Digital Image Processing

Project Report

Phase I

Synthetic Data for Text Localization in Natural Scene Images

Team : "~CNN"

Joyneel Misra

Sarthak Sharma

Introduction

The project[1] deals with a new method for generating synthetic images of text that naturally blends text in existing natural scenes. We use this method to automatically generate a new synthetic dataset of text in cluttered conditions. The key difference with existing synthetic text datasets[2] is that they only contain word level image regions,hence are unsuitable to train detectors.

The original paper extends the idea further, contributing a text detection deep architecture,Fully_convolutional Regression Network. It performs prediction at every image location,the prediction being parameters of a bounding box enclosing the word centered at that location, the latter idea being borrowed from YOLO technique[3].

Goals

To develop a synthetic text-scene image generation engine for building a large annotated dataset for text localisation.

Approach

The synthetic engine produces realistic scene-text images so that the trained models can generalize to real images(non-synthetic).

The engine is fast,fully automated and enables the generation of large quantities of data without supervision.

The text generation pipeline can be summarized as follows:

- **Acquiring suitable text and image samples:**

The synthetic text generation process starts by sampling some text and a background image. The text is extracted from the Newsgroup 20 dataset [4] in three ways — words, lines (up to 3 lines) and paragraphs (up to 7 lines). To favour variety, 8,000 background images are extracted from Google Image Search through queries related to different objects/scenes and indoor/outdoor and natural/artificial locales. To guarantee that all text occurrences are fully annotated, these images must not contain text of their own.

- **Segmentation and Geometry Estimation**

In real images, text tends to be contained in well defined regions (e.g. a sign). We approximate this constraint by requiring text to be contained in regions characterised by a uniform colour and texture. This also prevents text from crossing strong image discontinuities, which is unlikely to occur in practice. Regions are obtained by thresholding the gPb-UCM contour hierarchies [5] at 0.11 using the efficient graph-cut implementation of [6].

In natural images, text tends to be painted on top of surfaces (e.g. a sign or a cup). In order to approximate a similar effect in our synthetic data, the text is perspectively transformed according to local surface normals. The normals are estimated automatically by first predicting a dense depth map using the CNN of [7] for the regions segmented above, and then fitting a planar facet to it using RANSAC [8].

Text is aligned to the estimated region orientations as fol-lows: first, the image region contour is warped to a frontal-parallel view using the estimated plane normal; then, a rectangle is fitted to the fronto-parallel region; finally, the text is aligned to the larger side ("width") of this rectangle. When placing multiple instances of text in the same region, text masks are checked for collision against each other to avoid placing them on top of each other. Not all segmentation regions are suitable for text placement — regions should not be too small, have an extreme aspect ratio, or have surface normal orthogonal to the viewing direction; all such regions are filtered in this stage. Further, regions with too much texture are also filtered, where the degree of texture is measured by the strength of third derivatives in the RGB image.

- **Text Rendering and Image Composition**

Once the location and orientation of text has been de-cided, text is assigned a colour. Pixels in each cropped word images are partitioned into two sets using K-means,resulting in a colour pair, with one colour approximating the foreground (text) colour and the other the background. When rendering new text, the colour pair whose background colour matches the target image region the best (using L2-norm in the Lab colour space) is selected, and the corresponding foreground colour is used to render the text.The text is blended using Poisson image editing.

References

[1] Synthetic Data for Text Localisation in Natural Images

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. IJCV , 2015.

[3] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proc. CVPR , 2016. To appear.

[4] K. Lang and T. Mitchell. Newsgroup 20 dataset, 1999.

[5] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE PAMI ,33:898–916, 2011.

[6] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In Proc. CVPR,2014.

[7] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In Proc. CVPR ,2015.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM , 24(6):381–395, 1981.