International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
Volume 6, Issue2 Pages 145-150

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

DESIGN & IMPLEMENTATION OF CONVOLUTION NEURAL NETWORKS

[1]M A Sameer, [2]M. Iliyas, [3]Azeem Hussain, [4]Aaqib Ihraz
[2] Professor, [1,3,4] Assist. Professor
[1,2,3,4] Shadan College of Engineering and Technology, Hyderabad.
[1]sameerk87@gmail.com

Abstract Full end-to-end text recognition in natural images is a challenging problem that has received much attention recently. Traditional systems in this area have re- lied on elaborate models incorporating carefully hand-engineered features or large amounts of prior knowl- edge. In this paper, we take another method and combine the representative power of large, multilayer neural networks together with recent developments in unsupervised feature learning, which allows us to use a common framework to train highly-accurate text detec- tor and character recognizer modules. Then, using only simple off-the-shelf methods, we integrate these two modules into a full end-to-end, lexicon-driven, scene text recognition system that achieves state-of-the-art performance on standard benchmarks, and popular streets

1 INTRODUCTION

Extracting textual information from natural images is a challenging problem with many practical applica- tions. Unlike character recognition for scanned docu- ments, recognizing text in unconstrained images is com- plicated by a wide range of variations in backgrounds, textures, fonts, and lighting conditions. As a result, many text detection and recognition systems rely on cleverly hand-engineered features [5, 4, 14] to repre- sent the underlying data. Sophisticated models such as conditional random fields [11, 19] or pictorial structures
[18] are also often required to combine the raw detec- tion/recognition outputs into a complete system.

In this paper, we attack the problem from a differ- ent angle. For low-level data representation, we use an unsupervised feature learning algorithm that can auto- matically extract features from the given data. Such algorithms have enjoyed numerous successes in many

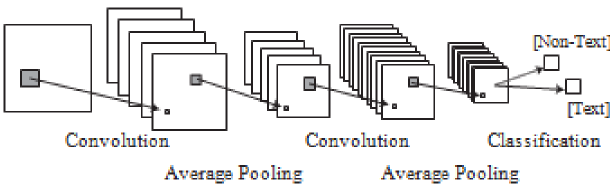

Figure 1. CNN used for text detection.

related fields such as visual recognition [3] and action recognition [7]. In the case of text recognition, the system in [2] achieves competitive results in both text detection and character recognition using a simple and scalable feature learning architecture incorporating very little hand-engineering and prior knowledge.

We integrate these learned features into a large, discriminatively-trained convolutional neural network (CNN). CNNs have enjoyed many successes in simi- lar problems such as handwriting recognition [8], visual object recognition [1], and character recognition [16]. By leveraging the representational power of these net- works, we are able to train highly accurate text detection and character recognition modules. Using these mod- ules, we can build an end-to-end system with only sim- ple post-processing techniques like non-maximal sup- pression (NMS)[13] and beam search [15]. Despite its simplicity, our system achieves state-of-the-art perfor- mance on standard test sets.

2. LEARNING ARCHITECTURE

In this section, we describe our text detector and character recognizer modules, which are the essential building blocks of our full end-to-end system. Given a 32-by-32 pixel window, the detector decides whether the window contains a

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

centered character. Similarly, the recognizer decides which of 62 characters (26 upper- case, 26 lowercase letters, and 10 digits) is in the win- dow. As described at length in Section 3, we slide the

Figure 2. Examples from our training set.
Left: from ICDAR. Right: synthetic data

detector across a full scene image to identify candidate lines of text, on which we perform word-level segmen- tation and recognition to obtain the end-to-end results.

For both detection and recognition, we use a multi- layer, convolutional neural network (CNN) similar to [8, 16]. Our networks have two convolutional layers
for detection with n1 = 96 and n2 = 256 is shown in with n1 and n2 filters respectively. The network we use (n1 = 115 and n2 = 720) is used for recognition. Figure 1, while a larger, but structurally identical one

We train the first layer of the network with an un-
supervised learning algorithm similar to [2, 3]. In par- ticular, given a set of 32-by-32 grayscale training im- ages1 as illustrated in Figure 2, we randomly extract
ZCA whitened [6] to form input vectors x(i) ∈ R64, i ∈ m 8-by-8 patches, which are contrast normalized and
{1, ..., m}. We then use the variant of K-means de-
scribed in [2] to learn a set of low-level filters D ∈
R64×n1 . For a single normalized and whitened 8-by-8
patch x, we compute its first layer responses z by per-
a scalar activation function: $z = \max\{0, |DTx| - \alpha\}$, where $\alpha = 0.5$ is a hyperparameter. forming inner product with the filter bank followed by

Given a 32-by-32 input image, we compute z for ev-
ery 8-by-8 sub-window to obtain a 25-by-25-by-n1 first layer response map. As is common in CNNs, we aver- age pool over the first layer response map to bring its dimensions to 5-by-5-by-n1. We stack another convo- lution and average pooling layer on top of the first layer to obtain a 2-by-2-by-n second layer response map. These outputs are fully connected to the classification layer. We discriminatively train the network by back-
propagating the L2-SVM classification error,2 but we
fix the filters in the first convolution layer (learned from K-means). Given the size of the networks, fine-tuning is performed using multiple GPUs.

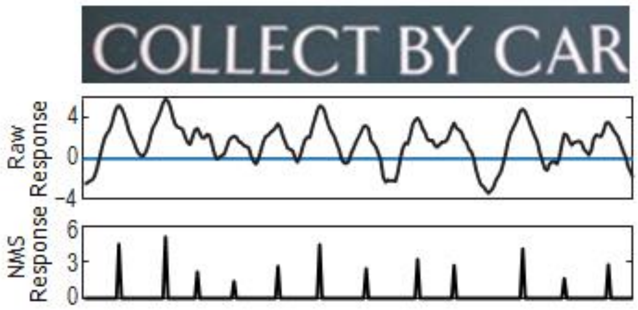

Figure 3. Detector responses in a line.

3. END-TO-END PIPELINE INTEGRATION

Our full end-to-end system combines a lexicon with our detection/recognition modules using post- processing techniques including NMS and beam search. Here we assume that we are given a lexicon (a list of tens to hundreds of candidate words) for a particular im- age. As argued in [18], this is often a valid assumption as we can use prior knowledge to constrain the search to just certain words in many applications. The pipeline mainly involves the following two stages:

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

(i) We run sliding window detection over high res- olution input images to obtain a set of candidate lines of text. Using these detector responses, we also estimate locations for the spaces in the line.

(ii) We integrate the character responses with the can- didate spacings using beam search [15] to obtain full end-to-end results.

First, given an input image, we identify horizontal lines of text using multiscale, sliding window detec- tion. At each scale s, we evaluate the detector response

Rs[x, y] at each point (x, y) in the scaled image. As

acters at the right scale produce positive Rs[x, y]. We apply NMS [13] to Rs[x, r] in each individual row r to shown in Figure 3, windows centered on single char-

estimate the character locations on a horizontal line. In

particular, we define the NMS response non-zero R˜s[x, r], we form a line-level bounding box where δ is some width parameter. For a row r with Lr with the same height as the sliding window at scale

s. The left and right boundaries of Lr are defined as min(x) and max(x), s.t. R˜s[x, r] > 0. This yields a

1Our dataset consists of examples from the ICDAR 2003 train- ing images [10], the English subset of the Chars74k dataset [4], and synthetically generated examples.

2In the form of a squared hinge loss: max{0, 1 − θTx}2 .

set of possibly overlapping line-level bounding boxes. We score each box by averaging the nonzero values of

R˜s[x, r]. We then apply standard NMS to remove all

L's that overlaps by more than 50% with another box

bounding boxes L˜. Since gaps between words produce of a higher score, and obtain the final set of line-level sharply negative responses, we also estimate possible

space locations within each Lr by applying the same NMS technique as above to the negative responses.

After identifying the horizontal lines of text, we jointly segment the lines of text into words and recog- nize each word in the line. Given a line-level bounding box L and its candidate space locations, we evaluate a number of possible word-level bounding boxes using a Viterbi-style algorithm and find the best segmentation scheme using a beam search technique similar to [9].

character recognizer across it and obtain a 62 × N score To evaluate a word-level bounding box B, we slide the matrix M , where N is the number of sliding windows

M (i, j) suggests a higher chance that the character with within the bounding box. Intuitively, a more positive index i is centered on the location of the jth window.

Similar to the detection phase, we perform NMS over

to be present. The other columns of M are set to −∞. M to select the columns where a character is most likely We then find the lexicon word w∗ that best matches a

score matrix M as follows: given a lexicon word w, compute the alignment score

Table 1. Cropped word recognition accu- racies on ICDAR 2003 and SVT

Benchmark	I-WD-50	I-WD	SVT-WD
Our approach	90%	84%	70%
Wang, *et al.* [18]	76%	62%	57%
Mishra, *et al.* [11]	82%	-	73%

4. EXPERIMENTAL RESULTS

In this section we present a detailed evaluation of our text recognition pipeline. We measure cropped charac- ter and word recognition accuracies, as well as end-to- end text recognition performance of our system on the ICDAR 2003 [10] and the Street View Text (SVT) [18] datasets. Apart from that, we also perform additional analysis to evaluate the importance of model size on dif- ferent stages of the pipeline.

First we evaluate our character recognizer module on the ICDAR 2003 dataset. Our 62-way character classifier achieves state-of-the-art accuracy of 83.9% on cropped characters from the ICDAR 2003 test set. The best known previous result on the same benchmark is

81.7% reported by [2]

Our word recognition sub-system is evaluated on im- ages of perfectly cropped words from the ICDAR 2003 and SVT datasets. We use the exact same test setup as [18]. More concretely, we measure word-level accuracy with a lexicon containing all the words from the where lw is the alignment vector3 between the characters in w and the columns of M . Sw can be com- ICDAR test set (called I-WD), and with lexicons consisting of the ground truth words for that image plus puted efficiently using a Viterbi-style alignment algorithm similar to [17].4 We compute Sw for all lexicon

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

50 random "distractor" words added from the test set the highest scoring word w∗. We take SB = Sw to be words and label the word-level bounding-box B with the recognition score of B.

Having defined the recognition score for a single bounding box, we can now systematically evaluate pos- sible word-level segmentations using beam search [15], a variant of breadth first search that explores the top N possible partial segmentations according to some heuristic score. In our case, the heuristic score of a can- didate segmentation is the sum of the SB's over all the resulting bounding boxes in a line of text L. In order to deal with possible false positives from the text detec- tion stage, we threshold individual segments based on their recognition scores. In that way, segments with low recognition scores are pruned out as being "non-text."

3For example, lw = 6 means the 4th character in w aligns with the 6th column of M , or the 6th sliding window in a line of text.

4In practice, we also augment Sw with additional terms that en- courage geometric consistency. For example, we penalize character spacings that are either too narrow or vary a lot within a single word. (called I-WD-50). For the SVT dataset, we used the provided lexicons to evaluate the accuracy (called SVT- WD). Table 1 compares our results with [18] and the very recent work of [11]. We evaluate our final end-to-end system on both the ICDAR 2003 and SVT datasets, where we locate and recognize words in full scene images given a lexicon. For the SVT dataset, we use the provided lexicons; for the ICDAR 2003 dataset, we used lexicons of 5, 20 and 50 distractor words provided by the authors of [18], as well as the "FULL" lexicon consisting of all words in the test set. We call these benchmarks I-5, I-20, I-50 and I-FULL respectively. Like [18], we only consider alphanumeric words with at least 3 characters. Figure 5 shows some sample outputs of our system. We fol- low the standard evaluation criterion described in [10] to compute the precision and recall. Figure 4 shows pre- cision and recall plots for the different benchmarks on the ICDAR 2003 dataset.

As a standard way of summarizing results, we also

Figure 5. Example output bounding boxes of our end-to-end system on I-FULL and SVT bench- marks. Green: correct detections. Red: false positives. Blue: misses.

Table 2. F-scores from end-to-end evalua- tion on ICDAR 2003 and SVT datasets.

Benchmark	I-5	I-20	I-50	I-FULL	SVT
Our approach	.76	.74	.72	.67	.46
Wang, *et al.* [18]	.72	.70	.68	.51	.38

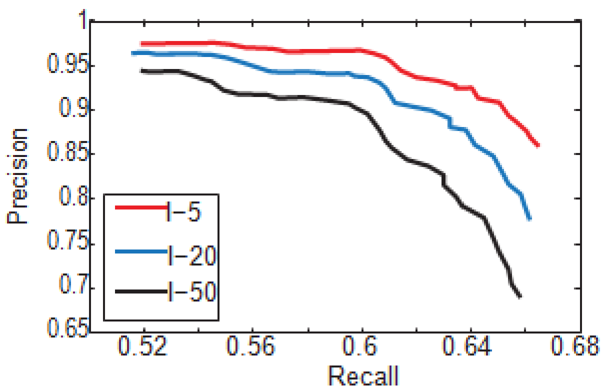

Figure 4. End-to-end PR curves on ICDAR 2003 dataset using lexicons with 5, 20, and 50 distractor words.

report the highest F-scores over the PR curves and com- pare with [18] in Table 2. Our system achieves higher F-scores in every case. Moreover, the margin of im- provement is much higher on the harder benchmarks (0.16 for I-FULL and 0.08 for SVT), suggesting that our system is robust in more general settings.

In addition to settings with a known lexicon, we also extend our system to the more general setting by using a large lexicon L of common words. Since it is infea- sible to search over all words in this case, we limit our search to a small subset P ∈ L of "visually plausible"

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

words. We first perform NMS on the score matrix M
across positions and character classes, and then thresh- old it with different values to obtain a set of raw strings. The raw strings are fed into Hunspell5 to yield a set of suggested words as our smaller lexicon P , Using this simple setup, we achieve scores of 0.54/0.30/0.38 (precision/recall/F-score) on the ICDAR dataset. This 5Hunspell is an open source spell checking software available at http://hunspell.sourceforge.net/. We augment its default lexicon with a corpus of English proper names to better handle text in scenes.

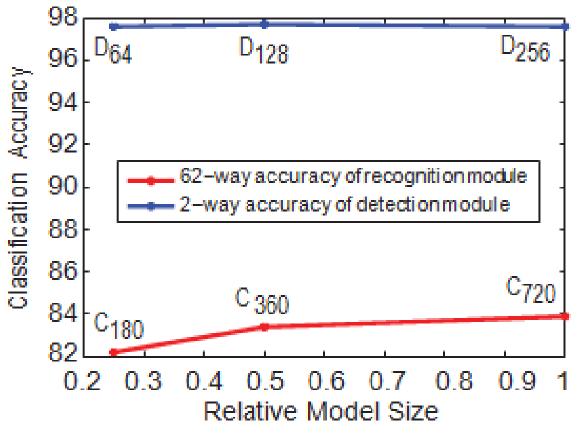

Figure 6. Accuracies of the detection and recognition modules on cropped patches

is comparable to the best known result 0.42/0.39/0.40 obtained with a general lexicon by [14].
In order to analyze the impact of model size on dif- ferent stages of the pipeline, we also train detection and
tional filters. The detection modules have n2 = 64 and recognition modules with fewer second layer convolu- 128 compared to 256 in our full model. We call the de-
tection modules D64, D128 and D256 respectively. Sim- ilarly, we call the recognition modules C180, C360 and C720, which corresponds to n2 = 180, 360 and 720.
The smaller models have about 1/4 and 1/2 number of
learnable parameters compared to the full models.
To evaluate the performance of the detection mod-

Table 3. Classification and end-to-end re- sults of different recognition modules

Recognition module	C_{180}	C_{360}	C_{720}
Classification accuracy	82.2%	83.4%	83.9%
End-to-end F-score	.6330	.6333	.6723

ules, we construct a 2-way (character vs. non-character) classification dataset by cropping patches from the IC- DAR test images. The recognition modules are eval- uated on cropped characters only. As shown in Fig- ure 6, the 62-way classification accuracy increases as model size gets larger, while the 2-way classification re- sults remain unchanged. This suggests that larger model sizes yield better recognition modules, but not necessar- ily better detection modules.
Finally, we evaluate the the 3 different recognition modules on the I-FULL benchmark, with D256 as the detector for all 3 cases. The end-to-end F-scores are listed against the respective classification accuracies in
Table 3. The results suggests that higher character clas- sification accuracy does give rise to better end-to-end results. This trend is consistent with the findings of [12] on house number recognition in natural images.

5. CONCLUSION

In this paper, we have considered a novel approach for end-to-end text recognition. By leveraging large, multi-layer CNNs, we train powerful and robust text detection and recognition modules. Because of this increase in representational power, we are able to use simple non-maximal suppression and beam search tech- niques to construct a complete system. This represents a departure from previous systems which have gener- ally relied on intricate graphical models or elaborately hand-engineered systems. As evidence of the power of this approach, we have demonstrated state-of-the- art results in character recognition as well as lexicon- driven cropped word recognition and end-to-end recog- nition. Even more, we can easily extend our model to the general-purpose setting by leveraging conventional open-source spell checkers and in doing so, achieve per- formance comparable to state-of-the-art.

REFERENCES

[1]D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. High performance neural net- works for visual object classification. Technical Report IDSIA-01-11, Dalle Molle Institute for Artificial Intel- ligence, 2011.

International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
UGC Approved Journal No: 45483
ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

[2]A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng. Text de- tection and character recognition in scene images with unsupervised feature learning. In ICDAR, 2011.
[3]A. Coates, H. Lee, and A. Y. Ng. An analysis of single- layer networks in unsupervised feature learning. In AIS- TATS, 2011.
[4]T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In VISAPP, 2009.
[5]B. Epshtein, E. Oyek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In CVPR, 2010.
[6]A. Hyvarinen and E. Oja. Independent component anal- ysis: algorithms and applications. Neural networks, 13(4-5):411–430, 2000.
[7]Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learn- ing hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In CVPR, 2011.
[8]Amairullah Khan Lodhi, M. S. S. Rukmini, Syed Abdulsattar "Lifetime and Performance Enhancement in WSN by Energy-Buffer Residual Status of Nodes and The Multiple Mobile Sink" TEST Engineering and Management (Unpaid Scopus), Vol.82, pp. 3835-3845 on 20th January 2020.
[9]Amairullah Khan Lodhi, M. S. S. Rukmini, Syed Abdulsattar "Performance Enhancement by Cluster Head Selection Based On Residual Status of Energy and Buffer in WSNs" is accepted in INTERSCIENCIA Journal (SCIE Journal) on 16th December 2019.
[10]Amairullah Khan Lodhi, M. S. S. Rukmini, Syed Abdulsattar " Efficient Energy Routing Protocol Based on Energy & Buffer Residual Status (EBRS) for Wireless Sensor Networks" International Journal of Engineering and Technology (Scopus), Vol. 9. Issue no. 1S5 pp. 33-37, in December 2019.
[11]Amairullah Khan Lodhi, M. S. S. Rukmini, Syed Abdulsattar "Energy-efficient Routing Protocol for Life Enhancement in Wireless Sensor Networks" Recent Patents on Computer Science (Unpaid Scopus), Vol.12, no.1. pp. 01-10, 2019.
[12]Amairullah Khan Lodhi, M. S. S. Rukmini, Syed Abdulsattar "Energy-Efficient Routing Protocol Based on Mobile Sink Node in Wireless Sensor Networks" International Journal of Innovative Technology and Exploring Engineering (Scopus), Vol.8, Issue-7, pp 1788-1792, 2019.
[13]Amairullah Khan Lodhi, M. S. S. Rukmini "Energy-Efficient Routing Protocol for Node Lifetime Enhancement in Wireless Sensor Networks" International Journal of Advanced Trends in Computer Science and Engineering (Scopus), Vol. 8. No.1.3. pp. 24-28, 2019.
[14]Amairullah Khan Lodhi, Mohammad Iliyas, Safiya Kaunain "Design of 8-4 and 9-4 Compressors For high-Speed Multiplication" Journal of Applied Science and Computations (JASC)–Volume.6. Issue.6. pp. 2056-2068 in June 2019.
[15]Amairullah Khan Lodhi, Sindhu S. Kale "Development on Gas Leak Detection and Location System Based on Wireless Sensor Networks: A" International Journal of Engineering Trends and Technology (IJETT)–Volume 12. in 2014.
[16]Amairullah Khan Lodhi, S. R. Madkar, Prof. A. V. Karande "Wireless Monitoring of Soil Moisture & Humidity using Zigbee in Agriculture" International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume.10. Issue.4. pp. 817-821 in October 2014.
[17]A. K. Lodhi, Prashant R. Gade "Review of Poly House Control System" International Journal of Advanced (IJAIEM), Volume.2. Issue.3. pp. 76 in February 2014.
[18]K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In ICCV, 2011.
[19]J. J. Weinman, E. Learned-Miller, and A. R. Hanson. A discriminative semi-markov model for robust scene text recognition. In ICPR, Dec. 2008.