Deep Kinematic Pose Regression

Xingyi Zhou[1], Xiao Sun[2], Wei Zhang[1], Shuang Liang[3], Yichen Wei[2]

[1]Fudan University, [2]Microsoft Research, [3]Tongji University

[1]{zhouxy13, weizh}@fudan.edu.cn, [2]{xias, yichenw}@microsoft.com, [3]shuangliang@tongji.edu.cn

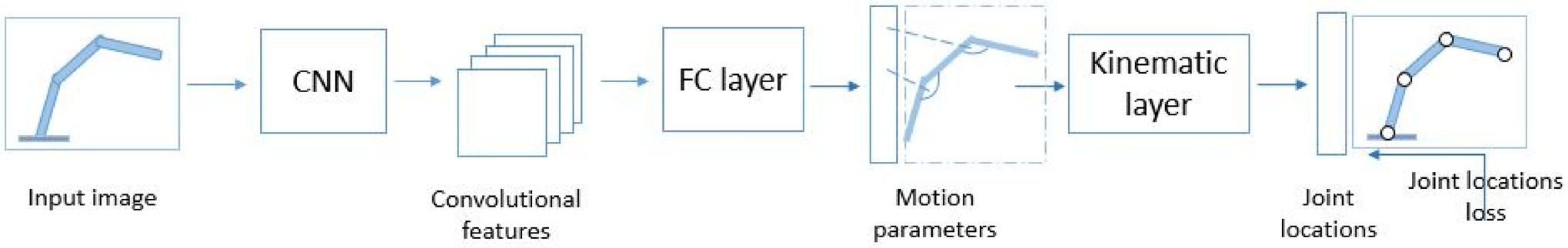


Figure 1: Illustration of our framework. The input image undergoes a convolutional neutral network and a fully connected layer to output model motion parameters (global position and rotation angles). The kinematic layer maps the motion parameters to joints. The joints are connected to ground truth joints to compute the joint loss that drives the network training.

Overview

Goal

Estimate object joint locations from a single image.

Pose Representation

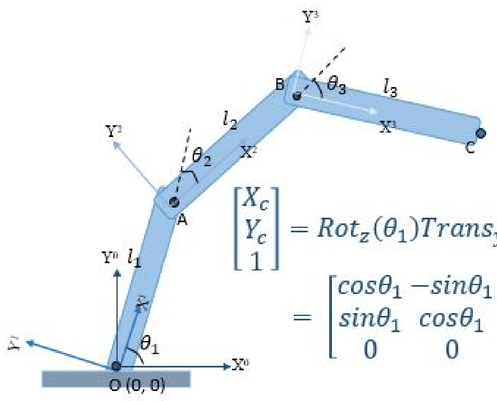
- Pictorial Structure Model
- Linear Dictionary
- Linear Feature Embedding
- Implicit Representation by Retrieval
- **Explicit Geometric Model**

Our Approach

We propose to directly embed a kinematic object model into the deep neutral network learning for general articulated object pose estimation [4].

Method

Kinematic Model



$$\begin{bmatrix} X_c \\ Y_c \\ 1 \end{bmatrix} = Rot_z(\theta_1)Trans_y(l_1)Rot_z(\theta_2)Trans_x(l_2)Rot_z(\theta_3)Trans_x(l_3)$$

$$= \begin{bmatrix} cos\theta_1 & -sin\theta_1 & 0 \\ sin\theta_1 & cos\theta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & l_1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos\theta_2 & -sin\theta_2 & 0 \\ sin\theta_2 & cos\theta_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & l_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos\theta_3 & -sin\theta_3 & 0 \\ sin\theta_3 & cos\theta_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & l_3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

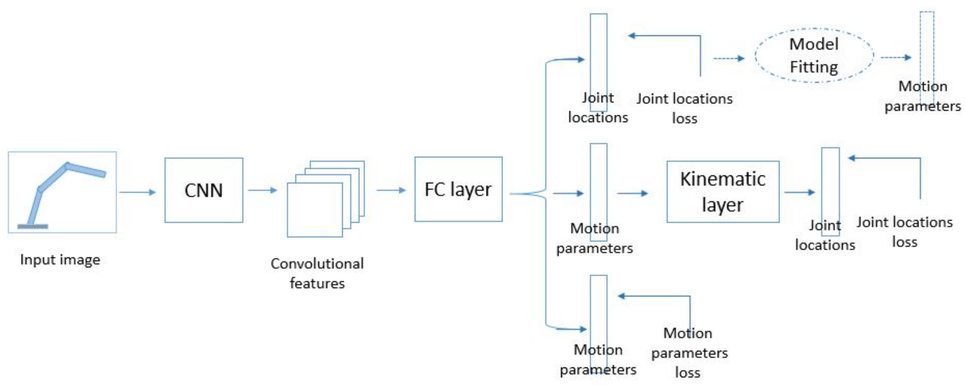
$$\Theta = \{\mathbf{p}, \mathbf{o}\} \cup \{\theta_i\}_{i=1}^{J} \quad (1)$$

$$p_u = (\prod_{v \in Pa(u)} Rot(\theta_v) \times Trans(l_v))\mathbf{O}^\top \quad (2)$$

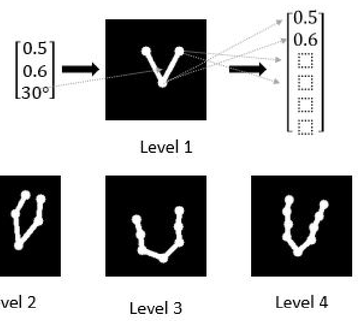
$$\mathcal{F} : \{\Theta\} \to \mathcal{Y} \quad (3)$$

$$L(\Theta) = \frac{1}{2}||\mathcal{F}(\Theta) - Y||^2 \quad (4)$$

Experiments



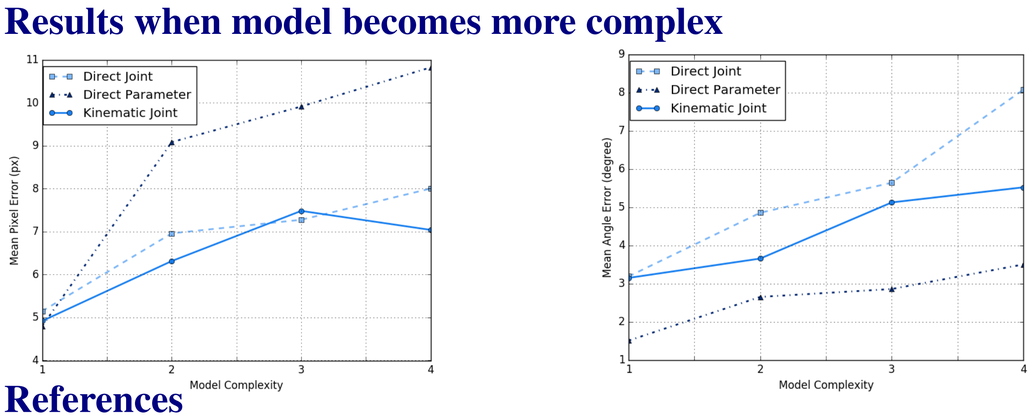
A Toy Example



Results

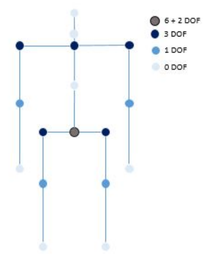
- All three methods achieve low pixel error.
- Bone length of **Direct joint** has large deviation.

Results when model becomes more complex



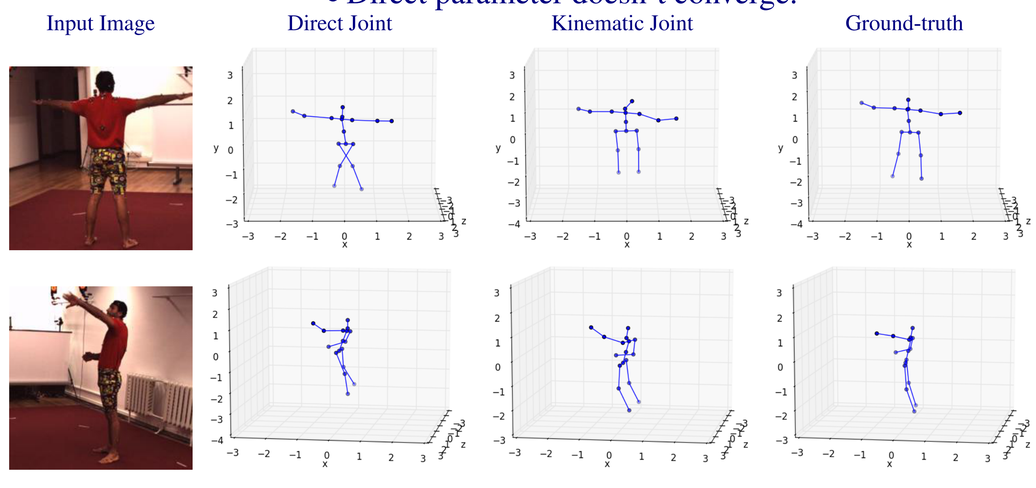
3D Human Pose Estimation

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
Li et al [1]	-	136.88	96.94	124.74	-	168.68	-	-
Tekin et al [2]	-	129.06	91.43	121.68	-	162.17	-	-
Zhou et al [3]	**87.36**	109.31	**87.05**	103.16	116.18	143.32	106.88	99.78
Ours(Direct)	106.38	104.68	104.28	107.80	115.44	**114.05**	103.80	109.03
Ours(ModelFit)	109.75	110.47	113.98	112.17	123.66	122.82	121.27	117.98
Ours(Kinematic)	91.83	**102.41**	96.95	**98.75**	**113.35**	125.22	**90.04**	**93.84**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkPair	Average
Li et al [1]	-	-	-	-	132.17	69.97	-	-
Tekin et al [2]	-	-	-	-	130.53	**65.75**	-	-
Zhou et al [3]	**124.52**	199.23	107.42	118.09	114.23	79.39	97.70	113.01
Ours(Direct)	125.87	**149.15**	112.64	105.37	**113.69**	98.19	110.17	112.03
Ours(ModelFit)	137.29	157.44	136.85	110.57	128.16	102.25	114.61	121.28
Ours(Kinematic)	132.16	158.97	**106.91**	**94.41**	126.04	79.02	98.96	**107.26**



Results

- Direct joint with ResNet50 already achieves state-of-the-art performance.
- Kinematic joint is slightly better then direct joint.
- Model fitting can not improve direct joint.
- Direct parameter doesn't converge.



References

[1] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *ICCV*, December 2015.

[2] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.

[3] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, June 2016.

[4] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. Model-based deep hand pose estimation. In *IJCAI*, 2016.

More Information
Visit Homepage!
https://goo.gl/WUC8ym