A Multi-layer Composite Model for Human Pose Estimation

Kun Duan[1], Dhruv Batra[2], David Crandall[1]

[1]Indiana University, Bloomington, IN; [2]Toyota Technological Institute at Chicago, Chicago, IL

INDIANA UNIVERSITY

1. Overview

- *Multi-layer composition of different tree-structured part-based models.*
- *Each layer captures human pose at a different scale.*
- *Dual Decomposition for efficient inference.*
- *Outperform state-of-the-art under different evaluation metrics.*

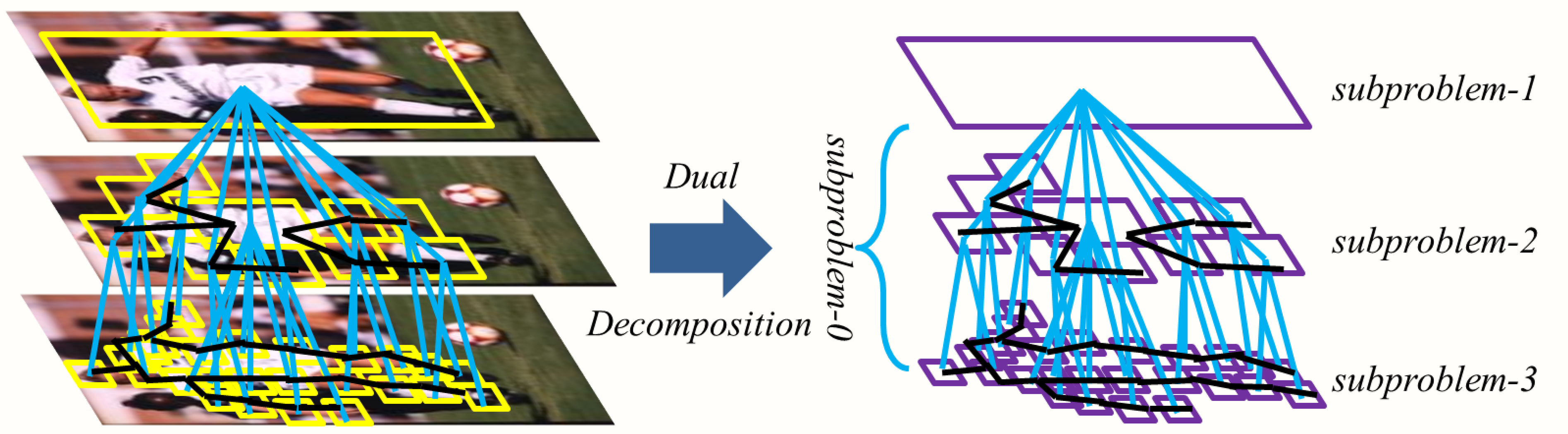


Figure 1: *Our multi-layer composite part-based model.*

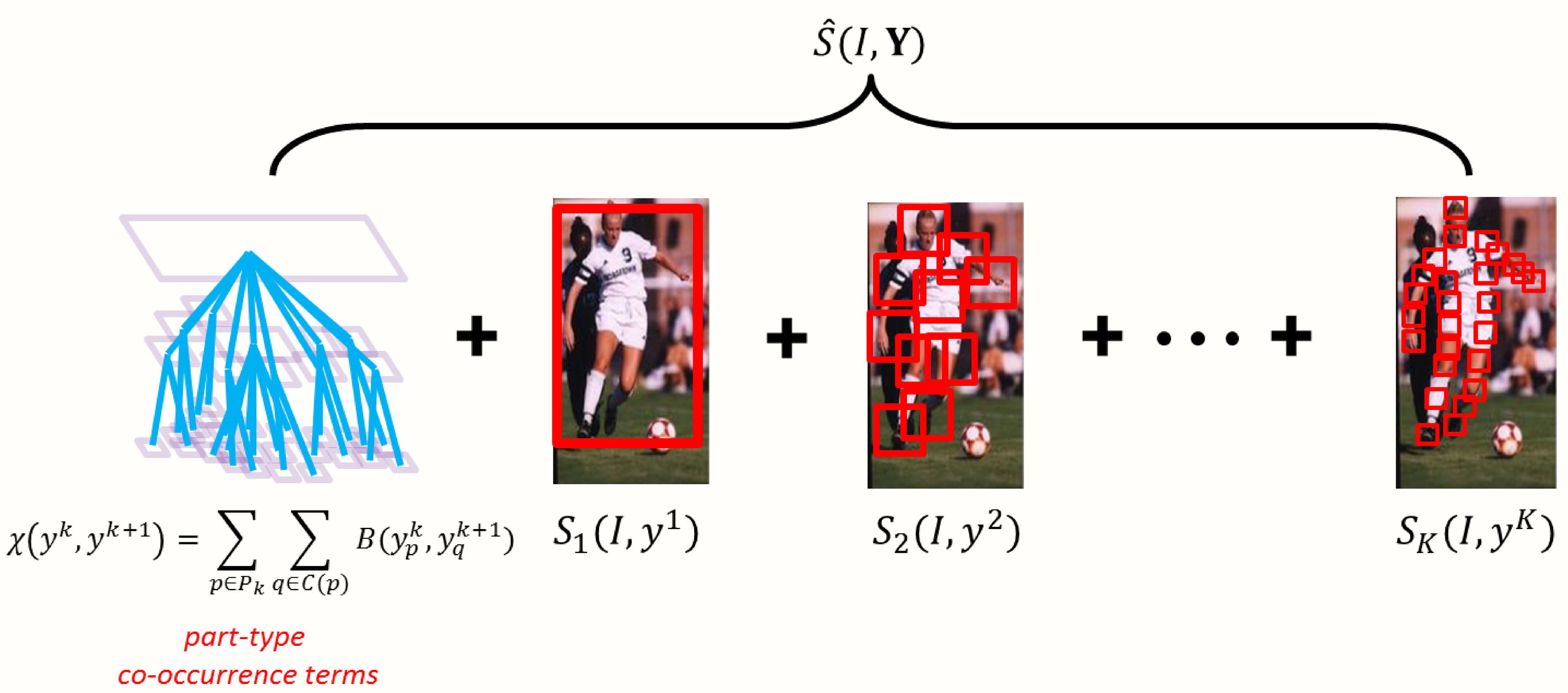
2. Model

- **Single layer model.** Our model is built on the *mixture of parts tree model* in Yang & Ramanan (CVPR11):

local filter score *deformation score* *part type co-occurrence score*

$$S(I,\mathbf{y}) = \sum_{p\in P} D(I,\mathbf{y}_p) + \sum_{(p,q)\in E} (L(\mathbf{y}_p,\mathbf{y}_q) + T(\mathbf{y}_p,\mathbf{y}_q))$$

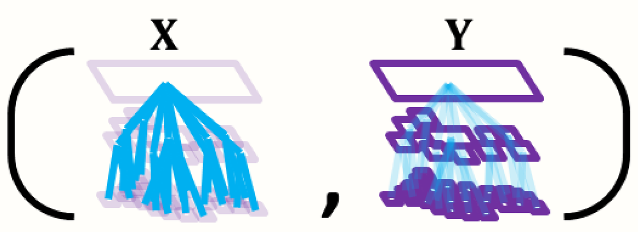
- **Multi-layer composite model.** The proposed model generalizes the above model with multiple layers:



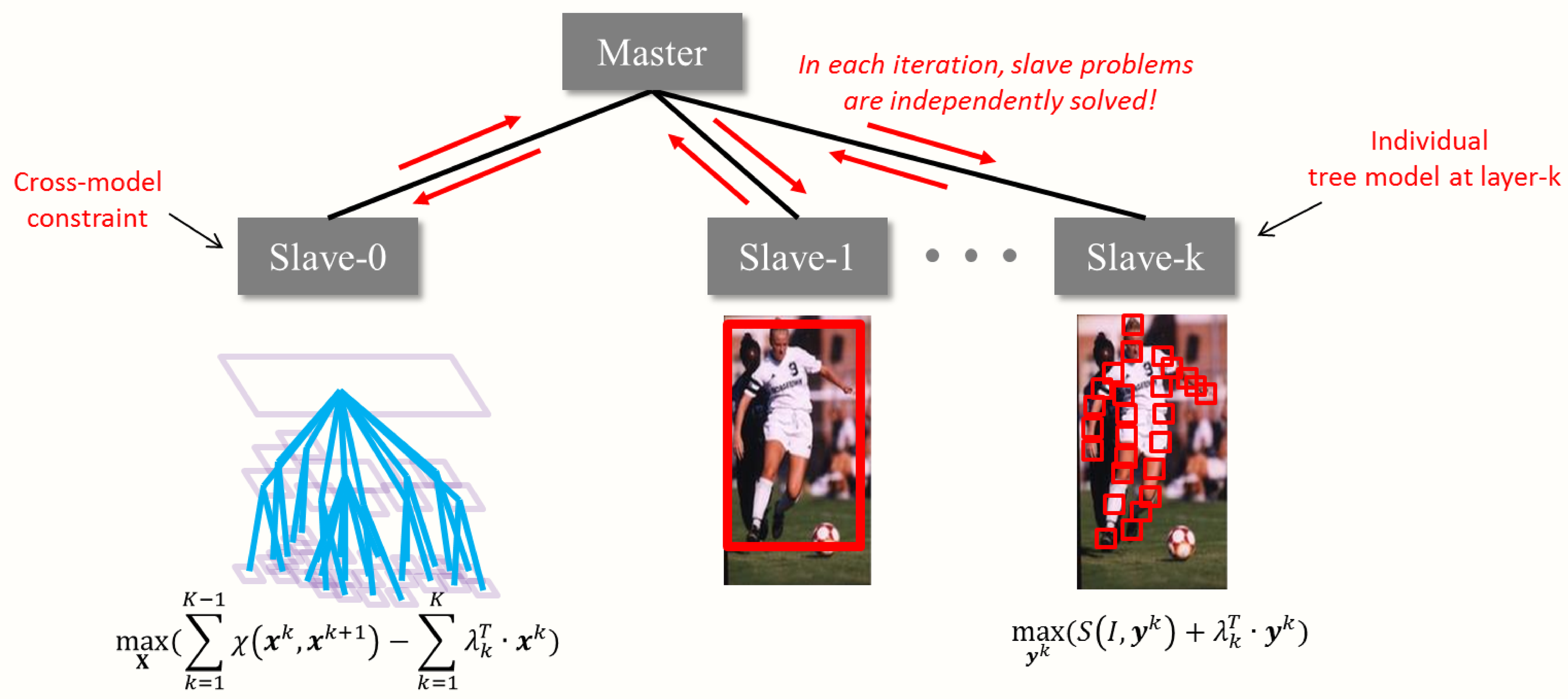
3. Inference

- We adopt *Dual Decomposition* for efficient inference, which naturally decomposes the original graphical structure into *multiple trees*.

$$\mathbf{Y} = (\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^K) \longrightarrow$$



$$g(\lambda) = \max_{\mathbf{X},\mathbf{Y}} \sum_{k=1}^{K} S(I,\mathbf{y}^k) + \sum_{k=1}^{K-1} \chi(\mathbf{x}^k,\mathbf{x}^{k+1}) + \sum_{k=1}^{K} \lambda_k \cdot (\mathbf{y}^k - \mathbf{x}^k)$$



- *Subgradient descent:*

Check agreement; Update dual variables

$$\lambda_k^{(t+1)} \leftarrow \lambda_k^{(t)} - \alpha^{(t)}(y^k(\lambda_k^{(t)}) - x^k(\lambda_k^{(t)}))$$

4. Learning

- *Structural SVM* formulation: $\hat{S}(I,\mathbf{Y}) = \beta\cdot\Phi$

weights *features*

$$\min_{\beta} \frac{1}{2}\|\beta\|^2 + C\sum_m \xi_m$$

$$\text{s.t.} \quad \beta\cdot\Phi(I_m,\mathbf{Y}_m) \geq 1-\xi_m \qquad \forall m \in \text{pos}$$

$$\beta\cdot\Phi(I_m,\mathbf{Y}) \leq -1+\xi_m \qquad \forall m \in \text{neg}, \forall \mathbf{Y}$$

5. Experiments

- *Datasets.* We use Parse [2] and UIUC Sport [1] in our experiments.
- *Evaluation criteria.*

How to define correct part localization	
1. Distance of *each* endpoint from ground truth endpoint is less than a threshold.	2. *Mean* distance between estimated and ground truth endpoints is less than a threshold.

How to compute final PCP score	
A. Calculated for *each* image; Averaged on *all*.	B. Calculated *only* for *correct* person detections; Averaged and multiplied by detection rate.

- *Quantitative results.*

(a) **Table 1**: PCP (1A) on Parse & UIUC Sport.

	Parse dataset							UIUC Sport dataset						
	Torso	UL	LL	UA	LA	Head	Total	Torso	UL	LL	UA	LA	Head	Total
Ramanan2006	52.1	37.5	31.0	29.0	17.5	13.6	27.2	28.7	7.3	19.2	7.5	20.6	12.9	15.1
Wang2011	–	–	–	–	–	–	–	75.3	49.2	39.5	25.2	11.2	47.5	37.3
Yang2011	82.9	69.0	63.9	55.1	35.4	77.6	60.7	85.3	61.3	55.5	49.7	35.5	73.5	56.3
Ours (**26+10**)	82.0	**72.4**	**67.8**	55.6	**36.6**	79.0	62.6	85.4	61.6	**57.9**	49.1	34.8	72.9	56.4
Ours (**26+1**)	**85.6**	71.7	65.6	**57.1**	**36.6**	**80.4**	**62.8**	86.0	**62.2**	57.5	**51.0**	**36.3**	73.7	**57.3**
Ours (**26+10+1**)	81.0	71.7	67.6	55.9	36.3	79.5	62.3	**86.2**	61.2	55.7	49.9	35.9	**73.8**	56.5
Pishchulin2012*	88.8	77.3	67.1	53.7	36.1	73.7	63.1	–	–	–	–	–	–	–
Johnson2011*	87.6	74.7	67.1	67.3	45.8	76.8	67.4	–	–	–	–	–	–	–

*Pishchulin2012 and Johnson2011 are not directly comparable due to the use of more annotations.

Our composite models outperform state-of-the-art

(b) **Table 2**: Effect of PCP 1A, 1B, 2B on Parse.

	PCP (variant 1A)								PCP 1B	PCP 2B
Threshold	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.5	0.5
Yang2011	33.4	47.2	56.0	60.7	64.4	67.2	69.7	71.5	56.0	74.9
Ours (**26+10**)	**34.5**	**49.2**	**57.6**	62.6	65.9	68.7	71.3	73.0	58.5	75.0
Ours (**26+1**)	**34.5**	48.3	56.5	**62.8**	**66.9**	**70.0**	**72.0**	**73.6**	59.3	75.8
Ours (**26+10+1**)	34.3	48.9	57.3	62.3	65.7	68.6	70.9	72.7	**59.5**	**75.9**

*Evaluation metrics **significantly** affect the final PCP scores*

- *Qualitative results.*

Figure 2: Sample results. *Left*: Examples in which [2] failed (top), but our 3-level model estimated poses correctly (bottom). *Right*: Some failure cases of our model.

6. Conclusions

- A general framework for combining different pose estimation models.
- Our model outperforms state-of-the-art methods on challenging datasets.

Acknowledgements: We thank Devi Parikh for helpful comments and discussions. This work was supported in part by the Lilly Endowment and the Indiana University Data-to-Insight Center.

References:

[1] Y. Wang, D. Tran, Z. Liao, Learning hierarchical poselets for human parsing, CVPR 2011
[2] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, CVPR 2011