Supplementary Material of “3D Local Convolutional Neural Networks for Gait Recognition”

Zhen Huang[1, 2]*, Dixiu Xue[2], Xu Shen[2], Xinmei Tian[1]†,
Houqiang Li[1], Jianqiang Huang[2], Xian-Sheng Hua[2]†
[1]University of Science and Technology of China, [2]Alibaba Group
hz13@mail.ustc.edu.cn, {xinmei,lihq}@ustc.edu.cn,
{dixiu.xdx,shenxu.sx,jianqiang.hjq,xiansheng.hxs}@alibaba-inc.com

1. Localization Module

Our 3D local convolutional networks (3D local CNN) can incorporate part-specific sequential information. Inside the novel building block, a localization module is designed to dynamically localize the local 3D volumes in a sequence with adaptive spatial and temporal scales, locations and lengths.

Specifically, we utilize a block with convolution, ReLU, batch normalization, max pooling and fully connected layers as the localization module. The detailed architecture of the localization module is presented in Table 1. The output of the localization module is a set of local volumes whose locations are denoted by 8 parameters $(\Delta_x, \Delta_y, \Delta_t, \delta_x, \delta_y, \delta_t, \sigma^2, \gamma)$.

Moreover, examples of part localization in more views (18° to 72°) are illustrated in Fig. 1. The colorful patches denote the localized 3D volumes for each human body part. Different 3D volumes have different spatial positions, scales and temporal lengths. The result show that given the predefined center (c_x, c_y, c_t) of the part, the localization module is able to precisely adjust the corresponding volume with respect to each view.

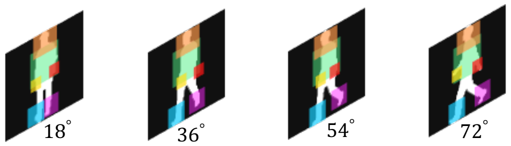
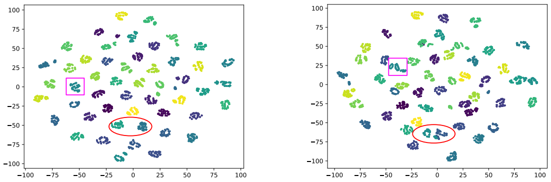


Figure 1. The result of the localization module in views of 18° to 72° (better viewed in color). The colorful patches denote the localized 3D volumes for each human body part. Different 3D volumes have different spatial positions, scales and temporal lengths. The result show that the localization module is able to precisely adjust the corresponding volume of local part with respect to each view.



(a) 3D local CNN (b) GaitPart

Figure 2. The tSNE visualizations of 3D local CNN and GaitPart on CASIA-B (better viewed in color). Each sequence is visualized as one point, and the colors of the points denote different subjects. The 3D local CNN points inside the magenta rectangle are more compact and the clusters inside the red ellipse are also more distant from each other. These demonstrate that the representations generated by 3D local CNN are semantically better grouped and more discriminative than GaitPart.

2. Visualization Analysis

Fig. 2 shows the t-SNE visualizations of the features produced by our method and GaitPart [3] on CASIA-B. Each sequence is visualized as one point, and the colors denote different subjects. In the Fig. 2(a), the points inside the magenta rectangle are more compact than that of the Fig. 2(b). This shows that the representations generated by 3D local CNN are semantically better grouped than GaitPart [3]. The clusters inside the red ellipse of 3D local CNN are also more distant from each other, meaning that the representations generated by 3D local CNN are more discriminative than GaitPart [3].

3. Results on More Challenging Protocol

In the main manuscript, we have demonstrated that our 3D local CNN outperforms other methods in the most challenging scenario (CL) with a large margin (exceeding GaitSet [2] by 13.0%, GaitPart [3] by 4.7%, GLN [4] by 6.8%, MT3D [5] by 1.9%). To further verify this, we conduct additional experiments on small training (ST) protocol, *i.e.*,

*This work was done when the author was visiting Alibaba as a research intern.

†Corresponding author.

Table 1. Architecture of the localization module. N is the batch size. C, T, H, and W denote the number of channels, frames, height and width of the input feature maps, respectively. All convolution and pooling layers are 3D operation. The output of the localization module is a set of local volumes whose locations denoted by 8 parameters $(\Delta_x, \Delta_y, \Delta_t, \delta_x, \delta_y, \delta_t, \sigma^2, \gamma)$.

Layer	Kernel	Stride	Pad	#Filters	Output
Input	-	-	-	-	$N \times C \times T \times H \times W$
Pool0	$3 \times 3 \times 3$	2	1	-	$N \times C \times T/2 \times H/2 \times W/2$
Conv1	$1 \times 1 \times 1$	1	0	$C/4$	$N \times C/4 \times T/2 \times H/2 \times W/2$
Conv2	$3 \times 3 \times 3$	1	1	$C/4$	$N \times C/4 \times T/2 \times H/2 \times W/2$
Pool3	$3 \times 3 \times 3$	2	1	-	$N \times C/4 \times T/4 \times T/4 \times W/4$
Conv4	$3 \times 3 \times 3$	1	1	$C/4$	$N \times C/4 \times T/4 \times H/4 \times W/4$
Pool5	$3 \times 3 \times 3$	2	1	-	$N \times C/4 \times T/8 \times T/8 \times W/8$
Fc6	-	-	-	-	$N \times 8$

Table 2. Averaged rank-1 accuracy on **CASIA-B (ST)**, identical views cases excluded. For GaitPart [3] and GLN [4], we use our replementation based on the paper. For GaitSet [2] and MT3D [5], we use the numbers from the original papers. The probe sequences are divided into three subsets (NM, BG and CL) according to the walking conditions. The ST protocol is defined by small-sample training (ST), which means that the model is trained by using only 24 subjects, while the rest 100 subjects are used to test.

Gallery NM #1-4		0° – 180°											Mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet	64.6	83.3	90.4	86.5	80.2	75.5	80.3	86.0	87.1	81.4	59.6	79.5
	GaitPart	66.5	85.2	91.7	88.1	81.8	76.8	83.6	86.6	88.9	83.1	61.3	81.2
	GLN	66.5	85.9	92.7	89.2	83.4	77.6	84.2	88.2	89.6	83.8	61.8	82.1
	MT3D	71.9	83.9	90.9	90.1	81.1	75.6	82.1	89.0	91.1	86.3	69.2	82.8
	3DLocal	**71.2**	**86.2**	**93.2**	**91.4**	**83.5**	**77.9**	**84.4**	**90.3**	**92.4**	**86.6**	**68.4**	**84.1**
BG #1-2	GaitSet	55.8	70.5	76.8	75.5	69.7	63.4	68.0	75.8	76.2	70.7	52.5	68.6
	GaitPart	58.9	73.4	79.7	78.4	72.8	66.3	70.5	78.7	79.5	73.6	55.0	71.5
	GLN	60.3	74.7	81.1	79.9	74.3	67.7	71.9	80.1	80.7	75.2	56.3	72.9
	MT3D	64.5	76.7	82.8	82.8	73.2	66.9	74.0	81.9	84.8	80.2	63.0	74.0
	3DLocal	**66.4**	**78.8**	**84.9**	**84.5**	**75.5**	**68.8**	**75.8**	**83.8**	**86.9**	**82.4**	**65.0**	**76.0**
CL #1-2	GaitSet	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
	GaitPart	34.4	47.7	54.3	53.6	47.1	45.5	49.6	51.9	47.9	40.4	30.3	45.7
	GLN	36.3	49.4	56.3	56.5	49.0	47.1	51.2	53.5	50.1	42.2	33.1	47.5
	MT3D	46.6	61.6	66.5	63.3	57.4	52.1	58.1	58.9	58.5	57.4	41.9	56.6
	3DLocal	**51.0**	**65.6**	**70.5**	**67.3**	**61.0**	**56.1**	**62.0**	**63.3**	**62.9**	**61.5**	**45.9**	**60.6**

the model is trained by using only 24 subjects, while the rest 100 subjects are used for test. Compared to the regular settings in Table 1 of the main manuscript, ST protocol is more challenging because there are fewer subjects in the training set (24 vs. 74) and more subjects in the testing set (100 vs. 50). As shown in Table 2, 3D local CNN again outperforms other methods with a significant margin, exceeding GaitSet [2] by 19.7%, GaitPart [3] by 14.9%, GLN [4] by 13.1% and MT3D [5] by 4.0%. More importantly, the improvement of our 3D local CNN is more significant than that in the main manuscript, indicating that our method becomes more superior than other methods when the protocol is more challenging.

This phenomenon confirms that the adaptive local volume sampling and processing mechanism is more powerful at handling large appearance changes of human body.

Table 3. Prior settings of the head, left-arm, right-arm, torso, left-leg and right-leg. p_H, p_W and p_T are the proportions of the height, width and length, respectively. c_x, c_y and c_z are the prior center of the sampling grid.

Path	p_H	p_W	p_T	c_x	c_y	c_t
Head	1/8	3/11	1/3	1/2	1/16	1/2
Left arm	3/16	3/11	2/3	3/10	1/2	1/2
Right arm	3/16	3/11	2/3	7/10	1/2	1/2
Torso	1/2	3/11	1/3	1/2	3/8	1/2
Left leg	3/8	5/11	2/3	3/10	4/5	1/2
Right leg	3/8	5/11	2/3	7/10	4/5	1/2

4. Prior Knowledge

For feature learning of gait recognition, it is quite natural to define six local paths corresponding to the head, left-

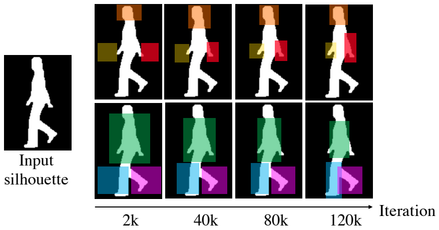


Figure 3. Localized body parts along with the training process (size for scale and center for location). Left: the input silhouette. Right: the focused patch of head, left arm, right arm, torso, left leg and right leg at iteration 2k, 40k, 80k and 120k, respectively.

arm, right-arm, torso, left-leg and right-leg. Following [1] and common sense knowledge, the general (height, width, length) proportions (p_H, p_W, p_L) and the prior centers of the sampling grid (c_x, c_y, c_t) of the head, left-arm, right-arm, torso, left-leg right-leg of the human body are summarized in Table 3.

Except the prior position of part center ($\{c_x, c_y, c_t\}$), our local operations learn in a totally unsupervised manner ($\{\Delta_x, \Delta_y, \Delta_t, \delta_x, \delta_y, \delta_t, \gamma\}$). As the location, scale and confidence are automatically determined for each input by our localization module, the prior position is only used to roughly initialize the focus of each part with little overlaps and cover the input for better training efficiency. Fig. 3 shows that starting with very coarse and inaccurate prior position (left arm, left leg, *etc.*), the localization module still learns to focus on each part in a suitable location and scale. In this paper, the prior center of each part is just the common proportion of the human body (*e.g.* $\{c_x = 1/2, c_y = 1/16\}$ for head, $c_t = 1/2$ for middle frame), which does not add any extra costs. When this work is applied to other tasks, it is easy to replace this proportion prior information in human body with coarse prior information in the new tasks. This prior information can also be determined by common knowledge or automatic pose/saliency detection.

5. Discussion

More speicifcally, we regard the limbs on the left/right side of the image as left/right limbs. In the feature fusion module (FS), the outputs of global and all local branches will be merged by a $1 \times 1 \times 1$ convolutional layer. Random horizontal flipping is applied to the input sequence during training, which helps the output feature of FS to be robust to left/right exchange (*e.g.* left arm in $0°$ vs. right arm in $180°$). Results in Table 2 (L648-668) indicate that this exchange brings in about 0.5% performance gap (*e.g.* $0°$ vs. $180°$, $15°$ vs. $195°$).

References

[1] Barry Bogin and Maria Inês Varela-Silva. Leg length, body proportion, and health: a review with a note on beauty. *International journal of environmental research and public health*, 7(3):1047–1075, 2010. 3

[2] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019. 1, 2

[3] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1, 2

[4] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398. Springer International Publishing, 2020. 1, 2

[5] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM MM*, pages 3054–3062, 2020. 1, 2