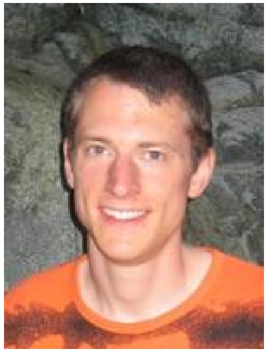
Multi-view Pictorial Structures for 3D Human Pose Estimation


Sikandar Amin[1]


Mykhaylo Andriluka[2]


Marcus Rohrbach[2]


Bernt Schiele[2]

[1]Intelligent Autonomous Systems Group
Technische Universität München (TUM)
Munich, Germany

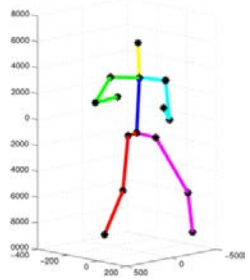
[2]Computer Vision and Multimodal Computing
Max Planck Institute for Informatics
Saarbrücken, Germany

Human Pose Estimation

State-of-the-art for 3D:

- 3D body models
  - ✘ Complex inference
- Tracking
  - ✘ Activity specific motion prior
- ✘ Controlled settings

Bregler et al. 2004
Sigal et al. 2004
Deutscher&Reid 2005
Andriluka et al. 2010
Bergholdt et al. 2010
Gall et al. 2010
Hasler et al., 2009
Taylor et al. 2010
Yao et al. 2011



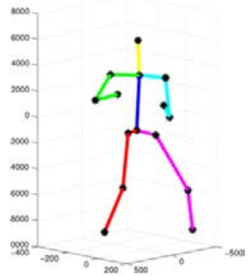
Human Pose Estimation

State-of-the-art for 3D:

- 3D body models
  - ✘ Complex inference
- Tracking
  - ✘ Activity specific motion prior

✘ Controlled settings

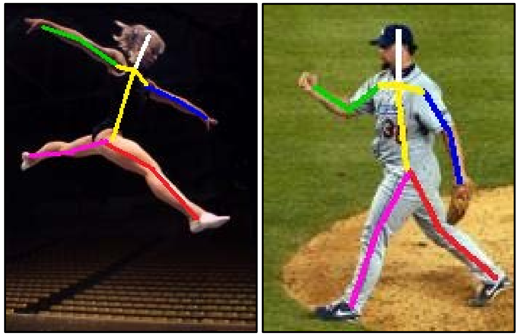
Bregler et al. 2004
Sigal et al. 2004
Deutscher&Reid 2005
Andriluka et al. 2010
Bergholdt et al. 2010
Gall et al. 2010
Hasler et al., 2009
Taylor et al. 2010
Yao et al. 2011



State-of-the-art for 2D:

- Pictorial structures
  - Discriminative part detectors
  - Simple generative body model
  - ✓ Exact & Efficient inference
  - ✓ Performs well in realistic settings
  - ✘ Does not model depth information

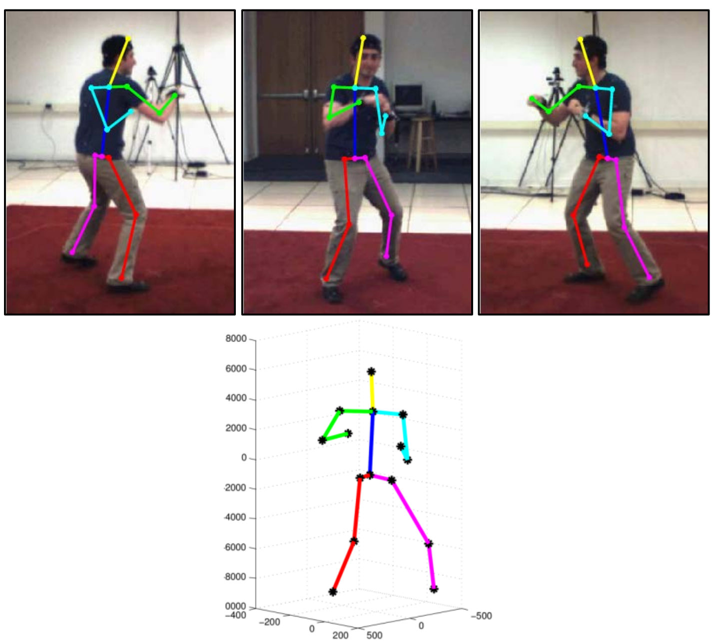
Lan & Huttenlocher 2005
Sigal & Black 2006
Andriluka et al. 2009
Ferrari et al. 2008
Sapp et al. 2010
Johnson&Everingham 2010
Yang&Ramanan 2011



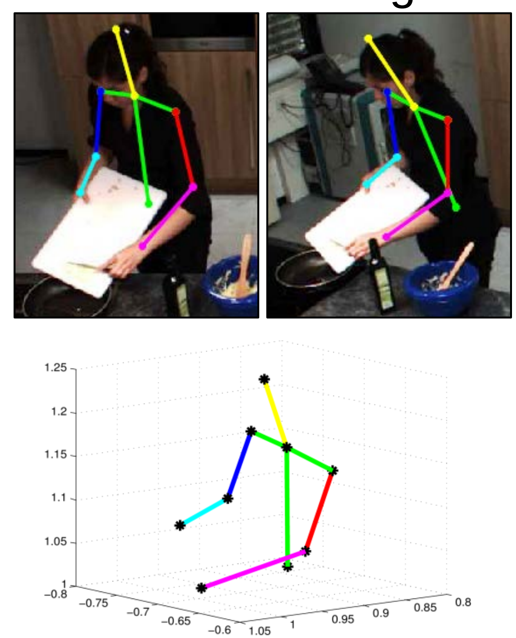
Goal

- Build upon state-of-the-art 2D pictorial structures model
- No temporal filtering / motion priors
- Pose estimation in
  - Laboratory settings (static background, single activity)
  - Realistic settings (non-static background, multiple activities)

HumanEva-I



MPII-Cooking



Related Work

Controlled laboratory conditions:

- Non-parameteric belief propagation [Sigal et al., 2004, 2006]
- Annealed particle filtering [Deutscher & Reid 2005]

Related Work

Controlled laboratory conditions:

- Non-parameteric belief propagation [Sigal et al., 2004, 2006]
- Annealed particle filtering [Deutscher & Reid 2005]

More realistic settings:

- Activity specific pose and motion priors [Gall et al., 2010]
- Structure from motion methods [Hasler et al., 2009]
- Combined visual and on-body inertial sensors [Pons-moll et al., 2011]

Related Work

Controlled laboratory conditions:

- Non-parameteric belief propagation [Sigal et al., 2004, 2006]
- Annealed particle filtering [Deutscher & Reid 2005]

More realistic settings:

- Activity specific pose and motion priors [Gall et al., 2010]
- Structure from motion methods [Hasler et al., 2009]
- Combined visual and on-body inertial sensors [Pons-moll et al., 2011]

- 3D Pictorial Structures for Multiple views [Burenius et al., 2013]
  - Simple tree graph, Higher complexity, Weaker likelihood

OUTLINE

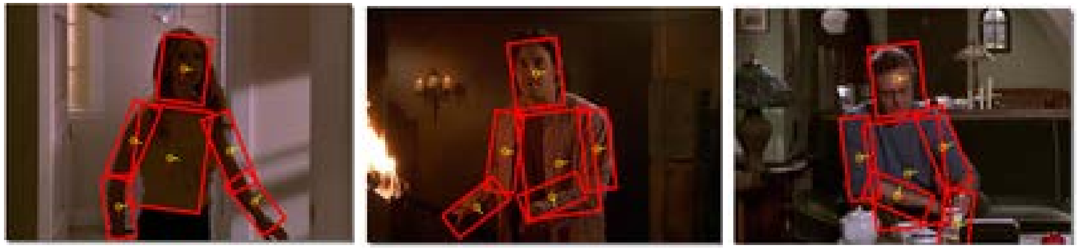
OUTLINE

OUTLINE

OUTLINE

likelihood *pose prior*

$$p(L|I) \propto p(I|L) \cdot p(L)$$

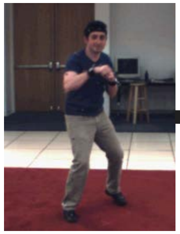


Pictorial structures implementation from [Andriluka et al., CVPR'09]

likelihood *pose prior*

$$p(L|I) \propto p(I|L) \cdot p(L)$$

Appearance features



Shape-context

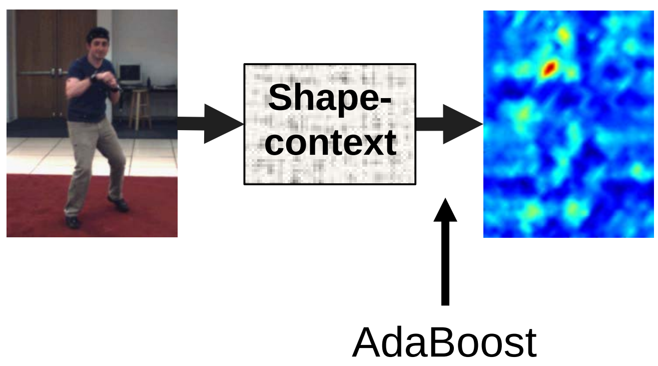
Pictorial structures implementation from [Andriluka et al., CVPR'09]

likelihood *pose prior*

$$p(L|I) \propto \boxed{p(I|L)} \cdot p(L)$$

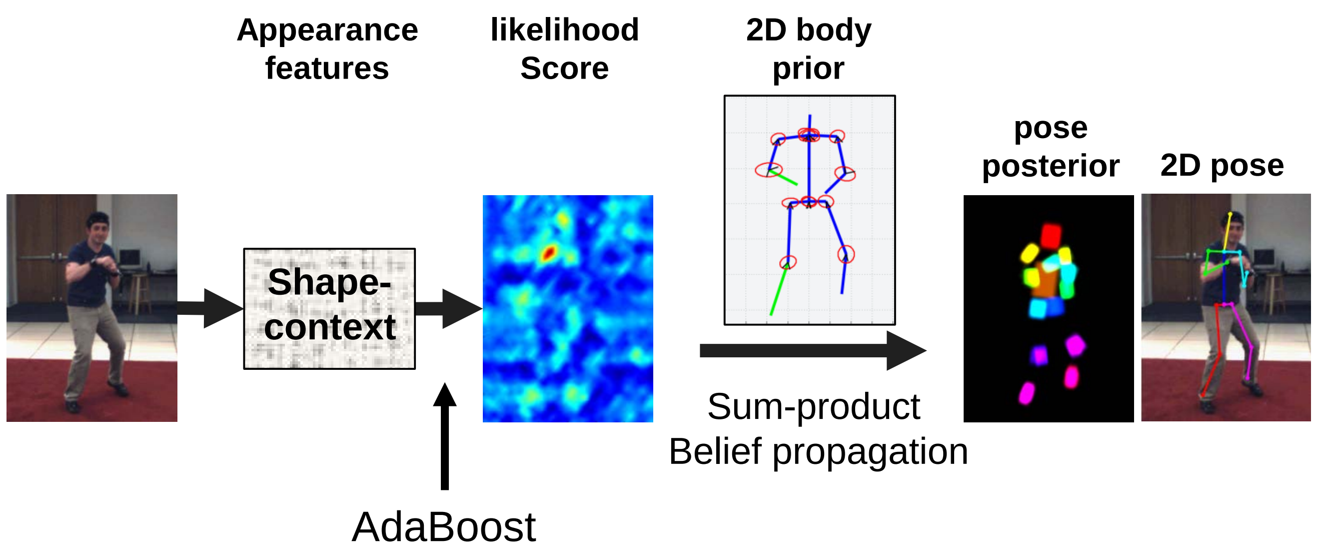
Appearance features

likelihood Score



Pictorial structures implementation from [Andriluka et al., CVPR’09]

$$p(L|I) \propto \overbrace{p(I|L)}^{\textit{likelihood}} \cdot \overbrace{p(L)}^{\textit{pose prior}}$$

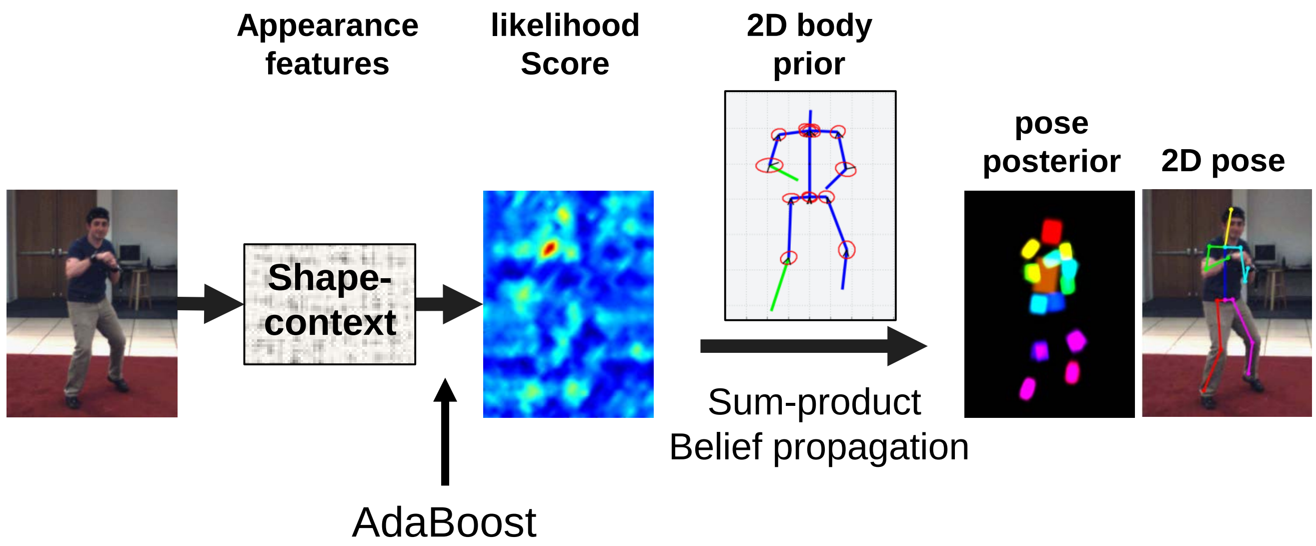


Pictorial structures implementation from [Andriluka et al., CVPR’09]

Foreshortening

likelihood *pose prior*

$$p(L|I) \propto p(I|L) \cdot p(L)$$



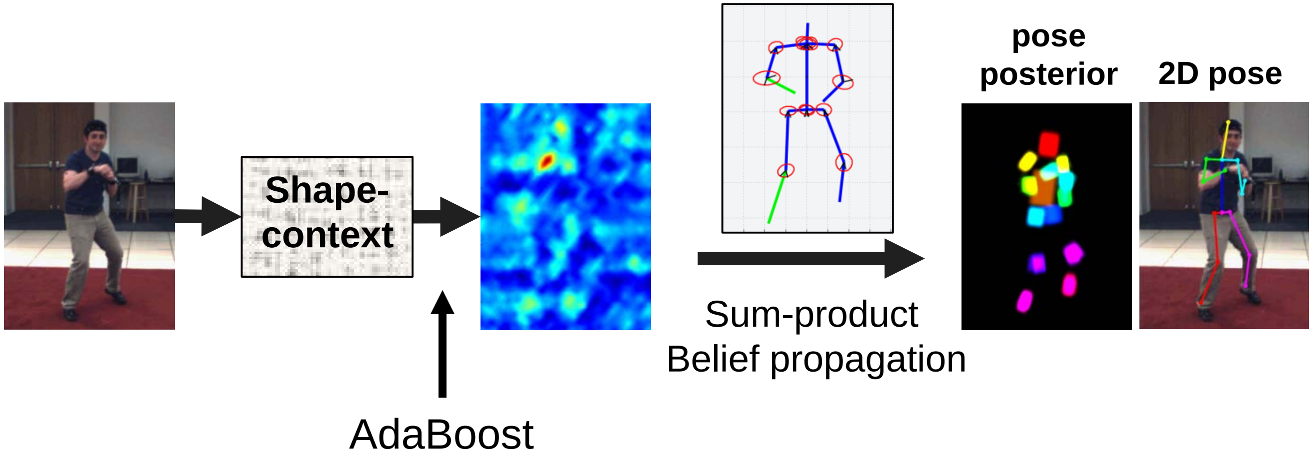
Pictorial structures implementation from [Andriluka et al., CVPR’09]

Foreshortening

Multi-modal part dependencies

$$p(L|I) \propto \underbrace{p(I|L)}_{\text{likelihood}} \cdot \underbrace{p(L)}_{\text{pose prior}}$$

Appearance features **likelihood Score** **2D body prior**



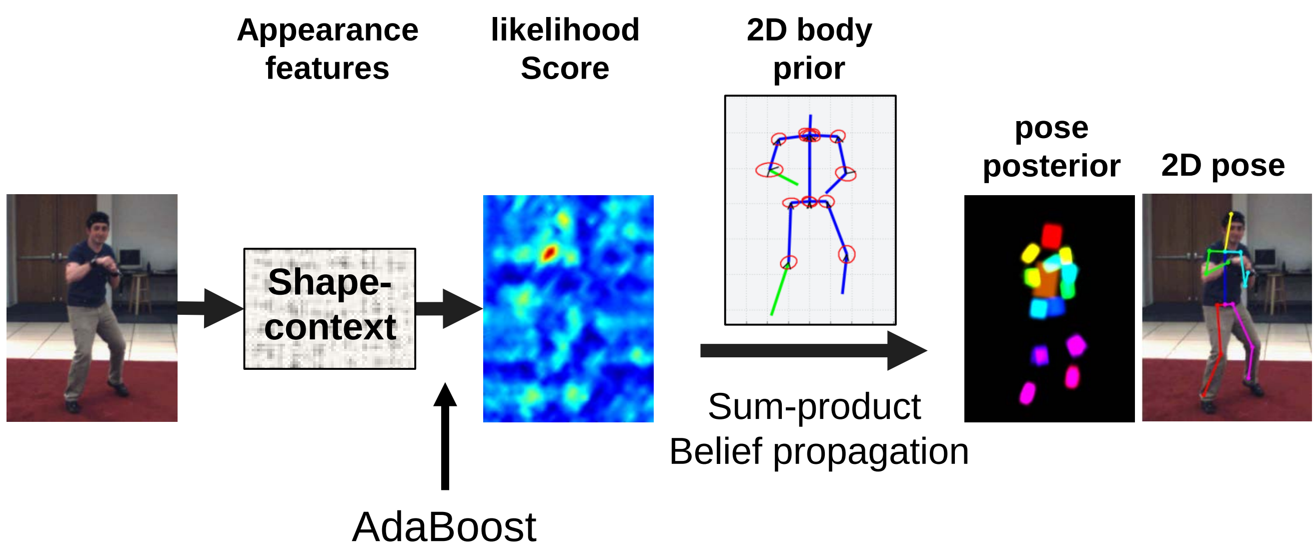
Pictorial structures implementation from [Andriluka et al., CVPR’09]

$$p(L|I) \propto \overbrace{p(I|L)}^{\textit{likelihood}} \cdot \overbrace{p(L)}^{\textit{pose prior}}$$

Foreshortening

Multi-modal part dependencies

Multi-modal part appearance



Pictorial structures implementation from [Andriluka et al., CVPR'09]

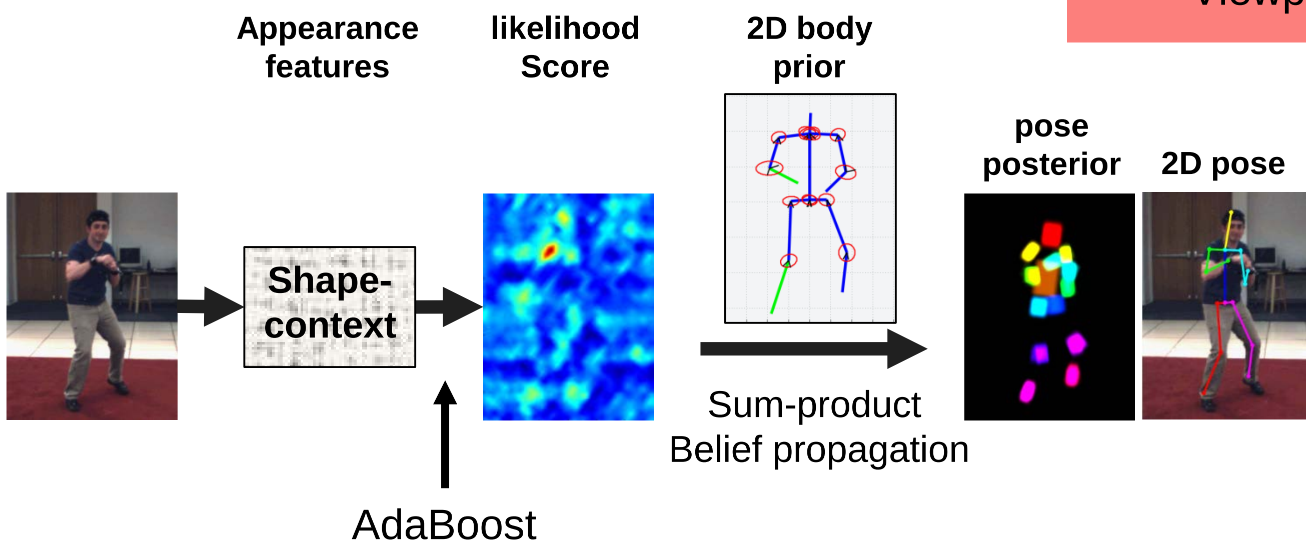
Foreshortening

Multi-modal part dependencies

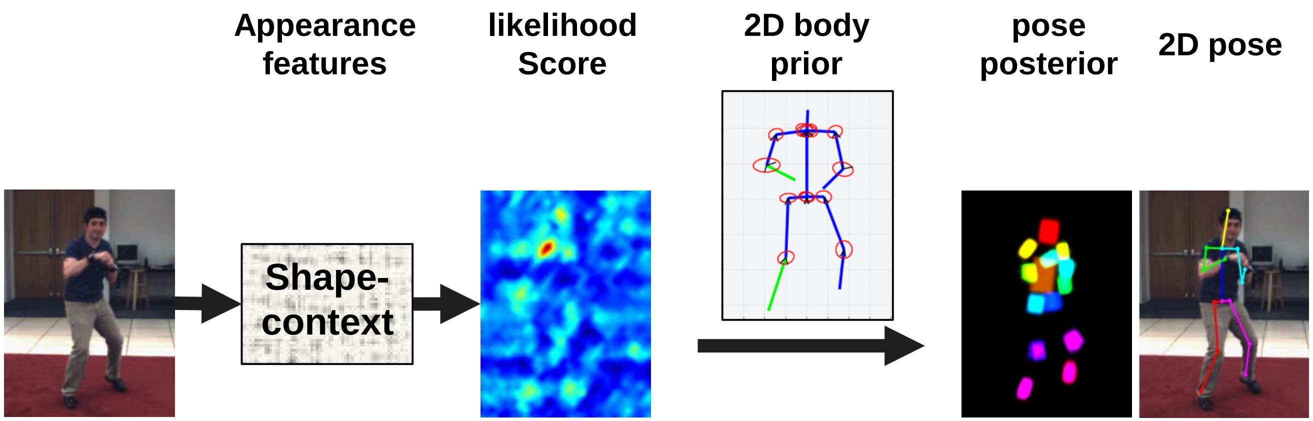
Multi-modal part appearance

Viewpoint variations

$$p(L|I) \propto \overset{\textit{likelihood}}{p(I|L)} \cdot \overset{\textit{pose prior}}{p(L)}$$



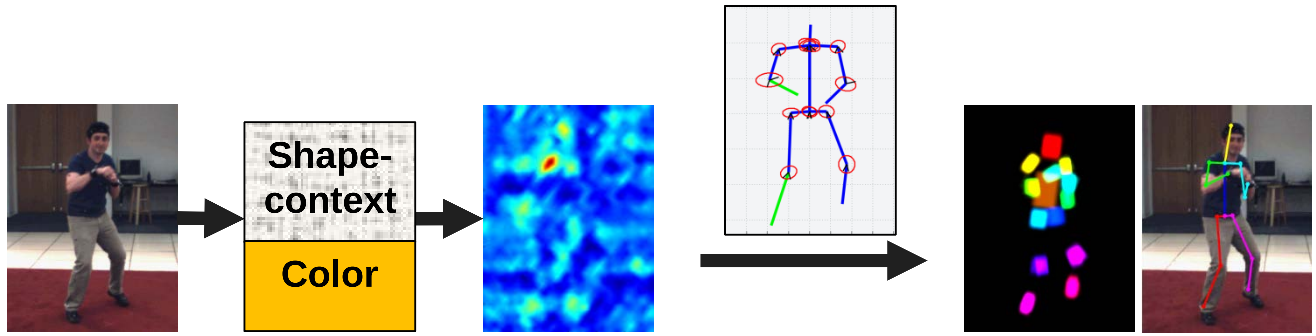
Pictorial structures implementation from [Andriluka et al., CVPR'09]


Appearance features
likelihood Score
2D body prior
pose posterior
2D pose
Shape-context

Joint feature representation

[Eichner & Ferrrai, ACCV’12]

Appearance features **likelihood Score** **2D body prior** **pose posterior** **2D pose**

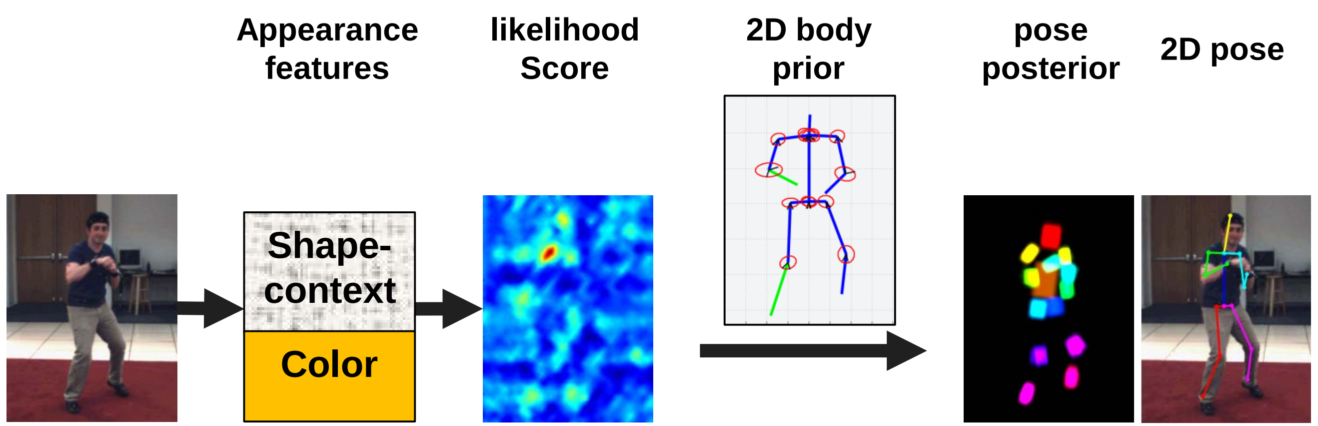


Joint feature representation

[Eichner & Ferrrai, ACCV’12]

Multi-modal pairwise terms

[Yang & Ramanan CVPR‘11]



Joint feature representation

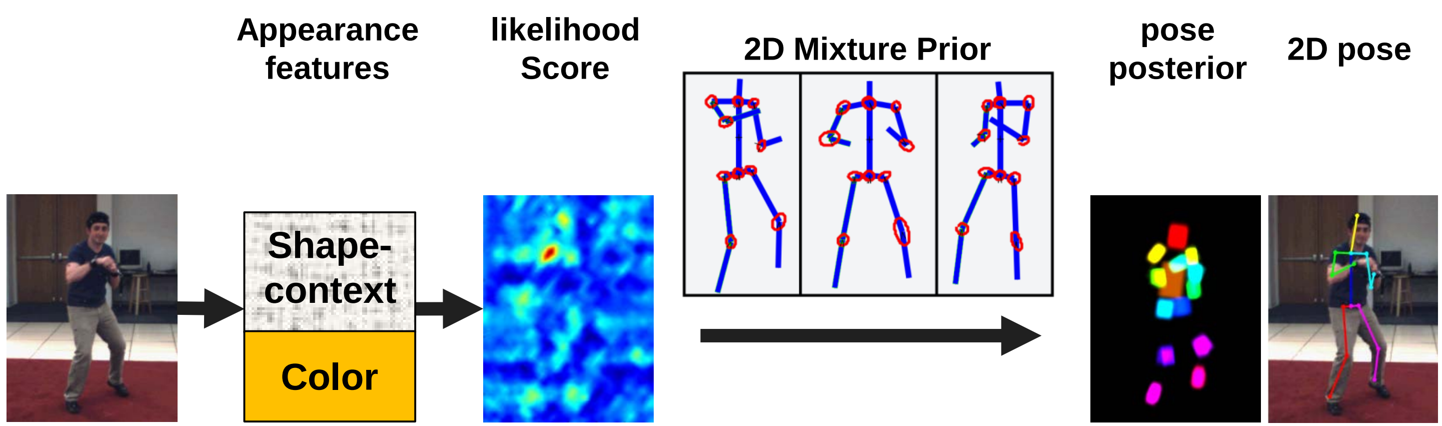
[Eichner & Ferrrai, ACCV'12]

Multi-modal pairwise terms

[Yang & Ramanan CVPR'11]

2D Mixture model

[Johnson & Everingham, BMVC'10]



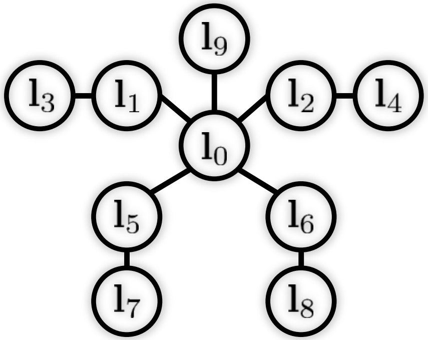
Joint feature representation

Multi-modal pairwise terms

2D Mixture model

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$



Joint feature representation | Multi-modal pairwise terms | 2D Mixture model

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$

$h_{n,t}$: weak classifier for feature t

$$f(\mathbf{l}_n; I) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}(e_n(\mathbf{l}_n))}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z} \boxed{\prod_{n=1}^{N} f_n(\mathbf{l}_n; I)} \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$

$$f(\mathbf{l}_n; I) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}(e_n(\mathbf{l}_n))}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

$h_{n,t}$: weak classifier for feature t

$\alpha_{n,t}$: learned weight for feature t

Joint feature representation | Multi-modal pairwise terms | 2D Mixture model

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$

$$f(\mathbf{l}_n; I) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}(e_n(\mathbf{l}_n))}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

$h_{n,t}$: weak classifier for feature t

$\alpha_{n,t}$: learned weight for feature t

$e_n(\mathbf{l}_n)$: feature vector for part n

shape-context

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z} \prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$

$h_{n,t}$: weak classifier for feature t

$\alpha_{n,t}$: learned weight for feature t

$e_n(\mathbf{l}_n)$: feature vector for part n

$$f(\mathbf{l}_n; I) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}(e_n(\mathbf{l}_n))}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

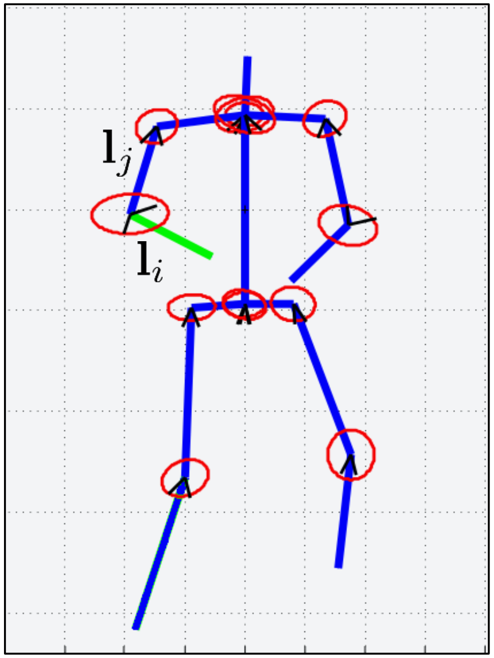
shape-context | color

AdaBoost

Discriminative classifier for part *n*

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z}\prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$



Pairwise Terms:

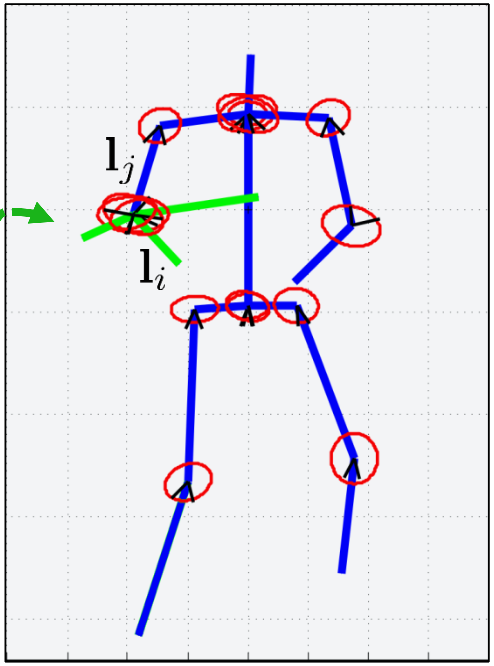
$$f_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \mathcal{N}(T_{ji}(\mathbf{l}_i) - T_{ij}(\mathbf{l}_j)|\mu_{ij}, \Sigma_{ij})$$

unary terms *pairwise terms*

$$p(L|I) = \frac{1}{Z}\prod_{n=1}^{N} f_n(\mathbf{l}_n; I) \cdot \prod_{(i,j)\in E} f_{ij}(\mathbf{l}_i, \mathbf{l}_j)$$

Multi-Modal Pairwise Terms:

$$f_{ij}(\mathbf{l}_i, \mathbf{l}_j) = \max_{k=1}^{K} \mathcal{N}(T_{ji}^{k}(\mathbf{l}_i) - T_{ij}^{k}(\mathbf{l}_j) | \mu_{ij}^{k}, \Sigma_{ij}^{k})$$



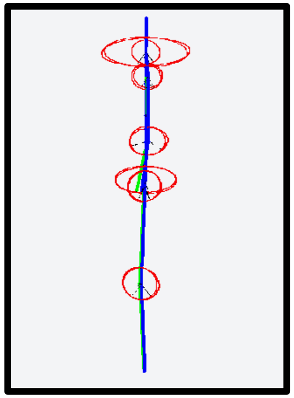
Explicitly models:
- ✓ Foreshortening
- ✓ Multiple articulations

Joint feature representation | Multi-modal pairwise terms | 2D Mixture model

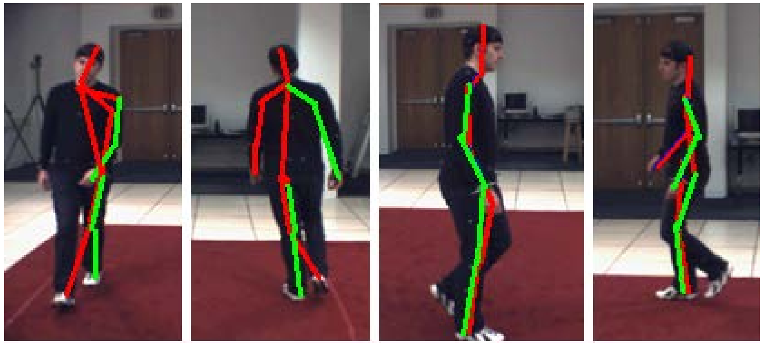
Ground-truth for ***Walking*** activity – Train set

Training
Single component

Very broad
2D body prior



2D Pose Results



Joint feature representation

Multi-modal pairwise terms

2D Mixture model

[Johnson & Everingham, BMVC'10]

Ground-truth for ***Walking*** activity – Train set

Joint feature representation

Multi-modal pairwise terms

2D Mixture model

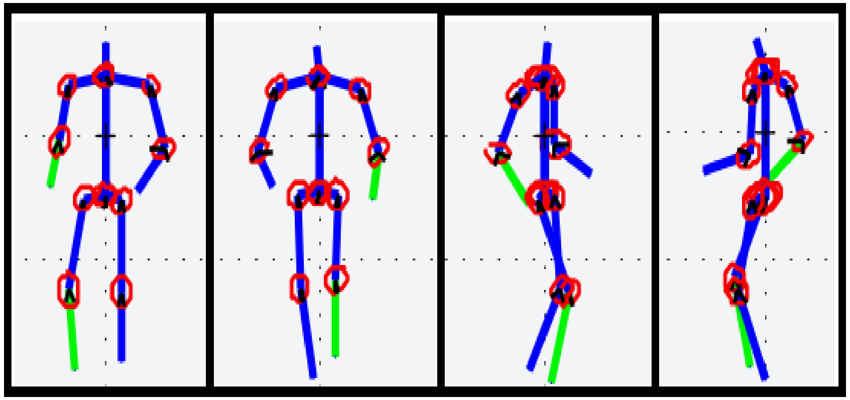
[Johnson & Everingham, BMVC'10]

Ground-truth for ***Walking*** activity – Train set

[Johnson & Everingham, BMVC‘10]

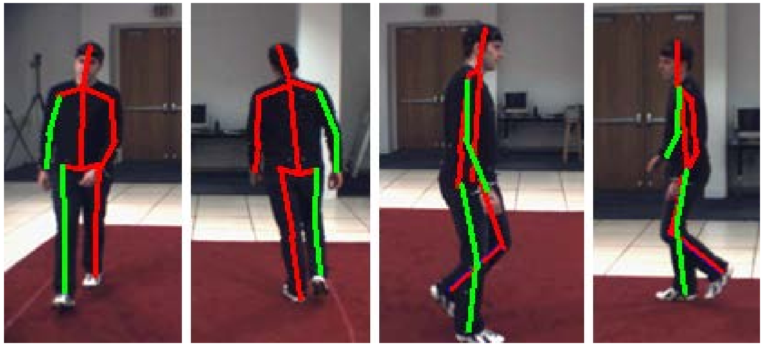
Ground-truth for ***Walking*** activity – Train set

2D Mixture prior models



front back right left

2D Pose Results



Component selection

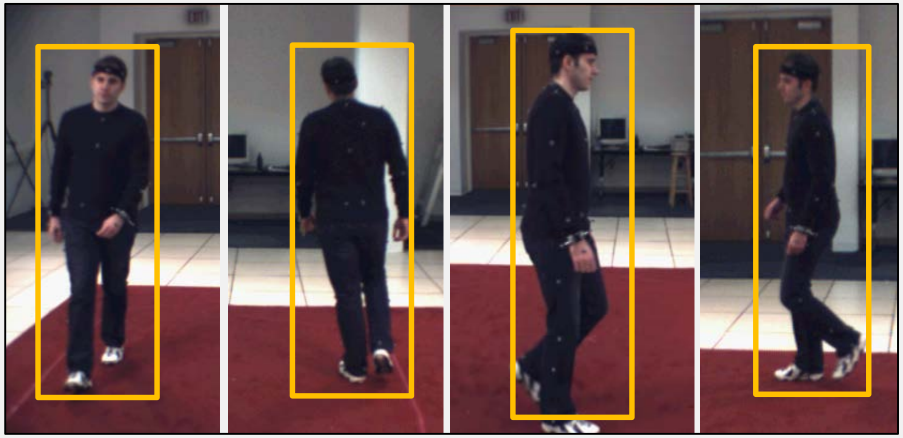
Classifier	Minimum variance

Component selection

Classifier | Minimum variance

1. Component Classifier
   [Pepik et al., CVPR‘12]

➢ Joint object detection & viewpoint classification



Component selection

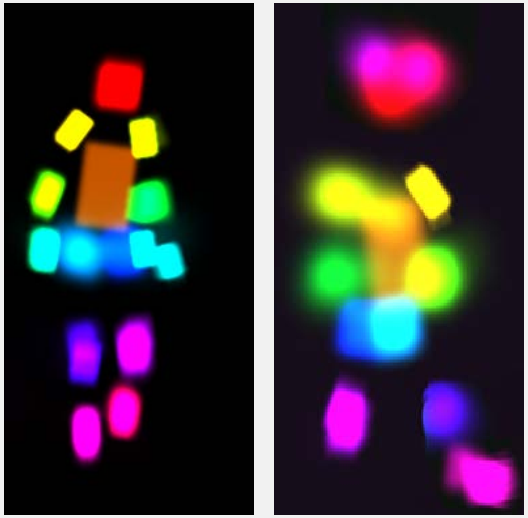
Classifier | Minimum variance

2. Minimum variance
inspired by [Jammalamadaka et al., ECCV'12]

un-certaininty

$$\hat{k} = \text{argmin}_k \, s(k, I)$$

$$s(k, I) = \sum_{n=1}^{N} \|\text{Cov}_n(k, I)\|_2$$



k=1 k=2

Component selection

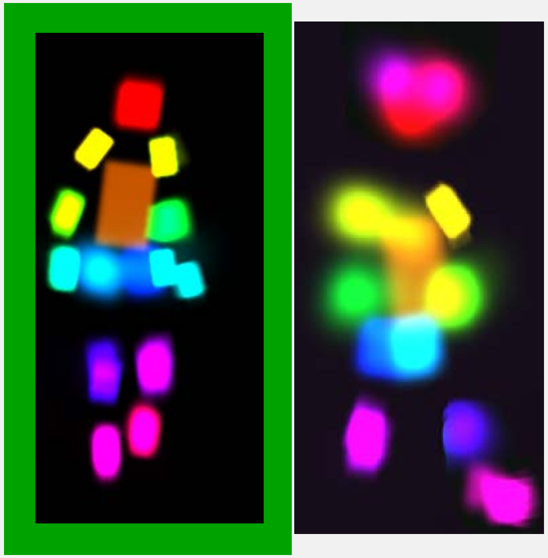
Classifier | **Minimum variance**

2. Minimum variance
   inspired by [Jammalamadaka et al., ECCV'12]

un-certaininty

$$\hat{k} = \text{argmin}_k \, s(k, I)$$

$$s(k, I) = \sum_{n=1}^{N} \|\text{Cov}_n(k, I)\|_2$$



k=1 k=2

3D Mixture model

3D Mixture model

Multi-view
pairwise factors

3D Mixture model

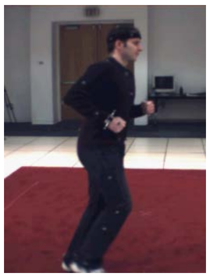
Multi-view
pairwise factors

2-stage inference

[Andriluka et al., IJCV'12]

3D Mixture model | Multi-view pairwise factors | 2-stage inference

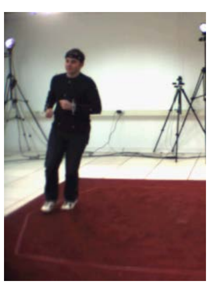
View 1

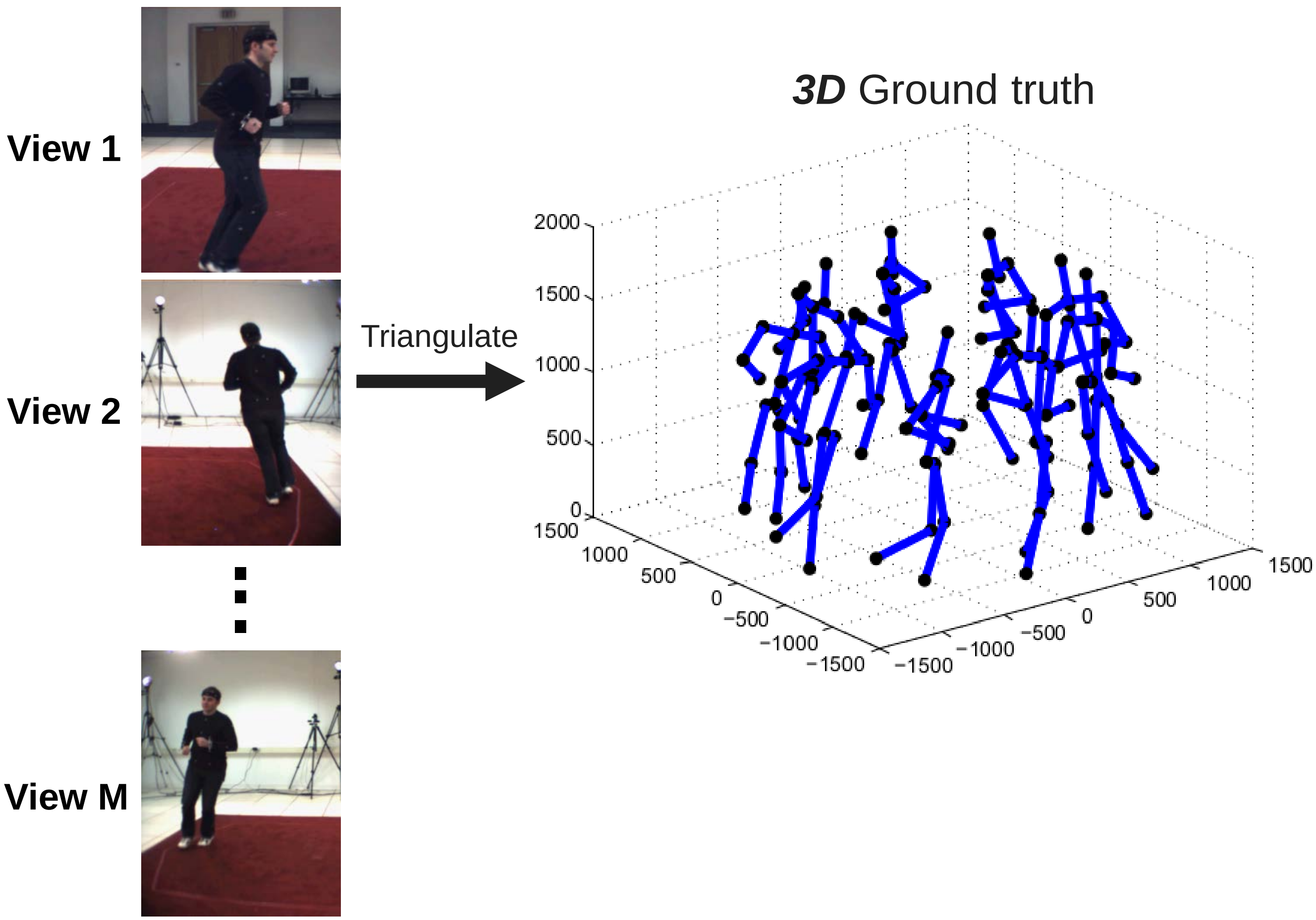


View 2

⋮

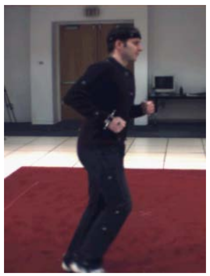
View M




View 1
View 2
View M
Triangulate
3D Ground truth
2000
1500
1000
500
0
1500
1000
500
0
−500
−1000
−1500
−1500
−1000
−500
0
500
1000
1500

3D Mixture model | Multi-view pairwise factors | 2-stage inference

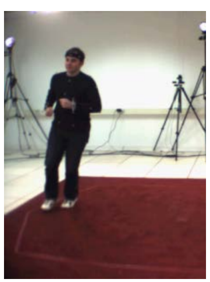
View 1



View 2

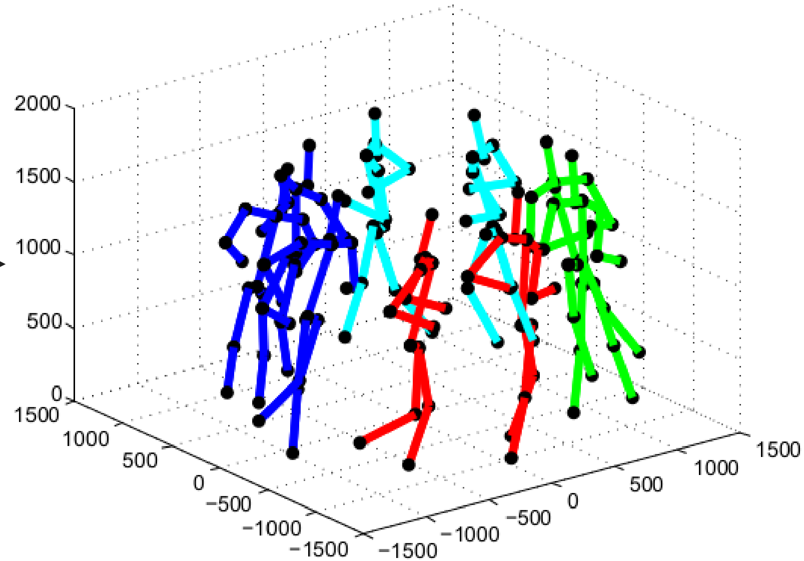
⋮

View M



Triangulate →

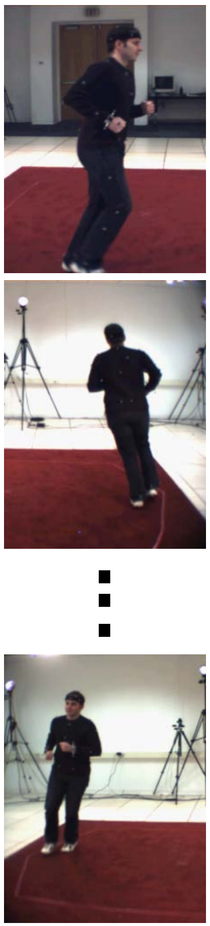
3D Ground truth



View 1

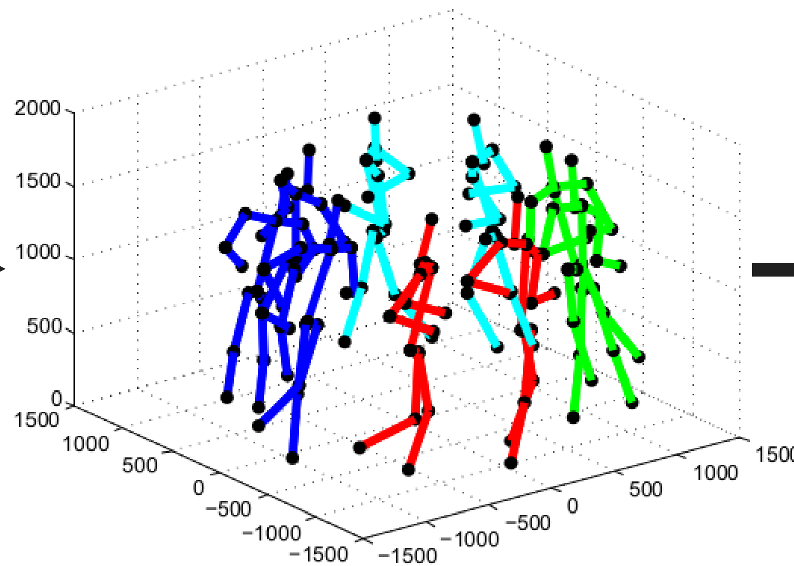
View 2

View M

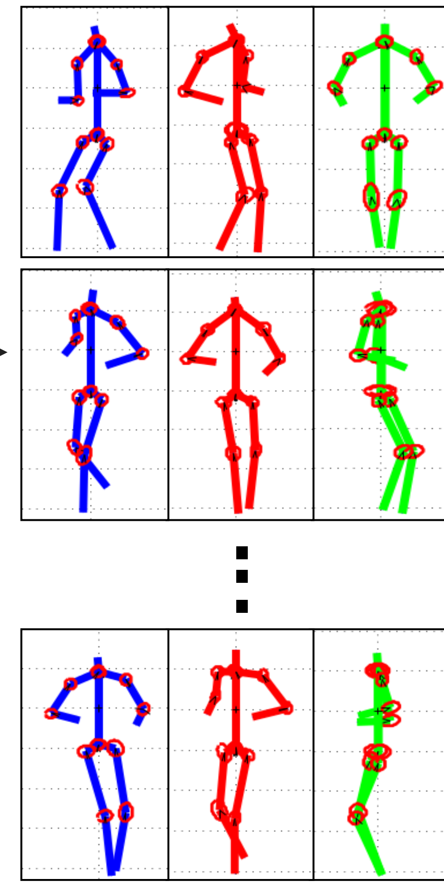


Triangulate

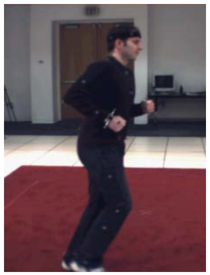
3D Ground truth



Consistent *2D* body priors



View 1



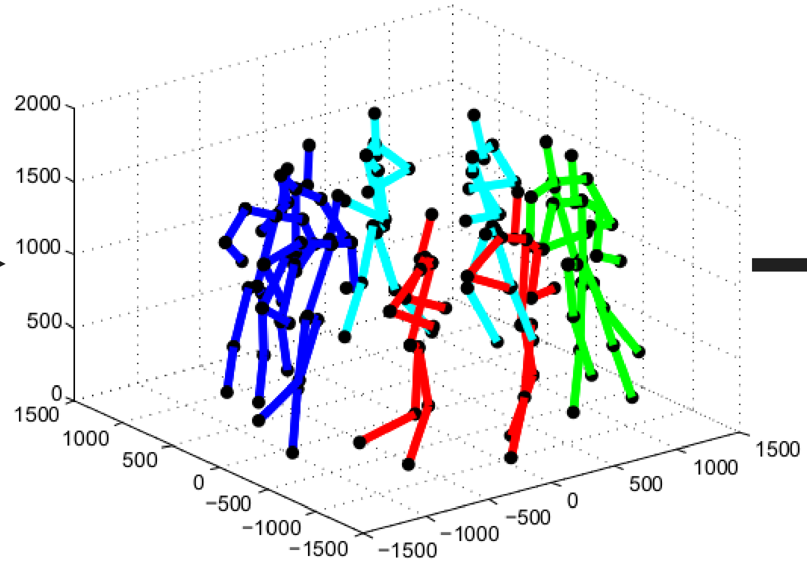
View 2

⋮

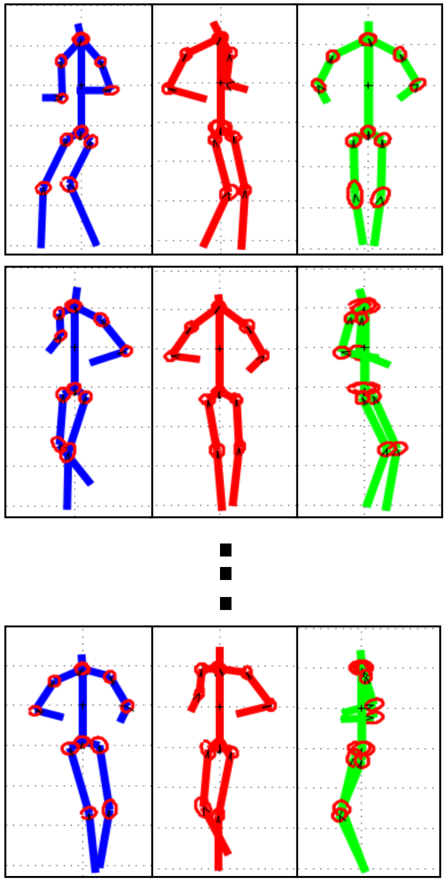
View M

Triangulate

3D Ground truth

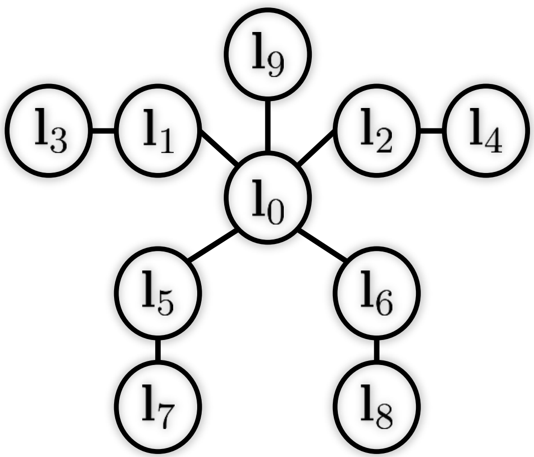


Consistent ***2D*** body priors

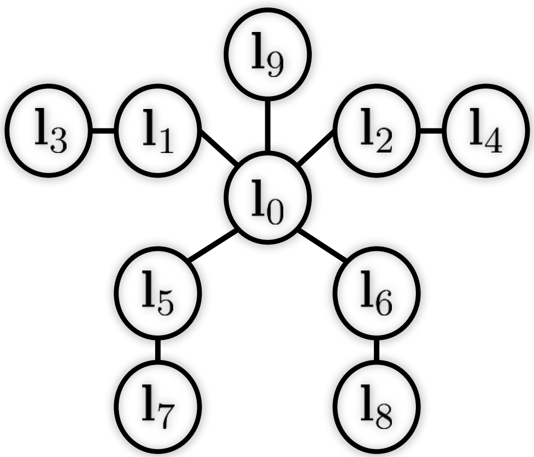


- ✓ Consistent *Multi-view* appearance
- ✓ Consistent 2d body priors
- ✓ No viewpoint ambiguities

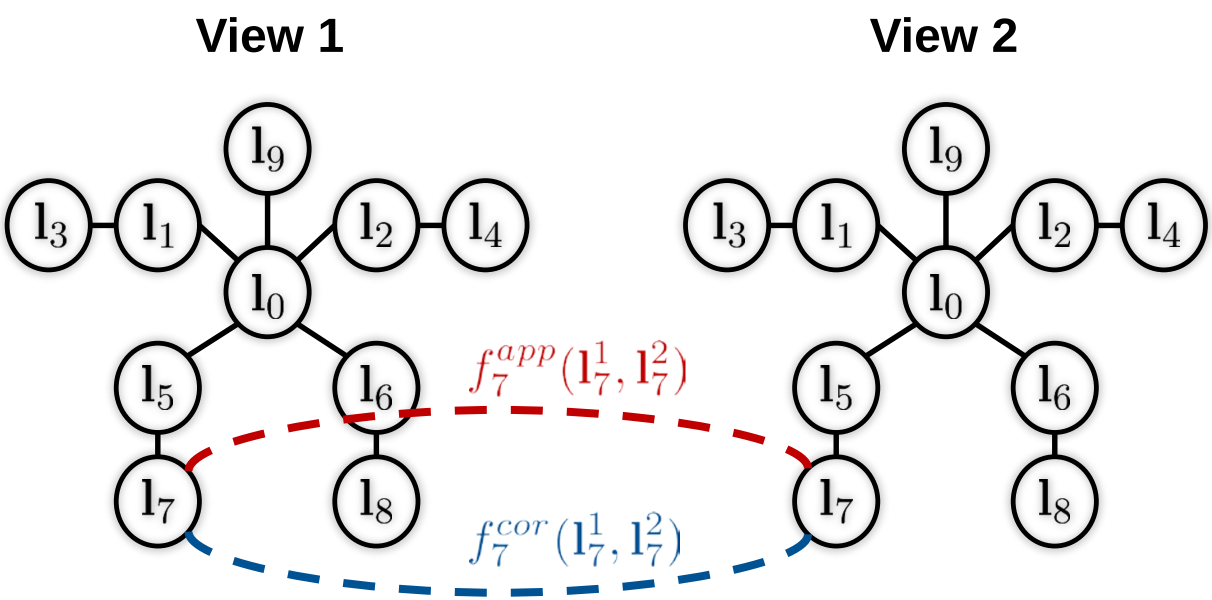
View 1



View 2



$$p(L_1, L_2 | I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2)$$



Single-view factors *Multi-view pairwise factors*

$$p(L_1, L_2 | I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2)$$

View-1 S C

View-2 S C

Multi-view Appearance:

$$f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}([e_n(\mathbf{l}_n^1), e_n(\mathbf{l}_n^2)])}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

Multi-view Correspondence:

$$f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2) = \exp(-(\|\mathbf{l}_n^1 - \hat{\mathbf{l}}_n^1\|^2 + \|\mathbf{l}_n^2 - \hat{\mathbf{l}}_n^2\|^2))$$

$$p(L_1, L_2 | I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2)$$

View-1		View-2	
S	C	S	C

2-view Joint feature vector

Multi-view Appearance:

$$f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) = \max\left(\frac{\sum_t \alpha_{n,t} h_{n,t}([e_n(\mathbf{l}_n^1), e_n(\mathbf{l}_n^2)])}{\sum_t \alpha_{n,t}}, \varepsilon_0\right)$$

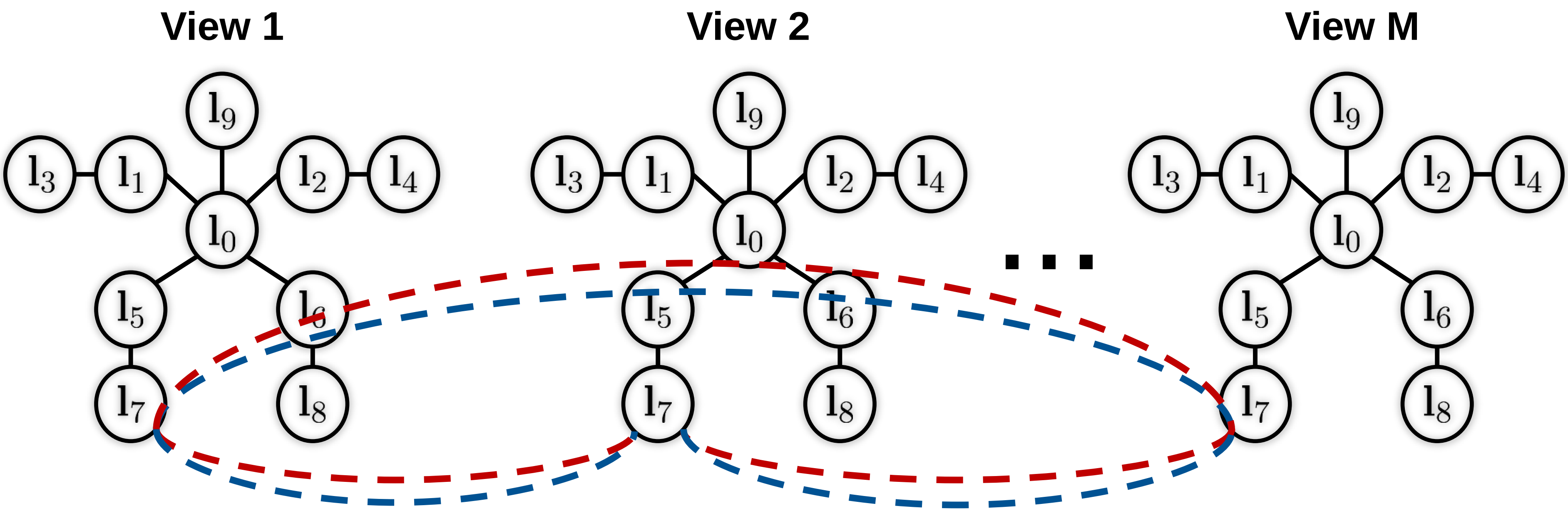
Reprojection error

Multi-view Correspondence:

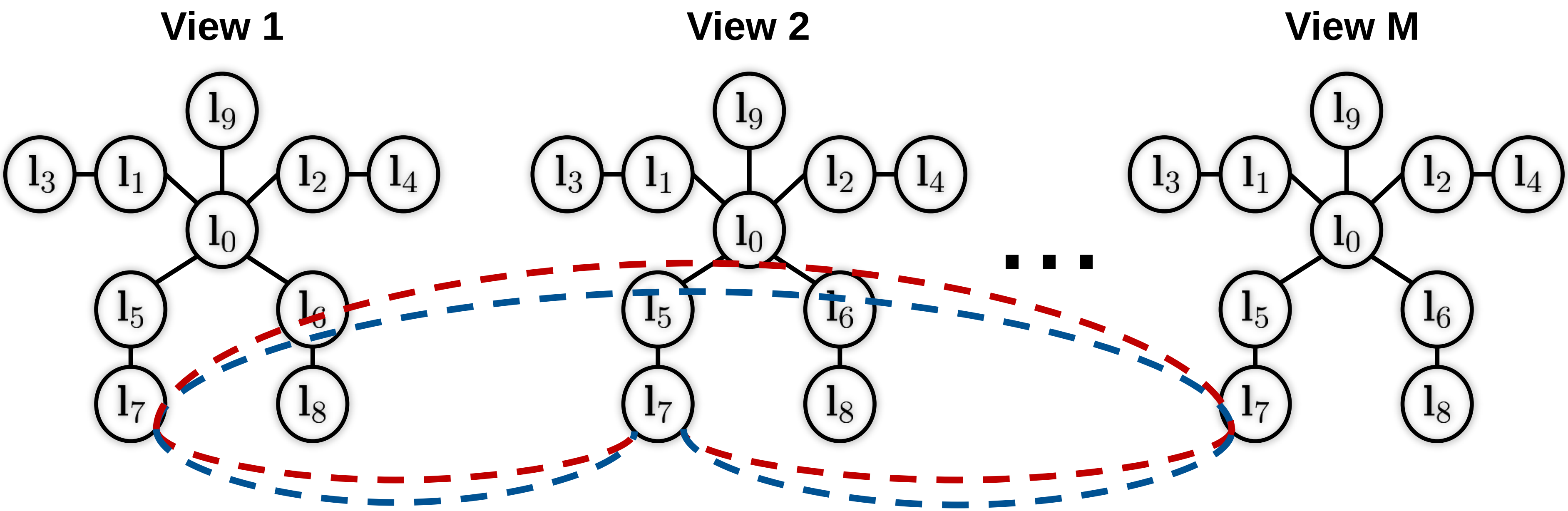
$$f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2) = \exp(-(\|\mathbf{l}_n^1 - \hat{\mathbf{l}}_n^1\|^2 + \|\mathbf{l}_n^2 - \hat{\mathbf{l}}_n^2\|^2))$$

$$p(L_1, L_2 | I_1, I_2) = \frac{1}{Z} f(L_1; I_1) f(L_2; I_2) \prod_n f_n^{app}(\mathbf{l}_n^1, \mathbf{l}_n^2; I_1, I_2) f_n^{cor}(\mathbf{l}_n^1, \mathbf{l}_n^2)$$

Pairwise terms between all *view-pairs*

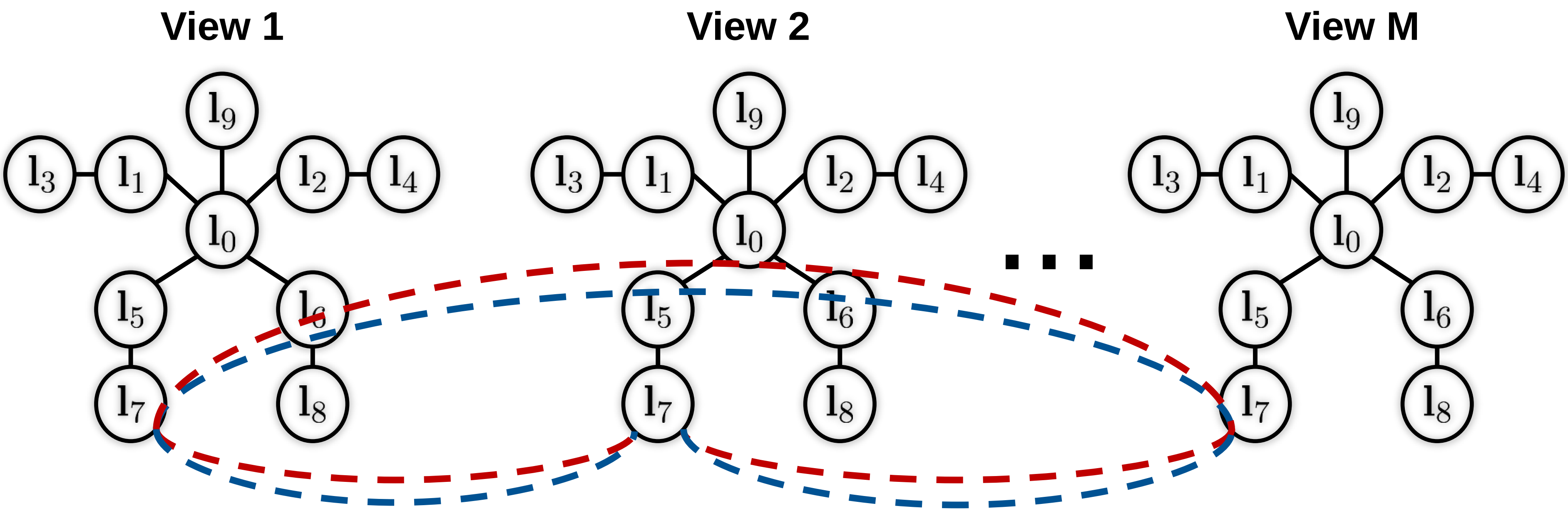


Pairwise terms between all *view-pairs*



- Loopy graphical model
- Non-Gaussian terms
  - multi-modal pairwise terms
  - multi-view pairwise factors

Pairwise terms between all *view-pairs*

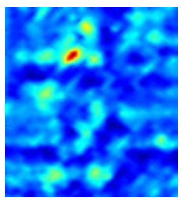


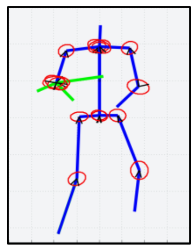
- Loypy graphical model
- Non-Gaussian terms
  - multi-modal pairwise terms
  - multi-view pairwise factors

✘ Exact Inference not tractable

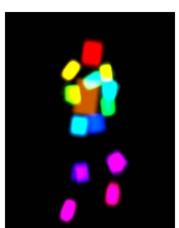
[Andriluka et al., IJCV'12]

Separate inference for each view



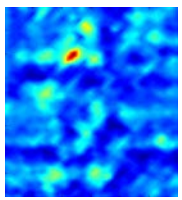


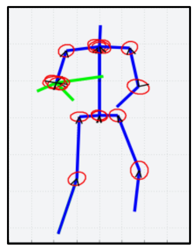
sum-product BP



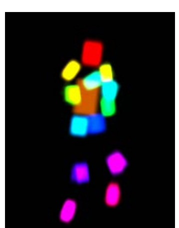
[Andriluka et al., IJCV‘12]

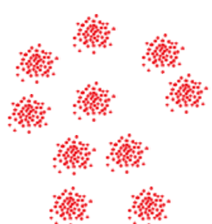
Separate inference for each view





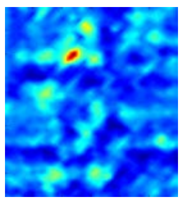
sum-product BP

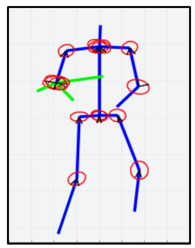




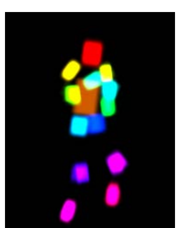
[Andriluka et al., IJCV'12]

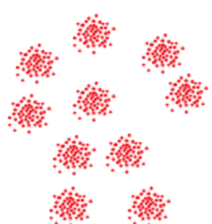
Separate inference for each view





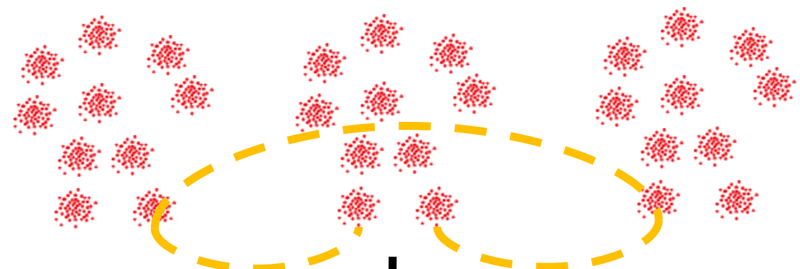
sum-product BP



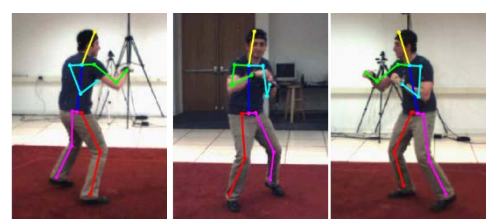


Joint Inference

view-1 view-2 view-3

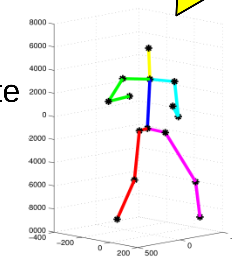


max-product BP



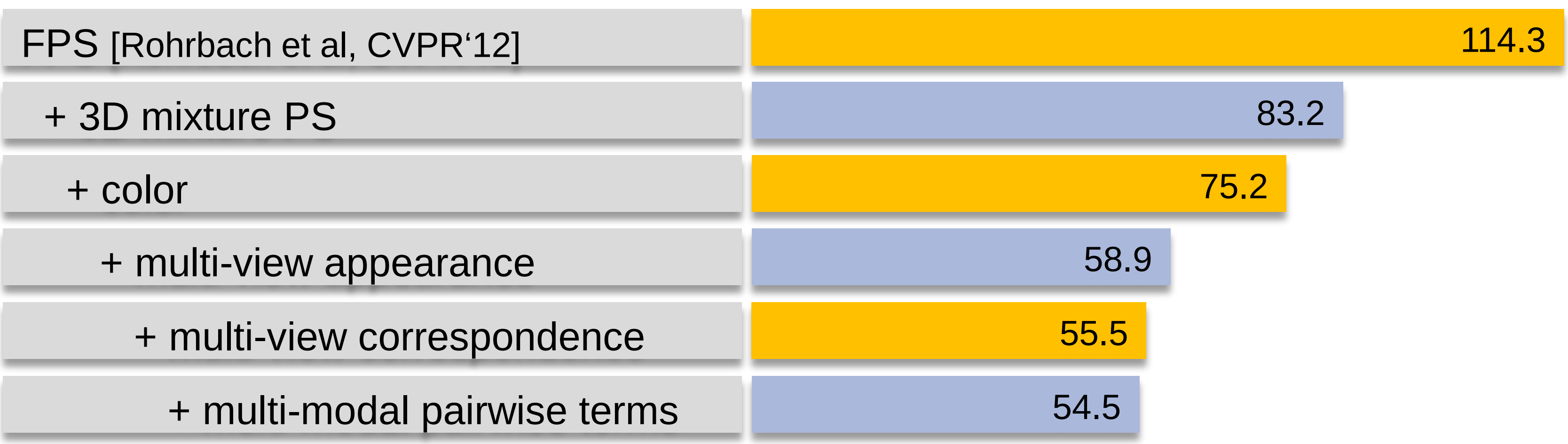
Triangulate

3D Pose



HumanEva-I

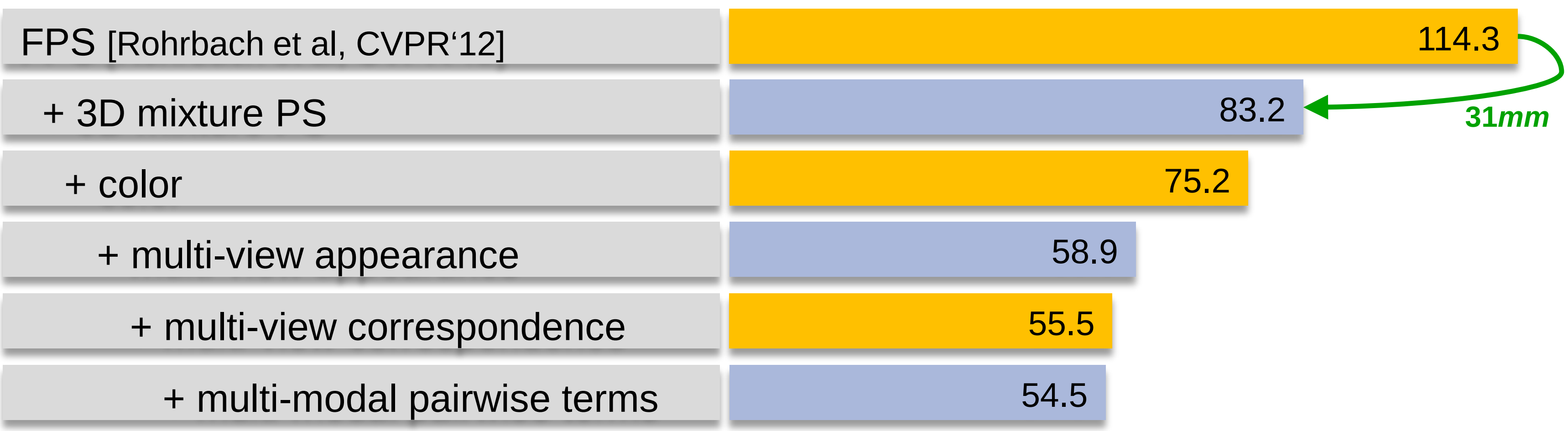
- Standard 3D human pose estimation dataset
- 3 color cameras
- Evaluated for 4 different activities as in [Taylor et al. 2010, Yao et al. 2011]
  - Walking, Jog, Box, Combo (Walking + Jog)



Improvements on walking sequence (S1) – 3D error in *mm*

HumanEva-I

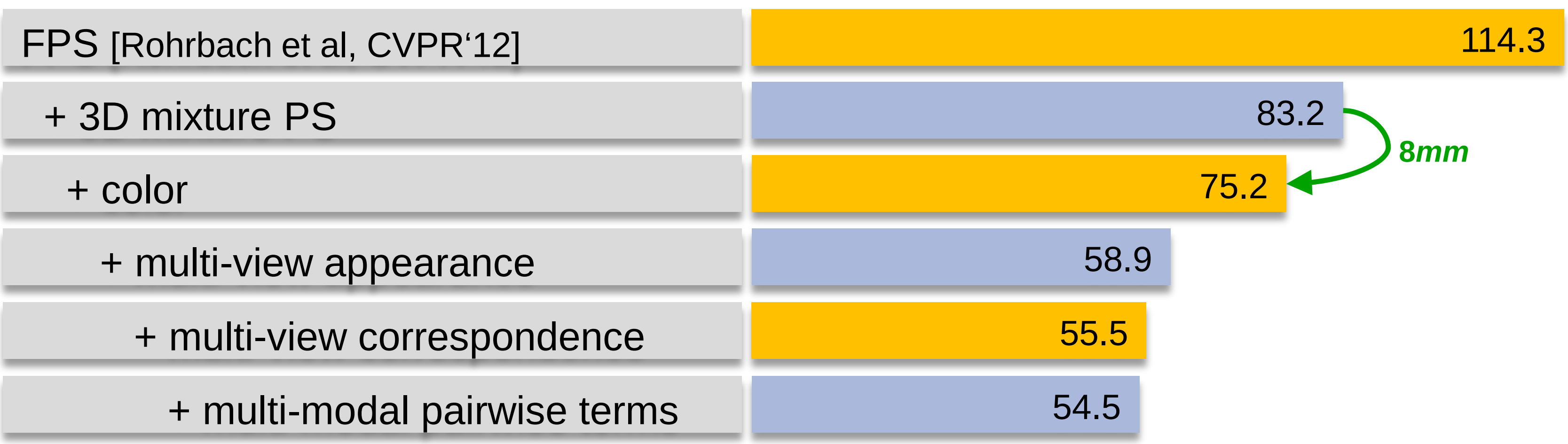
- Standard 3D human pose estimation dataset
- 3 color cameras
- Evaluated for 4 different activities as in [Taylor et al. 2010, Yao et al. 2011]
  - Walking, Jog, Box, Combo (Walking + Jog)



Improvements on walking sequence (S1) – 3D error in *mm*

HumanEva-I

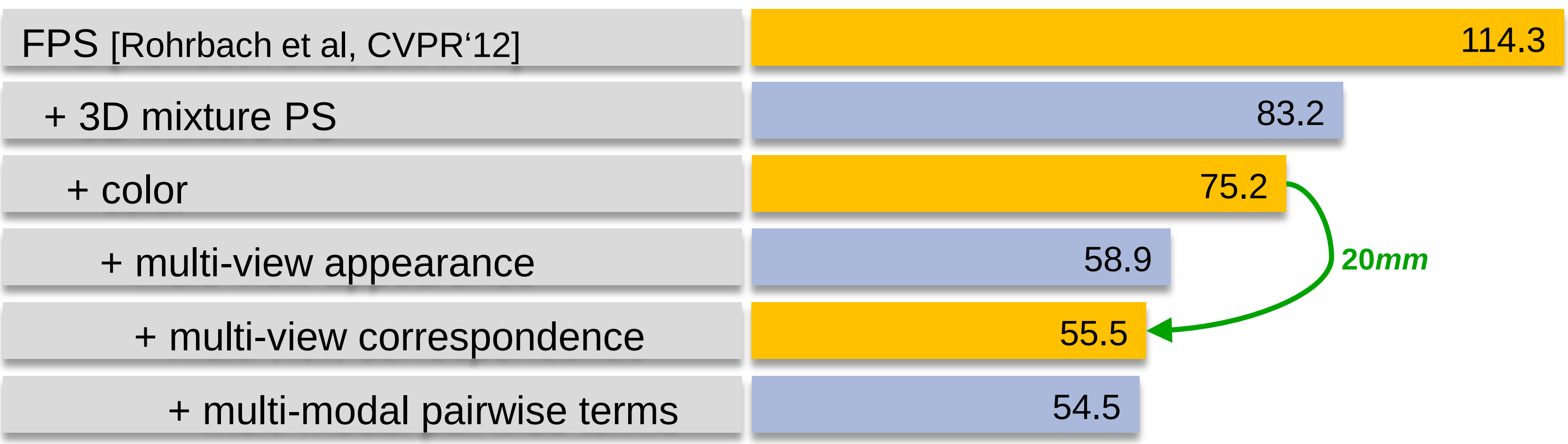
- Standard 3D human pose estimation dataset
- 3 color cameras
- Evaluated for 4 different activities as in [Taylor et al. 2010, Yao et al. 2011]
    - Walking, Jog, Box, Combo (Walking + Jog)



Improvements on walking sequence (S1) – 3D error in *mm*

HumanEva-I

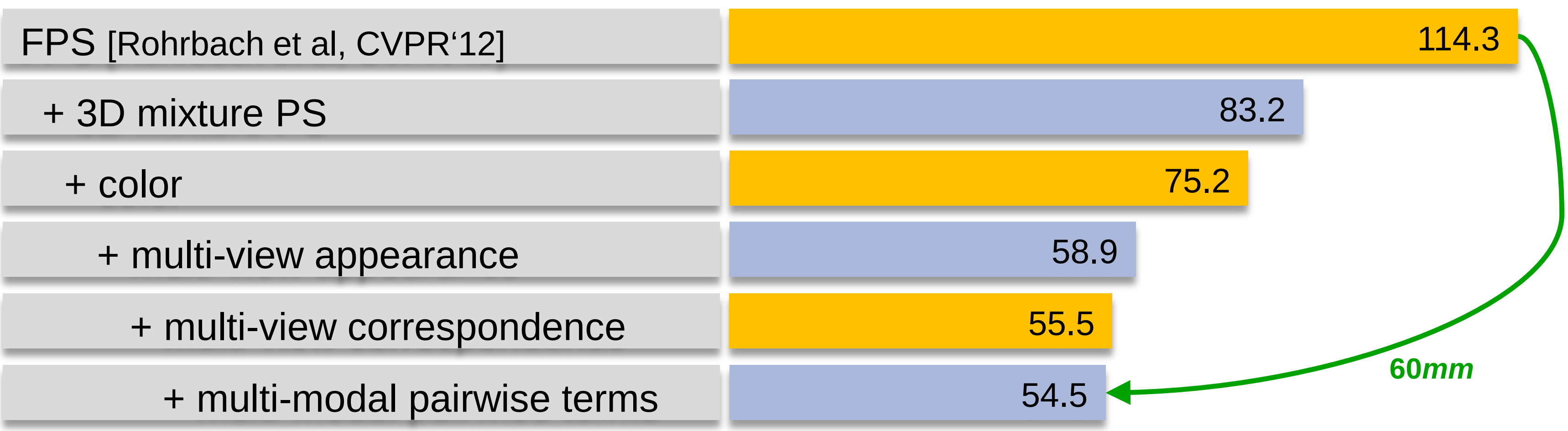
- Standard 3D human pose estimation dataset
- 3 color cameras
- Evaluated for 4 different activities as in [Taylor et al. 2010, Yao et al. 2011]
  - Walking, Jog, Box, Combo (Walking + Jog)



Improvements on walking sequence (S1) – 3D error in *mm*

HumanEva-I

- Standard 3D human pose estimation dataset
- 3 color cameras
- Evaluated for 4 different activities as in [Taylor et al. 2010, Yao et al. 2011]
  - Walking, Jog, Box, Combo (Walking + Jog)



Improvements on walking sequence (S1) – 3D error in *mm*

HumanEva-I

Combo

CRBM [Taylor et al, CVPR'10]	75.5
our	51.8

Test Subject S3 – 3D error in *mm*

Box

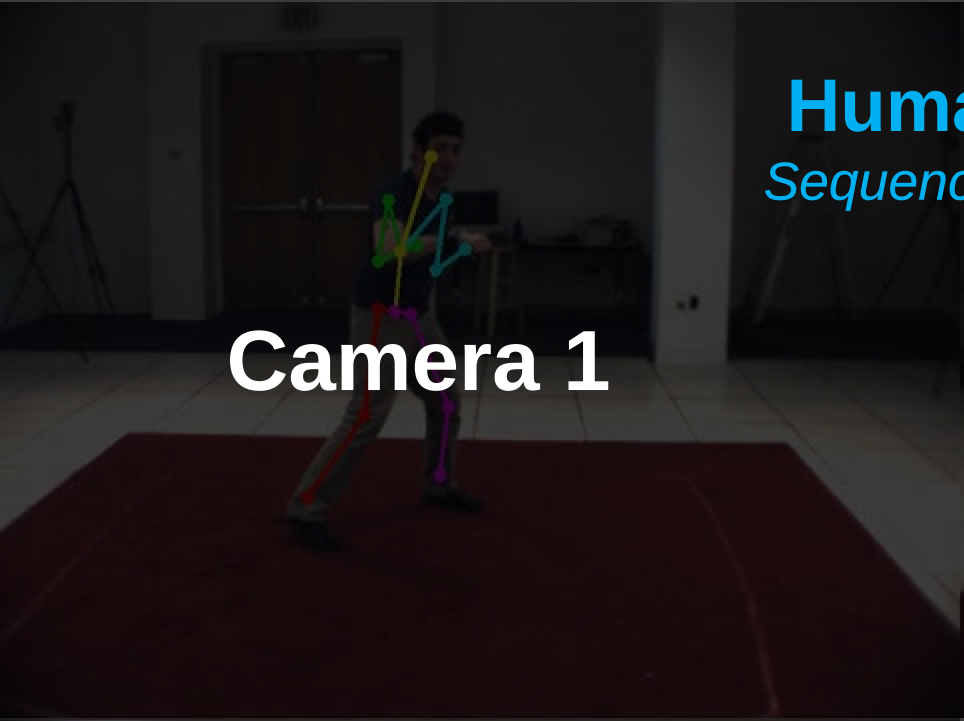
CRBM [Taylor et al, CVPR'10]	75.4
[Yao et al, NIPS'11]	74.1
our	47.7

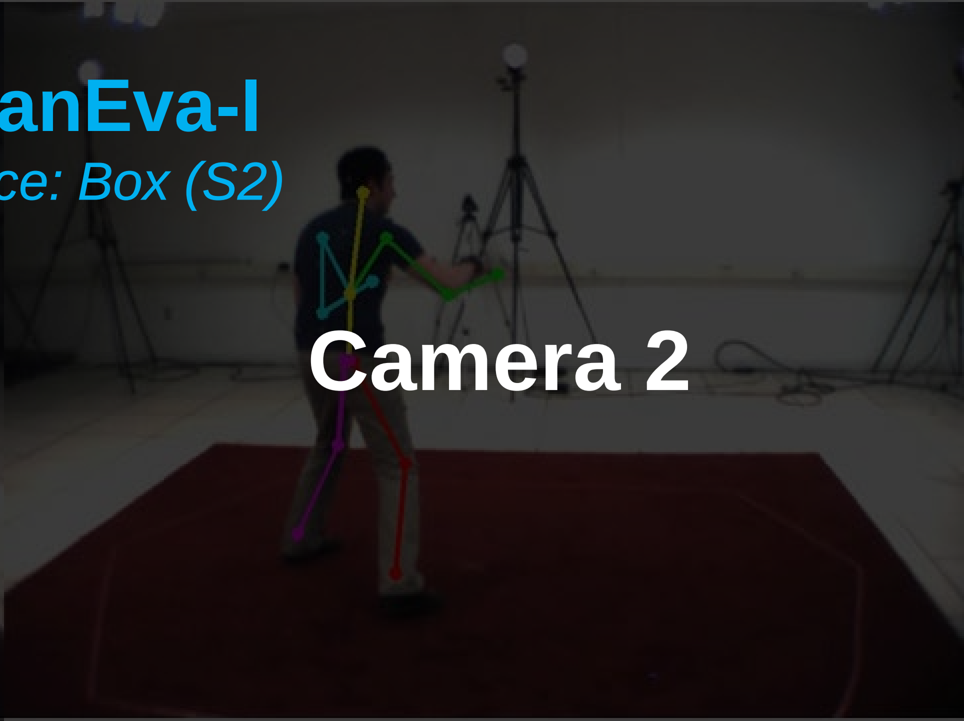
Test Subject S1 – 3D error in *mm*

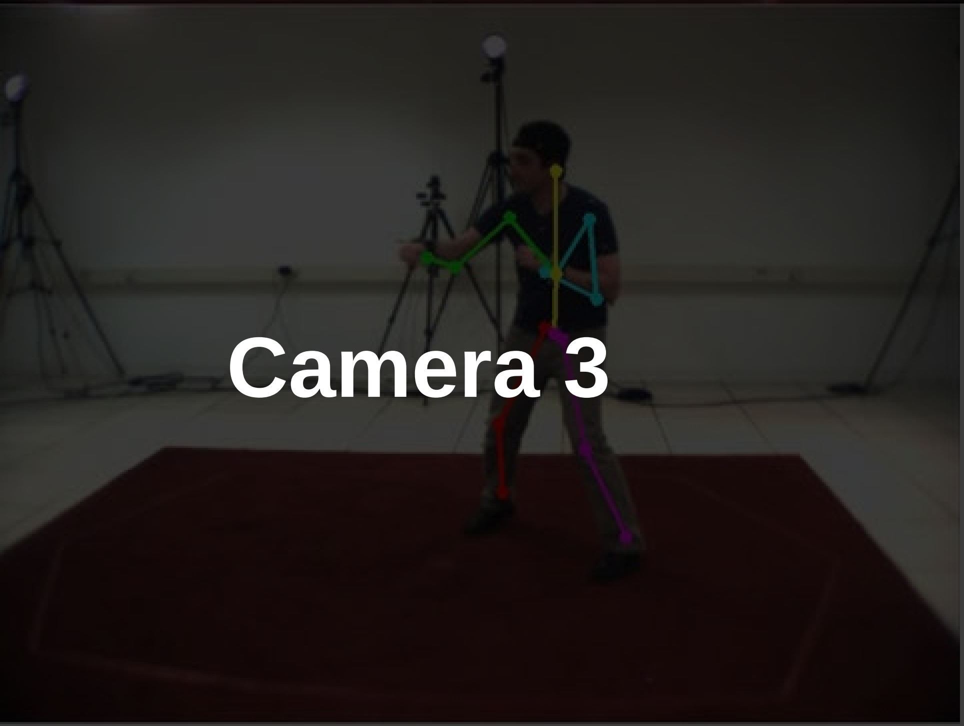
Quantitative comparison to state-of-the-art – 3D error in *mm*

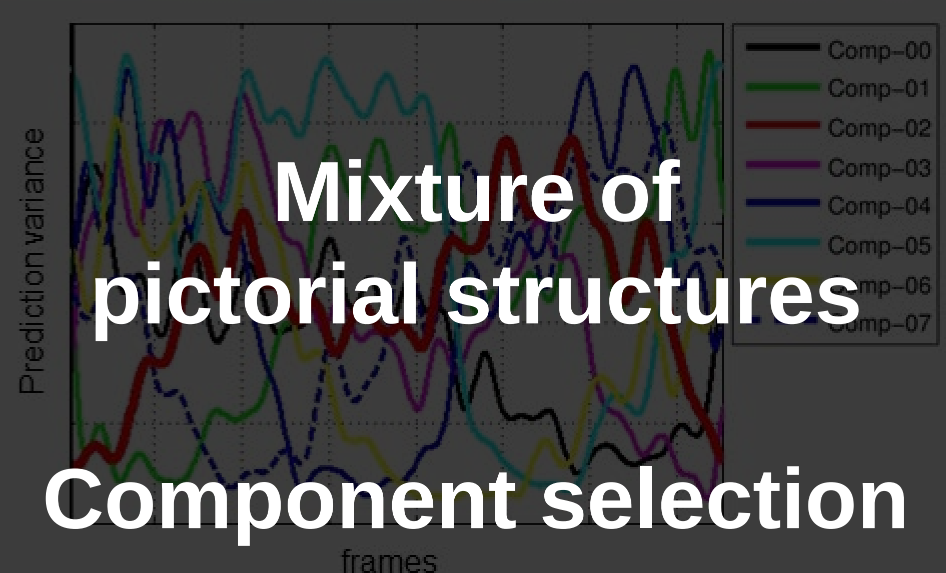
HumanEva-I

Sequence: Box (S2)

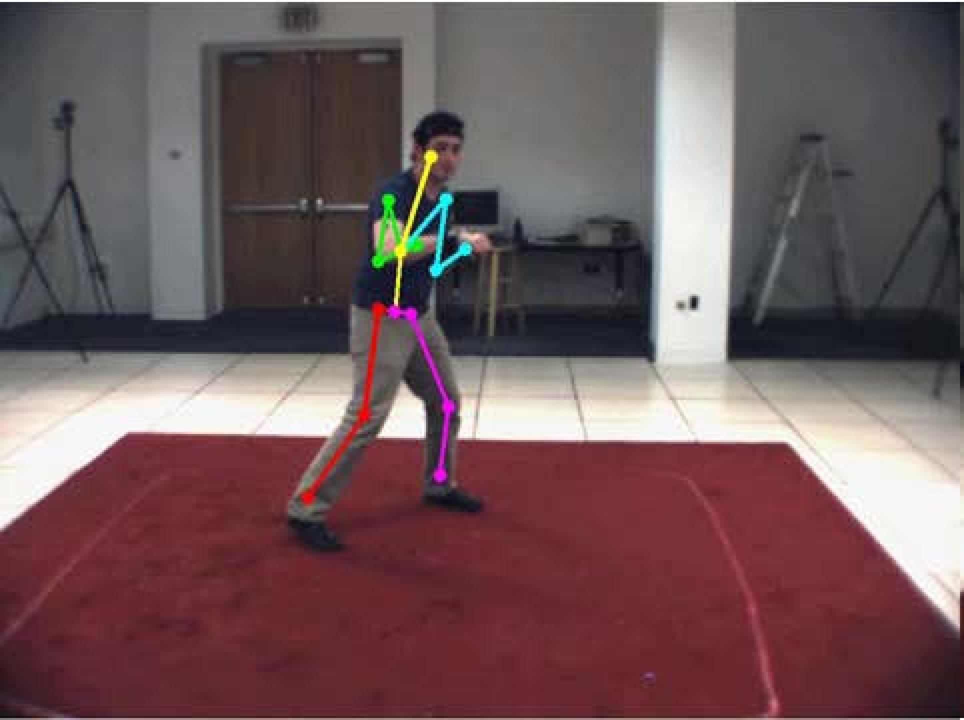


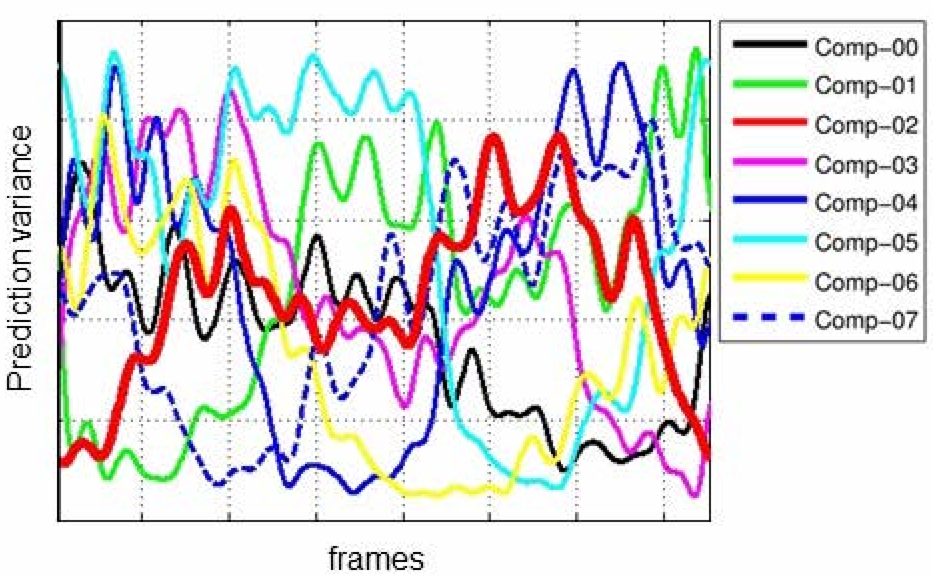






Component selection based on *min-variance*



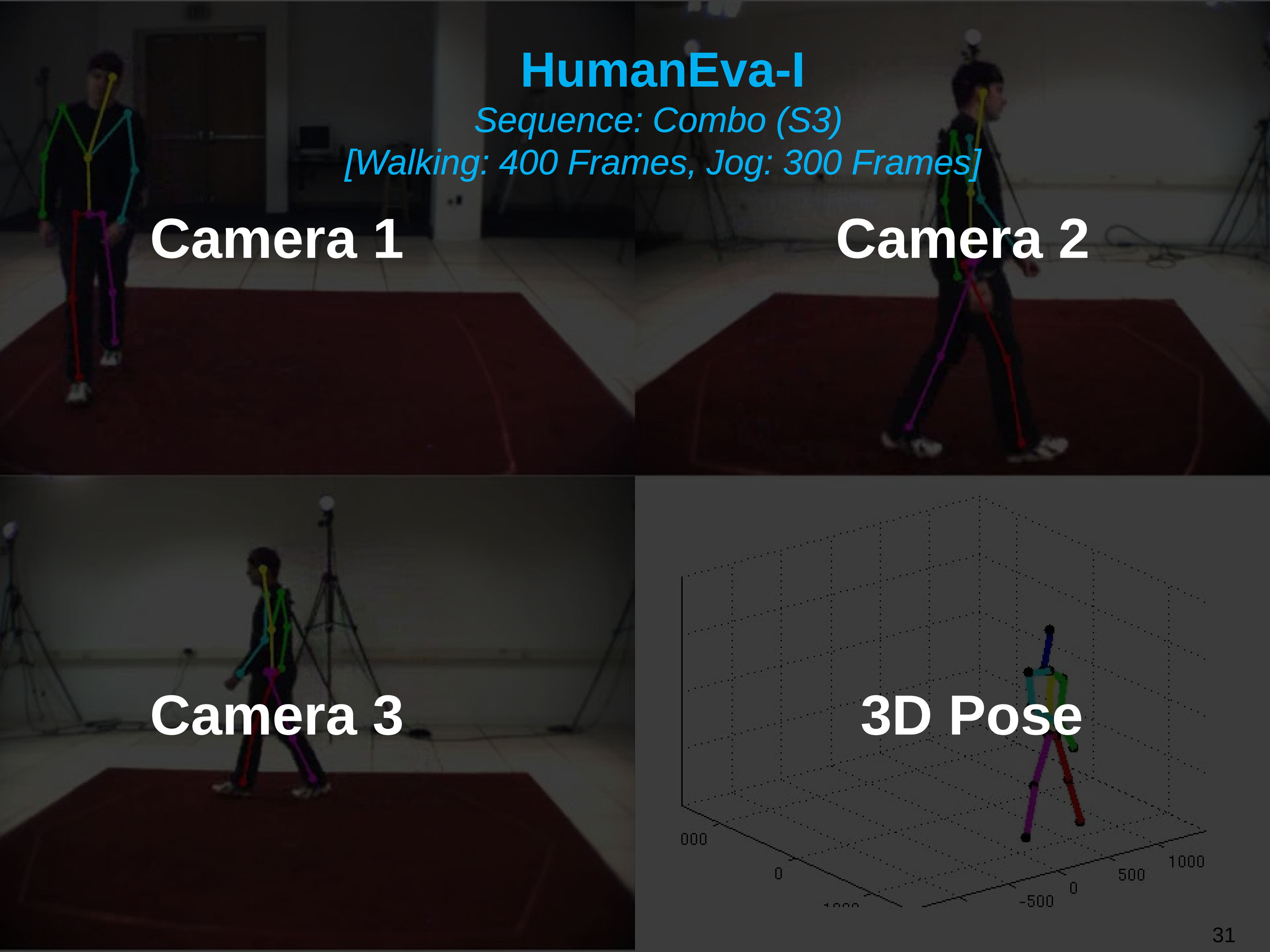


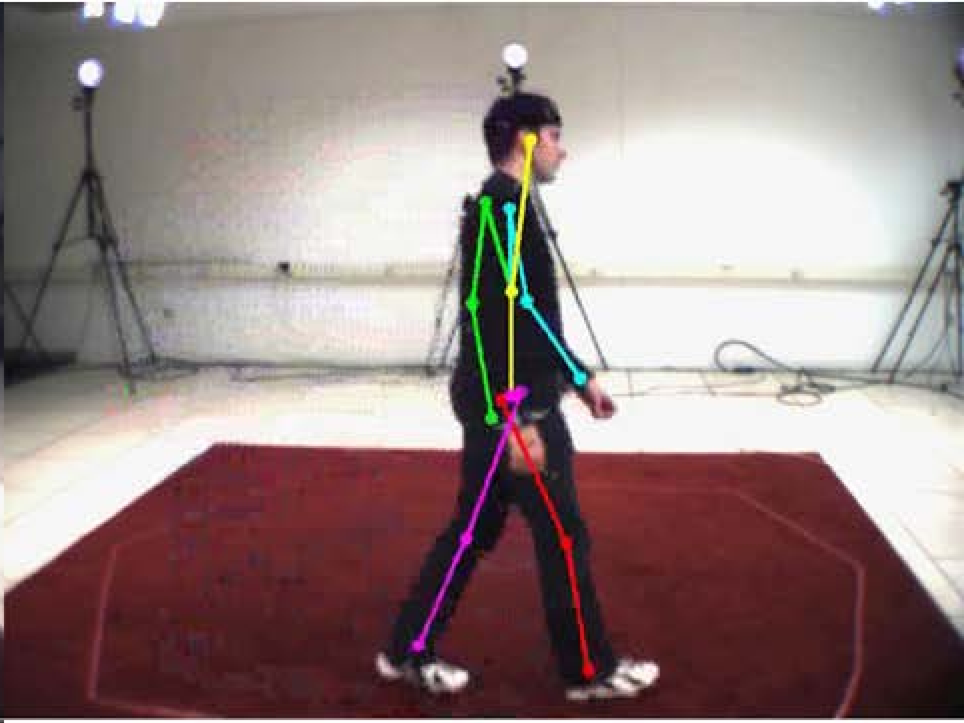
Component selection based on *min-variance*

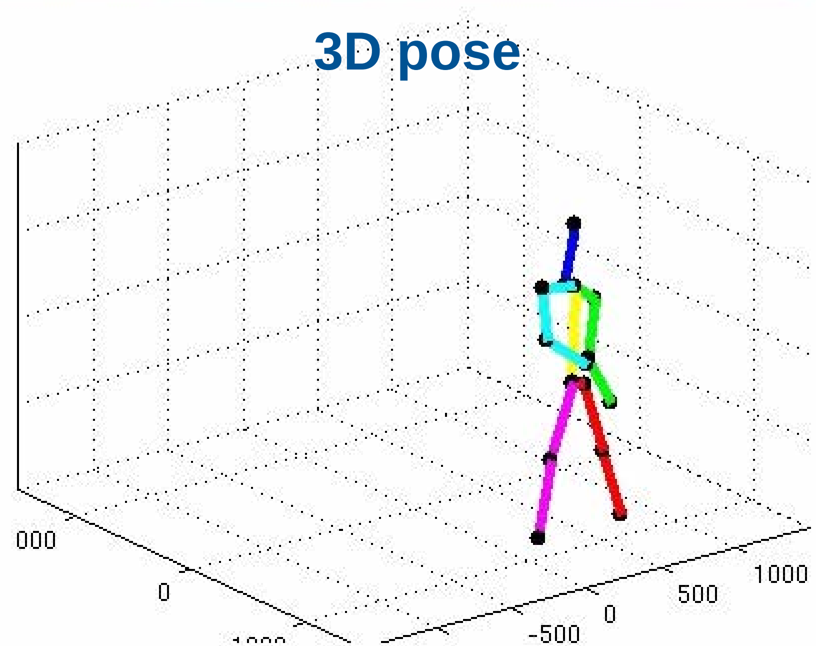
HumanEva-I

Sequence: Combo (S3)

[Walking: 400 Frames, Jog: 300 Frames]

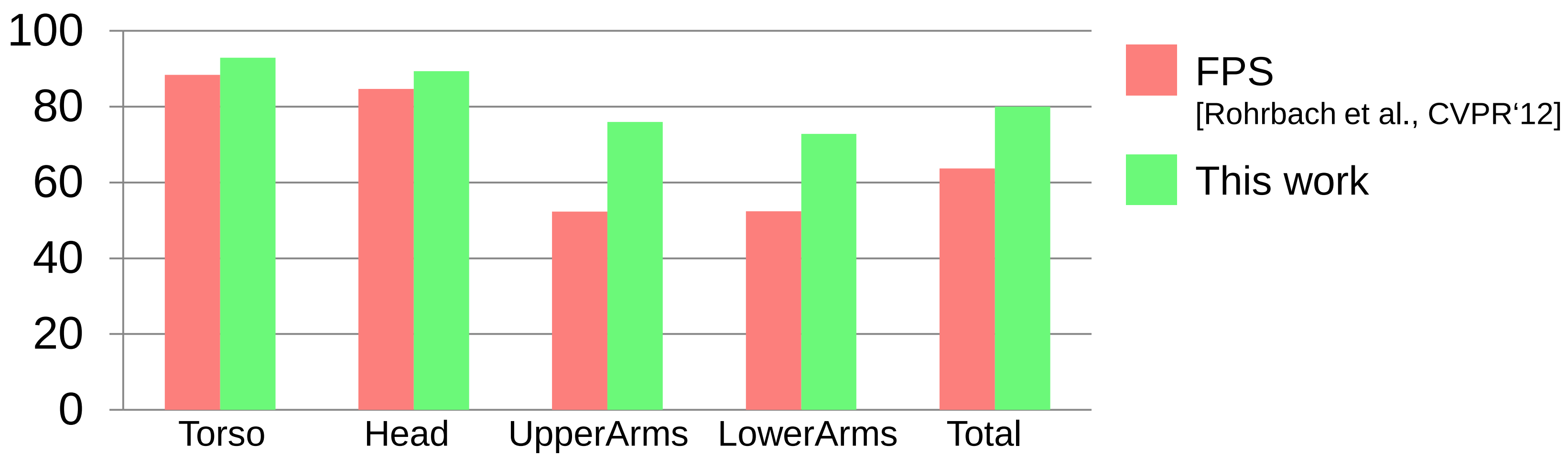





3D pose
000
0
1000
500
0
-500

MPII-Cooking Dataset

- 65 fine grained activities
- Disjoint training and test subjects
- Train set: 896 images x 2 views, 5 subjects
- Test set: 1154 images x 2 views, 7 subjects

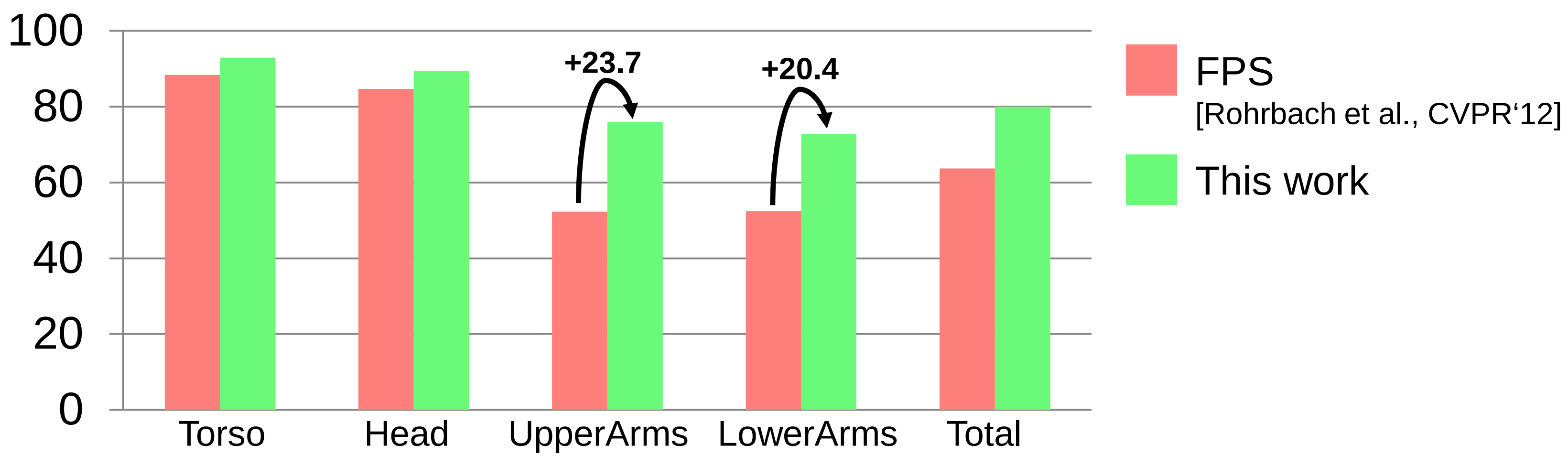


Quantitative comparison to state-of-the-art

percentage of correct parts (PCP)

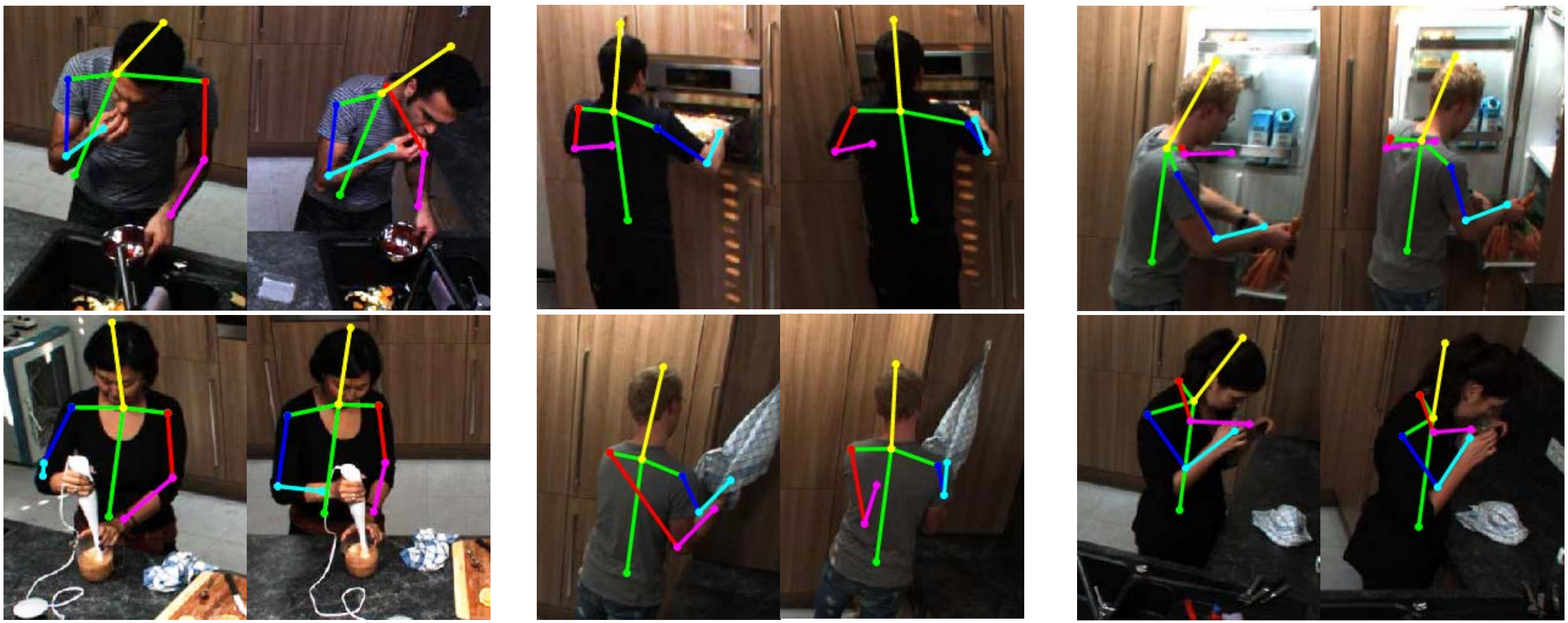
MPII-Cooking Dataset

- 65 fine grained activities
- Disjoint training and test subjects
- Train set: 896 images x 2 views, 5 subjects
- Test set: 1154 images x 2 views, 7 subjects



Quantitative comparison to state-of-the-art
percentage of correct parts (PCP)

MPII-Cooking Dataset



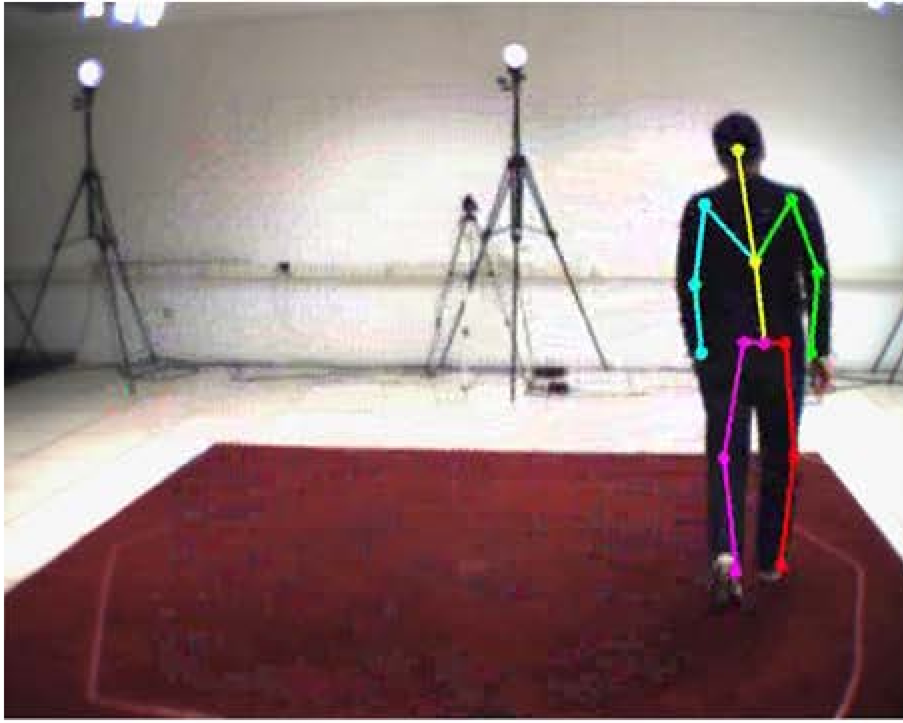
Frontal Back Side

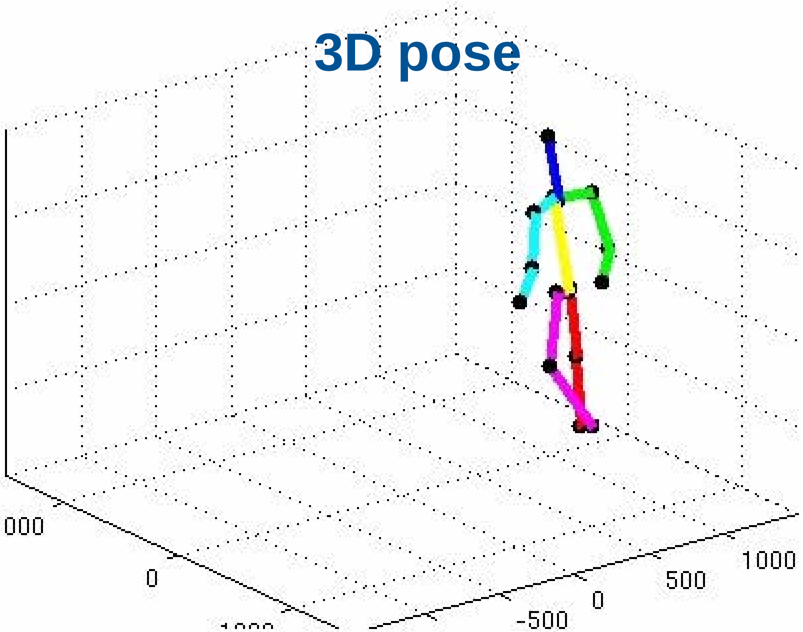
Summary

Multi-view pictorial structures

This work

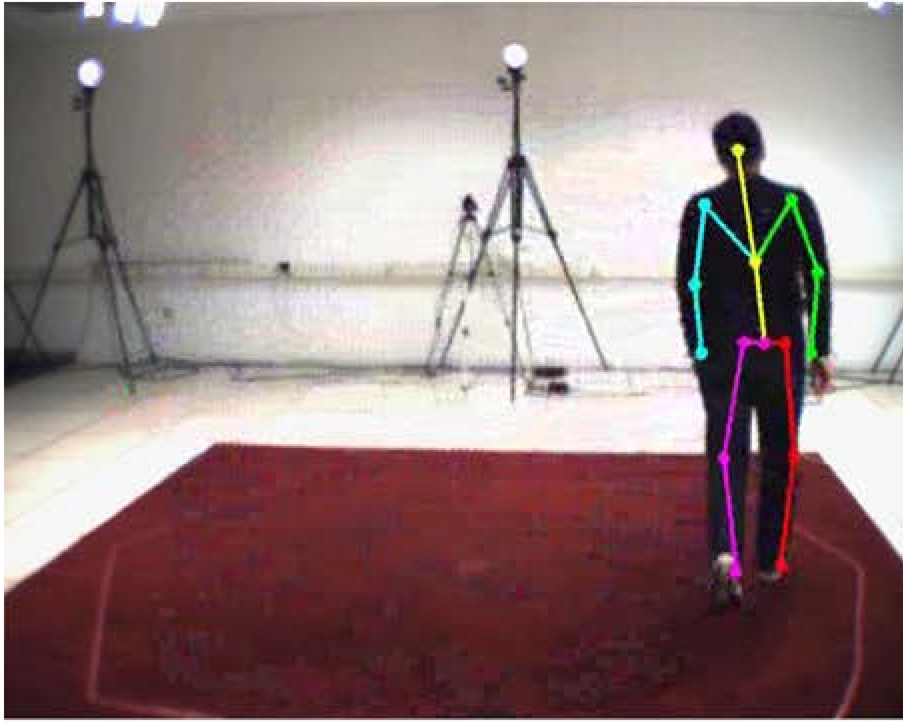
- Improved 2D pictorial structrues model
- Multi-view extension for 3D Human pose estimation
- Outperforms existing work significantly





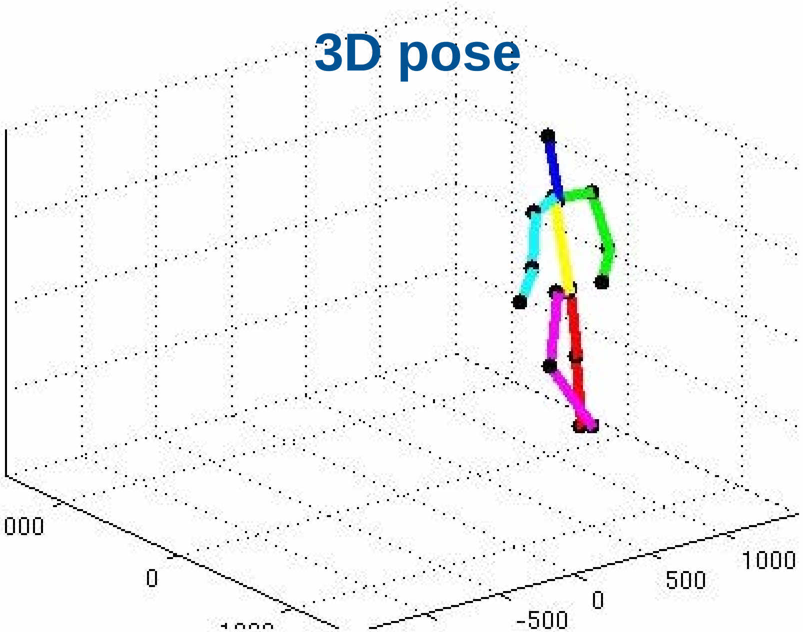
Summary

Multi-view pictorial structures



This work

- Improved 2D pictorial structrues model
- Multi-view extension for 3D Human pose estimation
- Outperforms existing work significantly



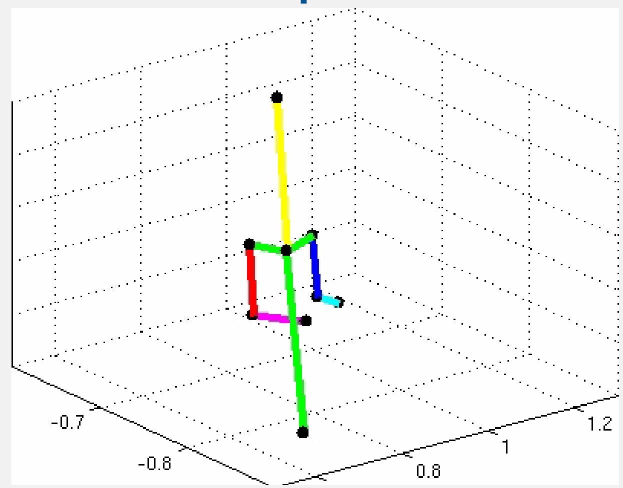
Future work

- Evaluate for non-static camera setup
- Multiple interacting subjects

MPII-Cooking Dataset available at:
http://www.d2.mpi-inf.mpg.de/mpii-cooking

2D estimates

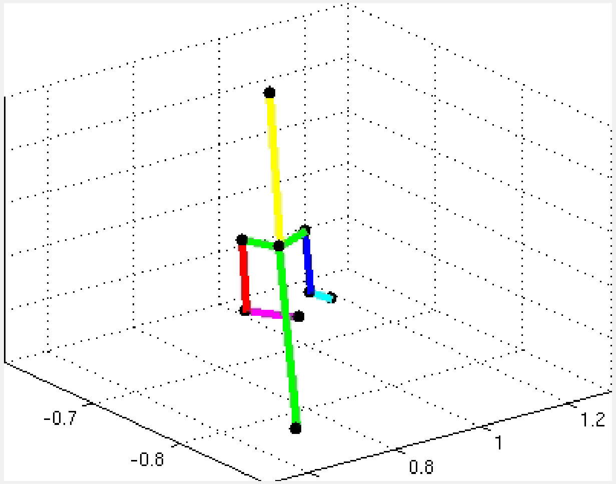
3D pose



MPII-Cooking Dataset available at:
http://www.d2.mpi-inf.mpg.de/mpii-cooking

2D estimates

3D pose



Thanks for your attention.