed (img/s)
Human performance	97.2	0.25
Anagnostopoulos *et al.* [5] *	86.0	3.60
Zhu [6] **	87.3	9.80
Proposed	90.2	10.0
Proposed (no augment)	83.3	10.4

* Intel Pentium IV at 3.0 GHz with 512 MB RAM.

** Intel Core 2 Duo at 2.4 GHz with 512 MB RAM.

Table 4: ICDAR 2015 Task 2 (233 test images).

Method	R (%)	P (%)	Hmean (%)
StradVision [8]	80.2	90.9	85.2
VGGMaxNet_cmb [9]	77.3	92.2	84.1
ABBYY OCR SDK v10	35.1	61.0	44.5
Proposed	67.0	83.2	74.1
Proposed (no augment)	47.9	82.4	60.5

References

[1] 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," in *Proc. ICLR*, 2014. 1, 2, 4

[2] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proc. ICCV*, 2013. 1

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 1, 2

[4] P. E. Forssén and D. G. Lowe, "Shape descriptors for maximally stable extremal regions," in *Proc. ICCV*, 2007. 1, 2, 4, 5

[5] C. Anagnostopoulos, I. Anagnostopoulos, V. Loumos, and E. Kayafas, "A license plate recognition algorithm for intelligent transportation system applications," *IEEE Trans. on ITS*, vol. 7, no. 3, pp. 377–392, 2006. 1, 3, 4, 5

[6] S. Zhu, "An end-to-end license plate localization and recognition system," 2015. RIT. 1, 4, 5

[7] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Bigorda, S. Mestre, J. Mas, D. Mota, J. Almaz, and L. Heras, "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, 2013. 1, 3, 4

[8] M. C. Sung, B. Jun, H. Cho, and D. Kim, "Scene text detection with robust character candidate extraction method," in *Proc. ICDAR*, 2015. 1, 5

[9] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Proc. NIPS*, 2014. 1, 2, 5

[10] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. ICCV*, 2011. 1, 3

[11] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. CVPR*, 2012. 1, 2, 4, 5

[12] Y. Li, W. Jia, C. Shen, and A. van den Hengel, "Characterness: An indicator of text in the wild," *IEEE Trans. on IP*, vol. 23, no. 4, pp. 1666–1677, 2014. 1, 2

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. BMVC*, 2002. 2, 4, 5