CMPE 537 Computer Vision

TERM PROJECT

Gait Recognition via Disentangled Representation Learning

—

İpek Erdoğan

What is the problem?

For sign language recognition tasks: the result gets affected from the identity attributes of signers.

“Focusing on the wrong side”

Solution: Learning representations in more specific ways for specific tasks.

Özdemir, Oğulcan & Kındıroğlu, Ahmet & Camgoz, Necati & Akarun, Lale. (2020). BosphorusSign22k Sign Language Recognition Dataset.

Literature

Disentangled Representation Learning GAN for Pose-Invariant Face Recognition[2]

Multi-task Adversarial Network for Disentangled Feature Learning [6]

Disentangling Latent Hands for Image Synthesis and Pose Estimation [4]

Recognize Actions by Disentangling Components of Dynamics [5]

Disentangled Representation Learning for 3D Face Shape [3]

Unsupervised Domain Adaptation by Backpropagation [7]

Learning signer-invariant representations with adversarial training [8]

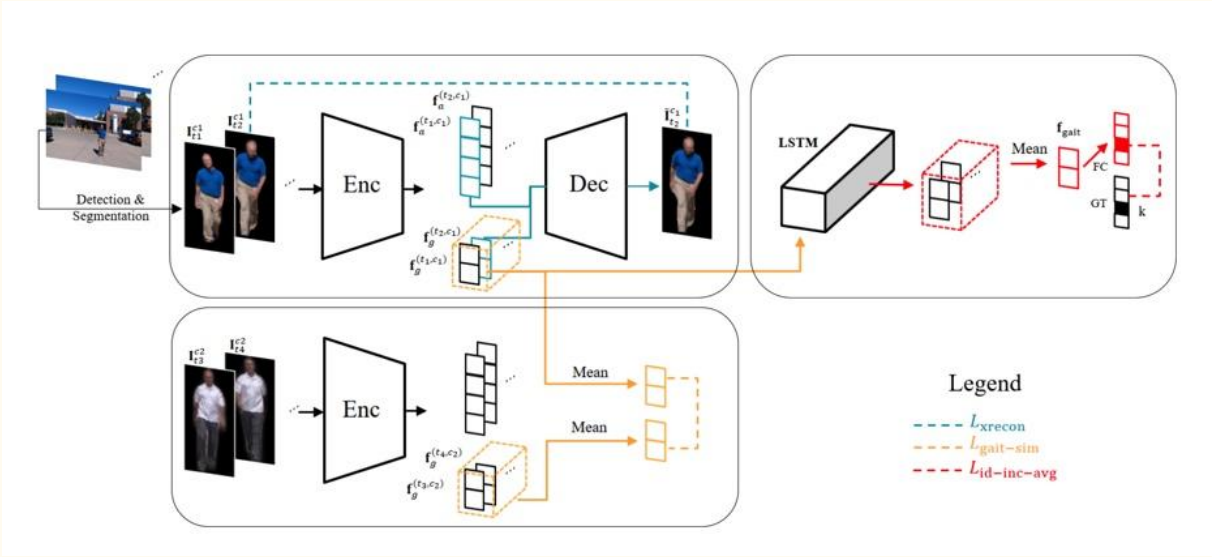
Proposed Approach

$$\mathcal{L} = \mathcal{L}_{\text{id-inc-avg}} + \lambda_r \mathcal{L}_{\text{xrecon}} + \lambda_s \mathcal{L}_{\text{gait-sim}}.$$

$$\mathcal{L}_{\text{gait-sim}} = \left\| \frac{1}{n_1} \sum_{t=1}^{n_1} \mathbf{f}_g^{(t,c_1)} - \frac{1}{n_2} \sum_{t=1}^{n_2} \mathbf{f}_g^{(t,c_2)} \right\|_2^2.$$

$$\mathcal{L}_{\text{id-inc-avg}} = \frac{1}{n} \sum_{t=1}^{n} -w_t \log \left(C_k \left(\frac{1}{t} \sum_{s=1}^{t} \mathbf{h}^s \right) \right)$$

$$\mathcal{L}_{\text{xrecon}} = \left\| \mathcal{D}(\mathbf{f}_a^{t_1}, \mathbf{f}_g^{t_2}) - \mathbf{I}_{t_2} \right\|_2^2,$$



Paper's Contribution and Critics: Novel Loss Functions

$$\max_D V_D(D,G) = E_{\mathbf{x}\sim p_d(\mathbf{x})}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))],$$
$$\max_G V_G(D,G) = E_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(D(G(\mathbf{z}))].$$

from DR-GAN[2]

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i y_i \log(\hat{y}_i)$$

Cross Entropy Loss

$$\mathcal{L}_{\text{xrecon}} = \left\| \mathcal{D}(\mathbf{f}_a^{t_1}, \mathbf{f}_g^{t_2}) - \mathbf{I}_{t_2} \right\|_2^2,$$

$$\mathcal{L}_{\text{gait-sim}} = \left\| \frac{1}{n_1} \sum_{t=1}^{n_1} \mathbf{f}_g^{(t,c_1)} - \frac{1}{n_2} \sum_{t=1}^{n_2} \mathbf{f}_g^{(t,c_2)} \right\|_2^2.$$

$$\mathcal{L}_{\text{id-inc-avg}} = \frac{1}{n} \sum_{t=1}^{n} -w_t \log\left(C_k \left(\frac{1}{t} \sum_{s=1}^{t} \mathbf{h}^s \right) \right)$$

Paper's Contribution and Critics: Results

Results in Gait Recognition task (This model is successful at the hardest tasks (0* and 180*), in this degrees view is almost full of appearance.)

Methods	0°	18°	36°	54°	72°	108°	126°	144°	162°	180°	Average
CPM [12]	13	14	17	27	62	65	22	20	15	10	24.1
GEI-SVR [29]	16	22	35	63	95	95	65	38	20	13	42.0
CMCC [28]	18	24	41	66	96	95	68	41	21	13	43.9
ViDP [26]	8	12	45	80	**100**	**100**	81	50	15	8	45.4
STIP+NN [30]	–	–	–	–	84.0	86.4	–	–	–	–	–
LB [46]	18	36	67.5	93	99.5	99.5	92	66	36	18	56.9
L-CRF [12]	38	**75**	68	93	98	99	**93**	67	76	39	67.8
GaitNet (ours)	**68**	74	**88**	91	99	98	84	**75**	76	**65**	**81.8**

(Rank 1 recognition accuracies)

Paper’s Contributions and Critics: Using LSTM

“Multi-layer LSTM has been used to obtain gait feature.”

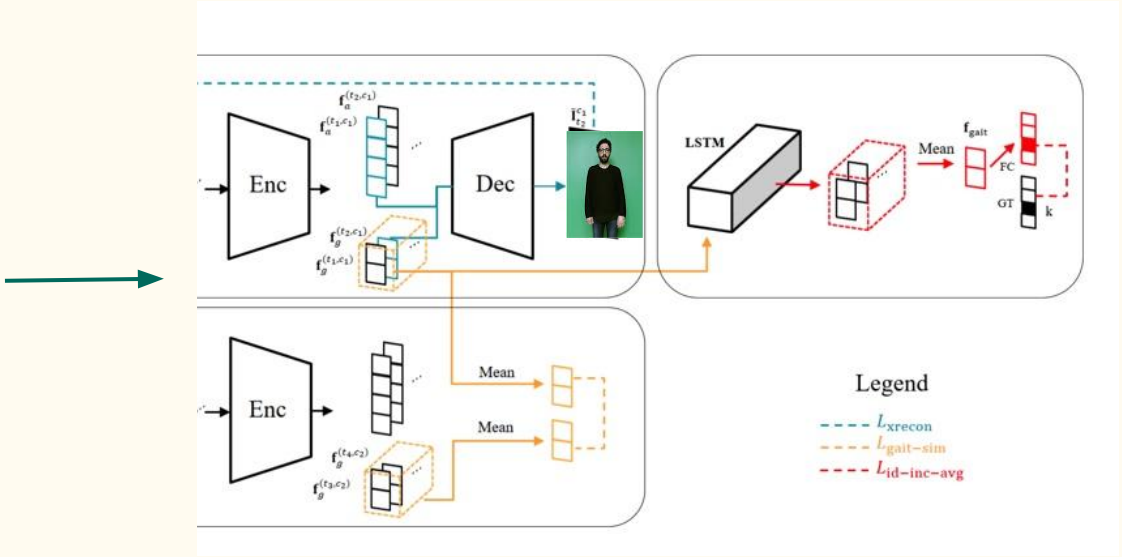
How about transformer networks? Benefit from the attention mechanism.

Paper’s Contributions and Critics: Encoder

Their encoder-decoder network is a typical CNN. Encoder consisting of 4 stride-2 convolution layers following by Batch Normalization and Leaky ReLU activation.

Is it enough? How about using deeper architectures?

Conclusion and Future Work



References

[1] Z. Zhang *et al.*, "Gait Recognition via Disentangled Representation Learning," 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, CA, USA, 2019, pp. 4705-4714, doi: 10.1109/CVPR.2019.00484.

[2] L. Tran, X. Yin and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition," 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, HI, 2017, pp. 1283-1292, doi: 10.1109/CVPR.2017.141.

[3] Zihang, Jiang & Wu, Qianyi & Chen, Keyu & Zhang, Juyong. (2019). Disentangled Representation Learning for 3D Face Shape.

[4] Yang, Linlin & Yao, Angela. (2019). Disentangling Latent Hands for Image Synthesis and Pose Estimation. 9869-9878. 10.1109/CVPR.2019.01011.

[5] Y. Zhao, Y. Xiong and D. Lin, "Recognize Actions by Disentangling Components of Dynamics," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 6566-6575, doi: 10.1109/CVPR.2018.00687.

References

[6] Liu, Yang & Wang, Zhaowen & Jin, Hailin & Wassell, Ian. (2018). Multi-task Adversarial Network for Disentangled Feature Learning. 3743-3751. 10.1109/CVPR.2018.00394.

[7] Ganin, Yaroslav & Lempitsky, Victor. (2014). Unsupervised Domain Adaptation by Backpropagation.

[8] Pedro M. Ferreira, Diogo Pernes, Ana Rebelo, Jaime S. Cardoso, "Learning signer-invariant representations with adversarial training," Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019), 114333D (31 January 2020)