Modeling Mutual Context of Object and Human Pose in Human-object Interaction Activities

- Bangpeng Yao
- Li Fei-Fei

Presented by Sahil Shah

Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

Introduction

- Note on author
  - Pioneer of ImageNet dataset
  - Must see TED talk in March 2015

Introduction

- Problem: Detecting objects in cluttered scenes and estimating articulated human body parts especially in human object interaction activities

Introduction

Introduction

Introduction

- Key insight: Mutual Context
  - Automatically discover relevant poses
  - Automatically discover spatial relationships
  - Optimize for mutual co-occurrence of object and pose

Introduction

- Contribution
  - Builds up on Prof. Gupta's work
  - First to use mutual context
  - Jointly solve object detection & pose estimation

Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

Problem Formulation

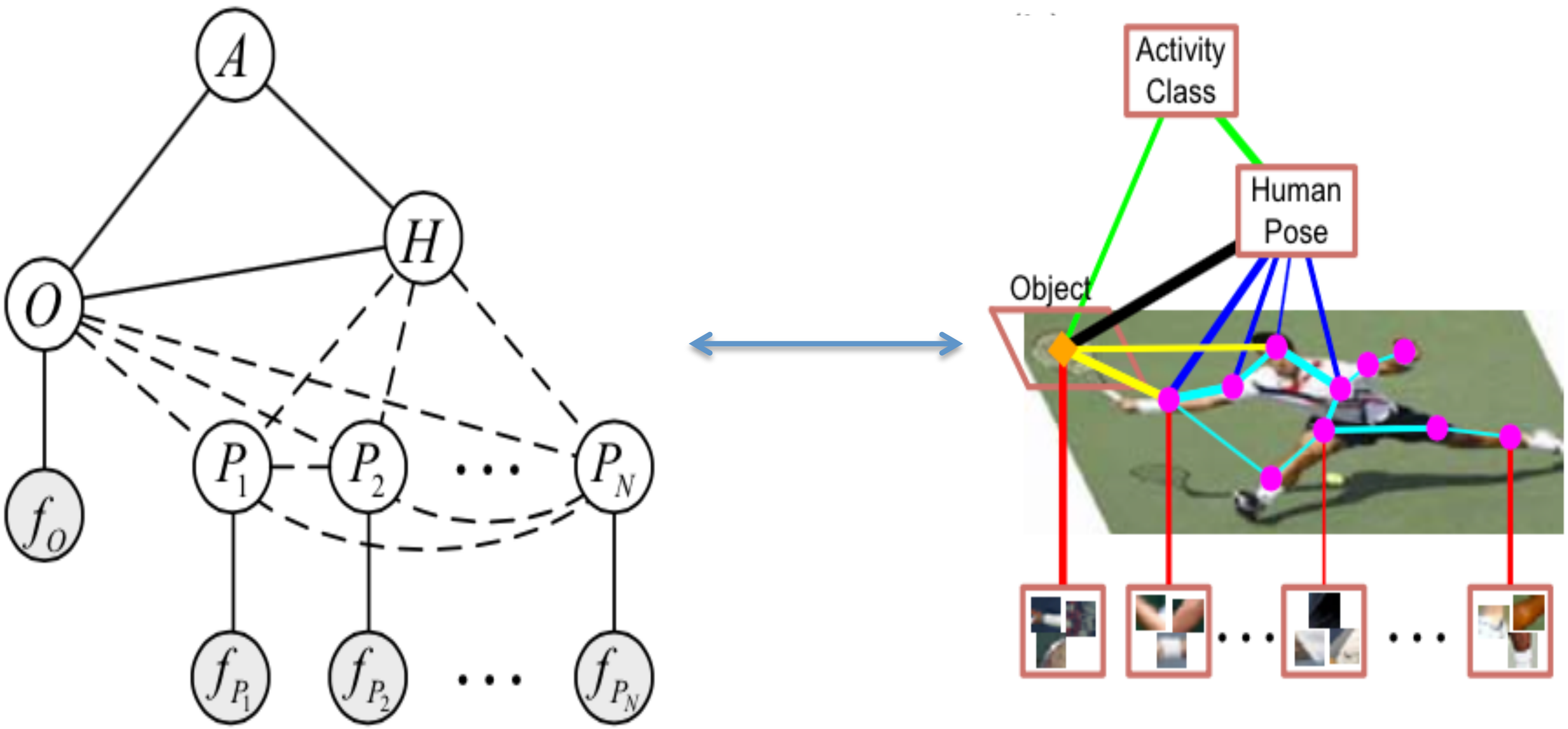
- Goal: Given an image of HOI activity we need to estimate human pose(H), detect the object(O) and classify HOI activity(A)
- Model
  - Hierarchical Random Field
  - A,O and H contribute to detection of each other
  - H is a hidden variable
  - Body parts $\{P_n\}$ are found using feature based detectors and they compose to form H

Problem Formulation

Golf Swing

Tennis Forehand

Problem Formulation



Problem Formulation

- Why need to learn structure?
  - The model captures important connections between object and the body parts
  - Which parts of the body should be connected to overall pose (H) and object (O)?

Problem Formulation

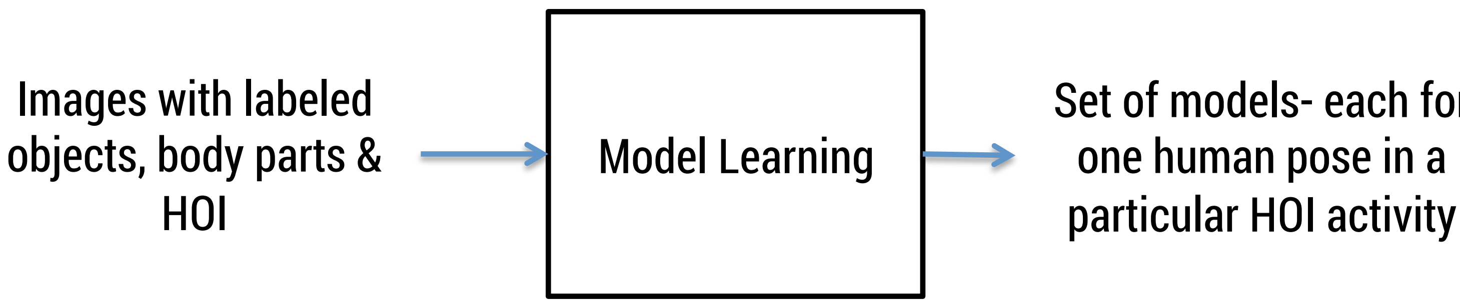
- Model
  - Overall model: $\Psi = \sum \boldsymbol{w_e \psi_e}$
  - A,O,H: $\boldsymbol{\psi_e(A, O)}$, $\boldsymbol{\psi_e(A, H)}$, and $\boldsymbol{\psi_e(O, H)}$
    - Counting co-occurrence frequencies
  - Spatial Relationships: $\boldsymbol{\psi_e(O, P_n)}$ & $\boldsymbol{\psi e\ (P_m, P_n)}$
    - $\mathrm{bin}(\mathbf{l}_O - \mathbf{l}_{Pn}) \cdot \mathrm{bin}(\theta_O - \theta_{Pn}) \cdot \mathcal{N}(s_O / s_{Pn})$
  - Compatibility: $\boldsymbol{\psi_e(H, P_n)}$
    - $\mathrm{bin}(\mathbf{l}_{Pn} - \mathbf{l}_{P1}) \cdot \mathrm{bin}(\theta_{Pn}) \cdot \mathcal{N}(s_{Pn})$
  - Object & Body parts: $\boldsymbol{\psi_e(O, f_O)}$ and $\boldsymbol{\psi_e(P_n, f_{Pn})}$
    - Shape context feature based detectors

Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

Learning

- Input and Output



Learning

- Overall Algorithm

Hill-climbing structure learning for each activity class.

foreach *Iteration* **do**

- Model parameter estimation by max-margin learning;
- Choose the model with the largest number of mis-classified images;
- Cluster the images in the selected model into two sub-classes;
- Structure learning for the two new sub-classes;

end

Learning

- Hill climbing structure learning
  - Each pose in each HOI activity class
  - Add/remove an edge and check for optima
  - Keep tabu list to avoid revisiting solutions
  - Randomly initialize thrice to avoid local optimas

Learning

- Max-margin for parameter estimation
  - Maximize discrimination between different A
  - Each A has subclasses, hence multiple models and multiple weight vectors
  - Training sample: (x_i, $\boldsymbol{c_i}$, $\boldsymbol{y}(\boldsymbol{c_i})$) $\boldsymbol{y}$: maps $\boldsymbol{c_i}$ to class label
  - F: $\boldsymbol{y}(F(x\boldsymbol{i})) = \boldsymbol{y}(\boldsymbol{ci})$ $F(x_i) = \text{argmax}_r\{w_r \cdot x_i\}$ w_r: weights for r^{th} sub-class.

$$\min_{\mathbf{w},\xi} \frac{1}{2} \sum_r \|\mathbf{w}_r\|_2^2 + \beta \sum_i \xi_i$$

subject to: $\forall i,\ \xi_i \geq 0$

$$\forall i, r \text{ where } y(r) \neq y(c_i),\ \mathbf{w}_{c_i} \cdot \mathbf{x}_i - \mathbf{w}_r \cdot \mathbf{x}_i \geq 1 - \xi_i$$

Learning

- Overall Algorithm

Hill-climbing structure learning for each activity class.

foreach *Iteration* **do**

- Model parameter estimation by max-margin learning;
- Choose the model with the largest number of mis-classified images;
- Cluster the images in the selected model into two sub-classes;
- Structure learning for the two new sub-classes;

end

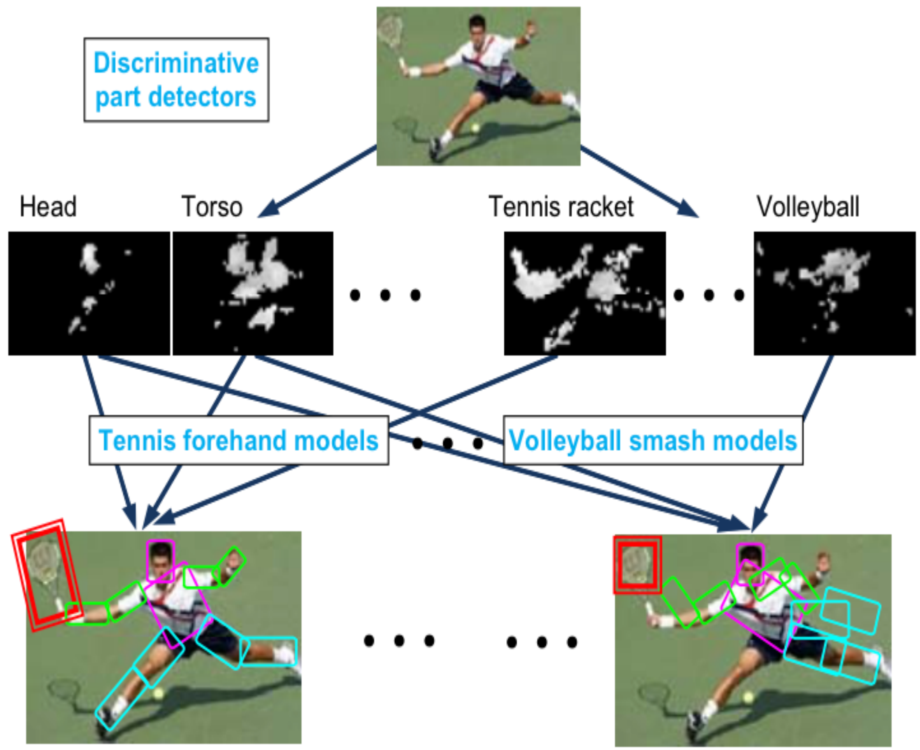
Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

Inference

- Given a test image(I), estimate pose and detect object and classify activity
  - To detect object (O) we maximize likelihood of the models given that object. Denoted as $\max_{\boldsymbol{O},\boldsymbol{H}} \Psi(\boldsymbol{A_k}, \boldsymbol{O}, \boldsymbol{H}, \mathrm{I})$
  - To detect human pose (H), compute $\max_{\boldsymbol{O},\boldsymbol{H}} \Psi(\boldsymbol{A_k}, \boldsymbol{O}, \boldsymbol{H}, \mathrm{I})$ for each A_k and select the one corresponding to the ML score

Inference



$$O_1, H_1 = \arg\max_{O,H} \Psi(A_1, O, H, I) \qquad O_K, H_K = \arg\max_{O,H} \Psi(A_K, O, H, I)$$

$$A^*, H^* = \arg\max_{k} \Psi(A_k, O_k, H_k, I)$$

Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results

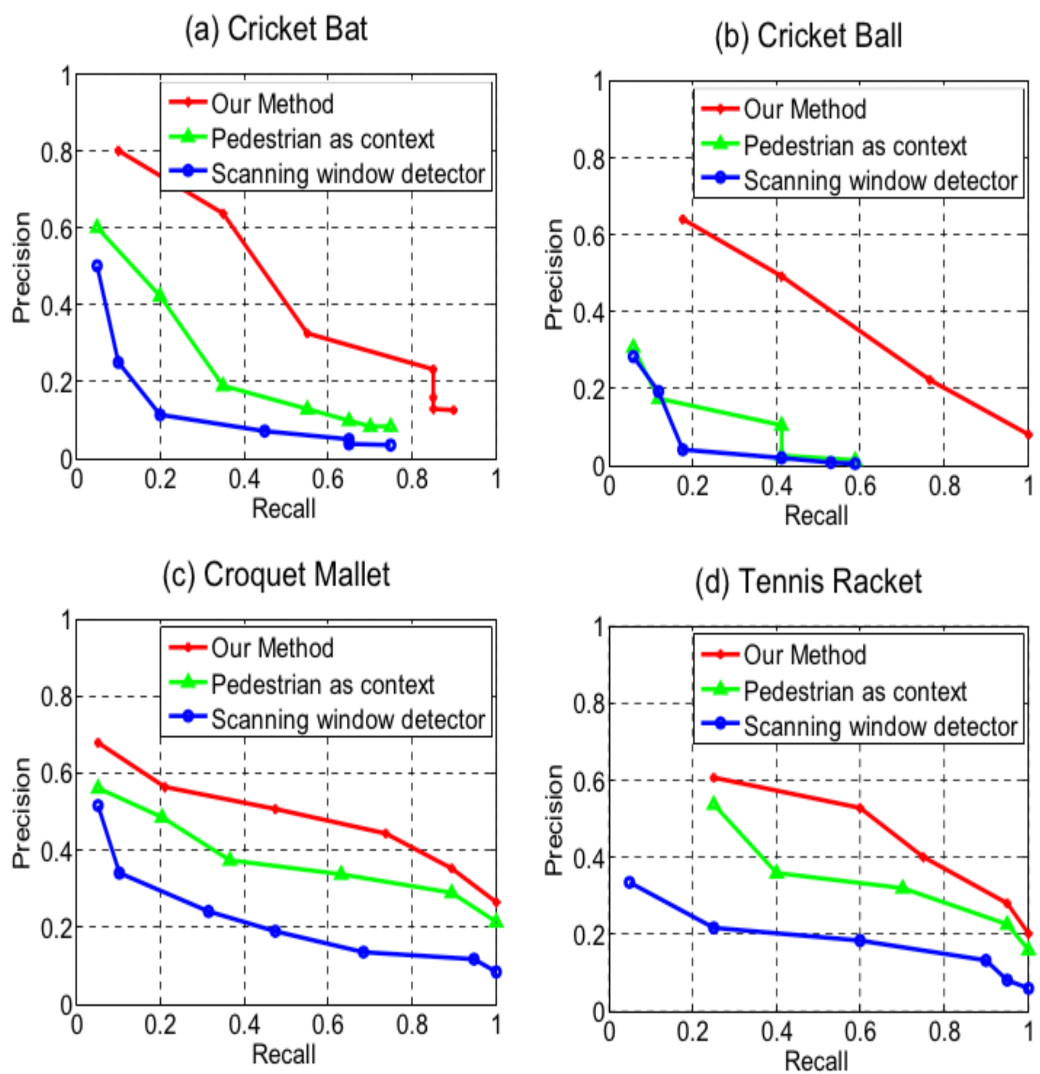
Results

Results

Results

- Object Detection
  - Compare with two experiments
    1. Sliding window as baseline
    2. Pedestrian detector for human’s location context

Results


(a) Cricket Bat
(b) Cricket Ball
(c) Croquet Mallet
(d) Tennis Racket
Our Method
Pedestrian as context
Scanning window detector
Precision
Recall
0
0.2
0.4
0.6
0.8
1

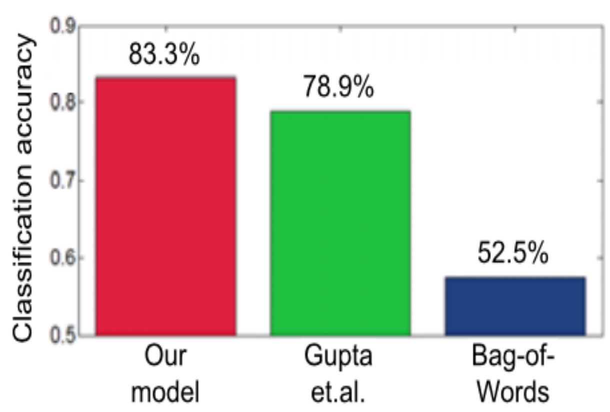
Results

- Pose Estimation

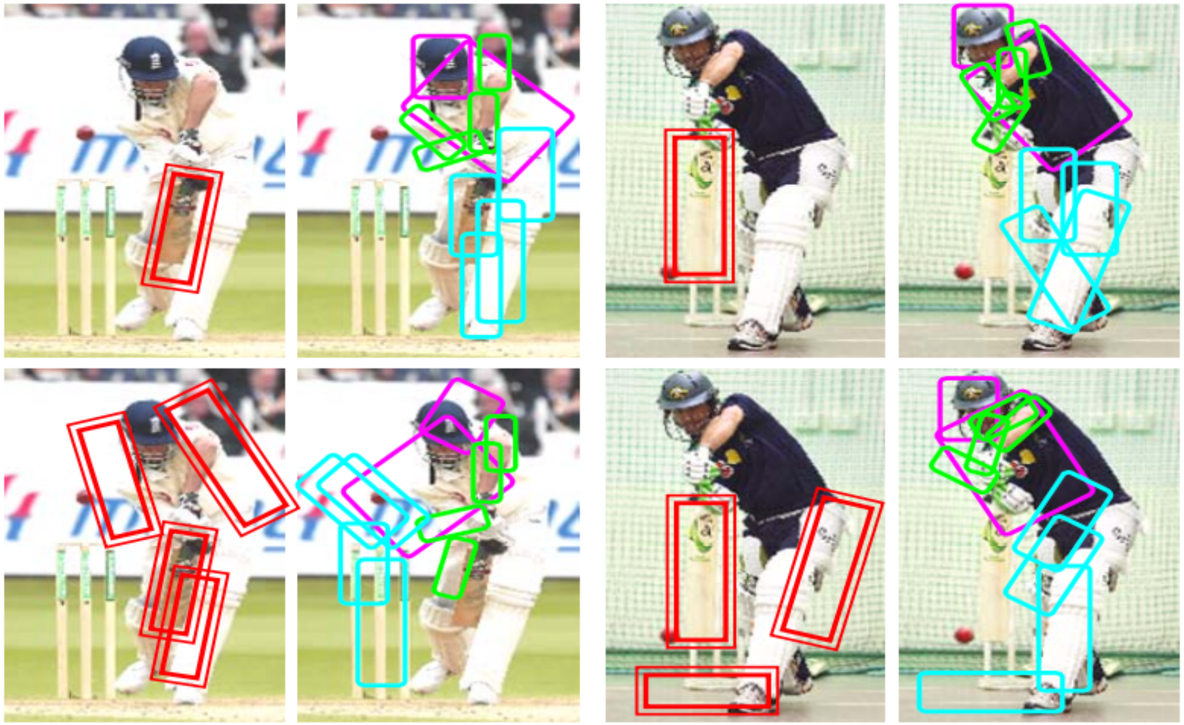
Method	Torso	Upper Leg	Lower Leg	Upper Arm	Fore Arm	Head
Iterative parsing [26]	52±19	22±11 22±10	21±9 28±16	24±16 28±17	17±11 14±10	42±18
Pictorial structure [1]	50±14	31±12 30±9	31±15 27±18	18±6 19±9	11±8 11±7	45±8
Class-based pictorial structure	59±9	36±11 26±17	39±9 27±9	30±12 31±12	13±6 18±14	46±11
Our model, only one pose per class	63±5	40±8 36±15	41±10 31±9	38±13 35±10	21±12 23±14	52±8
Our full model	**66±6**	**43±8 39±14**	**44±10 34±10**	**44±9 40±13**	**27±16 29±13**	**58±11**

Results

- HOI classification
  - Compare with SVM with BoW
  - Compare with Gupta et. al.



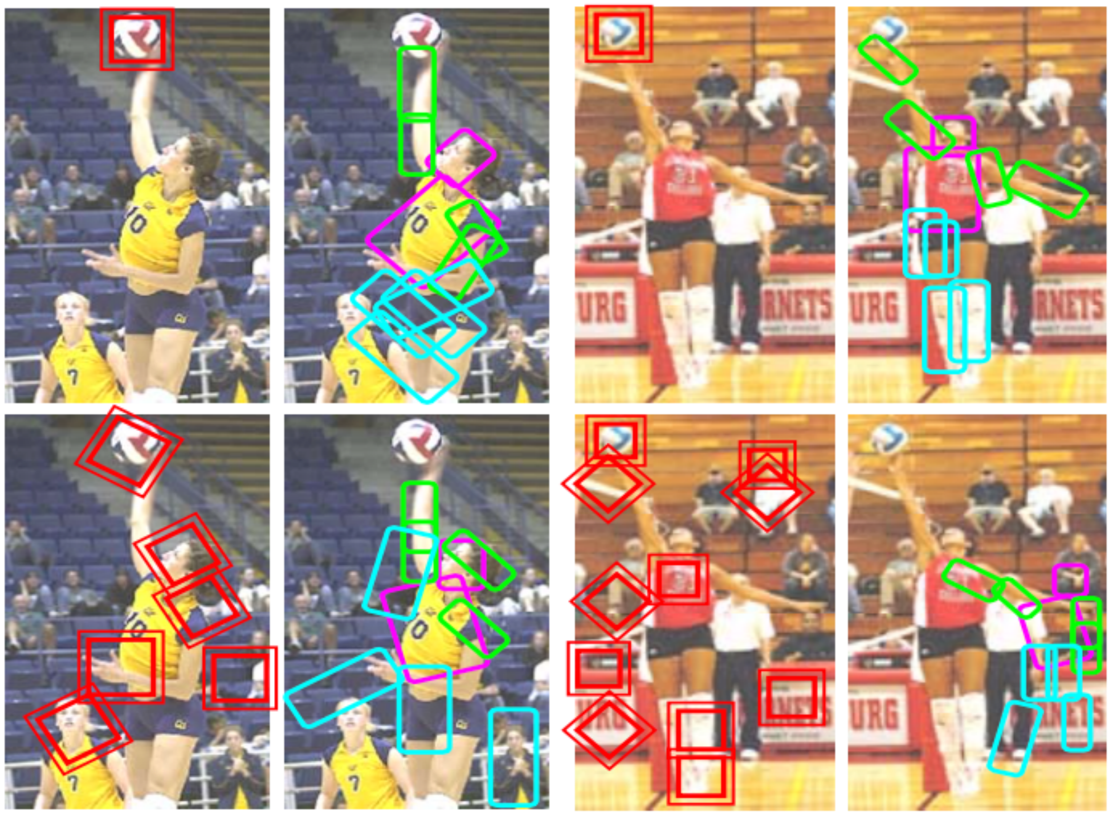
Results



(a) cricket defensive shot (b) cricket defensive shot

- Upper-left → object detection by mutual context
- Lower-left → object detection by a scanning window
- Upper-right → pose estimation by mutual context
- Lower-right → pose estimation by the state-of-the-art pictorial structure method

Results



(g) volleyball smash (h) volleyball smash

- Upper-left → object detection by mutual context
- Lower-left → object detection by a scanning window
- Upper-right → pose estimation by mutual context
- Lower-right → pose estimation by the state-of-the-art pictorial structure method

Thank you!