Human Pose Estimation in Static Images

Xianjie Chen

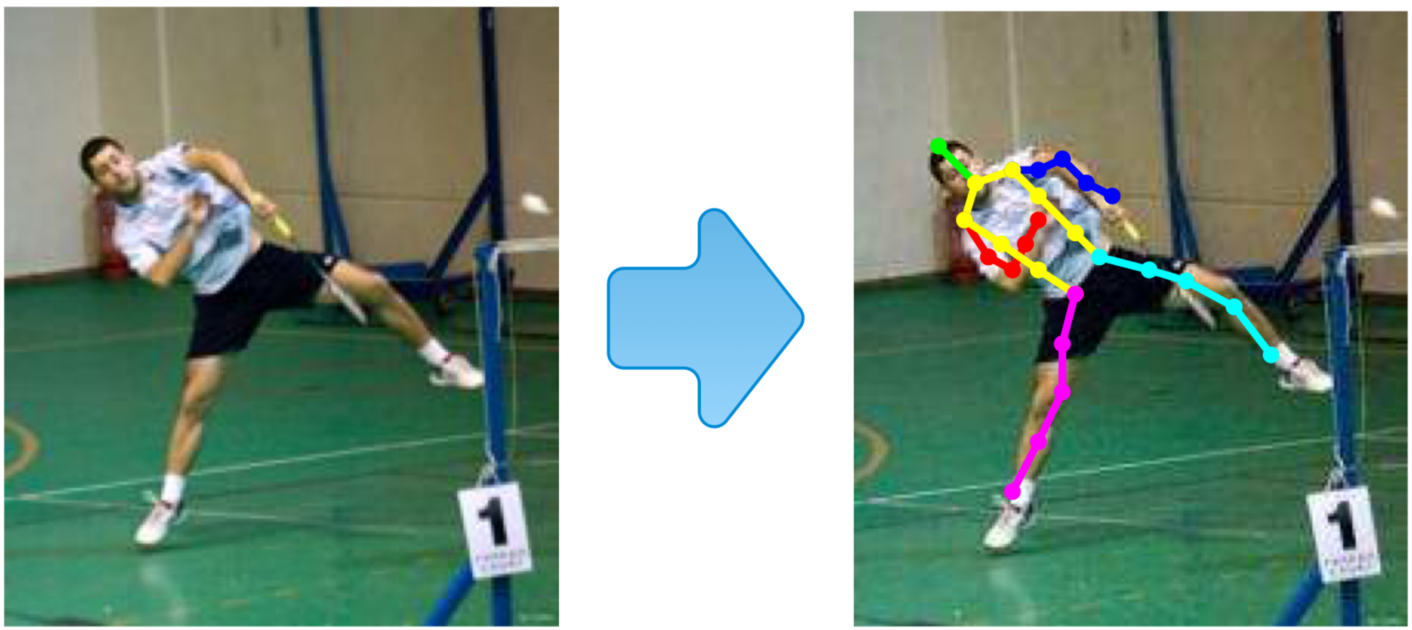
Advisor: Prof. Alan Yuille

Overview

- Introduction
  - Background & Related work
- Our model
  - Key ideas
  - Model formulation
  - Inference and Learning
  - Relationship to other models
- Experiment comparison
  - Benchmark performance
  - Diagnostic experiments
- The follow-up work: Parsing occluded people

Introduction

- Estimate articulated 2D human pose from a single static image.

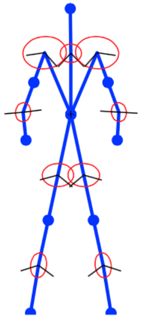


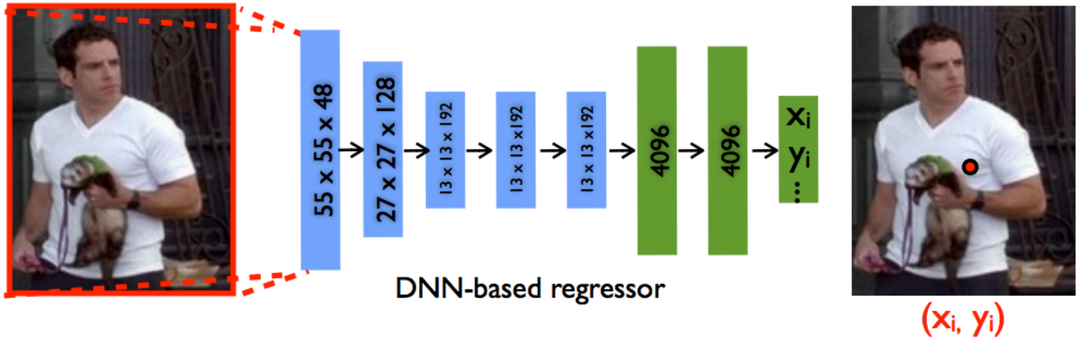
Introduction

- Fundamental task in computer vision.
  - Activity recognition, Image understanding.
- Applications
  - Video surveillance / analysis.
  - Fashion item localization etc.

Related work

- Most work has been based on graphical model
  - Pishchulin et. al., ICCV’13, CVPR’13
  - Yang & Ramanan, TPAMI’13.
- ConvNet based regression
  - DeepPose, CVPR’14.



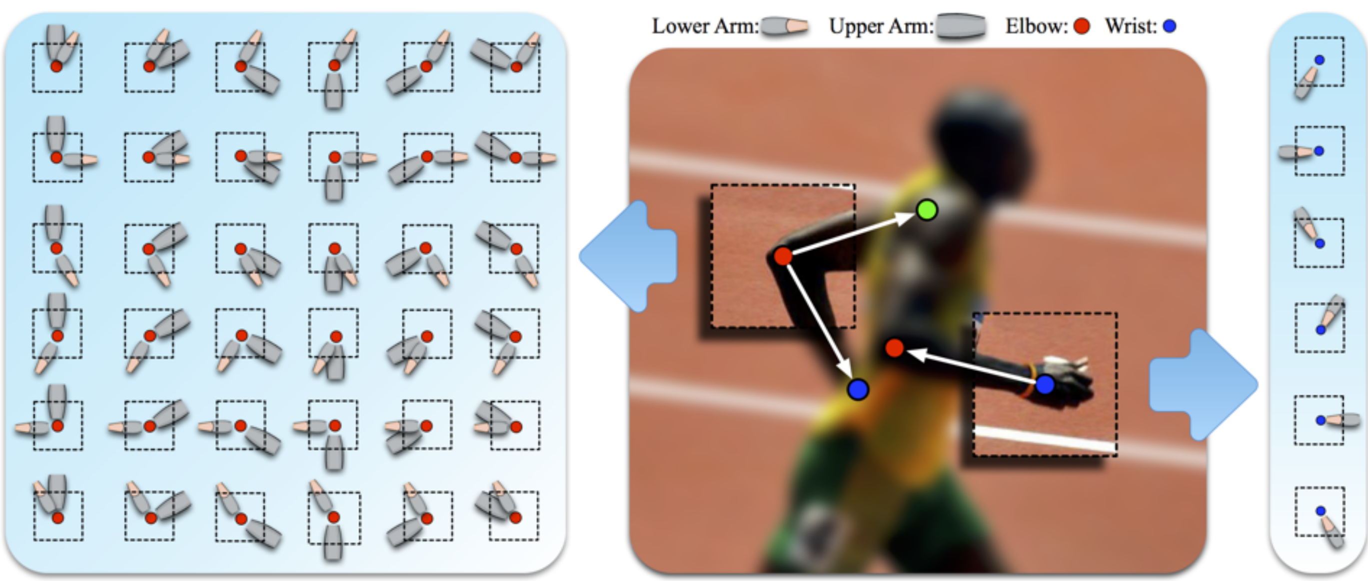


Pros & Cons

- Graphical model
  - Pro: Explicit and flexible representation.
  - Con: Data independent pairwise relations
    - too loose to be helpful
    - too strict to model highly variable poses.
- ConvNet
  - Pro: Large learning capacity, good at extracting image info.
  - Con: Implicit and hard to diagnose.
- Our method
  - Extend graphical model by stronger pairwise relations.
  - Use ConvNet to extract info from local image patches.

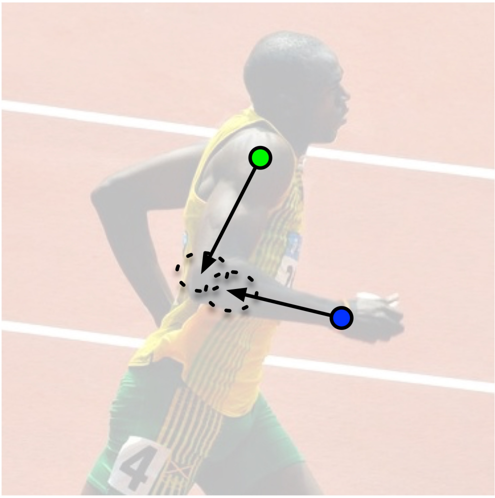
Our method

- Graphical model: Image dependent pairwise relations (IDPRs).
  - Local image measurements can reliably predict the relative positions of all its neighbors (as well as detect the part).

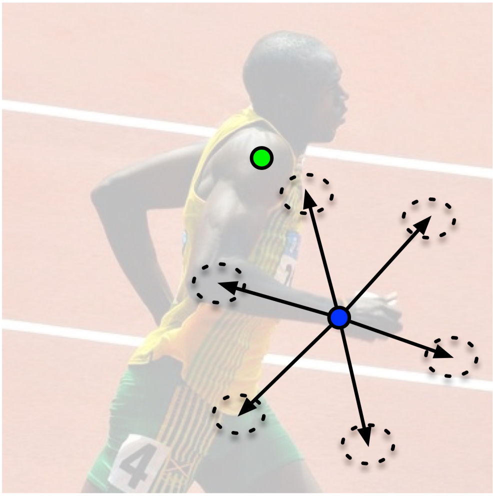


Our method

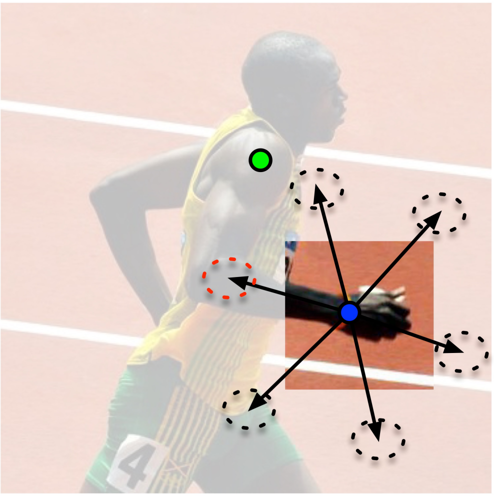
- Stronger pairwise term
  - Local image measurements give input to the pairwise terms (as well as the unary terms).



Too strict



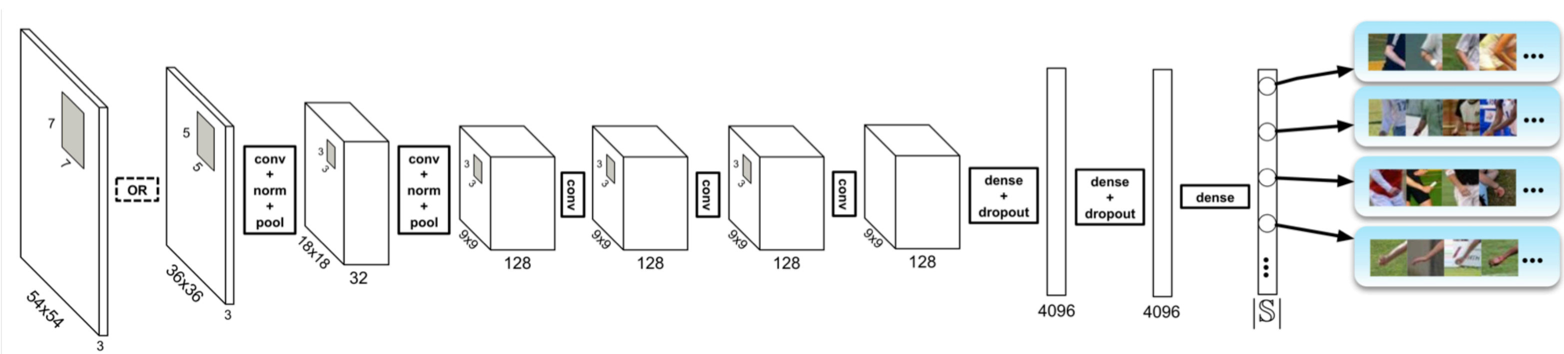
Too loose



Flexible & Helpful

Our method

- Require method to extract info from local image patches.
  - Part presence (appearance terms).
  - Pairwise part relations (IDPR terms).
- ConvNet is suitable.
  - Full supervised training.
  - We design a ConvNet to efficiently extract both info together.



State of the art

- Extend graphical model, and combine it with ConvNet.
  - Significantly outperforms the state of the art methods on benchmarks (LSP, FLIC).
  - Very good cross-dataset generalization (Buffy).
- The code is public online.

Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations.
Xianjie Chen, Alan Yuille
Neural Information Processing Systems (NIPS), 2014.

The Graphical Model

- Tree model: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - The pixel locations $\mathbf{l}_i = (x, y)$ of part $i \in \mathcal{V}$
  - Pairwise relation types $t_{ij} \in \{1, \ldots, T_{ij}\}, \forall (i,j) \in \mathcal{E}$
- Unary terms:

$$U(\mathbf{l}_i|\mathbf{I}) = w_i \phi(i|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta})$$

- Image Dependent Pairwise Relational (IDPR) Terms:

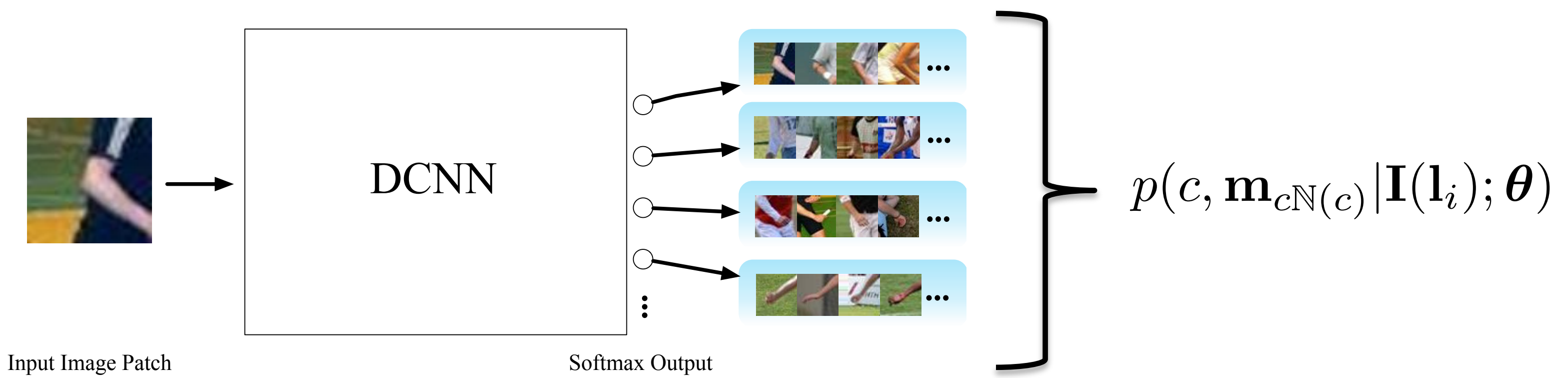
$$\begin{aligned} R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji}|\mathbf{I}) = \quad & \langle \mathbf{w}_{ij}^{t_{ij}}, \boldsymbol{\psi}(\mathbf{l}_j - \mathbf{l}_i - \mathbf{r}_{ij}^{t_{ij}}) \rangle + w_{ij} \varphi(t_{ij}|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) \\ + \quad & \langle \mathbf{w}_{ji}^{t_{ji}}, \boldsymbol{\psi}(\mathbf{l}_i - \mathbf{l}_j - \mathbf{r}_{ji}^{t_{ji}}) \rangle + w_{ji} \varphi(t_{ji}|\mathbf{I}(\mathbf{l}_j); \boldsymbol{\theta}) \end{aligned}$$

$$\boldsymbol{\psi}(\Delta \mathbf{l} = [\Delta x, \Delta y]) = [\Delta x \; \Delta x^2 \; \Delta y \; \Delta y^2]^\intercal$$

- The Full score： $F(\mathbf{l}, \mathbf{t}|\mathbf{I}) = \sum_{i \in \mathcal{V}} U(\mathbf{l}_i|\mathbf{I}) + \sum_{(i,j) \in \mathcal{E}} R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji}|\mathbf{I}) + w_0$

Image dependent terms

- ConvNet for Image dependent terms:
  - Appearance terms $\phi(.|.;\boldsymbol{\theta})$
  - IDPR terms $\varphi(.|.;\boldsymbol{\theta})$



$$\phi(i|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta}) = \log(p(c=i|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta}))$$

$$\varphi(t_{ij}|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta}) = \log(p(m_{ij}=t_{ij}|c=i,\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta}))$$

Inference

- Image Dependent Terms
  - Fully convolutional inference by a single ConvNet.
  - Computations common to overlapping regions are shared.
- Graphical model inference $O(T^2LK)$
  - Dynamic programming -> Linear in # of parts.
  - Distance Transform -> Linear in # of locations.

Learning

- Fully supervised learning
  - Annotated part locations.
  - Derive pairwise type labels by clustering.
- Three sets of parameters
  - Mean relative positions r of different pairwise relation types, by K-means clustering.
  - Parameters θ of image dependent terms, by ConvNet.
  - Weight parameters $\mathbf{w}$, by linear SVM.

Implementation Detail

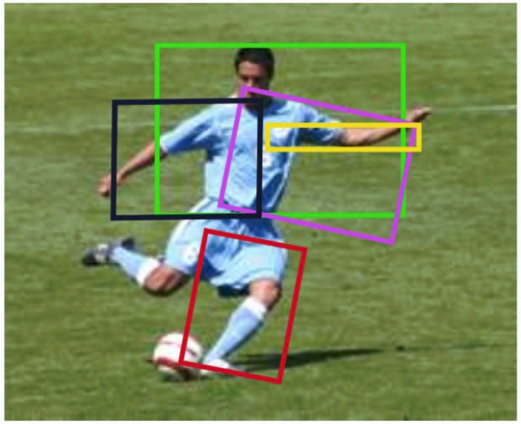
- Data Augmentation for ConvNet training
  - Local part patches (~20 parts) + Random background patches.
  - horizontally flipping + rotating -> ~1 million patches.
- Graphical model structure
  - Predefined tree structure.
  - Part size is roughly the size of the head.
- ConvNet structure is similar to the AlexNet.
  - Input: much smaller image patch (36 x 36 on the LSP)
  - Use the Caffe implmentation.

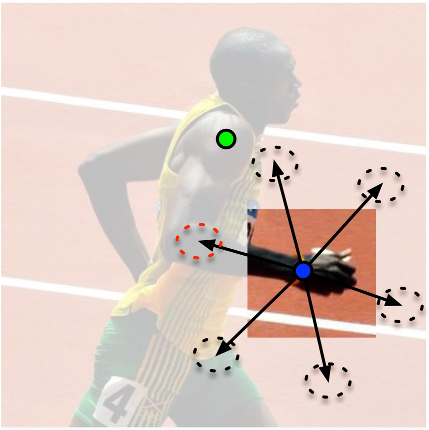
Relationship to other models

- Pictorial Structure (PS)
  - Recover by allowing one pairwise relation type.
- Yang and Ramanan’s Mixtures-of-parts (MOP), TPAMI’13.
  - MOP defines different “types” of part by its relative position with respect to its parent.
  - Recover by only allowing parent to predict child.
- DeepPose, CVPR’14
  - ConvNet based regression.
  - Does not give confidence of the estimation. Assume given bounding box of human.

Relationship to other models

- Pishchulin et. al., Poselet Conditioned Pictorial Structures, CVPR'13
  - Focus on capturing dependencies between non-connected body parts by mid-level representation (poselets).
  - We focus on extracting more info (pairwise relations) from local image measurements.





Benchmark Performance

- LSP
  - Using Observer-Centric annotation.
  - Percentage of Correct Parts (PCP)

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
Ours	**92.7**	**87.8**	**69.2**	**55.4**	**82.9**	**77.0**	**75.0**
Pishchulin et al. [16]	88.7	85.6	61.5	44.9	78.8	73.4	69.2
Ouyang et al. [14]	85.8	83.1	63.3	46.6	76.5	72.2	68.6
DeepPose* [23]	-	-	56	38	77	71	-
Pishchulin et al. [15]	87.5	78.1	54.2	33.9	75.7	68.0	62.9
Eichner&Ferrari [4]	86.2	80.1	56.5	37.4	74.3	69.3	64.3
Yang&Ramanan [26]	84.1	77.1	52.5	35.9	69.5	65.6	60.8

Table 1: Comparison of *strict* PCP results on the LSP dataset. Our method improves on all parts by a significant margin, and outperforms the best previously published result [1] by 5.8% on average. Note that DeepPose uses Person-Centric annotations and is trained with an extra 10,000 images.

Benchmark Performance

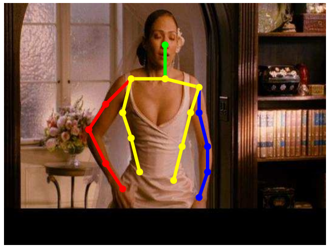
- LSP
  - Using Person-Centric annotation.
  - Percentage of Correct Parts (PCP).
  - Thanks Pishchulin et. al. for comparing different methods.

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	mPCP
Ours	**96.0**	**85.6**	**69.7**	**58.1**	**77.2**	**72.2**	**73.6**
Tompson et al. NIPS'14	90.3	83.7	63.0	51.2	70.4	61.1	66.6
Pishchulin et al., ICCV'13	88.7	85.1	46.0	35.2	63.6	58.4	58.0
Wang& Li, CVPR'13	87.5	79.1	43.1	32.1	56.0	55.8	54.1

Table 1: Comparison of *strict* PCP results on the LSP dataset using Person-Centric annotations.

Benchmark Performance

- FLIC
  - Upper-body human poses.
  - PCP & Percentage of Detected Joints (PDJ)



Method	U.arms	L.arms	Mean
Ours	**97.0**	**86.8**	**91.9**
Tompson, NIPS'14	93.7	80.9	87.3
MODEC, CVPR'13	84.4	52.1	68.3

Table 2: Comparison of *strict* PCP results on the FLIC dataset. Our method significantly outperforms state of the art.

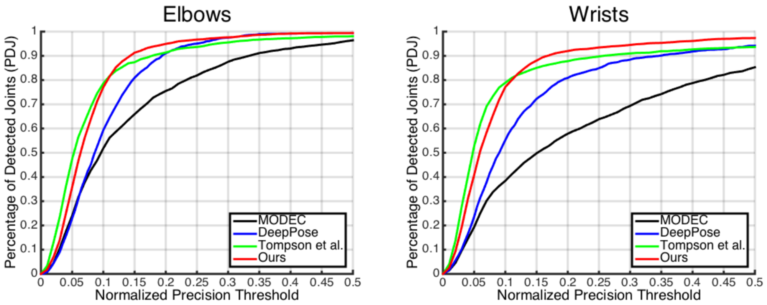


Figure 1: Comparison of PDJ curves of elbows and wrists on the FLIC dataset.

Diagnostic Experiments

- Term Analysis
  - ConvNet for extracting information from patches.
  - Stronger pairwise relations (IDPR).

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
Unary-Only	56.3	66.4	28.9	15.5	50.8	45.9	40.5
No-IDPRs	87.4	74.8	60.7	43.0	73.2	65.1	64.6
Full Model	**92.7**	**87.8**	**69.2**	**55.4**	**82.9**	**77.0**	**75.0**

Table 3: Diagnostic term analysis *strict* PCP results on the LSP dataset. The unary term alone is still not powerful enough to get good results, even though it's trained by a DCNN classifier. *No-IDPRs* method, whose pairwise terms are not dependent on the image, can get comparable performance with the state-of-the-art, and adding IDPR terms significantly boost our final performance to 75.0%.

Diagnostic Experiments

- Cross-dataset Generalization.
  – Apply model trained on FLIC to Buffy dataset.

Method	U.arms	L.arms	Mean
Ours*	96.8	89.0	92.9
Ours* *strict*	94.5	84.1	89.3
Yang [27]	97.8	68.6	83.2
Yang [27] *strict*	94.3	57.5	75.9
Sapp [21]	95.3	63.0	79.2
FLPM [11]	93.2	60.6	76.9
Eichner [5]	93.2	60.3	76.8

Table 3: Cross-dataset PCP results on Buffy test subset. The PCP numbers are *Buffy* PCP unless otherwise stated.

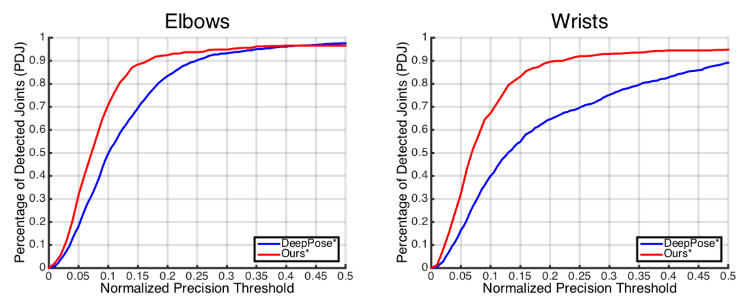
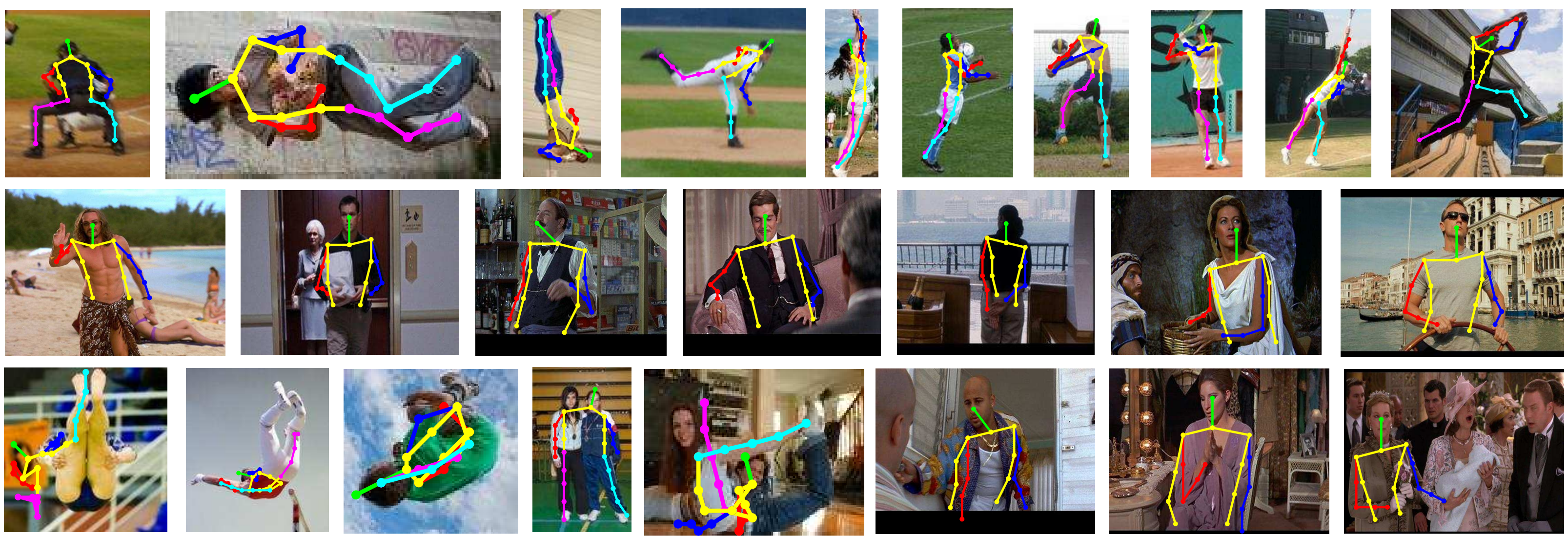


Figure 2: Cross-dataset PDJ curves on Buffy test subset. Note that both our method and DeepPose [23] are trained on the FLIC dataset.

Results

- The last row shows some failure cases
  - large foreshortening, occlusions.
  - distractions from clothing or overlapping people.

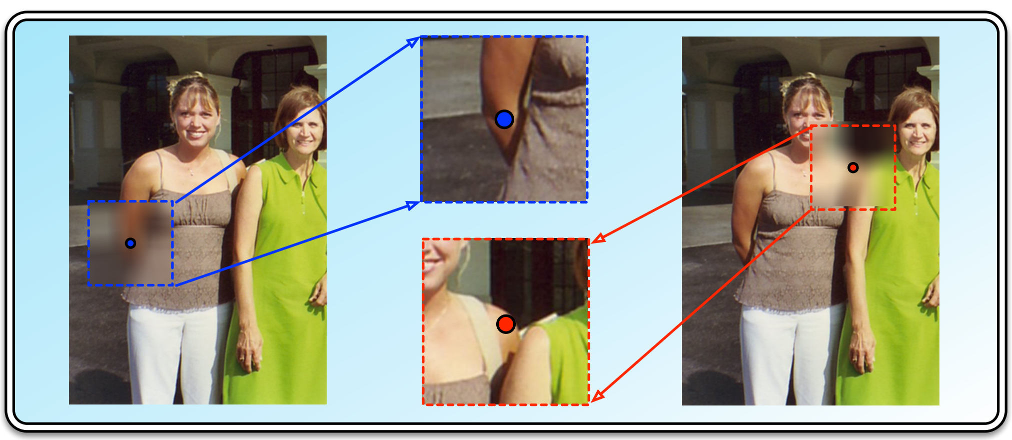


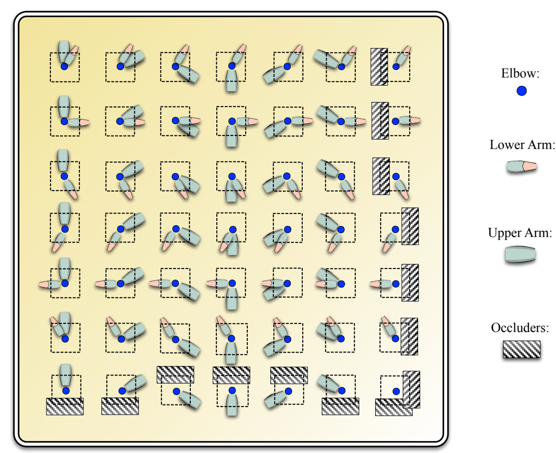
How about occlusion?

- People are often significantly occluded
  - Parse humans when there is significant occlusion.
  - Predict part occlusion & localize visible parts.

Key Idea – Occlusion Modeling

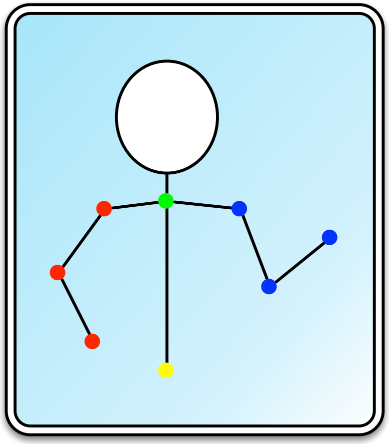
- Classical: Cue from absence of evidence for body part.
- local image measurements -> occlusion cue
  - Local patch around the occlusion boundary can reliably provide evidence of occlusion.



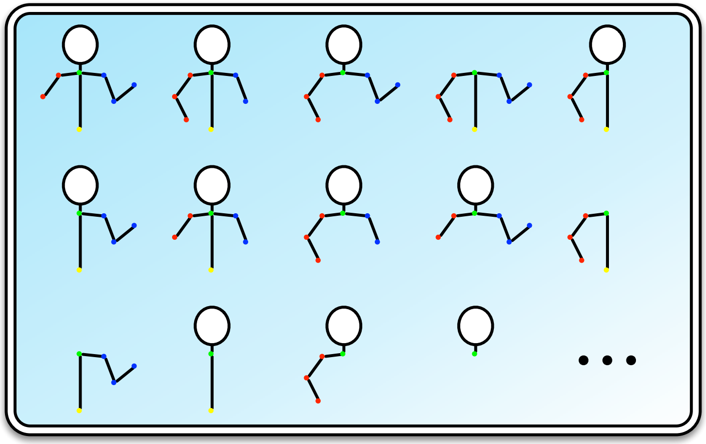


Key Idea – Flexible Compositions

- Occlusions often occur in regular patterns.
  - Connectivity prior: the visible parts of human tend to consist of a subset of connected parts.
  - Flexible compositions: all the possible connected sub-trees of the graph.



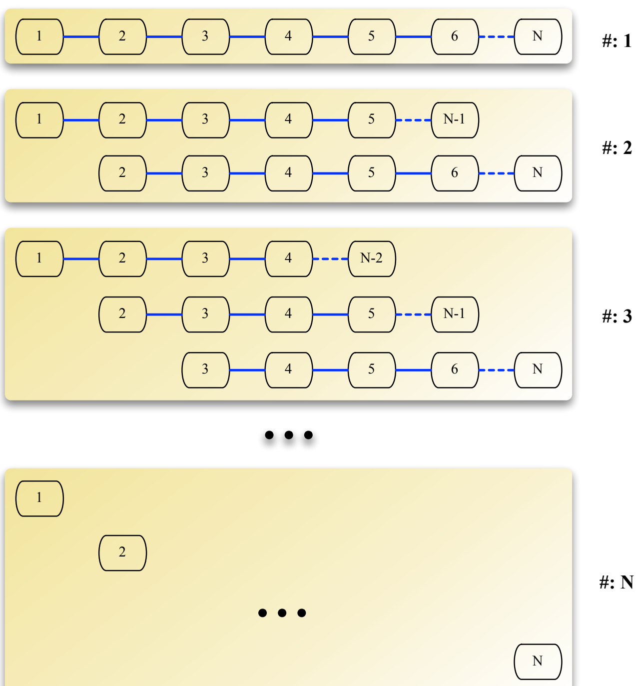
Full Graph



Flexible Compositions

Flexible compositions (FCs)

- Chain like model with N parts: # of FCs = N(N+1)/2.
  - # of FCs with K parts = N-K+1
- Exploit part sharing for efficient inference.

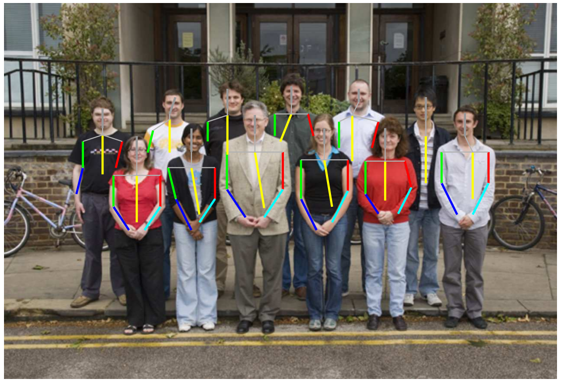


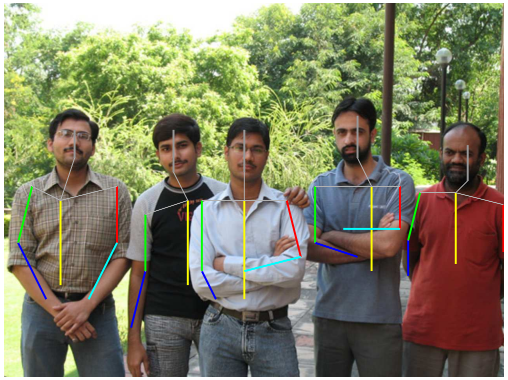
Connectivity Prior

- Experimental Verification
  - 95.1% of the people instances have their visible parts form a connected subtree.
- Hard to verify that some isolated pieces of body parts belong to the same person.

State of the art

- Significantly outperforms alternatives on benchmark dataset:
  - the state of the art methods.
  - and our base model (i.e., not modeling occlusion).





Parsing Occluded People by Flexible Compositions.
Xianjie Chen, Alan Yuille
Computer Vision and Pattern Recognition (CVPR), 2015.

The Graphical Model

- Tree model: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - The pixel locations $\mathbf{l}_i = (x, y)$ of part $i \in \mathcal{V}$
  - Pairwise relation types $t_{ij} \in \{1, \ldots, T_{ij}\}, \forall (i, j) \in \mathcal{E}$
  - Binary occlusion decoupling variable γ_{ij} on each edge
- Unary terms: $A(\mathbf{l}_i|\mathbf{I}) = w_i \phi(i|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta})$
- Image Dependent Pairwise Relational (IDPR) Terms:

$$
\begin{aligned}
R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji}|\mathbf{I}) = & \langle \mathbf{w}_{ij}^{t_{ij}}, \boldsymbol{\psi}(\mathbf{l}_j - \mathbf{l}_i - \mathbf{r}_{ij}^{t_{ij}}) \rangle \\
& + w_{ij} \varphi^s(t_{ij}, \gamma_{ij} = 0|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) \\
& + \langle \mathbf{w}_{ji}^{t_{ji}}, \boldsymbol{\psi}(\mathbf{l}_i - \mathbf{l}_j - \mathbf{r}_{ji}^{t_{ji}}) \rangle \\
& + w_{ji} \varphi^s(t_{ji}, \gamma_{ji} = 0|\mathbf{I}(\mathbf{l}_j); \boldsymbol{\theta})
\end{aligned},
$$

The Graphical Model

- Image Dependent Occlusion Decoupling (IDOD) Terms: $D_{ij}(\gamma_{ij}=1,\mathbf{l}_i|\mathbf{I}) = w_{ij}\varphi^d(\gamma_{ij}=1|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta})$
- Bias Terms for decoupling the subtree $\mathcal{T}_j = (\mathcal{V}(\mathcal{T}_j), \mathcal{E}(\mathcal{T}_j))$ at part i: $B_{ij} = \sum_{k \in \mathcal{V}(\mathcal{T}_j)} b_k$
- The model score for each flexible compositions $c \in \mathcal{C}_{\mathcal{G}}$

$$
\begin{aligned}
F(\mathbf{l},\mathbf{t},\mathcal{G}_c|\mathbf{I},\mathcal{G}) = & \sum_{i \in \mathcal{V}_c} A(\mathbf{l}_i|\mathbf{I}) \\
& + \sum_{(i,j) \in \mathcal{E}_c} R(\mathbf{l}_i,\mathbf{l}_j,t_{ij},t_{ji}|\mathbf{I}) \\
& + \sum_{(i,j) \in \mathcal{E}_c^d} (B_{ij} + D_{ij}(\gamma_{ij}=1,\mathbf{l}_i|\mathbf{I}))
\end{aligned}
$$

$\mathcal{E}_c^d = \{(i,j) \in \mathcal{E} | i \in \mathcal{V}_c, j \notin \mathcal{V}_c\}$ is the edges that are decoupled.

Efficient Inference

- Maximize the model score by searching
  - the flexible composition
  - the configurations of locations and types

$$(c^*, \mathbf{l}^*, \mathbf{t}^*) = \arg\max_{c,\mathbf{l},\mathbf{t}} F(\mathbf{l}, \mathbf{t}, \mathcal{G}_c | \mathbf{I}, \mathcal{G})$$

- Efficient Inference by exploiting part sharing
  - Proved: only twice as expensive as searching for the entire object (i.e., not modeling occlusion).

Learning

- Fully supervised learning
  - Annotated part locations / part occlusion.
  - Derive pairwise type labels by clustering.
- Three sets of parameters
  - Mean relative positions r of different pairwise relation types, by K-means clustering.
  - Parameters θ of image dependent terms, by ConvNet.
  - Weight parameters $\mathbf{w}$, by linear SVM.

Benchmark Performance

- "We Are Family" Dataset:
  - Accuracy of Occlusion Prediction (AOP)
  - Percentage of Correct Part (PCP)

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Ours	**84.9**	**88.5**	**98.5**	**77.2**	**71.3**	**80.7**
Multi-Person [11]	80.0	86.1	97.6	68.2	48.1	69.4
Ghiasi et. al. [17]	74.0	-	-	-	-	63.6
One-Person [11]	73.9	83.2	97.6	56.7	28.6	58.6

Table 1: Comparison of PCP and AOP on the WAF dataset. Our method improves the PCP performance on all parts, and significantly outperform the best previously published result [11] by 11.3% on mean PCP, and 4.9% on AOP.

Diagnostic Experiments

- Term Analysis
  - Flexible composition representation.
  - Cues from local image measurement around the occlusion boundary (IDOD term).

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Base Model [6]	73.9	81.4	92.6	63.6	47.6	66.1
FC	82.0	87.0	**98.6**	72.7	67.5	77.7
FC+IDOD	**84.9**	**88.5**	98.5	**77.2**	**71.3**	**80.7**

Table 1: Diagnostic Experiments PCP and AOP results on the WAF dataset. Using flexible compositions (*i.e.*, *FC*) significantly improves our base model [6] by 11.6% on PCP and 8.1% on AOP. Adding *IDOD* terms (*FC+IDODs*, *i.e.*, the full model) further improves our PCP performance to 80.7% and AOP performance to 84.9%, which is significantly higher than the state of the art methods.

Results

Questions & Suggestions