Detecting and Parsing Humans

Alan Yuille (JHU)

Summary

- This talk describes recent work on detection and parsing visual objects. The methods represent objects in terms of object parts encoding spatial relations between them.
- We use deep convolutional neural networks (DCNNs) to make proposals for detecting the object parts.
- We will use graphical models to reason about spatial relations.
- We extend to graphical models that deal with occlusion.

Compositional Strategy

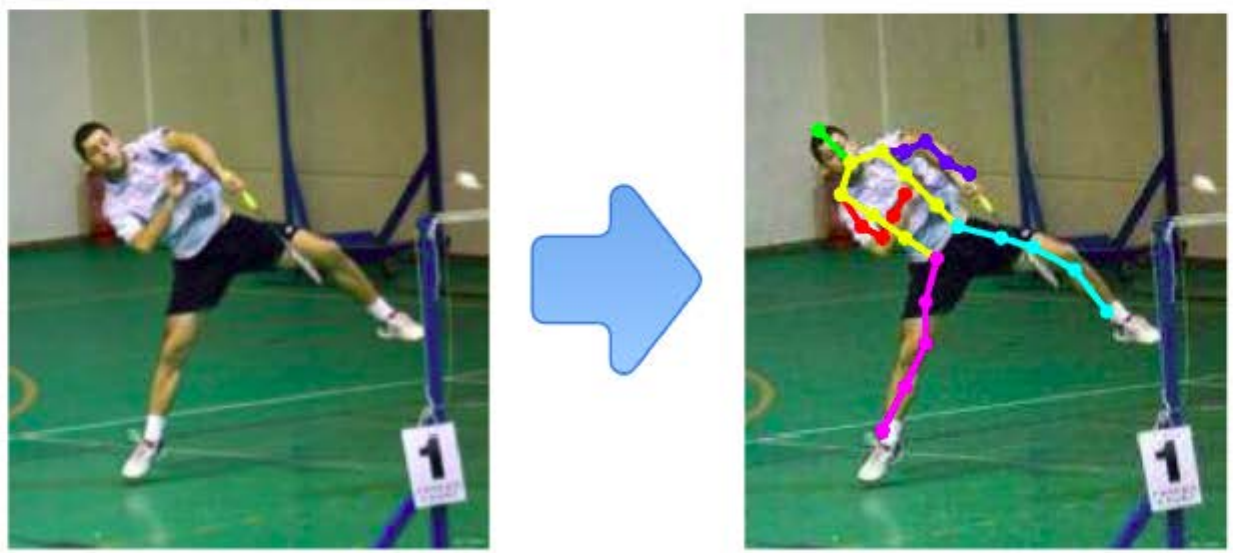
- Deep Convolutional Neural Networks (DCNNs) have been extremely successful for many visual tasks – such as object detection.
- But DCNNs are complicated “black boxes” and it is hard to understand what they are doing. They do not have explicit representations of object parts and the spatial relationships between them.
- Our strategy is to represent objects in terms of compositions of object parts. DCNNs are trained to detect parts. Then we use explicit graphical models – including AND/OR graphs – to encode spatial relations and to enable part sharing.
-

Parsing Human – Joint Detection

- In this project, the parts are joints (e.g., elbows, wrists, shoulders,…).
- Graphical models are used to represent spatial relationships between the parts.
- Part sharing is used to enable efficient inference when the human is occluded.
- X. Chen and A.L. Yuille (NIPS 2014, CVPR 2015).

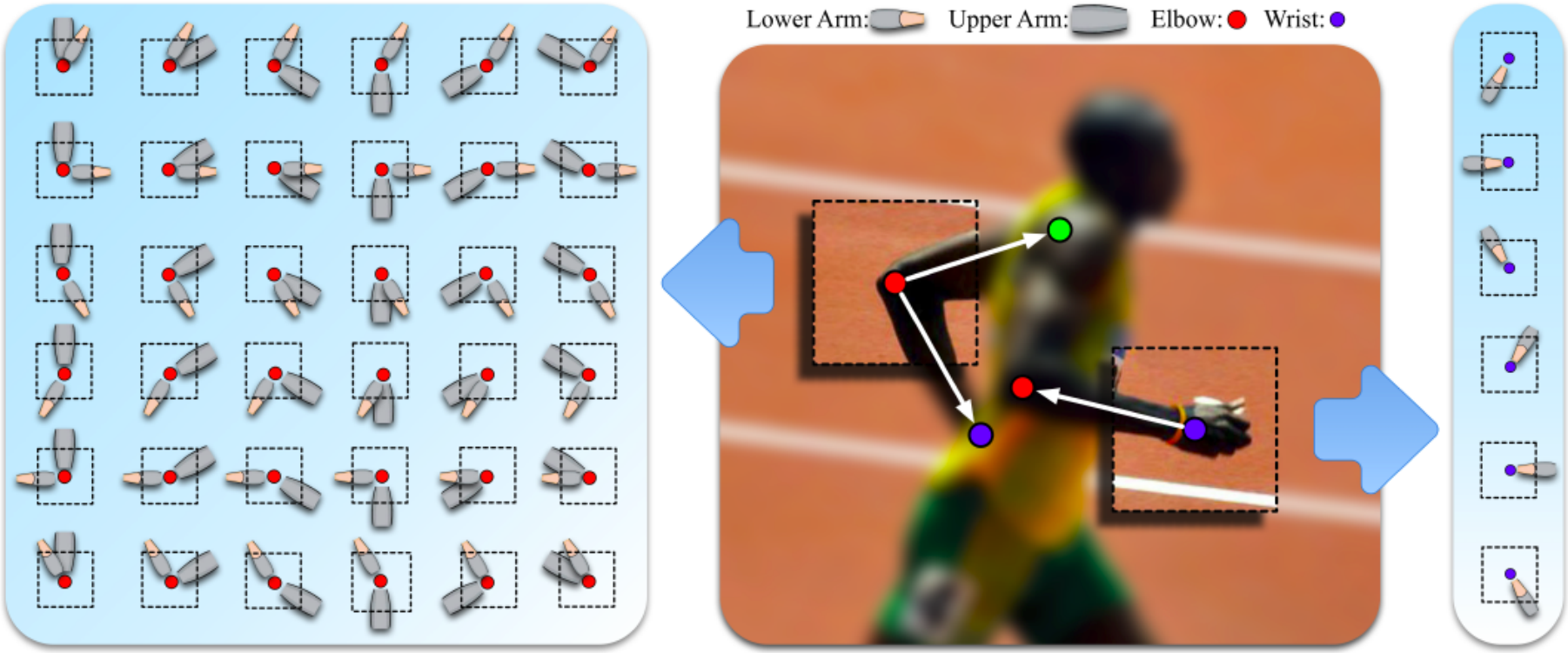
Introduction

- Task is to estimate articulated human pose from a single static image.



❑ Image Dependent Pairwise Relations (IDPRs)

- **Intuition**: We can reliably predict the relative positions of a part's neighbors (as well as the presence of the part itself) by *only* observing the local image patch around it.
- We specify a graphical model for human pose with novel pairwise relations that make adaptive use of local image measurements.



DCNN for Image Dependent Terms

- Require a method to extract information about *pairwise part relations*, as well as *part presence*, from local image patches.
- Deep Convolutional Neural Network (DCNN) is suitable for this, since it is efficient and share features between different parts and part relationships.

Performance Summary

State of the Art Performance

- Our model combines the representational flexibility of graphical models with the efficiency and statistical power of DCNNs.
- Significantly outperforms the state of the art methods on the **LSP** and **FLIC** datasets and also performs very well on the **Buffy** dataset without any training on it.

The Graphical Model

- Variables of the Tree Model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - The pixel location $\mathbf{l}_i = (x, y)$ 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:

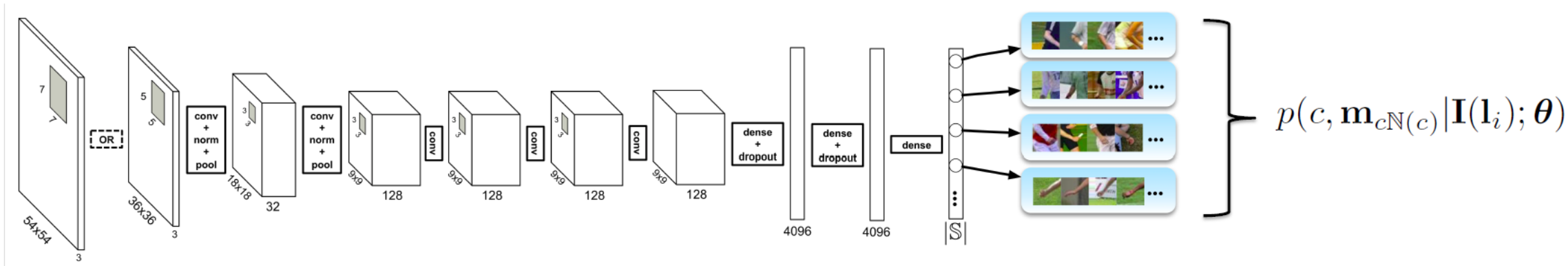
$$\begin{aligned} R(\mathbf{l}_i, \mathbf{l}_j, t_{ij}, t_{ji}|\mathbf{I}) = & \quad \langle \mathbf{w}_{ij}^{t_{ij}}, \boldsymbol{\psi}(\mathbf{l}_j - \mathbf{l}_i - \mathbf{r}_{ij}^{t_{ij}}) \rangle + w_{ij} \varphi(t_{ij}|\mathbf{I}(\mathbf{l}_i); \boldsymbol{\theta}) \\ + & \quad \langle \mathbf{w}_{ji}^{t_{ji}},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

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

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

DCNN for Image Dependent Terms

- Appearance terms $\phi(.|.;\boldsymbol{\theta})$ and IDPR terms $\varphi(.|.;\boldsymbol{\theta})$ depend on the image.
- We use DCNN to learn the conditional probability distribution $p(c, \mathbf{m}_{c\mathbb{N}(c)}|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta})$ defined on the space $|\mathbb{S}|$, where each element corresponds to a part with all the types of its pairwise relationships, or the background.



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

Inference

- Dynamic programming + Distance Transform: $O(T^2LK)$
  - L: # of locations, K: # of parts, T: # of pairwise types
- Image Dependent Terms are efficiently calculated by a single DCNN at all locations.
  - The computations common to overlapping regions are shared by considering fully-connected layers as 1x1 convolutions.

Relationship to other models

Pictorial Structure (PS)

- Recover PS by allowing one pairwise relation type, i.e., $T_{ij} = 1$
- We use DCNN to learn data term instead of HOG filters.

Yang and Ramanan's Mixtures-of-parts (MOP) [26]

- MOP defines different "types" of part by its relative position with respect to its parent.
- Recover MOP by restricting each part in our model to only predict the relative position of its parent, i.e., $T_{ij} = 1$ if j is not parent of i.

Conditional Random Fields (CRFs)

- Related to CRFs literature on data dependent priors.
- Efficiently model all the image dependent terms in a single DCNN.

Learning

- Supervised learning by deriving pairwise type labels from the annotated part (joint) locations by clustering.
- Learn three sets of parameters:
  - Mean relative positions **r** of different pairwise relation types, by K-means clustering.
  - Parameters θ of image dependent terms, by DCNN.
  - Weight parameters $\mathbf{W}$, by linear SVM.

Benchmark Performance

❑ LSP

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.

- Two recent ECCV'14 papers, Kiefel&Gehler and Ramakrishna et al., also report performance on the LSP dataset, and our performance is better.

❑ FLIC

Method	U.arms	L.arms	Mean
Ours	**97.0**	**86.8**	**91.9**
MODEC [20]	84.4	52.1	68.3

Table 2: Comparison of *strict* PCP results on the FLIC dataset. Our method significantly outperforms MODEC [20].

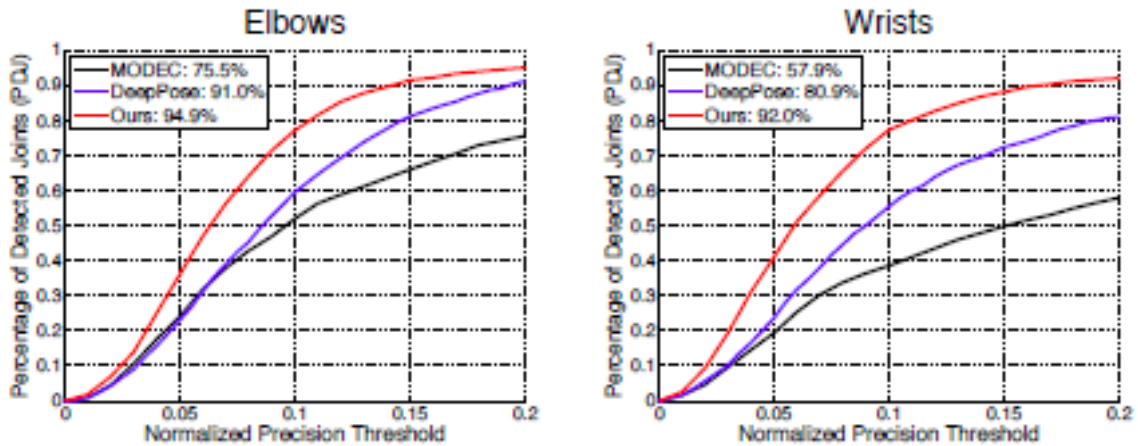


Figure 1: Comparison of PDJ curves of elbows and wrists on the FLIC dataset. The legend shows the PDJ numbers at the threshold of 0.2.

Datasets

- **Leeds Sports Poses (LSP) dataset**: 1000 training and 1000 testing full-body human poses.
- **Frames Labeled In Cinema (FLIC) dataset**: 3987 training and 1016 testing upper-body human poses.
- **Buffy Stickmen dataset**: 276 testing upper-body human poses. *We do not train on this dataset.*

Diagnostic Experiments

❑ Terms Analysis

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%.

❑ Cross-dataset Generalization

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.

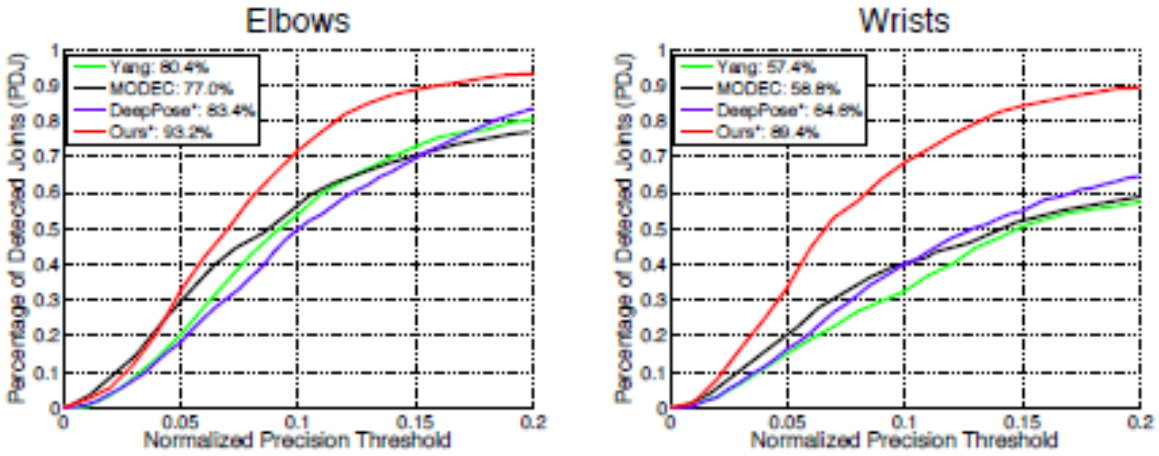


Figure 2: Cross-dataset PDJ curves on Buffy test subset. The legend shows the PDJ numbers at the threshold of 0.2. Note that both our method and DeepPose [23] are trained on the FLIC dataset.

- Compared with Figure 1., the margin between our method and DeepPose significantly increases in Figure 2., which implies that our model generalizes better to the Buffy dataset.

Parsing People by Flexible Compositions. (Chen and Yuille CVPR 2015).

- In realistic images many object parts are occluded.
- Previous graphical model are robust to only a few occlusions.
- Prior – observed nodes of graphical model are often connected.
- Strategy: extend the method used in NIPS 2014 to deal with occlusion.

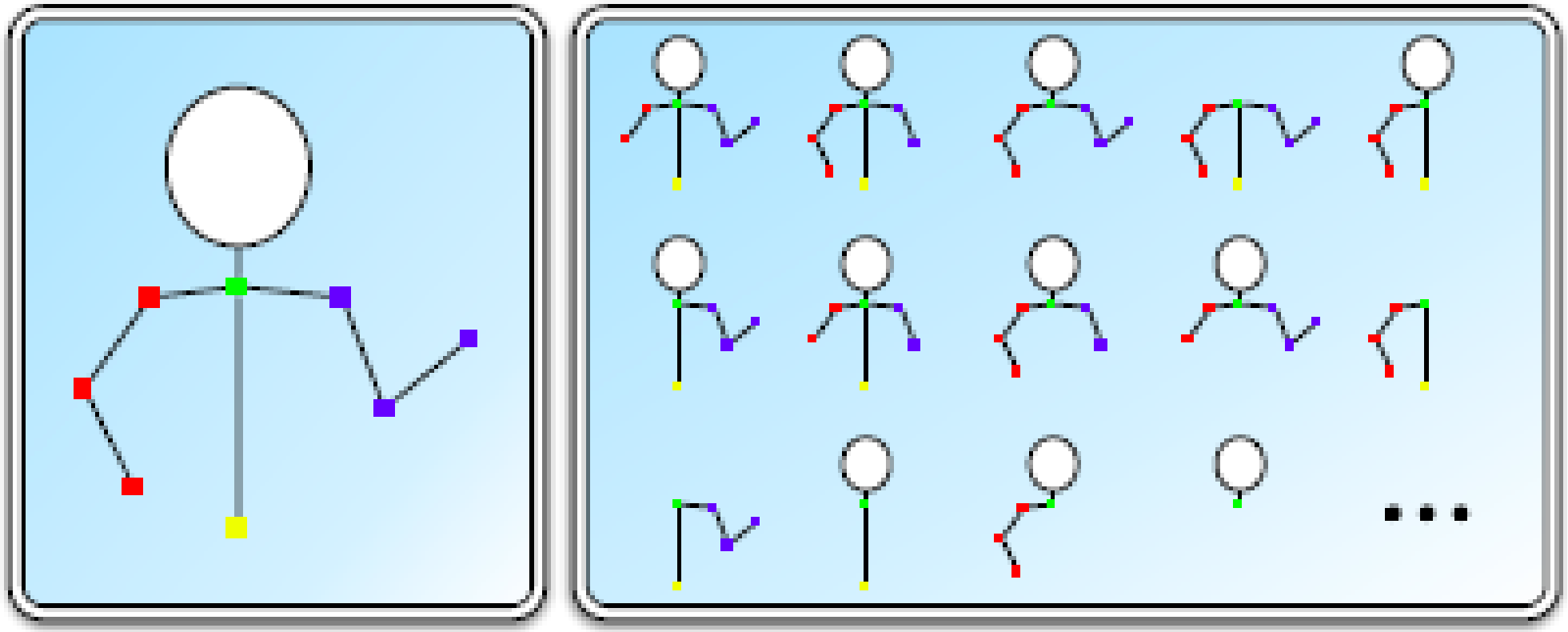
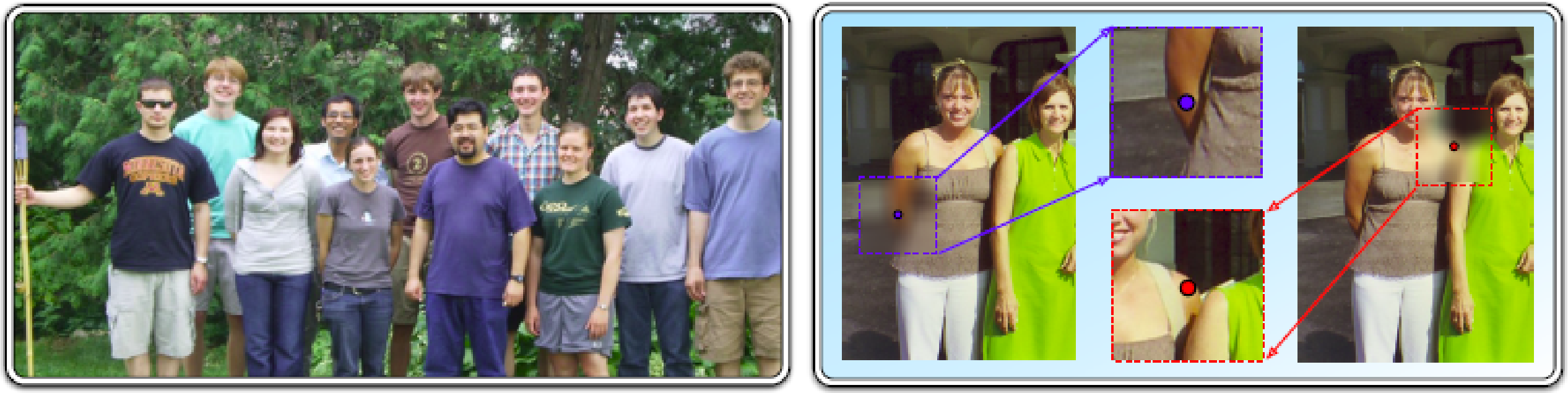


Figure 1: An illustration of the *flexible compositions*. Each connected subtree of the *full graph* (include the full graph itself) is a flexible composition. The flexible compositions that do not have certain parts is suitable for the people with those parts occluded.



(a) (b)

Figure 2: Motivation. (a): In real world scenes, people are usually significantly occluded (or truncated). Requiring the model to localize a fixed set of body parts while ignoring the fact that different people have different degrees of occlusion (or truncation) is problematic. (b): The absence of body parts evidence can help to predict occlusion, *e.g.*, the right wrist of the lady in brown can be inferred as occluded because of the absence of suitable wrist near the elbow. However, absence of evidence is not evidence of absence. It can fail in some challenging scenes, for example, even though the left arm of the lady in brown is completely occluded, there is still strong image evidence of suitable elbow and wrist at the plausible locations due to the confusion caused by nearby people (*e.g.*, the lady in green). In both situations, the local image measurements near the occlusion boundary (*i.e.*, around the right elbow and left shoulder), *e.g.*, in a image patch, can reliably provide evidence of occlusion.

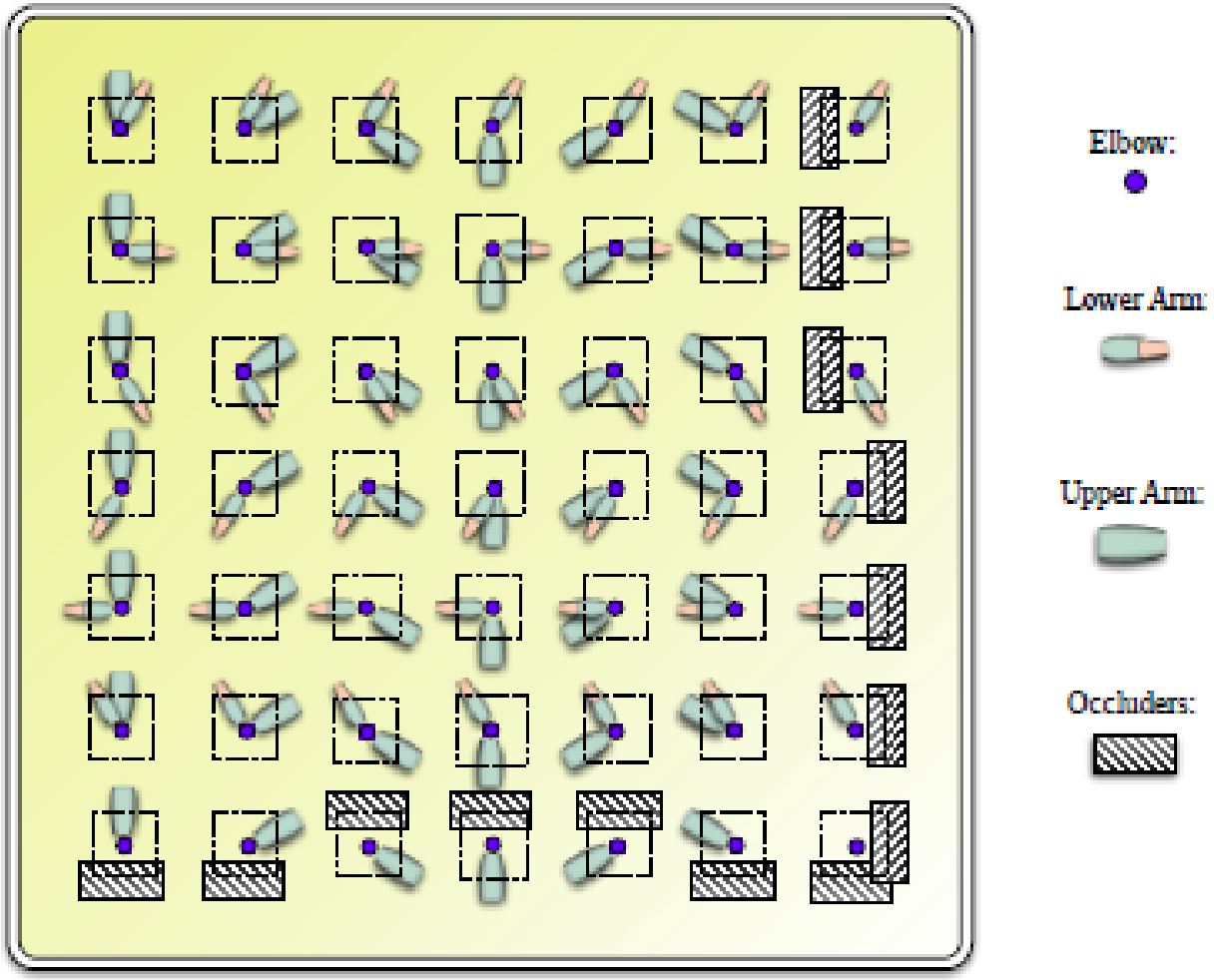


Figure 3: Different occlusion decoupling and spatial relationships between the elbow and its neighbors, *i.e.*, wrist and shoulder. The local image measurement around a part (*e.g.*, the elbow) can reliably predict the relative positions of its neighbors when they are not occluded, which is demonstrated in the base model [5]. In the case when the neighboring parts are occluded, the local image measurement can also reliably provide evidence for the occlusion.

Model

- Base Model: as before.
- Introduce decoupling terms
- Penalties for missing terms

$$D_{ij}(\gamma_{ij}=1,\mathbf{l}_i|\mathbf{I}) = w_{ij}\varphi^d(\gamma_{ij}=1|\mathbf{I}(\mathbf{l}_i);\boldsymbol{\theta}),$$

$$B_{ij} = \sum_{k\in\mathcal{V}(T_j)} b_k$$

$$\begin{aligned} F(\mathbf{l},\mathbf{t},\mathcal{G}_c|\mathbf{I},\mathcal{G}) = & \sum_{i\in\mathcal{V}_c} A(\mathbf{l}_i|\mathbf{I}) \\ & + \sum_{(i,j)\in\mathcal{E}_c} R(\mathbf{l}_i,\mathbf{l}_j,t_{ij},t_{ji}|\mathbf{I}) \\ & + \sum_{(i,j)\in\mathcal{E}_c^d} (B_{ij} + D_{ij}(\gamma_{ij}=1,\mathbf{l}_i|\mathbf{I})) \end{aligned} \tag{5}$$

Inference

- There are many different models – no. of connected subtrees of the graph.
- But inference is efficient because of part-sharing.
- Inference is only twice the complexity of the base model:
- $$O(2T^2LK)$$

Evaluation

- "We Are Family" (WAF) Dataset
- 525 images, six people per image on average. (350/175 train/test).

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Ours	**84.9**	**88.5**	**98.5**	**77.2**	**71.3**	**80.7**
Multi-Person [9]	80.0	86.1	97.6	68.2	48.1	69.4
Ghiasi et. al. [15]	74.0	-	-	-	-	63.6
One-Person [9]	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 [9] by 11.3% on mean PCP, and 4.9% on AOP.

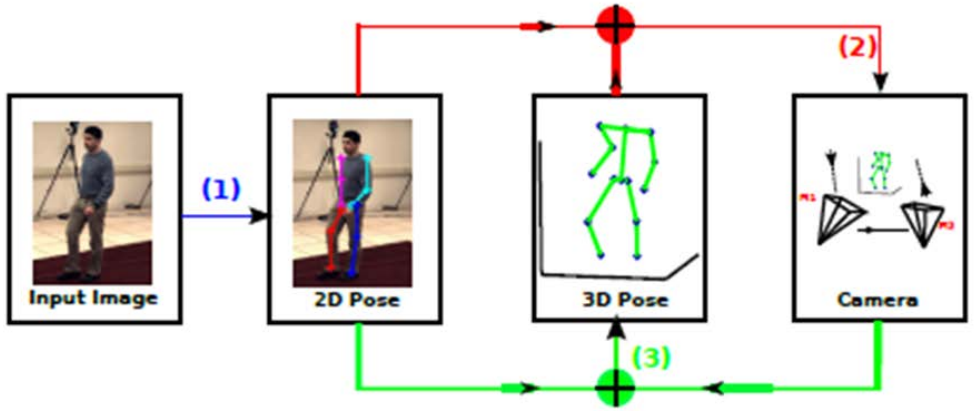
Diagnostics

Method	AOP	Torso	Head	U.arms	L.arms	mPCP
Base Model [5]	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 2: Diagnostic Experiments PCP and AOP results on the WAF dataset. Using flexible compositions (*i.e.*, *FC*) significantly improves our base model [5] 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.

Figure 5: Results on the WAF dataset. We show the parts that are inferred as visible, and thus have estimated configurations, by our model.

From 2D to 3D.



- Pose detection – with and with occlusion.
- Prior – connected parts – for occlusion (validated on WAF)
- Efficient inference despite occlusion – due to part sharing.

- Note: detection of pose is important for many applications.

E.g., estimating of 3D structure (C. Wang et al. 2014), action recognition (C. Wang et al, 2013, 2014).

Collaboration with Peking University.

Summary of Part I: Parsing Humans -- Joints

- Detection of object parts (joints) in presence of occlusion. DCNNs for detecting parts, graphical models to impose spatial relations, efficient inference using dynamic programming.
- The detected parts can be used to estimate 3D structure of humans from a single image and enable action recognition.

- Limitations. Objects are represented in terms of joints only. This becomes problematic in some human configurations.

Papers Cited.

- X. Chen and A.L. Yuille. Articulated Pose Estimation with Image-Dependent Preference on Pairwise Relations. NIPS 2014.
- X. Chen and A.L. Yuille. Parsing Occluded People by Flexible Compositions. CVPR 2015.
- C. Wang, Y. Wang, Z. Lin, A.L. Yuille, and W. Gao .Robust Estimation of 3D Human Poses from Single Images . CVPR. 2014.