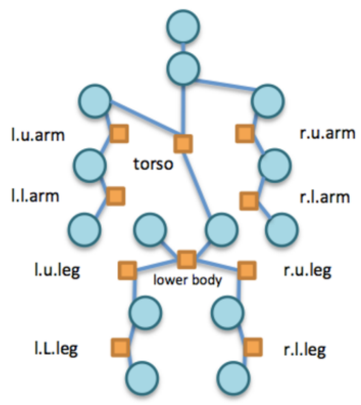


Beyond Physical Connections: Tree Models in Human Pose Estimation

Fang Wang[1, 2] and Yi Li[2]

1 Nanjing University of Science and Technology, China
2 Canberra Research Laboratory, NICTA, Canberra, Australia;
Yi.Li@nicta.com.au

Australian National University

NICTA

Problem summary

- Human pose estimation in images via tree models
- Attempt to answer the following critical questions:
  - Are simple tree models sufficient?
  - How to use tree models in human pose estimation?
  - How shall we use combined parts with single parts?
- Latent tree models for discovering graphical model structure
  - Exact inference
  - Visual categorization for combined parts
  - Better performance

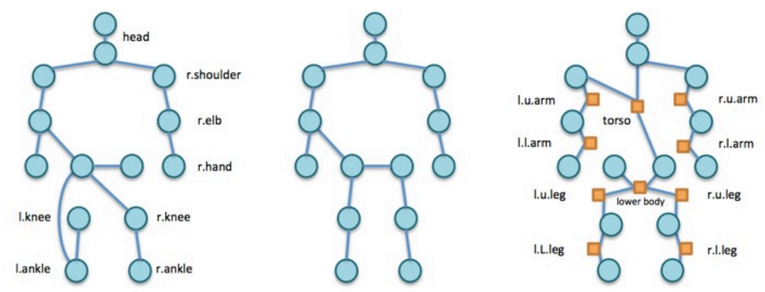
Latent tree models for human pose

- Learn a tree structure directly from our observations without making many assumptions of the physical constraints
- Information distance: $d_{ij} = -\log(\frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}})$

RG ⇒ ⇒

CLRG ⇒ ⇒

primitive parts
combined parts
hidden nodes

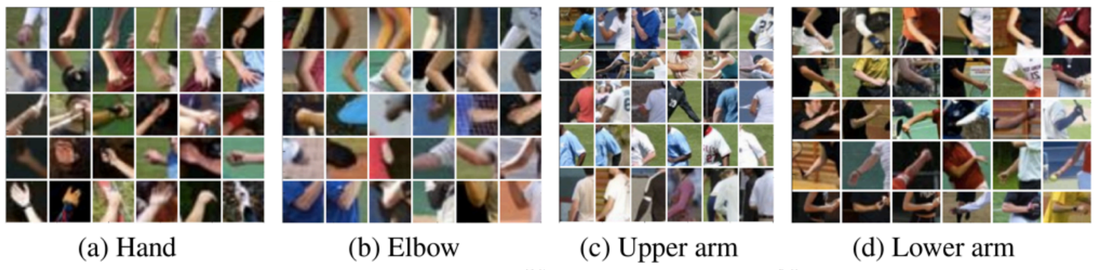


Our Approach

A framework for integrating primitive parts and combined parts [1]

- Primitive parts (non-oriented): geometric clustering [4]
- Combined parts: Visual Categorization SVM+HOG [3]
- Tree structured models Learned directly from data
- Textbook example of exact inference

Visual categorization for combined parts



(a) Hand (b) Elbow (c) Upper arm (d) Lower arm

Latent SVM [3]

$$\arg\min_{w} \frac{1}{2}\sum_{k=1}^{K} \| w_k \|^2 + C\sum_{i=1}^{N}\epsilon_i,$$

$$y_i w_{t_i}\phi(x_i) \geq 1 - \epsilon_i, \epsilon_i \geq 0, t_i = \arg\max_k w_k\phi(x_i)$$

Learned HOG Filters

Results

Dataset:

- LSP: 2000 images, subject-centric
- PARSE: 305 images, image-centric
- Pascal Dog dataset: subset

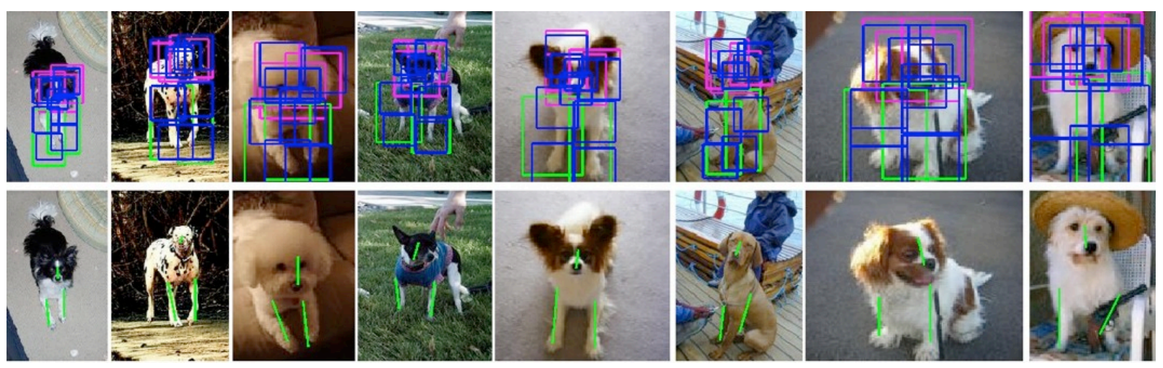
LSP

Exp.		Method	Torso	Head	U.Leg	L.Leg	U.Arm	L.Arm	Total
LSP	L	Yang & Ramanan	92.6	87.4	66.4	57.7	50.0	30.4	58.9
	L	Tian *et al.* (First 200)	93.7	86.5	68.0	57.8	49.0	29.2	58.8
	L	Tian *et al.* (5 models)	**95.8**	**87.8**	69.9	60.0	**51.9**	32.8	61.3
	L	Ours (First 200)	88.4	80.8	69.1	60.0	50.5	29.2	59.0
	L	Ours	91.9	86.0	**74.0**	**69.8**	48.9	32.2	**62.8**
	S	Johnson & Everingham	78.1	62.9	**65.8**	**58.8**	**47.4**	**32.9**	**55.1**
	S	Yang & Ramanan	82.0	75.8	54.4	51.6	41.0	28.4	50.9
	S	Ours (Restrict eval)	**88.3**	**81.4**	55.3	55.3	43.1	30.5	53.8
PARSE (cross dataset)	L	Yang & Ramanan	78.8	70.0	66.0	61.1	**61.0**	**37.4**	60.0
	L	Ours	**88.3**	**78.7**	**75.2**	**71.8**	60.0	35.9	**65.3**

L: Loose evaluation S: Strict evaluation

Dog pose

Method	Head	L.F.Leg	R.F.Leg	Legs	Total
Yang & Ramanan, CVPR 2011	**56.1**	52.8	58.3	55.6	55.7
Ours	52.8	**60.6**	**63.3**	**62.0**	**58.9**



Conclusion

- Tree models for human pose estimation are efficient
- Latent tree is an effective tool for recovering model structure
- Learning visual category of combined part becomes important.

References

[1] Fang Wang and Yi Li, "Beyond Physical Connections: Tree Models in Human Pose Estimation", CVPR 2013
[2] Fang Wang and Yi Li, "Learning Visual Symbols for Parsing Human Poses in Images", IJCAI 2013
[3] S. Divvala, A. Efros, and M. Hebert, "How important are deformable parts in the deformable parts model?," CoRR, vol. abs/1206.3714, 2012
[4] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in CVPR 2011

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council

From imagination to impact