Human Pose Estimation

Xiantong Xiang

Shandong University

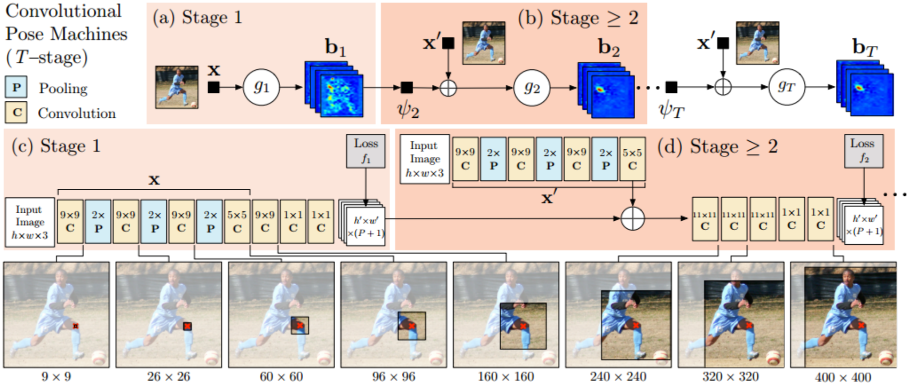
Wednesday, January 22rd, 2025

Contents

Background

- Top-down
  - CPM
  - Hourglass
  - CPN
  - ViTPose/ViTPose++
- Bottom-up
  - OpenPose
- One-Stage
  - RTMO(top-down)
- Multi-Stage
  - CPM
  - Hourglass

CPM



(e) Effective Receptive Field

Convolutional Pose Machines, Wei etc, CVPR 2016 DOI: 10.48550/arXiv.1602.00134

CPM

- A pose machine consists of a sequence of multi-class predictors $g_t(\cdot)$

$$g_1(x_z) \rightarrow \{b_{p1}(Y_p = z)\}_{p \in \{0,\dots,P\}}$$

where $b_{p1}(Y_p = z)$ is the score predicted by the classifier g_1 for assigning the p th part in the first stage at image location z.

- represent all the beliefs of part p evaluated at every location $z = (u, v)^T$ in the image as $b_{p1} \in \mathbb{R}^{w \times h}$

$$b_{pt}[u, v] = b_{pt}(Y_p = z)$$

-

$$g_t(x'_z, \psi_t(z, b_{t-1})) \rightarrow \{b_{pt}(Y_p = z)\}_{p \in \{0,\dots,P+1\}}$$

where $\psi_{t>1}(\cdot)$ is a mapping from the beliefs b_{t-1} to context features

CPM

- The evidence is local because the receptive field of the first stage of the network is constrained to a small patch around the output pixel location.
- composed of five convolutional layers followed by two 1×1 convolutional layers
- **Large receptive fields**:
  - pooling at the expense of precision
  - increasing the kernel size of the convolutional filters at the expense of increasing the number of parameters
  - increasing the number of convolutional layers at the risk of encountering vanishing gradients during training

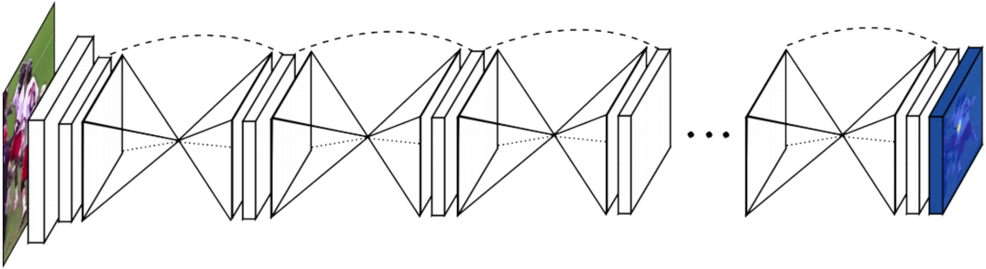
CPM

minimizes the l_2 distance between the predicted and ideal belief maps for each part

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in Z} \| b_p^t(z) - b_p^*(z) \|_2^2$$

$$F = \sum_{t=1}^{T} f_t$$

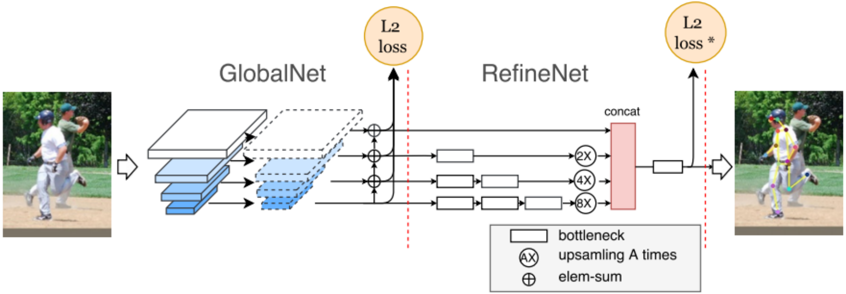
Hourglass



- The Hourglass is a multi-stage architecture, consisting of several stacked Hourglass modules (as the network resembles multiple stacked hourglasses).
- Each Hourglass module includes both a bottom-up process and a top-down process.

Stacked Hourglass Networks for Human Pose Estimation, Newell etc, ECCV 2016

CPN



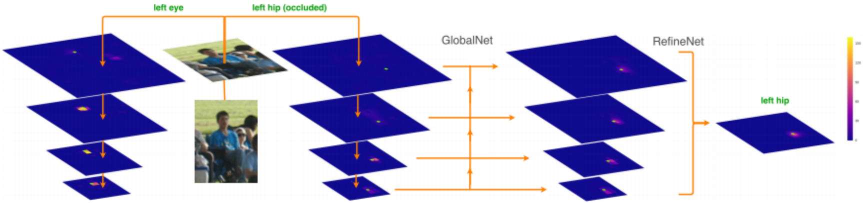
- Two stages:
  - GlobalNet learns a good feature representation based on feature pyramid network
  - RefineNet explicitly address the "hard" joints based on an online hard keypoints mining loss

Cascaded Pyramid Network for Multi-Person Pose Estimation, Chen etc, CVPR 2018

CPN - GlobalNet

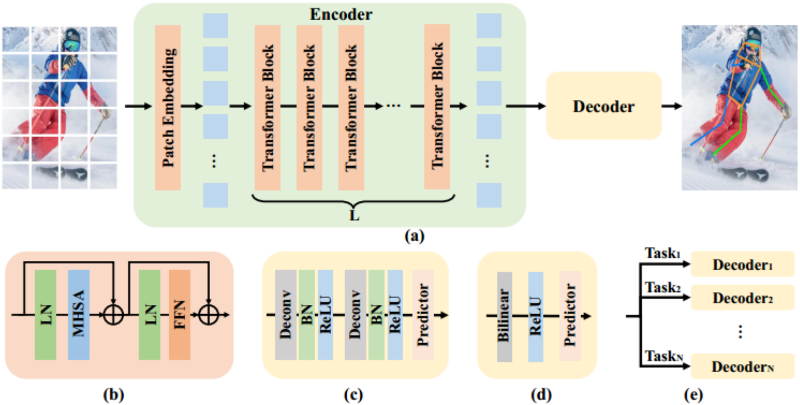
- GlobalNet based on the **ResNet** backbone
- The advantages and disadvantages of feature representation:
  - **the shallow features** have the high spatial resolution for localization but low semantic information for recognition
  - **deep feature** layers have more semantic information but low spatial resolution
- an **U-shape structure** similar to the Feature Pyramid Network (FPN)
- GlobalNet fuses feature maps from different layers through **upsampling and downsampling**

CPN - RefineNet



- Deep features (low resolution, strong semantics) are upsampled to a higher resolution
- These refined features are then fused with the corresponding shallow features, combining both semantic and detailed information.

ViTPose



Neurips 2022

ViTPose

- Simplicity
  - Patch Embedding
  - Transformer Encoder
  - Decoder
- Scalability
- Flexibility
  - transferability
  - Pre-training data flexibility
  - Attention type flexibility
  - Finetuning flexibility
  - Task flexibility
- Transferability

OpenPose

(a) Input Image (b) Part Confidence Maps (c) Part Affinity Fields (d) Bipartite Matching (e) Parsing Results

- 2D confidence maps S

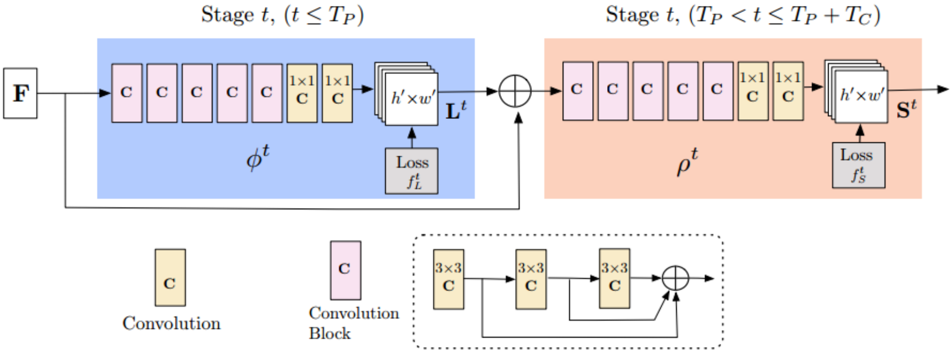
$$S_j \in \mathbb{R}^{w \times h}, \quad j \in 1, \ldots, J$$

- 2D vector fields L of part affinity fields (PAFs)

$$L_c \in \mathbb{R}^{w \times h \times 2}, \quad c \in 1, \ldots, C$$

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Cao etc, CVPR 2017

OpenPose


Stage t, $(t \leq T_P)$
Stage t, $(T_P < t \leq T_P + T_C)$
F
C
1×1 C
$h' \times w'$
$\mathbf{L}^t$
ϕ^t
Loss f_L^t
ρ^t
Loss f_S^t
$\mathbf{S}^t$
Convolution
Convolution Block
3×3 C

OpenPose

- concatenate the predictions from the previous stage and the original image features F

$$L^t = \phi_t(F, L^{t-1}), \forall 2 \leq t \leq T_P,$$

- the process is repeated for the confidence maps detection

$$S^{T_P} = \rho_t(F, L^{T_P}), \forall t = T_P$$

$$S^t = \rho_t(F, L^{T_P}, S^{t-1}), \forall T_P < t \leq T_P + T_C$$

- weight the loss functions spatially

$$f_L^{t_i} = \sum_p \sum_{c=1}^{C} W(p) \cdot \|L_c^{t_i}(p) - L_c^*(p)\|_2^2$$

$$f_S^{t_k} = \sum_p \sum_{j=1}^{J} W(p) \cdot \|S_j^{t_k}(p) - S_j^*(p)\|_2^2$$

OpenPose - the groundtruth part confidence map

- the groundtruth part confidence map

$$S^*_{j,k}(p) = \exp\left(-\frac{|p - x_{j,k}|_2^2}{\sigma^2}\right)$$

- the individual confidence maps via a max operator

$$S^*_j(p) = \max_k S^*_{j,k}(p)$$

OpenPose - the groundtruth PAF

- the unit vector in the direction of the limb

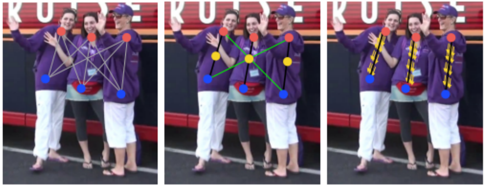
$$v = \frac{x_{j2,k} - x_{j1,k}}{\|x_{j2,k} - x_{j1,k}\|_2}$$

- p on limb

$$0 \leq v \cdot (p - x_{j1,k}) \leq l_{c,k}$$

$$|v^{\perp} \cdot (p - x_{j1,k})| \leq \sigma_l$$

- the limb width σ_l
the limb length $l_{c,k} = \|x_{j2,k} - x_{j1,k}\|_2$



(a) (b) (c)

OpenPose - the groundtruth PAF

- the confidence in their association

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j2} - d_{j1}}{\|d_{j2} - d_{j1}\|_2} du$$

- $p(u)$ interpolates the position of the two body parts

$$p(u) = (1-u)d_{j1} + ud_{j2}$$

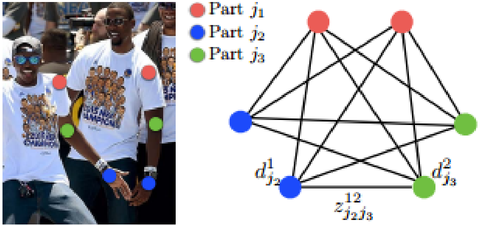
OpenPose

- a maximum weight bipartite graph matching problem

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m \in D_{j1}} \sum_{n \in D_{j2}} E_{mn} \cdot z_{j1j2}^{mn}$$

$$s.t. \quad \forall m \in D_{j1}, \quad \sum_{n \in D_{j2}} z_{j1j2}^{mn} \leq 1$$

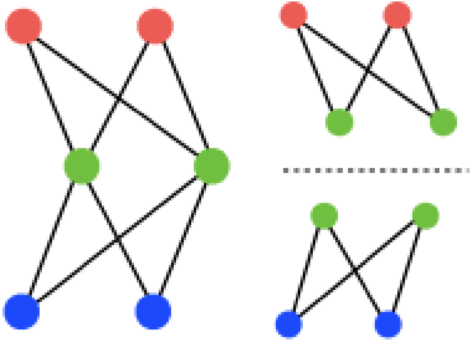
$$\forall n \in D_{j2}, \quad \sum_{m \in D_{j1}} z_{j1j2}^{mn} \leq 1$$



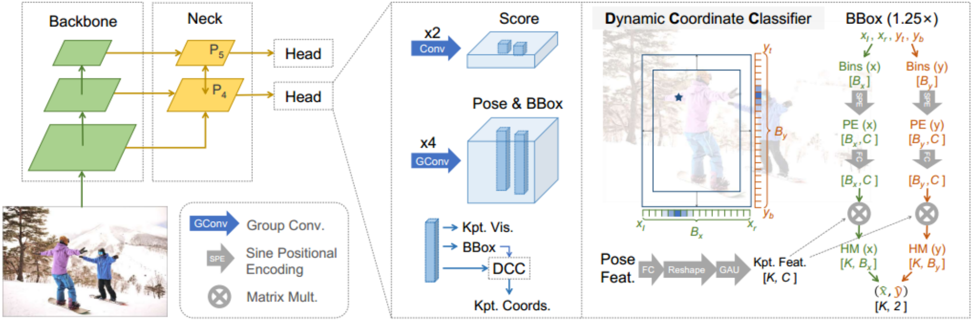
- the full body pose of multiple people $\rightarrow$ a K-dimensional matching problem(NP-Hard)

OpenPose

- Two relaxations to the optimization
  - Minimum Spanning Tree Relaxation
  - Local Matching Relaxation



RTMO



RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation, Lu etc, CVPR 2024

RTMO - Dynamic Coordinate Classifier

- DCC addresses these limitations by dynamically assigning bins to align with each instance' s bounding box, ensuring localized coverage.

$$x_i = x_l + \frac{(x_r - x_l) \cdot (i-1)}{B_x - 1}$$

- DCC generates tailored representations on-the-fly

$$[\mathrm{PE}(x_i)]_c = \begin{cases} \sin\left(\frac{x_i}{t^{c/C}}\right), & \text{for even c} \\ \cos\left(\frac{x_i}{t^{(c-1)/C}}\right), & \text{for odd c} \end{cases}$$

- The probability heatmap is generated by multiplying the keypoint features with the positional encodings of each bin

$$\hat{p}_k(x_i) = \frac{e^{f_k \cdot \phi(\mathrm{PE}(x_i))}}{\sum_{j=1}^{B_x} e^{f_k \cdot \phi(\mathrm{PE}(x_j))}}$$

RTMO - MLE for Coordinate Classification

- Gaussian label smoothing

$$p_k(x_i \mid \mu_x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu_x)^2}{2\sigma^2}} \sim \mathcal{N}(x_i; \mu_x, \sigma^2)$$

- the Gaussian distribution is symmetric with respect to its mean

$$P(\mu_x) = \sum_{i=1}^{B_x} P(\mu_x \mid x_i) P(x_i) = \sum_{i=1}^{B_x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu_x)^2}{2\sigma^2}} \hat{p}_k(x_i)$$

- a negative log-likelihood loss

$$L_{mle}^{(x)} = -\log \left[\sum_{i=1}^{B_x} \frac{1}{\sigma} e^{-\frac{|x_i-\mu_x|}{2\sigma s}} \hat{p}_k(x_i) \right]$$

Thanks

Happy Xiaonian !