THE ARM-SWING IS DISCRIMINATIVE IN VIDEO GAIT RECOGNITION FOR ATHLETE RE-IDENTIFICATION

Yapkan Choi, Yeshwanth Napolean, Jan C. van Gemert

Computer Vision Lab, Delft University of Technology, Delft, The Netherlands

ABSTRACT

In this paper we evaluate running gait as an attribute for video person re-identification in a long-distance running event. We show that running gait recognition achieves competitive performance compared to appearance-based approaches in the cross-camera retrieval task and that gait and appearance features are complementary to each other. For gait, the arm swing during running is less distinguishable when using binary gait silhouettes, due to ambiguity in the torso region. We propose to use human semantic parsing to create partial gait silhouettes where the torso is left out. Leaving out the torso improves recognition results by allowing the arm swing to be more visible in the frontal and oblique viewing angles, which offers hints that arm swings are somewhat personal. Experiments show an increase of 3.2% mAP on the CampusRun and increased accuracy with 4.8% in the frontal and rear view on CASIA-B, compared to using the full body silhouettes.

Index Terms— Video person re-identification, gait recognition, human semantic parsing

1. INTRODUCTION

Athletes in long-distance running events are typically identified and tracked using the number tag on their race bib, which may include a RFID tag for measuring split times at specific locations, or a GPS tracker for real-time tracking. In smaller events, usually only the start and finish time are registered, not intermediate locations and times. With increasing prevalence of smartphones/cameras, images and videos from race organizers or spectators provide an additional source of information for runner identification and tracking [1]. Vision-based methods for identifying distance runners include bib number detection [2, 3, 4, 5] and appearance-based person re-identification [6]. Potential issues with these methods arise when the bib number is partially or fully occluded, or when multiple athletes wear similar clothing styles and color. Additionally, avoiding the storage of RGB images would alleviate privacy concerns where people can easily be recognized by others. In this paper we investigate if identifying runners based on their running gait is possible and we explore gait recognition as a complementary alternative to runner re-identification with appearance features.

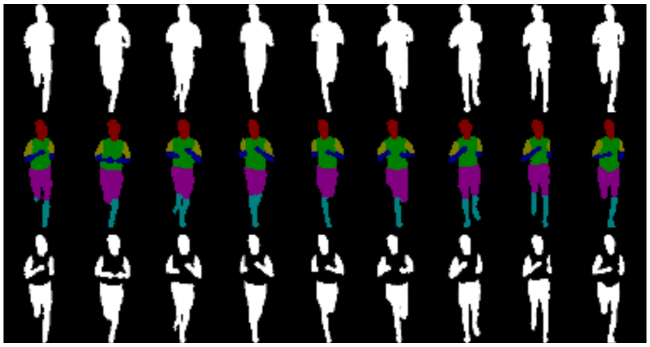


Fig. 1. With binary gait silhouettes, the arm swing is not visible (top). We reduce the ambiguity caused by self-occlusion with the torso with body-part-specific segmentation masks (middle). Partial binary gait silhouettes of a running gait cycle with visible arm swing by removing the torso (bottom).

Recently, research in gait recognition has focused on dealing with co-variates such as view angle [7, 8], clothing and carrying conditions [9]. Although speed-invariant gait recognition from treadmill sequences has been proposed before [10, 11], to the best of our knowledge, no previous research has evaluated running gait recognition with unconstrained running conditions. In this paper, we use the CampusRun dataset [6] of videos captured by hand-held cameras during a running event and evaluate models in a cross-camera setting.

Representing gait as a sequence of binary gait silhouettes has been widely adopted [7, 12, 13, 14]. The primary concern with silhouettes for running gait is self-occlusion of body parts. Specifically, a portion of the arm swing is lost due to the ambiguity of the torso region when using gait silhouettes. As the arm swing is informative, we propose to create partial silhouettes (from binary images) from body-part-specific segmentation masks generated by a semantic body-part parsing model [15]. The partial binary gait silhouettes without the torso reduces the ambiguity of the torso region and improves person re-identification results by allowing the arm swing to

This study is supported by NWO grant P16-28 project 8 Monitor and prevent thermal injuries in endurance and Paralympic sports of the Perspectief Citius Altius Sanius - Injury-free exercise for everyone program.

be much more visible (see figure 1). Additionally, we evaluate the partial silhouettes on walking sequences. As we walk with straight arms, ambiguity in the binary gait silhouettes is less prevalent. Our main contributions can be summarized as follows:

- We apply cross-camera video person re-identification in the long-distance running domain by extending the CampusRun video dataset [6] with 2,581 annotated tracklets of 257 recreational runners from 18 cameras.

- We compare and complement gait features with appearance features on the CampusRun. We demonstrate the feasibility and usefulness of gait as a feature for the cross-camera retrieval task.

- We show that removing the torso from the silhouettes provides 3.2% improved mAP on the CampusRun and we also achieved a 4.8% increase in performance with the CASIA-B dataset (frontal and rear view).

2. METHOD

2.1. Gait silhouette

Given a sequence of bounding boxes indicating a subjects position in an image/video frame (tracklet) of a runner, we construct silhouettes using bounding boxes from consecutive frames. Background subtraction [16] for extracting gait silhouettes requires a static camera for reliable results. Since we allow hand-held cameras, we use convolutional neural networks for segmenting gait silhouettes from the tracklets.

Pipeline. Figure 2 depicts the pipeline for constructing binary gait silhouettes. For each bounding box, we use a human semantic parsing model [15] to segment the input images into body-part-specific masks. As the human parsing model is on a semantic level and the bounding box can contain multiple identities, we use Mask R-CNN [17] to segment the person of interest and keep only the largest instance when multiple instances are found by Mask R-CNN. The body-part-specific masks are converted to binary silhouettes, aligned, and resized to a size of 64×44, following the GaitSet approach [13].

Partial silhouettes. The human semantic parsing model [15] in the pipeline is pre-trained on the PASCAL-Person-Part dataset [18]. Unlike other human semantic parsing datasets [19, 20], the PASCAL-Person-Part dataset does not have clothing-specific segmentation label categories. We use 7 labels: Background, Head, Torso, Upper Arms, Lower Arms, Upper Legs and Lower Legs. These body-part-specific labels suit the task of gait recognition, because the resulting segmentation masks are less dependent on the person's clothing. The partial gait silhouettes are composed of all body-part-specific segmentation masks without the torso.

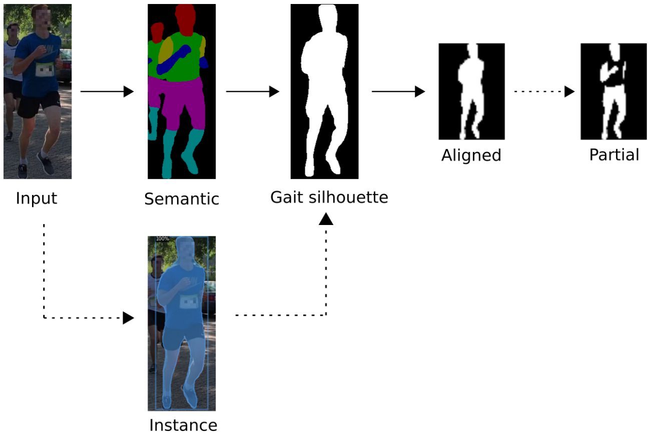


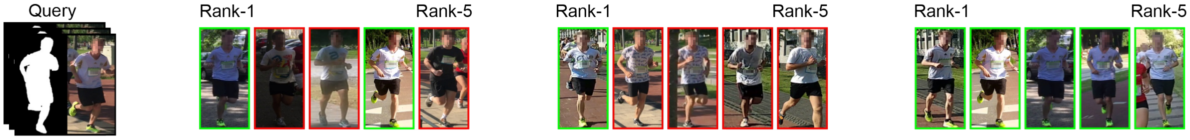
Fig. 2. Our pipeline with human semantic parsing [15] and instance segmentation [17] to create (partial) binary gait silhouettes from bounding boxes. Instance segmentation is only used when there are multiple persons in the bounding box.

2.2. Models.

We use a baseline gait recognition model and two appearance-based person re-identification models to compare gait and appearance features. For a fair comparison, all models use identical input sampling, input resolution and loss function.

Gait features. We use GaitSet [13] as our baseline gait recognition model. It achieves state-of-the-art performance on CASIA-B [9] and OU-MVLP [8] for cross-view gait recognition. In GaitSet, the identity of a person is learned from a set of gait silhouettes. The network first extracts frame-level features and then aggregates the feature maps of each silhouette using max pooling at the set-level. Horizontal Pyramid Pooling [21] slices the last set-level feature map into different horizontal strips of multiple pyramid scales, to learn feature representations with different receptive fields and spatial locations. For each set of silhouettes, the network outputs a discriminative representation, consisting of 62 feature map strips with 256 dimensions each. During training, the set of silhouettes is a subset of the sequence, where we randomly sample a fixed number of silhouettes from the tracklet. As human gait is a periodic movement, a representation can be learned if we sample sufficient frames. All silhouettes from the tracklet are used during evaluation.

Appearance features. For appearance-based person re-identification models, we explore 2D and 3D CNN models with a ResNet-50 backbone [22]. Like our baseline gait recognition model, both appearance-based models use a randomly sampled subset of bounding boxes during training. For evaluation, both models output a feature vector with 2,048 dimensions for each input tracklet. We use a 2D ResNet-50 [22] model, pre-trained on ImageNet [23]. The model aggregates frame-level features using average pooling to get one feature representation for the set of input bounding boxes. To



(a) Sample (b) GaitSet feature (c) 2D ResNet-50 feature (d) Combined feature

Fig. 3. (a) Sample query and corresponding rank-5 retrieval results for (b) GaitSet, (c) 2D ResNet-50 and (d) their combined feature. Green and red borders denote correct and incorrect matches respectively. Gait features (b) are more robust against co-variates such as background and clothing color. Multi-modal approach (d) improves retrieval results.

leverage features from both the temporal and spatial dimensions, we use a 3D ResNet-50 [24] model which is pre-trained on Kinetics [25] for the action recognition task. In contrast to GaitSet and 2D ResNet-50, we use randomly sampled sequences with consecutive frames for 3D ResNet-50 during training. We use the layer before the final classification layer as the person identity feature. During testing, a tracklet gets split into non-overlapping chunks with a fixed number of consecutive frames, followed by taking the mean of the person identity features from each chunk.

Triplet loss. Models are trained with Batch-All triplet loss [26], where all triplet combinations in a batch are used for calculating the loss. The triplet loss in GaitSet is calculated for each of the 62 feature strips individually, followed by taking the mean of the losses. The batch size is $p \times k \times c$, where p denotes the number of people, k the number of tracklets for each person and c the number of frames for each tracklet.

3. EXPERIMENTS

3.1. Comparison with appearance-based features

CampusRun dataset. The CampusRun [6] was a running event with 257 runners who were captured on video using 9 non-stationary hand-held smartphone cameras across the whole track, where each camera operator was allowed to move along the course. We use multi-object tracking [27] to extract tracklets and bounding boxes for each runner from the videos. After manually annotating the bib numbers, we obtain bounding boxes for 257 runners and 2,581 tracklets with an average sequence length of 77 frames. The 10 km runners have 13 tracklets on average.

Evaluation protocol. We use the 5 km runners for model training and validation, while the 10 km runners are only used for testing. The training set and validation set are constructed using a 60/40 split (5 km, 125 runners, 9 cameras, 860 tracklets). The test set (10 km, 132 runners, 18 cameras, 1,721 tracklets) and validation set are evaluated using a cross-camera setting, where the probe identity is captured from a different camera than the positive matches in the gallery. We have 1,721 test queries with a maximum of 17 positive matches for each query, as the runners do not appear more than once per camera.

Method	mAP	Rank-1
Appearance: 2D ResNet-50 [22]	**56.3**	74.6
Appearance: 3D ResNet-50 [24]	56.2	74.0
Gait: GaitSet [13]	52.2	**78.7**
GaitSet [13] + 2D ResNet-50 [22]	**81.1**	**93.9**
GaitSet [13] + 3D ResNet-50 [24]	71.1	90.1
2D ResNet-50 [22] + 3D ResNet-50 [24]	59.3	78.4

Table 1. **Exp. 1:** Comparison of appearance-based and gait-based methods on CampusRun. Gait features achieves comparable mean average precision and rank-1 accuracy to appearance-based methods. The pairwise model combinations show that gait features are complementary to appearance features for person re-identification.

Training details. We follow the training protocol of GaitSet [13] for all models, but use a smaller batch size ($p = 8$ persons, $k = 4$ tracklets) due to the CampusRun dataset having fewer sequences per identity than in CASIA-B. Additionally, the GaitSet model is pre-trained on CASIA-B. The learning rate is set to 1e-4, and we train the models for 80K iterations. We choose the best model checkpoint based on the mAP of the validation set. During training, we randomly sample $c = 30$, $c = 10$ and $c = 30$ frames for GaitSet, 2D ResNet-50 and 3D ResNet-50 respectively. For data augmentation, we randomly horizontal flip the entire tracklet. We compare the output vectors of two tracklets using Euclidean distance. We resize and align the silhouettes to 64×44, while the bounding boxes for the 2D ResNet-50 and 3D ResNet-50 are resized to 64×32.

Exp. 1: Results on CampusRun. Table 1 shows the comparison between appearance-based and gait-based models on the CampusRun dataset. We observe comparable mAP and rank-1 accuracy between all three models. We use pairwise combinations of the three models to analyze if the models learn different features. Before performing distance calculations, we concatenate the two feature vectors from each pair of models, after first ℓ_2 normalizing the individual vectors. The results for the pairwise combinations in table 1 show that a multi-modal approach using gait and appearance features leads to a more diverse and complementary ensemble

Instance seg.	Semantic parsing	Partial silhouettes	mAP	Rank-1
✓			37.4	64.9
✓	✓		51.3	78.2
✓	✓	✓	**54.5**	**81.1**

Table 2. **Exp. 2:** The contribution of each component of our proposed pipeline towards performance on CampusRun. The largest improvement is attributed to the increase in detail of human semantic parsing silhouettes.

than adding a spatio-temporal model from the same modality. Retrieval results in figure 3 highlight advantages and disadvantages of the respective methods.

3.2. Evaluating partial binary gait silhouettes

Datasets. For this experiment, we train the GaitSet model from scratch using partial gait silhouettes without the torso. We use the CampusRun dataset as described in section 3.1. Additionally, we explore partial gait silhouettes for walking sequences using the CASIA-B [9] dataset. It contains gait sequences of 124 persons with 3 walking conditions: normal (6 sequences NM#1-6), carrying a bag (2 sequences BG#1-2) and wearing a coat (2 sequences CL#1-2). The participants are captured from 11 views from 0° to 180° in 18° increments, resulting in $11 \times (6+2+2) = 110$ sequences for each person.

Evaluation protocol. We follow the same setup and evaluation protocol as in GaitSet [13]. The first 74 persons are used for training and the remaining 50 persons for testing. The models are evaluated with rank-1 accuracy, but we exclude identical-view cases.

Exp. 2: Results on CampusRun. Table 2 shows the contribution of each component of our proposed pipeline for constructing partial gait silhouettes. The segmentation masks generated by the human semantic parsing model [15] are more detailed than segmentation masks from the instance segmentation model [17], leading to a 13.9% improvement in mAP. Partial gait silhouettes without the torso increases the portion of arm swing that is visible, resulting in a 3.2% mAP improvement over using the full body silhouettes.

Exp. 3: Results on CASIA-B. The torso segmentation mask is subtracted from the original binary gait silhouettes provided by CASIA-B. Table 3 shows the average rank-1 accuracies for full body silhouettes and partial silhouettes without the torso. For all models and co-variates, oblique view angles (18°-72°, 108°-162°) achieve higher accuracy than frontal (0°, 180°) or lateral views (90°), because silhouettes observed from oblique view angles contain more motion information than the other two planes individually. Subtracting the torso from the original silhouettes increases accuracy for the frontal views, because contours of the arm swing become more perceptible. With binary gait silhouettes, it is difficult to discern the human gait in the frontal plane, as the arm and leg swing happen in the sagittal plane. Without the torso, the motion in the frontal plane is more visible, which improves accuracy for the frontal results. A similar pattern of results was obtained for the bag and clothing co-variates.

Gallery NM#1-4			0°-180°			
Probe	Method	Silhouette	Frontal	Oblique	Lateral	Mean
NM#5-6	GaitPart [14]	Full body	92.3	**97.7**	**92.3**	**96.2**
	GaitSet [13]	Full body	88.3	97.1	91.7	95.0
	GaitSet [13]	Partial	**93.1**	97.1	91.1	95.8
BG#1-2	GaitPart [14]	Full body	**87.5**	**93.4**	**84.9**	**91.5**
	GaitSet [13]	Full body	81.4	89.5	81.0	87.2
	GaitSet [13]	Partial	82.3	89.6	79.1	87.3
CL#1-2	GaitPart [14]	Full body	**68.6**	**82.0**	**72.5**	**78.7**
	GaitSet [13]	Full body	55.7	74.1	70.1	70.4
	GaitSet [13]	Partial	62.2	75.4	67.0	72.2

Table 3. **Exp. 3:** Averaged rank-1 accuracies on CASIA-B. The oblique (18°-72°, 108°-162°) and frontal (0°, 180°) probe views are grouped together in the table, but the mean accuracy is calculated over all 11 views. Subtracting the torso leads to increased accuracy in the frontal views, but decreased accuracy in the lateral view.

For CASIA-B, it remains unclear to which degree recognition performance is attributed to the pixel-level accuracy of the segmentation masks generated by the human semantic parsing model. We observe incorrect parsing results for the lateral view angle when the arms align with the torso. This may explain why the accuracy for the lateral view angle decreases for all three probe subsets (NM, BG, CL), when subtracting the torso from the silhouettes. Most sequences in the CampusRun dataset were captured from oblique angles between 10° and 45°. We did not find an increase in rank-1 accuracy for oblique angles in CASIA-B as was observed in the CampusRun, which suggests that the arm swing is more discriminative for recognizing running gait.

4. CONCLUSION

We extend the CampusRun dataset [6] with additional annotations, recorded at a long-distance running event for cross-camera video person re-identification. Experimental results using the CampusRun dataset show that runners can be identified based on their running gait. Furthermore, we demonstrate that gait features are both competitive and complementary to appearance features. Additionally, we investigate arm swing as a feature by extracting partial binary gait silhouettes using human semantic parsing and instance segmentation. We demonstrate that subtracting the torso from the gait silhouettes for runners leads to increased recognition performance by making the arm swing more visible.

5. REFERENCES

[1] M. D. Flintham, R. Velt, M. L. Wilson, et al., “Run spot run: capturing and tagging footage of a race by crowds of spectators,” in *CHI’15*, 2015, pp. 747–756.

[2] I. Ben-Ami, T. Basha, and S. Avidan, “Racing bib numbers recognition.,” in *BMVC*, 2012, pp. 1–10.

[3] P. Shivakumara, R. Raghavendra, L. Qin, et al., “A new multi-modal approach to bib number/text detection and recognition in marathon images,” *Pattern Recognition*, vol. 61, pp. 479–491, 2017.

[4] N. Boonsim, “Racing bib number localization on complex backgrounds,” *WSEAS Transactions on Systems and Control*, vol. 13, pp. 226–231, 2018.

[5] S. Karaoglu, R. Tao, J. C. van Gemert, and T. Gevers, “Con-text: Text detection for fine-grained object classification,” *IEEE transactions on image processing*, vol. 26, no. 8, pp. 3965–3980, 2017.

[6] Y. Napolean, P. T. Wibowo, and J. C. van Gemert, “Running event visualization using videos from multiple cameras,” in *MMSports’19*, 2019, pp. 82–90.

[7] Z. Wu, Y. Huang, L. Wang, et al., “A comprehensive study on cross-view gait based human identification with deep cnns,” *PAMI*, vol. 39, no. 2, pp. 209–226, 2017.

[8] N. Takemura, Y. Makihara, D. Muramatsu, et al., “Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition,” *IPSJ Transactions on Computer Vision and Applications*, vol. 10, no. 1, pp. 4, 2018.

[9] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in *ICPR’06*. IEEE, 2006, vol. 4, pp. 441–444.

[10] Y. Guan and C.-T. Li, “A robust speed-invariant gait recognition system for walker and runner identification,” in *2013 International Conference on Biometrics (ICB)*. IEEE, 2013, pp. 1–8.

[11] C. Xu, Y. Makihara, X. Li, et al., “Speed-invariant gait recognition using single-support gait energy image,” *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26509–26536, 2019.

[12] M. Z. Uddin, D. Muramatsu, N. Takemura, et al., “Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion,” *IPSJ Transactions on Computer Vision and Applications*, vol. 11, no. 1, pp. 9, 2019.

[13] H. Chao, Y. He, J. Zhang, and J. Feng, “Gaitset: Regarding gait as a set for cross-view gait recognition,” in *AAAI*, 2019, vol. 33, pp. 8126–8133.

[14] C. Fan, Y. Peng, C. Cao, et al., “Gaitpart: Temporal part-based model for gait recognition,” in *CVPR*, 2020, pp. 14225–14233.

[15] P. Li, Y. Xu, Y. Wei, and Y. Yang, “Self-correction for human parsing,” *PAMI*, 2020.

[16] L. Wang, T. Tan, H. Ning, and W. Hu, “Silhouette analysis-based gait recognition for human identification,” *PAMI*, vol. 25, no. 12, pp. 1505–1518, 2003.

[17] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask r-cnn,” in *ICCV*, 2017, pp. 2961–2969.

[18] X. Chen, R. Mottaghi, X. Liu, et al., “Detect what you can: Detecting and representing objects using holistic models and body parts,” in *CVPR*, 2014, pp. 1971–1978.

[19] X. Liang, S. Liu, X. Shen, et al., “Deep human parsing with active template regression,” *PAMI*, vol. 37, no. 12, pp. 2402–2414, 2015.

[20] K. Gong, X. Liang, D. Zhang, et al., “Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing,” in *CVPR*, 2017, pp. 932–940.

[21] Y. Fu, Y. Wei, Y. Zhou, et al., “Horizontal pyramid matching for person re-identification,” in *AAAI*, 2019, vol. 33, pp. 8295–8302.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[23] O. Russakovsky, J. Deng, H. Su, et al., “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[24] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?,” in *CVPR*, 2018, pp. 6546–6555.

[25] W. Kay, J. Carreira, K. Simonyan, et al., “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.

[26] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.

[27] Y. Zhang, C. Wang, X. Wang, et al., “A simple baseline for multi-object tracking,” *arXiv preprint arXiv:2004.01888*, 2020.