May 2022

Text Cutoff Detection for Document Images

Avisek Lahiri

Xinwei Yao

Tianli Yu

Text Cutoff Detection for Document Images

ABSTRACT

This disclosure describes techniques for detection of text cutoff in captured images of documents that include text. Optical character recognition (OCR) is applied to an input image. A bounding box for each text character (OCR symbol) is determined, defined by x and y coordinates of its four corners. A feature vector is determined and utilized to represent the spatial location of OCR symbols extracted from the image. The feature vector is constructed based on OCR symbol coordinates and is provided to a trained classifier to determine a class label for the input document, indicating whether the document includes text cutoff. Optionally, the area of an image that includes text is automatically determined and utilized to limit the area of the image utilized for downstream document processing.

KEYWORDS

- Document segmentation
- Optical character recognition (OCR)
- Feature vector
- Classifier
- Support Vector Machine (SVM)
- Text cutoff
- Symbol box
- Document corner detection
- Bounding box

BACKGROUND

Users commonly take pictures of documents, checks, identification documents (ID), etc. with their device cameras. Many such pictures are subsequently shared with various service providers for online processing and/or verification. In some cases, a portion of the document can be inadvertently excluded from the captured image due to camera positioning. This can lead to missing portions of text (text cutoff) in the captured image. The missing text can pose challenges to the automated and/or manual processing of the document, sometimes necessitating the user to retake the picture. A notification to the user at the time of image capture itself that the captured image excludes one or more portions, leading to text cutoff can be advantageous and allow a user to retake the image immediately to ensure smooth document processing based on the captured image.

DESCRIPTION

This disclosure describes techniques for real-time detection of text cutoff in captured images. Text cutoff in document images refers to document images that have one or more text characters positioned either very close to the image boundary or only partially visible in the image frame. Per techniques of this disclosure, with user permission, optical character recognition (OCR) is applied to an input image of a document such as that of an identification document (ID), etc. to make a determination of whether the image includes portions where the text is partially missing and/or truncated. A feature vector is determined and utilized to represent the spatial location of the OCR symbols extracted from the document. The feature vector is provided to a trained classifier that predicts whether the document image includes portions of text cutoff.

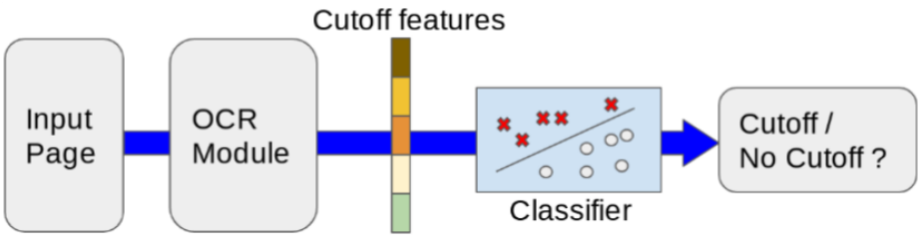

Fig. 1: Workflow to determine text cutoff in a document

Fig. 1 depicts an example workflow for the detection of document image text cutoff, per techniques of this disclosure. A captured image of a document (input page) is provided to an OCR module. A feature vector representative of cutoff features is determined based on the document and is provided to a classifier, which is utilized to make a determination whether the document includes text cutoff.

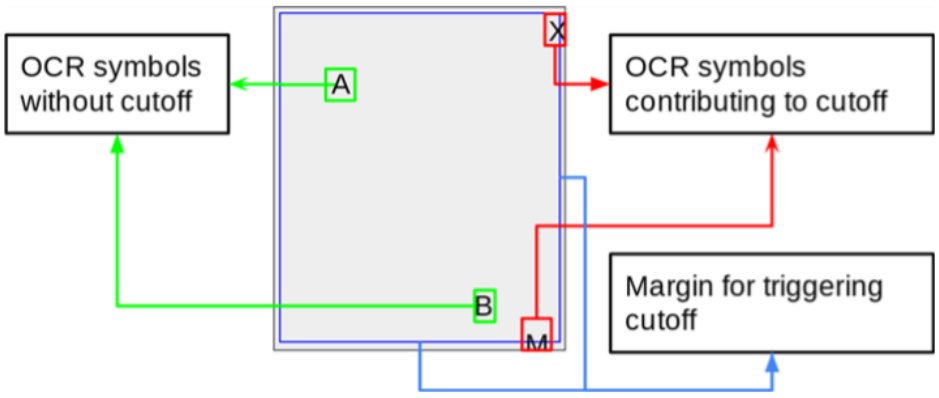

Fig. 2: Document margin is utilized to determine document text cutoff

Fig. 2 depicts an example document after it has been processed by an OCR module. A safe margin, $\Gamma_{cut} \epsilon [0, 1]$ is determined for the document image and utilized to denote a safe margin inside the image, beyond which an OCR symbol is considered to be cut off. The safe margin rectangle spanned by Γ_{cut} is depicted by the purple rectangle in Fig. 2. As seen in Fig. 2, the OCR module returns a bounding box for each text character (OCR symbol). The bounding box is defined by x (horizontal) and y (vertical) coordinates of four corners of the bounding

box. The height and width of the document can be normalized to a [0, 1] scale. The OCR bounding box is characterized by its coordinates $[x_{1:4}, y_{1:4}] \in [0, 1]$ that denote the location of each of its four corners.

The coordinates of the OCR bounding boxes are utilized to create a feature representation for text cutoff detection. A minimum and maximum of the horizontal and vertical coordinates is determined for the OCR characters detected in the document.

$$x_{min} = x_{i^*}; \quad i^* = \arg\min_i x_i$$
$$x_{max} = x_{i^*}; \quad i^* = \arg\max_i x_i$$
$$y_{min} = y_{i^*}; \quad i^* = \arg\min_i y_i$$
$$y_{max} = y_{i^*}; \quad i^* = \arg\max_i y_i$$

An OCR symbol can be considered to be cut off if its location is determined to be near to any one of the four image boundaries. An OCR bounding box is considered to be contributing towards cutoff if $x_{min} < \Gamma_{cut}$ or $y_{min} < \Gamma_{cut}$ or $x_{max} > 1-\Gamma_{cut}$ or $y_{max} > 1-\Gamma_{cut}$. In this illustrative example, the characters ‘A’ and ‘B’ are not cut off whereas the characters ‘X’ and ‘M’ are considered to be cut off based on their location relative to the safe margin.

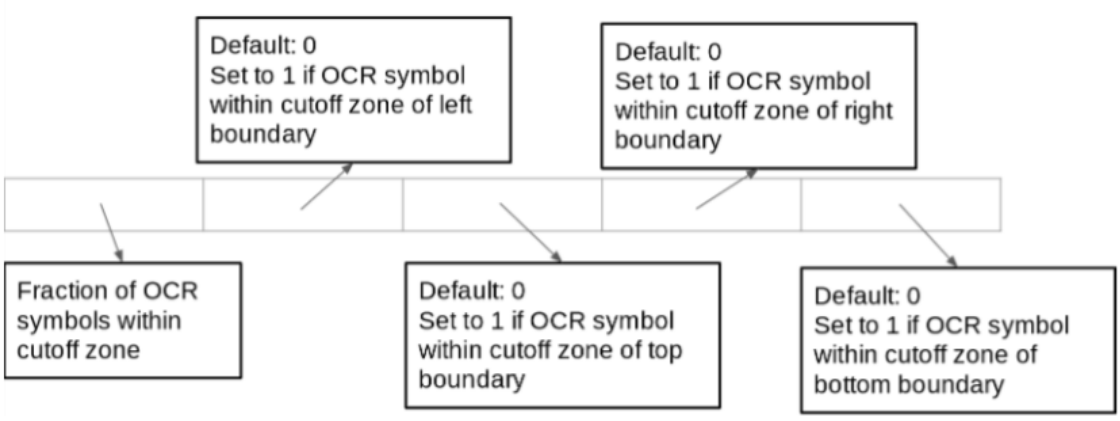

Fig. 3: Construction of feature vector based on OCR symbol coordinates

Fig. 3 depicts example feature vector construction based on OCR symbol coordinates, per techniques of this disclosure. A five-dimensional (5D) feature vector, f_{cut} , is utilized to represent cutoff. The components of f_{cut} are Boolean fields and are described below. A default value for all elements of the feature vector is set to 0.

- $f_{cut}[0]$: This element (field) represents the fraction of total OCR symbols which satisfy the cutoff condition.
- $f_{cut}[1]$: This element (field) is set to '1' if for any OCR box, $x_{min} < \Gamma_{cut}$
- $f_{cut}[2]$: This element (field) is set to '1' if for any OCR box, $y_{min} < \Gamma_{cut}$
- $f_{cut}[3]$: This element (field) is set to '1' if for any OCR box, $x_{max} > 1 - \Gamma_{cut}$
- $f_{cut}[4]$: This element (field) is set to '1' if for any OCR box, $y_{max} > 1 - \Gamma_{cut}$

The feature vector is provided to a trained classifier, e.g., a trained support vector machine (SVM) classifier in inference mode, to determine a class label for the input page, e.g., whether the page includes text cutoff. The classifier is trained based on training data that includes sets of text cutoff and non-cutoff documents. Feature vectors from documents that include document text cutoff and those without text cutoff are provided to the classifier which learns a decision boundary to distinguish text cutoff documents from non-cutoff documents. During the training phase, weights of the classifier are adjusted such that the classifier can correctly predict the class of an input document.

In some implementations, the area of an uploaded image that includes text, e.g., is covered by text, can be automatically determined. This information can be utilized for image enhancement operations, limiting the image area utilized in downstream document processing, etc. For example, a histogram equalization operation can be performed to increase the contrast

of an input image. With accurate textual area information, the enhancement is applied only to a portion of the entire image frame, thereby reducing an effective image area to be enhanced.

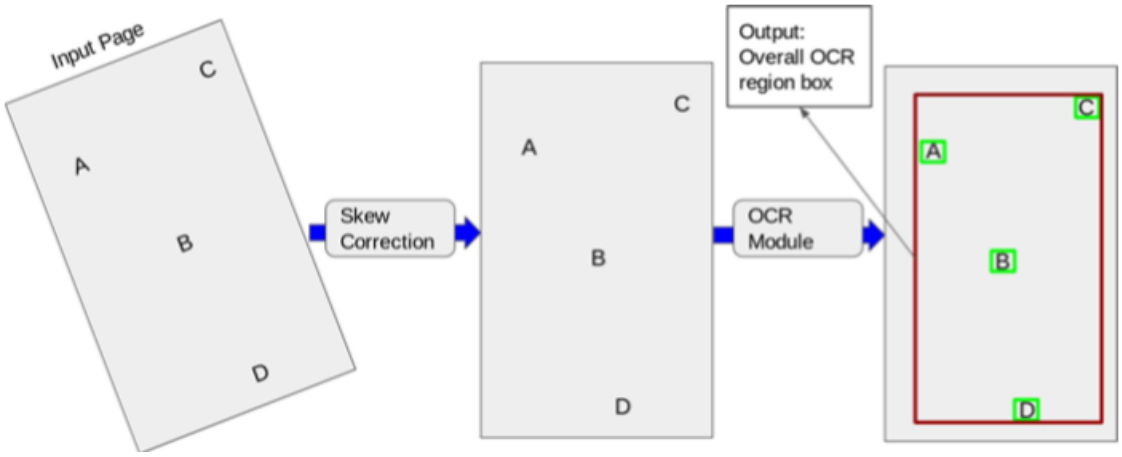

Fig. 4: Determination of portion of document containing text

Fig. 4 depicts example determination of the textual portion of a document based on the coordinates of the OCR bounding boxes, per techniques of this disclosure. Specifically, a rectangle is determined that covers all text regions of the image. The input document is de-skewed for frontalization. The de-skewed image is provided to an OCR module which returns OCR symbol bounding boxes for the detected text characters, which are utilized to fit a rectangle that encompasses all text in the document image.

In an illustrative example, there are a total of N OCR symbol bounding boxes detected. Each box, k, is the four corners, $[x^k_{1,2,3,4}, y^k_{1,2,3,4}]$. The overall OCR region bounding box, *B*, is defined by its top-left corner and the bottom-right corner (and depicted as the red rectangle in Fig.4), and whose coordinates are determined as provided below:

$$B_x^{top-left} = min(x_i^k);\ i \in \{1,2,3,4\}, k \in [1,2,..,N]$$
$$B_y^{top-left} = min(y_i^k);\ i \in \{1,2,3,4\}, k \in [1,2,..,N]$$
$$B_x^{bottom-right} = max(x_i^k);\ i \in \{1,2,3,4\}, k \in [1,2,..,N]$$
$$B_y^{bottom-right} = max(y_i^k);\ i \in \{1,2,3,4\}, k \in [1,2,..,N]$$

CONCLUSION

This disclosure describes techniques for detection of text cutoff in captured images of documents that include text. Optical character recognition (OCR) is applied to an input image. A bounding box for each text character (OCR symbol) is determined, defined by x and y coordinates of its four corners. A feature vector is determined and utilized to represent the spatial location of OCR symbols extracted from the image. The feature vector is constructed based on OCR symbol coordinates and is provided to a trained classifier to determine a class label for the input document, indicating whether the document includes text cutoff. Optionally, the area of an image that includes text is automatically determined and utilized to limit the area of the image utilized for downstream document processing.

REFERENCES

1. Kamola, Grzegorz, Michal Spytkowski, Mariusz Paradowski, and Urszula Markowska-Kaczmar. "Image-based logical document structure recognition." *Pattern Analysis and Applications* 18, no. 3 (2015): 651-665.