SHPD: SURVEILLANCE HUMAN POSE DATASET AND PERFORMANCE EVALUATION FOR COARSE-GRAINED POSE ESTIMATION

Qiuhui Chen[1], Chongyang Zhang[1,2], Weiwei Liu[1], and Dan Wang[1]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China [2]Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai 200240, China

INTRODUCTION

Motivation:

Pose estimation is highly valued in surveillance systems in the era of big data. However, current human pose datasets are limited in their coverage of the pose estimation challenges in outdoor surveillance scenarios. In this paper, we introduce a novel Surveillance Human Pose Dataset (SHPD).

Main Contributions:

- a more specialized human pose benchmark for surveillance tasks;
- proposing the concept of coarse-grained global pose estimation used for the pose recognition of small scale human targets in many practical surveillance applications;
- giving performance evaluation of global-pose estimation using four widely adopted baseline deep-learning networks.

DISTRIBUTION

train	test
18447	5000

bending	riding	falling	jumping	lying	running	sitting	squat	standing	walking
2598	2022	2056	2036	2024	2184	2557	2926	2026	3018

Table. 1. Numbers of train and test set and 10 pose categorizations of SHPD. Walking was the most frequently observed, with squat being the second most frequent in the dataset. Each pose categorization has over 2000 sample images.

Datasets	human height
MPII	257-690 pixels
MSCOCO	225-445 pixels
SHPD	**74-302 pixels**

Table. 2. human height distribution in centre 70% interval on MPII, MSCOCO, and SHPD.

THE SURVEILLANCE HUMAN POSE DATASET

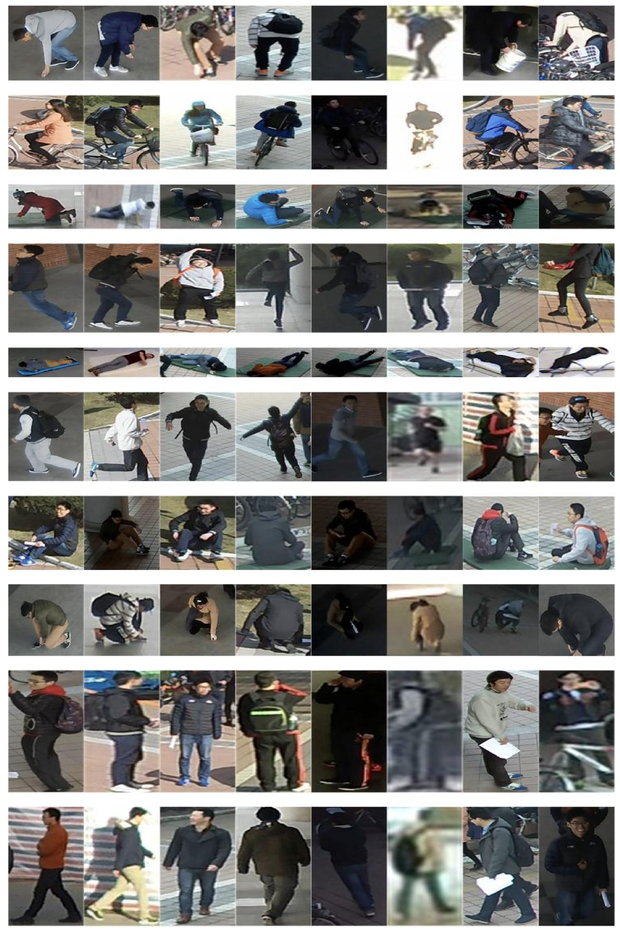
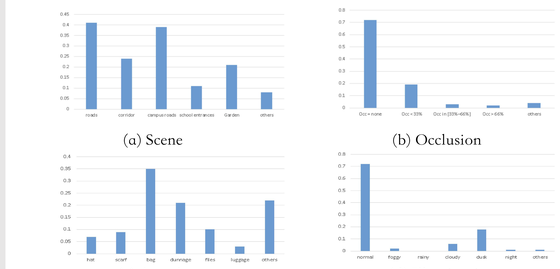


Fig. 2. Examples of diverse human poses in SHPD.
Ten rows show ten pose categories. From top to down: bending, riding, falling, jumping, lying, running, siting, squatting, standing and walking. The first four columns show the various view of human (left side, right side, front side and back side). The 5th column shows poses captured during night and illumination is bad. The 6th column shows the low resolution poses. The 7th column shows poses with various attachments. The last column shows poses with occlusion or truncation.

COLLECTION

- collected from on-using monitoring cameras.
- city roads, or beside the squares and highways.
- Most cameras are 3-5 meters high above the ground.
- 297 pieces of videos, totally about 61 hours long, images are extracted from the video pieces at intervals then human objects are selected, labeled.

PROPERTIES



(a) Scene (b) Occlusion

(c) Attachment (d) Illumination

Fig. 3. Statistics of different properties in SHPD. 6 typical scenes are included. Occlusion levels are divided by the ratio of occluded parts to the whole human body. Occluded ratio is grouped into five levels: Occ = none, Occ <33%, Occ in [33%, 66%], Occ >66%, others. Common attachments contain hat, files, scarf, bag, dunnage, luggage and others. 7 illumination situations are included: normal, foggy, rainy, cloudy, dusk, night and other.

EVALUATION

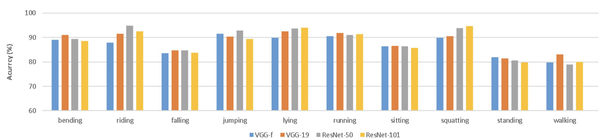


Fig. 4. Global-pose estimation performance of all models, per pose category

Models	Mean accuracy
VGG-f	86.96%
VGG-19	88.30%
ResNet-50	**88.56%**
ResNet-101	87.94%

Table. 3. Mean accuracy for all pose categories

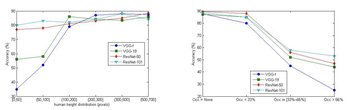


Fig. 5. Comparison of performance on scale and occlusion

DISCUSSION

- Pose. The best performance is achieved on pose categories with attachments(e.g. riding and lying). The poor performance is achieved on pose categories with slightly foreshortened torso(e.g. falling and sitting) or small interclass similarity(e.g. standing and walking).
- Scale. The presence or absence of small-scale has relatively large influence on the result. The best performing model ResNet-101 is the most robust to small-scale human.
- Occlusion and truncation. The performance is best for fully visible people. Truncation showed the least influence overall among the discussed factors because the number of images with truncation is limited in our dataset (about 10% of the test data).

Summary:

- 1) For small-scale human, performance degrades catastrophically.
- 2) Similarity of global structure features among Inter-classes increases the difficulties of identification.
- 3) Performance drops fast under heavy occlusion situations.

REFERENCES

[1] Shih En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in CVPR, 2016.

[2] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in ECCV, 2016.

[3] Wenjuan Gong, Xuena Zhang, Changhe Tu, and Elhadi Zahzah, "Human pose estimation from monocular images: A comprehensive survey," in Sensors, 2016.

[4] Xiao Chu, Wei Yang, Wanli Ouyang, and Xiaogang Wang, "Multi-context attentionfor human pose estimation," in CVPR, 2017.

[5] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in CVPR, 2017.

[6] George Papandreou, Tyler Zhu, Jonathan Tompson, and Kevin Murphy, "Towards accurate multi-person pose estimation in the wild," in CVPR, 2017.

To view the full dataset visit http://ivlab.sjtu.edu.cn/Data/List/Resources

Corresponding Author: Chongyang Zhang (sunny_zhang@sjtu.edu.cn)