Hand detection using multiple proposals

Arpit Mittal[1]
arpit@robots.ox.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk
Philip H. S. Torr[2]
philiptorr@brookes.ac.uk

[1] Department of Engineering Science,
University of Oxford, UK
[2] Department of Computing,
Oxford Brookes University, UK

The objective of this work is to detect and localise human hands in still images. This is a tremendously challenging task as hands can be very varied in shape and viewpoint, can be closed or open, can be partially occluded, can have different articulations of the fingers, can be grasping other objects or other hands, etc. Our motivation for this is that having a reliable hand detector facilitates many other tasks in human visual recognition, such as determining human layout and actions from static images. It also benefits human temporal analysis, such as recognizing sign language [2], gestures and activities in video.

In this paper we propose a detector using a two-stage hypothesize and classify framework. First, hand hypotheses are proposed from three independent methods: a sliding window hand-shape detector, a sliding window context-based detector, and a skin-based detector. The sliding window detectors employ the Felzenszwalb *et al.* [3]'s part based deformable model with three components. Then, the proposals are scored by all three methods and a discriminatively trained model is used to verify them. The three proposal mechanisms ensure good recall, and the discriminative classification ensures good precision. In addition, we develop a new method of non-maximum suppression based on super-pixels [1]. Figure 1 overviews the detector.

Hand dataset. We have collected a comprehensive dataset of hand images (Table 1) from various public image sources (`http://www.robots.ox.ac.uk/~vgg/data/hands/`). In each image, all the hands that can be perceived clearly by humans are annotated. The annotations consist of a bounding rectangle, which does not have to be axis aligned, oriented with respect to the wrist.

	Training set	Validation set	Test set
# hand instances	9163	1856	2031
# bigger hand instances	2861	649	660

Table 1: Statistics of hand dataset. A hand instance is 'big' if area of the bounding box is greater than 1500 sq. pixels. Bigger hand instances are used for experimental evaluations.

The multiple proposal model is trained on the hand train/val set of the hand dataset, and then evaluated on several others, including: the test set of the hand dataset; the Signer dataset [2]; and the PASCAL VOC 2010 person layout test dataset. We outperform existing methods, and beat our previous winning entry in the VOC 2010 competition by a factor of two. Figure 2 shows some sample results.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Proc. CVPR*, 2009.

[2] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC.*, 2008.

[3] P. F. Felzenszwalb, R. B. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010.

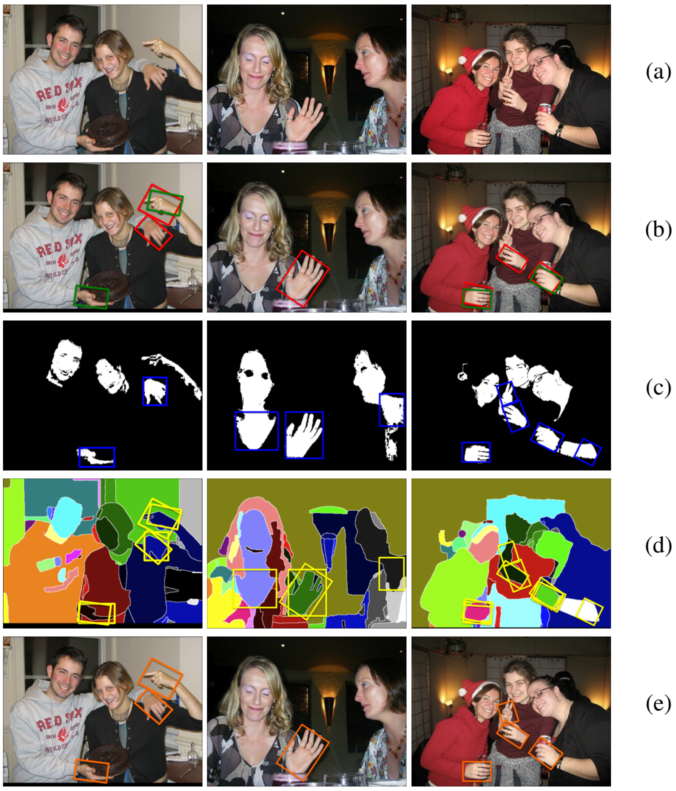

Figure 1: **Overview of the method.** (a) Original image. (b) Some of the hypotheses proposed by hand and context detector. Bounding boxes in 'Red' are proposed by the hand detector and 'Green' bounding boxes are proposed by the context detector. (c) Skin detection and hypotheses generation. (d) Super-pixel segmentation of the image with combined hypothesised bounding boxes from the three proposal schemes. Using super-pixel based non-maximum suppression (NMS), overlapping bounding boxes are suppressed. (e) Final detection after post-processing.

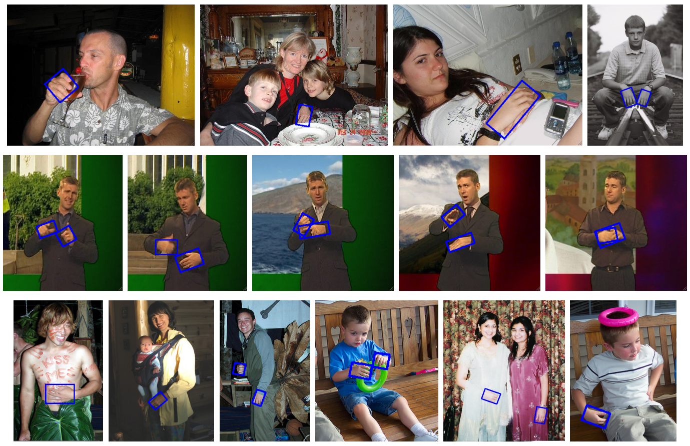

Figure 2: Examples of high-scoring detections on the three datasets. Top row consists of images from our hand dataset. Middle row has images from the Signer dataset. Bottom row represents the images from PASCAL VOC 2010 person layout test set.