June 2020

Multibox Human Detection for Images of Crowded Scenes

Anonymous

Multibox Human Detection for Images of Crowded Scenes

ABSTRACT

Images of crowded scenes typically have been challenging for human-detection and pose-estimation algorithms. Top-down approaches suffer from reliance on non-maximum suppression (NMS) algorithms, which often remove valid detections, while bottom-up approaches inconsistently associate body parts of different people into the same detection. This disclosure presents techniques that combine elements of both top-down and bottom-up approaches, by leveraging the observation that head-boxes overlap less with each other as compared to body-boxes. NMS algorithms are applied to head-boxes instead of body-boxes. Head boxes are detected jointly, and are matched to the corresponding body-boxes. The techniques improve detection and pose estimation results for images of crowded scenes.

KEYWORDS

- Human detection
- Body-part detection
- Non-maximum suppression
- Multibox detector
- Head detection
- Crowded scene detection

BACKGROUND

Images of crowded scenes typically have overlapping people or body parts. For such images, human-detection or pose-estimation algorithms are often unable to accurately draw bounding boxes around humans in the image.

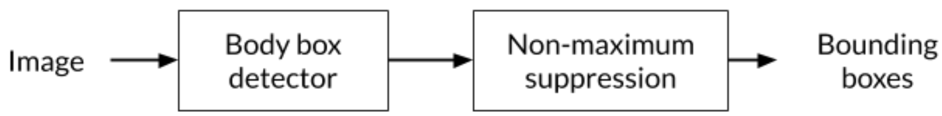
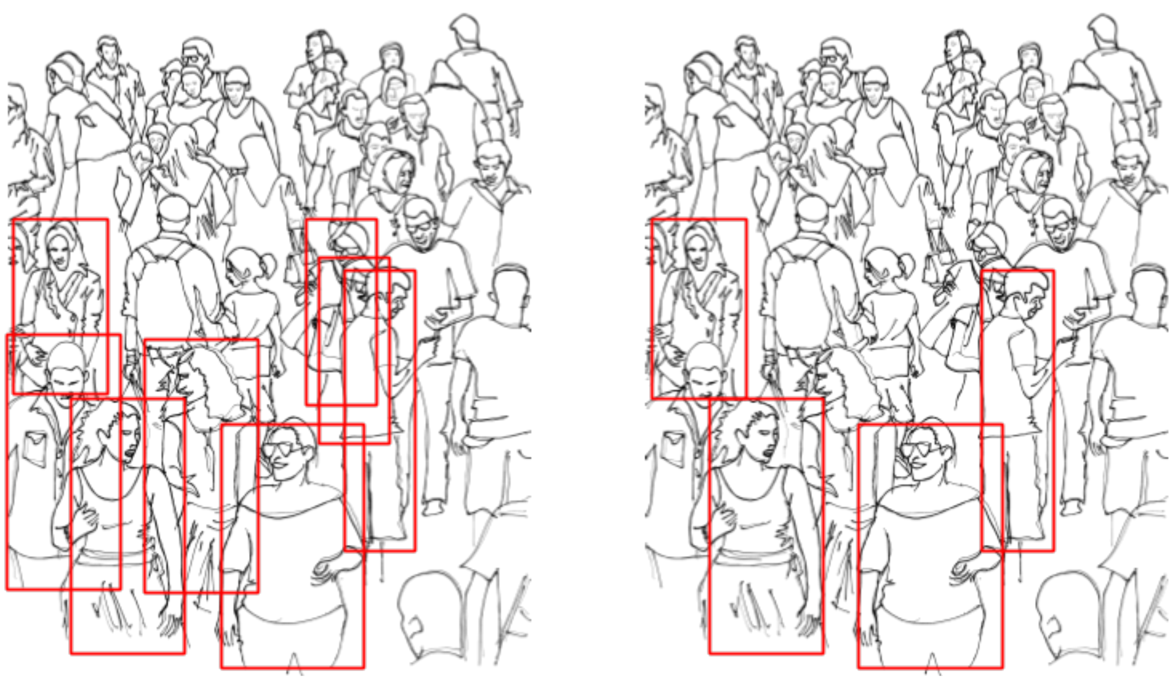


Fig. 1: A traditional top-down approach to human detection and pose estimation

As illustrated in Fig. 1, a follow-on algorithm, such as non-maximum suppression (NMS) is applied to detected body-boxes to remove highly overlapping detections to remove duplicate detections of the same person. As illustrated in Fig. 2, NMS, a top-down procedure, has the undesired side-effect of removing valid detections.



(a) **(b)**

Fig. 2: (a) Bounding boxes drawn across detected humans in a crowded scene (for clarity, not all detected humans are shown with bounding boxes); (b) Some valid bounding boxes are removed by NMS in an effort to remove duplicate detections of the same person.

DESCRIPTION

Per the techniques of this disclosure, an additional body-parts detector is utilized to detect body parts, e.g., head, torso, etc. within each body bounding-box. Rather than applying NMS on the body-boxes, NMS is applied to the body-part boxes, e.g., the head-boxes. The techniques leverage the observation that the head-boxes of people tend to overlap less than body-boxes.

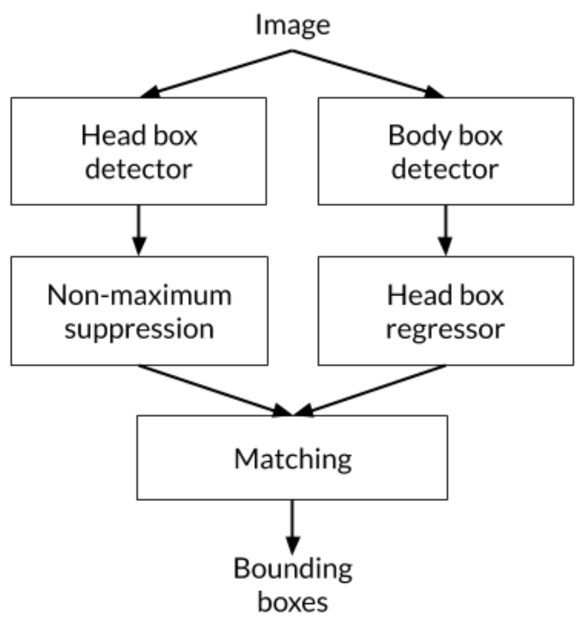
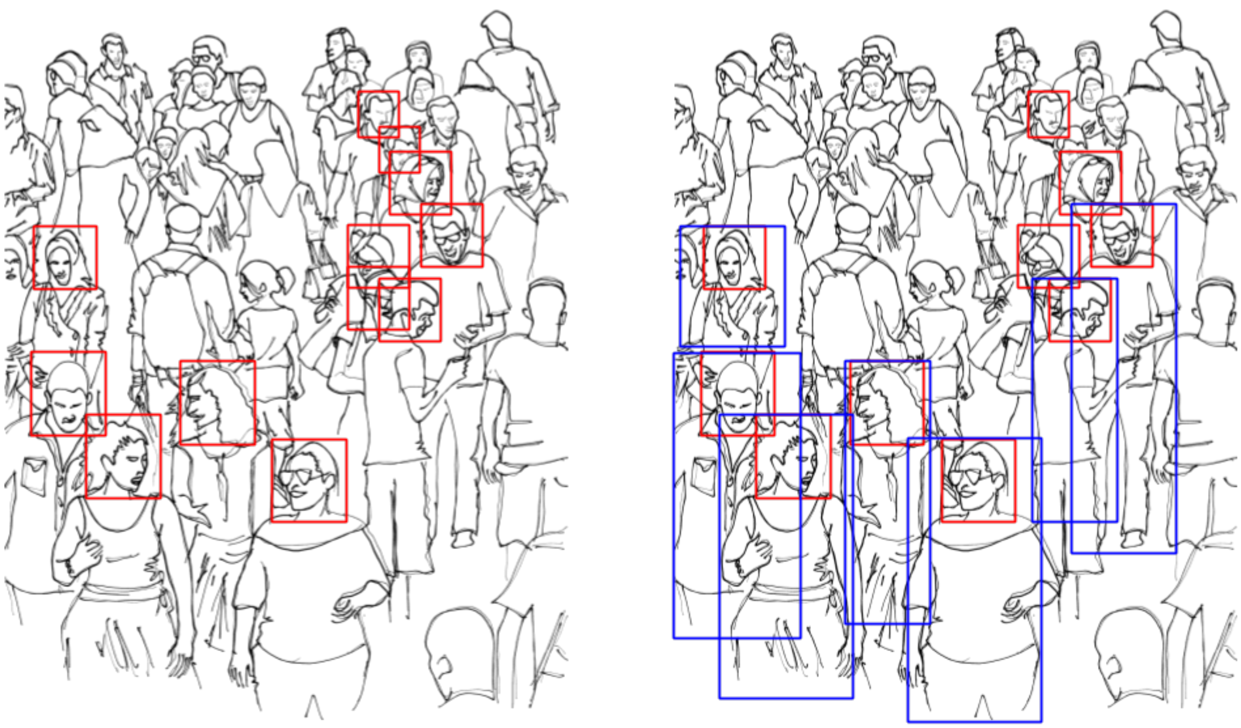


Fig. 3: Multibox Detector for Human Detection

As illustrated in Fig. 3, the techniques include the following:

- Given an input image, both head-boxes and body-boxes are detected using a multi-class object detector. At this point, there is no correspondence between the head-boxes and the body-boxes.
- A regressor is applied to the body-boxes to regress a body-box to a single head-box. As a body-box regresses to a head-box, a correspondence is established between the body-box and its head-box (as output by the head-box regressor).
- The head-boxes outputted by the multi-class object detector undergo non-maximum suppression to ensure that the remaining head-boxes don’t overlap by more than a certain threshold.

- The head-boxes outputted by the head-box regressor and the ones outputted by NMS are matched. In particular, head-boxes outputted by the head-box regressor (which have corresponding body-boxes), have their locations changed to the nearest (lowest intersection-over-union distance) head-box outputted by NMS.
- The final output bounding box is the head-box and the corresponding body-box, where the head-boxes assuredly do not overlap by more than a certain threshold.



(a) **(b)**

Fig. 4: (a) Detecting head-boxes (red); (b) Using NMS to remove head-boxes with an overlap greater than a certain threshold, and establishing a correspondence between head-boxes (red) and body-boxes (blue) (for clarity, not all body boxes are shown).

Fig. 4 illustrates an example image of a crowded scene where bounding boxes are determined using application of the techniques described herein. In Fig. 4(a) the head-boxes (red) are detected using a multi-class object detector. As expected given that the image depicts a crowded scene, the head-boxes do not overlap much with each other.

In Fig. 4(b), head-boxes that overlap beyond a certain threshold are removed using NMS. Body-boxes (blue) are attached to head-boxes with whom a correspondence has been established by the head-box regressor. In this manner, the described techniques result in robust human detection, especially for images that depict crowded scenes. Substantial improvements in average precision (mAP) and in average recall (mAR) can be achieved.

CONCLUSION

Images of crowded scenes typically have been challenging for human-detection and pose-estimation algorithms. Top-down approaches suffer from reliance on non-maximum suppression (NMS) algorithms, which often remove valid detections, while bottom-up approaches inconsistently associate body parts of different people into the same detection. This disclosure presents techniques that combine elements of both top-down and bottom-up approaches, by leveraging the observation that head-boxes overlap less with each other as compared to body-boxes. NMS algorithms are applied to head-boxes instead of body-boxes. Head boxes are detected jointly, and are matched to the corresponding body-boxes. The techniques improve detection and pose estimation results for images of crowded scenes.

REFERENCES

[1] Li, Xiaojie, Lu Yang, Qing Song, and Fuqiang Zhou. "Detector-in-Detector: Multi-Level Analysis for Human-Parts." *arXiv preprint arXiv:1902.07017* (2019).