Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation

Y.Fan & M.Andriushchenko

Saarland Univ.

30 January 2018

What are we doing?

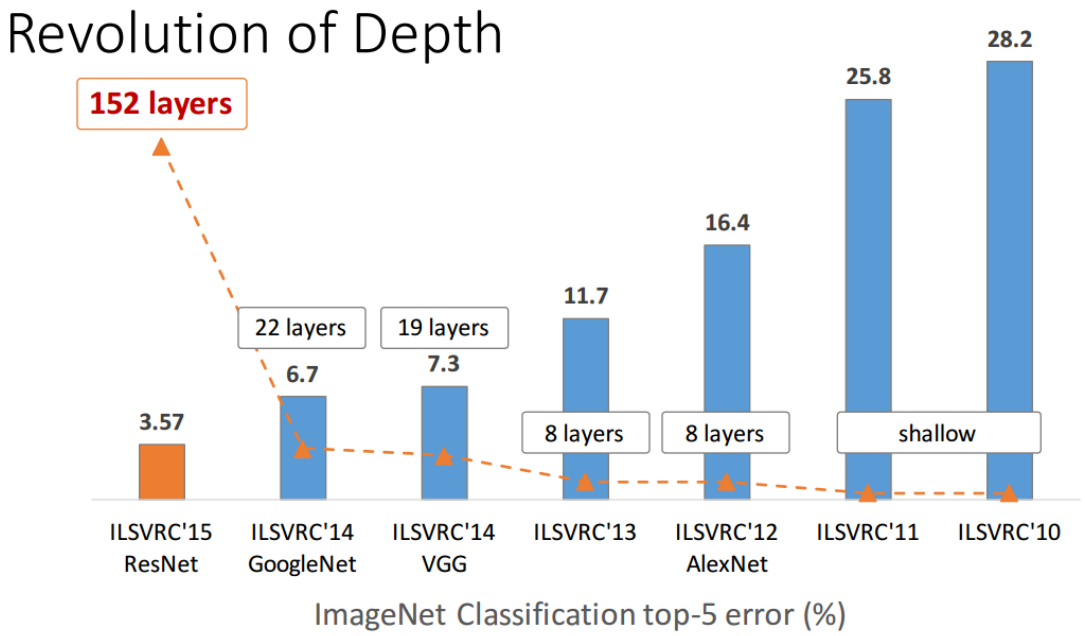
Task: **single person pose estimation** from 2D images

What are we doing?

Task: **single person pose estimation** from 2D images
Method: combination of a CNN and a PGM trained end-to-end jointly

Modern computer vision

CNNs is the method of choice for many computer vision tasks.

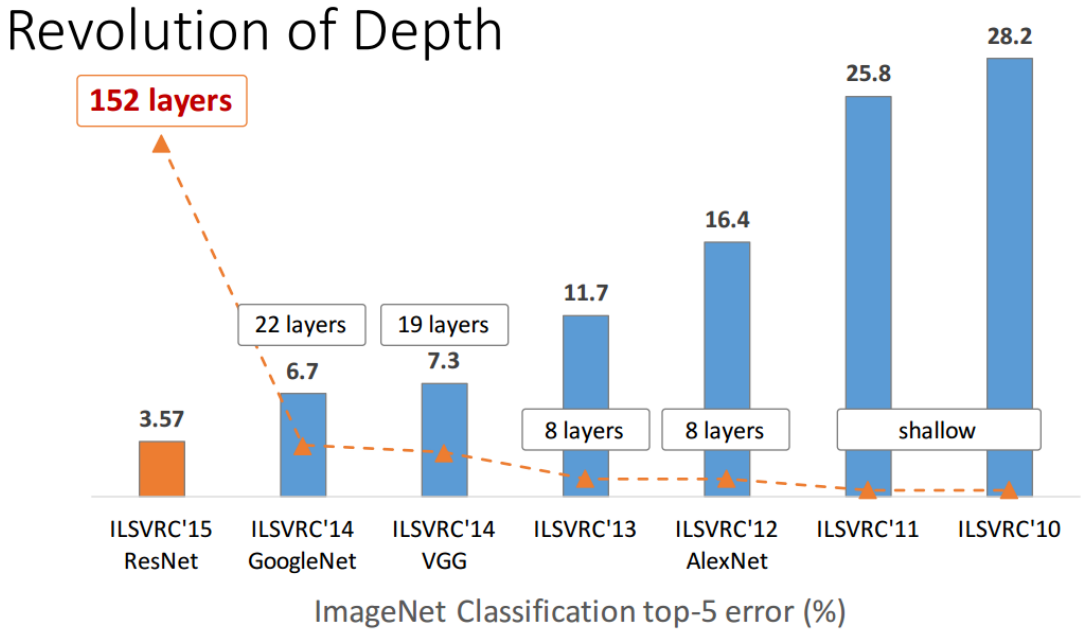


Source: Kaiming He, ICML 2016

`icml.cc/2016/tutorials/icml2016_tutorial_deep_residual_networks_kaiminghe.pdf`

Modern computer vision

CNNs is the method of choice for many computer vision tasks.
We cannot ignore advances in the recent CNN development!



Source: Kaiming He, ICML 2016
`icml.cc/2016/tutorials/icml2016_tutorial_deep_residual_networks_kaiminghe.pdf`

Modern computer vision

Wait a moment, but what about **PGMs**?

Modern computer vision

Wait a moment, but what about **PGMs**?

Do we actually need them?

Modern computer vision

Wait a moment, but what about **PGMs**?

Do we actually need them?

Can **CNNs** learn directly from data all needed relations between parts?

CNN part detector

Generally the results with a CNN part detector are quite **good**
(we trained a CNN *from scratch* on FLIC: 4000 images from movies)

Source: YF and MA

CNN part detector

Generally the results with a CNN part detector are quite **good**

Source: YF and MA

CNN part detector

But there are also quite **many false positives**!
(pink corresponds to a hip detection)

Source: YF and MA.

CNN part detector

But there are also quite **many false positives**!
Some parts are hard, e.g. **right wrist detection** is problematic.

Source: YF and MA.

Solution?

What should we do with these **false positives**?

Solution?

What should we do with these **false positives**?

Use a PGM to enforce kinematic constraints!

Solution?

What should we do with these **false positives**?

Use a PGM to enforce kinematic constraints!

So we can combine the best from 2 worlds: CNN and PGM.

Solution?

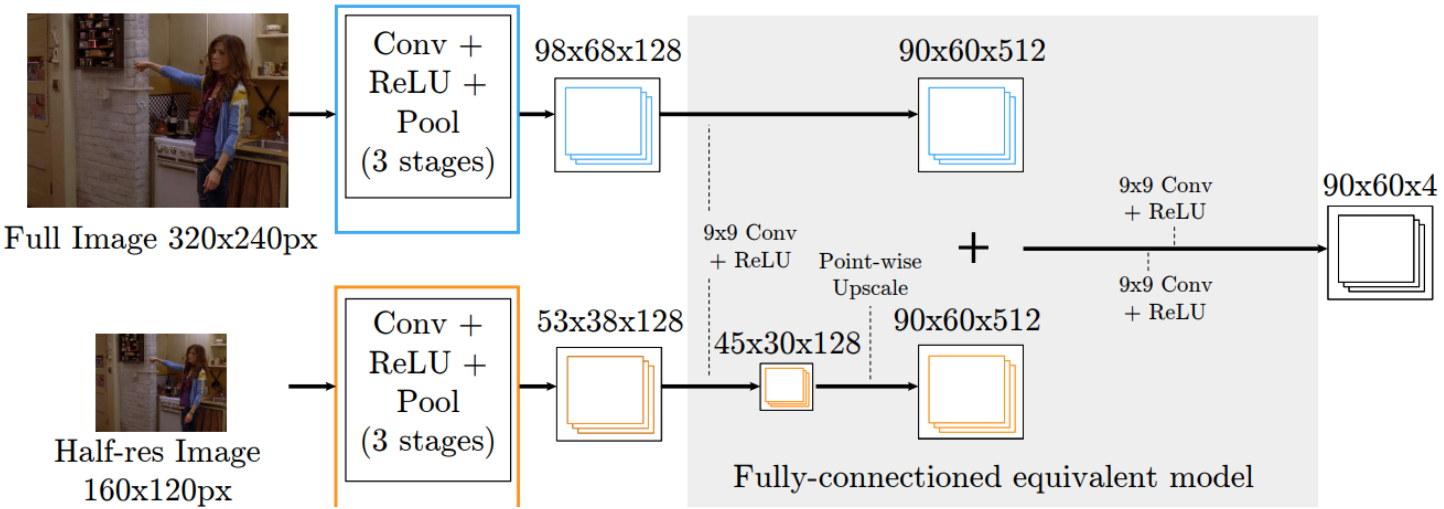
What should we do with these **false positives**?

Use a PGM to enforce kinematic constraints!

So we can combine the best from 2 worlds: CNN and PGM.

Moreover, we can train them jointly!

CNN part detector



Source: "Joint Training of a CNN and a Graphical Model for Human Pose Estimation".

How the **CNN part detector** is implemented? → **unary potentials**

- Coarse and fine resolution → 2 branches of the CNN
- In the end: **softmax** and **MSE loss** to compare with the ground truth
- We added **Batch Normalization**, which speeds up convergence **x10 times**! → makes development of the project much faster
- 6 convolutional layers is so 2014...

Higher-Level Spatial Model

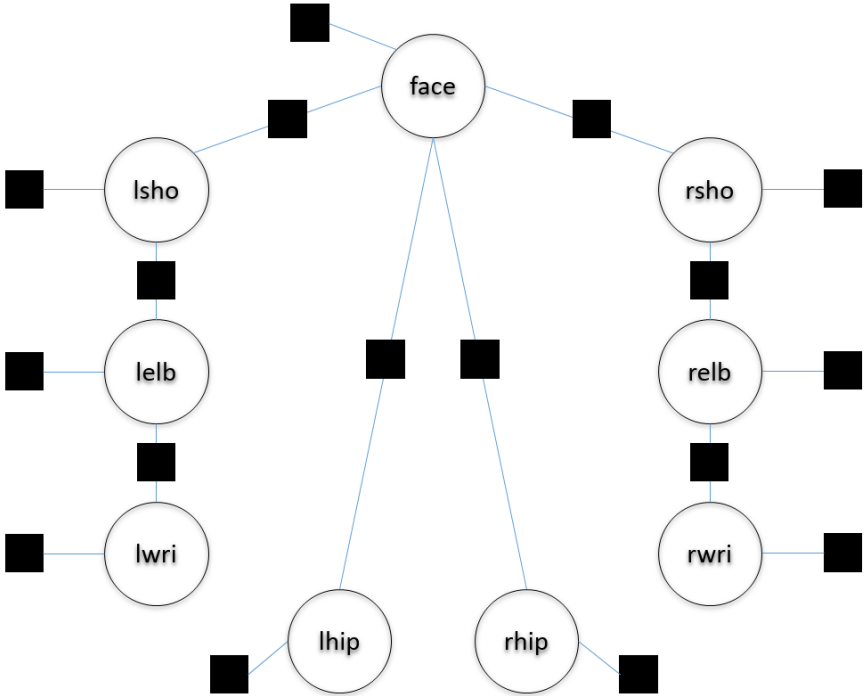
Problem: Part Detector produces many false positives.

Solution: use a Spatial Model to enforce the consistency.

Source: YF and MA.

Spatial Model as a PGM

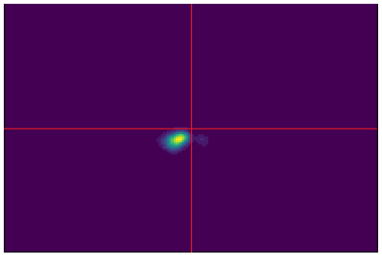
Traditional approach: use the star model!
What do we have for FLIC dataset?



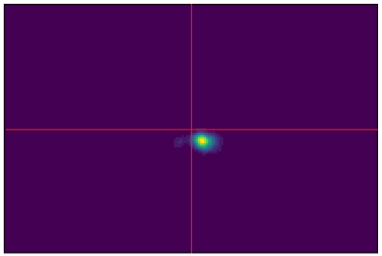
Source: YF and MA.

Pairwise Potentials

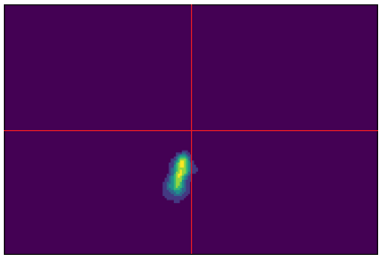
- Simple $\mathcal{N}(\mu, \Sigma)$ doesn't fit all the cases! (especially with diagonal Σ!)
- We can learn this distribution in a non-parametric way (parametrized by 180x120 heat maps of pairwise compatibilities) by **backprop**



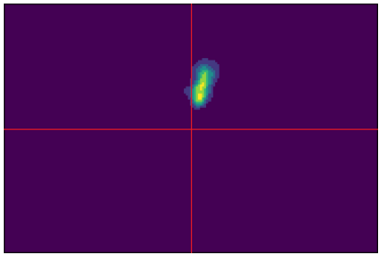
Left shoulder given face



Right shoulder given face



Left hip given face



Face given left hip

We can use empirical histogram of joint displacements as **good init**. Source: YF and MA.

Spatial Model as a trainable PGM

But can we learn a PGM from data?

$\rightarrow$ We should use **fully connected PGM** and train all potentials!

Star PGM:

- Computationally more efficient (during the train phase).
- Less parameters to train.
- Inference is exact.

Fully Connected PGM:

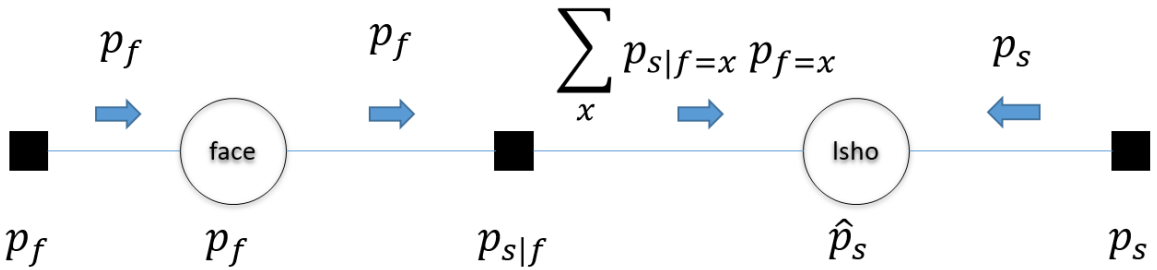
- More model capacity.
- **The model is learned from the data, no need of expert prior.**
- Loopy structure has no guarantee of convergence.

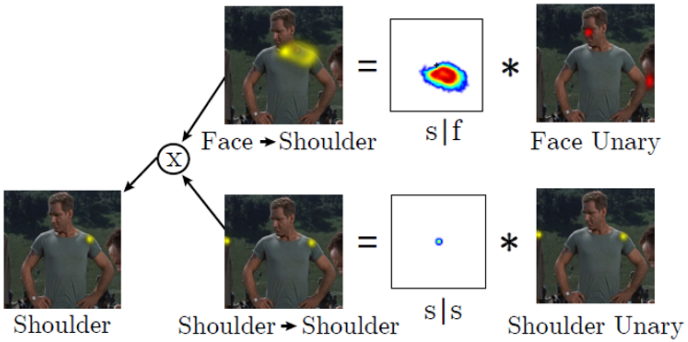
Fully connected PGM trained jointly with CNN is the main novelty!

Inference in the fully connected PGM

We do a single round of sum-product belief propagation to get marginals

Can be seen as approximation for Loopy Belief Propagation in MRF!





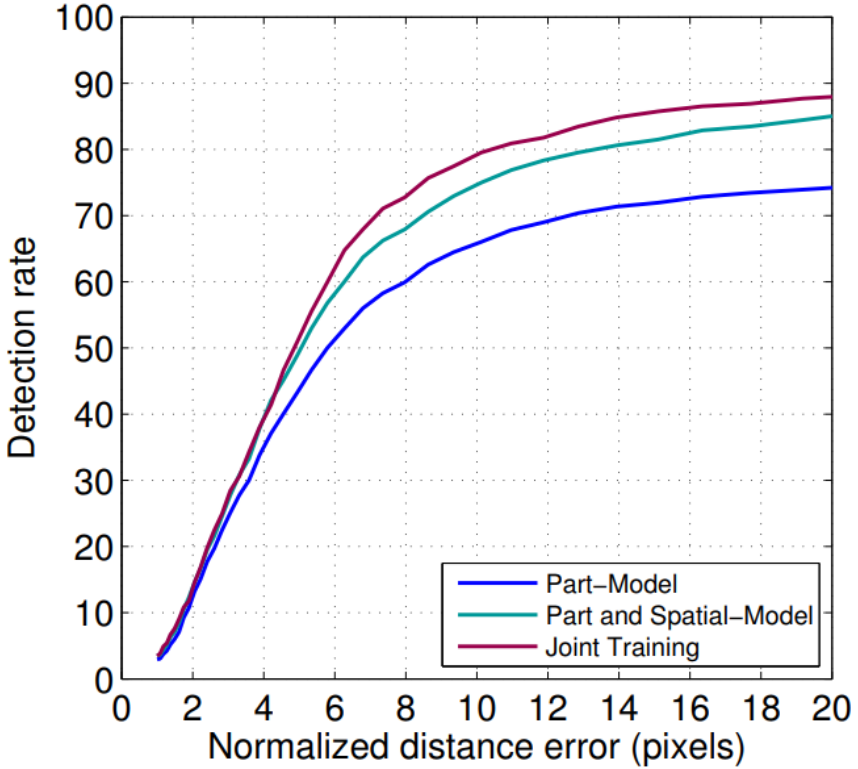
$$\hat{p}_i \propto p_i \prod_{u \in U} (p_{i|u} * p_u)$$

where U is a set of neighbouring nodes of body part i

Source: YF, MA + "Joint Training of a CNN and a Graphical Model for HPE".

Joint training

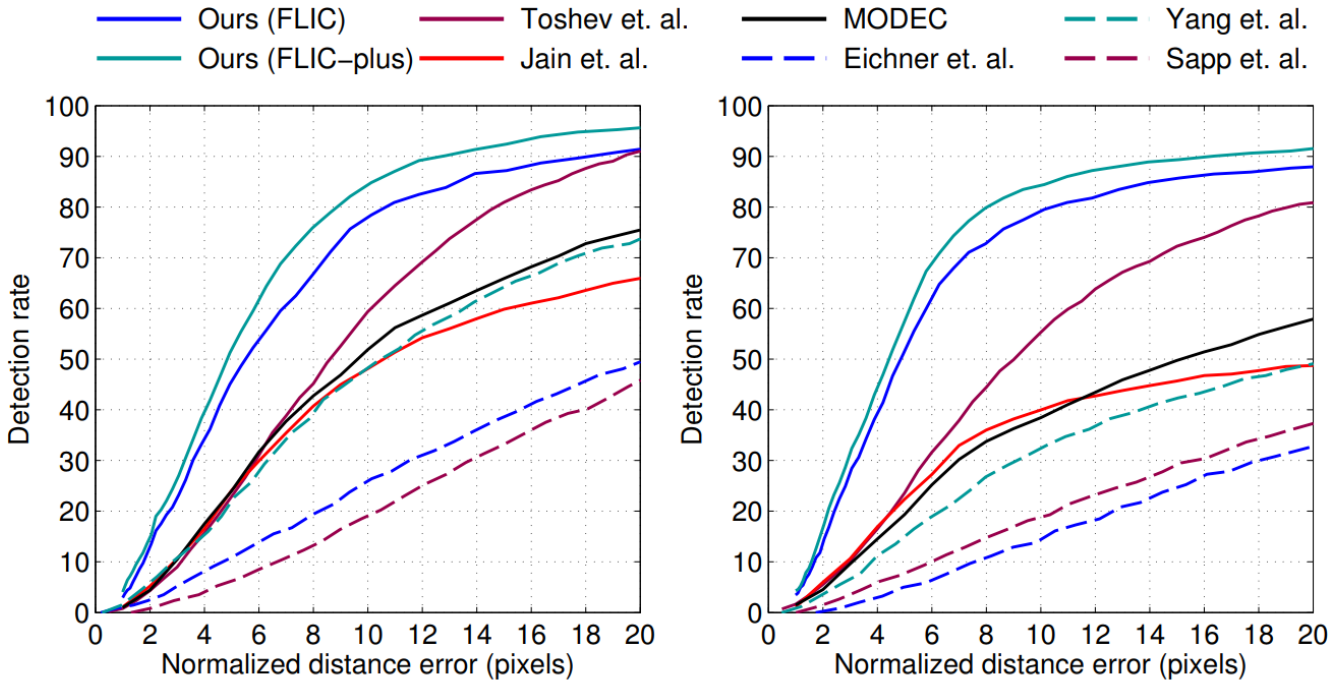
Joint training of the fully connected CNN and the PGM matters!



Source: "Joint Training of a CNN and a Graphical Model for Human Pose Estimation".

State-of-the-art for 2014

The technique described above achieves the best results for 2014!



(a) FLIC: Elbow (b) FLIC: Wrist

Source: "Joint Training of a CNN and a Graphical Model for Human Pose Estimation".

Final details

- We open sourced all our code in our Github repository: `https://github.com/max-andr/cnn_mrf_hybrid_for_hpe`!
- Up to our knowledge, this is the first implementation of the presented paper [1].
- We extensively use TensorBoard! Now small demo.

Thanks for your attention!

Any questions?

[1] Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation

[2] Learning Human Pose Estimation Features with Convolutional Networks