Stacked Hourglass Networks for Human Pose Estimation

Alejandro Newell, Kaiyu Yang, and Jia Deng

University of Michigan, Ann Arbor

ECCV 2016

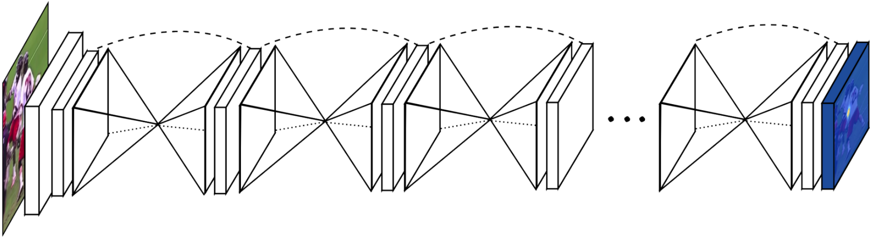
Presenter: Van-Thanh Hoang

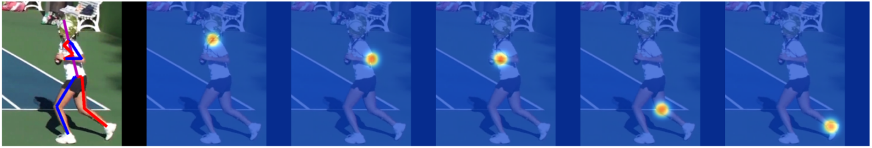
thanhhv@islab.ulsan.ac.kr

December, 16, 2017

Overview

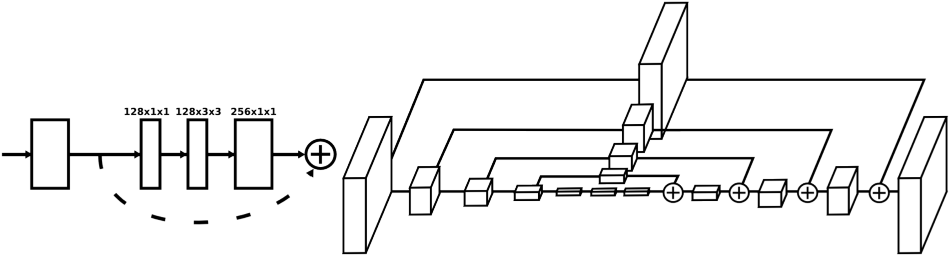
- Problem:
  - Detect position of human body joints in an image
- Proposed a novel architecture called Stacked Hourglass





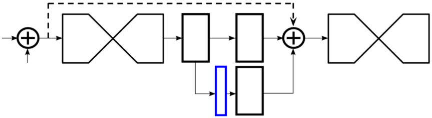
Hourglass Design

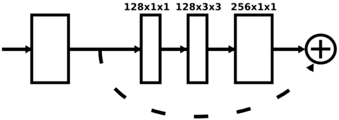
- Motivations: need to capture information at every scale
- Set up HG modules
  - Convolutional and max pooling layers are used to process features down to a very low resolution
  - After reaching the lowest resolution, the network begins the sequence of upsampling and combination of features across scales
  - No Conv layers have filter greater than 3×3

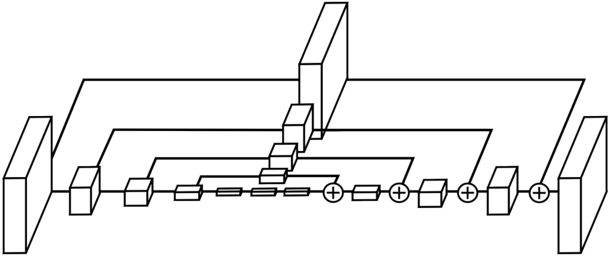


Stacked Hourglass with Intermediate Supervision

- Stacking multiple hourglasses
- Feeding the output of one as input into the next
- Loss is applied to the predictions of all hourglasses using the same ground truth.





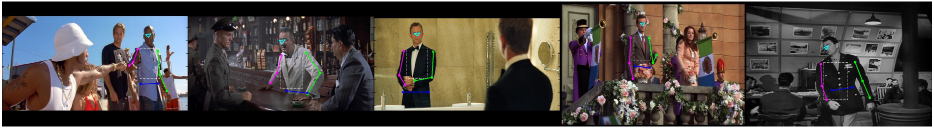


Configurations (1/4)

- Running Information
  - NVIDIA TitanX GPU with 12 GB
  - Network has 8 HG modules
  - Input images are resized to 256×256 pixels
  - Do data augmentation with
    - Rotation $(+/-30$ degrees)
    - Scaling $(.75 - 1.25)$
  - Using Torch7 framework
  - Training takes 3 days
  - A single forward pass takes 75 ms
  - Result of an image is the average of the heatmaps of origin input and the flipped version (1% improvement)

Configurations (2/4)

- Datasets
  - Frames Labeled In Cinema (FLIC) (https://bensapp.github.io/flic-dataset.html)
    - 5003 images (3987 training, 1016 testing)
    - Taken from films.



  - MPII Human Pose
    - 25k images
    - 40k annotated samples (28k training, 11k testing)

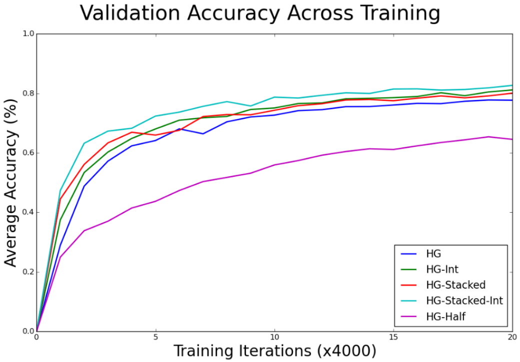
Configurations (3/4)

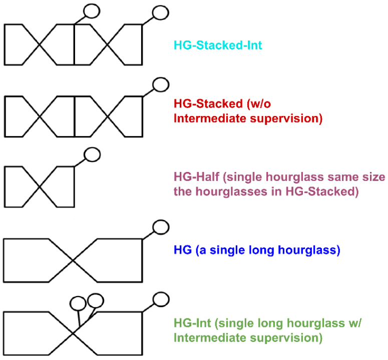
- MPII Human Pose examples

Configurations (4/4)

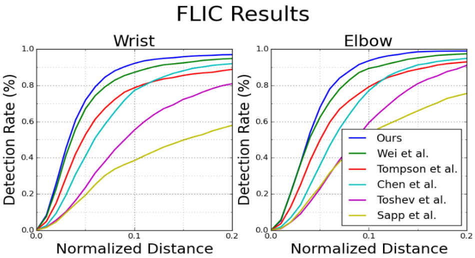
- Using Percentage of Correct Keypoints (PCK) metric
- A candidate keypoint to be correct if it falls within $\alpha \cdot \max(h, w)$ pixels of the groundtruth keypoint, where h and w are the height and width of the bounding box of human (usually use torse)
- PCKh: using head size instead of bounding box size

- Comparison of training with different types of HG network





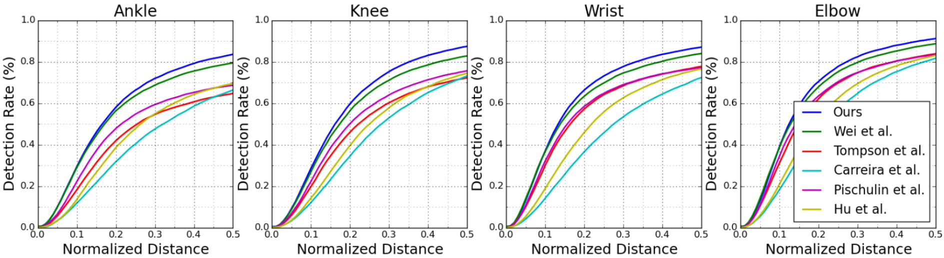
- Experiments on FLIC (PCK@0.2)



	Elbow	Wrist
Sapp et al. [1]	76.5	59.1
Toshev et al. [24]	92.3	82.0
Tompson et al. [16]	93.1	89.0
Chen et al. [25]	95.3	92.4
Wei et al. [18]	97.6	95.0
Our model	**99.0**	**97.0**

Evaluation (3/4)

- Experiments on MPII (PCKh@0.5)



	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Tompson et al. [16], CVPR'15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [19], CVPR'16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Pishchulin et al. [17], CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Hu et al. [27], CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Wei et al. [18], CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Our model	**98.2**	**96.3**	**91.2**	**87.1**	**90.1**	**87.4**	**83.6**	**90.9**

Evaluation (4/4)

- Failure in case of multiple people
- Can be fail if there is a slight translation and/or change of scale of the input image
- Reasons:
  - Network is trained for estimate pose of single person
  - Person is in the center of training images

Conclusion

- Proposed a new convolutional network architecture called Stacked Hourglass Network for human pose estimation task
  - Achieve state-of-the-art results
  - Can capture information in many scales

- Comments
  - Weakness:
    - Detect single person
    - Result depends on how good of people detector
  - Good idea for capture information in every scales

Thank you for your attention!