25th International Conference on Pattern Recognition. January 10-15, 2021

StrongPose: Bottom-up and Strong Keypoint Heat Map Based Pose Estimation

Niaz Ahmad and Jongwon Yoon

Hanyang University, South Korea

Abstract

- We present *StrongPose* system that detects strong keypoint heat maps and predicts their comparative displacements, allowing keypoints to be grouped into human instances.
- *StrongPose* utilizes the keypoints to generate body heat maps that can determine the position of the human body in the image.
- Evaluation results show that our framework achieves average precision of 0.708 using ResNet-101 and 0.725 using ResNet-152, outperforms prior bottom-up frameworks.

Motivation

- Human pose estimation allows for higher-level reasoning in the context of human-computer interaction and activity recognition.
- Accurate keypoint localization for pose estimation can increase the reliability of applications that require human detection.
- The following figures shows some challenging scenarios for keypoints identification e.g., overlapping, occlusion and tangled keypoints.

StrongPose System

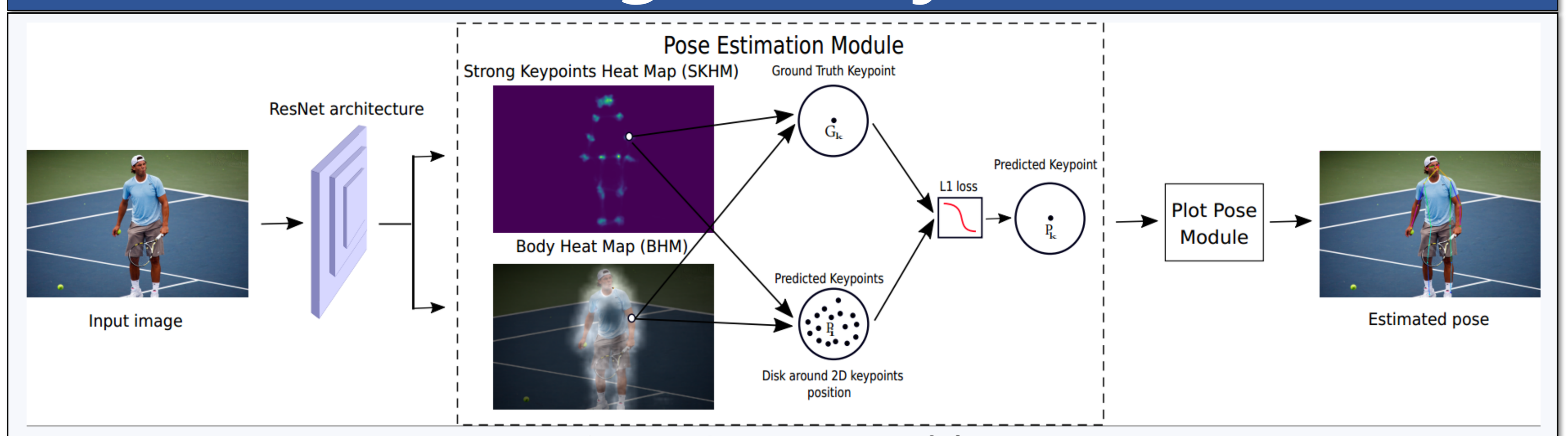
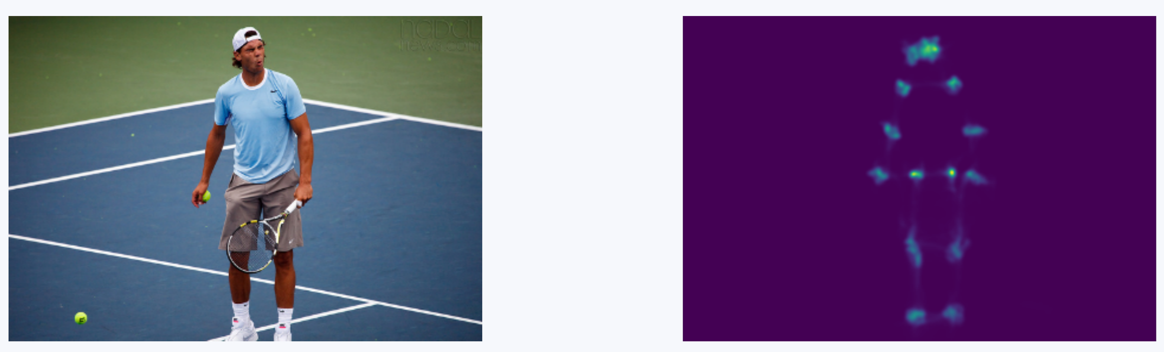


Figure 1. StrongPose model

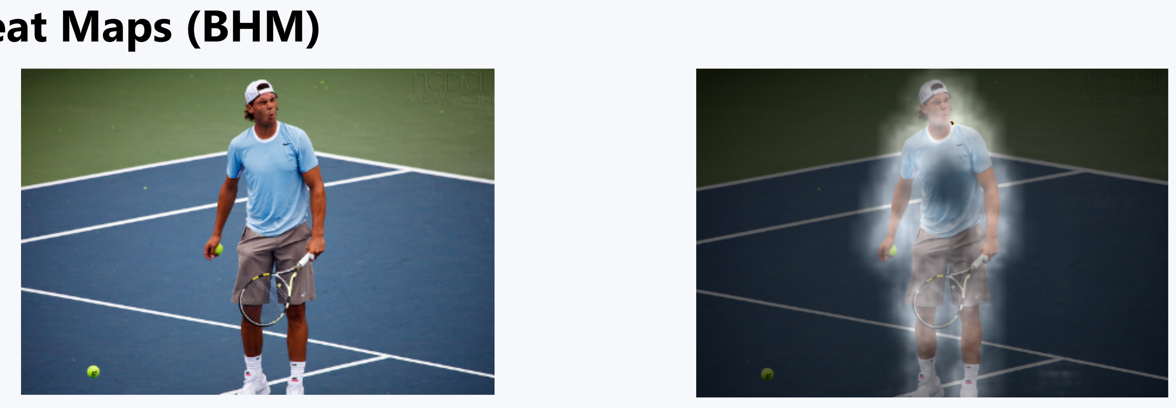
- Figure 1 depicts an overview of the *StrogPose* system consist of pose estimation module and Plot Pose Module.
- The system uses ResNet as a backbone network and follows bottom-up approach for keypoint identification.

Strong Keypoint Heat Map (SKHM)



- The SKHM is generated for all hard and soft keypoints.
- The role of the SKHM is to correctly localize and produce the heat map for each keypoint. Details are explained in bellow:
  - Suppose p_i (2D keypoint position)
  - Let $D_R(q) = p{:}\,||p - q|| \le R$(R is a disk radius centered around q)
  - $R = 16$ pixels
  - Let $q_{j,k}$(2D position of k-th keypoint of the j-th person instance)
  - Predicted Keypoint heatmap $pk(h)$ =1 if h ϵ $D_R(q_{j,k})$otherwise $pk(h)$=0

Body Heat Maps (BHM)



- The BHM is generated on the same manner as the SKHM.
- BHM helps to correctly localize the human body in the image.
- BHM also provides us the advantage of person detection without the concept of a bounding box.

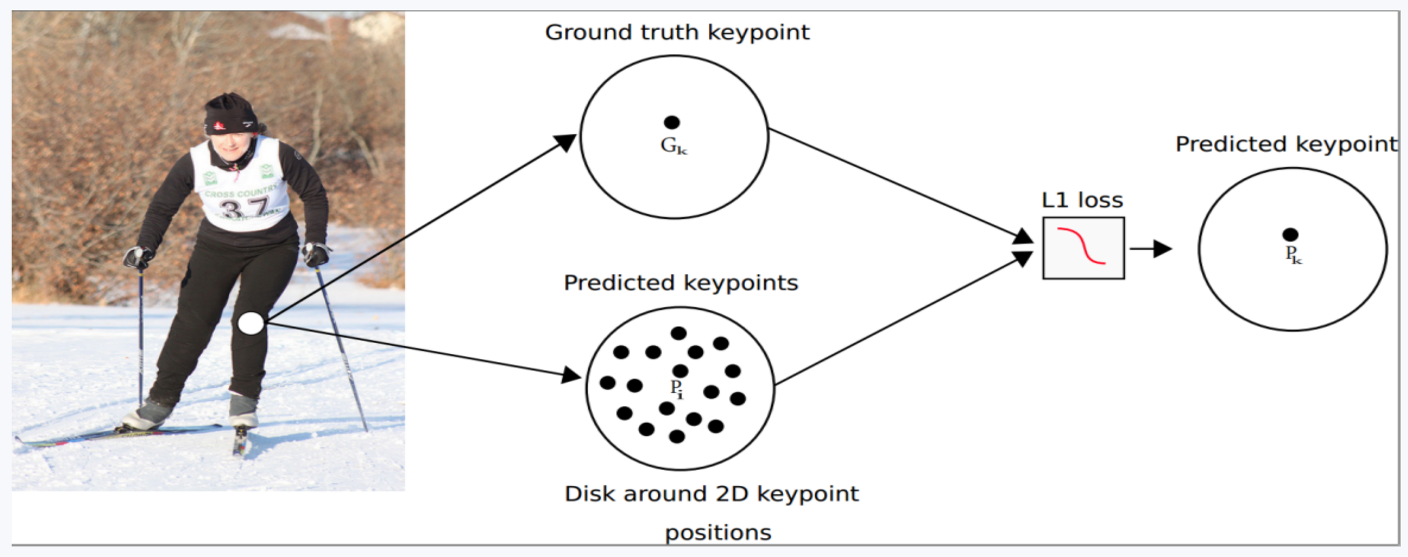
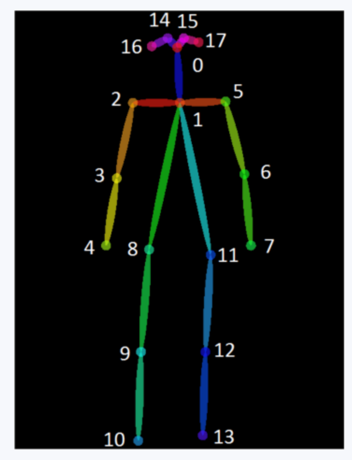


Figure 2. Keypoint disk around keypoint positions

- To increase the keypoint localization precision we predict keypoint offset vector $vk(x)$ for each keypoint.
- The purpose of $vk(x)$ is to compare the predicted 2D keypoint position P_i with the ground truth position G_k in the keypoint disk.
- Figure 2 illustrates the loss between P_i and G_k panelized by $L1$ loss.

Pose plot

- The pose plot module defines associations between the keypoints as tuples.
- It groups all the associated keypoints into a human instance.



Performance Evaluation

Table.1 shows evaluation on Val2017 Dataset.

- 5.7 % in AP compare to Hourglass
- 2.2 % in AP compare to CPN
- 1.2 % in AP compare to CPN (OHKM)
- 11.0 % in AP compare to CMU-Pose
- 4.1 % in AP compare to PersonLab

Table. 1 Performance on COCO keypoint Val2017 Dataset

Method	Backbone	Input Size	OHKM	AP	AR
Top-down:					
8-stage Hourglass	-	156 x 192	×	0.669	-
8-stage Hourglass	-	156 x 156	×	0.671	-
CPN	ResNet-50	256 x 192	×	0.686	-
CPN	ResNet-50	384 x 288	×	0.706	-
CPN	ResNet-50	256 x 192	✓	0.694	-
CPN	ResNet-50	384 x 288	✓	0.716	-
HRNet-W48	HRNet-W48	384 x 288	×	0.763	0.812
Bottom-up:					
CMU-Pose	-	-	×	0.618	-
PersonLab(single-Scale)	ResNet-152	-	×	0.665	0.707
PersonLab(multi-scale)	ResNet-152	-	×	0.687	-
StrongPose	ResNet-101	-	×	0.690	0.757
StrongPose	ResNet-152	-	×	0.728	0.800

Table.2 shows evaluation on Test2017 Dataset.

- 9.4 % in AP compare to Mask-RCNN
- 7.6 % in AP compare to G-RMI
- 0.4 % in AP compare to CPN
- 10.7 % in AP compare to CMU-Pose
- 7.0 % in AP compare to Assoc
- 3.8 % in AP compare to PersonLab
- 2.9 % in AP compare to MultiPoseNet

Table. 2 Performance on COCO keypoint Test2017 Dataset

Method	AP	$AP^{.50}$	$AP^{.75}$	AP^{M}	AP^{L}
Top-down:					
Mask-RCNN	0.631	0.873	0.687	0.578	0.714
G-RMI COCO-only	0.649	0.855	0.713	0.623	0.700
CPN	0.721	0.914	0.800	0.687	0.772
Bottom-up:					
CMU-Pose (+refine)	0.618	0.849	0.675	0.571	0.682
Assoc. Embed(single-Scale)	0.630	0.857	0.689	0.580	0.704
Assoc. Embed(mscale,refine)	0.655	0.879	0.777	0.690	0.752
PersonLab (single-scale)	0.665	0.880	0.726	0.624	0.723
PersonLab (multi-scale)	0.687	0.890	0.754	0.641	0.755
MultiPoseNet	0.696	0.863	0.766	0.650	0.763
StrongPose:					
ResNet101	0.708	0.889	0.752	0.652	0.753
ResNet152	0.725	0.891	0.778	0.671	0.762

Visualization Results

- Single-person pose estimation

- Multi-person pose estimation

Conclusion and Future Work

- We proposed a bottom-up model, *StrongPose*, that jointly tackle the problem of pose estimation and person detection.
- The effectiveness of the proposed model is evaluated using the COCO 2017 keypoint challenging dataset. Evaluatation results show significant increase in AP compare to other models.
- In the future work we will enhance the *StrongPose* system to understand human body language and capturing their actions in live environment.