1. INTRODUCTION

- An unconstrained (i.e., lexicon-free) end-to-end scene text localization and recognition method
- Individual characters detected and recognized as image regions which contain strokes of specific orientations in a specific relative position
- Character strokes are detected by a ridge filter in an image gradient scale space which is more robust to noise than SWT [2], because the assumption that a character is a single connected component is dropped
- The method combines the advantages of sliding-window methods (robustness to noise and blur) and connected component methods (effective enumeration [3])
- Character representation calculated effectively from values obtained in the stroke detection phase

2. STROKE DETECTION

- Stroke of direction $\alpha$ is detected as two opposing ridges in the gradient (approximately) perpendicular to the stroke direction. Note that the distance $w$ between the ridges is the stroke width [2]
- Assuming that the stroke width value is one ($w = 1$) in a certain scale, the method searches for all strokes of a unit width in a scale space by convolving the gradient projection in 4 orientations with a 5 x 5 ridge filter that responds to such strokes
- The response of the convolution filter is thresholded with a parameter $\theta$, which represents a trade-off between an ability to detect strokes of low-contrast and the number of candidate regions to classify
- The convolution is preformed twice, once in the original image and once in an inverted image to detect strokes with an opposite ridge orientation

3. CANDIDATE REGION DETECTION

- The method assumes that a character bounding-box can be found as a union of some of the character's strokes' bounding-boxes; consider the letter “E” - it consists of 4 strokes (3 horizontal and 1 vertical) and 4 out of 6 possible stroke pairs (and all 4 possible stroke triplets) induce an identical character bounding-box
- For each stroke (detected by the convolution filter in the previous stage) the method considers its bounding-box and then $K$ bounding-boxes created as a union of bounding-boxes of 1 to $K$ nearest neighboring strokes
- The problem is over-complete and exact position of the character bounding box is not crucial, therefore similar candidate bounding-boxes (> 95% overlap) are eliminated
- This approach reduces the number of target rectangles by three orders of magnitude when compared to the sliding-window methods which exhaustively search the image

4. CHARACTER RECOGNITION & WORD FORMATION

- The feature vector of each candidate region is created by concatenating responses of all 4 orientations, where each orientation pools convolution responses over a given interval of scales (which is determined by size of the bounding box)
- The representation is robust to shift at stroke level, which reduces intra-class variance and makes it less sensitive to positioning and scale variance
- The feature vector is classified by a nearest-neighbor classifier which either assigns a Unicode label or rejects the region as a non-character
- Character regions are agglomerated into text lines and final labeling is found by finding an optimal sequence in each text line which maximizes an objective function
- The objective function exploits character geometrical properties (consistent inter-character spacing, region positioning on a text line), character recognition confidence and a label adjacency probability based on a prior language model

5. EXPERIMENTS

- The ICDAR 2011 dataset [1] contains 1189 words in 255 images
- Achieved state-of-the-art end-to-end localization (recall 66.4%, precision 79.3%) and recognition (recall 45.4%, precision 44.8%) results

Text localization and recognition results on the ICDAR 2011 dataset

6. CONCLUSIONS

- The method combines benefits of connected-component and sliding-window based methods
- The assumption that a character is a connected component is dropped, which allows for detection of joint or disconnected characters
- On a standard PC the average running time on a 2M image is 35s in MATLAB

REFERENCES


Prepared at ICCV’13, Sydney, Australia, 3.12.2013