Learning from Synthetic Humans

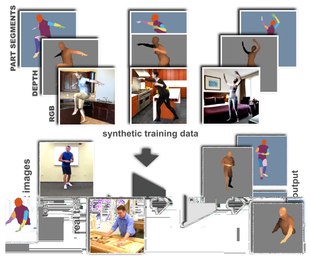
Gül Varol[1], Javier Romero[2], Xavier Martin[1], Naureen Mahmood[2], Michael J. Black[2], Ivan Laptev[1] and Cordelia Schmid[1]

[1]Inria, France [2]MPI for Intelligent Systems, Germany

IEEE 2017 Conference on Computer Vision and Pattern Recognition

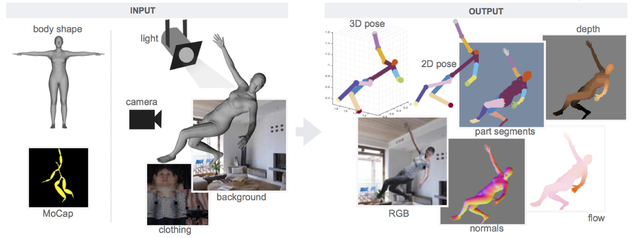
INTRODUCTION

- Goal & Contributions
  - Generating **synthetic** but **photo-realistic** videos of **people** for training CNNs.
  - Demonstrating advantages of this data for training human parts **segmentation** and **depth** estimation.
- Motivation
  - Manual labeling of 3D human attributes is impractical and pixel-level annotation is expensive.
  - Synthetic data comes with rich ground truth.



SURREAL DATASET: Synthetic hUmans foR REAL tasks

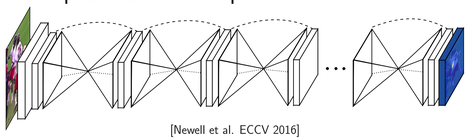
- Graphics pipeline for synthetic humans generation
  - **SMPL** body models posed with CMU **MoCap** and rendered on static background.
  - 1K clothings, 4K body shapes, 70K background images, random light and camera.
  - Ground truth segmentation, depth, optical flow, surface normals, 2D/3D pose.



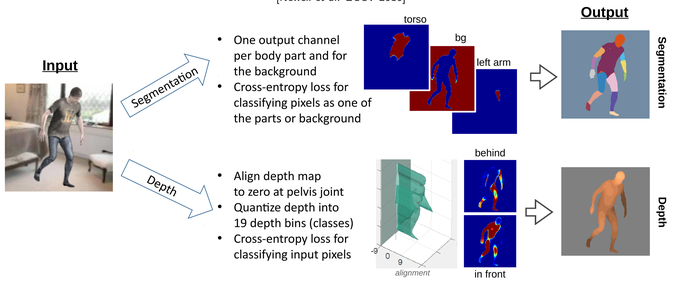
- Random samples from the **6M frames** in the SURREAL dataset

HUMAN PARTS SEGMENTATION AND DEPTH: TRAINING

- We build on the **stacked hourglass network** architecture introduced originally for 2D pose estimation problem.



[Newell et al. ECCV 2016]



- One output channel per body part and for the background
- Cross-entropy loss for classifying pixels as one of the parts or background

- Align depth map to zero at pelvis joint
- Quantize depth into 19 depth bins (classes)
- Cross-entropy loss for classifying input pixels

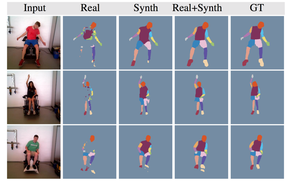
RESULTS: Synthetic

- Evaluation metrics
  - Segmentation: IOU, pixel accuracy
  - Depth: RMSE, st-RMSE

Segmentation	
IOU	69.13 %
Accuracy	80.61 %

Depth	
RMSE	72.9 mm
st-RMSE	56.3 mm

RESULTS: Freiburg Sitting People [Oliveira et al. ICRA 2016]



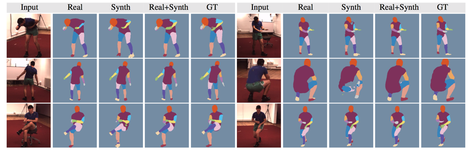
- Segmentation

Training data	Head IOU	Torso IOU	$Legs_{up}$ IOU	mean IOU	mean Acc.
[Oliveira 2016]	-	-	-	64.10	81.78
Real	58.44	24.92	30.15	28.77	38.02
Synth	73.20	65.55	39.41	40.10	51.88
Synth+Real	72.88	80.76	65.41	59.58	78.14
Synth+Real+up	**85.09**	**87.91**	**77.00**	**68.84**	**83.37**

RESULTS: Human3.6M [Ionescu et al. ICCV 2011]

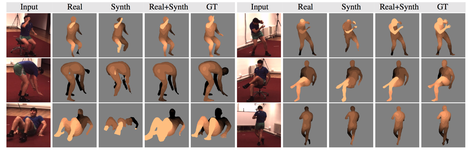
- Segmentation

	IOU		Accuracy	
Training data	fg+bg	fg	fg+bg	fg
Real	49.61	46.32	58.54	55.69
Synth	46.35	42.91	56.51	53.55
Synth+Real	**57.07**	**54.30**	**67.72**	**65.53**

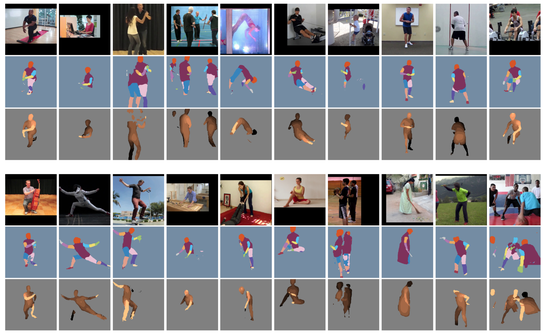


- Depth

	All pixels		Joints	
Training data	RMSE	st-RMSE	RMSE	st-RMSE
Real	96.3	75.2	122.6	94.5
Synth	111.6	98.1	152.5	131.5
Synth+Real	**90.0**	**67.1**	**92.9**	**82.8**

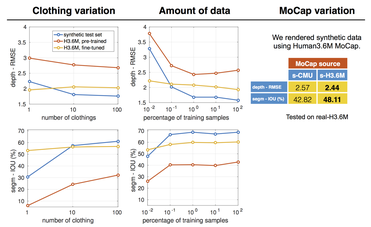


QUALITATIVE: MPII Human Pose [Andriluka et al. CVPR 2014]



- Challenging cases
  - Multi-person
  - Occlusion
  - Object interaction
  - Extreme poses
  - Clothing

DESIGN CHOICES



CONCLUSIONS

- It is possible to learn from synthetic images of people.
- CNNs trained on synthetic people can generalize.

Code & Data available!

gul.varol@inria.fr