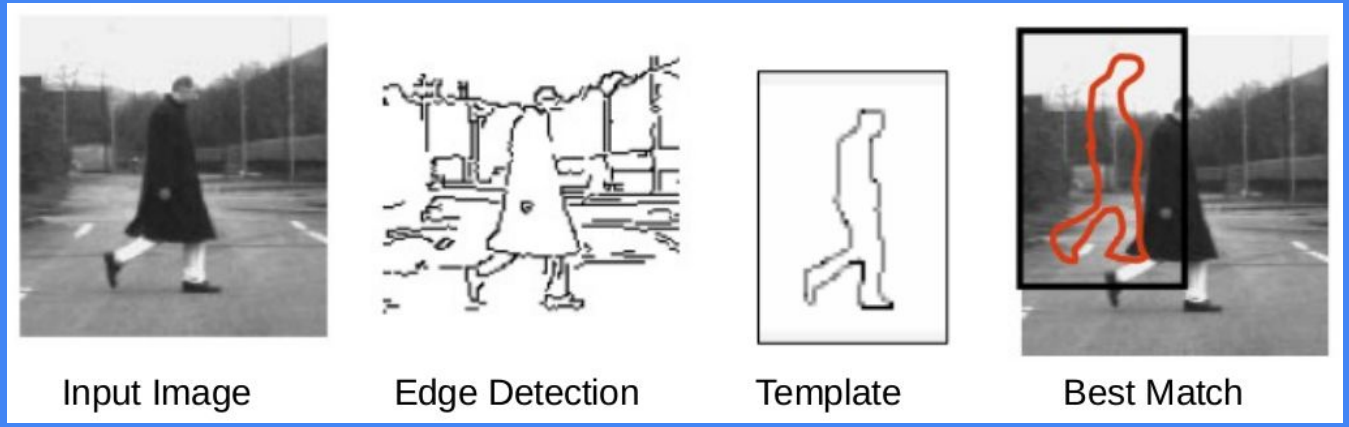
Machine Learning for Human Recognition in 2D: A brief history of pose estimation

Objective: 2D Human Pose Estimation

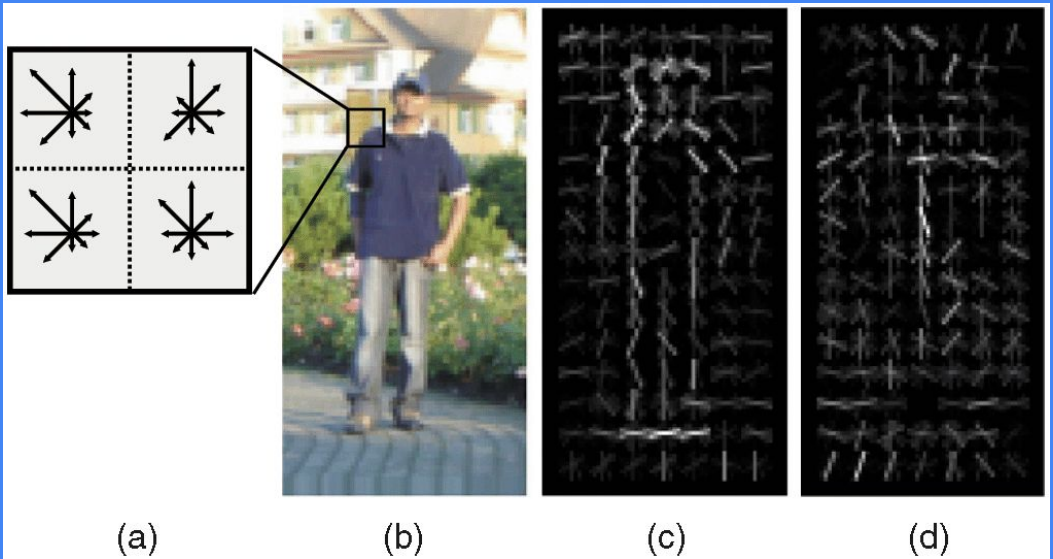
Newell, Yang and Deng. “Stacked Hourglass Networks for Human Pose Estimation”, Mar. 2016. https://arxiv.org/abs/1603.06937

Really old school (1999): Chamfer matching for pedestrian detection



From slides by K. Grauman and B. Leibe

The templates get a little smarter: HoG descriptors (2005, Dalal and Triggs)



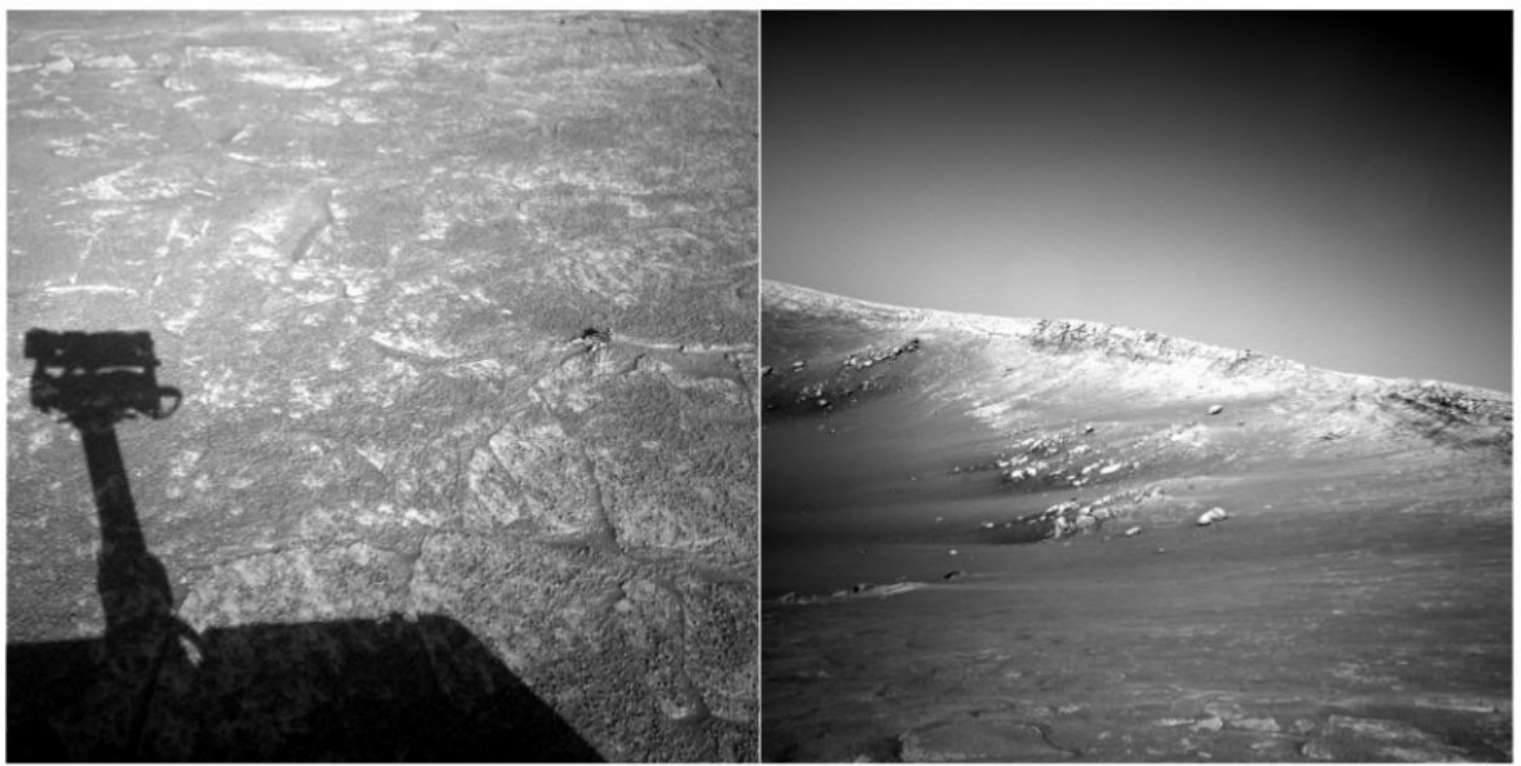
Histogram of Gradients (HoG):

The image is divided into 8x8 pixel cells. From each cell we accumulate a 1D histogram of gradient orientations over pixels in the cell.

The histograms capture local shape properties but are also somewhat invariant to small deformations.

Although they seem naïve, gradient-based descriptors can be quite robust (for instance, to changes in illumination). A quick example of the closely-related SIFT descriptor:

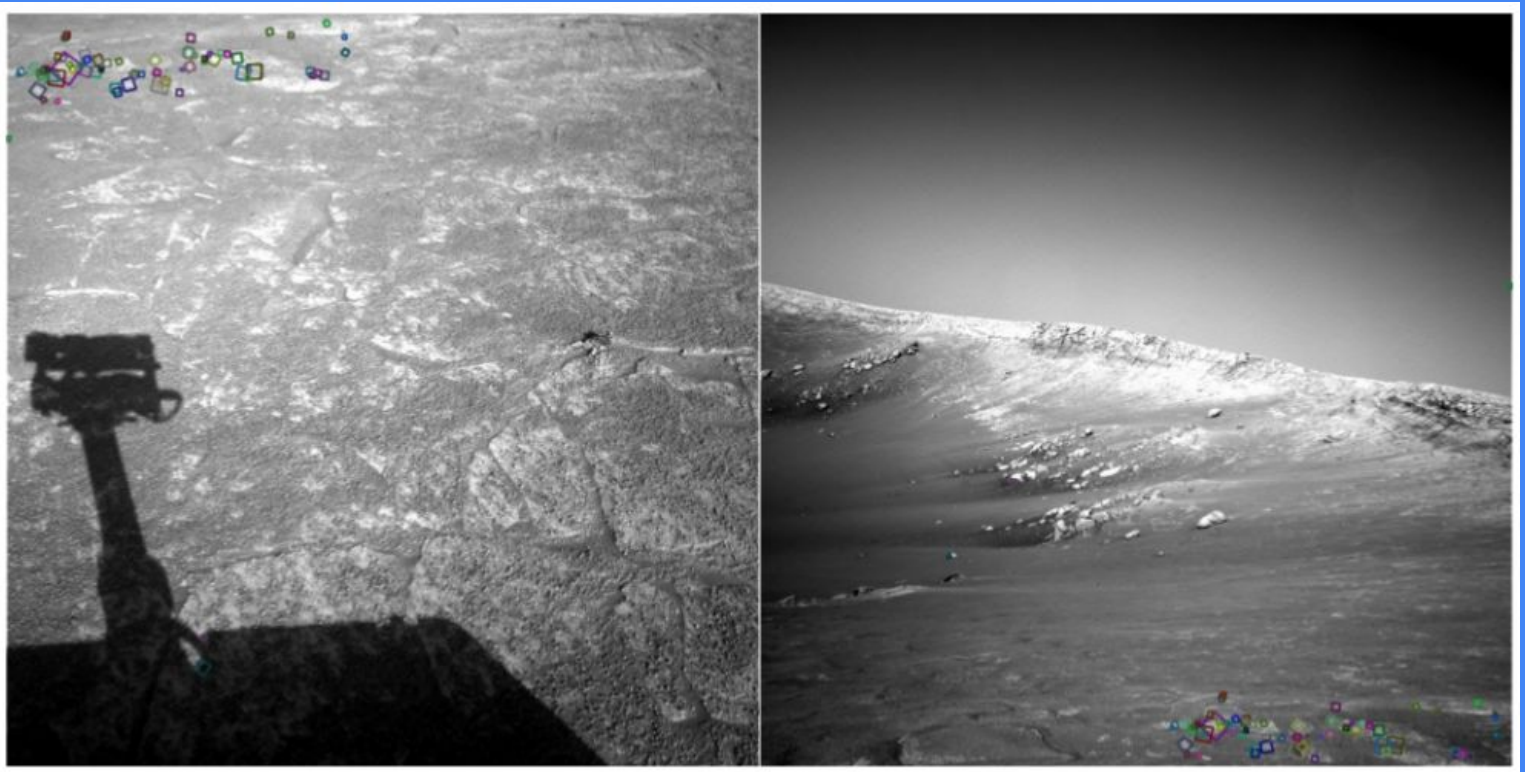
SIFT descriptors: matching landscape features



NASA Mars Rover images

Slide credit: S. Lazebnik

SIFT descriptors: matching landscape features



NASA Mars Rover images
with SIFT feature matches
Figure by Noah Snavely

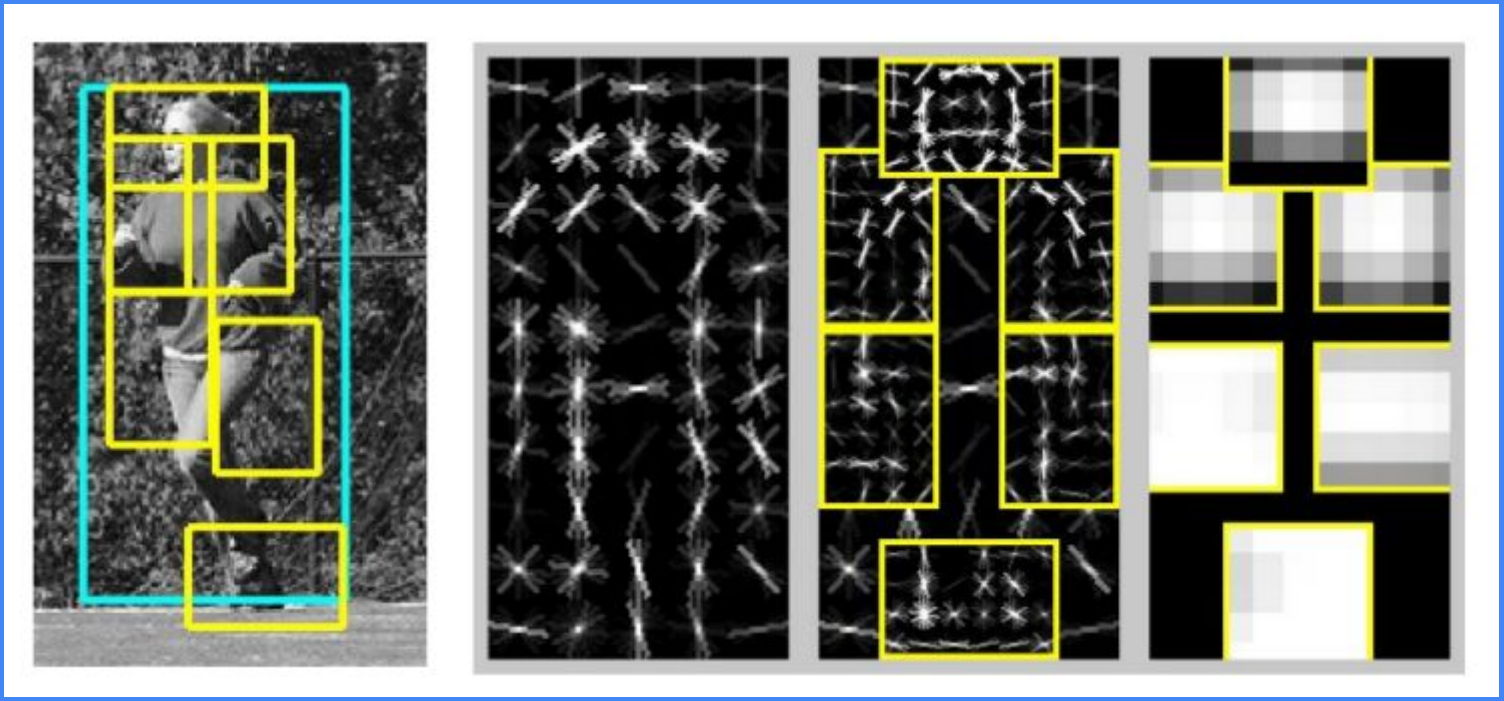
But, humans are deformable.

On challenging datasets composed of real-world images (PASCAL VOC 2007), rigid template-based models give poor results. Dalal and Triggs obtained an Average Precision (AP) score of only 12%.

Given limitations on annotated images, it's simply infeasible to construct enough rigid templates to handle a wide variety of people in different conditions.

Slide credit: Ross Girshick

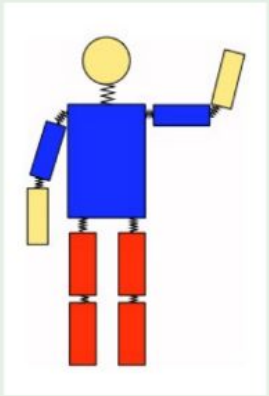
Adding flexibility:
Deformable Parts Models (DPM), 2008



R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, 2008.

Part-based Models

- Parts — local appearance templates
- "Springs" — spatial connections between parts (geom. prior)



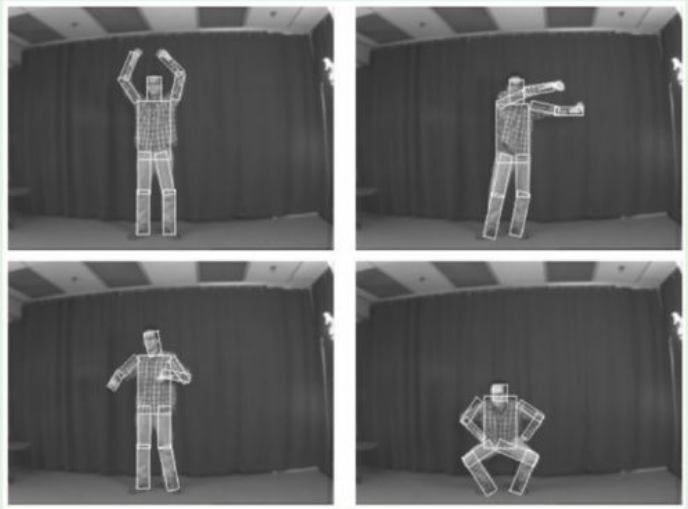


Image: [Felzenszwalb and Huttenlocher 05]

AP score jumps from 12% to 34%!

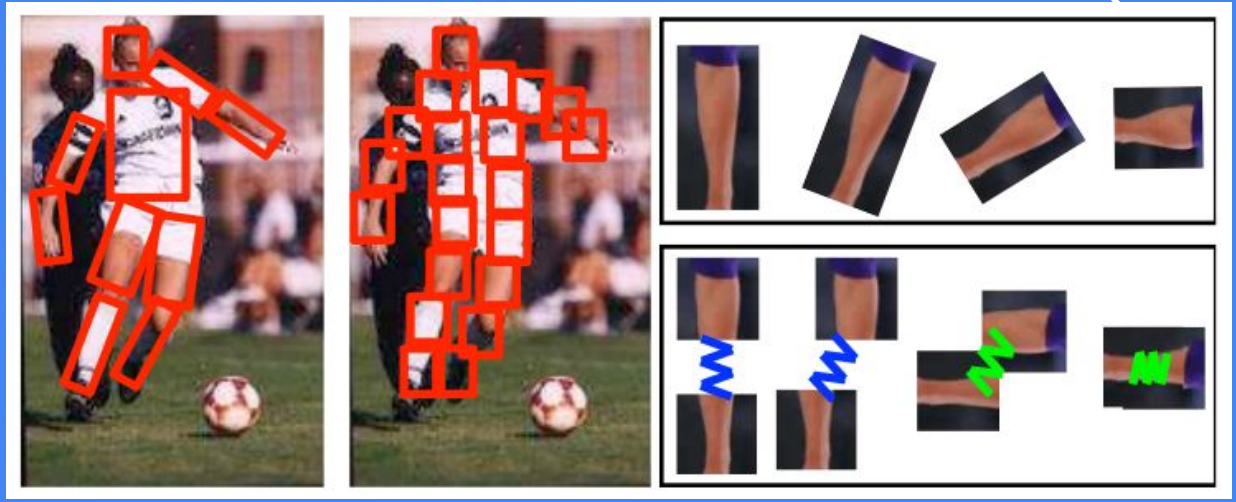
Some issues:

1) How do we deal with limb foreshortening? Do we warp templates to every possible orientation?

2) We're still not capturing **global** information.

Slide credit: Ross Girshick

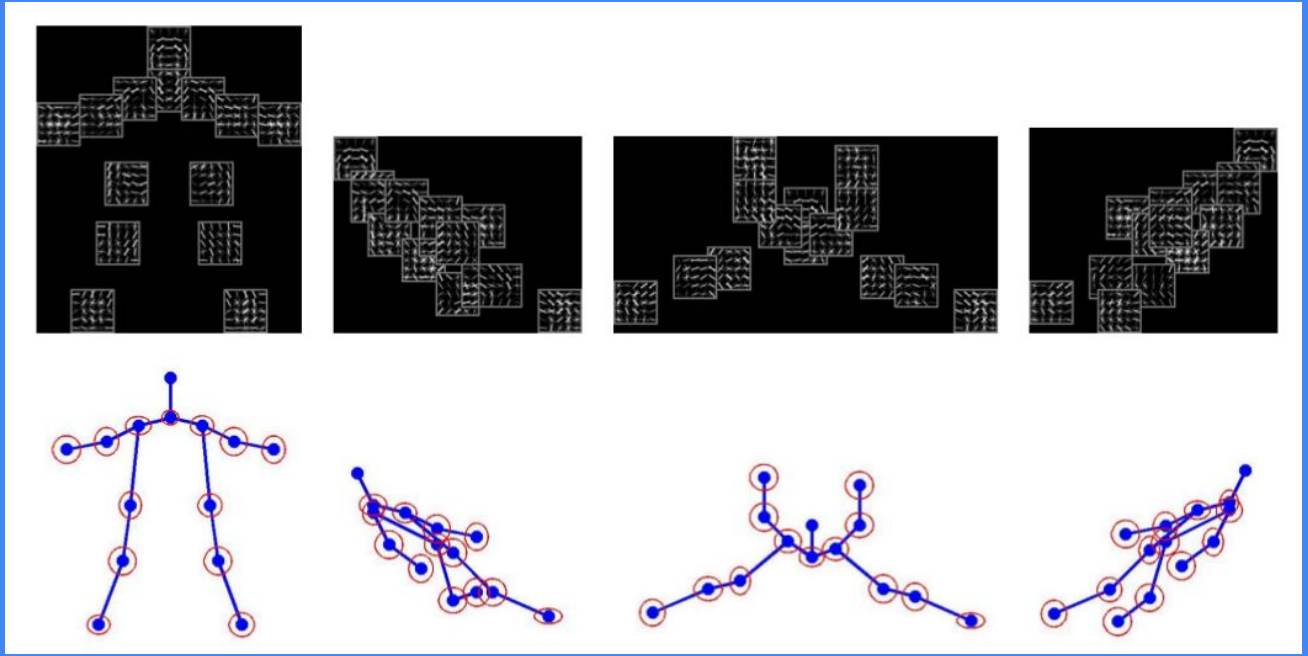
Flexible Mixtures-of-Parts Models (2012)



- **The problem**: We need to capture both local and global information. At the same time, it needs to be efficient!
- **Solution**: we can approximate warps of limbs using many smaller parts with spring-type connections. And, we'll capture global geometry by using different ***types*** of parts (orientations).
- The score function is similar to the DPM score, but accounts for likelihood of types of parts, as well as their co-occurrence likelihood.

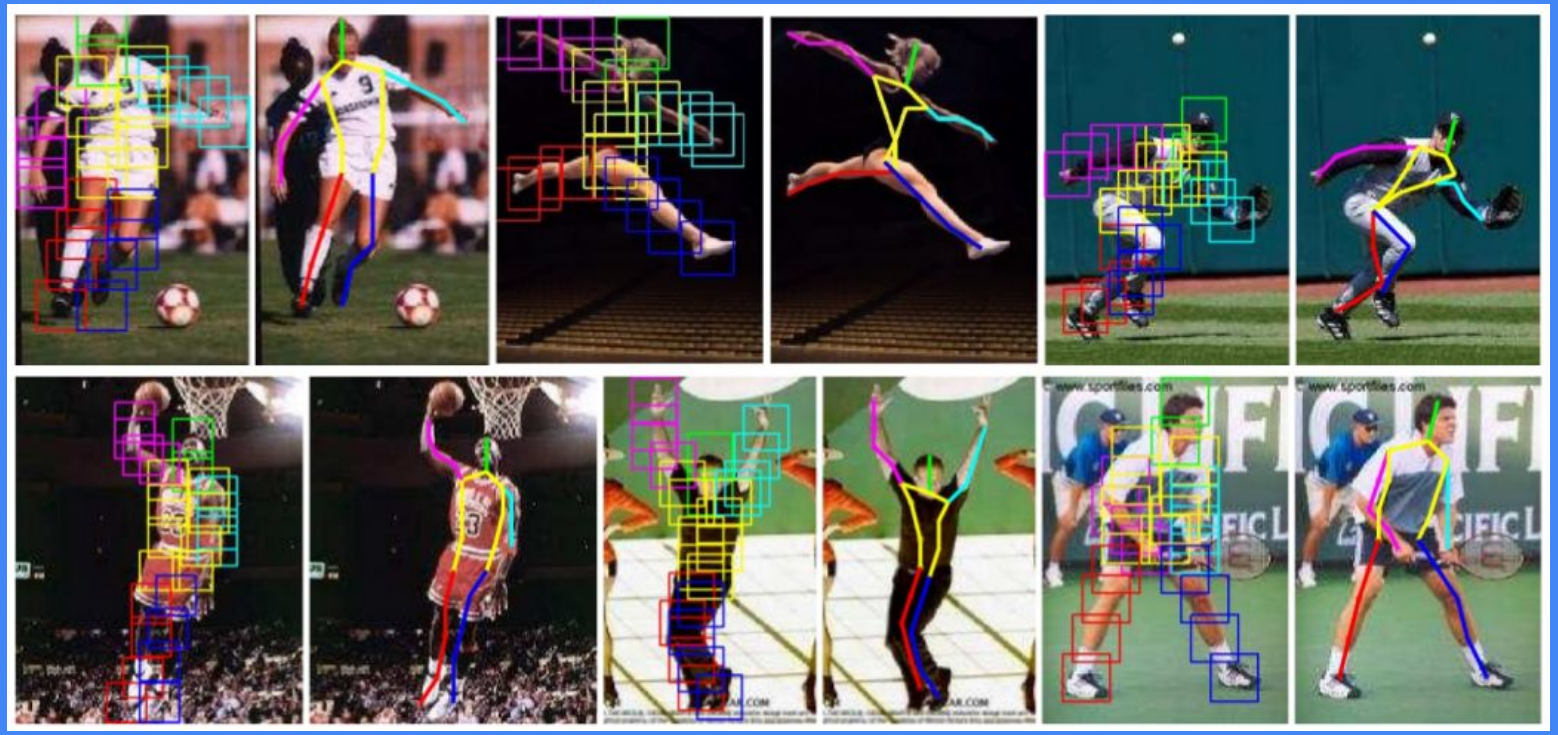
Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. TPAMI, 2012.

Flexible Mixtures-of-Parts Models (2012)



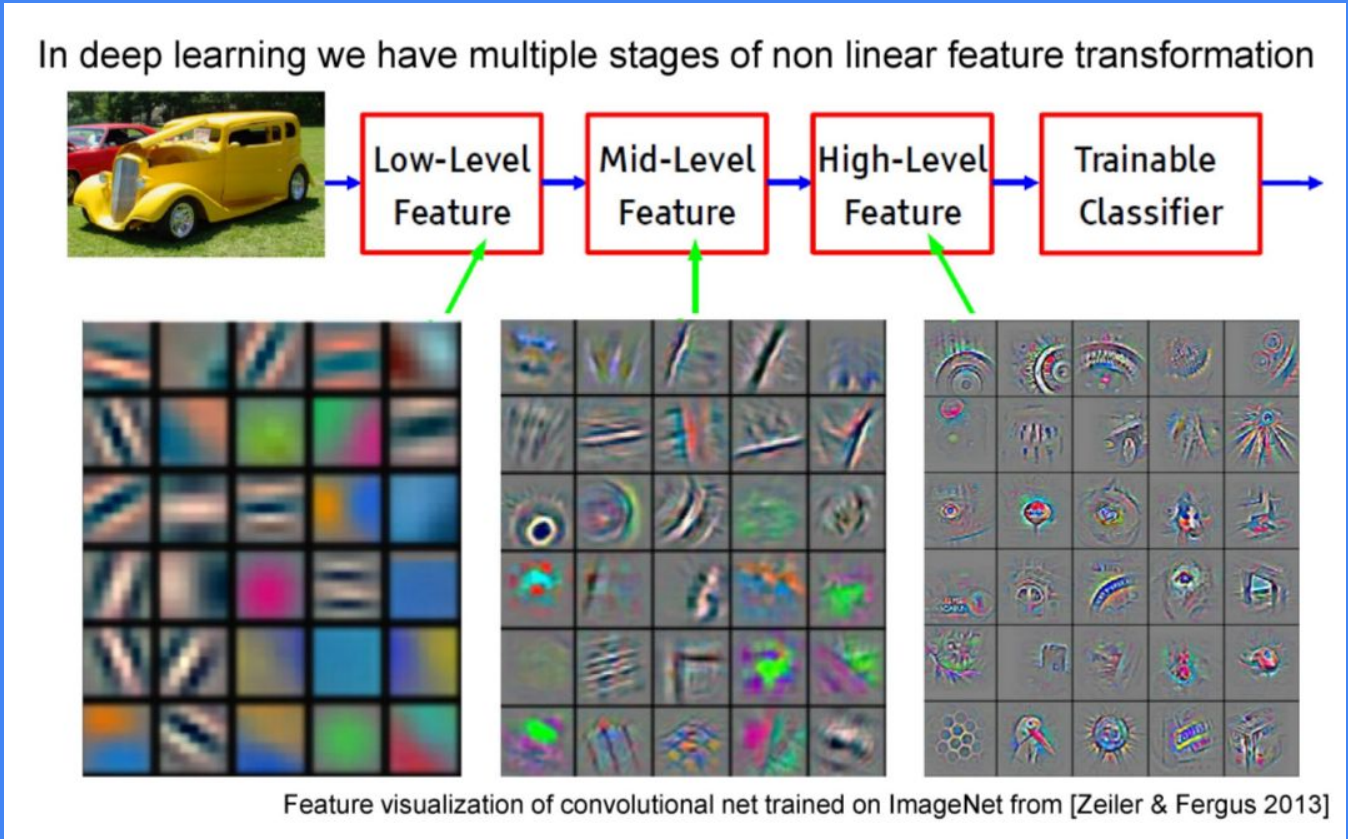
Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. TPAMI, 2012.

Flexible Mixtures-of-Parts Models (2012)



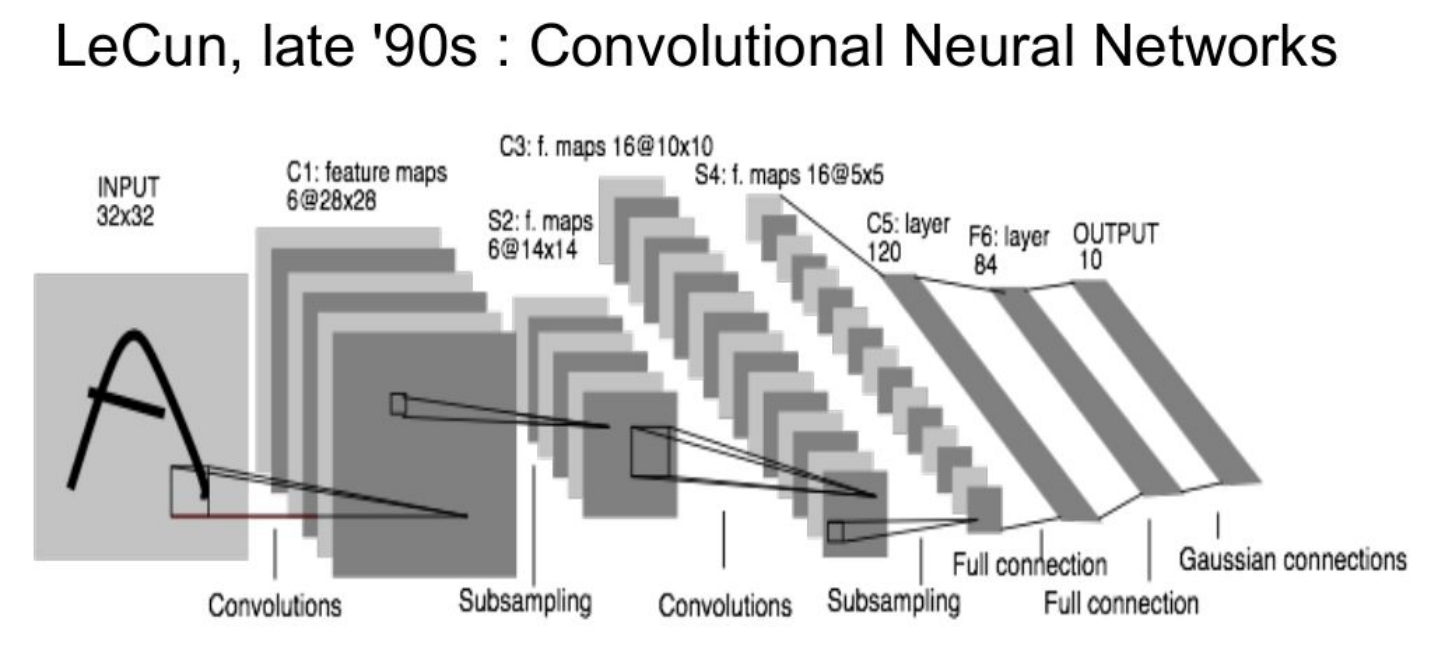
Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. TPAMI, 2012.

Deep Learning: Towards learning hierarchical representations within a single algorithm



Slide credit: Yann LeCun

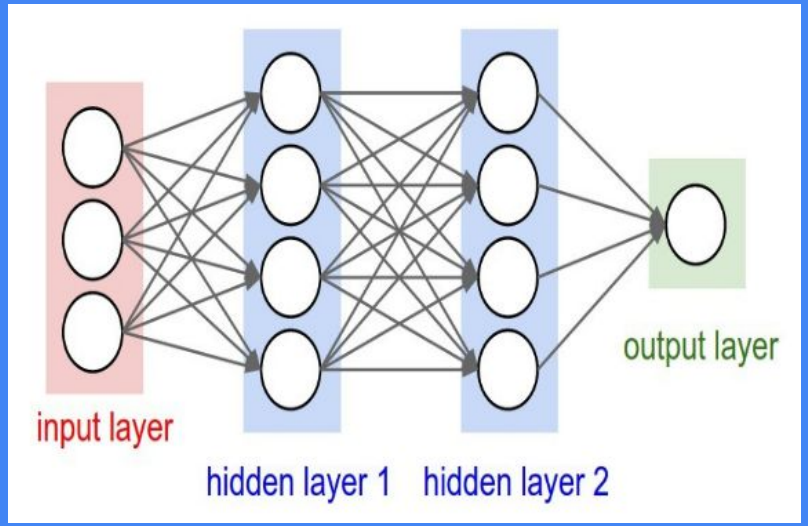
Convolutional Neural Networks

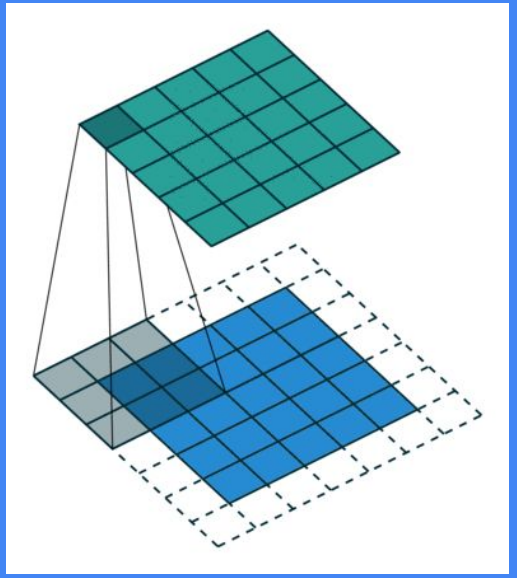


- 2012: a resurgence of interest in deep learning, following ImageNet classification results by Krizhevsky, Sutskever and Hinton

Slide credit: Yann LeCun

Convolutional Neural Networks

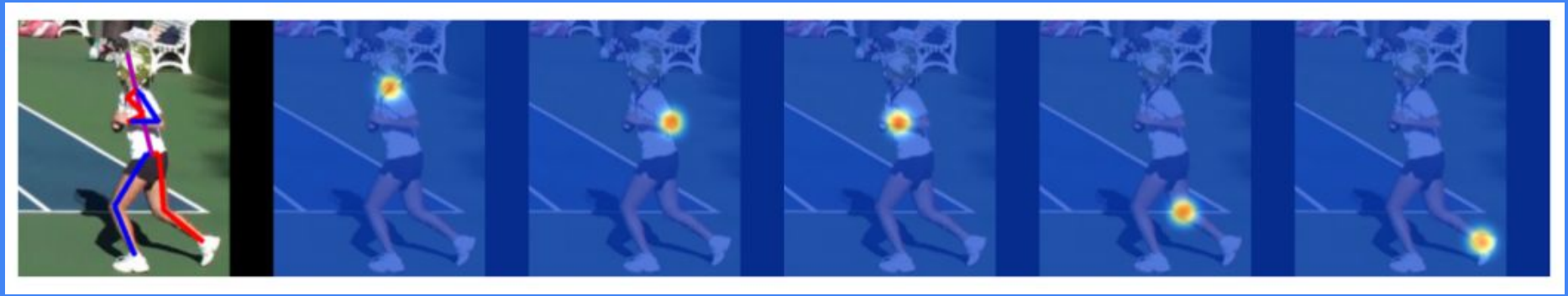




- Traditional neural networks are composed of ***fully-connected layers***, in which neurons are connected to every neuron in the previous layer. But this is impractical for use in images.
- Instead, we utilize a ***convolution*** operation, tiling regions with the same set of shared weights. The features learned are effectively filters on the input regions.

Convolutional Pose Machines (CPMs), 2016

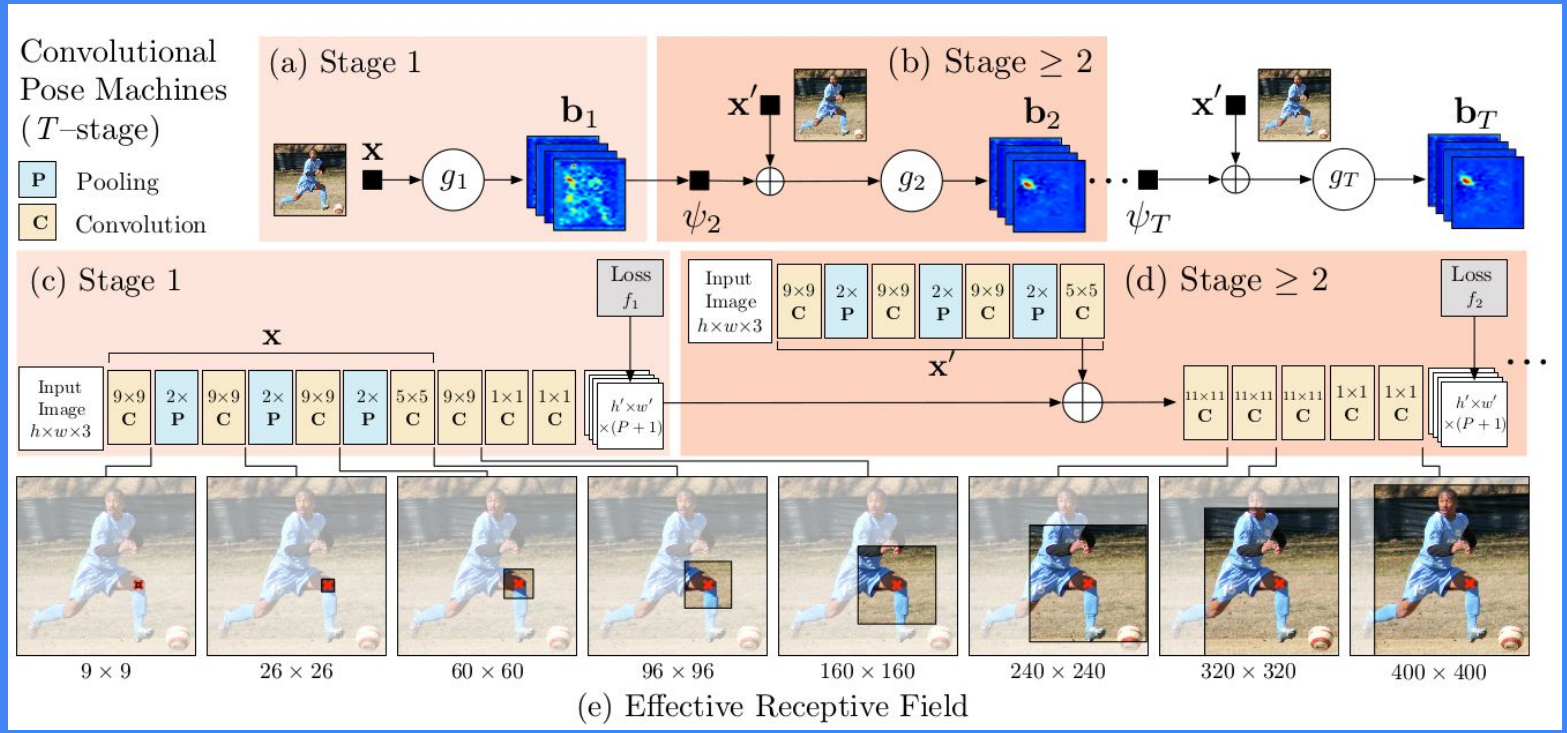
- How can we utilize convolutional neural networks for the task of human pose estimation?
  - Answer: output a heatmap comprised of a Gaussian at every joint location on the target person- allowing for some uncertainty.



Heatmap Image: Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. ECCV, 2016.
CPMs: Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (2016)

Convolutional Pose Machines (CPMs), 2016

The idea: Generate the initial heatmaps using local information, then refine with additional stages.



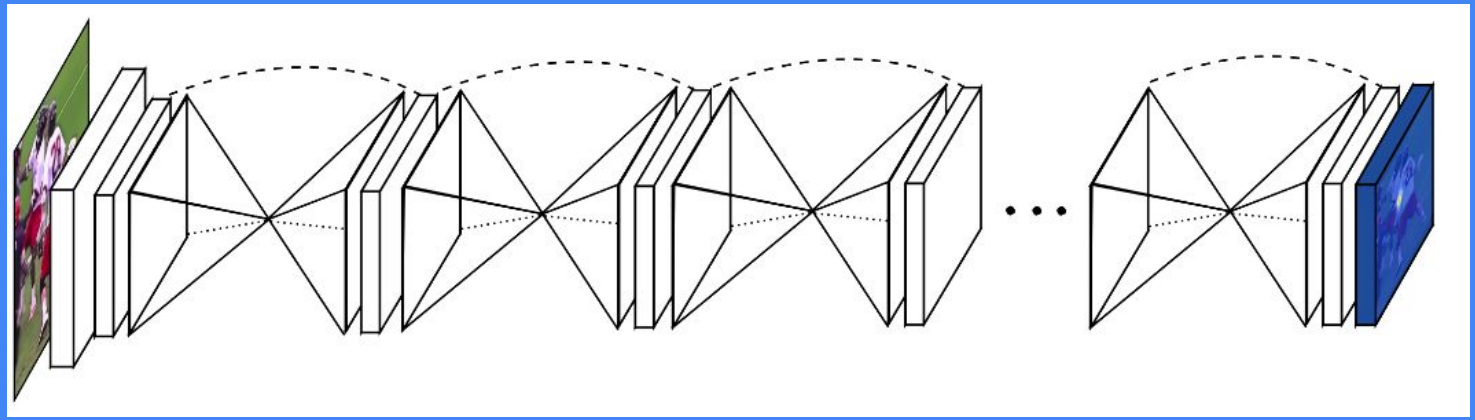
Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on (2016)

Convolutional Pose Machines (CPMs), 2016

- Information is captured and combined at both local and global scales.
- Rather than utilizing hand-crafted descriptors, we use convolutional layers to learn filter-like features (even if we gain little intuition from these features).
- Detection and pose estimation are handled simultaneously.
- Prior works improved performance by giving the algorithm more flexibility. Here, performance is improved by allowing some uncertainty in precise joint locations (Gaussian heatmaps).

Stacked Hourglass Networks (2 months later)

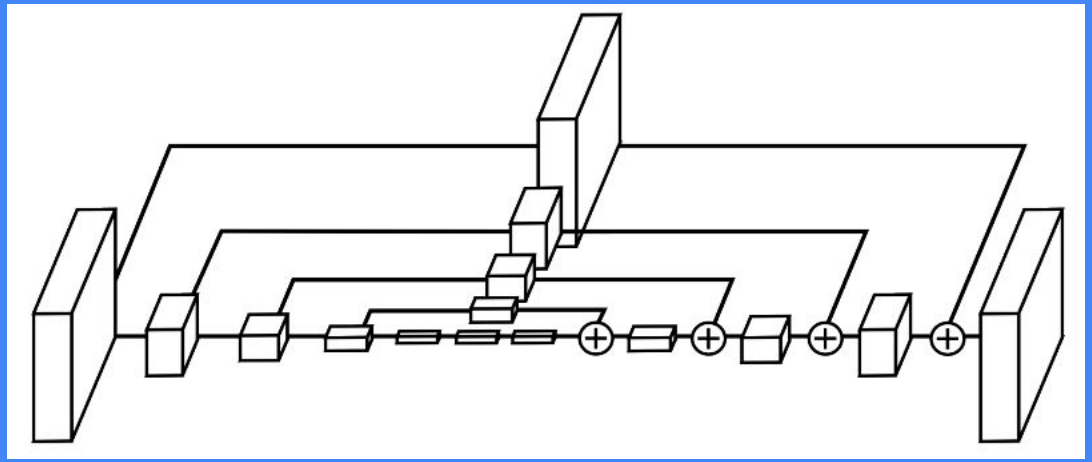
- The general structure is roughly the same as in Convolutional Pose Machines, but with more complex modules within each stage.
- Hourglass structure: consists of symmetric convolutional layers and upsampling, and residual connections across features at same spatial resolution.



Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. ECCV, 2016.

Stacked Hourglass Networks

- A single hourglass module:

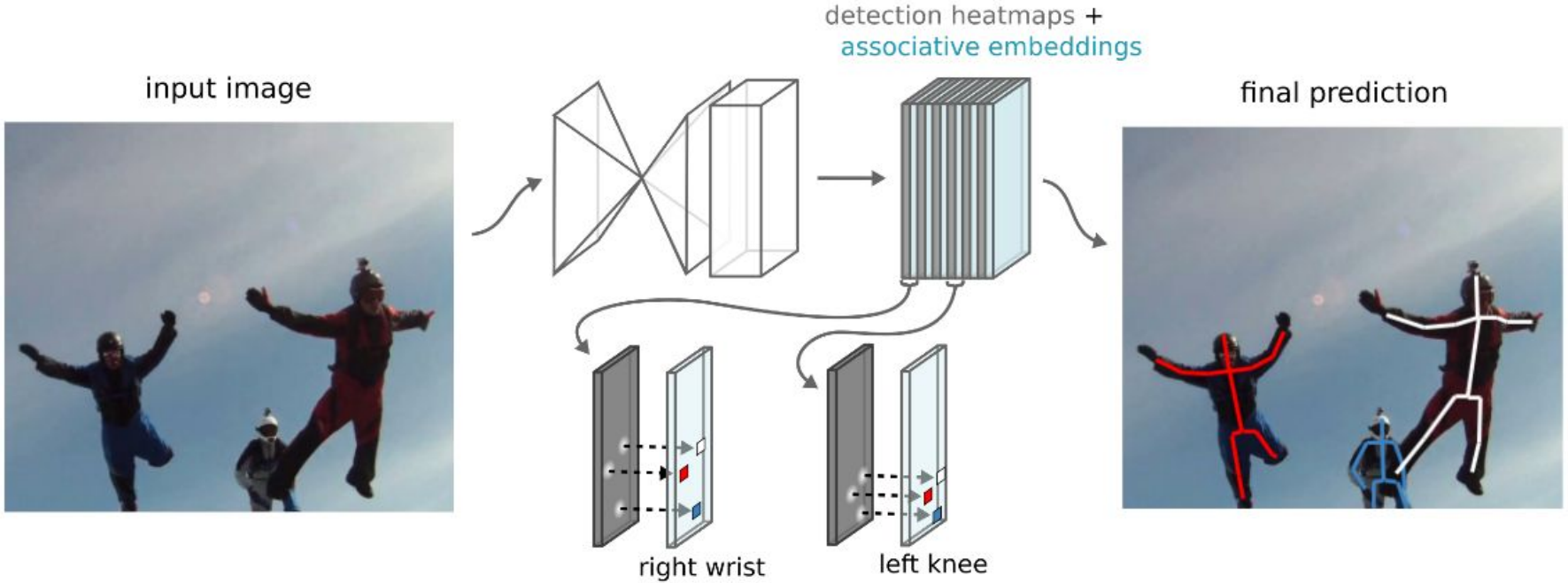


- The current state-of-the-art for an end-to-end trained network for the task of 2D human pose estimation.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. ECCV, 2016.

Extensions of Stacked Hourglass Networks (2017)

- Multi-person pose estimation by associative embeddings



Newell and Deng. “Pixels to Graphs by Associative Embedding”, Jun 2017. https://arxiv.org/abs/1706.07365

Takeaways:

- Deep nets are not just “black boxes”: the development of new architectures has been strongly guided by prior work.

- We have nearly-optimal architectures for collecting information across multiple spatial scales.
  What structures will be optimal when we include ***temporal*** information?