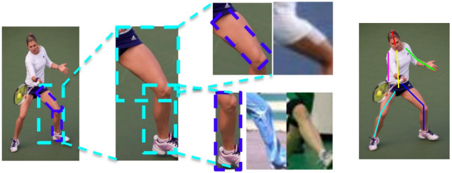
Beyond Physical Connections: Tree Models in Human Pose Estimation

Fang Wang[1,2] and **Yi Li**[2,3]

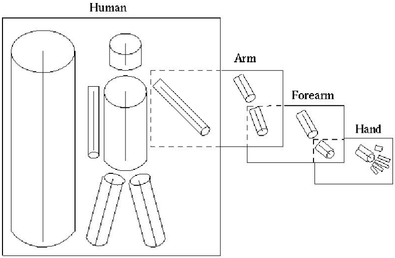
1. Nanjing University of Science and Technology, China
2. NICTA, Australia
3. Australian National University

yi.li@nicta.com.au

- Models for human body
  - Multiple granularity
  - Tree structure
  - Flexibility
  - Interaction
  - Latent structure

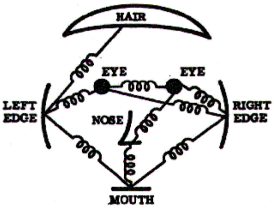


- Models for human body
  - Multiple granularity
  - Tree structure
  - Flexibility
  - Interaction
  - Latent structure



Marr, Vision

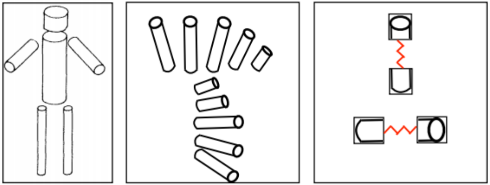
- Models for human body
  - Multiple granularity
  - Tree structure
  - Flexibility
  - Interaction
  - Latent structure



Felzenszwalb and Huttenlocher, IJCV 2005

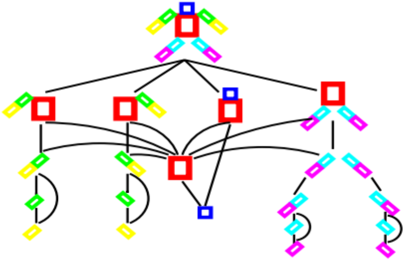
NICTA

- Models for human body
  - Multiple granularity
  - Tree structure
  - Flexibility
  - Interaction
  - Latent structure



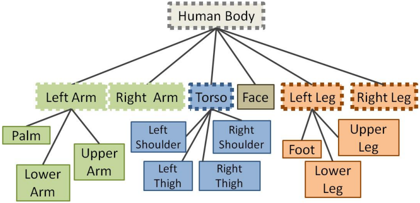
Yang and Ramanan, CVPR 2011

- Models for human body
  - Multiple granularity
  - Tree structure
  - Flexibility
  - Interaction
  - Latent structure



Wang et al, JMLR 12

- Models for human body
    - Multiple granularity
    - Tree structure
    - Flexibility
    - Interaction
    - Latent structure



Tian et al, ECCV 12

Manually defined structure

Learn the structure?

A model of **learned** structure

- handles compositional parts
- explores latent structure
- is still a tree
- captures dynamics beyond physical connections

A model of **learned** structure

- handles compositional parts
- explores latent structure
- is still a tree
- captures dynamics beyond physical connections

NICTA

A model of **learned** structure

- handles compositional parts
- explores latent structure
- is still a tree
- captures dynamics beyond physical connections

NICTA

A model of **learned** structure

- handles compositional parts
- explores latent structure
- is still a tree
- captures dynamics beyond physical connections

A **model** of learned structure

- handles compositional parts
- explores **latent** structure
- is still a **tree**
- captures dynamics beyond physical connections

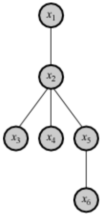
NICTA

Latent tree for pose estimation (1)

Latent Tree

To learn tree structured models
for approximating joint distribution of observable variables

- Tree building algorithms:
  - [Chow and Liu, 1968]
  - [Choi et al, JMLR 2011]
- Motivations
  - Novel latent models for human, or
  - Discover intrinsic structures



Wikipedia

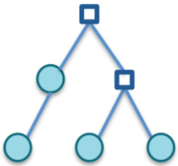
NICTA

Latent tree for pose estimation (1)

Latent Tree

To learn tree structured models
for approximating joint distribution of observable variables

- Tree building algorithms:
  - [Chow and Liu, 1968]
  - [Choi et al, JMLR 2011]
- Motivations
  - Novel latent models for human, or
  - Discover intrinsic structures



NICTA

Latent Tree

To learn tree structured models
for approximating joint distribution of observable variables

- Tree building algorithms:
  - [Chow and Liu, 1968]
  - [Choi et al, JMLR 2011]
- Motivations
  - Novel latent models for human, or
  - Discover intrinsic structures

Latent tree for pose estimation (1)

Latent Tree

To learn tree structured models
for approximating joint distribution of observable variables

- Tree building algorithms:
  - [Chow and Liu, 1968]
  - [Choi et al, JMLR 2011]
- Motivations
  - Novel latent models for human, or
  - Discover intrinsic structures

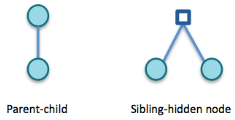
NICTA

Latent tree for pose estimation (2)

Definition

Information distance: $d_{ij} = -\log(\frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}})$

- Parent-Child relationship Test
  - For each triplet $i, j, k \in V$.
  - Define $\Phi_{ijk} \triangleq d_{jk} - d_{ik}$, take one of the two actions:
    - If $\Phi_{ijk} = d_{ij}$, j is set to be the parent of i.
    - If $-d_{ij} \leq \Phi_{ijk} = \Phi_{ijk'} \leq d_{ik}$ for all k and $k' \in V \backslash \{i, j\}$, add a hidden node as the parent of i and j.



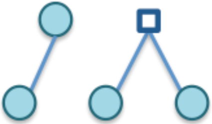
Recursive Grouping (RG)

- Initialize
- Test parent-child for pairs
- Repeat

[Choi et al, JMLR 2011]

Recursive Grouping (RG)

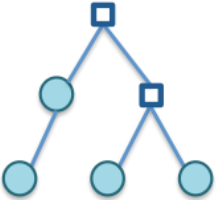
- Initialize
- Test parent-child for pairs
- Repeat



[Choi et al, JMLR 2011]

Recursive Grouping (RG)

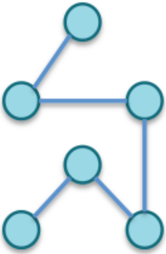
- Initialize
- Test parent-child for pairs
- Repeat



[Choi et al, JMLR 2011]

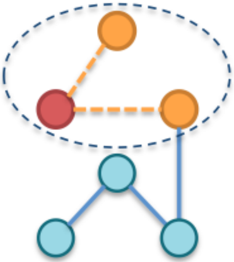
Chow-Liu Recursive Grouping (CLRG)

- Minimal spanning tree
- Select neighbor of an internal node
- Perform RG and update structure



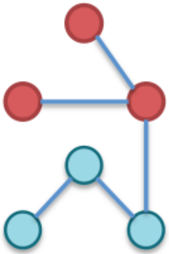
[Choi et al, JMLR 2011]

- Minimal spanning tree
- Select neighbor of an internal node
- Perform RG and update structure



[Choi et al, JMLR 2011]

- Minimal spanning tree
- Select neighbor of an internal node
- Perform RG and update structure



[Choi et al, JMLR 2011]

Building Latent Tree for Primitive Parts

Leeds Sport Pose from [Johnson and Everingham, BMVC 2010]

Our Approach

Learn a tree structured model for human pose estimation that integrates primitive parts and combined parts

- Primitive parts
    - Joints, non-oriented $\Rightarrow$ geometric clustering
    - [Yang and Ramanan, CVPR 2011]
- Combined parts
    - Distinctive $\Rightarrow$ Visual Categorization
    - SVM+HOG [Dalal and Triggs, CVPR 05]
- Tree structured models
    - Learned directly from data
    - Textbook example of exact inference and parameter learning

NICTA

Our Approach

Learn a tree structured model for human pose estimation
that integrates primitive parts and combined parts

- Primitive parts
  - Joints, non-oriented ⇒ geometric clustering
  - [Yang and Ramanan, CVPR 2011]
- Combined parts
  - Distinctive ⇒ Visual Categorization
  - SVM+HOG [Dalal and Triggs, CVPR 05]
- Tree structured models
  - Learned directly from data
  - Textbook example of exact inference and parameter learning

NICTA

Our Approach

Learn a tree structured model for human pose estimation that integrates primitive parts and combined parts

- Primitive parts
  - Joints, non-oriented $\Rightarrow$ geometric clustering
  - [Yang and Ramanan, CVPR 2011]
- Combined parts
  - Distinctive $\Rightarrow$ Visual Categorization
  - SVM+HOG [Dalal and Triggs, CVPR 05]
- Tree structured models
  - Learned directly from data
  - Textbook example of exact inference and parameter learning

NICTA

Our Approach

Learn a tree structured model for human pose estimation
that integrates primitive parts and combined parts

- Primitive parts
  - Joints, non-oriented $\Rightarrow$ geometric clustering
  - [Yang and Ramanan, CVPR 2011]
- Combined parts
  - Distinctive $\Rightarrow$ Visual Categorization
  - SVM+HOG [Dalal and Triggs, CVPR 05]
- Tree structured models
  - Learned directly from data
  - Textbook example of exact inference and parameter learning

NICTA

- Learn visual categories for combined parts
  - *k*-means algorithm on geometric config to find mean patch sizes
  - Latent SVM [Divvala et al, 2012] model for each combined part
  - Further info: [Wang and Li, IJCAI 2013]

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

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

$$t_i = \arg\max_{k}\; w_k\phi(x_i)$$

NICTA

Visual Categorization for Combined Parts

- Learn visual categories for combined parts
  - *k*-means algorithm on geometric config to find mean patch sizes
  - Latent SVM [Divvala et al, 2012] model for each combined part
  - Further info: [Wang and Li, IJCAI 2013]

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

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

$$t_i = \arg\max_{k} w_k \phi(x_i)$$

NICTA

Visual Categorization for Combined Parts

- Learn visual categories for combined parts
  - *k*-means algorithm on geometric config to find mean patch sizes
  - Latent SVM [Divvala et al, 2012] model for each combined part
  - Further info: [Wang and Li, IJCAI 2013]

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

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

$$t_i = \arg\max_{k} \; w_k\phi(x_i)$$

NICTA

Visual Categorization for Combined Parts

- Learn visual categories for combined parts
  - *k*-means algorithm on geometric config to find mean patch sizes
  - Latent SVM [Divvala et al, 2012] model for each combined part
  - Further info: [Wang and Li, IJCAI 2013]

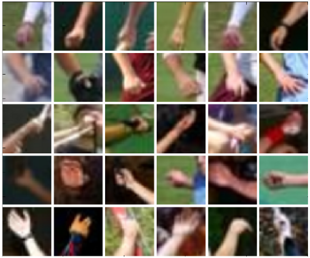
$$\arg\min_{w} \frac{1}{2}\sum_{k=1}^{K} \| w_k \|^2 + C\sum_{i=1}^{N} \epsilon_i,$$

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

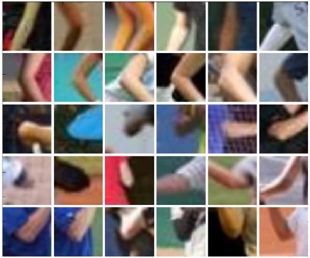
$$t_i = \arg\max_{k}\ w_k\phi(x_i)$$

NICTA

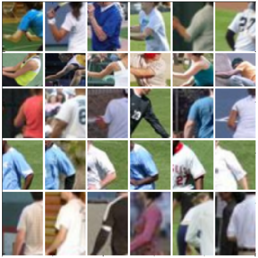
Results for Categorization



Hand



Elbow



Left arm

Left leg

Objective function

$$p = \arg\max_p S(t) + \sum_i S(I, p_i) + \sum_{i,j} S(I, p_i, p_j)$$

- Unary term
- Pairwise term
- Compatibility term

Defined as

$$S(I, p_i) = \omega_i^{t_i} \phi(I, loc_i)$$

NICTA

Objective Function for Inference

Objective function

$$p = \arg\max_p S(t) + \sum_i S(I, p_i) + \sum_{i,j} S(I, p_i, p_j)$$

- Unary term
- Pairwise term
- Compatibility term

DEFINED AS

$$S(I, p_i, p_j) = \omega_{ij}^{t_i t_j} \psi(p_i, p_j)$$

NICTA

Objective Function for Inference

Objective function

$$p = \arg\max_p S(t) + \sum_i S(I, p_i) + \sum_{i,j} S(I, p_i, p_j)$$

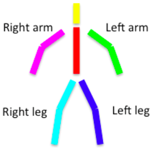
- Unary term
- Pairwise term
- Compatibility term

Defined as

$$S(t) = \sum b_i^{t_i} + \sum b_{ij}^{t_i t_j}$$

NICTA

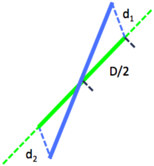
Experiments



human coordinate demo

image coord | human coord | image coord | human coord

PARSE dataset, from [Ramanan, NIPS 2006]

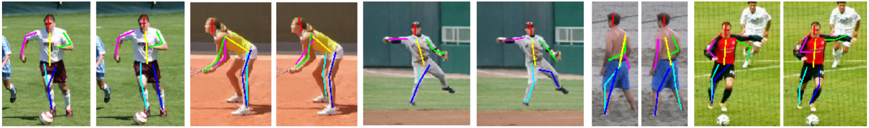


Strict evaluation: $d_1 < D/2$, $d_2 < D/2$

Loose evaluation: $(d_1 + d_2)/2 < D/2$

Percentage of Correct Parts (PCP)

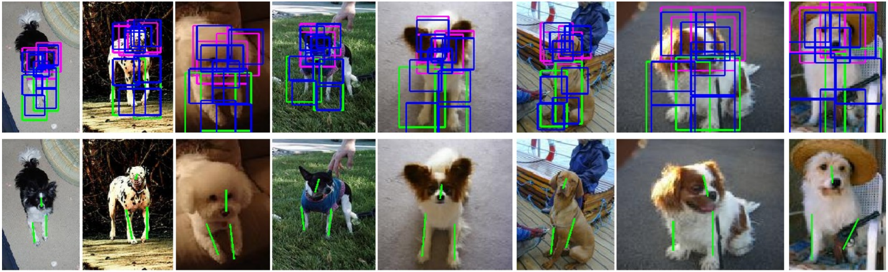
[Ferrari et al, CVPR 08]



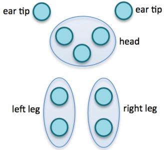
Exp.		Method	Torso	Head	U.Leg	L.Leg	U.Arm	L.Arm	Total
LSP	L	Yang & Ramanan	92.6	87.4	66.4	57.7	50.0	30.4	58.9
	L	Tian *et al.* (First 200)	93.7	86.5	68.0	57.8	49.0	29.2	58.8
	L	Tian *et al.* (5 models)	**95.8**	**87.8**	69.9	60.0	**51.9**	32.8	61.3
	L	Ours (First 200)	88.4	80.8	69.1	60.0	50.5	29.2	59.0
	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 (strict eval)	**88.3**	**81.4**	55.3	55.3	43.1	30.5	53.8
PARSE	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**

Table : Performance on the LSP dataset.

Experiments (2)



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 intrinsic structure
- Learning visual category of combined part

Thank you!

http://users.cecs.anu.edu.au/~yili/

yi.li@nicta.com.au

Funding support:
Bionic Eye (YL) and China Scholarship Council (FW)

Acknowledgement:
Prof. Yiannis Aloimonos and Dr. Cornelia Fermuller (Maryland), and Prof. Luciano Fadiga (IIT Italy)
Dr. Mathieu Salzmann and other NICTA folks.