Natural Scene Text Extraction

Akshay Dixit

February 23, 2015

1 Idea

Extracting text from natural images, as opposed to scans of printed pages, faxes and business cards, is an important step for a number of Computer Vision applications, such as computerized aid for visually impaired, automatic geocoding of businesses, and robotic navigation in urban environments. Retrieving texts in both indoor and outdoor environments provides contextual clues for a wide variety of vision tasks. Moreover, it has been shown that the performance of image retrieval algorithms depends critically on the performance of their text detection and extraction modules. Thus, we put forward a crowdsourced approach to extract text from natural images to tackle this problem.

2 Building Blocks

The following are the computer vision related building blocks that would make up parts of the whole system.

1. Localization of text in image
2. Understanding the text

The whole system architecture might consist of a service running on a machine which would accept images as uploads or through POST requests and forwards it to the Text Localization block. That image patches obtained from the Text Localization block would then be processed by the Text Extraction block, which would result in text that would be sent back to the user as a TXT or PDF file that he/she could download, or the service could keep the connection alive long enough for the pipeline to finish and send the extracted text back as a response*(This would depend on the execution times of the two building blocks, and since we are using HPUs, they could very well be the bottleneck of the servivce. Thus this method is not preferred).*

2.1 Text Localization

Although text with a limited scope can be successfully detected using existing technologies, such as in high quality printed documents, it is difficult to detect if it is merged in a noisy background such as a poster, a sign or an advertisement.

The textual data in the scene can vary with font, size, position, orientation, be blurred due to motion, or slightly occluded by other objects. Originating in a 3-D space, such text in scene images can be distorted by slant, tilt, and shape of objects on which they are found. There has already been a lot of work done in this field. Some various existing approaches include

1. Edge detection + Adaptive search + Layout analysis[1]
2. Texture segmentation[2]
3. Stroke Width Transform[3]
4. Background segmentation[4]

All of these above approaches, though fairly accurate, are prone to leaving out a few areas and end up generating false positives too. This problem can be solved by using human input to detect text regions in natural images. Consider the following images.

Figure 1: A noisy image. NOTE: *Due to scaling down of the image, the clarity has been reduced, although in full resolution, a person could easy make out the characters written even on the white posters*

In the above image, a normal Hindi literate person would easily be able to make out and understand all the text in the picture. But it would be hard for a machine due to the noisy background, unevenness of the posters, and random font sizes across the board. Similarly with this image below, a computer vision algorithm might detect "COUN", "HOUS" and "L" as separate word instances, although a human can easily figure out the real meaning even though the tree is blocking the poster.

Crowdsourcing

Thus, localizing words in natural scenes through crowdsourcing should easily beat regular computer vision algorithms at detection as shown by the two examples above. A simple implementation where users draw bounding boxes around word instances in the images would suffice. Another method would be to extract text patches using above stated algorithms and have humans vote on whether it contains text or not would be useful in eliminating false positives.

Figure 2: An image with occluded text

This building block would also be useful for the following projects as they both involve drawing bounding boxes around specific regions of interest(in our case, any text/characters)

1. Image Understanding
2. Number Plate Recognition

2.2 Extraction of text

Once we have the text patches extracted from the previous building block, we can attempt to interpret text from it. The following approaches have alreay been quite successful at extraction of textual data from image patches.

1. Synthetic data generation followed by whole word input CNNs(Convolutional Neural Networks)[5]
2. MSER(Maximally Stable Extremal Regions) based methods involving text candidate extraction and clustering[6, 7]
3. Recognition by Retrieval using synthetic word image generation and matching[8]

Crowdsourcing
Although the work done by Jaderberg *et al*[5] shows really high accuracy for datasets like ICDAR 2003[9] and Street View Text, crowdsourced input could still be used to enhance the accuracy of the detected text and check for corrections. The input of people could be used to either translate the words in the text patch verbatim, or they could be given a list of n most likely translations and asked to choose the correct one.

This building block could also be used by the following other projects in their optical character recognition processes.

1. Offline Intelligent Character Recognition for Critical Applications
2. Deciphering Inscriptions. This would a require a list of all possible characters that could be encountered to be known and provided to the HPU beforehand which is highly improbable.
3. Digitizing Handwritten Text
4. Notes Maker

3 Datasets

The following are a few of the datasets that could be used to implement and test the above discussed building blocks

1. ICDAR 2003 dataset[9]
2. ICDAR 2013 dataset[10]

4 Conclusion

Thus, we introduce two building blocks based on crowdsourcing that can be used to build a system which could successfully extract text from natural scenes with an accuracy that should be significantly greater than any of the existing computer vision approaches.

References

[1] Jie Yang, Xilin Chen, Jing Zhang, Ying Zhang, and Alex Waibel. Automatic detection and translation of text from natural scenes. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2101. IEEE, 2002.

[2] Victor Wu, R Manmatha, and Edward M Riseman. Finding text in images. In *ACM DL*, pages 3–12, 1997.

[3] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963–2970. IEEE, 2010.

[4] Katherine L Bouman, Golnaz Abdollahian, Mireille Boutin, and Edward J Delp. A low complexity method for detection of text area in natural images. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1050–1053. IEEE, 2010.

[5] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[6] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(5):970–983, 2014.

[7] Huizhong Chen, Sam S Tsai, Georg Schroth, David M Chen, Radek Grzeszczuk, and Bernd Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2609–2612. IEEE, 2011.

[8] V. Goel, A. Mishra, K. Alahari, and C. V. Jawahar. Whole is greater than sum of parts: Recognizing scene text words. In *Proceedings of International Conference on Document Analysis and Recognition*, 2013.

[9] Lucas Panaretos Sosa, S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *In Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 682–687. IEEE Press, 2003.

[10] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, L Gomez i Bigorda, S Robles Mestre, Joan Mas, D Fernandez Mota, J Almazan Almazan, and L-P de las Heras. Icdar 2013 robust reading competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1484–1493. IEEE, 2013.