Contextual Attention for Hand Detection in the Wild

Supreeth Narasimhaswamy* , Zhengwei Wei*, Yang Wang, Justin Zhang, and Minh Hoai
Department of Computer Science, Stony Brook University

Stony Brook University

ICCV 2019 Seoul, Korea

Contributions

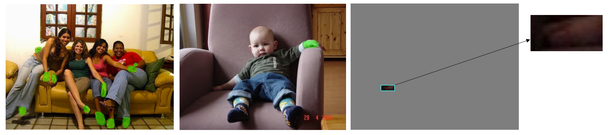
- A state-of-the-art Hand Detector
- Two large scale annotated hand datasets
- An attention method for object detection

Goal

- To improve the hand detection performance of OpenPose and Mask-RCNN on unconstrained images with large variation in appearance of hands (close-up shots, occlusions, motion blur)
- OpenPose:

- Ours:

- Mask-RCNN:



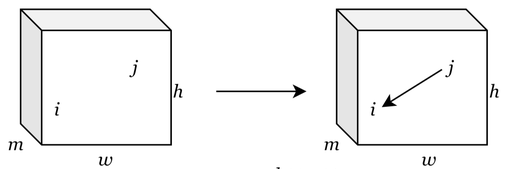
- Ours:

- Quantitative Comparison:

Method	AP
DPM (Girshik *et al.*, CVPR 2015)	36.8%
ST-CNN (Jaderberg *et al.*, NIPS 2015)	40.6%
RCNN (Girshik *et al.*, CVPR 2014)	42.3%
Context + Skin (Mittal *et al.*, BMVC 2011)	48.2%
RCNN + Skin (Roy *et al.*, ICCV 2017)	49.5%
FasterRCNN (Ren *et al.*, NIPS 2015)	55.7%
Rotation Network (Deng *et al.*, TIP 2018)	58.1%
Hand Keypoint (Simon *et al.*, CVPR 2017)	68.6%
Hand-CNN (proposed)	**78.8%**

Contextual Attention

- We propose a method to incorporate contextual cues during the detection process. This is based on two types of non-local contextual pooling: (1) feature similarity (2) spatial relationships between semantically related entities.



- Given a 3D feature map $\mathbf{X} \in \mathbb{R}^{h \times w \times m}$ our method computes a contextual feature map $\mathbf{Y} \in \mathbb{R}^{h \times w \times m}$ such that

$$\mathbf{y}_i = \sum_{j=1}^{hw} \left[\underbrace{\frac{f(\mathbf{x}_i, \mathbf{x}_j)}{C(\mathbf{x}_i)}}_{\text{similarity context}} + \underbrace{\sum_{k=1}^{K} \alpha_k \; p_k(\mathbf{x}_j) \; \exp\left(-\frac{(d_{ij} - \mu_k)^2}{\sigma_k^2} \right)}_{\text{semantics context}} \right] g(\mathbf{x}_j)$$

- The contextual attention module can be inserted in detection networks and the parameters of the attention module can be learned together with the other parameters of the detector end-to-end.

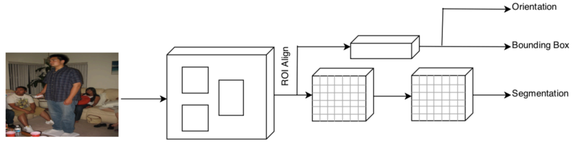
Orientation

- We extend Mask-RCNN to include an additional network branch to predict hand orientation
- Orientation branch shares weights with other branches, so it does not incur significant computational expenses
- Orientation Loss:

$$L_{ori}(\theta, \theta^*) = |arctan2(sin(\theta - \theta^*), cos(\theta - \theta^*))|$$

- Total Loss:

$$L = L_{RPN} + L_{BRN} + L_{mask} + \lambda L_{ori}$$



Datasets

- TV-Hand:
  - Image frames extracted from video clips of ActionThread dataset, and contains 8.5K images with 9.5K hands.
- COCO-Hand:
  - Images from a subset of Microsoft COCO dataset, and has around 26K images with 45K hands.

Name	Scope	# images	Label
EgoHands	Google glasses	4,800	Manual
Handseg	Color gloves	210,000	Auto
NYUHands	Three subjects	6,736	Auto
WorkingHands	Three subjects	7,905	Man.+Syn.
ColorHandPose	Specific poses	43,986	Synthetic
HandNet	Ten subjects	212,928	Auto
GTEA	Four subjects	663	Manual
Oxford-Hand	Unconstrained	2686	Manual
TV-Hand	Unconstrained	9498	Manual
COCO-Hand-S	Unconstrained	4534	Semiauto
COCO-Hand	Unconstrained	26499	Semiauto

Quantitative Results

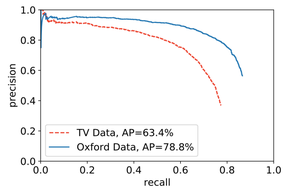
- Benefits of Context:

Method	Oxford-Hand	TV-Hand
MaskRCNN	69.9%	59.9%
Hand-CNN	73.0%	60.3%
Hand-CNN w/o semantic context	71.4%	59.4%
Hand-CNN w/o similarity context	70.8%	59.6%

- Benefits of Data:

	Test Data	
Train Data	Oxford-Hand	TV-Hand
TV-Hand	62.5%	55.4%
TV-Hand + COCO-Hand-S	69.9%	59.9%
TV-Hand + COCO-Hand	76.7%	63.5%

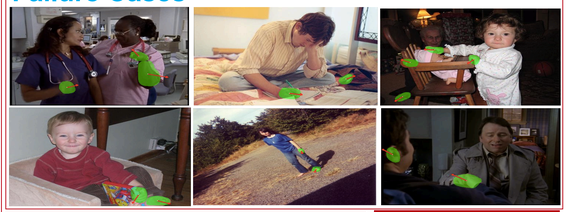
- Precision Recall, Orientation Performance:



	Prediction error in angle		
Test Data	$\leq 10°$	$\leq 20°$	$\leq 30°$
Oxford-Hand	41.26%	64.49%	75.97%
TV-Hand	37.65%	60.09%	73.50%

Qualitative Results

Failure Cases



Acknowledgements. This work was partially supported by the VinAI research and NSF IIS-1763981. The authors would also like to thank Tomas Simon for his suggestion about the COCO dataset and Rakshit Gautam for his contribution to the data annotation process.

Code, Data