AN IMPROVED LEARNED MULTI-SCALE SCENE TEXT RECOGNITION

Y N S Vamsi Mohan[1], Srinivas Karri[2], Siripalli Raghava Rao[3]
[1,2,3]Department of Electronics and Communication Engineering,
Bonam Venkata Chalamayya Institute of Technology & Science,
Amalapuram, Andhra Pradesh, India.

ABSTRACT: The main objective of this project is to design an efficient multi-scale representation for scene text recognition. This project proposes a novel multiscale representation, which leads to accurate, robust character identification and recognition. This representation consists of a set of mid-level primitives, termed strokelets, which capture the underlying substructures of characters at different granularities. In this paper, we propose a novel multi-scale representation for scene text recognition. This representation consists of a set of detectable primitives, termed as strokelets, which capture the essential substructures of characters at different granularities. Strokelets possess four distinctive advantages: (1) Usability: automatically learned from bounding box labels; (2) Robustness: insensitive to interference factors; (3) Generality: applicable to variant languages; and (4) Expressivity: effective at describing characters. Extensive experiments on standard benchmarks verify the advantages of strokelets and demonstrate the effectiveness of the proposed algorithm for text recognition.

INTRODUCTION: As an important carrier of human thoughts and emotions, text plays a crucial role in our daily lives. It is almost ubiquitous, especially in modern urban environments. For example, product tags, license plates, guideposts and billboards, all contain text. The rich information embedded in text can be very beneficial, but the rapid growth of text data has made it prohibitive to process, interpret and apply it manually. Consequently, automatic text detection and recognition have become an irresistible general trend. However, spotting and reading text in natural scenes are extremely difficult for computers. Though considerable progress has been achieved in recent years [8, 2, 3, 2, 2, 2, 4], detecting and recognizing text in uncontrolled environments are still open problems in computer vision. Various interference factors, such as variation, distortion, noise, blur, non-uniform illumination, local distractor and complex background, all may pose major challenges [3, 4]. To tackle these challenges, representation is lying in the middle and core. Excellent representations should be able to effectively describe the characteristics of characters in natu- ral images and meanwhile be robust to interference factors. In this work, we are concerned with the problem of text recognition in natural scenes (a.k.a. scene text recognition) and propose a novel multi-scale representation. This representation consists of a set of multi-scale mid-level primitives, termed as strokelets, each of which under ideal conditions represents a stroke shape. In particular, strokelets possess four distinctive advantages over conventional representations, which are called the "URGE" properties:

- Usability: automatically learned from bounding box labels, not requiring detailed annotations.
- Robustness: insensitive to interference factors, endowing the system with the ability to deal with real-world complexity.
- Generality: applicable to variant languages, as long as sufficient training examples are available.
- Expressivity: effective at describing characters in natural scenes, bringing high recognition accuracy.A subset of learned strokelets and several character recognition examples by a system operating on those strokelets are demonstrated in Fig. 1. Strokelets, as a universal representation for characters, faithfully seize the representative parts of characters at multiple scales; and characters in different fonts, scales, colors, and layouts can be successfully localized and read, even with the presence of noise, blur and distractor. Character identification1, the process of hunting each individual character and estimating the position and extent of these characters, is a critical stage in text recognition, as it constitutes the basis of subsequent feature computation, character classification and error correction. In this sense, the quality of character

identification largely determines the accuracy of text recognition. WRITING, considered as a hallmark of civilization [1], is one of the greatest inventions of humanity.

STROKELET GENERATION:

Given a set of training images S = {(Ii , Bi)}ni =1 containing characters, where Ii is an image and Bi is a set of bounding boxes specifying the location and extent of the characters in the image Ii , the goal of strokelet generation is to learn a set of universal part prototypes _ from the training set S. The part prototypes should be able to capture the essential sub-structures of characters and be distinctive from local background and against each other. As S only provides character level annotations, the part prototypes should be automatically discovered. The newly developed discriminative clustering algorithm proposed by Singh et al. [42] meets the requirements well, since it learns visual primitives that are both representative and discriminative from large image collections in an unsupervised manner.

RECOGNITION ALGORITHM:

The algorithmic pipeline for scene text recognition is fairly straightforward: Character candidates are first sought from the image via a voting based scheme for character identification (these candidates are then described by a histogram feature based on strokelets and a holistic descriptor ; and character classification is applied to assign the most probable class label to each character Optionally, the inferred word is replaced by the most similar item in a given dictionary, following [10], [28]. The algorithm described above is quite effective, even though without sophisticated approaches to error correction [8], [13].We believe better performance could be achieved if such error correction methods are incorporated.

CHARACTER IDENTIFICATION

Character identification is a key stage in scene text recognition. However, binarization based methods [14], [18] are sensitive to noise, blur and non- uniform illumination; connected component based methods [12], [13] are unable to handle connected characters and broken strokes; and sliding window based character detection based methods [8], [26] usually produce a lot of false alarms, mainly due to varying aspect ratios of characters and background clutters. In our previous work [24], we proposed a novel scheme to seek character candidates, via multi- scale strokelet detection and voting. The scheme shares the idea of estimating character centers through voting with the work of Yildirim et al. [29]. The work in [29] is essentially a patch based method, which does not explicitly infer character parts, but simply learns the mapping relations (multi-class Hough Forests) between local patches and character center; besides, it only performs voting at single scale, though multi-scale scanning is used during character detection. In contrast, the strategy in [24] casts votes from multiple scales. Firstly, the original word image) is resized to a standard height (64 pixels in this paper) with aspect ratio kept unchanged; since strokelets are naturally multiscale representation, a multi-scale sliding-window paradigm is performed to detect strokelets a Hough map is then generated by casting and accumulating the votes from the strokelets activations, similar to [5] and [21]

CHARACTER DESCRIPTION

Based on detection activations of strokelets, we introduce a histogram feature called Bag of Strokelets, in addition to the traditional feature HOG

1) Bag of Strokelets: For each identified character candidate, all the strokelets that have voted for it are sought via back-projection. A histogram feature is formed by binning the strokelets. Strokelets of all scales are assembled together. Each strokelet contributes to the histogram feature according to its detection score. To incorporate spatial information, the Spatial Pyramid strategy [62] (1 × 1 and 2 × 2 grids) is also adopted. Among all pooling methods, we found max-pooling [63] gave the highest accuracy, so maxpooling is employed in all the experiments in this paper.

2) HOG: Following [13], [28], we also adopt the HOG descriptor (the version proposed in [40]) to describe characters. A template with 5 × 7 cells is constructed for each character candidate. The Bag of Strokelets feature is complementary to HOG, as it conveys information from different levels and is robust to font variation, subtle deformation and partial occlusion. We will evaluate the effectiveness of these two types of features and compare their contributions to recognition accuracy

CHARACTER CLASSIFICATION

In this paper, we consider English letters (52 classes) and Arabic numbers (10 classes), i.e. the alphabet
{a, . . . , z; A, . . . , Z; 0, . . . , 9} and |_| = 62. To handle
invalid characters (e.g. punctuations and background components), we also introduce a special class, so there are 63 classes in total. We train 63 character recognizers (binary classifiers), one for each character class, in a one-vs-all manner. Random Forest [56] is adopted as the strong classifier because of its high performance and efficiency. Training examples are harvested by applying the strokelets to the images in the training set and compare the identified rectangles with the ground truth annotations. At runtime, the

character candidates are classified by the trained recognizers; for each character, the class label with the highest probability is assigned as the recognition consequence.

DETECTION ALGORITHM:

It has become an emerging trend in computer vision to adopt representation trained for one task to accomplish other tasks [9], [4].

HISTOGRAM ORIENTED GRADIENTS:

The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. Robert K. McConnell of Wayland Research Inc. first described the concepts behind HOG without using the term HOG in a patent application in 1986.[1] In 1994 the concepts were used by Mitsubishi Electric Research Laboratories.[2] However, usage only became widespread in 2005 when Navneet Dalal and Bill Triggs, researchers for the French National Institute for Research in Computer Science and Automation (INRIA), presented their supplementary work on HOG descriptors at the Conference on Computer Vision and Pattern Recognition (CVPR). In this work they focused on pedestrian detection in static images, although since then they expanded their tests to include human detection in videos, as well as to a variety of common animals and vehicles in static imagery.

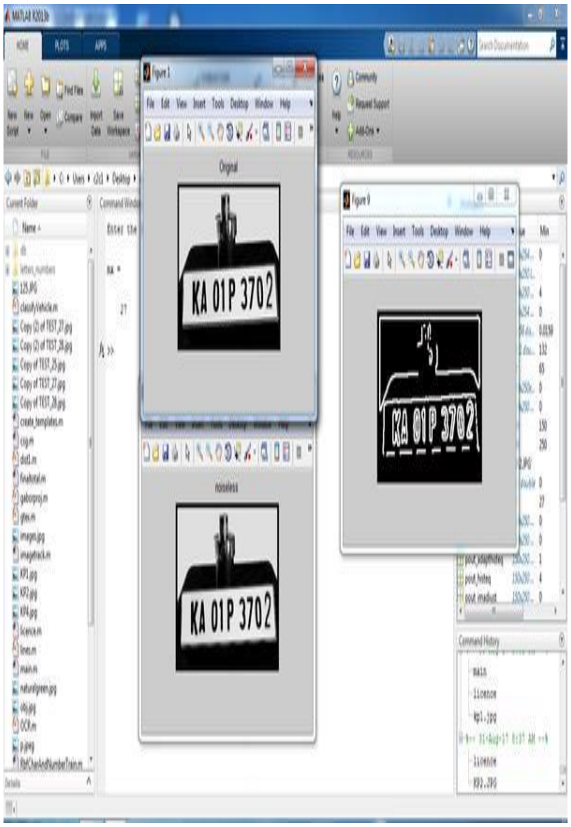
The essential thought behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is the concatenation of these histograms. For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing.

Algorithm overview

Compute a Histogram of Oriented Gradients (HOG) by

1. (optional) global image normalisation
2. computing the gradient image in x and y
3. computing gradient histograms
4. normalising across blocks
5. flattening into a feature vector

RESULT:



CONCLUSION:

We have introduced strokelets, a novel presentation automatically learned from bounding box labels, for the purpose of capturing the underlying substructures of characters at different granularities. Strokelets provide an alternative way to accurately identify individual characters and compose a histogram feature to effectively describe characters in natural scenes. The scene text recognition algorithm based on strokelets is both effective and robust. Extensive experiments on standard benchmarks verify the advantages of strokelets and demonstrate that the proposed algorithm consistently outperforms the current state-of-theart approaches in the literature. In this paper, we only demonstrated the strengths of strokelets on the task of text recognition in cropped images. The idea is actually quite general and can be employed to perform both text detection and recognition in full images. This is an ongoing work. Furthermore, we could extend the applicability of this idea by learning multi scale prototypes for other object classes (e.g. cars, persons, and faces) and using them to detect and recognize such object classes.

REFERENCES:

[1] ABBYY FineReader 9.0. http://www.abbyy.com/.

[2] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In Proc. ECCV, 2010.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In Proc. ICCV, 2009.

[4] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[5] Y. Cheng. Mean shift, mode seeking, and clustering. IEEE Trans. PAMI, 17(8):790–799, 1995.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. CVPR, 2005.

[7] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In Proc. of VISAPP, 2009.

[8] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In Proc. Of CVPR, 2010.

[9] L. Fei-Fei and P. Perona. A bayesian heirarcical model for learning natural scene categories.In Proc. of CVPR, 2005.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Trans. PAMI, 32(9):1627–1645, 2010.

[11] V. Goel, A. Mishra, K. Alahari, and C. V. Jawahar. Whole is greater than sum of parts: Recognizing scene text words. In Proc. of ICDAR, 2013.

[12] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classication. In Proc. of CVPR, 2013.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. of CVPR, 2006.

[14] S. H. Lee, K. J. M. S. Cho, and J. H. Kim. Scene text extraction with edge constraint andtext collinearity link. In Proc. of ICPR, 2010.

[15] B. Leibe, A. Leonardis, and B. Schiele. Robust object detectionwith interleaved categorization and segmentation. IJCV, 77(1-3):259–289, 2008.

[16] J. J. Lim, C. L. Zitnick, and P. Dollar. Sketch tokens: A learned mid-level representation for contour and object detection. In Proc. of CVPR, 2013.

[17] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S.Wong, and R. Young. ICDAR 2003 robust reading competitions. In Proc. of ICDAR, 2003.

[18] S. McCann and D. G. Lowe. Spatially local coding for object recognition. In Proc. ACCV, 2012.

[19] A. Mishra, K. Alahari, and C. V. Jawahar. An MRF model for binarization of natural scene text. In Proc. Of ICDAR, 2011.

[20] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In Proc. Of BMVC, 2012.

[21] A. Mishra, K. Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In Proc. Of CVPR, 2012.

[22] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. Computer Speech and Language, 16(1):69–88, 2013.

[23] L. Neumann and J. Matas. Real-time scene text localization and recognition. In Proc. Of CVPR, 2012.

[24] L. Neumann and J. Matas. Scene text localization and recognition with oriented stroke detection. In Proc. Of ICCV, 2013.