Gait Recognition using 3D View-Transformation Model*

Philipp Schwarz[1][0000−0002−8364−4850], Philipp Hofer[3][0000−0002−7705−9938], and Josef Scharinger[2][0000−0001−6502−7501]

[1] LIT Secure and Correct Systems Lab, Johannes Kepler University, Linz, Austria
philipp.schwarz@jku.at
[2] Institute of Networks and Security, Johannes Kepler University, Linz, Austria
philipp.hofer@ins.jku.at
[3] Institute of Computational Perception, Johannes Kepler University, Linz, Austria
josef.scharinger@jku.at

Abstract. When it comes to visual based gait recognition, one of the biggest problems is the variance introduced by different camera viewing angles. We generate 3D human models from single RGB person image frames, rotate these 3D models into the side view, and compute gait features used to train a convolutional neural network to recognize people based on their gait information. In our experiment we compare our approach with a method that recognizes people under different viewing angles and show that even for low-resolution input images, the applied view-transformation 1) preserves enough gait information for recognition purposes and 2) produces recognition accuracies just as high without requiring samples from each viewing angle. We believe our approach will produce even better results for higher resolution input images. As far as we know, this is the first appearance-based method that recreates 3D human models using only single RGB images to tackle the viewing-angle problem in gait recognition.

Keywords: 3D Gait Recognition · View-Transformation Model · Convolutional Neural Network

* The research reported in this paper has been partly supported by the LIT Secure and Correct Systems Lab funded by the State of Upper Austria and by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH. This work has been carried out within the scope of Digidow, the Christian Doppler Laboratory for Private Digital Authentication in the Physical World. We gratefully acknowledge financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, 3 Banken IT GmbH, Kepler Universitätsklinikum GmbH, NXP Semiconductors Austria GmbH & Co KG, Österreichische Staatsdruckerei GmbH, and the State of Upper Austria.

1 Introduction

In the past years, gait has become an increasingly important modality for person recognition. While there are various sensors that can capture gait information, predominantly visual-based data such as videos are used.
The human gait has some properties that make it very attractive for recognition purposes such as its long distance recognizability, not requiring physical contact or not requiring user cooperation. On the downside, there are a lot of factors which influence gait recognition negatively. For instance walking speed, viewing-angle, injuries, clothing variations or carrying objects. One of the biggest problems regarding visual based gait recognition is the variance introduced by different camera viewing angles. The similarity between two different persons viewed under the same viewing angle is often bigger than the similarity between two images of the same person under different viewing angles.
There exist several approaches that deal with the viewing angle problem. We differentiate between approaches based on view-invariant features and approaches based on View Transformation Models (VTM). The former attempts to compute features which are not or hardly affected by different viewing angles, while the latter seeks to reconstruct features of a different viewing angle. This work describes a VTM-based approach.

2 Proposed Method

We use low resolution RGB images of the commonly used CASIA-B [13] gait dataset as input for our approach. After preprocessing the images we compute a 3D model from a single RGB image for each image of this dataset by applying PIFuHD [10]. We rotate the 3D model into a 90 degree view as it contains the most discriminating person information and project it onto a 2D plane. The projection is only a means to reduce the size of our features since 3D objects are generally too big to fit in memory when it comes to training a neural network. Based on these 2D representations we compute a feature similar to the well-known Gait Energy Image (GEI)[4] which serves as input to a Convolutional Neural Network (CNN). The last layer in the CNN classifies the input from which we compute the recognition accuracy.
Generating accurate 3D human models with PIFuHD requires high resolution images. Contrarily, one of the advantages of gait as a biometric modality is its recognizability under low resolution conditions. In order to close the gap between those contradicting requirements, we perform several preprocessing steps on our low resolution input data before we can create 3D human models. To increase recognition accuracy we also perform postprocessing steps on the projection of the 3D model before computing gait features. An overview of the whole processing pipeline can be viewed in Fig. 1. Each process is described in more detail in Section 2.2. To illustrate the improvements gained by our preprocessing measures Fig. 2 shows an unprocessed RGB image of the CASIA-B dataset as well as the resulting 3D model seen from the original perspective and from the front.

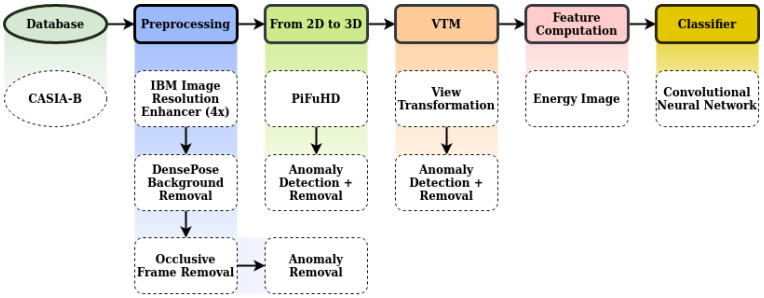

Fig. 1. Our proposed processing pipeline for 3D based Gait Recognition.

In the spirit of consistency and comparability we use the same frame for the following figures.

2.1 Dataset

In this work we use data from the publicly available dataset CASIA-B. It contains ten walking sequences for each of the 124 subjects. Six walks were recorded under normal conditions, two under carrying conditions and two under different clothing conditions. The walking sequences are captured by 11 cameras from 11 equidistant viewing angles. The images of this dataset have a resolution of 240x320px.

2.2 Preprocessing

To produce recognizable 3D models, the following processing steps are applied.

- **Image Resolution Enhancement**
  To artificially increase the resolution of our input images, we use IBM's Image

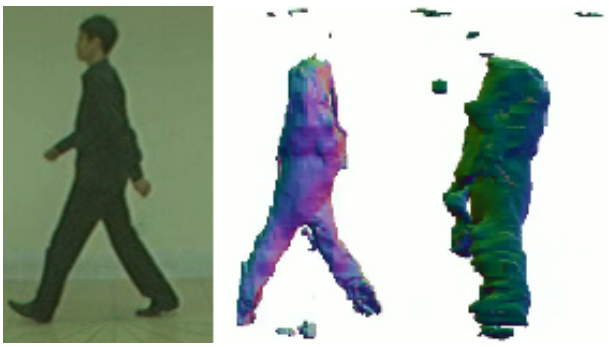

Fig. 2. Raw input image from CASIA-B dataset (left) with output of PIFuHD under same view (middle) and front view (right).

Resolution Enhancer [1]. It allows enhancing the resolution by a factor of up to four, which results in images of size 960x1280px. Fig. 3 shows the resolution enhanced image as well as the generated 3D model. Compared to Fig. 2 this preprocessing step does not seem to improve the quality of our model, unless combined with the next step in our pipeline.

Fig. 3. Upsampled RGB image (left) with output of PIFuHD under same view (middle) and front view (right).

- **DensePose Background Removal**
  Removing only the background (without the previous resolution enhancement step) already visibly improves the quality of the 3D model as is shown in Fig. 4. We perform background subtraction using the silhouette estimated by the person detection framework introduced in [3]. We can generate even

Fig. 4. Image with background removed (left) with output of PIFuHD under same view (middle) and front view (right).

more accurate 3D models when we both enhance the resolution and remove the background. This is illustrated in Fig. 5.

Fig. 5. Image upsampled and background removed (left) with output of PIFuHD under same view (middle) and front view (right).

- **Occlusive Frame Removal**
  Since we are not focusing on gait recognition under occlusion in this paper, we exclude frames that do not fully capture a persons silhouette. As we are working with data recorded under lab-conditions, we simply remove frames where a person 1) just enters or exits, 2) is too close to the camera or 3) is too far away to be recognized as such by the person detection framework mentioned above.
- **DensePose Anomaly Removal**
  Since the DensePose person detection framework is not extremely precise, we have to deal with artifacts. Therefore, we implement a threshold-based mechanism that removes small pixel islands that are unrealistically far away from a person but have falsely been labeled as part of a person.

2.3 From 2D to 3D

We let PIFuHD [10] process the preprocessed images to obtain 3D human models which we have already displayed in the figures above. PIFuHD is based on PIFu [9], which is a framework that produces high-resolution 3D models from either single 2D images or multiple 2D images from different perspectives. PIFuHD extends the coarse encoder of PIFu with an additional encoder for fine details. This work only leverages the single input image capability of PIFuHD.

2.4 View Transformation and Feature Computation

We let a camera rotate around the center of the 3D human models until it reaches the 90 degree view. No viewing-angle estimation is done because the original

viewing-angle is provided by CASIA-B. Rotating the 3D model can sometimes reveal artifacts which were not visible before, especially when the difference between the source view and the destination view (90 degree view) becomes larger. We differentiate between severe and non-severe artifacts. Accordingly, we implement an algorithm to get rid of the 3D model in case the artifacts are too severe and another algorithm to remove artifacts in the form of disconnected islands of pixel thereby preserving the 3D model.
From the person images, which are now viewed under a 90 degree angle, we compute a variation of the well-known GEI according to [11]. Additionally, no gait cycles are computed, we just take 20 consecutive frames which is roughly the same number of frames as one gait cycle.

3 Results

In our experiments, we only use the walking sequences recorded under normal walking conditions and disregard clothing and carrying variations. We partition the data according to [11] in training set and test set and perform six-fold cross-validation.

We use a custom shallow convolutional neural network and SGD as optimizer, similar to DPEI. In Table 1 we compare our approach with DPEI [11]. While DPEI can only accurately recognize persons that were seen under the same or a similar angle before (during training), our approach is not bound by this limitation, since we can transform a person under any viewing angle into any other viewing angle. Moreover, only a single frame of a person is required, even though more accurate 3D representations can be generated when multiple frames from different views are used [9].

Table 1. Comparison of Top-1 and Top-5 accuracies of our approach and DPEI.

		Top-1 Accuracy		**Top-5 Accuracy**	
Train walks	**Test walks**	**DPEI**	**pifuhdEI**	**DPEI**	**pifuhdEI**
01, 02, 03, 04	05, 06	**0.933**	0.847	0.969	**0.976**
02, 03, 04, 05	06, 01	**0.952**	0.931	0.969	**0.993**
03, 04, 05, 06	01, 02	0.955	**0.961**	0.982	**0.999**
04, 05, 06, 01	02, 03	**0.959**	0.942	0.986	0.996
05, 06, 01, 02	03, 04	0.953	**0.959**	0.994	**0.998**
06, 01, 02, 03	04, 05	**0.957**	0.884	**0.990**	0.981
Average		**0.951**	0.921	0.982	**0.991**

The Top-1 accuracy is on average slightly worse than that of DPEI and the average TOP-5 accuracy is slightly higher, despite the low resolution input images, anomalies introduced by PIFuHD and the low amount of training data in general (less than 60% of what was used in [11] due to stricter preprocessing methods).

Even though the quantitative results do not show an big increase in accuracy, we believe this approach in general bears much potential. Especially with the help of higher resolution cameras, we prognosticate more accurate 3D models and therefore higher accuracy in gait recognition.

4 Future Work

During this work, we already identified some starting points to improve this approach. Using gait data from higher resolution cameras is one improvement. A more accurate background subtraction algorithm might be useful, unless higher resolution data already suffices for accurately generating the 3D model. Adding gait data from cameras positioned at different angles improves the 3D model quality as already demonstrated in [9]. Viewing-angle estimation can be implemented to make this approach more applicable for real-world scenarios. Instead of computing hand-crafted features on which to train the neural net, either unprocessed 2D projections or compact 3D representations can be used as input to the neural net. More fine-tuned models as well as data-augmentation can also help in improving this approach.

5 Related Work

Various methods exist for acquiring 3D visual gait data. For instance, a setup of multiple calibrated cameras can be used [14, 8, 6]. Apart from multi-camera methods, there also exist RGBD-based single camera methods [5, 2] and LiDAR-based methods [12]. Most recently machine learning methods have been used to generate 3D gait information, most notably PIFu and PIFuHD. A completely different approach is described in [7]. They morph 3D parametric human body models based on 2D silhouettes.

References

1. Ibm image resolution enhancer. https://github.com/IBM/MAX-Image-Resolution-Enhancer, accessed: 2021-07-26
2. Ahmed, F., Paul, P.P., Gavrilova, M.L.: DTW-based kernel and rank-level fusion for 3D gait recognition using Kinect. Visual Computer **31**(6-8), 915–924 (2015)
3. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense Human Pose Estimation in the Wild. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 7297–7306 (2018)
4. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316–322 (2006)
5. Kondragunta, J., Jaiswal, A., Hirtz, G.: Estimation of gait parameters from 3d pose for elderly care. ACM International Conference Proceeding Series pp. 66–72 (2019)

6. López-Fernández, D., Madrid-Cuevas, F.J., Carmona-Poyato, A., Muñoz-Salinas, R., Medina-Carnicer, R.: A new approach for multi-view gait recognition on unconstrained paths. Journal of Visual Communication and Image Representation **38**, 396–406 (2016)
7. Luo, J., Tjahjadi, T.: Gait recognition and understanding based on hierarchical temporal memory using 3D gait semantic folding. Sensors (Switzerland) **20**(6) (2020)
8. Muramatsu, D., Shiraishi, A., Makihara, Y., Uddin, M.Z., Yagi, Y.: Gait-based person recognition using arbitrary view transformation model. IEEE Transactions on Image Processing **24**(1), 140–154 (2015)
9. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. Proceedings of the IEEE International Conference on Computer Vision pp. 2304–2314 (2019)
10. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 81–90 (2020)
11. Schwarz, P., Scharinger, J., Hofer, P.: Gait recognition with densepose energy images. In: International Conference on Systems, Signals and Image Processing. pp. 65–70. Springer (2021)
12. Yamada, H., Ahn, J., Mozos, O.M., Iwashita, Y., Kurazume, R.: Gait-based person identification using 3D LiDAR and long short-term memory deep networks. Advanced Robotics pp. 1201–1211 (2020)
13. Yu, S., Tan, D., Tan, T.: 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)
14. Zhao, G., Liu, G., Li, H., Pietikäinen, M.: 3D gait recognition using multiple cameras. FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition **2006**, 529–534 (2006)