Scene Text Aware Image Retargeting

Diptiben Patel, Shanmuganathan Raman

Electrical Engineering, Indian Institute of Technology Gandhinagar, Gujarat, India

IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2019, Ottawa, Canada

Motivation

- Extensive use of text labels and symbols available in the digital media for interpretation and communication of information has gained a lot of attention in the era of digital media.
- Access of the images with scene text in it through different display devices tend to deform the scene text region while resizing for better viewing experience.
- We introduce a scene text aware image retargeting operator in order to preserve the textual content present in the image.

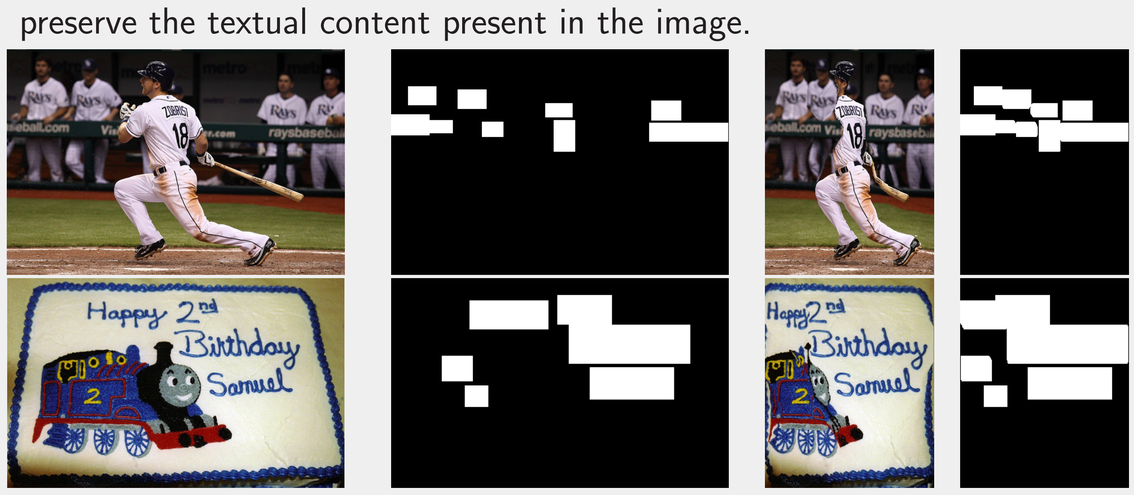
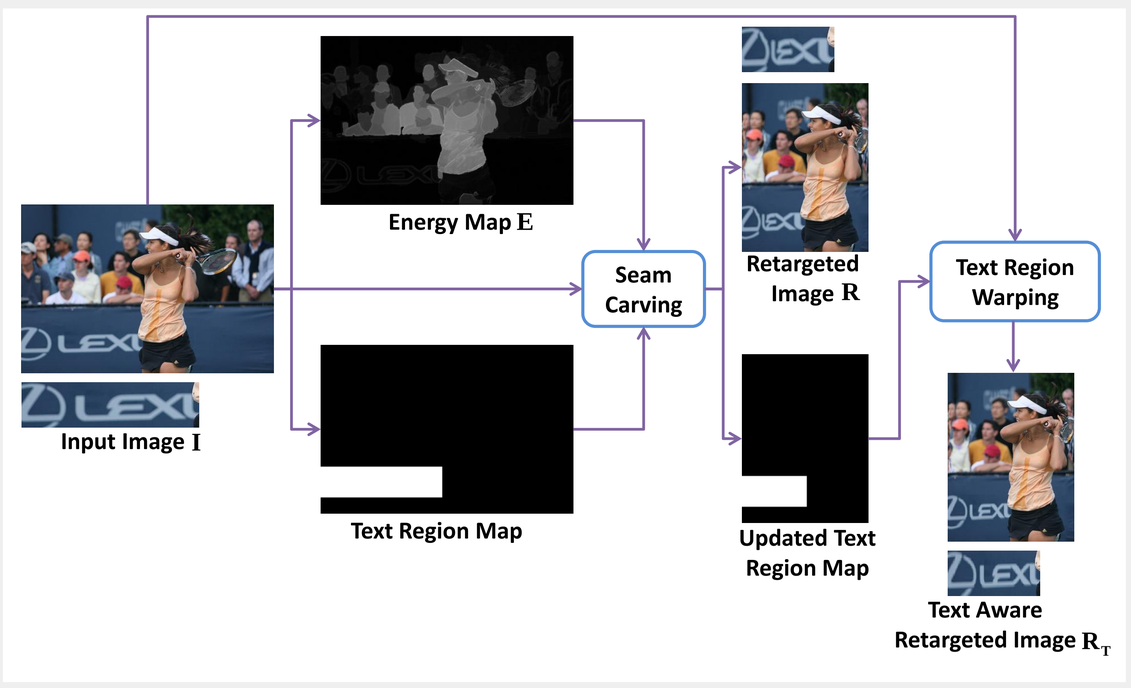
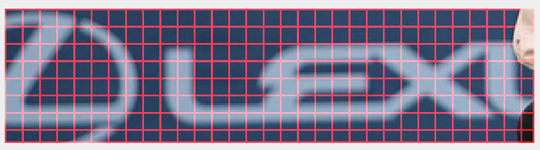


Figure 1:Image retargeting using scene text region as an energy map: (a) Original image, (b) Original energy map, (c) Retargeted image having width as 0.50 of the original image width, (d) Updated energy map.

Proposed Scene Text Aware Image Retargeting



Text Region Warping



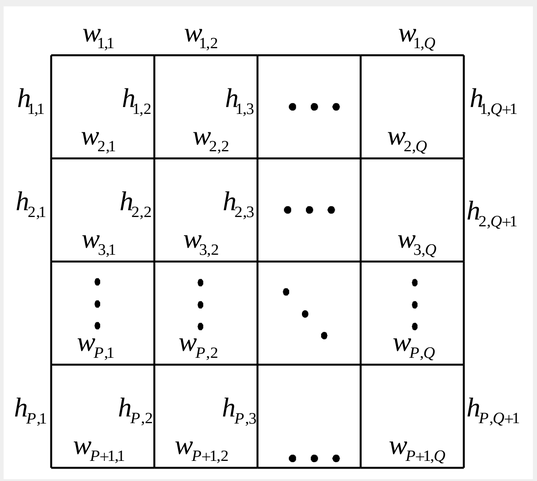


Figure 2:Quad edge-grid model and notations

- Quad edge-grid based optimization [1] problem which optimizes the grid height and the grid width as an optimization variable

$$\mathbf{k}^* = \underset{\mathbf{k}}{\operatorname{argmin}}\ \mathbf{E_T}$$

$$\text{s.t.} \sum_{x=1}^{P} h_{x,y} = m', \forall y \in \{1,2,\ldots,Q+1\}$$

$$\sum_{y=1}^{Q} w_{x,y} = n', \forall x \in \{1,2,\ldots,P+1\}$$

$$h_{x,y} \geq h_{min}, \forall x \in \{1,2,\ldots,P\}, \forall y \in \{1,2,\ldots,Q+1\}$$

$$w_{x,y} \geq w_{min}, \forall x \in \{1,2,\ldots,P+1\}, \forall y \in \{1,2,\ldots,Q\}$$

$\mathbf{k} = [h_{1,1}, h_{1,2}, \ldots, h_{P,Q+1}, w_{1,1}, w_{1,2}, \ldots, w_{P+1,Q}]^\top$.

- Energy factor

$$\mathbf{E_T} = \beta_1 \mathbf{E_{TEP}} + \beta_2 \mathbf{E_{GSP}} = \mathbf{k}^\top (\beta_1 \mathbf{Q_{TEP}} + \beta_2 \mathbf{Q_{GSP}}) \mathbf{k}$$

- $m \times n$ is textbox size in original image, $m' \times n'$ is textbox size in retargeted image.

Energy Factors of E_T

- Text Stroke Edge Preserving Energy $\mathbf{E_{TEP}}$

$$\mathbf{E_{TEP}} = \sum_{x=1}^{P}\sum_{y=1}^{Q}\left(\Omega_{x,y}\left(\frac{P}{m}h_{x,y} - \frac{Q}{n}w_{x,y}\right)\right)^2 + \sum_{x=1}^{P}\sum_{y=1}^{Q}\left(\Omega_{x,y}\left(\frac{P}{m}h_{x,y} - \frac{Q}{n}w_{x+1,y}\right)\right)^2 + \sum_{x=1}^{P}\sum_{y=1}^{Q}\left(\Omega_{x,y}\left(\frac{P}{m}h_{x,y+1} - \frac{Q}{n}w_{x,y}\right)\right)^2 + \sum_{x=1}^{P}\sum_{y=1}^{Q}\left(\Omega_{x,y}\left(\frac{P}{m}h_{x,y+1} - \frac{Q}{n}w_{x+1,y}\right)\right)^2$$

  The grid wise energy is calculated by averaging the absolute gradient value for each grid, and it is defined as $\mathbf{\Omega} \in \mathbb{R}^{P,Q}$.
- Grid Shearing Preventing Energy $\mathbf{E_{GSP}}$

$$\mathbf{E_{GSP}} = \sum_{x=1}^{P-1}\sum_{y=1}^{Q}\left(\sum_{z=1}^{x} h_{z,y} - \sum_{z=1}^{x} h_{z,y+1}\right)^2 + \sum_{x=1}^{Q-1}\sum_{y=1}^{P}\left(\sum_{z=1}^{x} w_{y,z} - \sum_{z=1}^{x} w_{y+1,z}\right)^2$$

- The warped textbox having optimized height and width of each grid is achieved using nearest-neighbor interpolation of each grid.

Qualitative Results

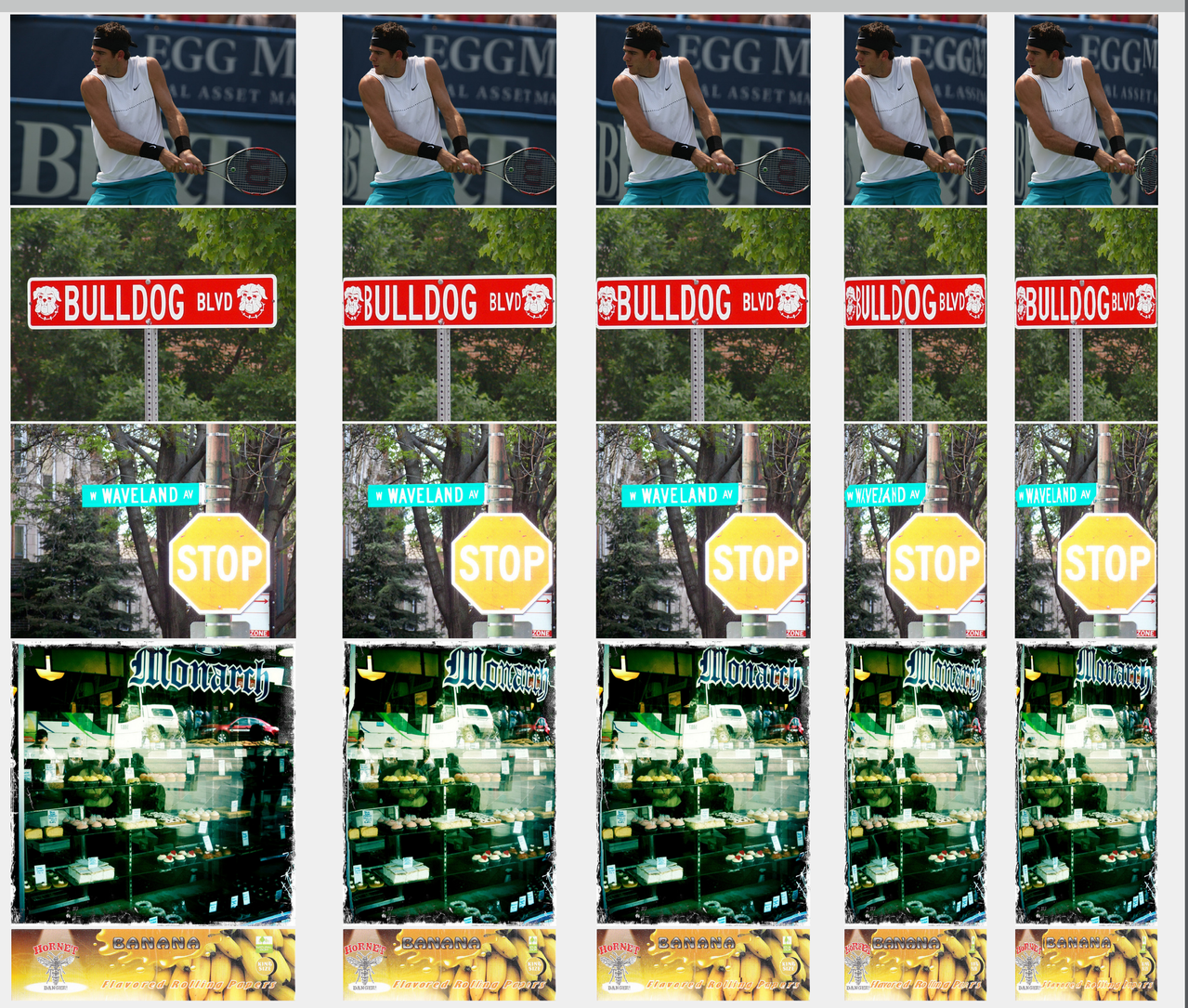


Figure 3:Image retargeting using COCO-Text dataset: (a) Original image, Retargeted image having width as 0.75 of the original image width using (b) Seam Carving [2] and (c) Proposed approach, Retargeted image having width as 0.50 of the original image width using (d) Seam Carving [2] and (e) Proposed approach.

Quantitative Evaluation

- The overall shape and structure information in a visually salient region is captured by low-level aspect ratio similarity (Q_{ARS}) measure. The deformation in the less salient region is captured by mid-level edge group similarity (Q_{EGS}) measure.
- The recognition score (Q_R) for an image

$$Q_R = 1 - \frac{\sum_{i=1}^{V} L^i}{\sum_{i=1}^{V} C^i}$$

  L^i is the edit distance between recognized text from the i^{th} textbox of the original image and that of the retargeted image. C^i is the number of characters in the recognized text from the i^{th} textbox of the original image. V is the number of detected textboxes in an image. $\{Q_{ARS}, Q_{EGS}, Q_R\} \in [0,1]$, grid height$=10$ pixels, grid width $=10$ pixels, $\beta_1 = 10^{-9}$ and $\beta_2 = 0.99$.

Table 1:Quantitative evaluation using different measures for the COCO-Text dataset

Width Reduction	0.75		0.50	
Measure/ Method	SC [2]	Proposed	SC [2]	Proposed
$Q_{ARS}(\uparrow)$	0.9584	**0.9586**	**0.8928**	**0.8928**
$Q_{EGS}(\uparrow)$	**0.8675**	0.8670	**0.8361**	0.8355
$Q_R(\uparrow)$	0.7115	**0.7282**	0.5043	**0.5430**

References

[1] Y. Kim, H. Eun, C. Jung, and C. Kim, "A quad edge-based grid encoding model for content-aware image retargeting," *IEEE transactions on visualization and computer graphics*, 2018.

[2] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," in *ACM transactions on graphics (TOG)*, vol. 27, no. 3. ACM, 2008, p. 16.

diptiben.patel@iitgn.ac.in, shanmuga@iitgn.ac.in

Thanks to IEEE SPS and SERB for providing the travel grant.