Running Gait Recognition Using Arm and Leg Swing for Video Person Re-Identification

Yapkan Choi

Running Gait Recognition Using Arm and Leg Swing for Video Person Re-Identification

by

Yapkan Choi

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday September 30, 2020 at 10:00 AM.

Student number:	4025067	
Project duration:	November, 2019 – September, 2020	
Thesis committee:	Dr. J. C. van Gemert,	TU Delft, supervisor
	Dr. S. L. Pintea,	TU Delft
	Dr. ir. S. E. Verwer,	TU Delft

TU Delft
Delft University of Technology

Preface

This master's thesis presents the results of the research conducted at the Computer Vision Lab of Delft University of Technology, under the supervision of Dr. Jan van Gemert. Yeshwanth Napolean has been my daily co-supervisor. The scientific paper contains the main contents of the report, which is followed by supplementary materials that include background information and additional experiments.

I would like to thank Dr. Jan van Gemert and Yeshwanth Napolean for their support, guidance and patience. They provided me with invaluable questions and suggestions throughout the project. I would also like to thank Dr. Silvia Pintea and Dr. Sicco Verwer for being part of my thesis committee and for taking the time to assess my work.

Yapkan Choi
Delft, September 2020

Contents

1

Scientific Paper

Running Gait Recognition Using Arm and Leg Swing for Video Person Re-Identification

Yapkan Choi
Delft University of Technology
Delft, the Netherlands
y.k.choi@student.tudelft.nl

Abstract

Person re-identification based on appearance is challenging due to varying views and lighting conditions in different cameras, or when multiple persons wear similar clothing styles and color. Considering these challenges, gait patterns provide an alternative to appearance, as gait can be captured from a distance and at a low resolution. In this paper we investigate and evaluate running gait as a unique attribute for video person re-identification in a recreational long-distance running event with 257 participants. We show that running gait recognition achieves competitive performance compared to video-based approaches in the cross-camera retrieval task and that gait and appearance features are complementary to each other. In addition, we compare gait recognition applied to walking and running sequences. An important difference is that we walk with straight arms, but run with bent arms. We propose to use human semantic parsing to create partial gait silhouettes from body parts to find the most discriminative combination. We demonstrate that the arm and leg swing are the most discriminative parts of the running gait. Our proposed method provides better recognition results by removing the torso from the silhouettes and allowing the arm swing to be more visible.

1. Introduction

Athletes in long-distance running events are identified and tracked using the number tag on their race bib. These bibs often include a RFID tag for measuring split times at specific locations, or incorporate a GPS tracker for real-time tracking. In smaller recreational distance running events, usually only the start and finish time are registered, while intermediate location and times are not recorded. With the increasing use of video cameras and smartphones, images and videos from the race organizers, photographers or spectators provide an additional source for runner identification and tracking [9]. Vision-based methods for identifying distance runners include bib number detection [2, 4, 38] and person re-identification [35]. 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. In this paper, we investigate if identifying runners based on their running gait is possible and we explore the use of gait recognition as an alternative to runner re-identification with appearance features.

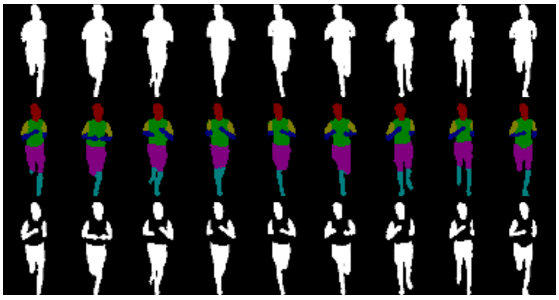


Figure 1. **Accentuated arm swing with partial gait silhouettes.** Running gait cycle for both full body silhouettes (top), human semantic parsing (middle), and partial silhouettes where the torso is removed (bottom). We use body-part-specific segmentation masks generated by a human semantic parsing model to remove the torso from the full body silhouettes.

Recently, research in gait recognition has mainly focused on dealing with covariates such as view angle [41, 48], clothing and carrying conditions [53]. Although speed-invariant gait recognition from treadmill sequences has been proposed before [12, 49], to the best of our knowledge, no previous research has investigated running gait recognition with unconstrained running conditions specifically. We construct a new large-scale video dataset of 257 recreational runners captured by hand-held cameras during a running event and evaluate the models in a cross-camera setting in which the runners are captured by 18 cameras. We compare walking and running sequences, where an important difference is that we walk with straight arms, but run with bent arms. Representing gait as a sequence of binary gait silhouettes has been widely adopted [6, 8, 42, 48]. The primary concerns with applying the silhouette-based representation to the running gait are self-occlusion of body parts, and that a portion of the arm swing is mostly lost due to the ambiguity of the torso region when using gait silhouettes. We propose to create partial binary gait silhouettes from body-part-specific segmentation masks generated by a human semantic parsing model [24]. Removing the torso from the silhouettes solves the ambiguity of the torso region and increases person re-identification results by allowing the arm swing to be more visible. In addition, the body-part-specific segmentation masks allow finding the discriminative contribution of each body part for running gait recognition and to compare it to previous studies which demonstrated that the upper body contains substantial discriminative power [5, 28, 31]. Our main contributions can be summarized as follows:

- We introduce a large-scale long-distance running video dataset for cross-camera video person re-identification. We extend the CampusRun video dataset [35] with 2581 annotated tracklets of 257 recreational runners from 18 cameras in the 5 km and 10 km distance races.
- We investigate if identifying runners based on their running gait is possible. In addition, we compare between gait features and appearance features for video person re-identification using the CampusRun dataset. We demonstrate the feasibility and usefulness of gait as a unique attribute for the cross-camera retrieval task.
- We propose a pipeline for creating partial binary gait silhouettes from bounding boxes, by using body-part-specific segmentation masks generated by a human semantic parsing model. We compare walking and running sequences and demonstrate that the arms and the legs are the most discriminative body parts for running gait recognition. Removing the torso from the silhouettes allows for the arm swing to be more visible.

2. Related work

Person re-identification. Person re-identification has been extensively studied in the cross-camera setting [1, 57], on benchmarks that represent a person's identity with images [25, 46, 56] or videos [36, 55]. For a given query, the task is to retrieve all moments the person (probe) appears in a set of images or videos (gallery), where the probe and gallery are captured from disjoint cameras. Prior research has addressed viewpoint variations [40, 52], illumination conditions [19] and occlusions [18, 58]. While most benchmarks consist of walking pedestrians, the research in person re-identification with running sequences remains limited due to lack of available datasets [39]. We argue that the long-distance running domain is particular suited for the cross-camera person re-identification task, because events often include a large number of participants, unconstrained outdoor environments, strictly defined courses, prolonged periods of activity, pace variations and clothing similarities. We apply a cross-camera evaluation protocol to a long-distance running event by extending the CampusRun video dataset [35] with annotated tracklets for 257 recreational runners.

Runner identification. Racing bib number detection is the primary method for identifying long-distance runners [20, 35]. As racing bibs are usually attached to the front of the athlete's shirt, a common approach is to reduce the search area by detecting the torso. Finding the torso region for bib number detection can be accomplished by first doing face detection [2, 4], upper body detection [38], using human pose estimation [33] or extracting skin portions [34]. Bib number detection faces challenges such as occlusion of the numbers and low resolution images. Alternatively, in our paper we explore appearance features [35] and gait recognition to match runners.

Gait recognition. There have been numerous studies on vision-based gait recognition [23, 43], with approaches primarily divided into model-based and appearance-based for extracting gait related features. Model-based methods model the human body to extract static and dynamic body parameters, such as stride parameters [3] or joint-angle trajectories [44]. Typical appearance-based methods use either a shape representation in the form of binary gait silhouettes [45], or a spatio-temporal variant: the gait energy image (GEI), which is the average over one gait cycle of silhouettes [13]. Binary gait silhouettes or GEI are then used for feature extraction [22, 27] and recognition. In our paper, we represent gait using a sequence of binary gait silhouettes.

Running gait recognition. Running gait recognition has been explored in prior studies with videos of persons walking and running on a treadmill [12, 32]. The participants are captured from a lateral view angle and their velocity range from 2 km/h to 10 km/h. Using a leg motion model for both gait modes [51], it was found that the recognition rate for running sequences was higher, suggesting that running gait has more discriminative power than walking due to larger variation in running gait. A modified GEI representation with silhouettes where the limbs are the most closed (single-support phase), was used in [49] for speed-invariant gait recognition, as this phase varies the least with speed changes. In our paper we also investigate running gait recognition, but instead use outdoor running sequences and with velocities up to 17 km/h.

Silhouette-based gait recognition. Recent works have focused on convolutional neural networks to tackle cross-view gait recognition [47, 48]. Representing a gait sequence as an unordered set of gait silhouettes was proposed by GaitSet [6], where the model extracts set-level features and maps the features to a final gait representation with Horizontal Pyramid Pooling [10]. In contrast to GaitSet [6], the GaitPart [8] model takes as input sorted silhouette sequences. Building upon GaitSet, GaitPart uses convolution on separate horizontal slices of the feature maps with different horizontal pyramid scales [10] to extract more fine-grained features for each horizontal part. Furthermore, GaitPart focuses on short-range motion by applying sliding windows over the sequence to aggregate frame-level features into short-range spatio-temporal representations. However, the limitations of representing gait with silhouettes are the ambiguity of the torso region when the arms are overlapping, and self-occlusion by other body parts. We propose to solve the ambiguity of the torso region using human semantic parsing to segment the arms from the torso.

Part-level gait silhouettes. Liu et al. [29] provided manual labeled part-level silhouettes with 8 body parts for 71 subjects captured from a lateral view angle. The Layered Deformable Model (LDM) [30] is used to recover body pose from gait silhouettes after estimating the model parameters from the manual labeled part-level silhouettes [29]. The LDM [30] was extended to include the full body model with 11 more parameters [31]. Besides leg movement, upper body dynamics of the arms, head and shoulders were found to improve recognition rates [31]. Boulgouris and Chi [5] used the same part-level silhouettes [29] and demonstrated that the human body parts have different discriminative power. Body parts between two identities were compared using Jaccard distance. The torso, occluded right arm and occluded right upper leg are found to be the most discriminative, due to pattern of appearance being more useful than the shape of the body parts [5]. In our paper we extend the part-level approach to multiple view angles, besides the lateral view. We investigate the relative importance of the human body components by creating partial binary gait silhouettes using combinations of one or more body parts.

3. Method

3.1. Gait silhouette

Pipeline. Figure 2 depicts the pipeline for constructing binary gait silhouettes. Given a tracklet of the runner, we construct silhouettes for the set of bounding boxes obtained from consecutive frames. For each bounding box, we use an off-the-shelf human semantic parsing model [24] 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 [15] 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 gait silhouettes and are aligned and resized to a size of 64×44 using the same procedure as in GaitSet [6].

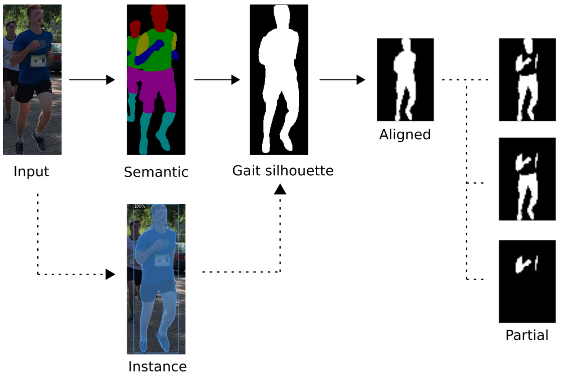


Figure 2. **From bounding box to binary gait silhouette.** Pipeline with human semantic parsing [24] and instance segmentation [15] to create (partial) gait silhouettes.

Partial silhouettes. The human semantic parsing model [24] in the pipeline is pre-trained on the PASCAL-Person-Part dataset [7]. Unlike other human semantic parsing datasets [11, 26], 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. We use the body-part-specific segmentation masks to create partial silhouettes from one or more body parts.

3.2. Models.

We use one baseline gait recognition model and two video-based person re-identification models to compare gait and appearance features. To enable a fair comparison, we train and evaluate all models using similar input sampling, input resolution and loss function.

Gait features. We use GaitSet [6] as our baseline gait recognition model. It achieves state-of-the-art performance on CASIA-B [53] and OU-MVLP [41] for the cross-view gait recognition task. 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 on the set-level. Horizontal Pyramid Pooling [10] 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 video-based person re-identification models, we explore 2D and 3D CNN models with a ResNet-50 backbone [16]. As with our baseline gait recognition model, both video-based models use a randomly sampled subset of bounding boxes during training. For evaluation, both models output a feature vector with 2048 dimensions for each input tracklet. We use a 2D ResNet-50 [16] model which is pre-trained on ImageNet [37]. The model aggregates frame-level features using average pooling to get one feature representation for the set of input bounding boxes. To leverage features from both the temporal and spatial dimensions, we use a 3D ResNet-50 [14] model which is pre-trained on Kinetics [21] 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. All three models are trained with Batch All triplet loss [17], where all possible combinations of triplets 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 persons, k the number of tracklets for each person and c the number of frames for each tracklet.

4. Experiments

The experimental section is divided into three parts. The first part compares our gait recognition baseline model with two video-based person re-identification models on the CampusRun dataset. In the second part, we evaluate the use of partial gait silhouettes for running sequences in the CampusRun dataset [35] and walking sequences from CASIA-B [53]. In the third part, we conduct ablation experiments for the GaitSet model [6] on the CampusRun dataset.

4.1. Comparison with video-based person re-identification

CampusRun dataset. The CampusRun [35] was a running event where 262 recreational runners simultaneously ran a 5 km course, while competing in various race distances. 128 runners participated in the 5 km distance race, while 134 runners ran two laps around the course as part of the 10 km distance. The runners were captured on video using 9 non-stationary hand-held smartphone cameras, where each camera operator was allowed to move along the course. The 5 km runners appear in at most 9 cameras, while the 10 km runners with two laps appear in at most 18 cameras. We use an off-the-shelf multi-object tracker [54] to extract tracklets and bounding boxes for each runner from the videos. After manual annotating the bib number of each tracklet, we obtain bounding boxes for 257 runners and 2581 tracklets with an average sequence length of 77 frames. The 10 km runners have 13 tracklets on average. Five registered participants were not found during the tracking and annotation process.

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, n=125, 9 cameras, 860 tracklets). The test set (10 km, n=132, 18 cameras, 1721 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. During evaluation, each tracklet is evaluated once as the probe subset, with every other tracklet in the gallery subset. We have 1721 test queries with a maximum of 17 positive matches for each query, as the runners do not appear more than once per camera. We report both mean average precision (mAP) and rank-1 accuracy, but mainly evaluate the models on mAP. With a maximum of 17 ground truths for each query, the quality of the ranked retrieval results is better reflected by mAP than rank-1 accuracy.

Training details. We follow the same training protocol of GaitSet [6] for all models, but use a smaller batch size ($p = 8$ persons, $k = 4$ tracklets) due to the CampusRun dataset having less 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 horizontal flip the entire tracklet. We compare the output vectors of two tracklets using Euclidean distance. For this experiment, the binary gait silhouettes are composed of all body parts. 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.

Results on CampusRun. Table 1 shows the comparison between video-based and gait-based models on the CampusRun dataset. We observe comparable mean average precision and rank-1 accuracy between all three models when using similar input resolutions. Retrieval results improve

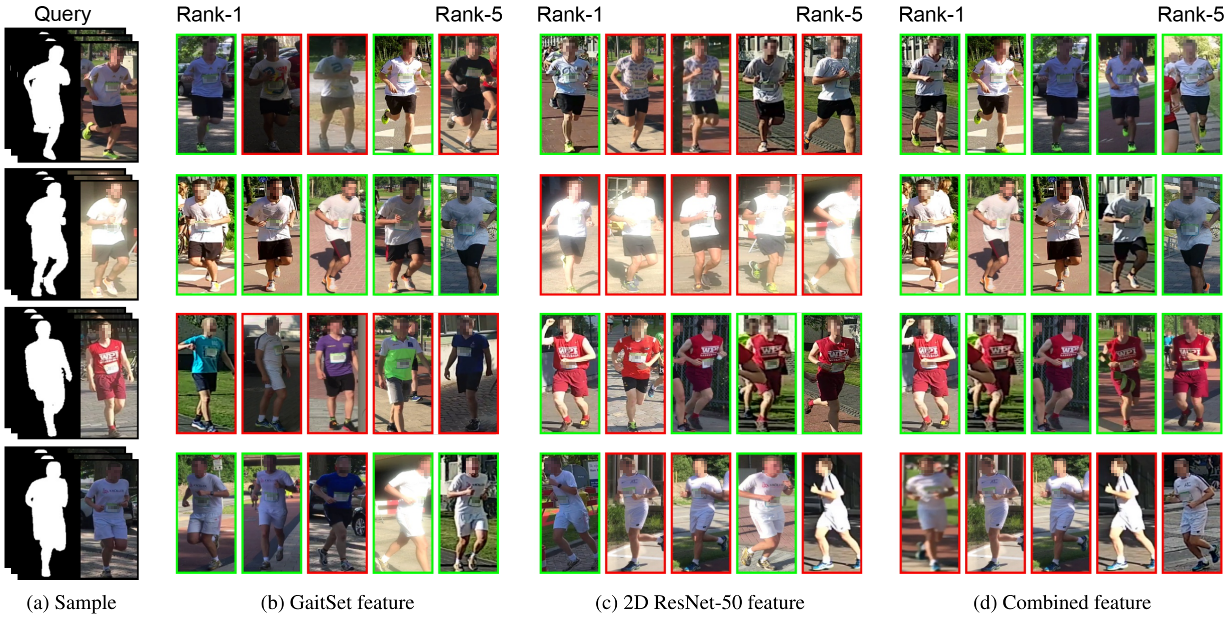


Figure 3. **Example retrieval results.** (a) Four query samples and their corresponding rank-5 retrieval results for (b) GaitSet, (c) 2D ResNet-50 and (d) their combined feature. Green and red borders denote correct and incorrect matches respectively. The queries highlight some advantages and disadvantages of the respective methods. The first query demonstrates that a multi-modal approach retrieves more correct matches than both gait or appearance models individually. In the second query, the gait model is not affected by difficult lighting conditions, whereas the appearance model retrieves five incorrect matches with the same overexposed lighting. In the third query, the gait model retrieves incorrect matches who are also walking, indicating the inability to distinguish between walking and running modes. In the fourth query, the combined feature performs worse than both individual models.

Method	Resolution	mAP	Rank-1
2D ResNet-50 [16]	64×32	56.3	**74.6**
+ re-ranking [59]	64×32	**64.4**	73.3
3D ResNet-50 [14]	64×32	56.2	74.0
+ re-ranking [59]	64×32	63.6	72.6
GaitSet [6]	64×44	52.2	**78.7**
+ re-ranking [59]	64×44	**66.1**	78.6

Table 1. **Comparison of video-based and gait-based methods on CampusRun dataset.** Gait-based person re-identification (GaitSet) achieves comparable mean average precision and rank-1 accuracy to video-based methods (2D ResNet-50, 3D ResNet-50). Retrieval results improve for all three methods when applying re-ranking using k-reciprocal neighbours [59]. The results show that the CampusRun dataset is challenging for single-modality models, where the ability to handle unconstrained conditions is tested.

for all three methods when applying re-ranking using k-reciprocal neighbors [59], but the gait-based approach benefits more from the re-ranking procedure than the video-based methods.

We use pairwise combinations of the three models to analyze if the models learn different features. Before performing distance calculations and re-ranking, 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 2 show that a multi-modal approach using gait and appearance features leads to a more diverse and complementary ensemble than adding a spatio-temporal model from the same modality.

Feature 1	Feature 2	mAP	Rank-1
GaitSet [6]	2D ResNet-50 [16]	**91.2**	**94.0**
GaitSet [6]	3D ResNet-50 [14]	85.3	90.6
2D ResNet-50 [16]	3D ResNet-50 [14]	68.8	77.3

Table 2. **Feature concatenation + re-ranking [59].** Feature concatenation of gait-based and video-based models before distance calculation and re-ranking [59]. We concatenate the two feature vectors after ℓ_2 normalization of each vector. The results show that gait features are complementary to appearance features for person re-identification. Whereas adding a spatio-temporal features from the same modality is less beneficial, compared to using a multi-modal approach.

Figure 3 shows four example queries and their corresponding rank-5 retrieval results for models with gait feature, 2D ResNet-50 feature and their combined features. All retrieval results are after re-ranking [59]. We observe scenarios where a gait-based approach is preferable to a video-based approach: visual similar clothing styles and color (first, second and fourth row) or difficult lighting conditions (second row). In the third query, the query sample is from a walking sequence. Although the models were mostly trained using running sequences, GaitSet retrieves five walking samples from the gallery. This demonstrates that GaitSet is not able to perform cross-mode gait recognition when trained on the CampusRun dataset. To conclude, the quantitative and qualitative results indicate that gait-based methods are competitive and complementary to appearance-based approaches for person re-identification.

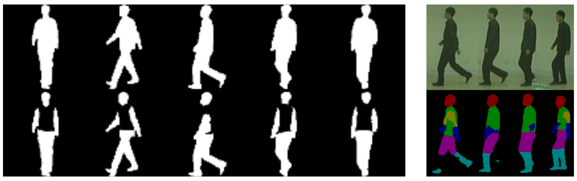
4.2. Partial binary gait silhouettes

Datasets. For this experiment, we train the GaitSet model from scratch using partial gait silhouettes composed of different body part combinations. We use the CampusRun dataset as described in section 4.1. Additionally, we explore partial gait silhouettes for walking sequences using the CASIA-B [53] dataset. CASIA-B is a popular dataset for cross-view gait recognition. 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 [6]. The first 74 persons are used for training and the remaining 50 persons for testing. During testing, the gallery consists of the first 4 walking sequences (NM#1-4), while the remaining sequences (NM#5-6, BG#1-2, CL#1-2) are used as the probe subsets. The models are evaluated with rank-1 accuracy, but we exclude identical-view cases, namely, when the probe and gallery samples are captured from the same view.

Results on CampusRun. Table 3 shows the cross-camera re-identification performance when using partial gait silhouettes compared to using the full body silhouette. The results for partial silhouettes with only one body part reveal that the upper legs are the most discriminative body component, followed by the lower and upper arms. We do not include the torso as it is most of the time affected by self-occlusion due to the arm swing.

We also present the results for combinations of multiple body parts, which show that the dynamic components of the running gait are the most discriminative. While the legs have greater recognition ability than the arms, we need to combine both the legs and arms to match the full body silhouette. The relative static components, such as the head or torso, do not contribute much to recognition performance. Furthermore, removing the torso increases the portion of arm swing that is visible, resulting in improvement in mAP and rank-1 accuracy over using the full body silhouette.


(a) 0°, 54°, 90°, 144°, 180°

(b) 90°

Figure 4. **Example CASIA-B silhouettes.** (a) Full body and partial binary gait silhouettes for different view angles. (b) Incorrect segmentation masks for 90° view.

These findings are consistent with prior studies [5, 31], where it was found that the upper legs were the most discriminative part [5] when considering individual body parts. Besides the legs, the dynamics of the arm swing are important for the re-identification of individuals, in line with the results from [31]. When comparing our results to those of older studies, it must be pointed out that previous studies only considered the lateral view, which is almost not found in the CampusRun dataset. Most sequences in the CampusRun dataset were captured from oblique angles between 10° and 45°.

Results on CASIA-B. We adopt the best performing partial silhouette combination from the CampusRun dataset. The torso segmentation mask is subtracted from the original binary gait silhouettes provided by CASIA-B, as seen in figure 4a. Table 4 shows the average rank-1 accuracies for full body silhouettes and partial silhouettes without the torso. We also include results for GaitPart [8], as it considers short-range micro-motion from subsequent frames.

For all models and covariates, oblique view angles (e.g. 18°, 36°, 54°, 72°) achieve higher accuracy than frontal (0°), rear (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 results in increased accuracy for the frontal and rear views, because contours of the arm swing are more perceptible. With binary gait silhouettes, it is generally hard to discern the human gait in the frontal plane, as the arm and leg swing happen in the sagittal plane. Without the torso, there is an overall increase of motion in the frontal plane, which leads to a substantial boost in accuracy for all frontal and rear results, achieving better accuracy than GaitPart [8] for the normal walking conditions. A similar pattern of results was obtained for the bag and clothing covariates.

It remains unclear to which degree recognition perfor-

Silhouette	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs	mAP	Rank-1
Full body	✓	✓	✓	✓	✓	✓	51.3	78.2
Head	✓						7.1	12.3
Upper arms			✓				16.9	31.3
Lower arms				✓			18.5	37.6
Upper legs					✓		**26.1**	**46.0**
Lower legs						✓	13.7	25.1
Arms			✓	✓			27.5	51.5
Upper body	✓	✓	✓	✓			30.1	52.5
Legs					✓	✓	38.9	65.1
Arms and legs			✓	✓	✓	✓	52.0	77.6
Head, arms and legs	✓		✓	✓	✓	✓	**54.5**	**81.1**

Table 3. **Partial binary gait silhouettes.** Performance comparison when using partial gait silhouettes on CampusRun dataset. The second block shows that the upper legs are the most discriminative when considering only one body part. The third block indicates that making the silhouette of the arms more visible by removing the torso, leads to increased re-identification performance.

Gallery NM#1-4			0°-180°					
Probe	Method	Silhouette	Frontal	Oblique	Lateral	Oblique	Rear	Mean
NM#5-6	GaitPart [8]	Full body	94.1	**97.6**	**92.3**	**97.8**	90.4	**96.2**
	GaitSet [6]	Full body	90.8	97.0	91.7	97.1	85.8	95.0
	GaitSet [6]	No torso	**95.0**	97.0	91.1	97.1	**91.1**	95.8
BG#1-2	GaitPart [8]	Full body	**89.1**	**93.7**	**84.9**	**93.1**	**85.8**	**91.5**
	GaitSet [6]	Full body	83.8	88.8	81.0	90.2	79.0	87.2
	GaitSet [6]	No torso	85.3	89.7	79.1	89.4	79.3	87.3
CL#1-2	GaitPart [8]	Full body	**70.7**	**83.2**	**72.5**	**80.8**	**66.5**	**78.7**
	GaitSet [6]	Full body	61.4	76.4	70.1	71.7	50.0	70.4
	GaitSet [6]	No torso	65.9	78.2	67.0	72.5	58.5	72.2

Table 4. **Comparison between full body silhouettes and without torso.** Averaged rank-1 accuracies on CASIA-B for normal sequences (NM#5-6), carrying a bag (BG#1-2), wearing a coat (CL#1-2), excluding identifical-view cases. We compare GaitPart [8], GaitSet [6] and GaitSet with torso subtracted from the silhouettes. Silhouettes from oblique views contain more motion cues than frontal (0°), lateral (90°) and rear (180°) views. Subtracting the torso leads to increased accuracy in the frontal and rear views, but decreased accuracy in the lateral view. Overall, the silhouettes without torso achieve better results than the full body silhouettes, approaching the results of GaitPart [8] for the normal walking condition. The oblique probe views are grouped together in the table: (18°, 36°, 54°, 72°) and (108°, 126°, 144°, 162°), but the mean accuracy is calculated over all 11 views.

mance is attributed to the pixel-level accuracy of the segmentation masks generated by the human semantic parsing model. In figure 4b, 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. We also did not find an increase in rank-1 accuracy for oblique angles for the normal walking conditions as in the CampusRun, which suggests that subtracting the torso is more useful for recognizing the running gait.

4.3. Ablation experiments

We perform ablation experiments to evaluate the contributions of individual factors to gait recognition performance. All ablation experiments are evaluated using the GaitSet model on the CampusRun dataset.

Impact of data augmentation. Popular large scale gait datasets such as CASIA-B [53] and OU-MVLP [41] only contain participants that are walking in one direction. Both datasets do not capture a full 360° view of the person, but only show one side of a person's body. Whereas the CampusRun dataset contains tracklets that were captured from both the left and right side of a person. Human gait is a

Method	mAP	Rank-1
Mask R-CNN [15]	30.6	58.5
+ horizontal flip	**39.3**	**68.1**

Table 5. **Impact of data augmentation.** Comparison when adding data augmentation during training. Horizontally flipping of the whole tracklet with a 50% probability during training improves the mAP by 8.7. Horizontally flipping the silhouettes resembles capturing the gait using a virtual camera view from the other side with a half period phase shift, due to the bilateral symmetry of the human gait [50].

Method	mAP	Rank-1
Mask R-CNN [15]	39.3	68.1
Human semantic parsing [24]	**52.2**	**78.7**

Table 6. **Impact of segmentation method.** Comparison between Mask R-CNN and human semantic parsing for extracting the gait silhouettes. The silhouettes from the human semantic parsing model are more detailed, which leads to better recognition results.

bilateral symmetric movement where the arm and leg of opposite sides are swinging towards the same direction, with a phase shift of half a period alternating the left and right arm or left and right leg [50]. We explore horizontal flipping of the entire tracklet with 50% probability during training as a data augmentation technique. Horizontally flipping the silhouettes resembles capturing the gait using a virtual camera view from the other side with a half period phase shift, due to the bilateral symmetry of the human gait. Table 5 shows that gait recognition performance is substantially higher with data augmentation.

Impact of segmentation method. Table 6 shows the comparison between binary gait silhouettes from Mask R-CNN [15] and binary gait silhouettes from the pipeline as described in section 3.1. Our pipeline delivers significantly better results, because the segmentation masks generated by the human semantic parsing model [24] are more detailed than the segmentation masks from Mask R-CNN.

Impact of pre-training. Table 7 shows that pre-training on CASIA-B [53] gives a minimal improvement in recognition performance compared to training from scratch. Although the silhouettes are size normalized using the same procedure, we speculate that there is no significant benefit due to the difference in gait velocity and clothing styles between the two datasets.

Impact of input resolution. We compare different input sizes of the binary gait silhouettes in table 8. The results show minimal improvement when using a input resolution of 128×88, which suggest that an input resolution of 64×44 is sufficient large to capture the important features. A similar pattern of results was obtained in [48], where a slight drop in performance was observed when downsampling the GEIs from 128×88 to 64×44 and 32×22.

Method	mAP	Rank-1
Training from scratch	51.3	78.2
Pre-training on CASIA-B [53]	**52.2**	**78.7**

Table 7. **Impact of pre-training.** Pre-training on CASIA-B dataset gives a minimal improvement compared to training the model from scratch. There is a large difference in clothing styles between the two datasets, with the participants of the CampusRun mostly wearing shorts and short sleeves, which could explain the minimal improvement when pre-training on CASIA-B.

Input resolution	mAP	Rank-1
64×44	52.2	78.7
128×88	**53.1**	**79.5**

Table 8. **Impact of input resolution.** An input resolution of 128×88 provides a minimal increase in mAP over using 64×44 gait silhouettes.

5. Conclusion

In this paper, we introduced the CampusRun dataset, a large-scale long-distance running event for cross-camera video person re-identification. Experimental results using the CampusRun dataset shows that runners can be identified based on their running gait. We demonstrate that gait features are both competitive and complementary to appearance features. Additionally, we introduced a pipeline for extracting partial binary gait silhouettes using human semantic parsing and instance segmentation. We analyzed the discriminative power of different human body parts and compared walking and running sequences with the CASIA-B and CampusRun datasets. We demonstrate that the arms and legs are the most discriminative parts of the running gait and we show that subtracting the torso from the gait silhouettes leads to increased recognition performance by making the arm swing more visible.

One limitation of our method is that it did not manage to retrieve the correct matches of running sequences when we used a walking sequence as the query, which demonstrates that our method cannot achieve cross-mode gait recognition. Our findings indicate that gait recognition performance is affected by segmentation method and mask accuracy. It is therefore recommended to fine-tune the human semantic parsing model for each domain, to make the partial gait silhouette pipeline more robust to view variations and occlusion.

References

[1] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, apr 2014.

[2] Idan Ben-Ami, Tali Basha, and Shai Avidan. Racing bib number recognition. In *BMVC 2012 - Electronic Proceedings of the British Machine Vision Conference 2012*, 2012.

[3] Aaron F. Bobick and Amos Y. Johnson. Gait recognition using static, activity-specific parameters. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2001.

[4] Noppakun Boonsim. Racing bib number localization on complex backgrounds. *WSEAS Transactions on Systems and Control*, 13:226–231, 2018.

[5] Nikolaos V. Boulgouris and Zhiwei X. Chi. Human gait recognition based on matching of body components. *Pattern Recognition*, 40(6):1763–1770, jun 2007.

[6] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(16):8126–8133, 2019.

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1979–1986. IEEE Computer Society, jun 2014.

[8] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221. IEEE, jun 2020.

[9] Martin D. Flintham, Raphael Velt, Max L. Wilson, Edward J. Anstead, Steve Benford, Anthony Brown, Timothy Pearce, Dominic Price, and James Sprinks. Run spot run: Capturing and tagging footage of a race by crowds of spectators. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2015-April, pages 747–756, New York, New York, USA, apr 2015. Association for Computing Machinery.

[10] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal Pyramid Matching for Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8295–8302, apr 2019.

[11] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into Person: Self-supervised Structure-sensitive Learning and a new benchmark for human parsing. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6757–6765, 2017.

[12] Yu Guan and Chang-Tsun Li. A robust speed-invariant gait recognition system for walker and runner identification. In *2013 International Conference on Biometrics (ICB)*, number Dcm, pages 1–8. IEEE, jun 2013.

[13] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555. IEEE, jun 2018.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, feb 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-Decem, pages 770–778. IEEE, jun 2016.

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, mar 2017.

[18] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. VRSTC: Occlusion-Free Video Person Re-Identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 7176–7185. IEEE, jun 2019.

[19] Yukun Huang, Xueyang Fu, Zheng Jun Zha, and Wei Zhang. Illumination-invariant person re-identification. In *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pages 365–373, New York, NY, USA, oct 2019. Association for Computing Machinery, Inc.

[20] Kamlesh, Pei Xu, Yang Yang, and Yongchao Xu. Person re-identification with end-to-end scene text recognition. In *Communications in Computer and Information Science*, volume 773, pages 363–374. Springer Verlag, oct 2017.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[22] Worapan Kusakunniran, Qiang Wu, Hongdong Li, and Jian Zhang. Multiple views gait recognition using view transformation model based on optimized gait energy image. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, pages 1058–1064, 2009.

[23] Tracey K. M. Lee, Mohammed Belkhatir, and Saeid Sanei. A comprehensive review of past and present vision-based techniques for gait recognition. *Multimedia Tools and Applications*, 72(3):2833–2869, oct 2014.

[24] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-Correction for Human Parsing. *arXiv preprint arXiv:1910.09777*, oct 2019.

[25] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[26] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep Human Parsing with Active Template Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2402–2414, mar 2015.

[27] Yushu Liu, Junping Zhang, Chen Wang, and Liang Wang. Multiple HOG templates for gait recognition. In *Proceedings - International Conference on Pattern Recognition*, pages 2930–2933, 2012.

[28] Zongyi Liu, Laura Malave, Adebola Osuntogun, Preksha Sudhakar, and Sudeep Sarkar. Toward understanding the limits of gait recognition. In Anil K. Jain and Nalini K. Ratha, editors, *Biometric Technology for Human Identification*, volume 5404, pages 195–205. SPIE, aug 2004.

[29] Zongyi Liu, Laura Malave, and Sudeep Sarkar. Studies on silhouette quality and gait recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2004.

[30] Haiping Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. A layered deformable model for gait analysis. In *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, volume 2006, pages 249–256, 2006.

[31] Haiping Lu, Konstantinos N. Plataniotis, and Anastasios N. Venetsanopoulos. A full-body layered deformable model for automatic model-based gait recognition. *Eurasip Journal on Advances in Signal Processing*, 2008(1):1–13, nov 2008.

[32] Yasushi Makihara, Hidetoshi Mannami, Akira Tsuji, Md Altab Hossain, Kazushige Sugiura, Atsushi Mori, and Yasushi Yagi. The OU-ISIR gait database comprising the treadmill dataset. *IPSJ Transactions on Computer Vision and Applications*, 4:53–62, 2012.

[33] Sauradip Nag, Raghavendra Ramachandra, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Mohan Kankanhalli. CRNN based jersey-bib number/text recognition in sports and marathon images. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pages 1149–1156. IEEE Computer Society, sep 2019.

[34] Sauradip Nag, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Michael Blumenstein. A new unified method for detecting text from marathon runners and sports players in video (PR-D-19-01078R2). *Pattern Recognition*, 107, 2020.

[35] Yeshwanth Napolean, Priadi T. Wibowo, and Jan C. Van Gemert. Running event visualization using videos from multiple cameras. In *MMSports 2019 - Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, co-located with MM 2019*, volume 19, pages 82–90, New York, New York, USA, oct 2019. Association for Computing Machinery, Inc.

[36] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9914 LNCS, pages 17–35. Springer Verlag, 2016.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, dec 2015.

[38] Palaiahnakote Shivakumara, R. Raghavendra, Longfei Qin, Kiran B. Raja, Tong Lu, and Umapada Pal. A new multimodal approach to bib number/text detection and recognition in Marathon images. *Pattern Recognition*, 61:479–491, 2017.

[39] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 7347–7354. AAAI press, 2018.

[40] Xiaoxiao Sun and Liang Zheng. Dissecting Person Re-Identification From the Viewpoint of Viewpoint. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2019-June, pages 608–617. IEEE, jun 2019.

[41] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1), 2018.

[42] Md Zasim Uddin, Daigo Muramatsu, Noriko Takemura, Md Atiqur Rahman Ahad, and Yasushi Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Transactions on Computer Vision and Applications*, 11(1):9, dec 2019.

[43] Jin Wang, Mary She, Saeid Nahavandi, and Abbas Kouzani. A review of vision-based gait recognition methods for human identification. In *Proceedings - 2010 Digital Image Computing: Techniques and Applications, DICTA 2010*, pages 320–327, 2010.

[44] Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2):149–158, feb 2004.

[45] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, dec 2003.

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[47] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3D convolutional neural networks. In *Proceedings - International Conference on Image Processing, ICIP*, volume 2016-Augus, pages 4165–4169. IEEE Computer Society, aug 2016.

[48] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, 2017.

[49] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Speed-Invariant Gait Recognition Using Single-Support Gait Energy Image. *Multimedia Tools and Applications*, pages 26509–26536, 2019.

[50] CY Yam, MS Nixon, and JN Carter. Gait recognition by walking and running: a model-based approach. *Asian Conference on Computer Vision (ACCV)*, (January):1–6, 2002.

[51] C. Y.Chew Yean Yam, Mark S. Nixon, and John N. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, may 2004.

[52] Hong Xing Yu, Ancong Wu, and Wei Shi Zheng. Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 994–1002, 2017.

[53] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *Proceedings - International Conference on Pattern Recognition*, 4:441–444, 2006.

[54] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *arXiv preprint arXiv:2004.01888*, apr 2020.

[55] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A Video Benchmark for Large-Scale Person Re-Identification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS of *Lecture Notes in Computer Science*, pages 868–884. Springer International Publishing, Cham, 2016.

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 1116–1124, 2015.

[57] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984*, oct 2016.

[58] Wei Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 4678–4686, 2015.

[59] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking Person Re-identification with k-reciprocal Encoding. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:3652–3661, jan 2017.

2
Introduction

We introduce a large-scale long-distance running video dataset of the CampusRun [23], a running event with 257 recreational runners participating in the 5 km or 10 km distance race. These runners were captured using hand-held smartphone cameras, where each camera operator was allowed to move along the course. This work investigates the use of gait recognition for video person re-identification.

2.1. Motivation

The long-distance running domain is particular suited for the cross-camera person re-identification task, because events often include a large number of participants, unconstrained outdoor environments, strictly defined courses, prolonged periods of activity, pace variations and clothing similarities. Currently, bib number detection is the primary method for (re-)identifying long-distance runners [2, 4, 17, 23, 27], but video person re-identification and vision-based gait recognition approaches could be viable alternatives.

Recent gait recognition models [5, 11] use binary gait silhouettes to learn the identity of a person. One specific difference between walking and running is that we walk with straight arms and run with bent arms close to the body. Distinctiveness of the arm swing is mostly lost when extracting a silhouette representation of the runner, due to the ambiguity of the torso region. This happens especially when the runner is captured from a frontal or oblique view.

2.2. Research questions

The two main research questions are:

1. **Are gait features consistent enough for cross-camera person re-identification in a running event?** Given a sequence of a runner, can we find all the other sequences of this person?

2. **Can gait recognition performance be improved through the use of body-part-specific segmentation masks?** Can we use body-part-specific segmentation masks generated by a human semantic parsing model, to accentuate discriminative body parts in the silhouettes?

Sub-questions:

1. How to extract individual gait features when there are multiple persons in a frame?

2. What do gait features add to person re-identification?

3. Do partial gait silhouettes contain sufficient discriminative information for gait recognition?

2.3. Supplementary materials

The outline for the supplementary materials is as follows: chapter 3 gives a brief overview on gait recognition and our baseline model. Chapter 4 covers the CampusRun dataset and describes the data annotation process. Extra experiments are included in chapter 5.

3

Gait Recognition

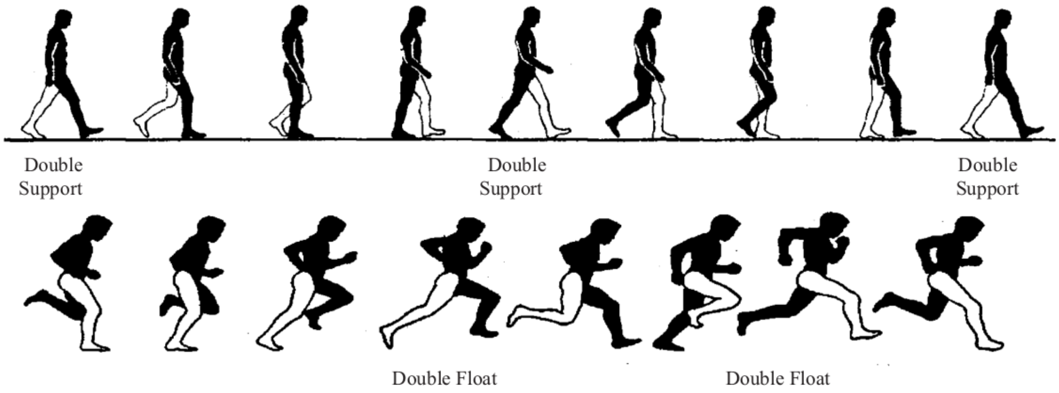


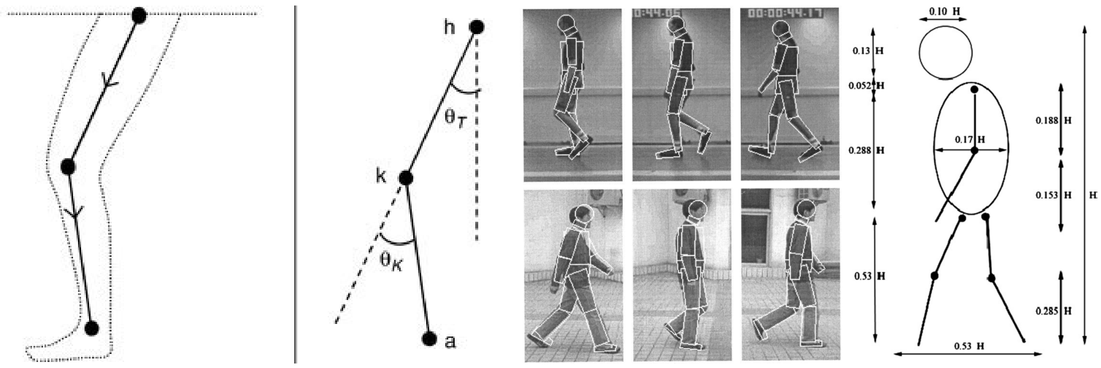
Figure 3.1: Comparison of a gait cycle for walking and running [36].

This chapter gives a brief overview of marker-free vision-based gait recognition and the baseline model.

3.1. Introduction

Human gait refers to the periodic movement and mechanics of walking and running. A gait cycle is defined as the distinctive phases before repetition occurs (figure 3.1). Gait is a bilateral symmetric movement where the arm and leg of opposite sides are swinging towards the same direction, with a phase shift of half a period between the left and right arm or left and right leg. A major difference between walking and running gait, is that a moment occurs during walking where both feet are touching the ground (double support), while in running both feet are of the ground momentarily (double float).

There have been numerous studies on vision-based gait recognition [18], as gait can be captured from a distance and at a low resolution. It is well acknowledged that human gait can be used as a biometric, as gait is found to be sufficiently unique for recognizing individuals [8]. Vision-based gait recognition is primarily divided into model-based and appearance-based approaches for extracting gait related features. Model-based methods model the human body to extract static and dynamic body parameters, such as stride parameters [3] or joint-angle trajectories [34]. Typical appearance-based methods use either a shape representation in the form of binary gait silhouettes [33], or a spatio-temporal variant: the gait energy image (GEI), which is the average over one gait cycle of silhouettes [13]. The gait representations are then used for further feature extraction or feature selection, followed by classification of the gait signature or metric learning to distinguish between identities.



(a) Structural model [7, 37] (b) Tracking of the limbs [34] (c) Model proportions [28]

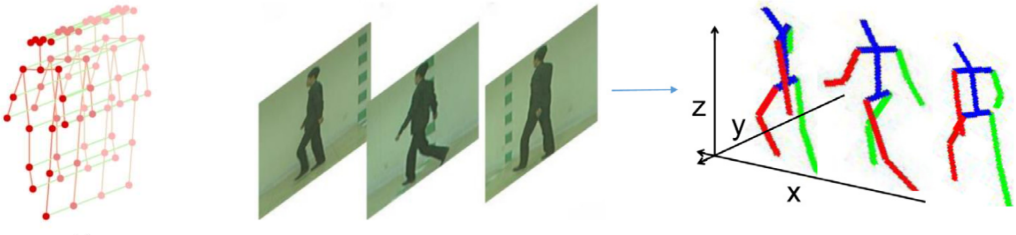
Figure 3.2: Model-based methods.

3.2. Model-based

Model-based approaches approximate gait using structural and motion models to extract gait related features. The model utilize static (e.g. height, stride length) and dynamic features (e.g. frequency, joint angle trajectories) and describes the relationships between them. An advantage of model-based approaches is the robustness against (self-)occlusion, noise and scale. Figure 3.2 shows some examples of model-based approaches.

Cunado et al. [7] uses Fourier series to describe the motion of the hip and thigh, while using the Velocity Hough Transform (VHT) [24, 25] to extract motion parameters from a sequence of images. Wang et al. [34] uses a human body model of 14 rigid body parts with walking specific constraints, to track the walker and extract the motion parameters. Joint-angle trajectories are compared with Euclidean distance between identities. Tafazzoli and Safabakhsh [28] extracts the movements of joint positions of the legs and arms, using a contour model [19] and Hough transform [10] to find body contours and line segments.

More recent deep-learning approaches extract gait features from human body joints with 2D [20] and 3D human pose estimation [22]. Li et al. [20] constructs gait graphs and graph convolutional neural networks [9] to extract spatio-temporal gait features. Liao et al. [22] estimates 3D pose from 2D pose to increase robustness against view variations.



(a) Gait graphs [20] (b) 3D pose estimation [22]

Figure 3.3: Deep learning model-based methods.

3.3. Appearance-based

Appearance-based approaches rely on shape or motion characteristics of human body silhouettes [33]. Early approaches used statistical features to compare two sequences of silhouettes [32]. Wang et al. [31] applied Procrustes shape analysis to silhouettes for obtaining the gait signature. Spatio-temporal variants such as Gait Energy Image (GEI) [13] calculated the average over one cycle of gait silhouettes. Aggregating the silhouettes into a GEI representation reduces the amount of computation while preserving the static and dynamic characteristics of the gait. The aggregated representation is also robust against small segmentation errors. Disadvantages of GEI is that recognition accuracy

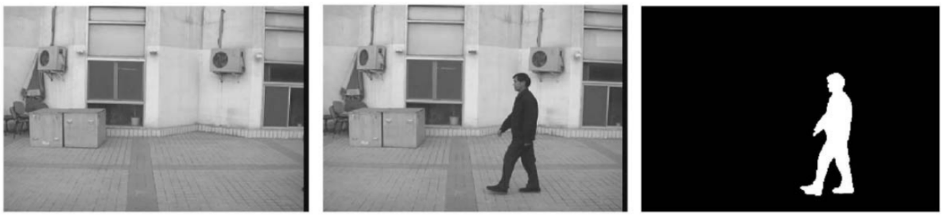


Figure 3.4: Background subtraction to obtain gait silhouette [33].

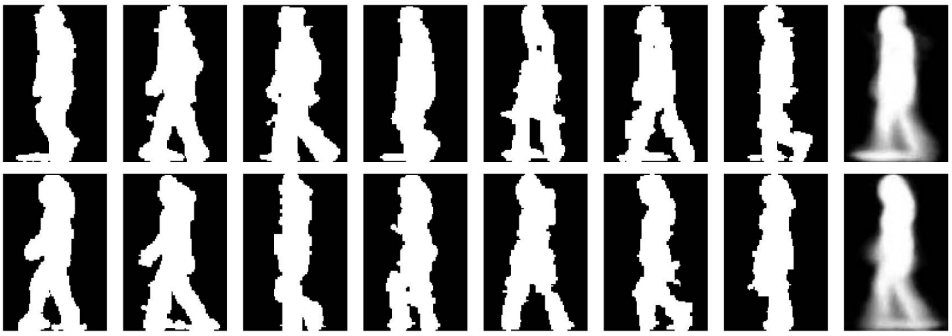


Figure 3.5: Averaging over one cycle of gait silhouettes to get the Gait Energy Image (GEI) [13].

drops considerable under covariate conditions [16], due to being sensitive to the shape of the static gait component.

Variations on the GEI were proposed to increase recognition performance when facing conditions such as occlusions [6] or covariates [1]. The Gait Entropy Image (GEnI) [1] was introduced to focus more on the dynamic components of the GEI, by measuring Shannon entropy at each pixel location in the GEI. This way, the dynamic components of the GEI are highlighted in the GEnI. With Chrono-Gait Image (CGI) [30], a temporal gait template is created using color mapped contours of the gait silhouettes, to preserve temporal information in the gait sequence.

Recent silhouette-based methods use convolutional neural networks with binary gait silhouettes [5] or GEI as input [35], to learn the discriminative gait features.

3.4. GaitSet

Representing a gait sequence as a set of gait silhouettes was proposed in GaitSet [5]. The authors argue that learning the identity of a person from a set of gait silhouettes has two advantages. Firstly, the gait sequence can have arbitrary length and secondly, the order of silhouettes does not matter as it can be assumed that the shape of a silhouette indicates its position in the gait cycle.

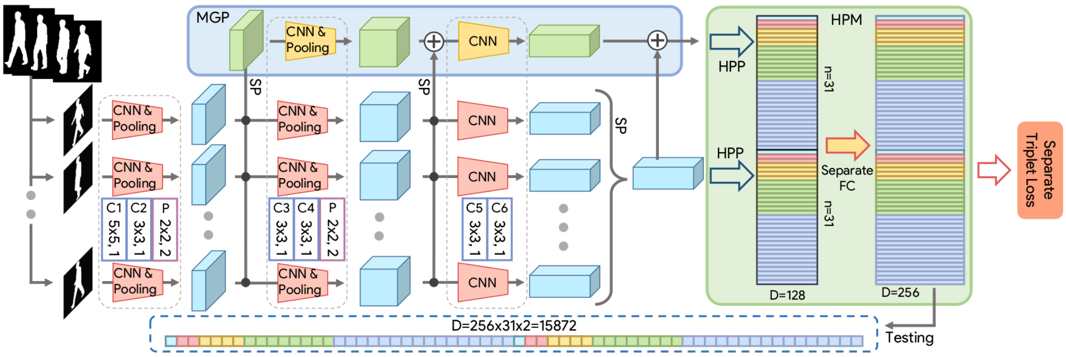


Figure 3.6: GaitSet model overview [5].

The network (figure 3.6) first extracts frame-level features and then aggregates the feature maps of

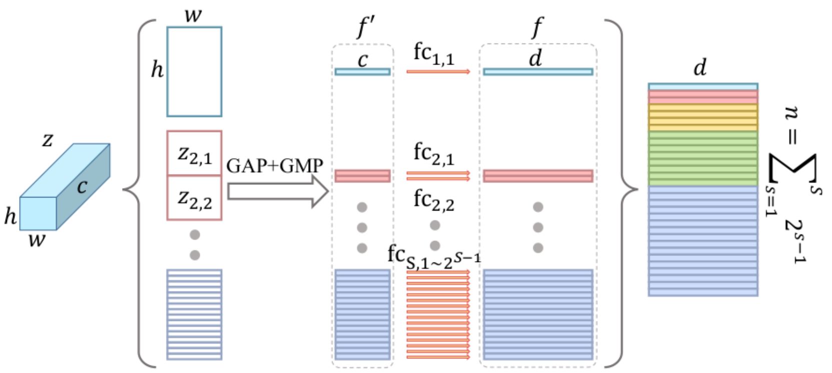


Figure 3.7: Horizontal Pyramid Pooling [5].

each silhouette using max pooling on the set-level (SP). The Multilayer Global Pipeline (MGP) takes feature maps from different layers in the network to learn features with different receptive fields. Horizontal Pyramid Pooling (HPP) [12] 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 (figure 3.7). For each set of silhouettes, the network outputs a discriminative representation, consisting of 62 feature map strips with 256 dimensions each.

GaitSet is trained with Batch All triplet loss [15], where all possible combinations of triplets 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 persons, k the number of tracklets for each person and c the number of frames for each tracklet. The trained model is robust against covariates such as varying view angle or carrying conditions when sampling batches during training with multiple covariate conditions.

3.5. Datasets

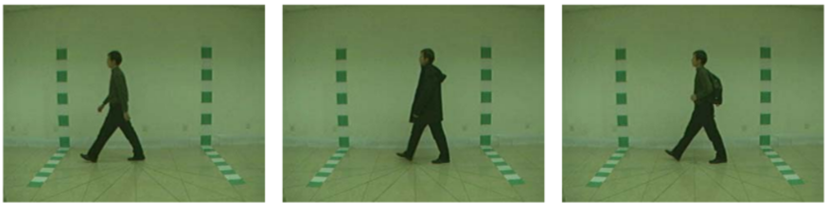


Figure 3.8: Different walking conditions: normal (NM), wearing a coat (CL) and carrying a bag (BG) [38].

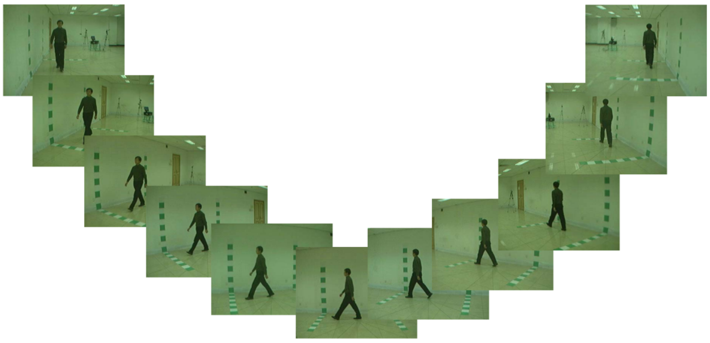


Figure 3.9: 11 views (0° - 180°) [38].

Two of the most popular gait datasets are CASIA-B [38] and OU-MVLP [29] for the cross-view gait recognition task. CASIA-B contains 124 participants who are captured from 11 view angles (figure 3.9) while walking with different walking conditions (figure 3.8). OU-MVLP [29] is a large population dataset with 14 view angles and 10307 participants.

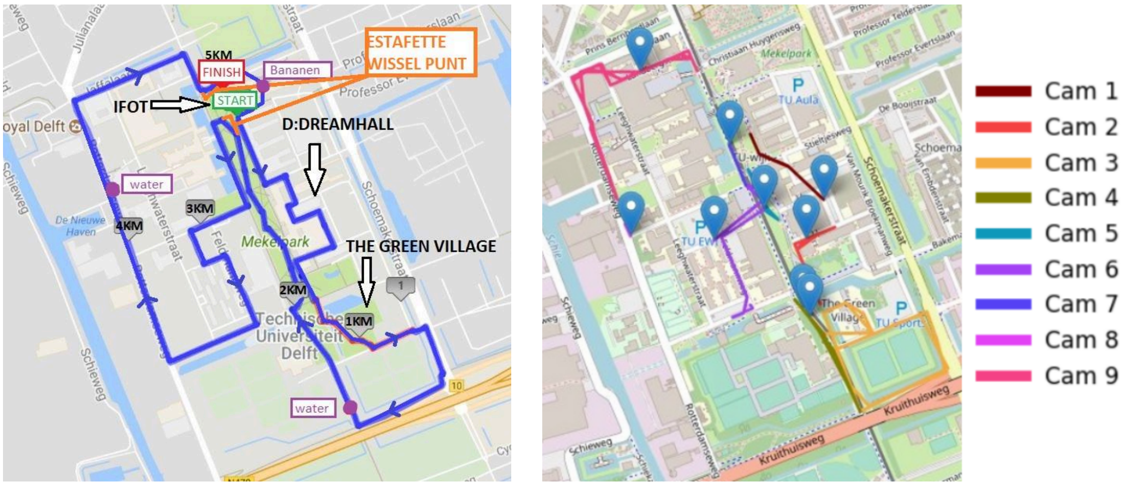
4

CampusRun Dataset

This chapter describes the CampusRun dataset and annotation process.

4.1. Introduction

The CampusRun video dataset was introduced in [23]. It is a long-distance running event where 262 recreational runners simultaneously ran a 5 km course, while competing in various race distances. 128 persons participated in the 5 km distance race, while 134 runners ran two laps around the course as part of the 10 km distance. The runners were captured on video using 9 non-stationary hand-held smartphone cameras, where each camera operator was allowed to move along the course. There is around 3 hours of video footage. We number each camera viewpoint as 1-9 and additionally give it the letter A for the first lap and the letter B for second lap. This gives us 18 camera points (A1-A9, B1-B9) for designating where a runner was observed. The course layout and camera trajectories are shown in figure 4.1.



(a) Course layout

(b) Camera trajectories [23]

Figure 4.1: CampusRun course layout and moving camera locations.

4.2. Multi-object tracking

Labels are provided by [23] in the form of start-frame and end-frame annotations that a runner appears in a specific video and some bounding boxes for the 5 km runners. To extend the dataset with the 10 km runners, we use an off-the-shelf multi-object tracker [39] on the raw videos to extract tracklets and bounding boxes for all the runners. The multi-object tracker finds 5204 local identities from 455 videos.

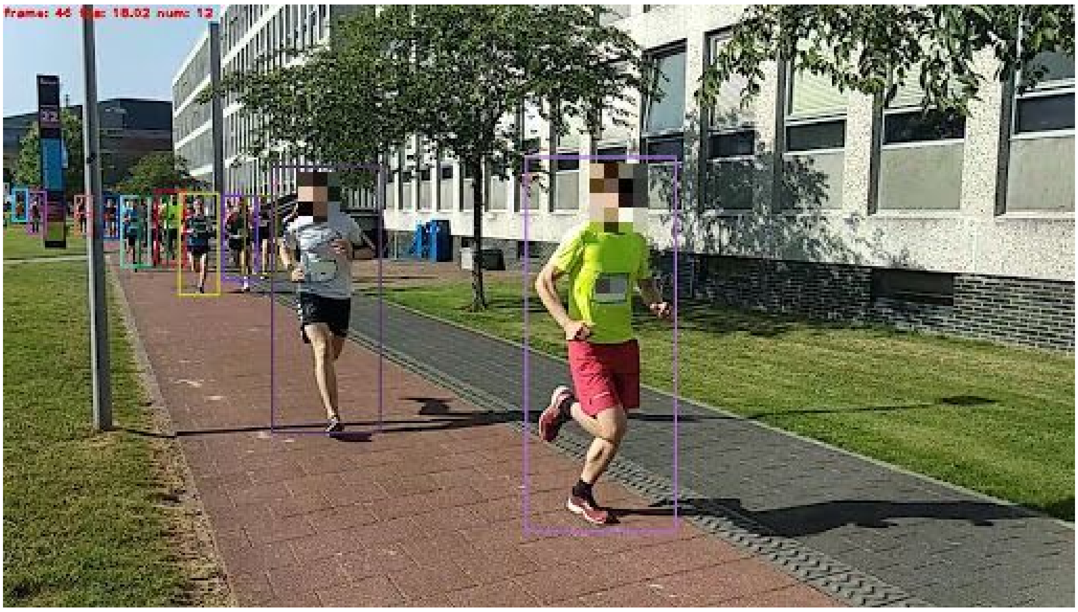
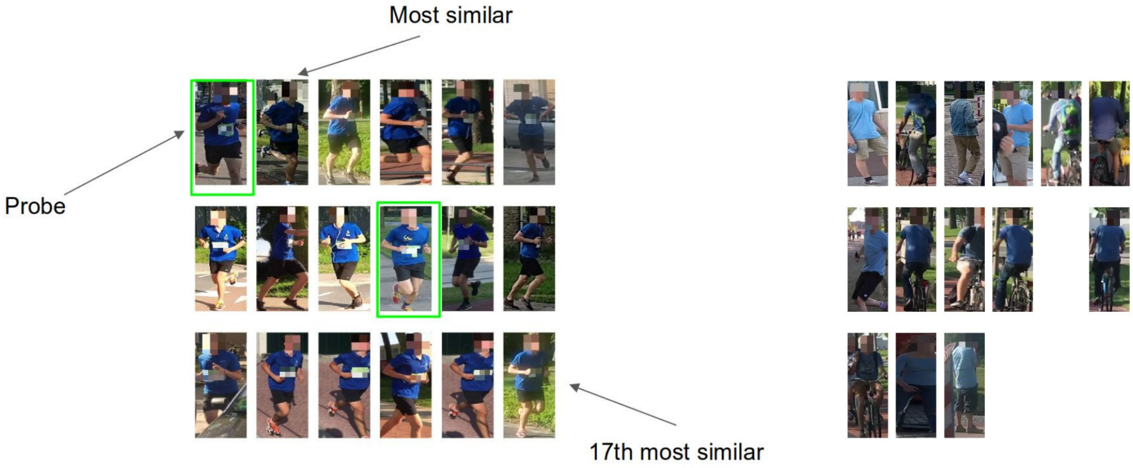


Figure 4.2: Example output from the multi-object tracker [39]

Many tracklets are relay runners, who are not part of the 5 km and 10 km race, and non-participants such as pedestrians.

4.3. Manual annotation of bib numbers

The 5204 local identities are from 18 cameras, and we need to match the local identities back to the bib numbers of the 262 participants. We first extract the longest consecutive sequence for each of the 5204 local identities. Then, we take the bounding box that has the highest confidence according to the multi-object tracker. The resulting 5204 bounding boxes are evaluated in a cross-camera setting using an off-the-shelf person re-identification model [40] to get the ranked retrieval results. We first fine-tune the model with the bounding boxes of the 5 km runners provided by [23].



(a) Grid with top-17 ranked list with one correct match.

(b) Grid with non-participants

Figure 4.3: Grid view for manual annotation. Manual annotation of the matches are shown in green.

The manual annotation setup is as follows: first we take the last camera (A9, B9) as the probe subset (n=654) and the other 8 cameras (A1-A8, B1-B8) in the gallery (n=4550). A participant can only appear in 17 cameras, therefore we show the top-17 most similar bounding boxes to our query in a grid view (figure 4.3) and manual annotate the matches. The person re-identification model and grid view makes the annotation process efficient. Despite the fact that the retrieval results are similar looking in clothing style and colour, it is easy to spot the correct matches to the query image when viewing multiple at the same time in a grid view. Furthermore, we assign labels to multiple instances at the same time and label non-participants to remove them later from the set of tracklets.

After manual annotation of the bib numbers, the labels of the bounding boxes are then assigned to their respective tracklet to end up with fully annotated sequences. We obtain bounding boxes for 257 runners and 2581 tracklets with an average sequence length of 77 frames. The 10 km runners have 13 tracklets on average. Five registered participants were not found during the tracking and annotation process.

4.4. 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, n=125, 9 cameras A1-A9, 860 tracklets). The test set (10 km, n=132, 18 cameras A1-A9, B1-B9, 1721 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. During evaluation, each tracklet is evaluated once as the probe subset, with every other tracklet in the gallery subset. We have 1721 test queries with a maximum of 17 positive matches for each query, as the runners do not appear more than once per camera. Because of the high number of ground truths, we evaluate the models using mean average precision (mAP) and rank-1 accuracy.

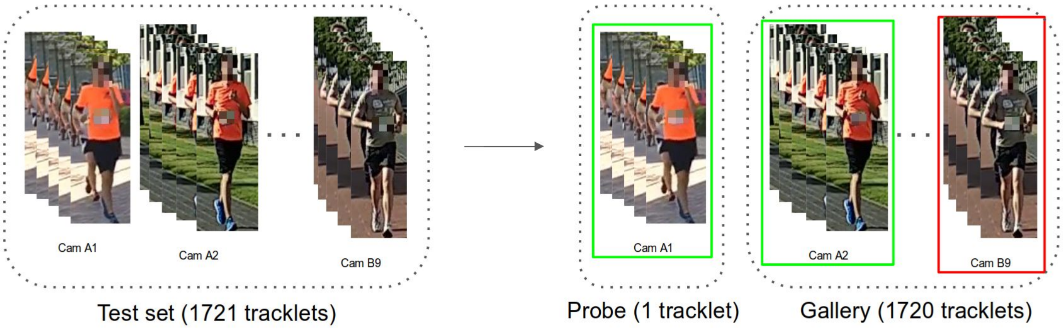


Figure 4.4: Cross-camera evalutation protocol using the 10 km runners.

5

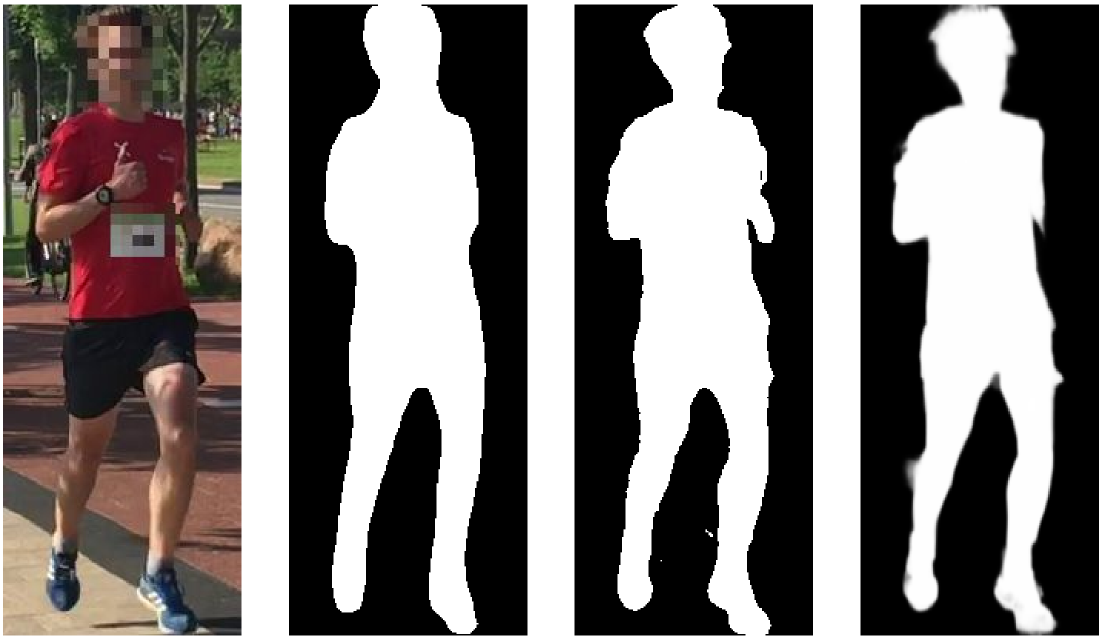
Experiments

This chapter describes experimental results and illustrations that did not make it into the scientific paper.

5.1. Segmentation method

5.1.1. CampusRun

Figure 5.1 shows a comparison between segmentation methods to create gait silhouettes. We use instance segmentation [14], human semantic parsing [21] and salient object detection [26]. In the scientific paper, it was demonstrated that mean average precision increased from 39.3 to 52.2 when using the silhouettes from human semantic parsing instead of instance segmentation. The resulting silhouettes from human semantic parsing contain more details than the instance segmentation method. Another option is to fine-tune the segmentation model for the intended domain before creating the gait silhouettes. Future research could examine the use of salient object detection, as it was able to extract fine details of the human body.

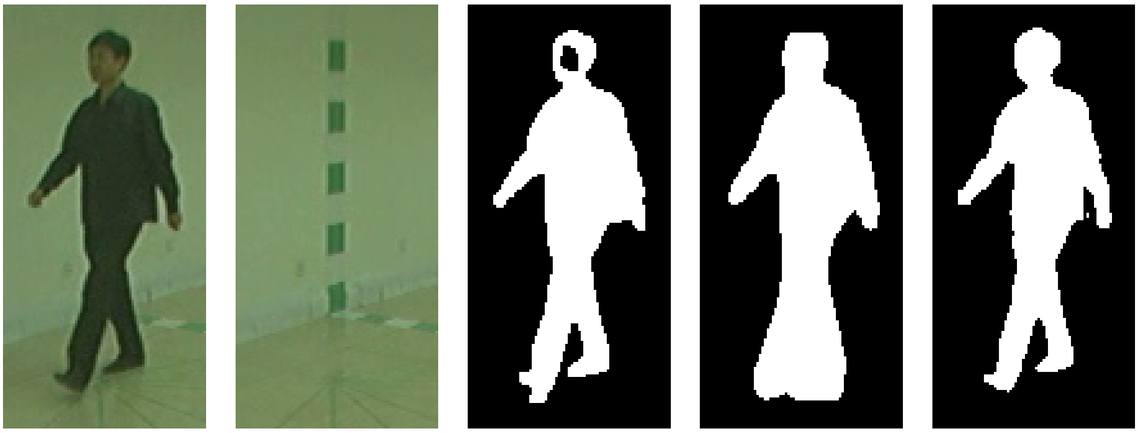


(a) Bounding box (b) Mask R-CNN [14] (c) SCHP [21] (d) U2-Net [26]

Figure 5.1: **Comparison between segmentation methods.**
(a) bounding box, (b) instance segmentation, (c) human semantic parsing, (d) salient object detection.

5.1.2. CASIA-B

Figure 5.2 shows a comparison between segmentation methods on the CASIA-B dataset [38]. The original CASIA-B silhouettes, which are included in the dataset, are created using background subtraction. It was observed for certain views that the calibration strip in the background, shown in figure 5.2(b), causes a hole in the head when using background subtraction. This observation was found in most silhouettes in the dataset, which can be corrected with morphological operators.



(a) (b) (c) (d) Mask R-CNN [14] (e) SCHP [21]

Figure 5.2: **Comparison between segmentation methods.**
(a) bounding box, (b) background, (c) original silhouette, (d) instance segmentation, (e) human semantic parsing.

We were interested in seeing if the GaitSet [5] model would pick up on this cue during training, which would mean that it exploits additional background information to identify certain views or identities. We compare the original silhouettes without pre-processing, with silhouettes consisting of the summation of the original silhouettes and human semantic parsing. Adding both silhouettes together results in clean silhouettes without artifacts. If the model exploits the background information, then we would expect the rank-1 accuracy to drop with the new silhouettes. The results are shown in table 5.1, which indicate that there is no large drop in accuracy, suggesting that the model does not exploit the background noise or the effect is minor.

Silhouettes	NM	BG	CL
Original	95.2	87.5	70.0
Original + human semantic parsing	94.1	84.7	72.1

Table 5.1: Comparison between the original silhouettes and adding human semantic parsing.

5.2. Activation maps

5.2.1. CASIA-B

We investigate activation maps [40] of the last convolutional layer of GaitSet to see the influence of certain regions in the silhouettes. The activation maps are computed by taking the sum of the absolute-valued feature maps along the channel dimension, followed by a spatial L2 normalization, upsampling and min-max scaling. The activation maps are shown in figure 5.3 for the silhouettes from the previous experiment. The model seems to utilize the shape and movement of the full body with some focus on the leg region. It is interesting that for the original silhouettes, there is a small spot around the shoulder and head region where there are higher activations, while this behavior is not present in the cleaner silhouettes. This is perhaps related to the artifacts in the head region due to the background (figure 5.2).

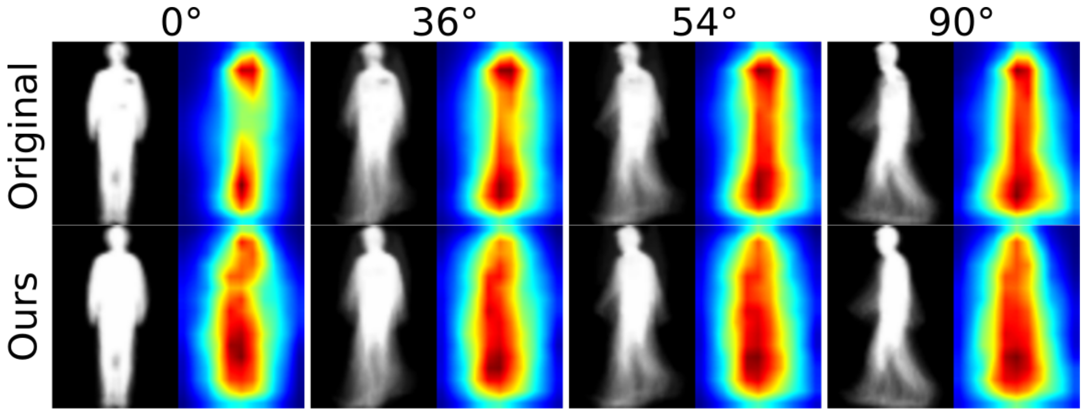
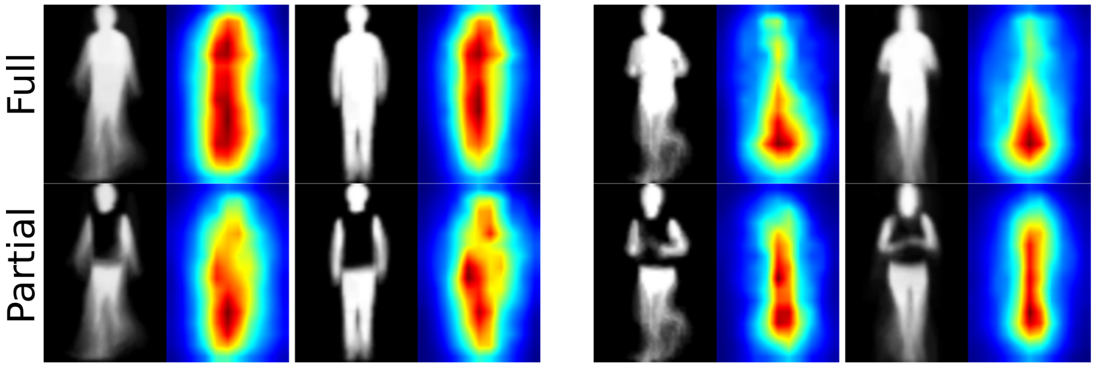


Figure 5.3: **Comparison between the original silhouettes (top) and adding human semantic parsing (bottom).** Activations of the last convolutional feature maps. Gait samples are depicted here as averaged silhouettes, but the input is a set of silhouettes.

5.2.2. Walking and running

Figure 5.4 shows the activation maps for both CASIA-B and CampusRun sequences with full body and partial silhouettes. The activation maps indicate that there are more activations in the leg region when running compared to walking. Removing the torso from the silhouettes results in higher activations in the torso and arm region.



(a) Walking

(b) Running

Figure 5.4: **Comparison between walking and running.**
Activations of the last convolutional feature maps for both walking and running sequences. All silhouettes are constructed using human semantic parsing. Gait samples are depicted here as averaged silhouettes, but the input is a set of silhouettes.

5.3. Retrieval results

Extra retrieval results for the GaitSet model on the CampusRun are shown in figure 5.5. For the first query, the person of interest is captured often with his hands waving in the air, which is why four samples are correctly retrieved. In the second query, the person of interest is captured with their back towards the camera, but the rank-5 results are all correct. This demonstrates that the model is robust against view changes. The third and fourth query demonstrates that crowded scenes and occlusions are challenges in unconstrained gait recognition. The gait silhouettes are incomplete for both queries due to being occluded by other runners. The retrieved ranked results are all occluded sequences and incorrect matches.

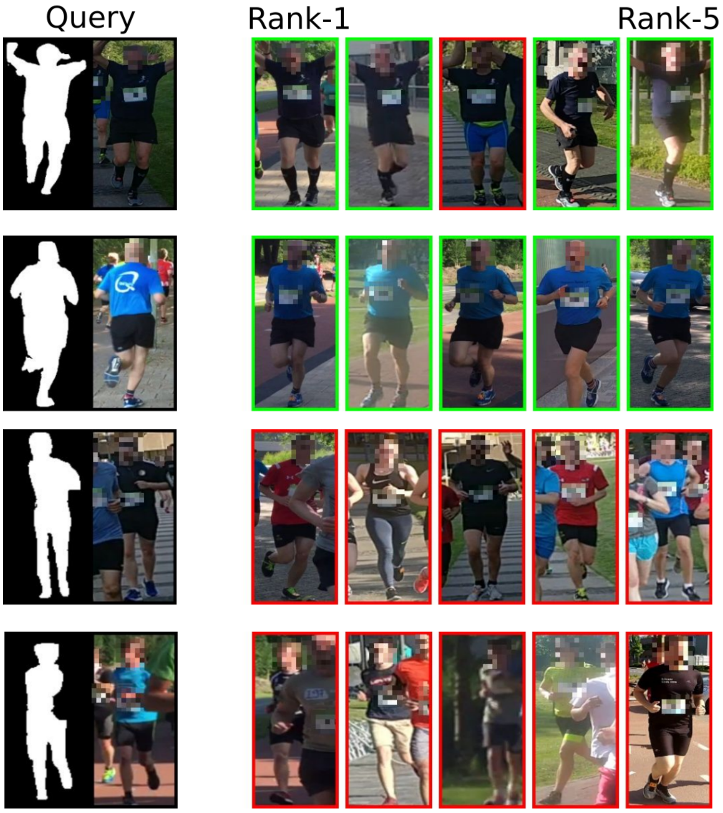


Figure 5.5: **Example retrieval results.**
Four queries and their corresponding rank-5 retrieval results for the GaitSet model.

Bibliography

[1] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, oct 2010. ISSN 01678655. doi: 10.1016/j.patrec.2010.05.027.

[2] Idan Ben-Ami, Tali Basha, and Shai Avidan. Racing bib number recognition. In *BMVC 2012 - Electronic Proceedings of the British Machine Vision Conference 2012*, 2012. doi: 10.5244/C.26.19.

[3] Aaron F. Bobick and Amos Y. Johnson. Gait recognition using static, activity-specific parameters. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2001. doi: 10.1109/cvpr.2001.990506.

[4] Noppakun Boonsim. Racing bib number localization on complex backgrounds. *WSEAS Transactions on Systems and Control*, 13:226–231, 2018. ISSN 22242856.

[5] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(16):8126–8133, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33018126.

[6] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, aug 2009. ISSN 01678655. doi: 10.1016/j.patrec.2009.04.012.

[7] David Cunado, Mark S. Nixon, and John N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, apr 2003. ISSN 10773142. doi: 10.1016/S1077-3142(03)00008-0.

[8] James E Cutting and Lynn T Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, 1977. ISSN 00905054. doi: 10.3758/BF03337021.

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016. URL https://github.com/mdeff/cnn{_}graph.

[10] Richard O. Duda and Peter E. Hart. Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Communications of the ACM*, 15(1):11–15, jan 1972. ISSN 15577317. doi: 10.1145/361237.361242. URL https://dl.acm.org/doi/10.1145/361237.361242.

[11] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221. IEEE, jun 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.01423. URL https://github.com/ChaoFan96/GaitPart.https://ieeexplore.ieee.org/document/9156784/.

[12] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal Pyramid Matching for Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8295–8302, apr 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33018295. URL http://arxiv.org/abs/1804.05275.

[13] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006. ISSN 01628828. doi: 10.1109/TPAMI.2006.38. URL https://ieeexplore.ieee.org/abstract/document/1561189/.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):386–397, feb 2020. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2844175. URL https://ieeexplore.ieee.org/document/8372616/.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, mar 2017. URL http://arxiv.org/abs/1703.07737.

[16] Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, and Yasushi Yagi. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521, 2012. ISSN 15566013. doi: 10.1109/TIFS.2012.2204253.

[17] Kamlesh, Pei Xu, Yang Yang, and Yongchao Xu. Person re-identification with end-to-end scene text recognition. In *Communications in Computer and Information Science*, volume 773, pages 363–374. Springer Verlag, oct 2017. ISBN 9789811073045. doi: 10.1007/978-981-10-7305-2_32. URL https://doi.org/10.1007/978-981-10-7305-2{_}32.

[18] Tracey K. M. Lee, Mohammed Belkhatir, and Saeid Sanei. A comprehensive review of past and present vision-based techniques for gait recognition. *Multimedia Tools and Applications*, 72(3): 2833–2869, oct 2014. ISSN 1380-7501. doi: 10.1007/s11042-013-1574-x. URL http://link.springer.com/10.1007/s11042-013-1574-x.

[19] Chunming Li, Chiu Yen Kao, John C. Gore, and Zhaohua Ding. Implicit active contours driven by local binary fitting energy. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. ISBN 1424411807. doi: 10.1109/CVPR.2007.383014.

[20] Na Li, Xinbo Zhao, and Chong Ma. A model-based Gait Recognition Method based on Gait Graph Convolutional Networks and Joints Relationship Pyramid Mapping. *arXiv preprint arXiv:2005.08625*, apr 2020. URL http://arxiv.org/abs/2005.08625.

[21] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-Correction for Human Parsing. *arXiv preprint arXiv:1910.09777*, oct 2019. URL http://arxiv.org/abs/1910.09777.

[22] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. ISSN 00313203. doi: 10.1016/j.patcog.2019.107069. URL https://doi.org/10.1016/j.patcog.2019.107069.

[23] Yeshwanth Napolean, Priadi T. Wibowo, and Jan C. Van Gemert. Running event visualization using videos from multiple cameras. In *MMSports 2019 - Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, co-located with MM 2019*, volume 19, pages 82–90, New York, New York, USA, oct 2019. Association for Computing Machinery, Inc. ISBN 9781450369114. doi: 10.1145/3347318.3355528. URL http://dl.acm.org/citation.cfm?doid=3347318.3355528.

[24] J. M. Nash, J. N. Carter, and M. S. Nixon. Extraction of Moving Articulated-Objects by Evidence Gathering. In *Procedings of the British Machine Vision Conference 1998*, pages 61.1–61.10. British Machine Vision Association, 1998. ISBN 1-901725-04-9. doi: 10.5244/C.12.61. URL https://eprints.soton.ac.uk/251959/1/P083.PDFhttp://www.bmva.org/bmvc/1998/papers/d083/h083.htm.

[25] Jason M. Nash, John N. Carter, and Mark S. Nixon. Velocity Hough Transform: A new technique for dynamic feature extraction. In *IEEE International Conference on Image Processing*, volume 2, pages 386–389. IEEE Comp Soc, 1997. doi: 10.1109/icip.1997.638786.

[26] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106:107404, oct 2020. ISSN 00313203. doi: 10.1016/j.patcog.2020.107404.

[27] Palaiahnakote Shivakumara, R. Raghavendra, Longfei Qin, Kiran B. Raja, Tong Lu, and Umapada Pal. A new multi-modal approach to bib number/text detection and recognition in Marathon images. *Pattern Recognition*, 61:479–491, 2017. ISSN 00313203. doi: 10.1016/j.patcog.2016.08.021. URL `http://dx.doi.org/10.1016/j.patcog.2016.08.021`.

[28] Faezeh Tafazzoli and Reza Safabakhsh. Model-based human gait recognition using leg and arm movements. *Engineering Applications of Artificial Intelligence*, 23(8):1237–1246, dec 2010. ISSN 09521976. doi: 10.1016/j.engappai.2010.07.004.

[29] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1), 2018. ISSN 18826695. doi: 10.1186/s41074-018-0039-6.

[30] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. Chrono-Gait Image: A Novel Temporal Template for Gait Recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6311 LNCS, pages 257–270. Springer Verlag, 2010. ISBN 3642155480. doi: 10.1007/978-3-642-15549-9_19. URL `https://link.springer.com/chapter/10.1007/978-3-642-15549-9{_}19http://link.springer.com/10.1007/978-3-642-15549-9{_}19`.

[31] Liang Wang, Huazhong Ning, Weiming Hu, and Tieniu Tan. Gait recognition based on procrustes shape analysis. In *IEEE International Conference on Image Processing*, volume 3, 2002. doi: 10.1109/icip.2002.1038998.

[32] Liang Wang, Tieniu Tan, Weiming Hu, and Huazhong Ning. Automatic gait recognition based on statistical shape analysis. *IEEE Transactions on Image Processing*, 12(9):1120–1131, sep 2003. ISSN 10577149. doi: 10.1109/TIP.2003.815251.

[33] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette Analysis-Based Gait Recognition for Human Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, dec 2003. ISSN 01628828. doi: 10.1109/TPAMI.2003.1251144.

[34] Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2):149–158, feb 2004. ISSN 10518215. doi: 10.1109/TCSVT.2003.821972.

[35] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2545669.

[36] C. Y.Chew Yean Yam, Mark S. Nixon, and John N. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, may 2004. ISSN 00313203. doi: 10.1016/j.patcog.2003.09.012.

[37] Chew-Yean Yam and Mark S. Nixon. Gait Recognition, Model-Based. In *Encyclopedia of Biometrics*, pages 633–639. Springer US, 2009. doi: 10.1007/978-0-387-73003-5_37. URL `https://link.springer.com/referenceworkentry/10.1007/978-0-387-73003-5{_}37`.

[38] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *Proceedings - International Conference on Pattern Recognition*, 4:441–444, 2006. ISSN 10514651. doi: 10.1109/ICPR.2006.67.

[39] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *arXiv preprint arXiv:2004.01888*, apr 2020. URL `http://arxiv.org/abs/2004.01888`.

[40] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 3701–3711, 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00380. URL `https://github.com/`.