Action Recognition with Exemplar Based 2.5D Graph Matching

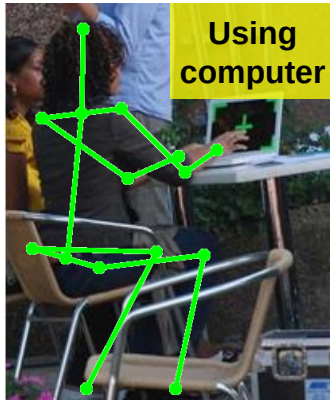
Bangpeng Yao and Li Fei-Fei

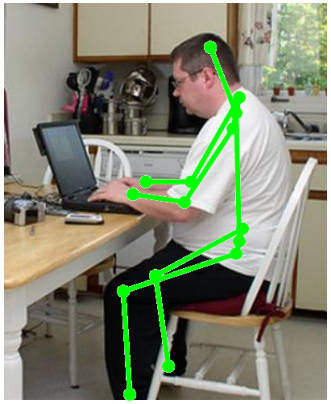
Computer Science Department

Stanford University

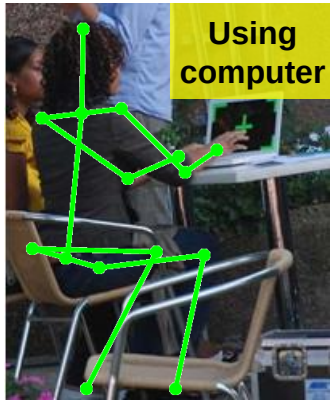
Action Recognition in Still Images

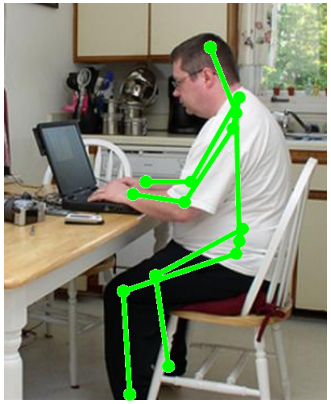
Action Recognition in Still Images





Action Recognition in Still Images





3D Recognition?

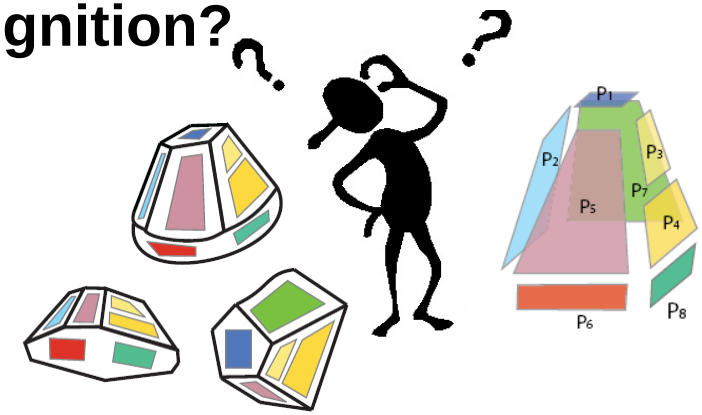
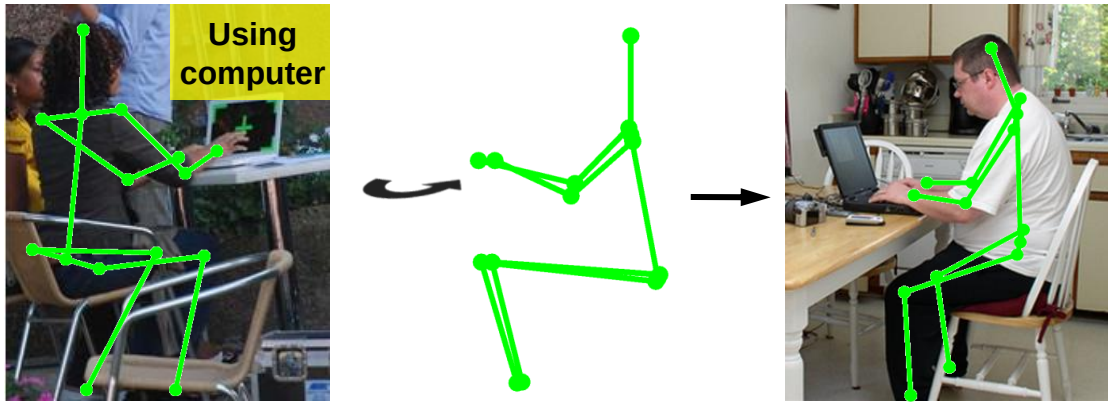
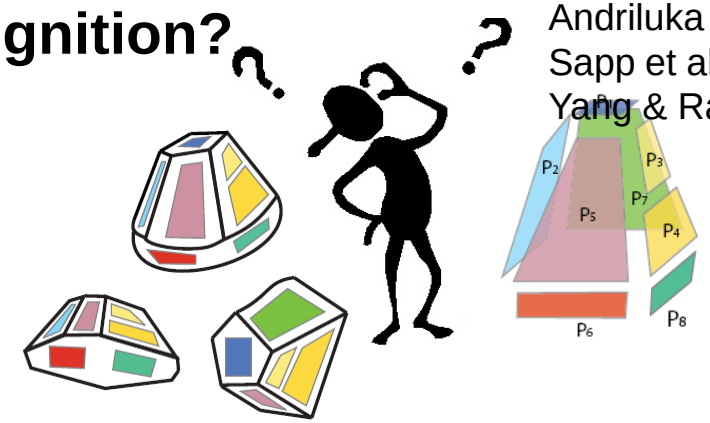


Figure credit: Savarese & Fei-Fei, 2010

Action Recognition in Still Images



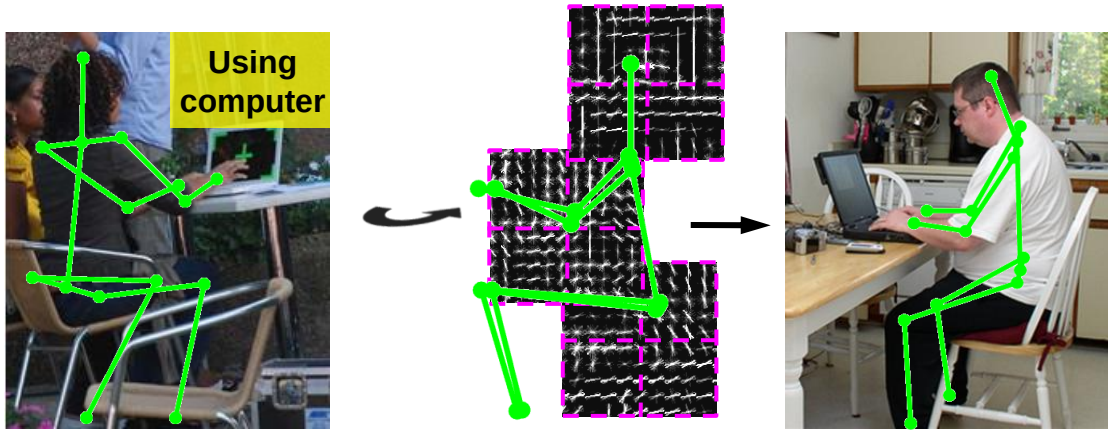
3D Recognition?



Felzenszwalb & Huttenlocher, 2005
Andriluka et al, 2009
Sapp et al, 2010
Yang & Ramanan, 2011

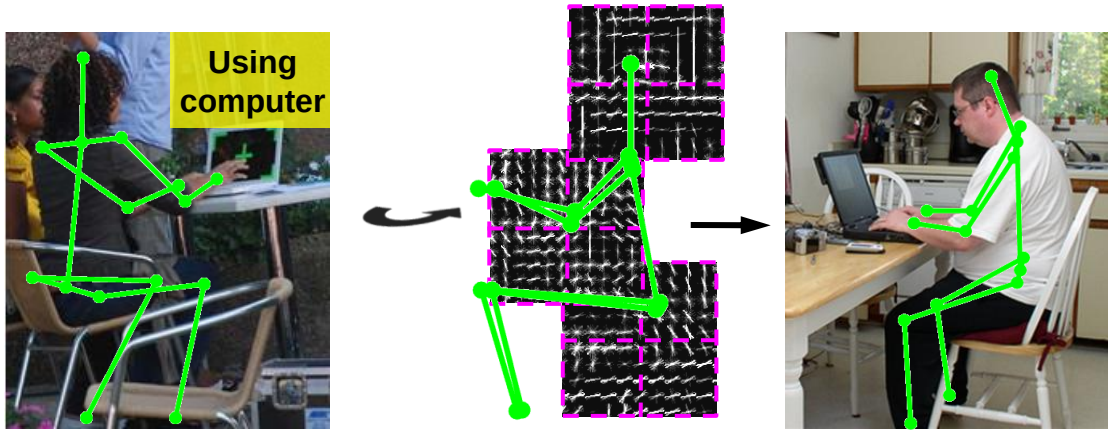
Figure credit: Savarese & Fei-Fei, 2010

Action Recognition in Still Images



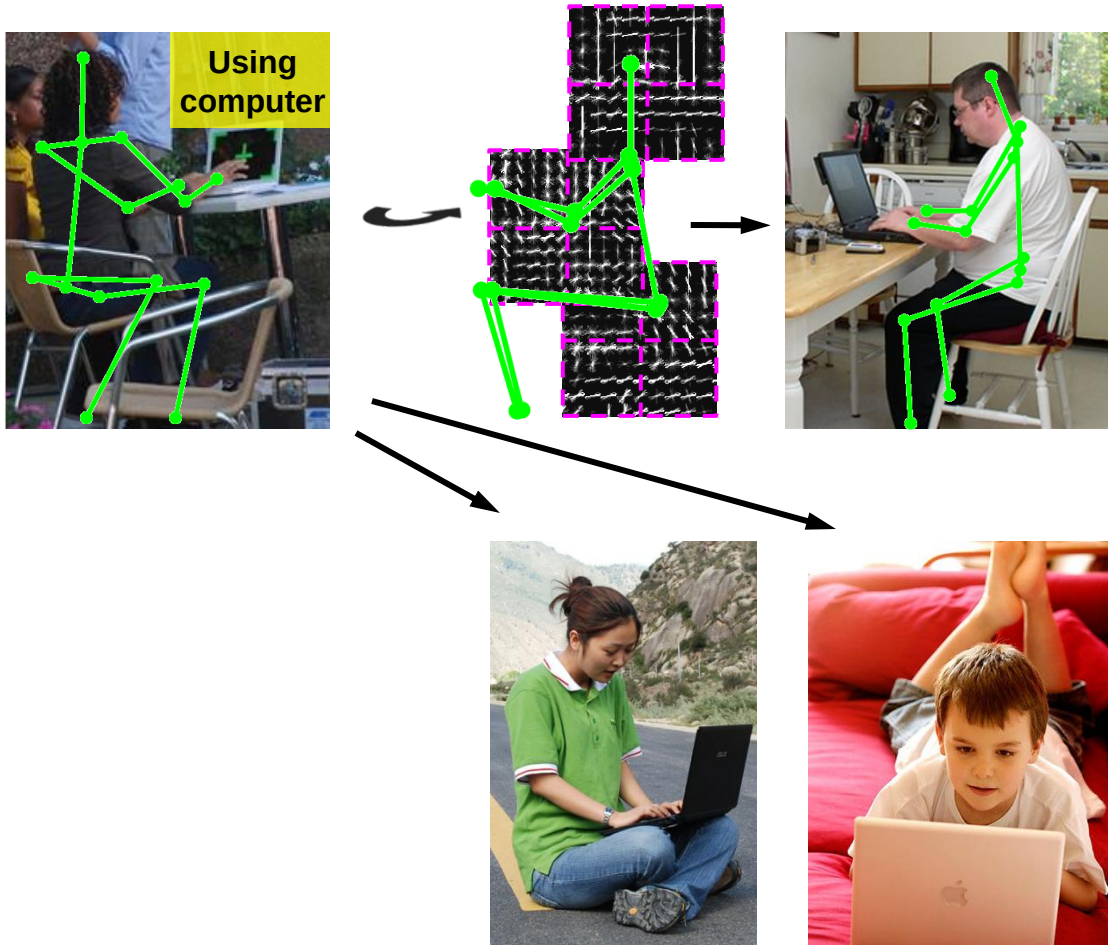
Lazebnik et al, 2006
Want et al, 2010
Delaitre et al., 2010
Yao et al., 2011

Action Recognition in Still Images



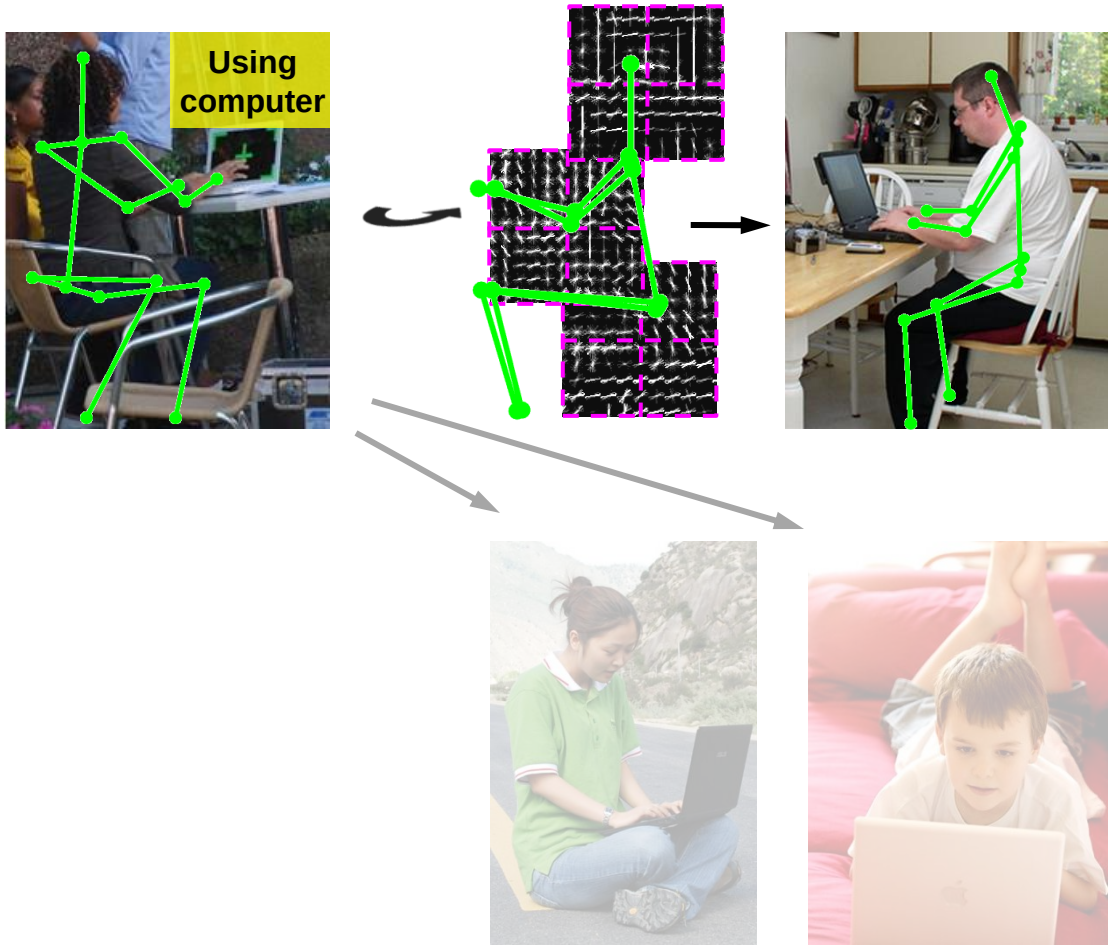
- 2.5D Representation
  - View independent 3D pose
  - Rich 2D appearance features

Action Recognition in Still Images



- 2.5D Representation
  - View independent 3D pose
  - Rich 2D appearance features

Action Recognition in Still Images



- 2.5D Representation
  - View independent 3D pose
  - Rich 2D appearance features

- Exemplar based recognition
  - Only consider nearest image

Outline

- 2.5D Representation of Human Actions
- Exemplar Based 2.5D Graph Matching
- Dataset & Experiments
- Conclusion

Outline

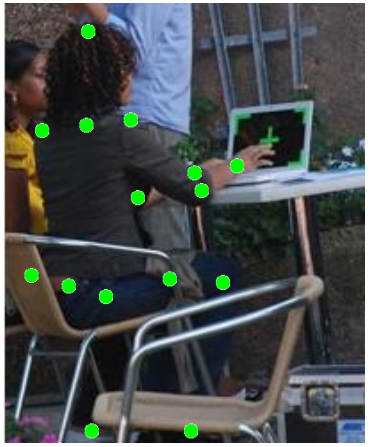
- **2.5D Representation of Human Actions**
- Exemplar Based 2.5D Graph Matching
- Dataset & Experiments
- Conclusion

2.5D Representation of Actions

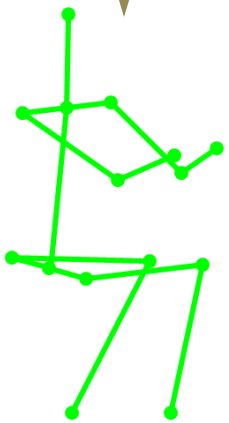
Original image

Original image

2.5D Representation of Actions



Original image



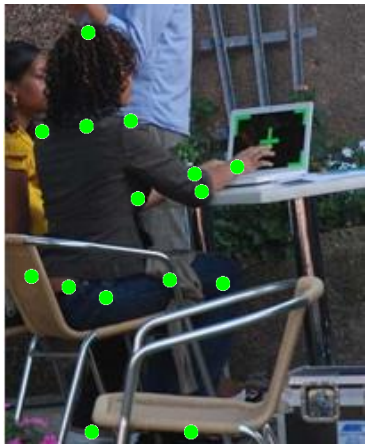
2D skeleton

Original image

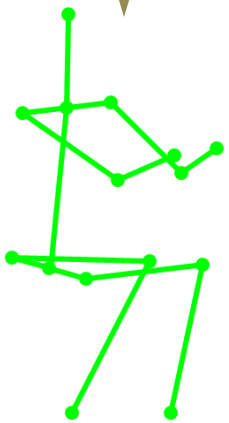
↓

2D key points

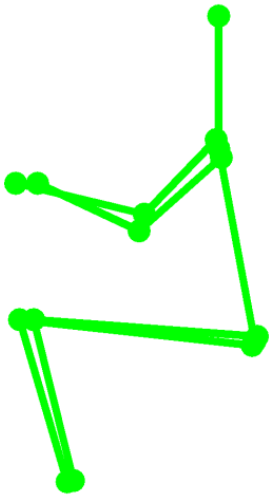
2.5D Representation of Actions



Original image



2D skeleton



3D skeleton

Original image

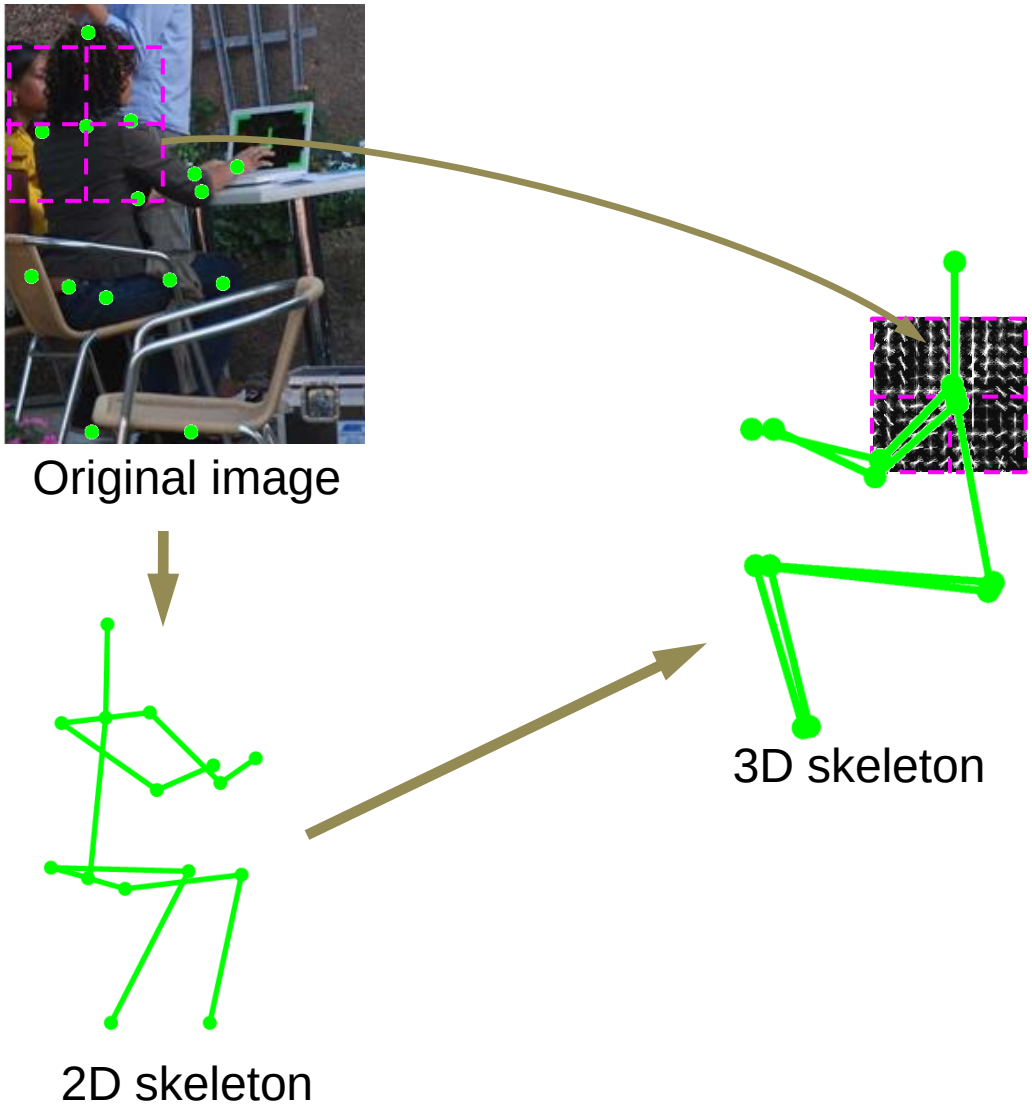
↓

2D key points

↓

3D key points

2.5D Representation of Actions



Original image

↓

2D key points

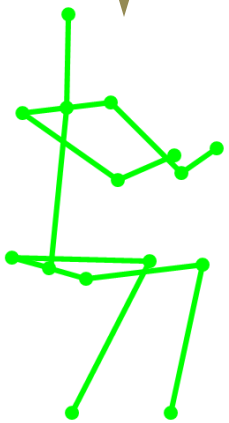
↓ ↓

3D key points

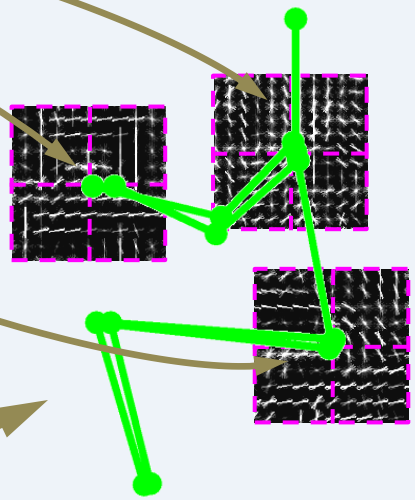
Appea -rance

2.5D Representation of Actions

Original image



2D skeleton



3D skeleton

2.5D Graph

Original image

↓

2D key points

↓ ↓

3D key points

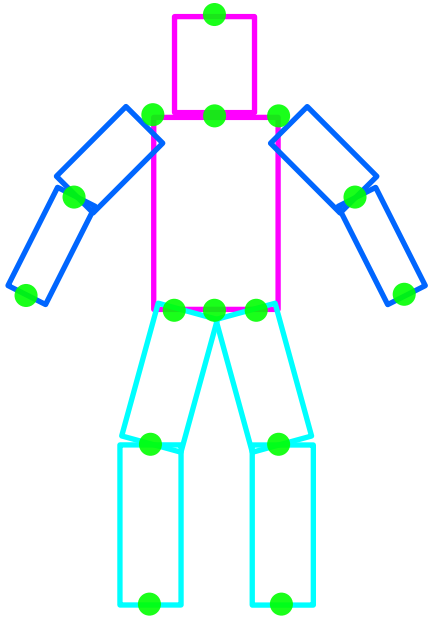
Appea-rance

↓ ↓

2.5D Graph

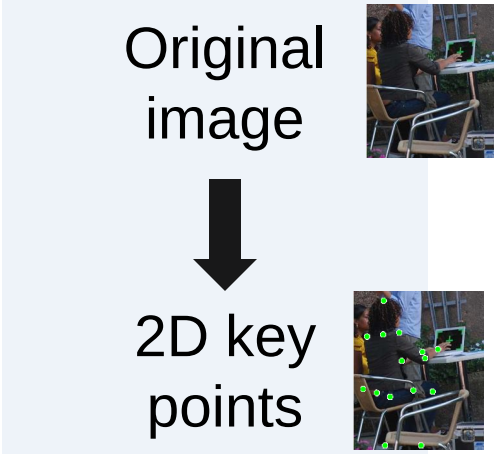
Estimating 2D Key Points

Pictorial Structure:



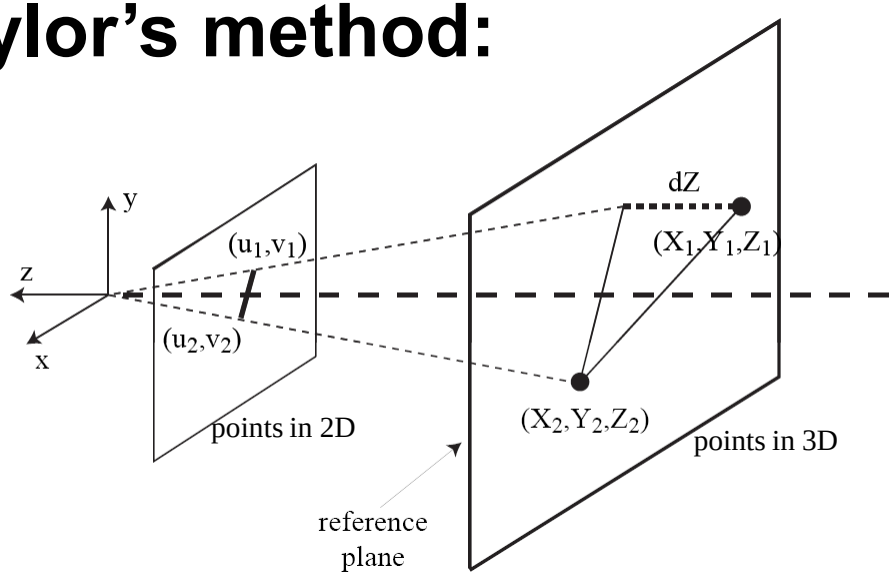
Not 100%?
See experiment …

Felzenszwalb & Hunttenlocher, 2005
Sapp et al, 2010



Converting 2D Key Points to 3D

Taylor's method:



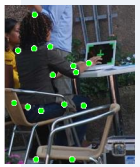
$$l^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2$$

$$dZ = \pm\sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)/s^2}$$

Resolving the "±" problem:

- Configuration constraints (Lee & Chen, 1985)
- Regression based refinement

Taylor, 2000



Original image → 2D key points → 3D key points → 2.5D Graph

Appea-rance → 2.5D Graph

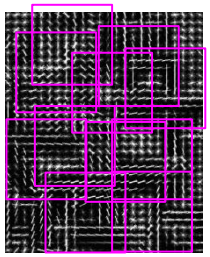
Outline

- 2.5D Representation of Human Actions
- **Exemplar Based 2.5D Graph Matching**
- Dataset & Experiments
- Conclusion

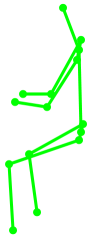
Matching Two 2.5D Graphs

$$Sim(G_1, G_2) = Sim(\mathrm{A}(G_1, G_2)) + \lambda \cdot Sim(\mathrm{S}(G_1, G_2))$$

G_1



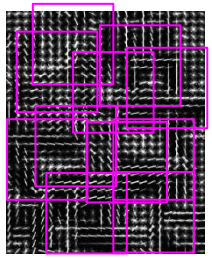
$\mathrm{A}(G_1)$



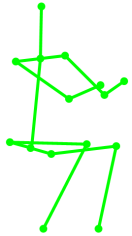
$\mathrm{S}(G_1)$

$$\mathbf{P}' = \begin{bmatrix} x_1' & x_2' & \cdots \\ y_1' & y_2' & \cdots \\ z_1' & z_2' & \cdots \end{bmatrix}$$

G_2



$\mathrm{A}(G_2)$



$\mathrm{S}(G_2)$

$$\mathbf{R}^*\mathbf{P}$$

$$\mathbf{P} = \begin{bmatrix} x_1 & x_2 & \cdots \\ y_1 & y_2 & \cdots \\ z_1 & z_2 & \cdots \end{bmatrix}$$

Original image

2D key points

3D key points

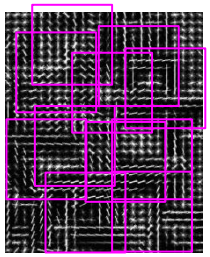
Appea-rance

2.5D Graph

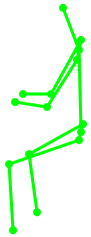
Matching Two 2.5D Graphs

$$Sim(G_1, G_2) = Sim(\mathrm{A}(G_1, G_2)) + \lambda \cdot Sim(\mathrm{S}(G_1, G_2))$$

G_1



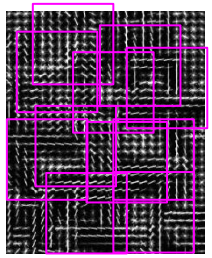
$\mathrm{A}(G_1)$



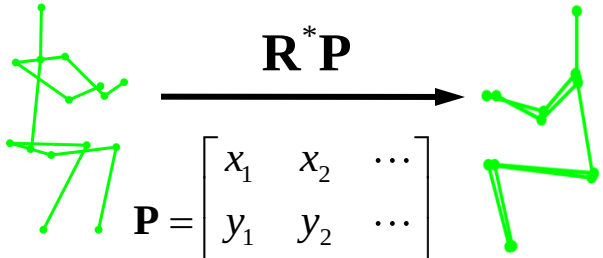
$\mathrm{S}(G_1)$

$$\mathbf{P}' = \begin{bmatrix} x_1' & x_2' & \cdots \\ y_1' & y_2' & \cdots \\ z_1' & z_2' & \cdots \end{bmatrix}$$

G_2



$\mathrm{A}(G_2)$



$\mathrm{S}(G_2)$

$$\mathbf{P} = \begin{bmatrix} x_1 & x_2 & \cdots \\ y_1 & y_2 & \cdots \\ z_1 & z_2 & \cdots \end{bmatrix}$$

$$\mathbf{R}^* = \arg\min_{\mathbf{R}} \left\| \mathbf{P}' - \mathbf{R}\mathbf{P} \right\|^2$$

Umeyama, 1991

Original image

2D key points

3D key points

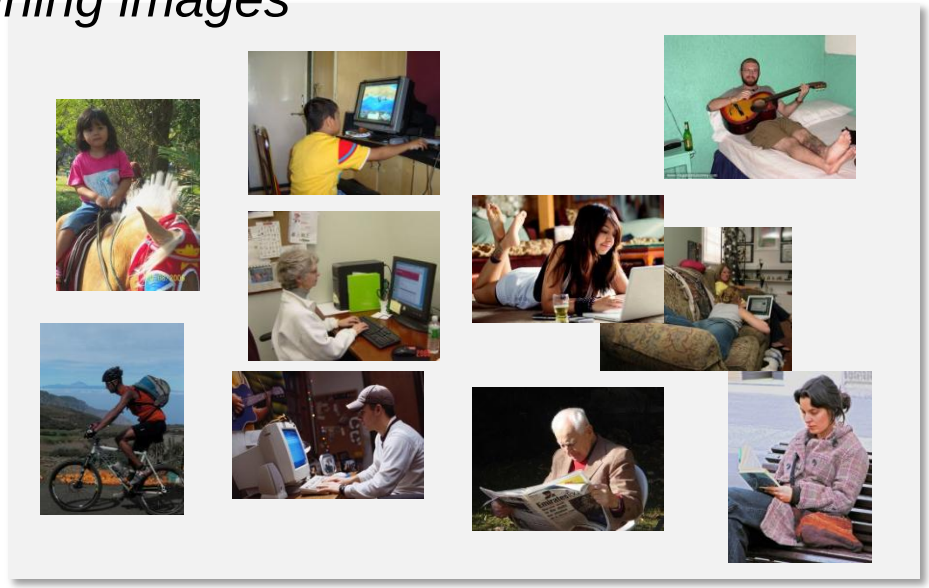
Appea-rance

2.5D Graph

Exemplar Based Action Recognition

Test image

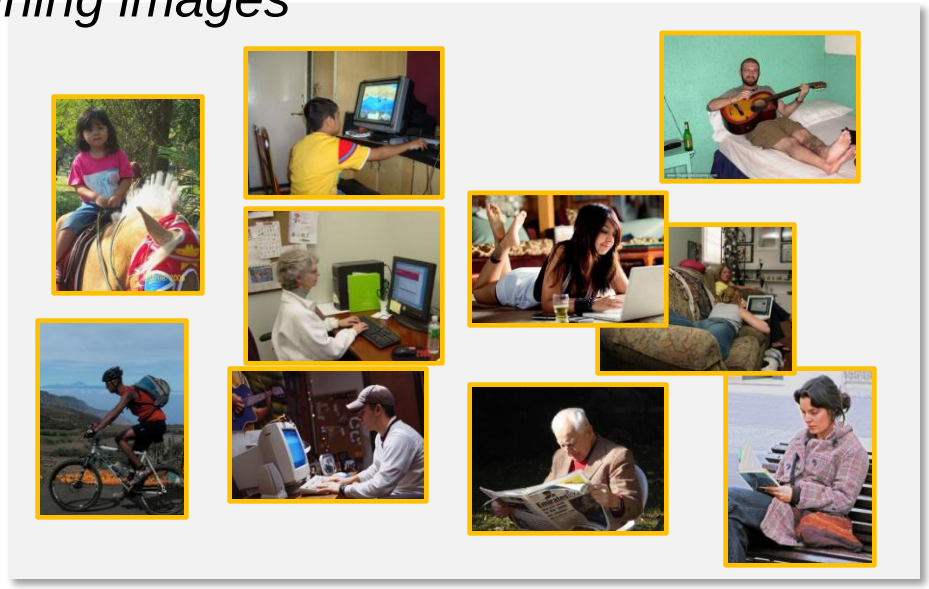
Training images



Exemplar Based Action Recognition

Test image

Training images

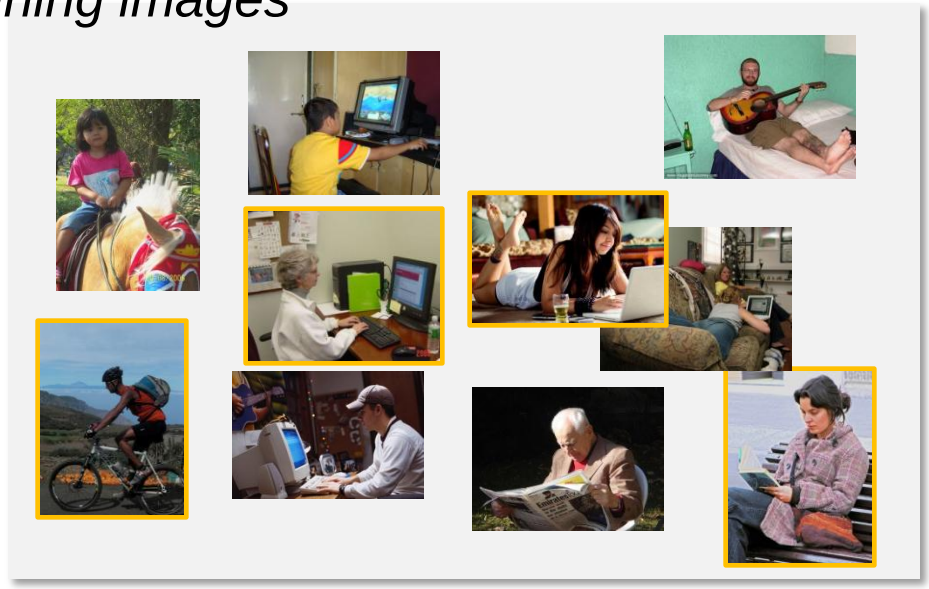


- Test image vs. each candidate image.
  - **Time-consuming**

Exemplar Based Action Recognition

Test image

Training images

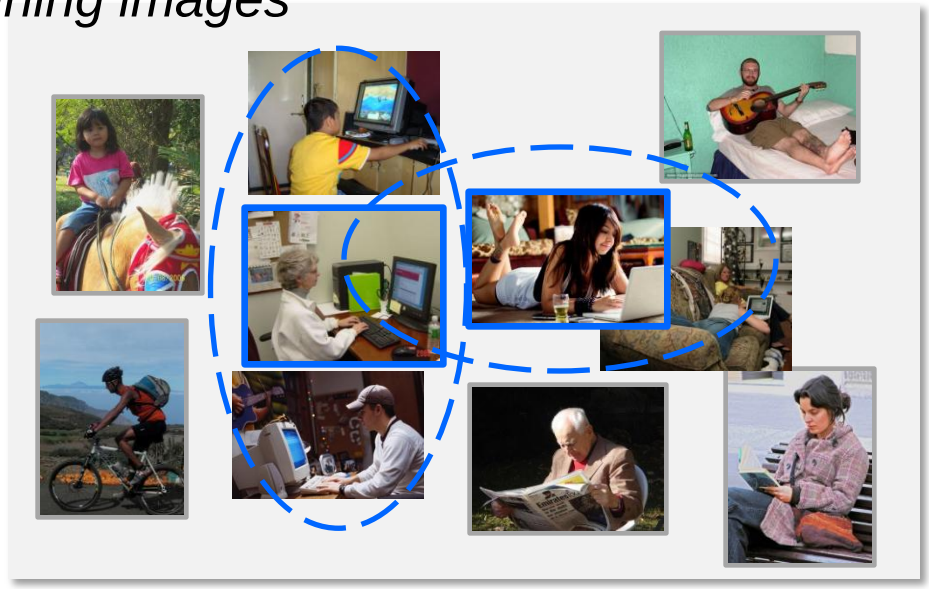


- Our approach: Test image vs. a subset of training images.

Exemplar Based Action Recognition

Test image

Training images

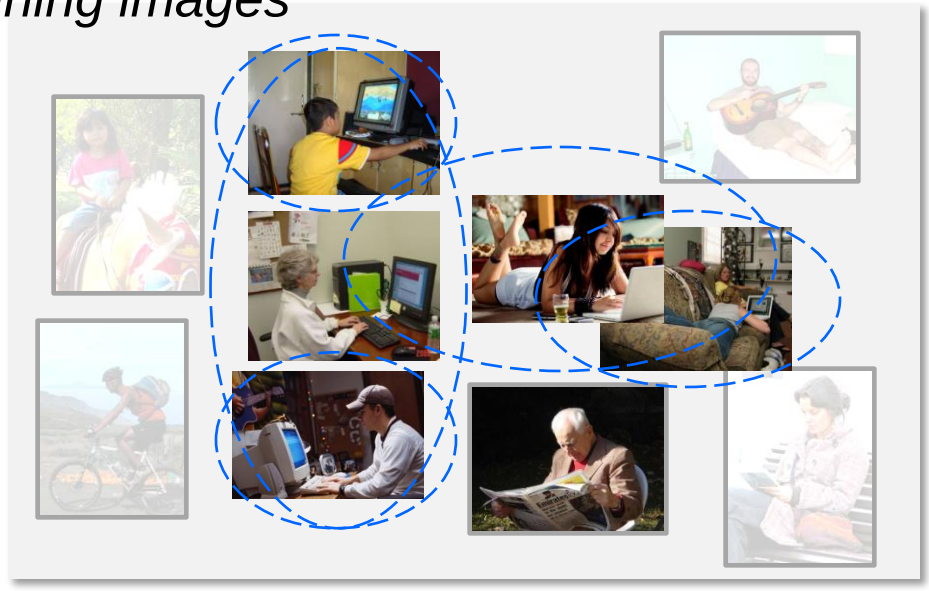


- Our approach: Test image vs. a subset of training images.
  - The smallest set of images that can recognize all within-class images in the exemplar based setting.
  - Dominating images.

Finding Dominating Images

Training images

Test image



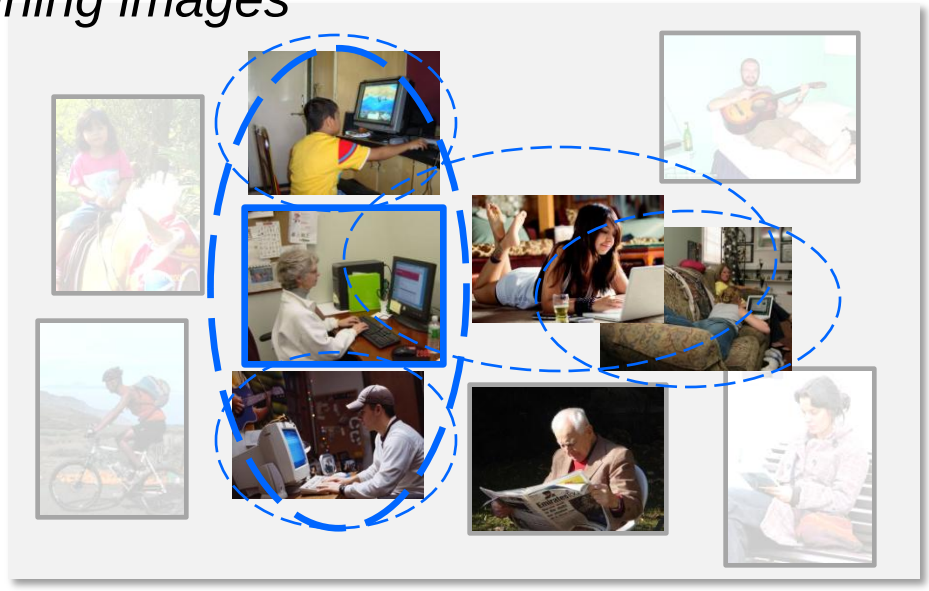
- Our approach: Test image vs. a subset of training images.

- An iterative approach to find dominating images:
  1. **Maximize Coverage(I) for each image I;**

Finding Dominating Images

Test image

Training images



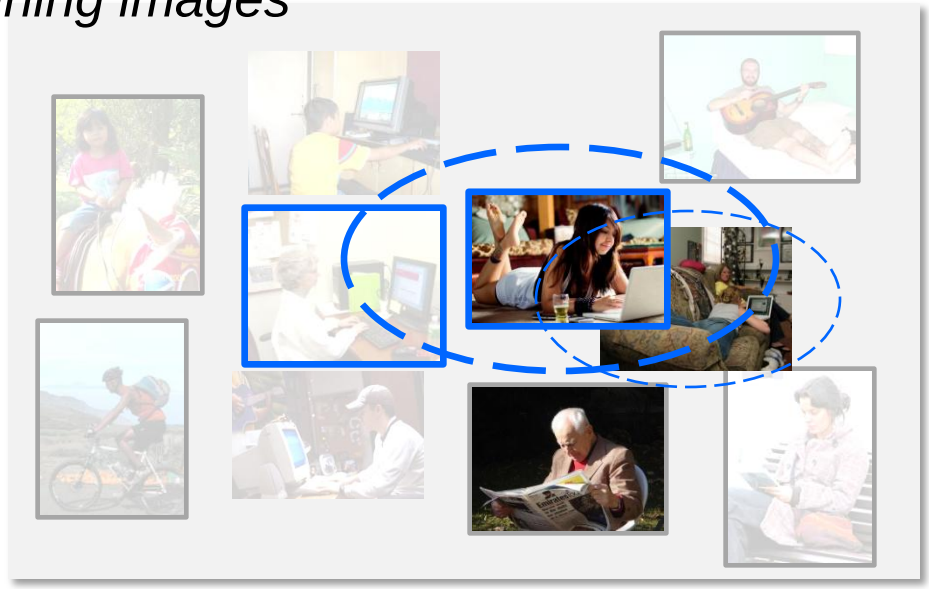
- Our approach: Test image vs. a subset of training images.

- An iterative approach to find dominating images:
  1. Maximize Coverage(I) for each image I;
  2. **Find I^* that maximizes Coverage(I^*).**

Finding Dominating Images

Test image

Training images



- Our approach: Test image vs. a subset of training images.

- An iterative approach to find dominating images:
  1. Maximize Coverage(I) for each image I;
  2. Find I^* that maximizes Coverage(I^*).
  3. **Remove I^* and Coverage(I^*), return to step 2.**

Finding Dominating Images

Test image

Training images

- Our approach: Test image vs. a subset of training images.

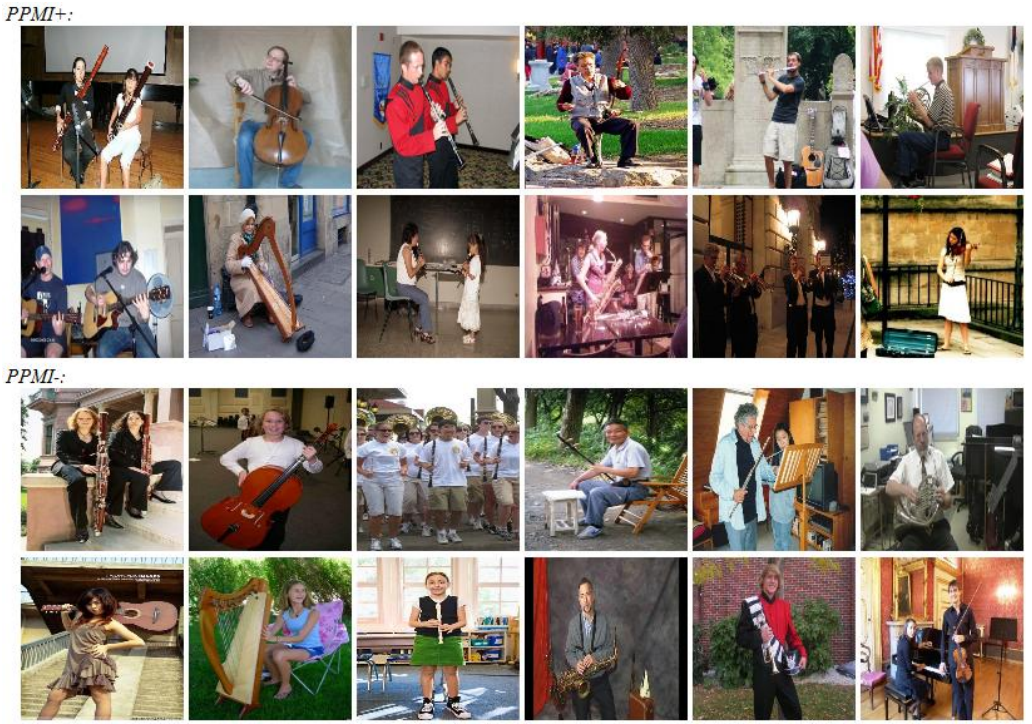
- An iterative approach to find dominating images:
  1. Maximize Coverage(I) for each image I;
  2. Find I* that maximizes Coverage(I*).
  3. Remove I* and Coverage(I*), return to step 2.

Outline

- 2.5D Representation of Human Actions
- Exemplar Based 2.5D Graph Matching
- **Dataset & Experiments**
- Conclusion

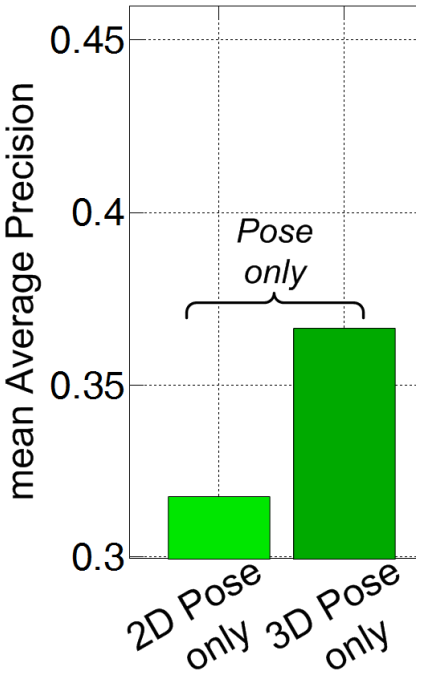
PPMI: Dataset

- PPMI: People Playing (Interacting with) Musical Instruments
- 24 classes of people interaction with different instruments.
  - 100 training & 100 testing for each class.

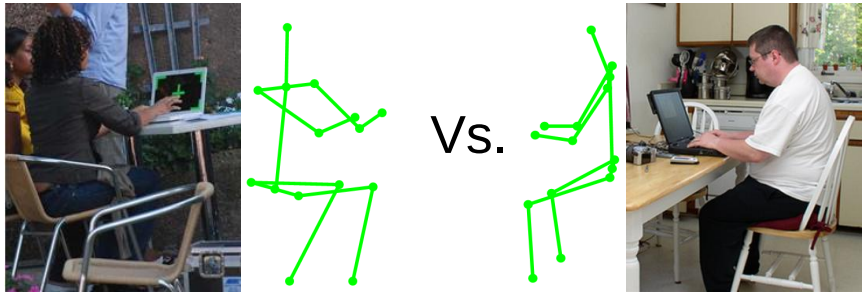


Yao & Fei-Fei, 2010

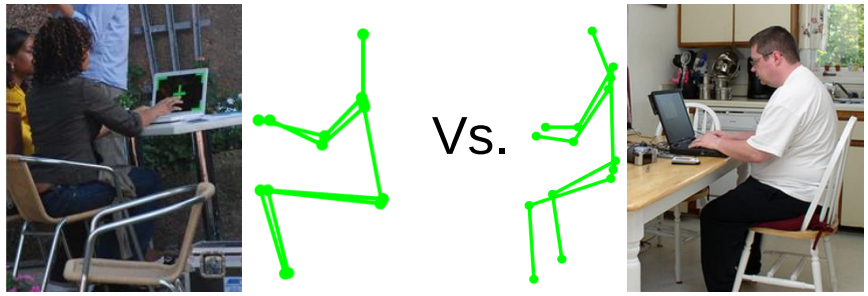
PPMI: Results



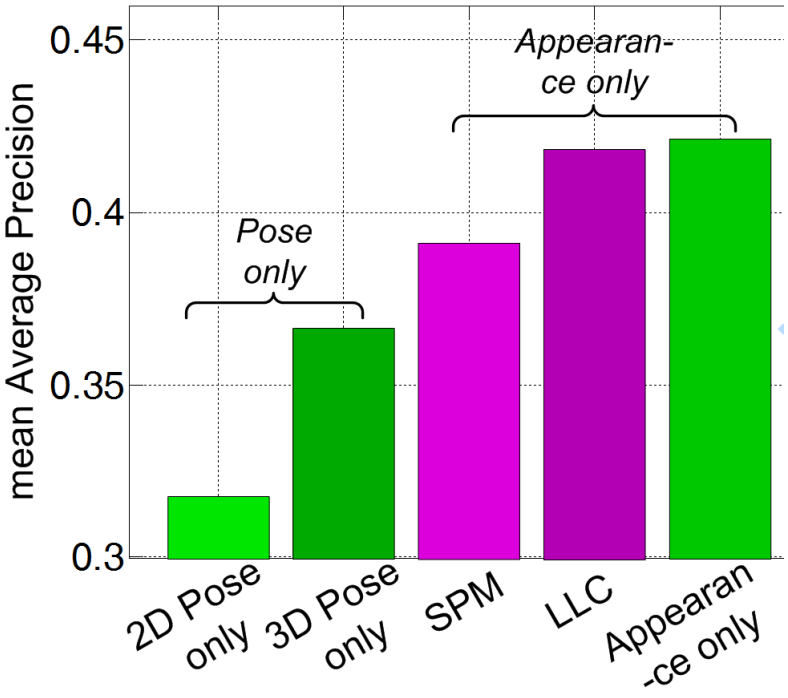
2D Pose only:



3D Pose only:

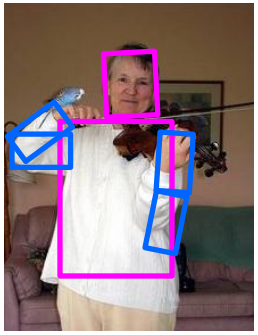


PPMI: Results



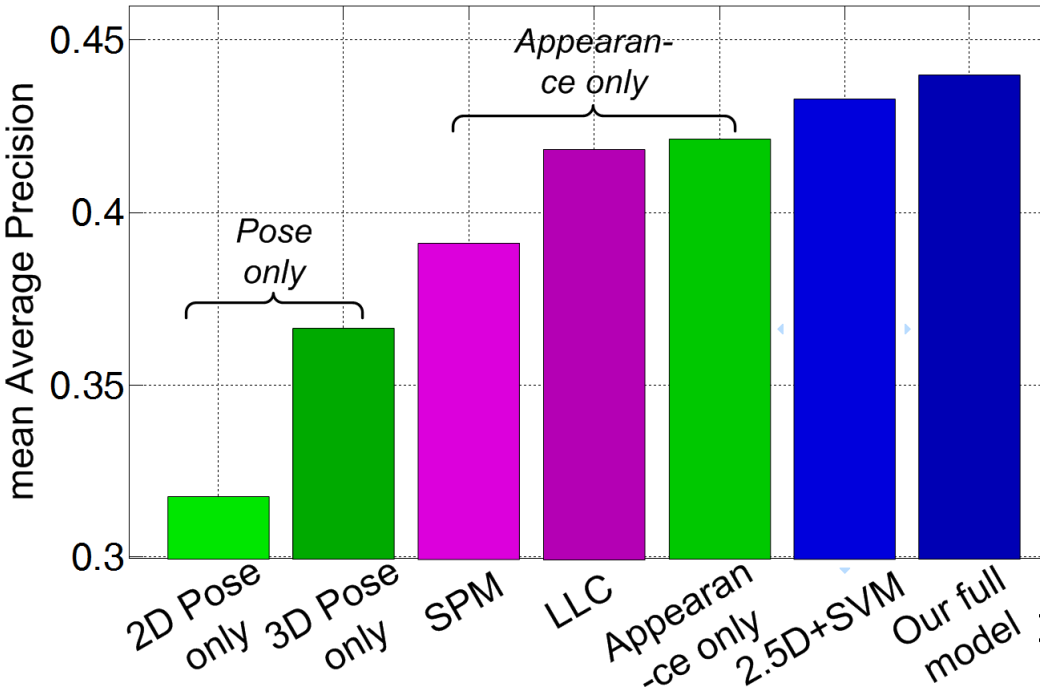
Lazebnik et al, 2006
Wang et al, 2010

- Rich appearance info.
- Pose can be wrong

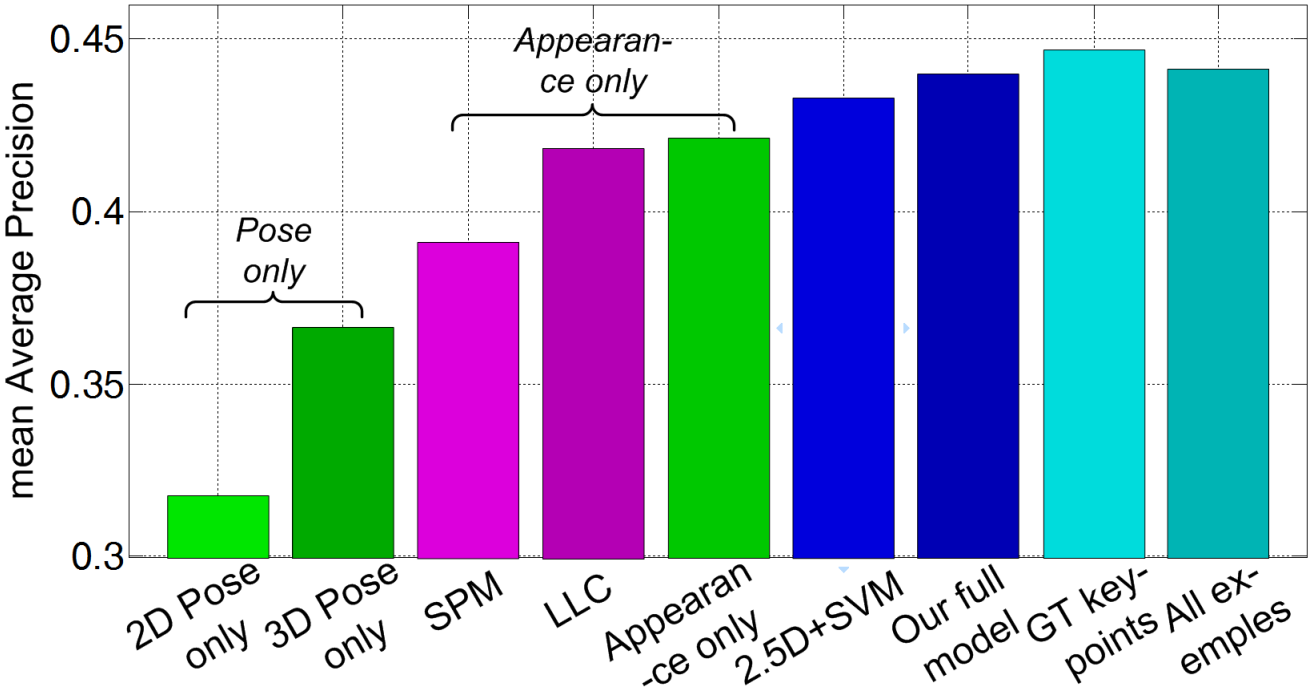


- Pose can be similar

PPMI: Results



PPMI: Results



PPMI: Dominating Images

PPMI: Dominating Images

PASCAL 2011: Dataset

Ten Action Classes + "Other"

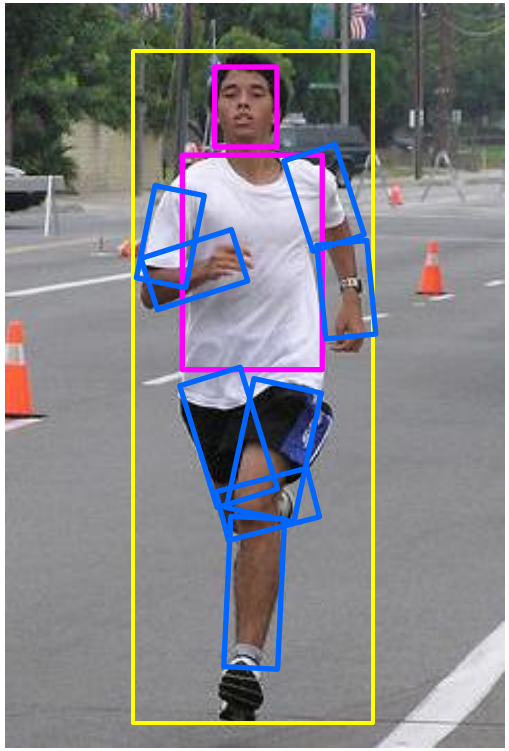
Figure credit: Everingham

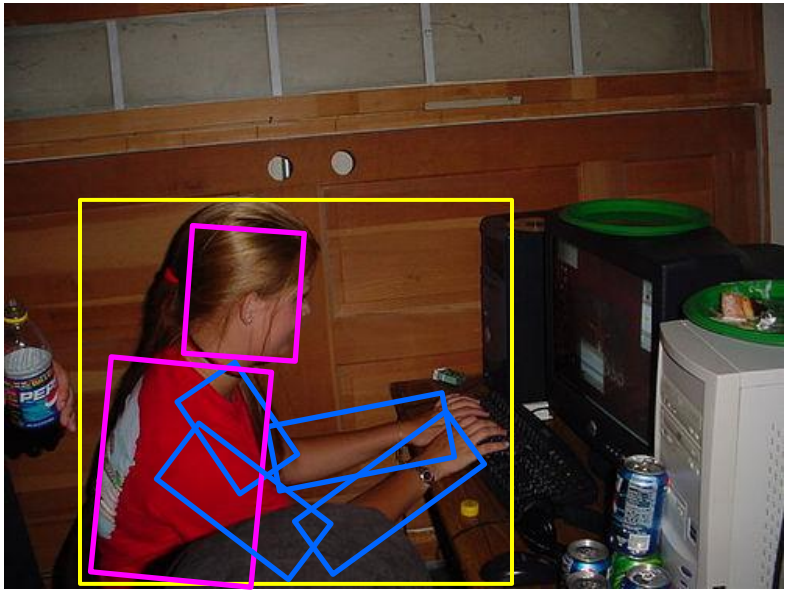
- 3000 training & 3000 testing images each class

Everingham et al, 2011

PASCAL 2011: Pose Estimation

- Human bounding box provided.
- Two PS models: Full body and upper body.





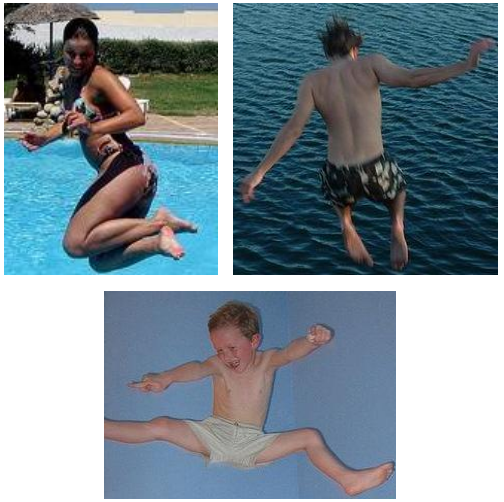
PASCAL 2011: Results

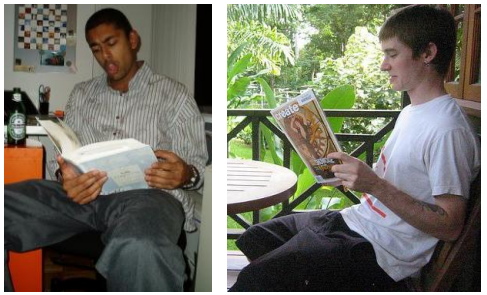
Action	HOBJ DSAL	CONT EXT	RF_S VM	POSE LETS	ATTR PART	Our Method		
						Pose	App.	Full
Jump	71.6	65.9	66.0	59.5	66.7	64.6	68.9	**72.4**
Phone	**50.7**	41.5	41.0	31.3	41.1	41.2	44.5	48.3
Play instrumnt	77.5	57.4	60.0	45.6	60.8	68.3	72.9	**77.7**
Read	37.8	34.7	41.5	27.8	42.2	36.0	39.2	**43.2**
Ride bike	86.5	88.8	90.0	84.4	**90.5**	81.4	86.6	89.0
Ride horse	89.5	90.2	92.1	88.3	**92.2**	80.4	87.1	90.0
Run	83.8	**87.9**	86.6	77.6	86.2	79.4	83.0	86.8
Take photo	25.1	25.7	28.8	**31.0**	28.8	21.6	25.1	27.9
Use computer	58.9	54.5	62.0	47.4	**63.5**	51.5	56.9	60.5
Walk	59.2	59.5	**65.9**	57.6	64.2	52.8	59.7	62.1
Mean	64.1	60.6	63.4	55.1	63.6	57.7	62.4	**65.8**

HOBJ_DSAL: Prest et al, 2011
RF_SVM: Yao et al, 2011a
POSELETS: Maji et al, 2011
ATTR_PART: Yao et al, 2011b

PASCAL 2011: Results

Action	HOBJ DSAL	CONT EXT	RF_S VM	POSE LETS	ATTR PART	Our Method		
						Pose	App.	Full
Jump	71.6	65.9	66.0	59.5	66.7	64.6	68.9	**72.4**
Phone	**50.7**	41.5	41.0	31.3	41.1	41.2	44.5	48.3
Play instrumnt	77.5	57.4	60.0	45.6	60.8	68.3	72.9	**77.7**
Read	37.8	34.7	41.5	27.8	42.2	36.0	39.2	**43.2**
Ride bike	86.5	88.8	90.0	84.4	**90.5**	81.4	86.6	89.0
Ride horse	89.5	90.2	92.1	88.3	**92.2**	80.4	87.1	90.0
Run	83.8	**87.9**	86.6	77.6	86.2	79.4	83.0	86.8
Take photo	25.1	25.7	28.8	**31.0**	28.8	21.6	25.1	27.9
Use computer	58.9	54.5	62.0	47.4	**63.5**	51.5	56.9	60.5
Walk	59.2	59.5	**65.9**	57.6	64.2	52.8	59.7	62.1
Mean	64.1	60.6	63.4	55.1	63.6	57.7	62.4	**65.8**

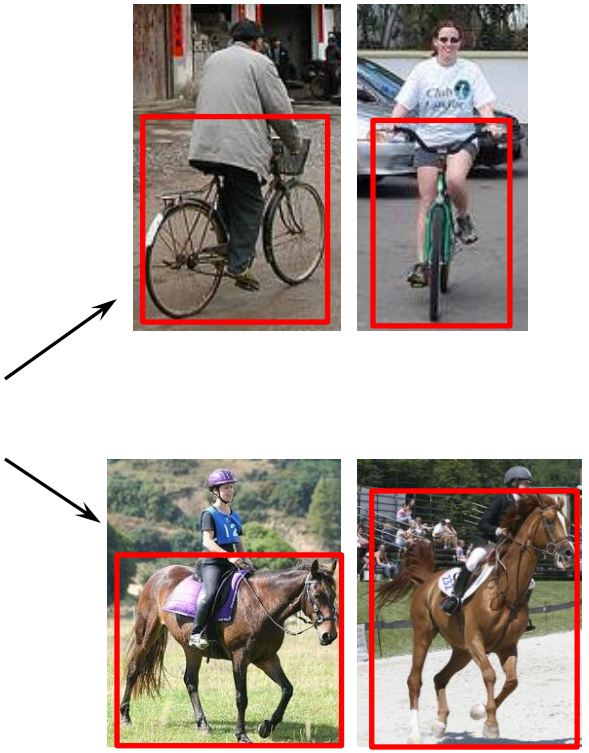




HOBJ_DSAL: Prest et al, 2011
RF_SVM: Yao et al, 2011a
POSELETS: Maji et al, 2011
ATTR_PART: Yao et al, 2011b

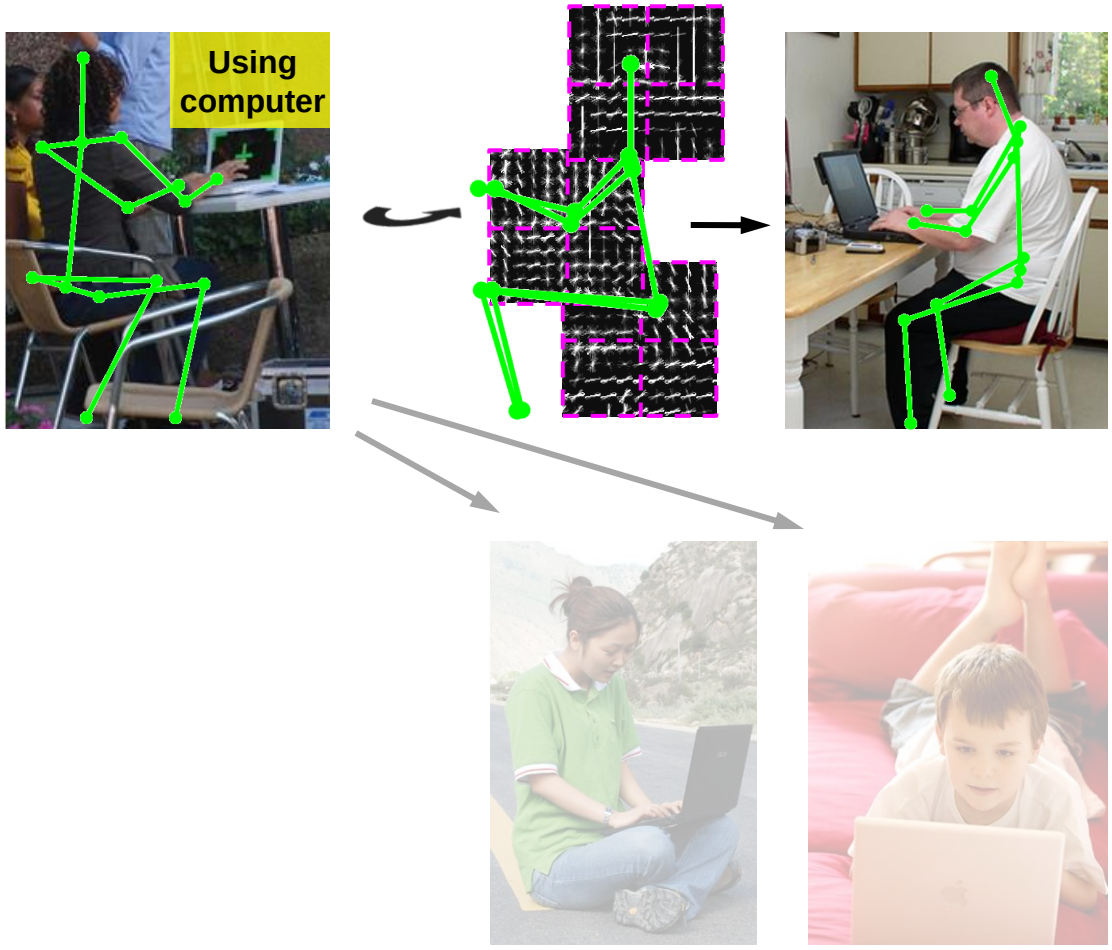
PASCAL 2011: Results

Action	HOBJ DSAL	CONT EXT	RF_S VM	POSE LETS	ATTR PART	Our Method		
						Pose	App.	Full
Jump	71.6	65.9	66.0	59.5	66.7	64.6	68.9	**72.4**
Phone	**50.7**	41.5	41.0	31.3	41.1	41.2	44.5	48.3
Play instrumnt	77.5	57.4	60.0	45.6	60.8	68.3	72.9	**77.7**
Read	37.8	34.7	41.5	27.8	42.2	36.0	39.2	**43.2**
Ride bike	86.5	88.8	90.0	84.4	**90.5**	81.4	86.6	89.0
Ride horse	89.5	90.2	92.1	88.3	**92.2**	80.4	87.1	90.0
Run	83.8	**87.9**	86.6	77.6	86.2	79.4	83.0	86.8
Take photo	25.1	25.7	28.8	**31.0**	28.8	21.6	25.1	27.9
Use computer	58.9	54.5	62.0	47.4	**63.5**	51.5	56.9	60.5
Walk	59.2	59.5	**65.9**	57.6	64.2	52.8	59.7	62.1
Mean	64.1	60.6	63.4	55.1	63.6	57.7	62.4	**65.8**



HOBJ_DSAL: Prest et al, 2011
RF_SVM: Yao et al, 2011a
POSELETS: Maji et al, 2011
ATTR_PART: Yao et al, 2011b

Conclusion



- 2.5D Representation
  - View independent 3D pose
  - Rich 2D appearance features

- Exemplar based recognition
  - Only consider nearest image

Acknowledgement